Relatório do Software Anti-plágio CopySpider
Para mais detalhes sobre o CopySpider, acesse: https://copyspider.com.br

Instruções
Este relatório apresenta na próxima página uma tabela na qual cada linha associa o conteúdo do arquivo de entrada com um documento encontrado na internet (para "Busca em arquivos da internet") ou do arquivo de entrada com outro arquivo em seu computador (para "Pesquisa em arquivos locais"). A quantidade de termos comuns representa um fator utilizado no cálculo de Similaridade dos arquivos sendo comparados. Quanto maior a quantidade de termos comuns, maior a similaridade entre os arquivos. É importante destacar que o limite de 3% representa uma estatística de semelhança e não um "índice de plágio". Por exemplo, documentos que citam de forma direta (transcrição) outros documentos, podem ter uma similaridade maior do que 3% e ainda assim não podem ser caracterizados como plágio. Há sempre a necessidade do avaliador fazer uma análise para decidir se as semelhanças encontradas caracterizam ou não o problema de plágio ou mesmo de erro de formatação ou adequação às normas de referências bibliográficas. Para cada par de arquivos, apresenta-se uma comparação dos termos semelhantes, os quais aparecem em vermelho.
Veja também:
Analisando o resultado do CopySpider
Qual o percentual aceitável para ser considerado plágio?

Relatório gerado por: caiomota802@gmail.com

Arquivos	Termos comuns	Similaridade
monografia-V7.docx.pdf X https://pubmed.ncbi.nlm.nih.gov/32802921	85	0,78
monografia-V7.docx.pdf X http://repositorio.unicamp.br/jspui/handle/REPOSIP/359544	39	0,38
monografia-V7.docx.pdf X https://www.marinha.mil.br/tm/?q=biblioteca_trabalhos_academicos	39	0,32
monografia-V7.docx.pdf X https://minerandodados.com.br/cross-validation-com-python	34	0,32
monografia-V7.docx.pdf X https://vitorborbarodrigues.medium.com/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-recall-quais-as-diferen%C3%A7as-c8f05e0a513c	33	0,31
monografia-V7.docx.pdf X https://medium.com/data-hackers/crossvalidation-de-maneira-did%C3%A1tica-79c9b080a6ec	24	0,22
monografia-V7.docx.pdf X http://www.cpdm.ufpr.br/testediaginostico.html	20	0,19
monografia-V7.docx.pdf X https://www.lume.ufrgs.br/bitstream/handle/10183/46711/Poster_11760.pdf?sequence=2	14	0,13
monografia-V7.docx.pdf X https://studylibpt.com/doc/4536178/hist%C3%B3ria---figure-b	2	0,02
monografia-V7.docx.pdf X https://www.youtube.com/watch?v=Jv6wdM7HXOQ	0	0,00

==========

Arquivo 1: monografia-V7.docx.pdf (9728 termos)
Arquivo 2: https://pubmed.ncbi.nlm.nih.gov/32802921 (1242 termos)
Termos comuns: 85
Similaridade: 0,78%
O texto abaixo é o conteúdo do documento monografia-V7.docx.pdf (9728 termos)
Os termos em vermelho foram encontrados no documento
https://pubmed.ncbi.nlm.nih.gov/32802921 (1242 termos)

==========

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO
CÂMPUS CAMPINAS
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
CAMPINAS
2021

CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
Trabalho de Conclusão de Curso apresentado como
exigência parcial para obtenção do diploma do
Curso de Tecnologia em Análise e
Desenvolvimento de Sistemas do Instituto Federal
de Educação, Ciência e Tecnologia Câmpus
Campinas.
Orientador: Prof. Me. Everton Josué da Silva.
CAMPINAS
2021

Dados Internacionais de Catalogação na Publicação

CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
Trabalho de Conclusão de Curso apresentado
como exigência parcial para obtenção do diploma
do Curso de Tecnologia em Análise e
Desenvolvimento de Sistemas do Instituto
Federal de Educação, Ciência e Tecnologia de
São Paulo Câmpus Campinas.
Aprovado pela banca examinadora em: _____ de _____ de _____.
BANCA EXAMINADORA

_____
Prof. Me. Everton Josué da Silva (orientador)

IFSP Câmpus Campinas

_____

Prof. Me. Carlos Eduardo Beluzo
IFSP Câmpus Campinas

_____

Prof. Dr. Ricardo Barz Sovat
IFSP Câmpus Campinas

Dedico este trabalho aos meus familiares,
colegas de classe, professores e servidores do Instituto
que colaboraram em minha jornada formativa.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pelo dom da vida e pela oportunidade de concluir mais uma etapa de minha experiência acadêmica.

Agradeço a todos os professores e servidores do IFSP Câmpus Campinas, que contribuíram direta e indiretamente para a conclusão deste trabalho.

Agradeço também à minha família, que deu todo o apoio necessário para que eu chegasse até aqui.

Agradeço ao meu orientador que me auxiliou a solucionar as dificuldades encontradas no caminho.

"Minha energia é o desafio, minha motivação é o impossível,
e é por isso que eu preciso
ser, à força e a esmo, inabalável.".
Augusto Branco

RESUMO

A mortalidade infantil pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível em uma população. O aprendizado de máquina é uma área da inteligência artificial que propõe métodos de análise de dados que automatizam a construção de modelos analíticos com base em reconhecimento de padrões, baseado no conceito de que sistemas podem aprender com dados, identificando padrões e tomando decisões com o mínimo de intervenção humana. O objetivo deste trabalho é testar e analisar dois tipos diferentes de algoritmos de aprendizado de máquina, utilizando a base de dados do SIM e SINASC do Brasil, do período de 2016 até 2018, para gerar modelos para predição de mortalidade neonatal. Os métodos utilizados foram Árvore de Decisão e de Regressão Logística e como métricas para avaliar os modelos resultantes foram utilizadas a AUC, Curva ROC e a Matriz de Confusão. Como resultado, apesar de terem alcançado valores de AUC 0.93, ambos modelos de predições acertaram muitas predições de vivos cerca de 671.000 e erraram muitas predições de mortos cerca de 2.600, principalmente devido ao fato de a base estar desbalanceada.

Palavras-chave: Mortalidade Neonatal, Predição, Aprendizado de Máquina.

ABSTRACT

Infant mortality can be used to analyze poverty and socioeconomic levels, as well as measure the quality of health and medical technology available in a population. Machine learning is an area of artificial intelligence that proposes data analysis methods that automate the

construction of analytical models based on pattern recognition, based on the concept that systems can learn from data, identifying patterns and making decisions with a minimum of human intervention. The objective of this work is to test and analyze two different types of machine learning algorithms, using the SIM and SINASC do Brasil database, from 2016 to 2018, to generate models for predicting neonatal mortality. The methods used were Decision Tree and Logistic Regression and as metrics to evaluate the resulting models, AUC, ROC Curve and Confusion Matrix were used. As a result, despite achieving values of AUC 0.93, both prediction models correct many live birth predictions around 671.000 and miss many stillbirth predictions around 2600, mainly due to the fact that the base is unbalanced.
Keywords: Neonatal Mortality, Prediction, Machine Learning.
LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS
SIM Sistema de Informação sobre Mortalidade
SINASC Sistema de Informações sobre Nascidos Vivos
TMI Taxa de Mortalidade Infantil
TMN Taxa de Mortalidade Neonatal
MN Mortalidade Neonatal
ML Machine Learning
DNV Declaração de Nascido Vivo
VN Verdadeiro Negativo
FN Falso Negativo
VP Verdadeiro Positivo
FP Falso Positivo
ROC Curva Característica de Operação do Receptor
NaN Não é um Número
RFE Eliminação Recursiva de Feature
RFECV Eliminação Recursiva de Feature com Validação Cruzada
SUMÁRIO

13

1 INTRODUÇÃO

A taxa de mortalidade infantil (TMI) é uma importante medida de saúde em uma população e é um grande problema no mundo todo. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível. Esta taxa é dividida em duas categorias: neonatal e pós-neonatal. É categorizada neonatal quando o óbito ocorre nos 28 primeiros dias de vida após o pós-parto, e o pós-neonatal é quando o óbito ocorre entre 29 e 364 dias de vida (BELUZO et al., 2020). Cerca de 46% das mortes no mundo acontecem entre os menores de cinco anos e grande parte está concentrada nos primeiros dias de vida (LANSKY, 2014). A diminuição dessas taxas é muito importante no mundo todo, refletindo nos indicadores de saúde pública e desenvolvimento do país.

Os fatores associados à mortalidade são profundamente influenciados pelas características biológicas maternas e neonatais, pelas condições sociais e pelos cuidados prestados pelos serviços de saúde (E. FRANÇA; S. LANSKY, 2009; R.M.D. NASCIMENTO, 2012). Em grande parte, o diagnóstico é altamente dependente da experiência adquirida pelo

profissional que o realiza. Apesar de indiscutivelmente necessário, não é perfeito, principalmente pelo fato de ser dependente de fatores humanos, por este motivo os profissionais sempre tiveram recursos tecnológicos para auxiliar nessa tarefa (BELUZO et al., 2020). Segundo estudos de 2010 no Brasil, calcula-se que aproximadamente 70% dos óbitos infantis ocorridos poderiam ter sido evitados por ações de saúde e que 60% dos óbitos neonatais ocorreram por situações que poderiam ser evitadas (BARRETO; SOUZA; CHAPMAN, 2015).

No presente trabalho foram utilizados dois algoritmos de aprendizado de máquina para criar modelos de predição de mortalidade neonatal. Além disso, foi realizada também uma análise exploratória da base de dados. Para este trabalho foram utilizadas duas fontes de dados: Sistema de Informação sobre Mortalidade (SIM) e Sistema de Informações sobre Nascidos Vivos (SINASC) do Brasil inteiro.

14

2 JUSTIFICATIVA

Com o avanço da tecnologia, podemos notar que ela está cada vez mais presente no nosso dia a dia e nos auxilia em diversas tarefas do cotidiano, e vem sendo aplicada em diversas áreas, dentre elas a saúde.

A mortalidade infantil é uma preocupação mundial na saúde pública, a ONU (Organizações das Nações Unidas) definiu a redução da mortalidade infantil como uma meta para o desenvolvimento global. A mortalidade neonatal é responsável por aproximadamente 60% da TMI nos países em desenvolvimento (A.K. SINGHA, 2016). Existem diversos fatores que podem contribuir com a redução de óbitos neonatais, como por exemplo acompanhamento médico à gestante no período de gravidez, acompanhamento médico ao recém-nascido nos primeiros dias de vida e a disponibilização deste serviço com qualidade e proficiência (WHO, 2009).

A diminuição dessa taxa é importante e de interesse para o mundo todo, e utilizar da tecnologia para auxiliar nessa diminuição é uma proposta funcional, e inovadora para a realidade brasileira (MOTA et al., 2019). Havendo disponibilidade de tecnologia para auxiliar nas decisões dos diagnósticos, os profissionais da saúde podem focar nos cuidados do recém-nascido com risco de vir a óbito, fornecendo tratamentos melhores.

Segundo o site Multiedro (2019), um grande problema que os médicos enfrentam ao realizar o diagnóstico, é analisar um grande volume de dados. Para a área médica obter diagnósticos rápidos e precisos pode salvar vidas, e a utilização de machine learning para realizar esses processos é importante, com a ML os dados podem ser processados em questões de segundos e resultar em um diagnóstico mais rápido, além disso utilizando bons modelos ML os diagnósticos serão mais precisos.

Diversos trabalhos vêm sendo desenvolvidos neste contexto. No trabalho realizado por BELUZO et al. (2020a), os autores utilizaram uma base de dados com informações da cidade de São Paulo para criação de modelos de aprendizado de máquina para predição de mortalidade neonatal. Este recorte foi utilizado também no presente trabalho para fins de exercício e definição de etapas da implementação de código. Na conclusão os autores discutem os fatores que tiveram mais influência nas predições, e na análise feita pelos autores, as consultas de pré-natal são muito importantes para a redução da mortalidade infantil, uma vez que algumas características muito importantes, como a malformação congênita, podem

ser detectadas ao longo das consultas de pré-natal.
Em um outro estudo realizado por BELUZO et al. (2020a), foi realizada uma análise exploratória da base de dados SPNeoDeath, onde e podemos observar detalhadamente cada
15
análise das colunas da tabela e diversos gráficos feito para um melhor entendimento dos dados.
Outro trabalho que foi realizado utilizando a base de dados com dados somente de São Paulo foi o de PELISSARI (2021), nesse trabalho é realizado uma análise exploratória na base de dados de São Paulo e é também analisado três algoritmos de aprendizado de máquina (Logistic Regression, Random Forest Classifier e XGBoost), nesse trabalho a autora conclui que os modelos de Logistic Regression e XGBoost foram os modelos que apresentaram os melhores desempenhos preditivos, e também mostra os problemas de estar utilizando uma base de dados desbalanceada.
O problema da mortalidade neonatal não é recente, e o mundo todo desenvolve iniciativas para lidar com ele. A ONU definiu como meta a redução da mortalidade infantil para o desenvolvimento global. Diversos fatores podem ajudar na diminuição da taxa de mortalidade neonatal como acompanhamento médico antes e após parto, tanto com a mãe quanto com o recém-nascido. Um serviço de qualidade e constante para os casos com maior risco de óbito pode salvar diversos recém-nascidos. Neste sentido, o desenvolvimento de ferramentas para a predição de mortalidade neonatal pode colaborar com esta iniciativa.

3 OBJETIVOS

3.1 OBJETIVO GERAL

O presente trabalho tem como objetivo testar e analisar os resultados da aplicação dos aprendizagem de máquina supervisionado Árvore de Decisão e Regressão Logística, utilizando dados do SIM e do SINASC, no período de 2016 até 2018, a fim de gerar modelos de predição de risco morte neonatal.

3.2 OBJETIVOS ESPECÍFICOS

? Realizar análise exploratória da base dados a ser utilizada na construção dos modelos;
? Implementar modelos de predição utilizando algoritmos Árvore de Decisão e Regressão Logística;
? Avaliar resultados e propor novas abordagens.
16

4 FUNDAMENTAÇÃO TEÓRICA

4.1 MORTALIDADE INFANTIL E NEONATAL

Como dito na monografia PELISSARI (2021), a TMI consiste no número de crianças que foram a óbito antes de concluir um ano de vida dividido por 1000 crianças nascidas vivas no tempo de um ano. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos e qualidade da saúde pública de um país (apud BELUZO et al., 2020).
Mencionado em PELISSARI (2021), a TMN é uma das duas categorias da TMI, é categorizado mortalidade neonatal quando o óbito ocorre do nascimento da criança até os 28 primeiros dias de vida. A mortalidade neonatal pode ser subdividida em mortalidade neonatal precoce que é quando o óbito ocorre entre 0 e 6 dias de vida, e o neonatal tardio quando o óbito ocorre entre 7 e 28 dias de vida (apud E. FRANÇA e S. LANSKY, 2009).
Segundo Lansky et al. (2014), a taxa de mortalidade neonatal é o principal

componente da TMI desde a década de 1990 no Brasil e vem se mantendo em níveis elevados, com taxa de 11,2 óbitos por mil nascidos vivos em 2010 (apud Maranhão et al., 2012). Também é dito em Lansky et al. (2014), que a TMI do Brasil em 2011 foi 15,3 por mil nascidos vivos, alcançando a meta 4 dos objetivos de desenvolvimento do milênio, compromisso dos governos integrantes das Nações Unidas de melhorar a saúde infantil e reduzir em 2/3 a mortalidade infantil entre 1990 e 2015. (apud Maranhão et al., 2012; apud Murray et al., 2015). Segundo Lansky et al. (2014) o principal componente da TMI em 2009 é a mortalidade neonatal precoce, e grande parte das mortes infantis acontece nas primeiras 24 horas (25%), indicando uma relação estreita com a atenção ao parto e nascimento (apud E. FRANÇA e S. LANSKY, 2009).

4.2 MÉTODOS DE APRENDIZADO DE MÁQUINA

A utilização de dados para a resolução de problemas, de modo a se identificar padrões ocultos e posteriormente auxiliar na tomada de decisões é definida como aprendizado de máquina (BRESAN, 2018 apud Brink et al. 2013).

O uso de técnicas de Aprendizado de Máquina se mostra altamente eficiente na resolução de tarefas que se apresentem difíceis de se resolver a partir de programas escritos por humanos (BRESAN, 2018 apud GOODFELLOW et al., 2016), bem como também possibilita uma maior compreensão sobre os funcionamentos dos princípios que compõem a inteligência humana.

17

Neste trabalho serão implementados modelos utilizando os algoritmos de Árvore de Decisão e Regressão Logística, os quais serão apresentados a seguir.

4.2.1 Árvore de Decisão

Segundo Batista e Filho (2019), os modelos preditivos baseados em árvores são geralmente utilizados para tarefas de classificação, embora também possam ser utilizados para tarefas de regressão. Considerado um dos mais populares algoritmos de predição, o algoritmo de árvore de decisão proporciona uma grande facilidade de interpretação.

As árvores de decisão utilizam a estratégia ?dividir-e-conquistar", na qual as árvores são construídas utilizando-se apenas alguns atributos. A árvores de decisão é uma das técnicas por meio da qual um problema complexo é decomposto em subproblemas mais simples. Recursivamente, a mesma estratégia é aplicada a cada subproblema (SILVA et al, 2008).

A árvore de decisões tem uma estrutura hierárquica onde cada nó da árvore representa uma decisão em uma das colunas do conjunto de dados, cada ramo da árvore representa uma tomada de decisão (BATISTA; FILHO, 2019).

O algoritmo segue algumas decisões para montar a árvore, o atributo mais importante é utilizado como nó raiz e será o primeiro nó, já os atributos menos importantes são mostrados nos ramos da árvore. Cada atributo é verificado para conferir se é possível gerar uma árvore menor e se é possível fazer uma classificação melhor, e assim é escolhido o nó que produz nós filhos (SILVA et al, 2008).

Existem diferentes tipos de algoritmo para a construção da árvore, como por exemplo, ID3 (Iterative Dichotomiser), C4.5 e CART (Classification and Regression Tree). O algoritmo ID3 escolhe os nós da árvore por meio da métrica do ganho de informação. Já o CART faz uso da equação de Gini (BATISTA; FILHO, 2019 apud KUHN; JOHNSON, 2013).

O algoritmo utilizado neste trabalho usará o cálculo da entropia e ganho de

informação, para decidir qual atributo será usado no nó, a entropia é uma medida da desorganização dos sistemas, maior é a incerteza do sistema, quanto maior a entropia maior está a desorganização, com a árvore de decisão a ideia é ir organizando o sistema e ir separando as decisões da melhor forma para que a entropia diminui e o ganho de informação aumente, para que a decisão do modelo fiquei mais assertiva. O cálculo da entropia utiliza a seguinte equação (ÁRVORE, 2020):

???????? =

?

?? ?? ???

2

??

18

Nesta equação o i representa possíveis classificações (rótulos), e o P é a probabilidade de cada uma. A ideia da árvore então é verificar qual é a entropia para cada uma das variáveis da base de dados, e verificar qual é a melhor organização de cada uma das variáveis. E isso é realizado através da técnica chamada de ganho de informação, essa técnica utiliza a equação a seguir (ÁRVORE, 2020):

????? = ????????(???) ? ? ????(??????) * ????????(??????)

Então o ganho de informação é calculado pela entropia do nó pai menos a soma do peso vezes a entropia dos nós filhos. Esse cálculo é realizado para todas as variáveis e a variável que tiver maior ganho de informação é utilizado para a decisão do nó. Esse processo é realizado recursivamente e a árvore vai sendo montada com base nesses cálculos (ÁRVORE, 2020).

Na Figura 1 temos a representação do início de uma árvore de decisão, no caso em questão o autor estava classificando exames positivos e negativos para o vírus SARS-CoV-2, ele utilizou uma base de dados que contém quase 6 mil exames realizados, contendo campos como o resultado do exame, idade do paciente, níveis de hemoglobina entre outras informações.

Figura 1. Exemplo de Árvore de Decisão. (Fonte: ÁRVORE 2020).

Pode-se observar que para o nó raiz foi utilizado a variável Leukocytes, logo foi o que apresentou a melhor organização para ser o nó raiz, e após ele temos os ramos da árvore. Cada retângulo da árvore é uma decisão do modelo e esses retângulos têm atributos importante como, a coluna da base de dados que será onde vai ser realizado a decisão, então pegando o nó raiz, pacientes que tiverem com os Leukocytes para menos que -0.43 seguiram o caminho

19

para a esquerda (true), já os pacientes que tiverem mais seguiram o caminho da direita (false). A entropia calculada do nó também é apresentada e a classificação do nó também, então se o resultado parar neste nó, a classificação que está nele será a classificação final, e temos o samples que é o número de amostras que se enquadram no na decisão. E esse processo de verificação do modelo vai se repetindo até não ter mais como dividir em subproblemas ou até chegar na profundidade máxima.

Uma característica importante da árvore de decisão é que ela é um modelo WhiteBox, ou seja, é possível visualizar a árvore montada e as decisões que o modelo terá que tomar na árvore. Na Figura 1 pode-se ver uma árvore de decisões montada, cada retângulo da árvore é

uma decisão que o modelo terá que tomar, conforme o modelo toma as decisões ele vai seguindo um caminho até chegar ao nó folha, nó folha é o nó que não contém nó filho, não tem mais ramos para baixo, chegando ao nó folha o modelo tem a classificação final.

4.2.2 Regressão Logística

O modelo de regressão logística estabelece uma relação entre a probabilidade de ocorrência da variável de interesse e as variáveis de entrado do modelo, sendo utilizada para tarefas de classificação, em que a variável de interesse é categórica, neste caso, a variável de interesse assume valores 0 ou 1, onde 1 é geralmente utilizado para indicar a ocorrência do evento de interesse (Batista e Filho, 2019).

De acordo com Mesquita (2014), a técnica de regressão logística, desenvolvida no século XIX, obteve maior visibilidade após 1950 ficando então mais conhecida. Ficou ainda mais difundida a partir dos trabalhos de Cox & Snell (1989) e Hosmer & Lemeshow (2000). Caracteriza-se por descrever a relação entre uma variável dependente qualitativa binária, associada a um conjunto de variáveis independentes qualitativas ou métricas.

Esta técnica inicialmente foi utilizada na área médica, porém a eficiência viabilizou sua implementação nas mais diversas áreas. Mesquita (2014) diz que o termo regressão logístico, tem sua origem na transformação usada com a variável dependente, que permite calcular diretamente a probabilidade da ocorrência do fenômeno em estudo.

Segundo Santos (2018), para estimar a probabilidade, é utilizado uma técnica chamada distribuição binomial para modelar a variável resposta do conjunto de treinamento, que tem como parâmetro a probabilidade de ocorrência de uma classe específica. Uma característica importante desse modelo é que a probabilidade estimada deve estar limitada ao intervalo de 0 a 1. O algoritmo de Regressão Logística gera um modelo de classificação binária (0 e 1). O modelo utiliza a Regressão Linear como base, a equação linear é (REGRESSÃO, 2020):

20

? = ? . ? + ?

O ?m? da equação é o coeficiente angular da reta, e o ?b? é o intercepto. Então, aplicando a equação linear obtemos um valor contínuo como resultado, após obter esse valor a Regressão Logística tem que realizar a classificação do resultado, então é aplicado uma função chamada sigmoide (REGRESSÃO, 2020):

? = 1

1+ ? ??

Essa função chamada de sigmoide, ou função logística, é responsável por ?achatar? o resultado da regressão linear, calculando a probabilidade de o resultado pertencer a classe 1, classificando assim o resultado da função linear em 0 ou 1.

Segundo Batista e Filho (2019), o modelo de regressão logística que tem como objetivo realizar uma classificação binária de uma variável de interesse irá predizer a probabilidade de pertencer à classe positiva. Tem-se, portanto, que é dado por:?

? = { 0, ?? ???? (? = 1 ??) < 0, 5

ou

? = { 1, ?? ???? (? = 1??) ? 0, 5

Então a decisão do modelo de regressão logística está relacionada à escolha de um ponto de corte para p(x). Caso o ponto de corte escolhido for p(x)=0,5, os resultados que forem p(x)>0,5 serão classificados como 1 e os resultados p(x)<5 serão classificados como 0

(SANTOS, 2018).
A Figura 2 ilustra um exemplo de como é o diagrama de Regressão Logística. Pelo diagrama então pode-se observar que o modelo possui uma representação gráfica em formato de ?S?, essa seria a curva logística. As predições do modelo sempre ficaram na linha 1 e 0 do eixo y, pois é o intervalo estabelecido pelo algoritmo, então a classificação sempre será nesse intervalo.
21
Figura 2. Exemplo de Diagrama de Regressão Linear. (Fonte: Batista e Filho 2019).
4.2.3 Treino, teste e validação do modelo de aprendizado de máquina
A criação de um modelo de aprendizado de máquina é realizada em duas etapas: treinamento do modelo, que é realizado a partir dos dados fornecidos para o algoritmo, e etapa de teste, onde os modelos recebem dados novos e sem o rótulo, para que seja avaliado a performance do modelo (PELISSARI, 2021).
Para o treinamento dos modelos foram utilizadas duas abordagens, a primeira onde a base foi particionada entre conjunto de dados para teste e para treino em uma porcentagem de 90% para treino e 10% para teste, porém dessa forma pode ocorrer problema de sobreajuste ou overfitting. Nesta situação, o que pode ocorrer é o modelo não aprender com os dados, e apenas ?decorar? os dados recebidos e quando é realizado o teste o modelo consegue predizer apenas situações parecidas, não sendo capaz de identificar padrões com diferenças mínimas.
Para evitar este problema, foi utilizado uma técnica de validação cruzada chamada K-folds cross-validation. Essa técnica de validação cruzada divide a base de dados em K subconjuntos, e então são realizadas várias rodadas de treinamento e teste alternando os subconjuntos utilizados para teste e treino, e o resultado do modelo baseia-se na média da acurácia observada em cada rodada (PELISSARI, 2021).
Pode-se observar na Tabela 1 uma ilustração da técnica K-fold cross-validation, onde temos um exemplo onde a base é dividida em 5 subconjuntos e cada subconjunto tem a parte de teste diferente dos outros subconjuntos, assim o modelo vai receber valores novos de teste e treino todas as vezes que executar um novo subconjunto.
22
Tabela 1. Exemplo da separação K-fold cross-validation. (Fonte: Adaptado de PELISSARI, 2021).

Predição 1	Predição 2	Predição 3	Predição 4	Predição 5
Teste	Treino	Treino	Treino	Treino
Treino	Teste	Treino	Treino	Treino
Treino	Treino	Teste	Treino	Treino
Treino	Treino	Treino	Teste	Treino
Treino	Treino	Treino	Treino	Teste

4.2.4 Métricas para avaliação de modelos
Para fins de avaliação, neste trabalho foram utilizadas três métricas para avaliar o desempenho dos modelos: a Matriz de Confusão e a AUC (Area under Receiver Operating Characteristic Curve - ROC). Na matriz de confusão pode-se observar o comportamento do modelo em relação a cada uma das classes a serem preditas pelo modelo, utilizando variáveis binárias (positivo: 1; e negativo: 0), para apresentar uma matriz que faz uma comparação entre as predições do modelo e os valores corretos do conjunto de dados (PELISSARI, 2021).

Na Tabela 2 temos um exemplo de uma matriz de confusão, a qual consiste em duas colunas e duas linhas, e os valores são atribuídos nos quadrantes com os seguintes resultados (PELISSARI, 2021):

? quando o valor real do conjunto de dados é 0, e a predição do modelo também classificou como 0, então temos um Verdadeiro Negativo (VN);

? quando o valor real é 1, e o modelo classificado como 0, temos um Falso Negativo (FN);

? quando o valor real é 0, e o modelo classificado como 1, então o resultado se enquadra no quadrante de Falso Positivo (FP);

? e por fim o quadrante Verdadeiro Positivo (VP), que são os resultados que tem o valor real 1, e o modelo classificado como 1.

23

Tabela 2. Modelo de Matriz de confusão. (Fonte: Adaptado de PELISSARI, 2021).

Valor Predito
Negativo (0) Positivo (1)
Classe
Real
Negativo
(0)
Verdadeiro
Negativo
Falso
Positivo
Positivo
(1)
Falso
Negativo
Verdadeiro
Positivo

A segunda métrica de avaliação utilizada foi a Curva ROC que permite avaliar o desempenho de um modelo com relação às predições efetuadas. A curva ROC é um gráfico de Sensibilidade (taxa de verdadeiros positivos) versus taxa de falsos positivos, a representação da curva ROC permite evidenciar os valores para os quais existe otimização da Sensibilidade em função da Especificidade (CABRAL, 2013).

Na Figura 3 temos um exemplo de uma curva ROC. A linha traçada na cor preta é uma linha de referência, ela representa hipoteticamente a curva ROC de um classificador puramente aleatório. Caso a linha laranja, que representa o resultado do modelo que está sendo avaliado, ficasse igual a linha tracejada preta, indicaria que o modelo avaliado estaria predizendo aleatoriamente os resultados.

Figura 3: Exemplo de Curva ROC. (Fonte: Elaboração Própria).

Por uma análise de inspeção visual, quanto mais próximo a linha laranja estiver do valor 1 no eixo Y, melhor está a capacidade do modelo para diferenciar as classes. O valor da

24

AUC representa a capacidade do modelo de diferenciar as classes, quanto maior o valor do

AUC, melhor é a predição realizada pelo modelo, uma vez que, os valores resultantes iguais a 0 são realmente 0, e os iguais a 1 são de fato 1 (PELISSARI, 2021).

5 METODOLOGIA

O desenvolvimento deste trabalho foi realizado em três etapas: (1) Pré-processamento e Análise Exploratória do conjunto de dados, onde foram aplicadas técnicas comuns para preparar o conjunto de dados como tratamento de valores nulos e seleção de variáveis de interesse; (2) Implementação de modelos de aprendizado de máquina supervisionado para predição das mortes neonatais, utilizando os algoritmos de árvore de decisão e regressão logística; e (3) Análise de Resultados, que será realizada uma análise do desempenho do modelo.

5.1 TECNOLOGIAS E FERRAMENTAS

Neste trabalho vamos utilizar a linguagem de programação Python (PYTHON, 2021) pela sua simplicidade de codificação e alto poder de processamento. Foi utilizado também a plataforma Google Colab Research (Colab, 2021), para o desenvolvimento dos algoritmos e treinamento dos modelos. Além disso, foi utilizado também o Jupyter Notebook (Jupyter, 2021), instalado localmente em computador pessoal, pois o Google Colab Research possui limitação de recursos de memória e processamento, então as duas plataformas foram usadas para realização deste trabalho. As principais bibliotecas utilizadas foram: numpy (NumPy, 2021), e pandas (Pandas, 2021), para a manusear os dados, matplotlib (Matplotlib, 2021), e seaborn (Seaborn, 2021), para a plotagem dos gráficos, e por fim a biblioteca sklearn (Scikit-Learn, 2021), que é a biblioteca responsável pelos algoritmos de aprendizado de máquina.

5.2 BASE DE DADOS

As bases de dados a serem utilizadas serão o SIM e SINASC, que são as duas principais fontes de informações sobre nascimentos e óbitos no Brasil. O SINASC é alimentado com base na Declaração de Nascido Vivo (Declaração de Nascido Vivo - DNV) (BELUZO et al.,2020b apud Oliveira et al., 2015). O SIM tem como objetivo principal apoiar a coleta, armazenamento e processo de gestão de registros de óbitos no Brasil (BELUZO et

25

al.,2020b apud Morais et al., 2017), e foi usado para rotular o óbito registrado no SIM, utilizando o campo DNV como chave de associação. O SIM será utilizado para rotular os registros do SINASC, pois o SIM coletas informações sobre mortalidade e é usado como base para o cálculo de estatísticas vitais, como a taxa de mortalidade neonatal e o SINASC reúne informações sobre dados demográficos e epidemiológicos do bebê, da mãe, do pré-natal e do parto (BELUZO et al.,2020b).

A Tabela 3 apresenta as variáveis da base de dados. Essa base tem variáveis que contêm valores quantitativos e categóricos. A coluna ?num_live_births? por exemplo contém o número de nascidos vivos anteriores da mãe e é composta por valores quantitativos contínuos, e a coluna ?tp_pregnancy? utiliza valores nominais categóricos que indicam o tipo de gravidez.

Tabela 3. Colunas da base de dados e suas descrições. (Fonte: Adaptado de BELUZO et al.,2020b).

Nome da coluna	Descrição	Domínio de dados
Variáveis	demográficas e socioeconômicas	

maternal_age Idade da mãe Quantitativo Contínuo (inteiro)
tp_maternal_race Raça / cor da pele da
mãe
Nominal categórico (inteiro)
1 - branco; 2 - Preto; 3 -
amarelo; 4 - Pele morena; 5 -
Indígena.
tp_marital_status Estado civil da mãe Nominal categórico (inteiro)
1 - Único; 2 - Casado; 3 - Viúva;
4 - Separados / divorciados
judicialmente; 5 - Casamento
por união estável; 9 - Ignorado.
tp_maternal_schooling Anos de escolaridade da
mãe
Nominal categórico (inteiro)
1 - nenhum; 2 - de 1 a 3 anos; 3 -
de 4 a 7 anos; 4 - de 8 a 11 anos;
5 - 12 e mais; 9 - Ignorado.
Variáveis obstétricas maternas
num_live_births Número de nascidos
vivos
Quantitativo Contínuo (inteiro)
num_fetal_losses Número de perdas fetais Quantitativo Contínuo (inteiro)
num_previous_gestations Número de gestações Quantitativo Contínuo (inteiro)
26
anteriores
num_normal_labors Número de partos
normais (trabalhos de
parto)
Quantitativo Contínuo (inteiro)
num_cesarean_labors Número de partos
cesáreos (partos)
Quantitativo Contínuo (inteiro)
tp_pregnancy Tipo de gravidez Nominal categórico (inteiro)
1 - Singleton; 2 - Twin; 3 -
Trigêmeo ou mais; 9 - Ignorado.
Variáveis relacionadas a cuidados anteriores
tp_labor Tipo de parto infantil
(tipo de parto)
Categórico Nominal (inteiro)
1 - Vaginal; 2 - Cesariana;
num_prenatal_appointments Número de consultas de
pré-natal
Quantitativo Contínuo (inteiro)

tp_robson_group Classificação do grupo
Robson
Ordinal categórico (inteiro)
Variáveis relacionadas ao recém-nascido
tp_newborn_presentation Tipo de apresentação de
recém-nascido
Categórico Nominal (inteiro)
1 - Cefálico; 2 - Pélvico ou
culatra; 3 - Transversal; 9 -
Ignorado.
has_congenital_malformation Presença de
malformação congênita
Nominal categórico (inteiro)
1 - Sim; 2 - Não; 9 - Ignorado
newborn_weight Peso ao nascer em
gramas
Quantitativo Contínuo (inteiro)
cd_apgar1 Pontuação de Apgar de 1
minuto
Ordinal categórico (inteiro)
cd_apgar5 Pontuação de Apgar de 5
minuto
Ordinal categórico (inteiro)
gestaional_week Semana de gestação Quantitativo Contínuo (inteiro)
tp_childbirth_care Assistência ao parto Categórico Nominal (inteiro)
1 - Doutor; 2 - enfermeira ou
obstetra; 3 - Parteira; 4 - outros;
9 - Ignorado.
was_cesarean_before_labor Foi cesáreo antes do
parto.
27
was_labor_induced Foi induzido ao trabalho
de parto.
is_neonatal_death Morte antes de 28 dias
(rótulo)
Nominal categórico (inteiro)
0 - sobrevivente; 1 - morto.

Neste trabalho foi utilizado dados do período de 2014 à 2016 devido sua melhor qualidade. Este recorte possui 6.760.222 registros dos quais 99.4% são amostras de recém-nascidos vivos e 0.6% são amostras onde os recém-nascidos vieram a óbito conforme pode ser observado na Figura 4. Com essas informações consegue-se observar que a base de dados é desbalanceada.

Figura 4. Distribuição da base de dados considerando se o recém-nascido viveu ou morreu durante o período neonatal. (Fonte: Elaboração Própria).

5.3 PREPARAÇÃO DA BASE DE DADOS

Nessa etapa foi realizada a preparação da base de dados para o treinamento dos modelos que consiste na limpeza, transformação e preparação dos dados a serem utilizados na análise exploratória e na criação dos modelos preditivos. A base de dados pode conter informações desnecessárias para o modelo que pode acabar atrapalhando as predições,
28
algumas colunas podem ser retiradas ou seus valores podem ser modificados, por exemplo valores reais são modificados para inteiros.

5.4 DESENVOLVIMENTO DOS MODELOS DE APRENDIZADO DE MÁQUINA

Como já dito anteriormente na seção 4 os algoritmos de aprendizado de máquina escolhidos para esse trabalho foram os: Regressão Logística e Árvore de Decisão que utilizam a lógica mostrada na seção 4 , Fundamentação Teórica. Esses dois foram escolhidos por alguns motivos, ambos são algoritmos de classificação e supervisionados, atendendo os critérios necessários para a predição de mortalidade neonatal, ambos os algoritmos são bons para realizar predições, a Árvore de Decisão inclusive é um ótimo modelo para analisar as decisões que estão sendo feitas pelo modelo, afinal esse algoritmo tem a característica de ser um WhiteBox, ou seja conseguimos ver o que o modelo aprendeu e o que ele está decidindo. Esses algoritmos selecionados são algoritmos que se encaixam na necessidade deste trabalho e são muito utilizados na comunidade acadêmica.

Para o algoritmo de Regressão Logística, como foi dito anteriormente na seção ?Preparação da base de dados? a base de dados foi tratada e particionada, no particionamento foi utilizado 90% da base para o treino e 10% para os testes e então foi realizado o treinamento do modelo, também foi aplicada a função RFECV (Eliminação Recursiva de Feature com Validação Cruzada), esse função executa a RFE (Eliminação Recursiva de Feature), que tem como ideia selecionar features considerando recursivamente conjuntos cada vez menores de features, isso ocorre da seguinte forma. Primeiro, o estimador é treinado no conjunto inicial de features e a importância de cada feature é obtida por meio de um atributo "coef" ou por meio de um atributo "feature_importances". Em seguida, as features menos importantes são removidas do conjunto atual de features. Esse procedimento é repetido recursivamente no conjunto removido até que o número desejado de features a serem selecionados seja finalmente alcançado. Porém o diferencial da RFECV é que essa função executa a RFE em um loop de validação cruzada para encontrar o número ideal ou o melhor número de features.

Após essa seleção, foi realizado o treinamento do modelo usando todas as features e também treinamos usando as features que o RFECV selecionou, para compararmos os resultados no final. Também foi utilizado o método de K-folds cross-validation em um novo treinamento para conferirmos se o modelo estava sofrendo overfitting (sobreajuste). E por fim aplicamos um método chamado GridSearchCV, esse método permite que seja realizado testes
29
alterando algumas combinação de parâmetros nos nossos modelos, facilitando para achar a melhor combinação, esse método é parecido com o cross validation, o diferencial do Grid Search é que ele faz os cross validation de vários modelos com hiperparâmetros diferentes de uma só vez.

Para o algoritmo de Árvore de Decisão também foi utilizado o mesmo

particionamento de 90% para treino e 10% para teste, na criação do modelo foi colocado a entropy como critério de decisão e uma profundidade máxima de 4, pois com números maiores o modelo ficava muito complexo e ele errava mais as predições, também utilizamos um algoritmo para mostrar a importância de cada Feature para o modelo.
O algoritmo de Regressão Logística e análise exploratória foi baseado em códigos disponível na plataforma web Kaggle (MNASSRI, 2020), no exemplo do Kaggle o autor faz os códigos para predizer mortes no navio Titanic, para o trabalho de mortalidade neonatal os códigos foram alterados para realizar predição de mortes neonatais. Já no algoritmo de Árvore de Decisão foi baseado em uma vídeo aula do professor Diogo Cortiz (ÁRVORE..., 2020), nessa vídeo aula o professor explica sobre os algoritmos de Árvore de Decisão e depois implementa o um exemplo de código, o código usado neste trabalho usou o codigo do professor Diogo Cortiz, e foi alterado para realizar predições de mortalidade neonatal.

5.5 ANÁLISE DOS RESULTADOS

Para a análise de resultados foi utilizado dois métodos que é o de Matriz de confusão e a curva ROC esses métodos são explicados na seção 4, a Fundamentação Teórica. Com esses métodos pode-se realizar uma avaliação do modelo, e assim será possível analisar os valores de acurácia, precisão e recall do modelo, essas métricas são utilizadas avaliar o desempenho do modelo, com a acurácia é possível calcular a proximidade entre o valor obtido na predição dos modelos e os valores esperados, com a precisão é possível analisar as amostras que o modelo classificou como 0 eram realmente 0 na classe real e com o recall é possível analisar as amostras que o modelo conseguiu reconhecer como a classe correta.

5.6 ACESSO A BASE DE DADOS E CÓDIGOS

A base de dados utilizada neste projeto pode ser encontrada no link: ... . E os códigos podem ser encontrados no link: ... .

30

6 RESULTADOS

Os mesmos experimentos mostrados abaixo foram aplicados para uma base de dados que contém somente dados de São Paulo, essa base é um recorte da base do Brasil todo em contém cerca de 1.400.000 amostras, esse recorte da base também é bem desbalanceado. Os resultados obtidos nesse experimento usando o recorte de São Paulo foram bem parecidos com os experimentos da base do Brasil todo, e os códigos desse experimento podem ser visualizados no apêndice X.

6.1 ANÁLISE EXPLORATÓRIA

O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

6.1.1 Apresentação dos Dados

Como dito anteriormente na seção 5.2 ?Base de Dados? deste trabalho, a base de dados é formada pelas informações do SIM e do SINASC contém 6.760.222 amostras e 29 colunas, porém serão utilizadas somente 23 colunas sendo elas as colunas descritas na Tabela 3. Na seção 5.2 também tem a Figura 4 que mostra que a base dados é desbalanceada tendo 99.4% das amostras de recém-nascidos vivos e 0.6% das amostras onde os recém-nascidos vieram a óbito no período neonatal.

6.1.1.1 Caracteristicas demográficas e socioeconômicas maternas

O apêndice X contém uma tabela que possui as estatísticas sumarizadas das variáveis

demográficas e socioeconômicas maternas. Nesta tabela por exemplo podemos observar a coluna ?maternal_age? que contém a idade da mãe, e na tabela mostra que a média da idade das mães é de 26.4 anos, e os dados possuem uma variação de no mínimo 8 anos e no máximo 55 anos, e a mediana é 26 anos. Na tabela também tem as colunas: ?tp_maternal_schooling? que mostra a categoria de anos de escolaridade da mãe, ?tp_marital_status? que mostra o estado civil da mãe em dados categóricos e a coluna ?tp_maternal_race? que mostra a raça da mãe também em dados categóricos.

31

6.1.1.2 Variávies obstétricas maternas

No apêndice X pode-se observar uma tabela que possui as estatísticas sumarizadas das variáveis obstétricas maternas. Nesta tabela por exemplo temos a coluna ?num_live_births?, essa coluna mostra o número de nascidos vivos anteriores que a mãe obteve, a média desta coluna é 0.94 , a mediana é 1, o número mínimo é 0 e o máximo é 10 nascidos vivos. Na tabela podemos observar também a coluna, ?num_fetal_losses?, esta coluna mostra o número de perdas fetais anteriores da mãe, a média desta coluna é 0.21, a medina é 0, o número mínimo é 0 e o máximo é 5, esses valores mostram que poucas mães tiveram perdas fetais antes da gravidez atual. Também na tabela tem estatísticas da coluna ?num_previous_gestations? esta coluna mostra o número de gestações anteriores da mãe, esta coluna tem a média de 1.14, a mediana é 1, o número mínimo é 0 e o máximo 10. Essa tabela também contém as colunas: ?num_normal_labors? que é o número de partos normais da mãe, ?num_cesarean_labor? que mostra o número de partos cesáreos da mãe e por fim mostra a coluna ?tp_pregnancy? que é o tipo da gravidez.

6.1.1.3 Variávies de histórico de gravidez

No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis do histórico de gravidez. Nesta tabela temos as colunas ?num_prenatal_appointments? que mostra o número de consultas de pré-natal, a média dessa coluna é 7.8, a mediana é 8, a variação dos dados é de no mínimo 0 e no máximo 40. A tabela também contém as colunas: ?tp_labor? que mostra o tipo de parto, e também contém a coluna ?tp_robson_group? que mostra a classificação do grupo Robson.

6.1.1.4 Variávies relacionadas ao recém-nascido

No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis relacionadas ao recém-nascido. Nesta tabela por exemplo podemos observar a coluna ?newborn_weight? que armazena o peso ao nascer dos recém-nascidos em gramas, a média dos dados dessa coluna é 3187.3, a mediana é 3215, a variação dos dados é de no mínimo 0 e no máximo 6000 gramas. Também podemos observar a coluna ?gestacional_week? que mostra o número de semanas de gestação, a média é 38.4, a mediana é 39, a variação dos dados é de no mínimo 15 e no máximo 45 semanas. Além dessas colunas a tabela também

32

contém as colunas: ?cd_apgar1? que mostra a pontuação de Apgar de 1 minuto, ?cd_apgar5? que mostra a pontuação de Apgar de 5 minuto, ?has_congenital_malformation? que mostra se o recém-nascido contém a presença de alguma malformação, ?tp_newborn_presentation? que mostra o tipo de apresentação de recém-nascido, ?tp_labor? que mostra o tipo de parto, ?was_cesarean_before_labor? que mostra se o recém-nascido foi cesáreo antes do parto, ?was_labor_induced ? que mostra se o recém-nascido foi induzido ao trabalho de parto,

?tp_childbirth_care? que mostra se ouve assistência ao parto, e por fim temos a coluna ?is_neonatal_death? que mostra se o recém-nascido veio a óbito ou não.

6.1.2 Análise das variáveis

Nesta seção serão mostrados a parte de análise de variáveis com diferentes tipos de gráficos.

6.1.2.1 Variáveis contínuas

Na Figura 5 temos dois gráficos boxplot, no boxplot à esquerda temos a distribuição da variável peso ao nascer e nele é possível observar duas caixas onde a caixa azul são os vivos e o laranja são os mortos. Nas duas caixas mostradas no gráfico podemos observar pequenos círculos nas pontas, esses círculos são outliers (dados fora da curva), então para a caixa de vivos os outliers são recém-nascidos com pesos acima de 4500 gramas ou abaixo de 2000 gramas, e para a caixa de mortos os outliers são os recém-nascidos com mais de 5500 gramas. Para os vivos temos a mediana de aproximadamente 3200 gramas e para os mortos a mediana é 1300 gramas. Cada caixa representa 50% da base de dados, a caixa azul de vivos mostra que o peso de recém-nascidos que sobreviveram está aproximadamente entre 2700 a 3600 gramas, e para a caixa laranja de mortos o peso de recém-nascidos que vieram a óbito estão aproximadamente entre 700 a 2500 gramas. Então o gráfico mostra para nós que os recém-nascidos que têm maior risco de vir a óbito tem pesos menores que 2000 gramas.

33

Figura 5. BoxPlot das features Peso ao nascer e Idade da mãe. (Fonte: Elaboração Própria).

No gráfico à direita da Figura 5 temos um boxplot da distribuição da variável idade da mãe, como foi dito anteriormente a caixa azul são os vivos e o laranja são os mortos. Nesse boxplot então podemos ver que os outliers da caixa de vivos são mães com idade acima de 46 anos aproximadamente, para os mortos os outliers são mães com idade acima de 50 anos. Para os vivos temos a mediana de aproximadamente 26 anos e para os mortos a mediana é de 26 anos. Nos vivos a idade das mães está aproximadamente entre 21 a 32 anos, e para os mortos a idade das mães está aproximadamente entre 20 a 34 anos. É importante ressaltar que o gráfico mostra as duas caixas bem parecidas, então de acordo com os dados não contém diferença significativa entre vivos e mortos usando a idade da mãe para distribuir.

Já na Figura 6 temos um boxplot da distribuição da variável semanas de gestação.

Nesse boxplot então podemos ver que os outliers da caixa de vivos são gestações com mais de 44 semanas aproximadamente, e gestações com menos de 35 semanas. Para os vivos temos a mediana de aproximadamente 39 semanas e para os mortos a mediana é de 32 semanas. Nos vivos as semanas de gestação estão aproximadamente entre 36 a 40 semanas, e para os mortos a semanas de gestação estão aproximadamente entre 26 a 36 semanas. Nesse boxplot os dados mostraram que para as gestações que têm menos de 36 semanas os recém-nascidos têm maior risco de vir a óbito.

34

Figura 6. BoxPlot da feature semana de gestação. (Fonte: Elaboração Própria).

Na Figura 7 temos um histograma da distribuição de peso ao nascer por classe, podemos observar no gráfico que grande parte dos vivos (linha azul) estão distribuídos entre 2000 e 4000 gramas, a área entre esses valores tem muitos dados de vivos, já entre os valores 0 e 2000 gramas temos uma concentração dos dados de mortos.

Figura 7. Histograma de distribuição do peso entre as classes. (Fonte: Elaboração Própria)

6.1.2.2 Variáveis categoricas

Observando a Figura 8 podemos ver um gráfico de barras da distribuição da variável escolaridade da mãe, este gráfico mostra a contagem de registros por escolaridade da mãe, esse gráfico mostra para nós que a maior parte das mães estão na categoria 4 que representa de 8 a 11 anos de escolaridade.

35

Figura 8. Gráfico de barras da distribuição da variável Escolaridade da mãe. (Fonte: Elaboração Própria).

Observando a Figura 9 podemos ver um gráfico de barras da distribuição da variável número de vivos, este gráfico mostra a contagem de registros por número de vivos, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho.

Figura 9. Gráfico de barras da distribuição da variável Número de nascidos vivos. (Fonte: Elaboração Própria).

Observando a Figura 10 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 1 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 1 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem

36

apgar de 8 ou 9, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.

Figura 10. Gráfico de barras da distribuição da variável Pontuação Apgar de 1 minuto. (Fonte: Elaboração Própria).

Observando a Figura 11 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 5 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 5 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem apgar de 9 ou 10, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.

Figura 11. Gráfico de barras da distribuição da variável Pontuação Apgar de 5 minutos. (Fonte: Elaboração Própria).

37

Observando a Figura 12 podemos ver um gráfico de barras da distribuição da variável presença de malformação congênita, este gráfico mostra a contagem de registros por presença de malformação congênita, esse gráfico mostra para nós que a grande maioria dos dados estão na categoria 2 que mostra que o recém-nascido não possui malformação, esse grande volume para a categoria 2 é causada pelo desbalanceamento da base.

Figura 12. Gráfico de barras da distribuição da variável Presença de malformação congênita. (Fonte: Elaboração Própria).

Observando a Figura 13 podemos ver um gráfico de barras da distribuição da variável número de gestações, este gráfico mostra a contagem de registros por número de gestações, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho, da mesma forma mostrada na Figura 9.

38

Figura 13. Gráfico de barras da distribuição da variável Número de gestações. (Fonte: Elaboração Própria).

Observando a Figura 14 podemos ver um gráfico de barras da distribuição da variável assistência ao parto, este gráfico mostra a contagem de registros por assistência ao parto, esse gráfico mostra para nós que a grande maioria dos dados mostram que o parto teve assistência de uma Enfermeira ou obstetra ou essa informação foi ignorada na hora do resgistro.
Figura 14. Gráfico de barras da distribuição da variável Assistência ao parto. (Fonte: Elaboração Própria).
6.1.2.3 Analise bi-variada
A Figura 15 mostra para nós a importância da feature peso ao nascer com o gráfico mostrado é possível observar claramente a diferença de pesos entre vivos e mortos, recém-nascidos com peso acima de 2000 gramas sobreviveram, já os recém-nascidos com menos de 2000 gramas vieram a óbito.
39
Figura 15. Gráfico de densidade de peso entre as classes. (Fonte: Elaboração Própria).
Na Figura 16 podemos observar um gráfico de correlação entre algumas features de valores contínuos, algumas dessas features possuem correlações positivas, como pontuação apgar de 1 minuto e pontuação apgar de 5 minutos, essas colunas possuem uma correlação de 0.7, então nascidos que recebem um valor alto de apgar de 1 minuto tendem a receber um valor alto no apgar de 5 minutos. O gráfico também mostra outras features com correlações positivas como, número de partos normais e número de gestações anteriores que contém uma correlação de 0.81, número de gestações anteriores com idade da mãe que tem uma correção de 0.39, número de partos de cesáreos com idade da mãe que possui correlação de 0.24, número de partos normais com idade da mãe com a correlação de 0.26, semanas de gestação com peso ao nascer em gramas com correção de 0.52, e as apgar de 1 e 5 minutos com semanas de gestação. E grande parte das features, possuem correlações baixas então elas são inversamente correlacionadas pelo que o gráfico mostra, como por exemplo número de partos normais com número de consultas pré-natais possuem uma correção de -0.17, e número de partos cesáreos com número de partos normais que contém uma correlação de -0.15. Então a correlação dessas features estão bem baixas tirando as correlações acima de 0.7.
40
Figura 16. Gráfico de correlação. (Fonte: Elaboração Própria)
6.1.2.4 Distribuição por raça
Na Figura 17 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por raça, e podemos observar que a raça 5 (indigena), é o boxplot que contem a menor variação de idade, já as raças 1 (branco) e 3 (amarelo) são as caixas que possuem maior variação com a média em torno de 20 e 30 anos.
41
Figura 17. Distribuição da idade da mãe por raça. (Fonte: Elaboração Própria).
No boxplot mostrado na Figura 18 podemos observar a distribuição de peso ao nascer por raça, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre as raças para peso ao nascer.
Figura 18. Distribuição de peso ao nascer por raça. (Fonte: Elaboração Própria).
Na Figura 19 temos o boxplot da a distribuição de semanas de gestação por raça, e praticamente todas as caixas são iguais, somente a caixa da raça 1 (branco) possui menor variação que as outras.

42
Figura 19. Distribuição de semanas de gestação por raça. (Fonte: Elaboração Própria).
6.1.2.5 Distribuição por estado civil
A Figura 20 mostra o boxplot da a distribuição de peso ao nascer por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre os estados civis para peso ao nascer.
Figura 20. Distribuição de peso ao nascer por estado civil. (Fonte: Elaboração Própria).
A Figura 21 mostra o boxplot da a distribuição de semanas de gestação por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Porém os estados civis 2 (Casado) e 4 (Separados/divorciados judicialmente), contém variações menores.
43
Figura 21. Distribuição de semanas de gestação por estado civil. (Fonte: Elaboração Própria).
Na Figura 22 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por estado civil, e podemos observar que o estado civil 3 (Viúva), possui a maior variação, já o estado civil 4 (Separados/divorciados judicialmente) é a caixa que possuem menor variação
Figura 22. Distribuição da idade da mãe por estado civil. (Fonte: Elaboração Própria).
6.1.2.6 Distribuição por escolaridade da mãe
Na Figura 23 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por escolaridade da mãe, e podemos observar que a escolaridade 2 (1 a 3 anos) e 3 (4 a 7
44
anos), possui as maiores variações, já escolaridade 5 (12 ou mais anos) é a caixa que possuem menor variação
Figura 23. Distribuição da idade da mãe por escolaridade da mãe. (Fonte: Elaboração Própria).
A Figura 24 mostra o boxplot da a distribuição de peso ao nascer por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então, pouca variação é observada entre a escolaridade da mãe para peso ao nascer.
Figura 24. Distribuição de peso ao nascer por escolaridade da mãe. (Fonte: Elaboração Própria).
A Figura 25 mostra o boxplot da a distribuição de semanas de gestação por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas,
45
todas são praticamente iguais. Porém a escolaridade 5 (12 ou mais anos) contém variação menor que as outras.
Figura 25. Distribuição de semanas de gestação por escolaridade da mãe. (Fonte: Elaboração Própria).
6.2 ÁRVORE DE DECISÃO
A Figura 26 mostra um trecho do algoritmo utilizado, nesse trecho é realizado a divisão do DataFrame das features de entrada , e o DataFrame que contém o rótulo.Neste caso o rótulo será o DataFrame ?target?, esse Dataframe contém a feature

?is_neonatal_death?, essa feature informa se o recém nascido veio a óbito ou não, sendo o foco da predição que desejamos realizar essa feature entra como rótulo para o modelo, já no Dataframe ?nome_features? temos as features que vão servir de entrada para o modelo, é a partir dessas features que o modelo tentará encontrar padrões para predizer o risco de óbito do recém nascido. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

46

Figura 26. Separação dos DataFrames de entrada e rótulo. (Fonte: Elaboração Própria).

6.2.1 Modelo de Árvore de Decisão Utilizando Particionamento 90/10

O algoritmo de árvore de decisão foi treinado utilizando duas formas diferentes: usando as partições 90% para treino e 10% para os testes e utilizando o método K-folds cross-validation. E para os experimentos iniciais foi utilizado o particionamento 90/10.

Tabela 4. Resultados do modelo de árvore de decisão. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	671807
Morto(1)	0.71	0.29	0.41	4216

Analisando a Tabela 4 pode-se ver os seguintes resultados, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 71% que realmente pertenciam a classe 1, então cerca de 29% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 29% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem a classe 1 que seria dos recém nascidos que poderiam vir a óbito, já para a classe de vivos a

47

classe 0 o modelo está identificando muito bem e classificando de forma correta. A tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.

Esse modelo teve uma acurácia de 0.99 que é uma acurácia bem alta, porém somente pela acurácia não é possível dizer que o modelo está excelente pelo motivo da base que está sendo usada é altamente desbalanceada, então para a classe 0 que é a de vivos temos muito mais dados que para a classe de mortos, e o modelo está acertando bastante a classe de vivos logo a acurácia fica alta por esses acertos, enquanto para a classe de mortos o modelo não está acertando tanto.

A Figura 27 mostra a curva ROC do modelo, pode-se observar que a AUC da curva ROC ficou em 0.92, a AUC mostra que o modelo está acertando bastante as predições, pois quanto mais perto de 1 melhor está no nosso modelo, mas no caso desse modelo vimos acima que as predições para a classe de mortos(1) está ruim, o modelo está acertando poucas classificações como 1. Então a AUC assim como a acurácia está alta porque a base de dados utilizada é altamente desbalanceada tendo muito mais amostras de vivos(1), e como o modelo está bom para essa classificação e está acertando bastante os valores de AUC e acurácia ficam altos .

Figura 27. Curva ROC modelo de Árvore de decisão. (Fonte: Elaboração Própria).
A Figura 28 mostra a importância de cada feature para o modelo sendo assim a feature 10 - Peso ao nascer, a mais importante para o modelo, pois dela o modelo conseguiu tirar mais informações, seguido das features, 13 - Apgar dos 5 primeiros minutos, 11 - Semanas de 48
Gestação, 14 - Presença de malformação 12 - Apgar do 1 primeiro minutos. Então essas são as features que foram mais decisivas para a montagem da árvore e para as predições do modelo.
Figura 28. Gráfico de importância das features. (Fonte: Elaboração Própria).
Na Figura 29 temos uma imagem reduzida da árvore de decisão que foi montada com o treinamento utilizando 5 como profundidade máxima para a árvore, a imagem da árvore completa pode ser visualizada no apêndice A. É possível ver que as features marcadas como importante para o modelo apareceu diversas vezes na árvore, e a feature peso ao nascer foi escolhida para ser o nó raiz da árvore, de todas as features ela foi a que teve a melhor entropia para ser o nó raiz, e depois aparece mais vezes para outras decisões em outros níveis da árvore e isso faz com que essa feature se torne a mais importante para o modelo, também podemos ver que a feature ?Apgar do 1 primeiro minuto? mesmo sendo a menos importante para o modelo ela aparece somente no nó de número 42 isso mostra que a entropia e o ganho de informação dessa feature nas outras decisões não foram boas fazendo ela ser a feature com menos importância utilizada no modelo.
49
Figura 29. Árvore de decisão montada a partir do algoritmo. (Fonte: Elaboração Própria).
6.2.2 Modelo de Árvore de Decisão Utilizando K-folds cross-validation
Após realizar esse experimento com a partição 90/10, foi realizado outro experimento com esse mesmo algoritmo porém agora utilizando o método K-folds cross-validation para o treinamento. O K-folds cross-validation foi configurado para separar a base de dados em 10 partes iguais, e assim o algoritmo foi treinado novamente gerando um novo modelo. O método de K-folds cross-validation é utilizado para conferir se o modelo não está sofrendo problemas de overfitting, é importante garantir que o modelo está aprendendo com os dados e não está somente decorando.
Na Tabela 5 temos então o resultados desse modelo, pode-se observar que os resultados para a classificação de vivos(0) não teve alteração, o modelo continua tendo 100% na precisão e no recall, mas na classificação de mortos(1) o modelo teve uma diminuição na precisão que agora está em 65% e também teve uma diminuição no recall que agora está em 23%, porém essa diminuição é justificada porque para o teste deste modelo foi utilizado a base completa e não somente a partição de teste que contém 10% da base total. Logo os resultados não tiveram uma alteração drástica e tiveram um comportamento bem semelhante, incluindo a acurácia que se manteve em 0.99 e a AUC que teve uma diminuição pequena também e ficou com 0.89.
50
Tabela 5. Resultados do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	6719191

Morto(1) 0.65 0.23 0.34 41031
Na Tabela 6 é possível ver a matriz de confusão do modelo que mostra o número exato de erros e acertos do modelo, pode-se observar que para a classe de vivos o modelo classificou corretamente 6.714.117 amostras da base de dados completa, o modelo classificou como 0 as amostras que nos rótulos realmente eram 0, errou apenas 5074 amostras. Já para a classe de mortos o modelo acertou somente 9552 amostras, classificou como 1 as amostras que no rótulo eram 1 também, e errou 31479. Isso mostra que o modelo está muito desbalanceado, pois está errado muito mais do que acertando as predições de mortos, sendo isso um reflexo da base de dados desbalanceada também. E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.
Tabela 6. Matriz de confusão do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6714117 5074
Mortos (1) 31479 9552
6.3 REGRESSÃO LOGÍSTICA
Para o algoritmo de Regressão Logística também foi utilizado métodos diferentes para o treinamento, O algoritmo foi treinado utilizando três formas diferentes: usando as partições 90% para treino e 10% para os testes, utilizando o método K-folds cross-validation e foi treinada utilizando o método RFECV que seleciona as melhores features para o modelo, juntamente com o método GridSearchCV. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.
51
6.3.1 Modelo de Regressão Logística Utilizando Particionamento 90/10
Para o primeiro experimento do algoritmo de regressão logística foi utilizado o particionamento 90/10 para o treinamento do algoritmo, sendo 90% da base de dados para realizar o treinamento do modelo e 10% da base para realizar os testes.
Na Tabela 7 temos os resultados do modelo de regressão logística utilizando o método de particionamento 90/10, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 65% que realmente pertenciam a classe 1, então cerca de 35% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 24% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem as amostras da classe 1 que seria dos recém-nascidos que poderiam vir a óbito, já para a classe de vivos a classe 0 o modelo está identificando muito bem e classificando de forma correta.E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.

Tabela 7. Resultados do modelo de regressão logística usando particionamento 90/10.
(Fonte: Elaboração Própria)
Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 671885
Morto(1) 0.65 0.24 0.35 4138
A Figura 30 abaixo mostra a curva ROC do modelo de regressão logística usando o particionamento 90/10, analisando a curva ROC podemos ver que a AUC da curva ficou em 0.93 e a acurácia do modelo ficou 0.99, mostrando que o modelo está acertando bastante as predições, porém esses números para o nosso modelo não mostra que ele está ótimo, com a análise da Tabela 7 acima podemos observar que o modelo está acertando muitas predições da classe de vivos e poucas predições da classe de mortos, como a base de dados é
52
desbalanceada, como a maior quantidade de dados está na classe de vivos e o modelo está acerto quase todos dessa classe a acurácia e a AUC ficam altas por esses acertos.Logo é preciso analisar mais resultados além da acurácia e da AUC do modelo.
Figura 30. Curva ROC modelo regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria).
Na Tabela 8 podemos analisar os número exatos que o modelo acertou de cada classe, como é mostrado na matriz de confusão o modelo acertou 671.435 da classe de vivos e acertou apenas 915 da classe de mortos, e errou mais que o triplo da classe de mortos. Então as predições do modelo estão muito desbalanceadas, para a classe de vivos o modelo está predizendo bem, porém para as classes de mortos o modelo está errando muitas predições.
Tabela 8. Matriz de confusão do modelo de regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 671435 504
Mortos (1) 3169 915
6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation
Para o segundo experimento do algoritmo de regressão logística foi utilizado o método K-folds cross-validation para o treinamento do modelo com a intenção de conferir e confirmar
53
que o modelo não esteja tendo problemas com overfitting e realmente esteja aprendendo com os dados que está sendo usado para o treinamento. O método K-folds cross-validation foi configurado para realizar uma separação de 10 partes iguais.
Na Tabela 9 podemos observar os resultados do modelo, e os resultados foram parecidos com o experimento anterior usando o particionamento 90/10, única diferença nos resultado é que no experimento anterior a precisão da classe de mortos ficou em 65% e no caso desse experimento usando o K-folds cross-validation a precisão ficou em 64%, mas os valores de recall e F1-Score se mantiveram, assim como todos os resultados da classe de mortos se mantiveram em 100%. A execução desse experimento mostra que o modelo realmente está aprendendo com os dados e não está apenas decorando, retirando o risco do modelo estar com overfitting. Mesmo sem alterações no resultado é importante saber que o

modelo não está com overfitting e está conseguindo aprender e encontrar padrões nos dados.
Tabela 9. Resultados do modelo de regressão logística usando K-folds cross-validation.
(Fonte: Elaboração Própria)
Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 6719191
Morto(1) 0.64 0.24 0.35 41031
Já na Figura 31 temos a curva ROC do modelo, essa curva ROC mostra a AUC de
todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do
método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas
para todas as vezes que foi realizado os testes, a AUC se manteve em 0.93 e 0.94, sendo a
média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não
se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua
acertando muito a classe de vivos que é a classe predominante na base de dados. Já era
esperado os resultados não sofrerem alterações grandes, pois a base de dados não sofreu
nenhuma alteração, esse experimento foi somente para conferir se a modelo não estava
sofrendo overfitting, e pelos resultados aparentemente a base não está com problemas de
sobreajuste.
54
Figura 31. Curva ROC modelo regressão logística usando K-folds cross-validation. (Fonte:
Elaboração Própria).
Na Tabela 10 temos a matriz de confusão do modelo, onde mostra que o modelo
acertou 6.713.658 amostras da classe de vivos e errou apenas 5.533, já para a classe de mortos
o modelo acertou somente 9.847 e novamente errou mais que o triplo errando 31.182
amostras. Como os resultados mostrados anteriormente não tiveram alteração era de se
esperar que as predições corretas e incorretas do modelo também não sofressem alterações, o
modelo continua acertando muito a classe de vivos e errando muito a classe de mortos,
mantendo as predições bem desbalanceadas.
55
Tabela 10. Matriz de confusão do modelo de regressão logística usando K-folds
cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6713658 5533
Mortos (1) 31182 9847
6.3.3 Modelo de Regressão Logística Utilizando GridSearchCV e RFECV
Para o terceiro e último experimento de regressão logística foi utilizado técnicas
diferentes juntas para tentar melhorar as predições do modelo, para esse experimento usamos
o K-folds cross-validation para o particionamento da base de dados assim como foi utilizado
no experimento acima, e também foi utilizado duas técnicas que ainda não foram usadas nos
experimentos anteriores que são as funções RFECV e GridSearchCV. A função RFECV
seleciona a quantidade ideal de features para ser usadas no treinamento do modelo e também
mostra quais são as melhores features para essa predição, então essa função ela utiliza a
validação cruzada para ir treinando e testando N vezes o modelo, e sempre vai alterando a

quantidade e as features utilizadas para o teste, então cada vez que a função testa os modelos ela grava a pontuação dos acertos e assim depois mostra o gráfico da Figura 32 e assim é possível ver quais são as features ideais para o modelo.
56
Figura 32. Resultado da função RFECV Base Brasil. (Fonte: Elaboração Própria).
Então os resultados mostrados na Figura 32 apresenta que o número ideal para treinar o modelo de features é 6, e as features sao: 'tp_maternal_schooling', 'gestaional_week', 'cd_apgar1', 'cd_apgar5', 'has_congenital_malformation', 'tp_labor'. Essas são as features que tiveram o melhor desempenho no RFECV, então para o treinamento do modelo desse experimento as features que vão ser utilizadas serão somente essas 6.
Após a execução da função de RFECV foi executado a função de GridSearchCV, o GridSearchCV é utilizado para descobrir quais são os melhores parâmetros para utilizar no algoritmo de regressão logística e criar os modelos, foi utilizado também a validação cruzada para testar qual seria os melhores parâmetros para melhorar o desempenho do modelo. Na Figura 33 pode-se ver que os parâmetros escolhidos pela função foram:
Figura 33. Resultado da função GridSearchCV. (Fonte: Elaboração Própria).
Na Tabela 11 podemos observar os resultados do modelo, e os resultados foram parecidos com os experimentos anteriores, para a classe de mortos a precisão teve um aumento e foi para 69% e o recall diminuiu chegando em 20%, então utilizando as novas
57
funções o modelo ganhou precisão e perdeu recall. Para a classe de vivos o modelo se manteve em 100% e não teve alterações para os resultados de precisão e recall. Mesmo com a utilização das funções de RFECV e GridSearchCV o modelo não teve uma melhora significativa para as predições de mortos.
Tabela 11. Resultados do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	6719191
Morto(1)	0.69	0.20	0.31	41031

Na Figura 34 temos a curva ROC do modelo, essa curva ROC mostra a AUC de todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas para todas as vezes que foi realizado os testes, a AUC se manteve em 0.92 e 0.93, sendo a média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua acertando muito a classe de vivos que é a classe predominante na base de dados.
58
Figura 34. Curva ROC modelo regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria).
Na Tabela 12 temos a matriz de confusão do modelo, onde mostra que o modelo acertou 6.715.535 amostras da classe de vivos e errou apenas 3.656, comparando com os experimentos anteriores de regressão logística as predições de vivos teve uma pequena piora e teve uma pequena diminuição nos acertos. Já para a classe de mortos, o modelo acertou 8.296 e errou 32.735 amostras, logo as predições de vivos tiveram uma uma piora e acertou um

pouco menos do que os outros experimentos, e na classe de vivos teve uma melhora acertando mais predições, porém como o problema está nas predições de mortos e para essa classe teve uma piora pode-se dizer que a função de GridSearchCV não teve um ganho significativo nos resultados. Então como modelo final foi escolhido o modelo utilizando somente o K-folds cross-validation, pois foi o modelo que mostrou o melhor desempenho dos experimentos de regressão logística.

59

Tabela 12. Matriz de confusão do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

Predito

Vivos (0) Mortos (1)

Classe Real

Vivos (0) 6715535 3656

Mortos (1) 32735 8296

60

7 CONCLUSÃO

O principal objetivo deste trabalho foi analisar os resultados das predições de mortalidade neonatal dos algoritmos de aprendizado de máquina usando uma base de dados com dados do Brasil inteiro. A partir destes resultados apresentados na secção anterior ?Resultados?, é possível concluir que analisar somente a AUC e a acurácia do modelo não é o suficiente para garantir que o modelo está excelente, sendo assim temos que ter mais atenção a precisão, recall e as predições mostradas na matriz de confusão.

Ambos os modelos ficaram desbalanceados nas predições, os modelos acertaram mais predições de vivos do que mortos, e as predições corretas de mortos foram mais baixas do que as incorretas, para os modelos ficarem melhor precisam passar por algoritmos que possam balancear essas predições, uma outra alternativa para melhorar o modelo é utilizar uma base de dados mais balanceada, pois essa base que foi utilizada é altamente desbalanceada.

Embora os dois modelos tenham tido resultados parecidos , o modelo de regressão logística utilizando somente o K-folds cross-validation se saiu melhor que os outros experimentos, o modelo criado na seção ?6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation? manteve a acurácia, AUC, precisão e recall dos demais modelos e acertou mais predições, talvez com algoritmos para melhorar as predições da classe de mortos, balanceando mais as predições, esse modelo fique com ótimas taxas preditivas.

Além disso podes-se afirmar que o peso ao nascer do recém nascido é um fator muito importante para as decisões dos modelos, ambos os modelos usaram muito essa informação para as predições, também podemos colocar, as colunas de presença de malformação e semana de gestação, como colunas que foram importantes para os modelos, essas causas podem ser evitadas se houver um acompanhamento médico com qualidade e eficiência.

61

REFERÊNCIAS

ÁRVORE de Decisão - Aula 3. Gravação de Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=ecYpXd4WREk&t=4089s. Acesso em: 1 jul. 2020.

BARRETO, J. O. M.; SOUZA, N. M.; CHAPMAN, E. Síntese de evidências para políticas de saúde: reduzindo a mortalidade perinatal. Ministério da Saúde, p. 1?44,2015.

BATISTA, André Filipe de Moraes; FILHO, Alexandre Dias Porto Chiavegatto. Machine Learning aplicado à Saúde. In: ZIVIANI, Artur; FERNANDES, Natalia Castro; SAADE, Débora Christina Muchaluat. SBCAS 2019: 19 Simpósio Brasileiro de Computação Aplicada à Saúde. [S. l.: s. n.], 2019. cap. Capítulo 1.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. Towards neonatal mortality risk classification: A data-driven approach using neonatal, maternal, and social factors, [s. l.], 23 out. 2020.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. SPNeoDeath: A demographic and epidemiological dataset having infant, mother, prenatal care and childbirth data related to births and neonatal deaths in São Paulo city Brazil ? 2012?2018, [s. l.], 19 jul. 2020.
BRESAN, Rodrigo. Um método para a detecção de ataques de apresentação em sistemas de reconhecimento facial através de propriedades intrínsecas e Deep Learning. Orientador: Me. Carlos Eduardo Beluzo. 2018. Trabalho de Conclusão de Curso (Superior de Tecnologia em Análise e Desenvolvimento de Sistemas) - Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas., [S. l.], 2018.
CABRAL, Cleidy Isolete Silva. Aplicação do Modelo de Regressão Logística num Estudo de Mercado. Orientador: João José Ferreira Gomes. 2013. Projeto Mestrado em Matemática Aplicada à Economia e à Gestão (Mestrado) - Universidade de Lisboa Faculdade de Ciências Departamento de Estatística e Investigação Operacional, [S. l.], 2013. Disponível em: https://repositorio.ul.pt/bitstream/10451/10671/1/ulfc106455_tm_Cleidy_Cabral.pdf. Acesso em: 1 maio 2021.
CAMPOS, Raphael. Árvores de Decisão: Então diga-me, como construo uma?. [S. l.], 28 nov. 2017. Disponível em: https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69. Acesso em: 23 jan. 2021.
Colab. 2021. Disponível em: https://colab.research.google.com/. Acesso em: 20 mar. 2021.
COX, D.R.;SNELL, E.J. Analysis of Binary Data.London: Chapman & Hall, 2ª Edição, 1989. HOSMER, D.W.; LEMESHOW, S.Applied Logistic Regression. New York: John Wiley, 2ª Edição, 2000.
E.França, S.Lansky. Mortalidade infantil neonatal no brasil: situação, tendências e perspectivas Rede Interagencial de Informações para Saúde - demografia e saúde: contribuição para análise de situação e tendências Série Informe de Situação e Tendências(2009), pp.83-112.
62
FREIRE, Sergio Miranda. Bioestatística Básica: Regressão Linear. [S. l.], 20 out. 2020. Disponível em: http://www.lampada.uerj.br/arquivosdb/_book/bioestatisticaBasica.html. Acesso em: 23 jan. 2021.
Jupyter. 2021. Disponível em: https://jupyter.org/. Acesso em: 20 jun. 2021.
GOODFELLOW, I. et al. Deep learning. [S.l.]: MIT press Cambridge, 2016.
Pandas. 2021. Disponível em: https://pandas.pydata.org/. Acesso em: 20 jun. 2021.
Python. 2021. Disponível em: https://www.python.org/. Acesso em: 20 jun. 2021.

KUHN, M.; JOHNSON, K. Applied predictive modeling. [S.l.]: Springer, 2013. v. 26.
LANSKY, S. et al. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante . Cadernos de Saúde Pública, Scielo, v. 30, p. S192 ? S207, 00 2014.
LANSKY, Sônia; FRICHE, Amélia Augusta de Lima; SILVA, Antônio Augusto Moura; CAMPOS, Deise; BITTENCOURT, Sonia Duarte de Azevedo; CARVALHO, Márcia Lazaro; FRIAS, Paulo Germano; CAVALCANTE, Rejane Silva; CUNHA, Antonio José Ledo Alves. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante e ao recém-nascido. [s. l.], Ago 2014. Disponivel em: https://www.scielosp.org/article/csp/2014.v30suppl1/S192-S207/pt/
Matplotlib. 2021. Disponível em: https://matplotlib.org/. Acesso em: 20 mar. 2021.
Maranhão AGK, Vasconcelos AMN, Trindade CM, Victora CG, Rabello Neto DL, Porto D,et al. Mortalidade infantil no Brasil: tendências, componentes e causas de morte no período de 2000 a 2010 . In: Departamento de Análise de Situação de Saúde, Secretaria de Vigilância em Saúde, Ministério da Saúde, organizador. Saúde Brasil 2011: uma análise da situação de saúde e a vigilância da saúde da mulher. v. 1. Brasília: Ministério da Saúde; 2012. p. 163-82.
MESQUITA, PAULO. UM MODELO DE REGRESSÃO LOGÍSTICA PARA AVALIAÇÃO DOS PROGRAMAS DE PÕS-GRADUAÇÃO NO BRASIL. Orientador: Prof. RODRIGO TAVARES NOGUEIRA. 2014. Dissertação (Mestre em Engenharia de Produção) - Universidade Estadual do Norte Fluminense, [S. l.], 2014.
MNASSRI , Baligh. Titanic: logistic regression with python. [S. l.], 1 ago. 2020. Disponível em: https://www.kaggle.com/mnassrib/titanic-logistic-regression-with-python. Acesso em: 14 jan. 2021.
Morais, R.M.d., Costa, A.L: Uma avaliação do sistema de informações sobre mortalidade. Saúde em Debate 41, 101?117 (2017). Disponível em: https://doi.org/10.1109/SIBGRAPI.2012.38.
MOTA, Caio Augusto de Souza; BELUZO , Carlos Eduardo; TRABUCO, Lavinia Pedrosa; SOUZA , Adriano; ALVES, Luciana; CARVALHO, Tiago. DESENVOLVIMENTO DE UMA PLATAFORMA WEB PARA CIÊNCIA DE DADOS APLICADA À SAÚDE MATERNO INFANTIL , [s. l.], 28 nov. 2019.
MULTIEDRO (Brasil). Conheça as aplicações do machine learning na área da saúde. [S. l.], 28 nov. 2019. Disponível em: https://blog.multiedro.com.br/machine-learning-na-area-da-saude/#:~:text=O%20machine%263
0learning%20na%20%C3%A1rea,diagn%C3%B3sticos%20e%20tratamentos%20mais%20precisos. Acesso em: 13 jun. 2021.
MUKHIYA,S.K.;AHMED,U. Hands-On Exploratory Data Analysis with Python: Perform EDA Techniques to Understand, Summarize, and Investigate Your Data. [S.l.]: PacktPublishingLtd,2020.ISBN978-1-78953-725-3.22,32
Murray CJ, Laakso T, Shibuya K, Hill K, Lopez AD. Can we achieve Millennium Development Goal 4? New analysis of country trends and forecasts of under-5 mortality to 2015. Lancet 2007; 370:1040-54.
NumPy. 2021. Disponível em:https://numpy.org/. Acesso em: 20 jun. 2021.
Oliveira, M.M.d., de Araújo, A.S.S.C., Santiago, D.G., Oliveira, J.a.C.G.d., Carvalho, M.D.,

de Lyra et al, R.N.D.: Evaluation of the national information system on live births in brazil , 2006-2010. Epidemiol. Serv. Saúde 24(4), 629?640 (2015).
PELISSARI, ANA CAROLINA CHEBEL. MORTALIDADE NEONATAL DA CIDADE DE SÃO PAULO: UMA ABORDAGEM UTILIZANDO APRENDIZADO DE MÁQUINA SUPERVISIONADO. 2021. Trabalho de Conclusão de Curso (Superior) - Instituto Federal de Educação, Ciência e Tecnologia Campus Campinas, [S. l.], 2021.
REGRESSÃO logística e binary cross entropy - Aula 7. Direção: Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=3J-LBtHVsm4. Acesso em: 21 jan. 2021.
Rmd Nascimento, AJM Leite, NMGSd Almeida, Pcd Almeida, Cfd Silva Determinantes da mortalidade neonatal: estudo caso-controle em fortaleza, ceará, brasil Cad Saúde Pública, 28(2012), pp.559 - 572.
SANTOS, Hellen Geremias. Machine learning e análise preditiva: considerações metodológicas: métodos lineares para classificação. In: SANTOS, Hellen Geremias. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina. 2018. Tese (Pós-Graduação em Epidemiologia) - Faculdade de Saúde Pública da Universidade de São Paulo, [S. l.], 2018.
Scikit-learn. 2021. Disponível em: https://scikit-learn.org/stable/. Acesso em: 20 jun. 2021.
Seaborn. 2021. Disponível em: https://seaborn.pydata.org/. Acesso em: 20 jun. 2021.
SILVA, Wesley; CORSO, Jansen; WELGACZ, Hanna; PEIXE, Julinês. AVALIAÇÃO DA ESCOLHA DE UM FORNECEDOR SOB CONDIÇÃO DE RISCOS A PARTIR DO MÉTODO DE ÁRVORE DE DECISÃO. ARTIGO ? MÉTODOS QUANTITATIVOS, [s. l.], 12 ago. 2008.
WHO, W. H. O.Women and health: today?s evidence, tomorrow?s agenda. [S.l.]: UNWorld Health Organization, 2009. ISBN 9789241563857.
64
APÊNDICE A

==========

Arquivo 1: monografia-V7.docx.pdf (9728 termos)
Arquivo 2: http://repositorio.unicamp.br/jspui/handle/REPOSIP/359544 (373 termos)
Termos comuns: 39
Similaridade: 0,38%
O texto abaixo é o conteúdo do documento monografia-V7.docx.pdf (9728 termos)
Os termos em vermelho foram encontrados no documento
http://repositorio.unicamp.br/jspui/handle/REPOSIP/359544 (373 termos)

==========

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO
CÂMPUS CAMPINAS
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
CAMPINAS
2021

CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS

Trabalho de Conclusão de Curso apresentado como
exigência parcial para obtenção do diploma do
Curso de Tecnologia em Análise e
Desenvolvimento de Sistemas do Instituto Federal
de Educação, Ciência e Tecnologia Câmpus
Campinas.
Orientador: Prof. Me. Everton Josué da Silva.
CAMPINAS
2021

Dados Internacionais de Catalogação na Publicação

CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS

Trabalho de Conclusão de Curso apresentado
como exigência parcial para obtenção do diploma
do Curso de Tecnologia em Análise e
Desenvolvimento de Sistemas do Instituto
Federal de Educação, Ciência e Tecnologia de
São Paulo Câmpus Campinas.
Aprovado pela banca examinadora em: _____ de _____ de _____.
BANCA EXAMINADORA

_____
Prof. Me. Everton Josué da Silva (orientador)

IFSP Câmpus Campinas

_____

Prof. Me. Carlos Eduardo Beluzo
IFSP Câmpus Campinas

_____

Prof. Dr. Ricardo Barz Sovat
IFSP Câmpus Campinas

Dedico este trabalho aos meus familiares,
colegas de classe, professores e servidores do Instituto
que colaboraram em minha jornada formativa.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pelo dom da vida e pela oportunidade de concluir mais uma etapa de minha experiência acadêmica.

Agradeço a todos os professores e servidores do IFSP Câmpus Campinas, que contribuíram direta e indiretamente para a conclusão deste trabalho.

Agradeço também à minha família, que deu todo o apoio necessário para que eu chegasse até aqui.

Agradeço ao meu orientador que me auxiliou a solucionar as dificuldades encontradas no caminho.

"Minha energia é o desafio, minha motivação é o impossível,
e é por isso que eu preciso
ser, à força e a esmo, inabalável.".
Augusto Branco

RESUMO

A mortalidade infantil pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível em uma população. O aprendizado de máquina é uma área da inteligência artificial que propõe métodos de análise de dados que automatizam a construção de modelos analíticos com base em reconhecimento de padrões, baseado no conceito de que sistemas podem aprender com dados, identificando padrões e tomando decisões com o mínimo de intervenção humana. O objetivo deste trabalho é testar e analisar dois tipos diferentes de algoritmos de aprendizado de máquina, utilizando a base de dados do SIM e SINASC do Brasil, do período de 2016 até 2018, para gerar modelos para predição de mortalidade neonatal. Os métodos utilizados foram Árvore de Decisão e de Regressão Logística e como métricas para avaliar os modelos resultantes foram utilizadas a AUC, Curva ROC e a Matriz de Confusão. Como resultado, apesar de terem alcançado valores de AUC 0.93, ambos modelos de predições acertaram muitas predições de vivos cerca de 671.000 e erraram muitas predições de mortos cerca de 2.600, principalmente devido ao fato de a base estar desbalanceada.

Palavras-chave: Mortalidade Neonatal, Predição, Aprendizado de Máquina.

ABSTRACT

Infant mortality can be used to analyze poverty and socioeconomic levels, as well as measure the quality of health and medical technology available in a population. Machine learning is an area of artificial intelligence that proposes data analysis methods that automate the

construction of analytical models based on pattern recognition, based on the concept that systems can learn from data, identifying patterns and making decisions with a minimum of human intervention. The objective of this work is to test and analyze two different types of machine learning algorithms, using the SIM and SINASC do Brasil database, from 2016 to 2018, to generate models for predicting neonatal mortality. The methods used were Decision Tree and Logistic Regression and as metrics to evaluate the resulting models, AUC, ROC Curve and Confusion Matrix were used. As a result, despite achieving values of AUC 0.93, both prediction models correct many live birth predictions around 671.000 and miss many stillbirth predictions around 2600, mainly due to the fact that the base is unbalanced.
Keywords: Neonatal Mortality, Prediction, Machine Learning.
LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

SIM Sistema de Informação sobre Mortalidade
SINASC Sistema de Informações sobre Nascidos Vivos
TMI Taxa de Mortalidade Infantil
TMN Taxa de Mortalidade Neonatal
MN Mortalidade Neonatal
ML Machine Learning
DNV Declaração de Nascido Vivo
VN Verdadeiro Negativo
FN Falso Negativo
VP Verdadeiro Positivo
FP Falso Positivo
ROC Curva Característica de Operação do Receptor
NaN Não é um Número
RFE Eliminação Recursiva de Feature
RFECV Eliminação Recursiva de Feature com Validação Cruzada

SUMÁRIO

13

1 INTRODUÇÃO

A taxa de mortalidade infantil (TMI) é uma importante medida de saúde em uma população e é um grande problema no mundo todo. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível. Esta taxa é dividida em duas categorias: neonatal e pós-neonatal. É categorizada neonatal quando o óbito ocorre nos 28 primeiros dias de vida após o pós-parto, e o pós-neonatal é quando o óbito ocorre entre 29 e 364 dias de vida (BELUZO et al., 2020). Cerca de 46% das mortes no mundo acontecem entre os menores de cinco anos e grande parte está concentrada nos primeiros dias de vida (LANSKY, 2014). A diminuição dessas taxas é muito importante no mundo todo, refletindo nos indicadores de saúde pública e desenvolvimento do país.

Os fatores associados à mortalidade são profundamente influenciados pelas características biológicas maternas e neonatais, pelas condições sociais e pelos cuidados prestados pelos serviços de saúde (E. FRANÇA; S. LANSKY, 2009; R.M.D. NASCIMENTO, 2012). Em grande parte, o diagnóstico é altamente dependente da experiência adquirida pelo

profissional que o realiza. Apesar de indiscutivelmente necessário, não é perfeito, principalmente pelo fato de ser dependente de fatores humanos, por este motivo os profissionais sempre tiveram recursos tecnológicos para auxiliar nessa tarefa (BELUZO et al., 2020). Segundo estudos de 2010 no Brasil, calcula-se que aproximadamente 70% dos óbitos infantis ocorridos poderiam ter sido evitados por ações de saúde e que 60% dos óbitos neonatais ocorreram por situações que poderiam ser evitadas (BARRETO; SOUZA; CHAPMAN, 2015).

No presente trabalho foram utilizados dois algoritmos de aprendizado de máquina para criar modelos de predição de mortalidade neonatal. Além disso, foi realizada também uma análise exploratória da base de dados. Para este trabalho foram utilizadas duas fontes de dados: Sistema de Informação sobre Mortalidade (SIM) e Sistema de Informações sobre Nascidos Vivos (SINASC) do Brasil inteiro.

14

2 JUSTIFICATIVA

Com o avanço da tecnologia, podemos notar que ela está cada vez mais presente no nosso dia a dia e nos auxilia em diversas tarefas do cotidiano, e vem sendo aplicada em diversas áreas, dentre elas a saúde.

A mortalidade infantil é uma preocupação mundial na saúde pública, a ONU (Organizações das Nações Unidas) definiu a redução da mortalidade infantil como uma meta para o desenvolvimento global. A mortalidade neonatal é responsável por aproximadamente 60% da TMI nos países em desenvolvimento (A.K. SINGHA, 2016). Existem diversos fatores que podem contribuir com a redução de óbitos neonatais, como por exemplo acompanhamento médico à gestante no período de gravidez, acompanhamento médico ao recém-nascido nos primeiros dias de vida e a disponibilização deste serviço com qualidade e proficiência (WHO, 2009).

A diminuição dessa taxa é importante e de interesse para o mundo todo, e utilizar da tecnologia para auxiliar nessa diminuição é uma proposta funcional, e inovadora para a realidade brasileira (MOTA et al., 2019). Havendo disponibilidade de tecnologia para auxiliar nas decisões dos diagnósticos, os profissionais da saúde podem focar nos cuidados do recém-nascido com risco de vir a óbito, fornecendo tratamentos melhores.

Segundo o site Multiedro (2019), um grande problema que os médicos enfrentam ao realizar o diagnóstico, é analisar um grande volume de dados. Para a área médica obter diagnósticos rápidos e precisos pode salvar vidas, e a utilização de machine learning para realizar esses processos é importante, com a ML os dados podem ser processados em questões de segundos e resultar em um diagnóstico mais rápido, além disso utilizando bons modelos ML os diagnósticos serão mais precisos.

Diversos trabalhos vêm sendo desenvolvidos neste contexto. No trabalho realizado por BELUZO et al. (2020a), os autores utilizaram uma base de dados com informações da cidade de São Paulo para criação de modelos de aprendizado de máquina para predição de mortalidade neonatal. Este recorte foi utilizado também no presente trabalho para fins de exercício e definição de etapas da implementação de código. Na conclusão os autores discutem os fatores que tiveram mais influência nas predições, e na análise feita pelos autores, as consultas de pré-natal são muito importantes para a redução da mortalidade infantil, uma vez que algumas características muito importantes, como a malformação congênita, podem

ser detectadas ao longo das consultas de pré-natal.
Em um outro estudo realizado por BELUZO et al. (2020a), foi realizada uma análise exploratória da base de dados SPNeoDeath, onde e podemos observar detalhadamente cada
15
análise das colunas da tabela e diversos gráficos feito para um melhor entendimento dos dados.
Outro trabalho que foi realizado utilizando a base de dados com dados somente de São Paulo foi o de PELISSARI (2021), nesse trabalho é realizado uma análise exploratória na base de dados de São Paulo e é também analisado três algoritmos de aprendizado de máquina (Logistic Regression, Random Forest Classifier e XGBoost), nesse trabalho a autora conclui que os modelos de Logistic Regression e XGBoost foram os modelos que apresentaram os melhores desempenhos preditivos, e também mostra os problemas de estar utilizando uma base de dados desbalanceada.
O problema da mortalidade neonatal não é recente, e o mundo todo desenvolve iniciativas para lidar com ele. A ONU definiu como meta a redução da mortalidade infantil para o desenvolvimento global. Diversos fatores podem ajudar na diminuição da taxa de mortalidade neonatal como acompanhamento médico antes e após parto, tanto com a mãe quanto com o recém-nascido. Um serviço de qualidade e constante para os casos com maior risco de óbito pode salvar diversos recém-nascidos. Neste sentido, o desenvolvimento de ferramentas para a predição de mortalidade neonatal pode colaborar com esta iniciativa.

3 OBJETIVOS

3.1 OBJETIVO GERAL

O presente trabalho tem como objetivo testar e analisar os resultados da aplicação dos aprendizagem de máquina supervisionado Árvore de Decisão e Regressão Logística, utilizando dados do SIM e do SINASC, no período de 2016 até 2018, a fim de gerar modelos de predição de risco morte neonatal.

3.2 OBJETIVOS ESPECÍFICOS

? Realizar análise exploratória da base dados a ser utilizada na construção dos modelos;
? Implementar modelos de predição utilizando algoritmos Árvore de Decisão e Regressão Logística;
? Avaliar resultados e propor novas abordagens.
16

4 FUNDAMENTAÇÃO TEÓRICA

4.1 MORTALIDADE INFANTIL E NEONATAL

Como dito na monografia PELISSARI (2021), a TMI consiste no número de crianças que foram a óbito antes de concluir um ano de vida dividido por 1000 crianças nascidas vivas no tempo de um ano. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos e qualidade da saúde pública de um país (apud BELUZO et al., 2020).
Mencionado em PELISSARI (2021), a TMN é uma das duas categorias da TMI, é categorizado mortalidade neonatal quando o óbito ocorre do nascimento da criança até os 28 primeiros dias de vida. A mortalidade neonatal pode ser subdividida em mortalidade neonatal precoce que é quando o óbito ocorre entre 0 e 6 dias de vida, e o neonatal tardio quando o óbito ocorre entre 7 e 28 dias de vida (apud E. FRANÇA e S. LANSKY, 2009).
Segundo Lansky et al. (2014), a taxa de mortalidade neonatal é o principal

componente da TMI desde a década de 1990 no Brasil e vem se mantendo em níveis elevados, com taxa de 11,2 óbitos por mil nascidos vivos em 2010 (apud Maranhão et al., 2012). Também é dito em Lansky et al. (2014), que a TMI do Brasil em 2011 foi 15,3 por mil nascidos vivos, alcançando a meta 4 dos objetivos de desenvolvimento do milênio, compromisso dos governos integrantes das Nações Unidas de melhorar a saúde infantil e reduzir em 2/3 a mortalidade infantil entre 1990 e 2015. (apud Maranhão et al., 2012; apud Murray et al., 2015). Segundo Lansky et al. (2014) o principal componente da TMI em 2009 é a mortalidade neonatal precoce, e grande parte das mortes infantis acontece nas primeiras 24 horas (25%), indicando uma relação estreita com a atenção ao parto e nascimento (apud E. FRANÇA e S. LANSKY, 2009).

4.2 MÉTODOS DE APRENDIZADO DE MÁQUINA

A utilização de dados para a resolução de problemas, de modo a se identificar padrões ocultos e posteriormente auxiliar na tomada de decisões é definida como aprendizado de máquina (BRESAN, 2018 apud Brink et al. 2013).

O uso de técnicas de Aprendizado de Máquina se mostra altamente eficiente na resolução de tarefas que se apresentem difíceis de se resolver a partir de programas escritos por humanos (BRESAN, 2018 apud GOODFELLOW et al., 2016), bem como também possibilita uma maior compreensão sobre os funcionamentos dos princípios que compõem a inteligência humana.

17

Neste trabalho serão implementados modelos utilizando os algoritmos de Árvore de Decisão e Regressão Logística, os quais serão apresentados a seguir.

4.2.1 Árvore de Decisão

Segundo Batista e Filho (2019), os modelos preditivos baseados em árvores são geralmente utilizados para tarefas de classificação, embora também possam ser utilizados para tarefas de regressão. Considerado um dos mais populares algoritmos de predição, o algoritmo de árvore de decisão proporciona uma grande facilidade de interpretação.

As árvores de decisão utilizam a estratégia ?dividir-e-conquistar", na qual as árvores são construídas utilizando-se apenas alguns atributos. A árvores de decisão é uma das técnicas por meio da qual um problema complexo é decomposto em subproblemas mais simples. Recursivamente, a mesma estratégia é aplicada a cada subproblema (SILVA et al, 2008).

A árvore de decisões tem uma estrutura hierárquica onde cada nó da árvore representa uma decisão em uma das colunas do conjunto de dados, cada ramo da árvore representa uma tomada de decisão (BATISTA; FILHO, 2019).

O algoritmo segue algumas decisões para montar a árvore, o atributo mais importante é utilizado como nó raiz e será o primeiro nó, já os atributos menos importantes são mostrados nos ramos da árvore. Cada atributo é verificado para conferir se é possível gerar uma árvore menor e se é possível fazer uma classificação melhor, e assim é escolhido o nó que produz nós filhos (SILVA et al, 2008).

Existem diferentes tipos de algoritmo para a construção da árvore, como por exemplo, ID3 (Iterative Dichotomiser), C4.5 e CART (Classification and Regression Tree). O algoritmo ID3 escolhe os nós da árvore por meio da métrica do ganho de informação. Já o CART faz uso da equação de Gini (BATISTA; FILHO, 2019 apud KUHN; JOHNSON, 2013).

O algoritmo utilizado neste trabalho usará o cálculo da entropia e ganho de

informação, para decidir qual atributo será usado no nó, a entropia é uma medida da desorganização dos sistemas, maior é a incerteza do sistema, quanto maior a entropia maior está a desorganização, com a árvore de decisão a ideia é ir organizando o sistema e ir separando as decisões da melhor forma para que a entropia diminui e o ganho de informação aumente, para que a decisão do modelo fiquei mais assertiva. O cálculo da entropia utiliza a seguinte equação (ÁRVORE, 2020):

???????? =

?

?? ?? ???

2

??

18

Nesta equação o i representa possíveis classificações (rótulos), e o P é a probabilidade de cada uma. A ideia da árvore então é verificar qual é a entropia para cada uma das variáveis da base de dados, e verificar qual é a melhor organização de cada uma das variáveis. E isso é realizado através da técnica chamada de ganho de informação, essa técnica utiliza a equação a seguir (ÁRVORE, 2020):

????? = ????????(???) ? ? ????(??????) * ????????(??????)

Então o ganho de informação é calculado pela entropia do nó pai menos a soma do peso vezes a entropia dos nós filhos. Esse cálculo é realizado para todas as variáveis e a variável que tiver maior ganho de informação é utilizado para a decisão do nó. Esse processo é realizado recursivamente e a árvore vai sendo montada com base nesses cálculos (ÁRVORE, 2020).

Na Figura 1 temos a representação do início de uma árvore de decisão, no caso em questão o autor estava classificando exames positivos e negativos para o vírus SARS-CoV-2, ele utilizou uma base de dados que contém quase 6 mil exames realizados, contendo campos como o resultado do exame, idade do paciente, níveis de hemoglobina entre outras informações.

Figura 1. Exemplo de Árvore de Decisão. (Fonte: ÁRVORE 2020).

Pode-se observar que para o nó raiz foi utilizado a variável Leukocytes, logo foi o que apresentou a melhor organização para ser o nó raiz, e após ele temos os ramos da árvore. Cada retângulo da árvore é uma decisão do modelo e esses retângulos têm atributos importante como, a coluna da base de dados que será onde vai ser realizado a decisão, então pegando o nó raiz, pacientes que tiverem com os Leukocytes para menos que -0.43 seguiram o caminho

19

para a esquerda (true), já os pacientes que tiverem mais seguiram o caminho da direita (false). A entropia calculada do nó também é apresentada e a classificação do nó também, então se o resultado parar neste nó, a classificação que está nele será a classificação final, e temos o samples que é o número de amostras que se enquadram no na decisão. E esse processo de verificação do modelo vai se repetindo até não ter mais como dividir em subproblemas ou até chegar na profundidade máxima.

Uma característica importante da árvore de decisão é que ela é um modelo WhiteBox, ou seja, é possível visualizar a árvore montada e as decisões que o modelo terá que tomar na árvore. Na Figura 1 pode-se ver uma árvore de decisões montada, cada retângulo da árvore é

uma decisão que o modelo terá que tomar, conforme o modelo toma as decisões ele vai seguindo um caminho até chegar ao nó folha, nó folha é o nó que não contém nó filho, não tem mais ramos para baixo, chegando ao nó folha o modelo tem a classificação final.

4.2.2 Regressão Logística

O modelo de regressão logística estabelece uma relação entre a probabilidade de ocorrência da variável de interesse e as variáveis de entrado do modelo, sendo utilizada para tarefas de classificação, em que a variável de interesse é categórica, neste caso, a variável de interesse assume valores 0 ou 1, onde 1 é geralmente utilizado para indicar a ocorrência do evento de interesse (Batista e Filho, 2019).

De acordo com Mesquita (2014), a técnica de regressão logística, desenvolvida no século XIX, obteve maior visibilidade após 1950 ficando então mais conhecida. Ficou ainda mais difundida a partir dos trabalhos de Cox & Snell (1989) e Hosmer & Lemeshow (2000). Caracteriza-se por descrever a relação entre uma variável dependente qualitativa binária, associada a um conjunto de variáveis independentes qualitativas ou métricas.

Esta técnica inicialmente foi utilizada na área médica, porém a eficiência viabilizou sua implementação nas mais diversas áreas. Mesquita (2014) diz que o termo regressão logístico, tem sua origem na transformação usada com a variável dependente, que permite calcular diretamente a probabilidade da ocorrência do fenômeno em estudo.

Segundo Santos (2018), para estimar a probabilidade, é utilizado uma técnica chamada distribuição binomial para modelar a variável resposta do conjunto de treinamento, que tem como parâmetro a probabilidade de ocorrência de uma classe específica. Uma característica importante desse modelo é que a probabilidade estimada deve estar limitada ao intervalo de 0 a 1. O algoritmo de Regressão Logística gera um modelo de classificação binária (0 e 1). O modelo utiliza a Regressão Linear como base, a equação linear é (REGRESSÃO, 2020):

20

? = ? . ? + ?

O ?m? da equação é o coeficiente angular da reta, e o ?b? é o intercepto. Então, aplicando a equação linear obtemos um valor contínuo como resultado, após obter esse valor a Regressão Logística tem que realizar a classificação do resultado, então é aplicado uma função chamada sigmoide (REGRESSÃO, 2020):

? = 1

1+ ? ??

Essa função chamada de sigmoide, ou função logística, é responsável por ?achatar? o resultado da regressão linear, calculando a probabilidade de o resultado pertencer a classe 1, classificando assim o resultado da função linear em 0 ou 1.

Segundo Batista e Filho (2019), o modelo de regressão logística que tem como objetivo realizar uma classificação binária de uma variável de interesse irá predizer a probabilidade de pertencer à classe positiva. Tem-se, portanto, que é dado por:?

? = { 0, ?? ???? (? = 1 ??) < 0, 5

ou

? = { 1, ?? ???? (? = 1??) ? 0, 5

Então a decisão do modelo de regressão logística está relacionada à escolha de um ponto de corte para p(x). Caso o ponto de corte escolhido for p(x)=0,5, os resultados que forem p(x)>0,5 serão classificados como 1 e os resultados p(x)<5 serão classificados como 0

(SANTOS, 2018).
A Figura 2 ilustra um exemplo de como é o diagrama de Regressão Logística. Pelo diagrama então pode-se observar que o modelo possui uma representação gráfica em formato de ?S?, essa seria a curva logística. As predições do modelo sempre ficaram na linha 1 e 0 do eixo y, pois é o intervalo estabelecido pelo algoritmo, então a classificação sempre será nesse intervalo.
21
Figura 2. Exemplo de Diagrama de Regressão Linear. (Fonte: Batista e Filho 2019).
4.2.3 Treino, teste e validação do modelo de aprendizado de máquina
A criação de um modelo de aprendizado de máquina é realizada em duas etapas: treinamento do modelo, que é realizado a partir dos dados fornecidos para o algoritmo, e etapa de teste, onde os modelos recebem dados novos e sem o rótulo, para que seja avaliado a performance do modelo (PELISSARI, 2021).
Para o treinamento dos modelos foram utilizadas duas abordagens, a primeira onde a base foi particionada entre conjunto de dados para teste e para treino em uma porcentagem de 90% para treino e 10% para teste, porém dessa forma pode ocorrer problema de sobreajuste ou overfitting. Nesta situação, o que pode ocorrer é o modelo não aprender com os dados, e apenas ?decorar? os dados recebidos e quando é realizado o teste o modelo consegue predizer apenas situações parecidas, não sendo capaz de identificar padrões com diferenças mínimas.
Para evitar este problema, foi utilizado uma técnica de validação cruzada chamada K-folds cross-validation. Essa técnica de validação cruzada divide a base de dados em K subconjuntos, e então são realizadas várias rodadas de treinamento e teste alternando os subconjuntos utilizados para teste e treino, e o resultado do modelo baseia-se na média da acurácia observada em cada rodada (PELISSARI, 2021).
Pode-se observar na Tabela 1 uma ilustração da técnica K-fold cross-validation, onde temos um exemplo onde a base é dividida em 5 subconjuntos e cada subconjunto tem a parte de teste diferente dos outros subconjuntos, assim o modelo vai receber valores novos de teste e treino todas as vezes que executar um novo subconjunto.
22
Tabela 1. Exemplo da separação K-fold cross-validation. (Fonte: Adaptado de PELISSARI, 2021).

Predição 1	Predição 2	Predição 3	Predição 4	Predição 5
Teste	Treino	Treino	Treino	Treino
Treino	Teste	Treino	Treino	Treino
Treino	Treino	Teste	Treino	Treino
Treino	Treino	Treino	Teste	Treino
Treino	Treino	Treino	Treino	Teste

4.2.4 Métricas para avaliação de modelos
Para fins de avaliação, neste trabalho foram utilizadas três métricas para avaliar o desempenho dos modelos: a Matriz de Confusão e a AUC (Area under Receiver Operating Characteristic Curve - ROC). Na matriz de confusão pode-se observar o comportamento do modelo em relação a cada uma das classes a serem preditas pelo modelo, utilizando variáveis binárias (positivo: 1; e negativo: 0), para apresentar uma matriz que faz uma comparação entre as predições do modelo e os valores corretos do conjunto de dados (PELISSARI, 2021).

Na Tabela 2 temos um exemplo de uma matriz de confusão, a qual consiste em duas colunas e duas linhas, e os valores são atribuídos nos quadrantes com os seguintes resultados (PELISSARI, 2021):

? quando o valor real do conjunto de dados é 0, e a predição do modelo também classificou como 0, então temos um Verdadeiro Negativo (VN);

? quando o valor real é 1, e o modelo classificado como 0, temos um Falso Negativo (FN);

? quando o valor real é 0, e o modelo classificado como 1, então o resultado se enquadra no quadrante de Falso Positivo (FP);

? e por fim o quadrante Verdadeiro Positivo (VP), que são os resultados que tem o valor real 1, e o modelo classificado como 1.

23

Tabela 2. Modelo de Matriz de confusão. (Fonte: Adaptado de PELISSARI, 2021).

Valor Predito
Negativo (0) Positivo (1)
Classe
Real
Negativo
(0)
Verdadeiro
Negativo
Falso
Positivo
Positivo
(1)
Falso
Negativo
Verdadeiro
Positivo

A segunda métrica de avaliação utilizada foi a Curva ROC que permite avaliar o desempenho de um modelo com relação às predições efetuadas. A curva ROC é um gráfico de Sensibilidade (taxa de verdadeiros positivos) versus taxa de falsos positivos, a representação da curva ROC permite evidenciar os valores para os quais existe otimização da Sensibilidade em função da Especificidade (CABRAL, 2013).

Na Figura 3 temos um exemplo de uma curva ROC. A linha traçada na cor preta é uma linha de referência, ela representa hipoteticamente a curva ROC de um classificador puramente aleatório. Caso a linha laranja, que representa o resultado do modelo que está sendo avaliado, ficasse igual a linha tracejada preta, indicaria que o modelo avaliado estaria predizendo aleatoriamente os resultados.

Figura 3: Exemplo de Curva ROC. (Fonte: Elaboração Própria).

Por uma análise de inspeção visual, quanto mais próximo a linha laranja estiver do valor 1 no eixo Y, melhor está a capacidade do modelo para diferenciar as classes. O valor da

24

AUC representa a capacidade do modelo de diferenciar as classes, quanto maior o valor do

AUC, melhor é a predição realizada pelo modelo, uma vez que, os valores resultantes iguais a 0 são realmente 0, e os iguais a 1 são de fato 1 (PELISSARI, 2021).

5 METODOLOGIA

O desenvolvimento deste trabalho foi realizado em três etapas: (1) Pré-processamento e Análise Exploratória do conjunto de dados, onde foram aplicadas técnicas comuns para preparar o conjunto de dados como tratamento de valores nulos e seleção de variáveis de interesse; (2) Implementação de modelos de aprendizado de máquina supervisionado para predição das mortes neonatais, utilizando os algoritmos de árvore de decisão e regressão logística; e (3) Análise de Resultados, que será realizada uma análise do desempenho do modelo.

5.1 TECNOLOGIAS E FERRAMENTAS

Neste trabalho vamos utilizar a linguagem de programação Python (PYTHON, 2021) pela sua simplicidade de codificação e alto poder de processamento. Foi utilizado também a plataforma Google Colab Research (Colab, 2021), para o desenvolvimento dos algoritmos e treinamento dos modelos. Além disso, foi utilizado também o Jupyter Notebook (Jupyter, 2021), instalado localmente em computador pessoal, pois o Google Colab Research possui limitação de recursos de memória e processamento, então as duas plataformas foram usadas para realização deste trabalho. As principais bibliotecas utilizadas foram: numpy (NumPy, 2021), e pandas (Pandas, 2021), para a manusear os dados, matplotlib (Matplotlib, 2021), e seaborn (Seaborn, 2021), para a plotagem dos gráficos, e por fim a biblioteca sklearn (Scikit-Learn, 2021), que é a biblioteca responsável pelos algoritmos de aprendizado de máquina.

5.2 BASE DE DADOS

As bases de dados a serem utilizadas serão o SIM e SINASC, que são as duas principais fontes de informações sobre nascimentos e óbitos no Brasil. O SINASC é alimentado com base na Declaração de Nascido Vivo (Declaração de Nascido Vivo - DNV) (BELUZO et al.,2020b apud Oliveira et al., 2015). O SIM tem como objetivo principal apoiar a coleta, armazenamento e processo de gestão de registros de óbitos no Brasil (BELUZO et 25
al.,2020b apud Morais et al., 2017), e foi usado para rotular o óbito registrado no SIM, utilizando o campo DNV como chave de associação. O SIM será utilizado para rotular os registros do SINASC, pois o SIM coletas informações sobre mortalidade e é usado como base para o cálculo de estatísticas vitais, como a taxa de mortalidade neonatal e o SINASC reúne informações sobre dados demográficos e epidemiológicos do bebê, da mãe, do pré-natal e do parto (BELUZO et al.,2020b).

A Tabela 3 apresenta as variáveis da base de dados. Essa base tem variáveis que contêm valores quantitativos e categóricos. A coluna ?num_live_births? por exemplo contém o número de nascidos vivos anteriores da mãe e é composta por valores quantitativos contínuos, e a coluna ?tp_pregnancy? utiliza valores nominais categóricos que indicam o tipo de gravidez.

Tabela 3. Colunas da base de dados e suas descrições. (Fonte: Adaptado de BELUZO et al.,2020b).

Nome da coluna	Descrição	Domínio de dados
Variáveis	demográficas e socioeconômicas	

maternal_age Idade da mãe Quantitativo Contínuo (inteiro)
tp_maternal_race Raça / cor da pele da
mãe
Nominal categórico (inteiro)
1 - branco; 2 - Preto; 3 -
amarelo; 4 - Pele morena; 5 -
Indígena.
tp_marital_status Estado civil da mãe Nominal categórico (inteiro)
1 - Único; 2 - Casado; 3 - Viúva;
4 - Separados / divorciados
judicialmente; 5 - Casamento
por união estável; 9 - Ignorado.
tp_maternal_schooling Anos de escolaridade da
mãe
Nominal categórico (inteiro)
1 - nenhum; 2 - de 1 a 3 anos; 3 -
de 4 a 7 anos; 4 - de 8 a 11 anos;
5 - 12 e mais; 9 - Ignorado.
Variáveis obstétricas maternas
num_live_births Número de nascidos
vivos
Quantitativo Contínuo (inteiro)
num_fetal_losses Número de perdas fetais Quantitativo Contínuo (inteiro)
num_previous_gestations Número de gestações Quantitativo Contínuo (inteiro)
26
anteriores
num_normal_labors Número de partos
normais (trabalhos de
parto)
Quantitativo Contínuo (inteiro)
num_cesarean_labors Número de partos
cesáreos (partos)
Quantitativo Contínuo (inteiro)
tp_pregnancy Tipo de gravidez Nominal categórico (inteiro)
1 - Singleton; 2 - Twin; 3 -
Trigêmeo ou mais; 9 - Ignorado.
Variáveis relacionadas a cuidados anteriores
tp_labor Tipo de parto infantil
(tipo de parto)
Categórico Nominal (inteiro)
1 - Vaginal; 2 - Cesariana;
num_prenatal_appointments Número de consultas de
pré-natal
Quantitativo Contínuo (inteiro)

tp_robson_group Classificação do grupo
Robson
Ordinal categórico (inteiro)
Variáveis relacionadas ao recém-nascido
tp_newborn_presentation Tipo de apresentação de
recém-nascido
Categórico Nominal (inteiro)
1 - Cefálico; 2 - Pélvico ou
culatra; 3 - Transversal; 9 -
Ignorado.
has_congenital_malformation Presença de
malformação congênita
Nominal categórico (inteiro)
1 - Sim; 2 - Não; 9 - Ignorado
newborn_weight Peso ao nascer em
gramas
Quantitativo Contínuo (inteiro)
cd_apgar1 Pontuação de Apgar de 1
minuto
Ordinal categórico (inteiro)
cd_apgar5 Pontuação de Apgar de 5
minuto
Ordinal categórico (inteiro)
gestaional_week Semana de gestação Quantitativo Contínuo (inteiro)
tp_childbirth_care Assistência ao parto Categórico Nominal (inteiro)
1 - Doutor; 2 - enfermeira ou
obstetra; 3 - Parteira; 4 - outros;
9 - Ignorado.
was_cesarean_before_labor Foi cesáreo antes do
parto.
27
was_labor_induced Foi induzido ao trabalho
de parto.
is_neonatal_death Morte antes de 28 dias
(rótulo)
Nominal categórico (inteiro)
0 - sobrevivente; 1 - morto.

Neste trabalho foi utilizado dados do período de 2014 à 2016 devido sua melhor qualidade. Este recorte possui 6.760.222 registros dos quais 99.4% são amostras de recém-nascidos vivos e 0.6% são amostras onde os recém-nascidos vieram a óbito conforme pode ser observado na Figura 4. Com essas informações consegue-se observar que a base de dados é desbalanceada.

Figura 4. Distribuição da base de dados considerando se o recém-nascido viveu ou morreu durante o período neonatal. (Fonte: Elaboração Própria).

5.3 PREPARAÇÃO DA BASE DE DADOS

Nessa etapa foi realizada a preparação da base de dados para o treinamento dos modelos que consiste na limpeza, transformação e preparação dos dados a serem utilizados na análise exploratória e na criação dos modelos preditivos. A base de dados pode conter informações desnecessárias para o modelo que pode acabar atrapalhando as predições,
28
algumas colunas podem ser retiradas ou seus valores podem ser modificados, por exemplo valores reais são modificados para inteiros.

5.4 DESENVOLVIMENTO DOS MODELOS DE APRENDIZADO DE MÁQUINA

Como já dito anteriormente na seção 4 os algoritmos de aprendizado de máquina escolhidos para esse trabalho foram os: Regressão Logística e Árvore de Decisão que utilizam a lógica mostrada na seção 4 , Fundamentação Teórica. Esses dois foram escolhidos por alguns motivos, ambos são algoritmos de classificação e supervisionados, atendendo os critérios necessários para a predição de mortalidade neonatal, ambos os algoritmos são bons para realizar predições, a Árvore de Decisão inclusive é um ótimo modelo para analisar as decisões que estão sendo feitas pelo modelo, afinal esse algoritmo tem a característica de ser um WhiteBox, ou seja conseguimos ver o que o modelo aprendeu e o que ele está decidindo. Esses algoritmos selecionados são algoritmos que se encaixam na necessidade deste trabalho e são muito utilizados na comunidade acadêmica.

Para o algoritmo de Regressão Logística, como foi dito anteriormente na seção ?Preparação da base de dados? a base de dados foi tratada e particionada, no particionamento foi utilizado 90% da base para o treino e 10% para os testes e então foi realizado o treinamento do modelo, também foi aplicada a função RFECV (Eliminação Recursiva de Feature com Validação Cruzada), esse função executa a RFE (Eliminação Recursiva de Feature), que tem como ideia selecionar features considerando recursivamente conjuntos cada vez menores de features, isso ocorre da seguinte forma. Primeiro, o estimador é treinado no conjunto inicial de features e a importância de cada feature é obtida por meio de um atributo "coef" ou por meio de um atributo "feature_importances". Em seguida, as features menos importantes são removidas do conjunto atual de features. Esse procedimento é repetido recursivamente no conjunto removido até que o número desejado de features a serem selecionados seja finalmente alcançado. Porém o diferencial da RFECV é que essa função executa a RFE em um loop de validação cruzada para encontrar o número ideal ou o melhor número de features.

Após essa seleção, foi realizado o treinamento do modelo usando todas as features e também treinamos usando as features que o RFECV selecionou, para compararmos os resultados no final. Também foi utilizado o método de K-folds cross-validation em um novo treinamento para conferirmos se o modelo estava sofrendo overfitting (sobreajuste). E por fim aplicamos um método chamado GridSearchCV, esse método permite que seja realizado testes
29
alterando algumas combinação de parâmetros nos nossos modelos, facilitando para achar a melhor combinação, esse método é parecido com o cross validation, o diferencial do Grid Search é que ele faz os cross validation de vários modelos com hiperparâmetros diferentes de uma só vez.

Para o algoritmo de Árvore de Decisão também foi utilizado o mesmo

particionamento de 90% para treino e 10% para teste, na criação do modelo foi colocado a entropy como critério de decisão e uma profundidade máxima de 4, pois com números maiores o modelo ficava muito complexo e ele errava mais as predições, também utilizamos um algoritmo para mostrar a importância de cada Feature para o modelo.
O algoritmo de Regressão Logística e análise exploratória foi baseado em códigos disponível na plataforma web Kaggle (MNASSRI, 2020), no exemplo do Kaggle o autor faz os códigos para predizer mortes no navio Titanic, para o trabalho de mortalidade neonatal os códigos foram alterados para realizar predição de mortes neonatais. Já no algoritmo de Árvore de Decisão foi baseado em uma vídeo aula do professor Diogo Cortiz (ÁRVORE..., 2020), nessa vídeo aula o professor explica sobre os algoritmos de Árvore de Decisão e depois implementa o um exemplo de código, o código usado neste trabalho usou o codigo do professor Diogo Cortiz, e foi alterado para realizar predições de mortalidade neonatal.

5.5 ANÁLISE DOS RESULTADOS

Para a análise de resultados foi utilizado dois métodos que é o de Matriz de confusão e a curva ROC esses métodos são explicados na seção 4, a Fundamentação Teórica. Com esses métodos pode-se realizar uma avaliação do modelo, e assim será possível analisar os valores de acurácia, precisão e recall do modelo, essas métricas são utilizadas avaliar o desempenho do modelo, com a acurácia é possível calcular a proximidade entre o valor obtido na predição dos modelos e os valores esperados, com a precisão é possível analisar as amostras que o modelo classificou como 0 eram realmente 0 na classe real e com o recall é possível analisar as amostras que o modelo conseguiu reconhecer como a classe correta.

5.6 ACESSO A BASE DE DADOS E CÓDIGOS

A base de dados utilizada neste projeto pode ser encontrada no link: ... . E os códigos podem ser encontrados no link: ... .

30

6 RESULTADOS

Os mesmos experimentos mostrados abaixo foram aplicados para uma base de dados que contém somente dados de São Paulo, essa base é um recorte da base do Brasil todo em contém cerca de 1.400.000 amostras, esse recorte da base também é bem desbalanceado. Os resultados obtidos nesse experimento usando o recorte de São Paulo foram bem parecidos com os experimentos da base do Brasil todo, e os códigos desse experimento podem ser visualizados no apêndice X.

6.1 ANÁLISE EXPLORATÓRIA

O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

6.1.1 Apresentação dos Dados

Como dito anteriormente na seção 5.2 ?Base de Dados? deste trabalho, a base de dados é formada pelas informações do SIM e do SINASC contém 6.760.222 amostras e 29 colunas, porém serão utilizadas somente 23 colunas sendo elas as colunas descritas na Tabela 3. Na seção 5.2 também tem a Figura 4 que mostra que a base dados é desbalanceada tendo 99.4% das amostras de recém-nascidos vivos e 0.6% das amostras onde os recém-nascidos vieram a óbito no período neonatal.

6.1.1.1 Caracteristicas demográficas e socioeconômicas maternas

O apêndice X contém uma tabela que possui as estatísticas sumarizadas das variáveis

demográficas e socioeconômicas maternas. Nesta tabela por exemplo podemos observar a coluna ?maternal_age? que contém a idade da mãe, e na tabela mostra que a média da idade das mães é de 26.4 anos, e os dados possuem uma variação de no mínimo 8 anos e no máximo 55 anos, e a mediana é 26 anos. Na tabela também tem as colunas: ?tp_maternal_schooling? que mostra a categoria de anos de escolaridade da mãe, ?tp_marital_status? que mostra o estado civil da mãe em dados categóricos e a coluna ?tp_maternal_race? que mostra a raça da mãe também em dados categóricos.
31
6.1.1.2 Variávies obstétricas maternas
No apêndice X pode-se observar uma tabela que possui as estatísticas sumarizadas das variáveis obstétricas maternas. Nesta tabela por exemplo temos a coluna ?num_live_births?, essa coluna mostra o número de nascidos vivos anteriores que a mãe obteve, a média desta coluna é 0.94 , a mediana é 1, o número mínimo é 0 e o máximo é 10 nascidos vivos. Na tabela podemos observar também a coluna, ?num_fetal_losses?, esta coluna mostra o número de perdas fetais anteriores da mãe, a média desta coluna é 0.21, a medina é 0, o número mínimo é 0 e o máximo é 5, esses valores mostram que poucas mães tiveram perdas fetais antes da gravidez atual. Também na tabela tem estatísticas da coluna ?num_previous_gestations? esta coluna mostra o número de gestações anteriores da mãe, esta coluna tem a média de 1.14, a mediana é 1, o número mínimo é 0 e o máximo 10. Essa tabela também contém as colunas: ?num_normal_labors? que é o número de partos normais da mãe, ?num_cesarean_labor? que mostra o número de partos cesáreos da mãe e por fim mostra a coluna ?tp_pregnancy? que é o tipo da gravidez.
6.1.1.3 Variávies de histórico de gravidez
No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis do histórico de gravidez. Nesta tabela temos as colunas ?num_prenatal_appointments? que mostra o número de consultas de pré-natal, a média dessa coluna é 7.8, a mediana é 8, a variação dos dados é de no mínimo 0 e no máximo 40. A tabela também contém as colunas: ?tp_labor? que mostra o tipo de parto, e também contém a coluna ?tp_robson_group? que mostra a classificação do grupo Robson.
6.1.1.4 Variávies relacionadas ao recém-nascido
No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis relacionadas ao recém-nascido. Nesta tabela por exemplo podemos observar a coluna ?newborn_weight? que armazena o peso ao nascer dos recém-nascidos em gramas, a média dos dados dessa coluna é 3187.3, a mediana é 3215, a variação dos dados é de no mínimo 0 e no máximo 6000 gramas. Também podemos observar a coluna ?gestacional_week? que mostra o número de semanas de gestação, a média é 38.4, a mediana é 39, a variação dos dados é de no mínimo 15 e no máximo 45 semanas. Além dessas colunas a tabela também
32
contém as colunas: ?cd_apgar1? que mostra a pontuação de Apgar de 1 minuto, ?cd_apgar5? que mostra a pontuação de Apgar de 5 minuto, ?has_congenital_malformation? que mostra se o recém-nascido contém a presença de alguma malformação, ?tp_newborn_presentation? que mostra o tipo de apresentação de recém-nascido, ?tp_labor? que mostra o tipo de parto, ?was_cesarean_before_labor? que mostra se o recém-nascido foi cesáreo antes do parto, ?was_labor_induced ? que mostra se o recém-nascido foi induzido ao trabalho de parto,

?tp_childbirth_care? que mostra se ouve assistência ao parto, e por fim temos a coluna ?is_neonatal_death? que mostra se o recém-nascido veio a óbito ou não.

6.1.2 Análise das variáveis

Nesta seção serão mostrados a parte de análise de variáveis com diferentes tipos de gráficos.

6.1.2.1 Variáveis contínuas

Na Figura 5 temos dois gráficos boxplot, no boxplot à esquerda temos a distribuição da variável peso ao nascer e nele é possível observar duas caixas onde a caixa azul são os vivos e o laranja são os mortos. Nas duas caixas mostradas no gráfico podemos observar pequenos círculos nas pontas, esses círculos são outliers (dados fora da curva), então para a caixa de vivos os outliers são recém-nascidos com pesos acima de 4500 gramas ou abaixo de 2000 gramas, e para a caixa de mortos os outliers são os recém-nascidos com mais de 5500 gramas. Para os vivos temos a mediana de aproximadamente 3200 gramas e para os mortos a mediana é 1300 gramas. Cada caixa representa 50% da base de dados, a caixa azul de vivos mostra que o peso de recém-nascidos que sobreviveram está aproximadamente entre 2700 a 3600 gramas, e para a caixa laranja de mortos o peso de recém-nascidos que vieram a óbito estão aproximadamente entre 700 a 2500 gramas. Então o gráfico mostra para nós que os recém-nascidos que têm maior risco de vir a óbito tem pesos menores que 2000 gramas.

33

Figura 5. BoxPlot das features Peso ao nascer e Idade da mãe. (Fonte: Elaboração Própria).

No gráfico à direita da Figura 5 temos um boxplot da distribuição da variável idade da mãe, como foi dito anteriormente a caixa azul são os vivos e o laranja são os mortos. Nesse boxplot então podemos ver que os outliers da caixa de vivos são mães com idade acima de 46 anos aproximadamente, para os mortos os outliers são mães com idade acima de 50 anos. Para os vivos temos a mediana de aproximadamente 26 anos e para os mortos a mediana é de 26 anos. Nos vivos a idade das mães está aproximadamente entre 21 a 32 anos, e para os mortos a idade das mães está aproximadamente entre 20 a 34 anos. É importante ressaltar que o gráfico mostra as duas caixas bem parecidas, então de acordo com os dados não contém diferença significativa entre vivos e mortos usando a idade da mãe para distribuir.

Já na Figura 6 temos um boxplot da distribuição da variável semanas de gestação.

Nesse boxplot então podemos ver que os outliers da caixa de vivos são gestações com mais de 44 semanas aproximadamente, e gestações com menos de 35 semanas. Para os vivos temos a mediana de aproximadamente 39 semanas e para os mortos a mediana é de 32 semanas. Nos vivos as semanas de gestação estão aproximadamente entre 36 a 40 semanas, e para os mortos a semanas de gestação estão aproximadamente entre 26 a 36 semanas. Nesse boxplot os dados mostraram que para as gestações que têm menos de 36 semanas os recém-nascidos têm maior risco de vir a óbito.

34

Figura 6. BoxPlot da feature semana de gestação. (Fonte: Elaboração Própria).

Na Figura 7 temos um histograma da distribuição de peso ao nascer por classe, podemos observar no gráfico que grande parte dos vivos (linha azul) estão distribuídos entre 2000 e 4000 gramas, a área entre esses valores tem muitos dados de vivos, já entre os valores 0 e 2000 gramas temos uma concentração dos dados de mortos.

Figura 7. Histograma de distribuição do peso entre as classes. (Fonte: Elaboração Própria)

6.1.2.2 Variáveis categoricas

Observando a Figura 8 podemos ver um gráfico de barras da distribuição da variável escolaridade da mãe, este gráfico mostra a contagem de registros por escolaridade da mãe, esse gráfico mostra para nós que a maior parte das mães estão na categoria 4 que representa de 8 a 11 anos de escolaridade.

35

Figura 8. Gráfico de barras da distribuição da variável Escolaridade da mãe. (Fonte: Elaboração Própria).

Observando a Figura 9 podemos ver um gráfico de barras da distribuição da variável número de vivos, este gráfico mostra a contagem de registros por número de vivos, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho.

Figura 9. Gráfico de barras da distribuição da variável Número de nascidos vivos. (Fonte: Elaboração Própria).

Observando a Figura 10 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 1 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 1 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem

36

apgar de 8 ou 9, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.

Figura 10. Gráfico de barras da distribuição da variável Pontuação Apgar de 1 minuto. (Fonte: Elaboração Própria).

Observando a Figura 11 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 5 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 5 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem apgar de 9 ou 10, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.

Figura 11. Gráfico de barras da distribuição da variável Pontuação Apgar de 5 minutos. (Fonte: Elaboração Própria).

37

Observando a Figura 12 podemos ver um gráfico de barras da distribuição da variável presença de malformação congênita, este gráfico mostra a contagem de registros por presença de malformação congênita, esse gráfico mostra para nós que a grande maioria dos dados estão na categoria 2 que mostra que o recém-nascido não possui malformação, esse grande volume para a categoria 2 é causada pelo desbalanceamento da base.

Figura 12. Gráfico de barras da distribuição da variável Presença de malformação congênita. (Fonte: Elaboração Própria).

Observando a Figura 13 podemos ver um gráfico de barras da distribuição da variável número de gestações, este gráfico mostra a contagem de registros por número de gestações, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho, da mesma forma mostrada na Figura 9.

38

Figura 13. Gráfico de barras da distribuição da variável Número de gestações. (Fonte: Elaboração Própria).

Observando a Figura 14 podemos ver um gráfico de barras da distribuição da variável assistência ao parto, este gráfico mostra a contagem de registros por assistência ao parto, esse gráfico mostra para nós que a grande maioria dos dados mostram que o parto teve assistência de uma Enfermeira ou obstetra ou essa informação foi ignorada na hora do resgistro.
Figura 14. Gráfico de barras da distribuição da variável Assistência ao parto. (Fonte: Elaboração Própria).
6.1.2.3 Analise bi-variada
A Figura 15 mostra para nós a importância da feature peso ao nascer com o gráfico mostrado é possível observar claramente a diferença de pesos entre vivos e mortos, recém-nascidos com peso acima de 2000 gramas sobreviveram, já os recém-nascidos com menos de 2000 gramas vieram a óbito.
39
Figura 15. Gráfico de densidade de peso entre as classes. (Fonte: Elaboração Própria).
Na Figura 16 podemos observar um gráfico de correlação entre algumas features de valores contínuos, algumas dessas features possuem correlações positivas, como pontuação apgar de 1 minuto e pontuação apgar de 5 minutos, essas colunas possuem uma correlação de 0.7, então nascidos que recebem um valor alto de apgar de 1 minuto tendem a receber um valor alto no apgar de 5 minutos. O gráfico também mostra outras features com correlações positivas como, número de partos normais e número de gestações anteriores que contém uma correlação de 0.81, número de gestações anteriores com idade da mãe que tem uma correção de 0.39, número de partos de cesáreos com idade da mãe que possui correlação de 0.24, número de partos normais com idade da mãe com a correlação de 0.26, semanas de gestação com peso ao nascer em gramas com correção de 0.52, e as apgar de 1 e 5 minutos com semanas de gestação. E grande parte das features, possuem correlações baixas então elas são inversamente correlacionadas pelo que o gráfico mostra, como por exemplo número de partos normais com número de consultas pré-natais possuem uma correção de -0.17, e número de partos cesáreos com número de partos normais que contém uma correlação de -0.15. Então a correlação dessas features estão bem baixas tirando as correlações acima de 0.7.
40
Figura 16. Gráfico de correlação. (Fonte: Elaboração Própria)
6.1.2.4 Distribuição por raça
Na Figura 17 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por raça, e podemos observar que a raça 5 (indigena), é o boxplot que contem a menor variação de idade, já as raças 1 (branco) e 3 (amarelo) são as caixas que possuem maior variação com a média em torno de 20 e 30 anos.
41
Figura 17. Distribuição da idade da mãe por raça. (Fonte: Elaboração Própria).
No boxplot mostrado na Figura 18 podemos observar a distribuição de peso ao nascer por raça, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre as raças para peso ao nascer.
Figura 18. Distribuição de peso ao nascer por raça. (Fonte: Elaboração Própria).
Na Figura 19 temos o boxplot da a distribuição de semanas de gestação por raça, e praticamente todas as caixas são iguais, somente a caixa da raça 1 (branco) possui menor variação que as outras.

42
Figura 19. Distribuição de semanas de gestação por raça. (Fonte: Elaboração Própria).
6.1.2.5 Distribuição por estado civil
A Figura 20 mostra o boxplot da a distribuição de peso ao nascer por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre os estados civis para peso ao nascer.
Figura 20. Distribuição de peso ao nascer por estado civil. (Fonte: Elaboração Própria).
A Figura 21 mostra o boxplot da a distribuição de semanas de gestação por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Porém os estados civis 2 (Casado) e 4 (Separados/divorciados judicialmente), contém variações menores.
43
Figura 21. Distribuição de semanas de gestação por estado civil. (Fonte: Elaboração Própria).
Na Figura 22 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por estado civil, e podemos observar que o estado civil 3 (Viúva), possui a maior variação, já o estado civil 4 (Separados/divorciados judicialmente) é a caixa que possuem menor variação
Figura 22. Distribuição da idade da mãe por estado civil. (Fonte: Elaboração Própria).
6.1.2.6 Distribuição por escolaridade da mãe
Na Figura 23 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por escolaridade da mãe, e podemos observar que a escolaridade 2 (1 a 3 anos) e 3 (4 a 7
44
anos), possui as maiores variações, já escolaridade 5 (12 ou mais anos) é a caixa que possuem menor variação
Figura 23. Distribuição da idade da mãe por escolaridade da mãe. (Fonte: Elaboração Própria).
A Figura 24 mostra o boxplot da a distribuição de peso ao nascer por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então, pouca variação é observada entre a escolaridade da mãe para peso ao nascer.
Figura 24. Distribuição de peso ao nascer por escolaridade da mãe. (Fonte: Elaboração Própria).
A Figura 25 mostra o boxplot da a distribuição de semanas de gestação por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas,
45
todas são praticamente iguais. Porém a escolaridade 5 (12 ou mais anos) contém variação menor que as outras.
Figura 25. Distribuição de semanas de gestação por escolaridade da mãe. (Fonte: Elaboração Própria).
6.2 ÁRVORE DE DECISÃO
A Figura 26 mostra um trecho do algoritmo utilizado, nesse trecho é realizado a divisão do DataFrame das features de entrada , e o DataFrame que contém o rótulo.Neste caso o rótulo será o DataFrame ?target?, esse Dataframe contém a feature

?is_neonatal_death?, essa feature informa se o recém nascido veio a óbito ou não, sendo o foco da predição que desejamos realizar essa feature entra como rótulo para o modelo, já no Dataframe ?nome_features? temos as features que vão servir de entrada para o modelo, é a partir dessas features que o modelo tentará encontrar padrões para predizer o risco de óbito do recém nascido. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

46

Figura 26. Separação dos DataFrames de entrada e rótulo. (Fonte: Elaboração Própria).

6.2.1 Modelo de Árvore de Decisão Utilizando Particionamento 90/10

O algoritmo de árvore de decisão foi treinado utilizando duas formas diferentes:
usando as partições 90% para treino e 10% para os testes e utilizando o método K-folds cross-validation. E para os experimentos iniciais foi utilizado o particionamento 90/10.

Tabela 4. Resultados do modelo de árvore de decisão. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	671807
Morto(1)	0.71	0.29	0.41	4216

Analisando a Tabela 4 pode-se ver os seguintes resultados, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 71% que realmente pertenciam a classe 1, então cerca de 29% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 29% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem a classe 1 que seria dos recém nascidos que poderiam vir a óbito, já para a classe de vivos a

47

classe 0 o modelo está identificando muito bem e classificando de forma correta. A tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.

Esse modelo teve uma acurácia de 0.99 que é uma acurácia bem alta, porém somente pela acurácia não é possível dizer que o modelo está excelente pelo motivo da base que está sendo usada é altamente desbalanceada, então para a classe 0 que é a de vivos temos muito mais dados que para a classe de mortos, e o modelo está acertando bastante a classe de vivos logo a acurácia fica alta por esses acertos, enquanto para a classe de mortos o modelo não está acertando tanto.

A Figura 27 mostra a curva ROC do modelo, pode-se observar que a AUC da curva ROC ficou em 0.92, a AUC mostra que o modelo está acertando bastante as predições, pois quanto mais perto de 1 melhor está no nosso modelo, mas no caso desse modelo vimos acima que as predições para a classe de mortos(1) está ruim, o modelo está acertando poucas classificações como 1. Então a AUC assim como a acurácia está alta porque a base de dados utilizada é altamente desbalanceada tendo muito mais amostras de vivos(1), e como o modelo está bom para essa classificação e está acertando bastante os valores de AUC e acurácia ficam altos .

Figura 27. Curva ROC modelo de Árvore de decisão. (Fonte: Elaboração Própria).
A Figura 28 mostra a importância de cada feature para o modelo sendo assim a feature 10 - Peso ao nascer, a mais importante para o modelo, pois dela o modelo conseguiu tirar mais informações, seguido das features, 13 - Apgar dos 5 primeiros minutos, 11 - Semanas de 48
Gestação, 14 - Presença de malformação 12 - Apgar do 1 primeiro minutos. Então essas são as features que foram mais decisivas para a montagem da árvore e para as predições do modelo.
Figura 28. Gráfico de importância das features. (Fonte: Elaboração Própria).
Na Figura 29 temos uma imagem reduzida da árvore de decisão que foi montada com o treinamento utilizando 5 como profundidade máxima para a árvore, a imagem da árvore completa pode ser visualizada no apêndice A. É possível ver que as features marcadas como importante para o modelo apareceu diversas vezes na árvore, e a feature peso ao nascer foi escolhida para ser o nó raiz da árvore, de todas as features ela foi a que teve a melhor entropia para ser o nó raiz, e depois aparece mais vezes para outras decisões em outros níveis da árvore e isso faz com que essa feature se torne a mais importante para o modelo, também podemos ver que a feature ?Apgar do 1 primeiro minuto? mesmo sendo a menos importante para o modelo ela aparece somente no nó de número 42 isso mostra que a entropia e o ganho de informação dessa feature nas outras decisões não foram boas fazendo ela ser a feature com menos importância utilizada no modelo.
49
Figura 29. Árvore de decisão montada a partir do algoritmo. (Fonte: Elaboração Própria).
6.2.2 Modelo de Árvore de Decisão Utilizando K-folds cross-validation
Após realizar esse experimento com a partição 90/10, foi realizado outro experimento com esse mesmo algoritmo porém agora utilizando o método K-folds cross-validation para o treinamento. O K-folds cross-validation foi configurado para separar a base de dados em 10 partes iguais, e assim o algoritmo foi treinado novamente gerando um novo modelo. O método de K-folds cross-validation é utilizado para conferir se o modelo não está sofrendo problemas de overfitting, é importante garantir que o modelo está aprendendo com os dados e não está somente decorando.
Na Tabela 5 temos então o resultados desse modelo, pode-se observar que os resultados para a classificação de vivos(0) não teve alteração, o modelo continua tendo 100% na precisão e no recall, mas na classificação de mortos(1) o modelo teve uma diminuição na precisão que agora está em 65% e também teve uma diminuição no recall que agora está em 23%, porém essa diminuição é justificada porque para o teste deste modelo foi utilizado a base completa e não somente a partição de teste que contém 10% da base total. Logo os resultados não tiveram uma alteração drástica e tiveram um comportamento bem semelhante, incluindo a acurácia que se manteve em 0.99 e a AUC que teve uma diminuição pequena também e ficou com 0.89.
50
Tabela 5. Resultados do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	6719191

Morto(1) 0.65 0.23 0.34 41031
Na Tabela 6 é possível ver a matriz de confusão do modelo que mostra o número exato de erros e acertos do modelo, pode-se observar que para a classe de vivos o modelo classificou corretamente 6.714.117 amostras da base de dados completa, o modelo classificou como 0 as amostras que nos rótulos realmente eram 0, errou apenas 5074 amostras. Já para a classe de mortos o modelo acertou somente 9552 amostras, classificou como 1 as amostras que no rótulo eram 1 também, e errou 31479. Isso mostra que o modelo está muito desbalanceado, pois está errado muito mais do que acertando as predições de mortos, sendo isso um reflexo da base de dados desbalanceada também. E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.
Tabela 6. Matriz de confusão do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6714117 5074
Mortos (1) 31479 9552
6.3 REGRESSÃO LOGÍSTICA
Para o algoritmo de Regressão Logística também foi utilizado métodos diferentes para o treinamento, O algoritmo foi treinado utilizando três formas diferentes: usando as partições 90% para treino e 10% para os testes, utilizando o método K-folds cross-validation e foi treinada utilizando o método RFECV que seleciona as melhores features para o modelo, juntamente com o método GridSearchCV. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.
51
6.3.1 Modelo de Regressão Logística Utilizando Particionamento 90/10
Para o primeiro experimento do algoritmo de regressão logística foi utilizado o particionamento 90/10 para o treinamento do algoritmo, sendo 90% da base de dados para realizar o treinamento do modelo e 10% da base para realizar os testes.
Na Tabela 7 temos os resultados do modelo de regressão logística utilizando o método de particionamento 90/10, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 65% que realmente pertenciam a classe 1, então cerca de 35% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 24% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem as amostras da classe 1 que seria dos recém-nascidos que poderiam vir a óbito, já para a classe de vivos a classe 0 o modelo está identificando muito bem e classificando de forma correta.E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.

Tabela 7. Resultados do modelo de regressão logística usando particionamento 90/10.
(Fonte: Elaboração Própria)
Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 671885
Morto(1) 0.65 0.24 0.35 4138
A Figura 30 abaixo mostra a curva ROC do modelo de regressão logística usando o particionamento 90/10, analisando a curva ROC podemos ver que a AUC da curva ficou em 0.93 e a acurácia do modelo ficou 0.99, mostrando que o modelo está acertando bastante as predições, porém esses números para o nosso modelo não mostra que ele está ótimo, com a análise da Tabela 7 acima podemos observar que o modelo está acertando muitas predições da classe de vivos e poucas predições da classe de mortos, como a base de dados é
52
desbalanceada, como a maior quantidade de dados está na classe de vivos e o modelo está acerto quase todos dessa classe a acurácia e a AUC ficam altas por esses acertos.Logo é preciso analisar mais resultados além da acurácia e da AUC do modelo.
Figura 30. Curva ROC modelo regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria).
Na Tabela 8 podemos analisar os número exatos que o modelo acertou de cada classe, como é mostrado na matriz de confusão o modelo acertou 671.435 da classe de vivos e acertou apenas 915 da classe de mortos, e errou mais que o triplo da classe de mortos. Então as predições do modelo estão muito desbalanceadas, para a classe de vivos o modelo está predizendo bem, porém para as classes de mortos o modelo está errando muitas predições.
Tabela 8. Matriz de confusão do modelo de regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 671435 504
Mortos (1) 3169 915
6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation
Para o segundo experimento do algoritmo de regressão logística foi utilizado o método K-folds cross-validation para o treinamento do modelo com a intenção de conferir e confirmar
53
que o modelo não esteja tendo problemas com overfitting e realmente esteja aprendendo com os dados que está sendo usado para o treinamento. O método K-folds cross-validation foi configurado para realizar uma separação de 10 partes iguais.
Na Tabela 9 podemos observar os resultados do modelo, e os resultados foram parecidos com o experimento anterior usando o particionamento 90/10, única diferença nos resultado é que no experimento anterior a precisão da classe de mortos ficou em 65% e no caso desse experimento usando o K-folds cross-validation a precisão ficou em 64%, mas os valores de recall e F1-Score se mantiveram, assim como todos os resultados da classe de mortos se mantiveram em 100%. A execução desse experimento mostra que o modelo realmente está aprendendo com os dados e não está apenas decorando, retirando o risco do modelo estar com overfitting. Mesmo sem alterações no resultado é importante saber que o

modelo não está com overfitting e está conseguindo aprender e encontrar padrões nos dados.
Tabela 9. Resultados do modelo de regressão logística usando K-folds cross-validation.
(Fonte: Elaboração Própria)
Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 6719191
Morto(1) 0.64 0.24 0.35 41031
Já na Figura 31 temos a curva ROC do modelo, essa curva ROC mostra a AUC de
todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do
método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas
para todas as vezes que foi realizado os testes, a AUC se manteve em 0.93 e 0.94, sendo a
média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não
se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua
acertando muito a classe de vivos que é a classe predominante na base de dados. Já era
esperado os resultados não sofrerem alterações grandes, pois a base de dados não sofreu
nenhuma alteração, esse experimento foi somente para conferir se a modelo não estava
sofrendo overfitting, e pelos resultados aparentemente a base não está com problemas de
sobreajuste.
54
Figura 31. Curva ROC modelo regressão logística usando K-folds cross-validation. (Fonte:
Elaboração Própria).
Na Tabela 10 temos a matriz de confusão do modelo, onde mostra que o modelo
acertou 6.713.658 amostras da classe de vivos e errou apenas 5.533, já para a classe de mortos
o modelo acertou somente 9.847 e novamente errou mais que o triplo errando 31.182
amostras. Como os resultados mostrados anteriormente não tiveram alteração era de se
esperar que as predições corretas e incorretas do modelo também não sofressem alterações, o
modelo continua acertando muito a classe de vivos e errando muito a classe de mortos,
mantendo as predições bem desbalanceadas.
55
Tabela 10. Matriz de confusão do modelo de regressão logística usando K-folds
cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6713658 5533
Mortos (1) 31182 9847
6.3.3 Modelo de Regressão Logística Utilizando GridSearchCV e RFECV
Para o terceiro e último experimento de regressão logística foi utilizado técnicas
diferentes juntas para tentar melhorar as predições do modelo, para esse experimento usamos
o K-folds cross-validation para o particionamento da base de dados assim como foi utilizado
no experimento acima, e também foi utilizado duas técnicas que ainda não foram usadas nos
experimentos anteriores que são as funções RFECV e GridSearchCV. A função RFECV
seleciona a quantidade ideal de features para ser usadas no treinamento do modelo e também
mostra quais são as melhores features para essa predição, então essa função ela utiliza a
validação cruzada para ir treinando e testando N vezes o modelo, e sempre vai alterando a

quantidade e as features utilizadas para o teste, então cada vez que a função testa os modelos ela grava a pontuação dos acertos e assim depois mostra o gráfico da Figura 32 e assim é possível ver quais são as features ideais para o modelo.
56
Figura 32. Resultado da função RFECV Base Brasil. (Fonte: Elaboração Própria).
Então os resultados mostrados na Figura 32 apresenta que o número ideal para treinar o modelo de features é 6, e as features sao: 'tp_maternal_schooling', 'gestaional_week', 'cd_apgar1', 'cd_apgar5', 'has_congenital_malformation', 'tp_labor'. Essas são as features que tiveram o melhor desempenho no RFECV, então para o treinamento do modelo desse experimento as features que vão ser utilizadas serão somente essas 6.
Após a execução da função de RFECV foi executado a função de GridSearchCV, o GridSearchCV é utilizado para descobrir quais são os melhores parâmetros para utilizar no algoritmo de regressão logística e criar os modelos, foi utilizado também a validação cruzada para testar qual seria os melhores parâmetros para melhorar o desempenho do modelo. Na Figura 33 pode-se ver que os parâmetros escolhidos pela função foram:
Figura 33. Resultado da função GridSearchCV. (Fonte: Elaboração Própria).
Na Tabela 11 podemos observar os resultados do modelo, e os resultados foram parecidos com os experimentos anteriores, para a classe de mortos a precisão teve um aumento e foi para 69% e o recall diminuiu chegando em 20%, então utilizando as novas
57
funções o modelo ganhou precisão e perdeu recall. Para a classe de vivos o modelo se manteve em 100% e não teve alterações para os resultados de precisão e recall. Mesmo com a utilização das funções de RFECV e GridSearchCV o modelo não teve uma melhora significativa para as predições de mortos.
Tabela 11. Resultados do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	6719191
Morto(1)	0.69	0.20	0.31	41031

Na Figura 34 temos a curva ROC do modelo, essa curva ROC mostra a AUC de todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas para todas as vezes que foi realizado os testes, a AUC se manteve em 0.92 e 0.93, sendo a média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua acertando muito a classe de vivos que é a classe predominante na base de dados.
58
Figura 34. Curva ROC modelo regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria).
Na Tabela 12 temos a matriz de confusão do modelo, onde mostra que o modelo acertou 6.715.535 amostras da classe de vivos e errou apenas 3.656, comparando com os experimentos anteriores de regressão logística as predições de vivos teve uma pequena piora e teve uma pequena diminuição nos acertos. Já para a classe de mortos, o modelo acertou 8.296 e errou 32.735 amostras, logo as predições de vivos tiveram uma uma piora e acertou um

pouco menos do que os outros experimentos, e na classe de vivos teve uma melhora acertando mais predições, porém como o problema está nas predições de mortos e para essa classe teve uma piora pode-se dizer que a função de GridSearchCV não teve um ganho significativo nos resultados. Então como modelo final foi escolhido o modelo utilizando somente o K-folds cross-validation, pois foi o modelo que mostrou o melhor desempenho dos experimentos de regressão logística.

59

Tabela 12. Matriz de confusão do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6715535 3656
Mortos (1) 32735 8296

60

7 CONCLUSÃO

O principal objetivo deste trabalho foi analisar os resultados das predições de mortalidade neonatal dos algoritmos de aprendizado de máquina usando uma base de dados com dados do Brasil inteiro. A partir destes resultados apresentados na secção anterior ?Resultados?, é possível concluir que analisar somente a AUC e a acurácia do modelo não é o suficiente para garantir que o modelo está excelente, sendo assim temos que ter mais atenção a precisão, recall e as predições mostradas na matriz de confusão.

Ambos os modelos ficaram desbalanceados nas predições, os modelos acertaram mais predições de vivos do que mortos, e as predições corretas de mortos foram mais baixas do que as incorretas, para os modelos ficarem melhor precisam passar por algoritmos que possam balancear essas predições, uma outra alternativa para melhorar o modelo é utilizar uma base de dados mais balanceada, pois essa base que foi utilizada é altamente desbalanceada.

Embora os dois modelos tenham tido resultados parecidos , o modelo de regressão logística utilizando somente o K-folds cross-validation se saiu melhor que os outros experimentos, o modelo criado na seção ?6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation? manteve a acurácia, AUC, precisão e recall dos demais modelos e acertou mais predições, talvez com algoritmos para melhorar as predições da classe de mortos, balanceando mais as predições, esse modelo fique com ótimas taxas preditivas.

Além disso podes-se afirmar que o peso ao nascer do recém nascido é um fator muito importante para as decisões dos modelos, ambos os modelos usaram muito essa informação para as predições, também podemos colocar, as colunas de presença de malformação e semana de gestação, como colunas que foram importantes para os modelos, essas causas podem ser evitadas se houver um acompanhamento médico com qualidade e eficiência.

61

REFERÊNCIAS

ÁRVORE de Decisão - Aula 3. Gravação de Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=ecYpXd4WREk&t=4089s. Acesso em: 1 jul. 2020.

BARRETO, J. O. M.; SOUZA, N. M.; CHAPMAN, E. Síntese de evidências para políticas de saúde: reduzindo a mortalidade perinatal. Ministério da Saúde, p. 1?44,2015.

BATISTA, André Filipe de Moraes; FILHO, Alexandre Dias Porto Chiavegatto. Machine Learning aplicado à Saúde. In: ZIVIANI, Artur; FERNANDES, Natalia Castro; SAADE, Débora Christina Muchaluat. SBCAS 2019: 19 Simpósio Brasileiro de Computação Aplicada à Saúde. [S. l.: s. n.], 2019. cap. Capítulo 1.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. Towards neonatal mortality risk classification: A data-driven approach using neonatal, maternal, and social factors, [s. l.], 23 out. 2020.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. SPNeoDeath: A demographic and epidemiological dataset having infant, mother, prenatal care and childbirth data related to births and neonatal deaths in São Paulo city Brazil ? 2012?2018, [s. l.], 19 jul. 2020.
BRESAN, Rodrigo. Um método para a detecção de ataques de apresentação em sistemas de reconhecimento facial através de propriedades intrínsecas e Deep Learning. Orientador: Me. Carlos Eduardo Beluzo. 2018. Trabalho de Conclusão de Curso (Superior de Tecnologia em Análise e Desenvolvimento de Sistemas) - Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas., [S. l.], 2018.
CABRAL, Cleidy Isolete Silva. Aplicação do Modelo de Regressão Logística num Estudo de Mercado. Orientador: João José Ferreira Gomes. 2013. Projeto Mestrado em Matemática Aplicada à Economia e à Gestão (Mestrado) - Universidade de Lisboa Faculdade de Ciências Departamento de Estatística e Investigação Operacional, [S. l.], 2013. Disponível em: https://repositorio.ul.pt/bitstream/10451/10671/1/ulfc106455_tm_Cleidy_Cabral.pdf. Acesso em: 1 maio 2021.
CAMPOS, Raphael. Árvores de Decisão: Então diga-me, como construo uma?. [S. l.], 28 nov. 2017. Disponível em: https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69. Acesso em: 23 jan. 2021.
Colab. 2021. Disponível em: https://colab.research.google.com/. Acesso em: 20 mar. 2021.
COX, D.R.;SNELL, E.J. Analysis of Binary Data.London: Chapman & Hall, 2ª Edição, 1989. HOSMER, D.W.; LEMESHOW, S.Applied Logistic Regression. New York: John Wiley, 2ª Edição, 2000.
E.França, S.Lansky. Mortalidade infantil neonatal no brasil: situação, tendências e perspectivas Rede Interagencial de Informações para Saúde - demografia e saúde: contribuição para análise de situação e tendências Série Informe de Situação e Tendências(2009), pp.83-112.
62
FREIRE, Sergio Miranda. Bioestatística Básica: Regressão Linear. [S. l.], 20 out. 2020. Disponível em: http://www.lampada.uerj.br/arquivosdb/_book/bioestatisticaBasica.html. Acesso em: 23 jan. 2021.
Jupyter. 2021. Disponível em: https://jupyter.org/. Acesso em: 20 jun. 2021.
GOODFELLOW, I. et al. Deep learning. [S.l.]: MIT press Cambridge, 2016.
Pandas. 2021. Disponível em: https://pandas.pydata.org/. Acesso em: 20 jun. 2021.
Python. 2021. Disponível em: https://www.python.org/. Acesso em: 20 jun. 2021.

KUHN, M.; JOHNSON, K. Applied predictive modeling. [S.l.]: Springer, 2013. v. 26.
LANSKY, S. et al. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante . Cadernos de Saúde Pública, Scielo, v. 30, p. S192 ? S207, 00 2014.
LANSKY, Sônia; FRICHE, Amélia Augusta de Lima; SILVA, Antônio Augusto Moura; CAMPOS, Deise; BITTENCOURT, Sonia Duarte de Azevedo; CARVALHO, Márcia Lazaro; FRIAS, Paulo Germano; CAVALCANTE, Rejane Silva; CUNHA, Antonio José Ledo Alves. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante e ao recém-nascido. [s. l.], Ago 2014. Disponivel em: https://www.scielosp.org/article/csp/2014.v30suppl1/S192-S207/pt/
Matplotlib. 2021. Disponível em: https://matplotlib.org/. Acesso em: 20 mar. 2021.
Maranhão AGK, Vasconcelos AMN, Trindade CM, Victora CG, Rabello Neto DL, Porto D,et al. Mortalidade infantil no Brasil: tendências, componentes e causas de morte no período de 2000 a 2010 . In: Departamento de Análise de Situação de Saúde, Secretaria de Vigilância em Saúde, Ministério da Saúde, organizador. Saúde Brasil 2011: uma análise da situação de saúde e a vigilância da saúde da mulher. v. 1. Brasília: Ministério da Saúde; 2012. p. 163-82.
MESQUITA, PAULO. UM MODELO DE REGRESSÃO LOGÍSTICA PARA AVALIAÇÃO DOS PROGRAMAS DE PÕS-GRADUAÇÃO NO BRASIL. Orientador: Prof. RODRIGO TAVARES NOGUEIRA. 2014. Dissertação (Mestre em Engenharia de Produção) - Universidade Estadual do Norte Fluminense, [S. l.], 2014.
MNASSRI , Baligh. Titanic: logistic regression with python. [S. l.], 1 ago. 2020. Disponível em: https://www.kaggle.com/mnassrib/titanic-logistic-regression-with-python. Acesso em: 14 jan. 2021.
Morais, R.M.d., Costa, A.L: Uma avaliação do sistema de informações sobre mortalidade. Saúde em Debate 41, 101?117 (2017). Disponível em: https://doi.org/10.1109/SIBGRAPI.2012.38.
MOTA, Caio Augusto de Souza; BELUZO , Carlos Eduardo; TRABUCO, Lavinia Pedrosa; SOUZA , Adriano; ALVES, Luciana; CARVALHO, Tiago. DESENVOLVIMENTO DE UMA PLATAFORMA WEB PARA CIÊNCIA DE DADOS APLICADA À SAÚDE MATERNO INFANTIL , [s. l.], 28 nov. 2019.
MULTIEDRO (Brasil). Conheça as aplicações do machine learning na área da saúde. [S. l.], 28 nov. 2019. Disponível em: https://blog.multiedro.com.br/machine-learning-na-area-da-saude/#:~:text=O%20machine%263
0learning%20na%20%C3%A1rea,diagn%C3%B3sticos%20e%20tratamentos%20mais%20precisos. Acesso em: 13 jun. 2021.
MUKHIYA,S.K.;AHMED,U. Hands-On Exploratory Data Analysis with Python: Perform EDA Techniques to Understand, Summarize, and Investigate Your Data. [S.l.]: PacktPublishingLtd,2020.ISBN978-1-78953-725-3.22,32
Murray CJ, Laakso T, Shibuya K, Hill K, Lopez AD. Can we achieve Millennium Development Goal 4? New analysis of country trends and forecasts of under-5 mortality to 2015. Lancet 2007; 370:1040-54.
NumPy. 2021. Disponível em:https://numpy.org/. Acesso em: 20 jun. 2021.
Oliveira, M.M.d., de Araújo, A.S.S.C., Santiago, D.G., Oliveira, J.a.C.G.d., Carvalho, M.D.,

de Lyra et al, R.N.D.: Evaluation of the national information system on live births in brazil , 2006-2010. Epidemiol. Serv. Saúde 24(4), 629?640 (2015).
PELISSARI, ANA CAROLINA CHEBEL. MORTALIDADE NEONATAL DA CIDADE DE SÃO PAULO: UMA ABORDAGEM UTILIZANDO APRENDIZADO DE MÁQUINA SUPERVISIONADO. 2021. Trabalho de Conclusão de Curso (Superior) - Instituto Federal de Educação, Ciência e Tecnologia Campus Campinas, [S. l.], 2021.
REGRESSÃO logística e binary cross entropy - Aula 7. Direção: Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=3J-LBtHVsm4. Acesso em: 21 jan. 2021.
Rmd Nascimento, AJM Leite, NMGSd Almeida, Pcd Almeida, Cfd Silva Determinantes da mortalidade neonatal: estudo caso-controle em fortaleza, ceará, brasil Cad Saúde Pública, 28(2012), pp.559 - 572.
SANTOS, Hellen Geremias. Machine learning e análise preditiva: considerações metodológicas: métodos lineares para classificação. In: SANTOS, Hellen Geremias. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina. 2018. Tese (Pós-Graduação em Epidemiologia) - Faculdade de Saúde Pública da Universidade de São Paulo, [S. l.], 2018.
Scikit-learn. 2021. Disponível em: https://scikit-learn.org/stable/. Acesso em: 20 jun. 2021.
Seaborn. 2021. Disponível em: https://seaborn.pydata.org/. Acesso em: 20 jun. 2021.
SILVA, Wesley; CORSO, Jansen; WELGACZ, Hanna; PEIXE, Julinês. AVALIAÇÃO DA ESCOLHA DE UM FORNECEDOR SOB CONDIÇÃO DE RISCOS A PARTIR DO MÉTODO DE ÁRVORE DE DECISÃO. ARTIGO ? MÉTODOS QUANTITATIVOS, [s. l.], 12 ago. 2008.
WHO, W. H. O.Women and health: today?s evidence, tomorrow?s agenda. [S.l.]: UNWorld Health Organization, 2009. ISBN 9789241563857.
64
APÊNDICE A

==========

Arquivo 1: monografia-V7.docx.pdf (9728 termos)
Arquivo 2: https://www.marinha.mil.br/tm/?q=biblioteca_trabalhos_academicos (2333 termos)
Termos comuns: 39
Similaridade: 0,32%
O texto abaixo é o conteúdo do documento monografia-V7.docx.pdf (9728 termos)
Os termos em vermelho foram encontrados no documento
https://www.marinha.mil.br/tm/?q=biblioteca_trabalhos_academicos (2333 termos)

==========

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO
CÂMPUS CAMPINAS
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
CAMPINAS
2021
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas.
Orientador: Prof. Me. Everton Josué da Silva.
CAMPINAS
2021
Dados Internacionais de Catalogação na Publicação
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo Câmpus Campinas.
Aprovado pela banca examinadora em: _____ de _____ de _____.
BANCA EXAMINADORA

_____
Prof. Me. Everton Josué da Silva (orientador)

IFSP Câmpus Campinas

_____

Prof. Me. Carlos Eduardo Beluzo
IFSP Câmpus Campinas

_____

Prof. Dr. Ricardo Barz Sovat
IFSP Câmpus Campinas

Dedico este trabalho aos meus familiares,
colegas de classe, professores e servidores do Instituto
que colaboraram em minha jornada formativa.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pelo dom da vida e pela oportunidade de concluir mais uma etapa de minha experiência acadêmica.

Agradeço a todos os professores e servidores do IFSP Câmpus Campinas, que contribuíram direta e indiretamente para a conclusão deste trabalho.

Agradeço também à minha família, que deu todo o apoio necessário para que eu chegasse até aqui.

Agradeço ao meu orientador que me auxiliou a solucionar as dificuldades encontradas no caminho.

"Minha energia é o desafio, minha motivação é o impossível,
e é por isso que eu preciso
ser, à força e a esmo, inabalável.".
Augusto Branco

RESUMO

A mortalidade infantil pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível em uma população. O aprendizado de máquina é uma área da inteligência artificial que propõe métodos de análise de dados que automatizam a construção de modelos analíticos com base em reconhecimento de padrões, baseado no conceito de que sistemas podem aprender com dados, identificando padrões e tomando decisões com o mínimo de intervenção humana. O objetivo deste trabalho é testar e analisar dois tipos diferentes de algoritmos de aprendizado de máquina, utilizando a base de dados do SIM e SINASC do Brasil, do período de 2016 até 2018, para gerar modelos para predição de mortalidade neonatal. Os métodos utilizados foram Árvore de Decisão e de Regressão Logística e como métricas para avaliar os modelos resultantes foram utilizadas a AUC, Curva ROC e a Matriz de Confusão. Como resultado, apesar de terem alcançado valores de AUC 0.93, ambos modelos de predições acertaram muitas predições de vivos cerca de 671.000 e erraram muitas predições de mortos cerca de 2.600, principalmente devido ao fato de a base estar desbalanceada.

Palavras-chave: Mortalidade Neonatal, Predição, Aprendizado de Máquina.

ABSTRACT

Infant mortality can be used to analyze poverty and socioeconomic levels, as well as measure the quality of health and medical technology available in a population. Machine learning is an area of artificial intelligence that proposes data analysis methods that automate the

construction of analytical models based on pattern recognition, based on the concept that systems can learn from data, identifying patterns and making decisions with a minimum of human intervention. The objective of this work is to test and analyze two different types of machine learning algorithms, using the SIM and SINASC do Brasil database, from 2016 to 2018, to generate models for predicting neonatal mortality. The methods used were Decision Tree and Logistic Regression and as metrics to evaluate the resulting models, AUC, ROC Curve and Confusion Matrix were used. As a result, despite achieving values of AUC 0.93, both prediction models correct many live birth predictions around 671.000 and miss many stillbirth predictions around 2600, mainly due to the fact that the base is unbalanced.
Keywords: Neonatal Mortality, Prediction, Machine Learning.
LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS
SIM Sistema de Informação sobre Mortalidade
SINASC Sistema de Informações sobre Nascidos Vivos
TMI Taxa de Mortalidade Infantil
TMN Taxa de Mortalidade Neonatal
MN Mortalidade Neonatal
ML Machine Learning
DNV Declaração de Nascido Vivo
VN Verdadeiro Negativo
FN Falso Negativo
VP Verdadeiro Positivo
FP Falso Positivo
ROC Curva Característica de Operação do Receptor
NaN Não é um Número
RFE Eliminação Recursiva de Feature
RFECV Eliminação Recursiva de Feature com Validação Cruzada
SUMÁRIO

13

1 INTRODUÇÃO

A taxa de mortalidade infantil (TMI) é uma importante medida de saúde em uma população e é um grande problema no mundo todo. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível. Esta taxa é dividida em duas categorias: neonatal e pós-neonatal. É categorizada neonatal quando o óbito ocorre nos 28 primeiros dias de vida após o pós-parto, e o pós-neonatal é quando o óbito ocorre entre 29 e 364 dias de vida (BELUZO et al., 2020). Cerca de 46% das mortes no mundo acontecem entre os menores de cinco anos e grande parte está concentrada nos primeiros dias de vida (LANSKY, 2014). A diminuição dessas taxas é muito importante no mundo todo, refletindo nos indicadores de saúde pública e desenvolvimento do país.

Os fatores associados à mortalidade são profundamente influenciados pelas características biológicas maternas e neonatais, pelas condições sociais e pelos cuidados prestados pelos serviços de saúde (E. FRANÇA; S. LANSKY, 2009; R.M.D. NASCIMENTO, 2012). Em grande parte, o diagnóstico é altamente dependente da experiência adquirida pelo

profissional que o realiza. Apesar de indiscutivelmente necessário, não é perfeito, principalmente pelo fato de ser dependente de fatores humanos, por este motivo os profissionais sempre tiveram recursos tecnológicos para auxiliar nessa tarefa (BELUZO et al., 2020). Segundo estudos de 2010 no Brasil, calcula-se que aproximadamente 70% dos óbitos infantis ocorridos poderiam ter sido evitados por ações de saúde e que 60% dos óbitos neonatais ocorreram por situações que poderiam ser evitadas (BARRETO; SOUZA; CHAPMAN, 2015).

No presente trabalho foram utilizados dois algoritmos de aprendizado de máquina para criar modelos de predição de mortalidade neonatal. Além disso, foi realizada também uma análise exploratória da base de dados. Para este trabalho foram utilizadas duas fontes de dados: Sistema de Informação sobre Mortalidade (SIM) e Sistema de Informações sobre Nascidos Vivos (SINASC) do Brasil inteiro.

14

2 JUSTIFICATIVA

Com o avanço da tecnologia, podemos notar que ela está cada vez mais presente no nosso dia a dia e nos auxilia em diversas tarefas do cotidiano, e vem sendo aplicada em diversas áreas, dentre elas a saúde.

A mortalidade infantil é uma preocupação mundial na saúde pública, a ONU (Organizações das Nações Unidas) definiu a redução da mortalidade infantil como uma meta para o desenvolvimento global. A mortalidade neonatal é responsável por aproximadamente 60% da TMI nos países em desenvolvimento (A.K. SINGHA, 2016). Existem diversos fatores que podem contribuir com a redução de óbitos neonatais, como por exemplo acompanhamento médico à gestante no período de gravidez, acompanhamento médico ao recém-nascido nos primeiros dias de vida e a disponibilização deste serviço com qualidade e proficiência (WHO, 2009).

A diminuição dessa taxa é importante e de interesse para o mundo todo, e utilizar da tecnologia para auxiliar nessa diminuição é uma proposta funcional, e inovadora para a realidade brasileira (MOTA et al., 2019). Havendo disponibilidade de tecnologia para auxiliar nas decisões dos diagnósticos, os profissionais da saúde podem focar nos cuidados do recém-nascido com risco de vir a óbito, fornecendo tratamentos melhores.

Segundo o site Multiedro (2019), um grande problema que os médicos enfrentam ao realizar o diagnóstico, é analisar um grande volume de dados. Para a área médica obter diagnósticos rápidos e precisos pode salvar vidas, e a utilização de machine learning para realizar esses processos é importante, com a ML os dados podem ser processados em questões de segundos e resultar em um diagnóstico mais rápido, além disso utilizando bons modelos ML os diagnósticos serão mais precisos.

Diversos trabalhos vêm sendo desenvolvidos neste contexto. No trabalho realizado por BELUZO et al. (2020a), os autores utilizaram uma base de dados com informações da cidade de São Paulo para criação de modelos de aprendizado de máquina para predição de mortalidade neonatal. Este recorte foi utilizado também no presente trabalho para fins de exercício e definição de etapas da implementação de código. Na conclusão os autores discutem os fatores que tiveram mais influência nas predições, e na análise feita pelos autores, as consultas de pré-natal são muito importantes para a redução da mortalidade infantil, uma vez que algumas características muito importantes, como a malformação congênita, podem

ser detectadas ao longo das consultas de pré-natal.
Em um outro estudo realizado por BELUZO et al. (2020a), foi realizada uma análise exploratória da base de dados SPNeoDeath, onde e podemos observar detalhadamente cada
15
análise das colunas da tabela e diversos gráficos feito para um melhor entendimento dos dados.
Outro trabalho que foi realizado utilizando a base de dados com dados somente de São Paulo foi o de PELISSARI (2021), nesse trabalho é realizado uma análise exploratória na base de dados de São Paulo e é também analisado três algoritmos de aprendizado de máquina (Logistic Regression, Random Forest Classifier e XGBoost), nesse trabalho a autora conclui que os modelos de Logistic Regression e XGBoost foram os modelos que apresentaram os melhores desempenhos preditivos, e também mostra os problemas de estar utilizando uma base de dados desbalanceada.
O problema da mortalidade neonatal não é recente, e o mundo todo desenvolve iniciativas para lidar com ele. A ONU definiu como meta a redução da mortalidade infantil para o desenvolvimento global. Diversos fatores podem ajudar na diminuição da taxa de mortalidade neonatal como acompanhamento médico antes e após parto, tanto com a mãe quanto com o recém-nascido. Um serviço de qualidade e constante para os casos com maior risco de óbito pode salvar diversos recém-nascidos. Neste sentido, o desenvolvimento de ferramentas para a predição de mortalidade neonatal pode colaborar com esta iniciativa.

3 OBJETIVOS

3.1 OBJETIVO GERAL

O presente trabalho tem como objetivo testar e analisar os resultados da aplicação dos aprendizagem de máquina supervisionado Árvore de Decisão e Regressão Logística, utilizando dados do SIM e do SINASC, no período de 2016 até 2018, a fim de gerar modelos de predição de risco morte neonatal.

3.2 OBJETIVOS ESPECÍFICOS

? Realizar análise exploratória da base dados a ser utilizada na construção dos modelos;
? Implementar modelos de predição utilizando algoritmos Árvore de Decisão e Regressão Logística;
? Avaliar resultados e propor novas abordagens.
16

4 FUNDAMENTAÇÃO TEÓRICA

4.1 MORTALIDADE INFANTIL E NEONATAL

Como dito na monografia PELISSARI (2021), a TMI consiste no número de crianças que foram a óbito antes de concluir um ano de vida dividido por 1000 crianças nascidas vivas no tempo de um ano. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos e qualidade da saúde pública de um país (apud BELUZO et al., 2020).
Mencionado em PELISSARI (2021), a TMN é uma das duas categorias da TMI, é categorizado mortalidade neonatal quando o óbito ocorre do nascimento da criança até os 28 primeiros dias de vida. A mortalidade neonatal pode ser subdividida em mortalidade neonatal precoce que é quando o óbito ocorre entre 0 e 6 dias de vida, e o neonatal tardio quando o óbito ocorre entre 7 e 28 dias de vida (apud E. FRANÇA e S. LANSKY, 2009).
Segundo Lansky et al. (2014), a taxa de mortalidade neonatal é o principal

componente da TMI desde a década de 1990 no Brasil e vem se mantendo em níveis elevados, com taxa de 11,2 óbitos por mil nascidos vivos em 2010 (apud Maranhão et al., 2012). Também é dito em Lansky et al. (2014), que a TMI do Brasil em 2011 foi 15,3 por mil nascidos vivos, alcançando a meta 4 dos objetivos de desenvolvimento do milênio, compromisso dos governos integrantes das Nações Unidas de melhorar a saúde infantil e reduzir em 2/3 a mortalidade infantil entre 1990 e 2015. (apud Maranhão et al., 2012; apud Murray et al., 2015). Segundo Lansky et al. (2014) o principal componente da TMI em 2009 é a mortalidade neonatal precoce, e grande parte das mortes infantis acontece nas primeiras 24 horas (25%), indicando uma relação estreita com a atenção ao parto e nascimento (apud E. FRANÇA e S. LANSKY, 2009).

4.2 MÉTODOS DE APRENDIZADO DE MÁQUINA

A utilização de dados para a resolução de problemas, de modo a se identificar padrões ocultos e posteriormente auxiliar na tomada de decisões é definida como aprendizado de máquina (BRESAN, 2018 apud Brink et al. 2013).

O uso de técnicas de Aprendizado de Máquina se mostra altamente eficiente na resolução de tarefas que se apresentem difíceis de se resolver a partir de programas escritos por humanos (BRESAN, 2018 apud GOODFELLOW et al., 2016), bem como também possibilita uma maior compreensão sobre os funcionamentos dos princípios que compõem a inteligência humana.

17

Neste trabalho serão implementados modelos utilizando os algoritmos de Árvore de Decisão e Regressão Logística, os quais serão apresentados a seguir.

4.2.1 Árvore de Decisão

Segundo Batista e Filho (2019), os modelos preditivos baseados em árvores são geralmente utilizados para tarefas de classificação, embora também possam ser utilizados para tarefas de regressão. Considerado um dos mais populares algoritmos de predição, o algoritmo de árvore de decisão proporciona uma grande facilidade de interpretação.

As árvores de decisão utilizam a estratégia ?dividir-e-conquistar", na qual as árvores são construídas utilizando-se apenas alguns atributos. A árvores de decisão é uma das técnicas por meio da qual um problema complexo é decomposto em subproblemas mais simples. Recursivamente, a mesma estratégia é aplicada a cada subproblema (SILVA et al, 2008).

A árvore de decisões tem uma estrutura hierárquica onde cada nó da árvore representa uma decisão em uma das colunas do conjunto de dados, cada ramo da árvore representa uma tomada de decisão (BATISTA; FILHO, 2019).

O algoritmo segue algumas decisões para montar a árvore, o atributo mais importante é utilizado como nó raiz e será o primeiro nó, já os atributos menos importantes são mostrados nos ramos da árvore. Cada atributo é verificado para conferir se é possível gerar uma árvore menor e se é possível fazer uma classificação melhor, e assim é escolhido o nó que produz nós filhos (SILVA et al, 2008).

Existem diferentes tipos de algoritmo para a construção da árvore, como por exemplo, ID3 (Iterative Dichotomiser), C4.5 e CART (Classification and Regression Tree). O algoritmo ID3 escolhe os nós da árvore por meio da métrica do ganho de informação. Já o CART faz uso da equação de Gini (BATISTA; FILHO, 2019 apud KUHN; JOHNSON, 2013).

O algoritmo utilizado neste trabalho usará o cálculo da entropia e ganho de

informação, para decidir qual atributo será usado no nó, a entropia é uma medida da desorganização dos sistemas, maior é a incerteza do sistema, quanto maior a entropia maior está a desorganização, com a árvore de decisão a ideia é ir organizando o sistema e ir separando as decisões da melhor forma para que a entropia diminui e o ganho de informação aumente, para que a decisão do modelo fiquei mais assertiva. O cálculo da entropia utiliza a seguinte equação (ÁRVORE, 2020):

???????? =

?

?? ?? ???

2

??

18

Nesta equação o i representa possíveis classificações (rótulos), e o P é a probabilidade de cada uma. A ideia da árvore então é verificar qual é a entropia para cada uma das variáveis da base de dados, e verificar qual é a melhor organização de cada uma das variáveis. E isso é realizado através da técnica chamada de ganho de informação, essa técnica utiliza a equação a seguir (ÁRVORE, 2020):

????? = ????????(???) ? ? ????(??????) * ????????(??????)

Então o ganho de informação é calculado pela entropia do nó pai menos a soma do peso vezes a entropia dos nós filhos. Esse cálculo é realizado para todas as variáveis e a variável que tiver maior ganho de informação é utilizado para a decisão do nó. Esse processo é realizado recursivamente e a árvore vai sendo montada com base nesses cálculos (ÁRVORE, 2020).

Na Figura 1 temos a representação do início de uma árvore de decisão, no caso em questão o autor estava classificando exames positivos e negativos para o vírus SARS-CoV-2, ele utilizou uma base de dados que contém quase 6 mil exames realizados, contendo campos como o resultado do exame, idade do paciente, níveis de hemoglobina entre outras informações.

Figura 1. Exemplo de Árvore de Decisão. (Fonte: ÁRVORE 2020).

Pode-se observar que para o nó raiz foi utilizado a variável Leukocytes, logo foi o que apresentou a melhor organização para ser o nó raiz, e após ele temos os ramos da árvore. Cada retângulo da árvore é uma decisão do modelo e esses retângulos têm atributos importante como, a coluna da base de dados que será onde vai ser realizado a decisão, então pegando o nó raiz, pacientes que tiverem com os Leukocytes para menos que -0.43 seguiram o caminho

19

para a esquerda (true), já os pacientes que tiverem mais seguiram o caminho da direita (false). A entropia calculada do nó também é apresentada e a classificação do nó também, então se o resultado parar neste nó, a classificação que está nele será a classificação final, e temos o samples que é o número de amostras que se enquadram no na decisão. E esse processo de verificação do modelo vai se repetindo até não ter mais como dividir em subproblemas ou até chegar na profundidade máxima.

Uma característica importante da árvore de decisão é que ela é um modelo WhiteBox, ou seja, é possível visualizar a árvore montada e as decisões que o modelo terá que tomar na árvore. Na Figura 1 pode-se ver uma árvore de decisões montada, cada retângulo da árvore é

uma decisão que o modelo terá que tomar, conforme o modelo toma as decisões ele vai seguindo um caminho até chegar ao nó folha, nó folha é o nó que não contém nó filho, não tem mais ramos para baixo, chegando ao nó folha o modelo tem a classificação final.

4.2.2 Regressão Logística

O modelo de regressão logística estabelece uma relação entre a probabilidade de ocorrência da variável de interesse e as variáveis de entrado do modelo, sendo utilizada para tarefas de classificação, em que a variável de interesse é categórica, neste caso, a variável de interesse assume valores 0 ou 1, onde 1 é geralmente utilizado para indicar a ocorrência do evento de interesse (Batista e Filho, 2019).

De acordo com Mesquita (2014), a técnica de regressão logística, desenvolvida no século XIX, obteve maior visibilidade após 1950 ficando então mais conhecida. Ficou ainda mais difundida a partir dos trabalhos de Cox & Snell (1989) e Hosmer & Lemeshow (2000). Caracteriza-se por descrever a relação entre uma variável dependente qualitativa binária, associada a um conjunto de variáveis independentes qualitativas ou métricas.

Esta técnica inicialmente foi utilizada na área médica, porém a eficiência viabilizou sua implementação nas mais diversas áreas. Mesquita (2014) diz que o termo regressão logístico, tem sua origem na transformação usada com a variável dependente, que permite calcular diretamente a probabilidade da ocorrência do fenômeno em estudo.

Segundo Santos (2018), para estimar a probabilidade, é utilizado uma técnica chamada distribuição binomial para modelar a variável resposta do conjunto de treinamento, que tem como parâmetro a probabilidade de ocorrência de uma classe específica. Uma característica importante desse modelo é que a probabilidade estimada deve estar limitada ao intervalo de 0 a 1. O algoritmo de Regressão Logística gera um modelo de classificação binária (0 e 1). O modelo utiliza a Regressão Linear como base, a equação linear é (REGRESSÃO, 2020):

20

? = ? . ? + ?

O ?m? da equação é o coeficiente angular da reta, e o ?b? é o intercepto. Então, aplicando a equação linear obtemos um valor contínuo como resultado, após obter esse valor a Regressão Logística tem que realizar a classificação do resultado, então é aplicado uma função chamada sigmoide (REGRESSÃO, 2020):

? = 1

1+ ? ??

Essa função chamada de sigmoide, ou função logística, é responsável por ?achatar? o resultado da regressão linear, calculando a probabilidade de o resultado pertencer a classe 1, classificando assim o resultado da função linear em 0 ou 1.

Segundo Batista e Filho (2019), o modelo de regressão logística que tem como objetivo realizar uma classificação binária de uma variável de interesse irá predizer a probabilidade de pertencer à classe positiva. Tem-se, portanto, que é dado por:?

? = { 0, ?? ???? (? = 1 ??) < 0, 5

ou

? = { 1, ?? ???? (? = 1??) ? 0, 5

Então a decisão do modelo de regressão logística está relacionada à escolha de um ponto de corte para p(x). Caso o ponto de corte escolhido for p(x)=0,5, os resultados que forem p(x)>0,5 serão classificados como 1 e os resultados p(x)<5 serão classificados como 0

(SANTOS, 2018).
A Figura 2 ilustra um exemplo de como é o diagrama de Regressão Logística. Pelo diagrama então pode-se observar que o modelo possui uma representação gráfica em formato de ?S?, essa seria a curva logística. As predições do modelo sempre ficaram na linha 1 e 0 do eixo y, pois é o intervalo estabelecido pelo algoritmo, então a classificação sempre será nesse intervalo.
21
Figura 2. Exemplo de Diagrama de Regressão Linear. (Fonte: Batista e Filho 2019).
4.2.3 Treino, teste e validação do modelo de aprendizado de máquina
A criação de um modelo de aprendizado de máquina é realizada em duas etapas: treinamento do modelo, que é realizado a partir dos dados fornecidos para o algoritmo, e etapa de teste, onde os modelos recebem dados novos e sem o rótulo, para que seja avaliado a performance do modelo (PELISSARI, 2021).
Para o treinamento dos modelos foram utilizadas duas abordagens, a primeira onde a base foi particionada entre conjunto de dados para teste e para treino em uma porcentagem de 90% para treino e 10% para teste, porém dessa forma pode ocorrer problema de sobreajuste ou overfitting. Nesta situação, o que pode ocorrer é o modelo não aprender com os dados, e apenas ?decorar? os dados recebidos e quando é realizado o teste o modelo consegue predizer apenas situações parecidas, não sendo capaz de identificar padrões com diferenças mínimas.
Para evitar este problema, foi utilizado uma técnica de validação cruzada chamada K-folds cross-validation. Essa técnica de validação cruzada divide a base de dados em K subconjuntos, e então são realizadas várias rodadas de treinamento e teste alternando os subconjuntos utilizados para teste e treino, e o resultado do modelo baseia-se na média da acurácia observada em cada rodada (PELISSARI, 2021).
Pode-se observar na Tabela 1 uma ilustração da técnica K-fold cross-validation, onde temos um exemplo onde a base é dividida em 5 subconjuntos e cada subconjunto tem a parte de teste diferente dos outros subconjuntos, assim o modelo vai receber valores novos de teste e treino todas as vezes que executar um novo subconjunto.
22
Tabela 1. Exemplo da separação K-fold cross-validation. (Fonte: Adaptado de PELISSARI, 2021).

Predição 1	Predição 2	Predição 3	Predição 4	Predição 5
Teste	Treino	Treino	Treino	Treino
Treino	Teste	Treino	Treino	Treino
Treino	Treino	Teste	Treino	Treino
Treino	Treino	Treino	Teste	Treino
Treino	Treino	Treino	Treino	Teste

4.2.4 Métricas para avaliação de modelos
Para fins de avaliação, neste trabalho foram utilizadas três métricas para avaliar o desempenho dos modelos: a Matriz de Confusão e a AUC (Area under Receiver Operating Characteristic Curve - ROC). Na matriz de confusão pode-se observar o comportamento do modelo em relação a cada uma das classes a serem preditas pelo modelo, utilizando variáveis binárias (positivo: 1; e negativo: 0), para apresentar uma matriz que faz uma comparação entre as predições do modelo e os valores corretos do conjunto de dados (PELISSARI, 2021).

Na Tabela 2 temos um exemplo de uma matriz de confusão, a qual consiste em duas colunas e duas linhas, e os valores são atribuídos nos quadrantes com os seguintes resultados (PELISSARI, 2021):

? quando o valor real do conjunto de dados é 0, e a predição do modelo também classificou como 0, então temos um Verdadeiro Negativo (VN);

? quando o valor real é 1, e o modelo classificado como 0, temos um Falso Negativo (FN);

? quando o valor real é 0, e o modelo classificado como 1, então o resultado se enquadra no quadrante de Falso Positivo (FP);

? e por fim o quadrante Verdadeiro Positivo (VP), que são os resultados que tem o valor real 1, e o modelo classificado como 1.

23

Tabela 2. Modelo de Matriz de confusão. (Fonte: Adaptado de PELISSARI, 2021).

Valor Predito
Negativo (0) Positivo (1)
Classe
Real
Negativo
(0)
Verdadeiro
Negativo
Falso
Positivo
Positivo
(1)
Falso
Negativo
Verdadeiro
Positivo

A segunda métrica de avaliação utilizada foi a Curva ROC que permite avaliar o desempenho de um modelo com relação às predições efetuadas. A curva ROC é um gráfico de Sensibilidade (taxa de verdadeiros positivos) versus taxa de falsos positivos, a representação da curva ROC permite evidenciar os valores para os quais existe otimização da Sensibilidade em função da Especificidade (CABRAL, 2013).

Na Figura 3 temos um exemplo de uma curva ROC. A linha traçada na cor preta é uma linha de referência, ela representa hipoteticamente a curva ROC de um classificador puramente aleatório. Caso a linha laranja, que representa o resultado do modelo que está sendo avaliado, ficasse igual a linha tracejada preta, indicaria que o modelo avaliado estaria predizendo aleatoriamente os resultados.

Figura 3: Exemplo de Curva ROC. (Fonte: Elaboração Própria).

Por uma análise de inspeção visual, quanto mais próximo a linha laranja estiver do valor 1 no eixo Y, melhor está a capacidade do modelo para diferenciar as classes. O valor da

24

AUC representa a capacidade do modelo de diferenciar as classes, quanto maior o valor do

AUC, melhor é a predição realizada pelo modelo, uma vez que, os valores resultantes iguais a 0 são realmente 0, e os iguais a 1 são de fato 1 (PELISSARI, 2021).

5 METODOLOGIA

O desenvolvimento deste trabalho foi realizado em três etapas: (1) Pré-processamento e Análise Exploratória do conjunto de dados, onde foram aplicadas técnicas comuns para preparar o conjunto de dados como tratamento de valores nulos e seleção de variáveis de interesse; (2) Implementação de modelos de aprendizado de máquina supervisionado para predição das mortes neonatais, utilizando os algoritmos de árvore de decisão e regressão logística; e (3) Análise de Resultados, que será realizada uma análise do desempenho do modelo.

5.1 TECNOLOGIAS E FERRAMENTAS

Neste trabalho vamos utilizar a linguagem de programação Python (PYTHON, 2021) pela sua simplicidade de codificação e alto poder de processamento. Foi utilizado também a plataforma Google Colab Research (Colab, 2021), para o desenvolvimento dos algoritmos e treinamento dos modelos. Além disso, foi utilizado também o Jupyter Notebook (Jupyter, 2021), instalado localmente em computador pessoal, pois o Google Colab Research possui limitação de recursos de memória e processamento, então as duas plataformas foram usadas para realização deste trabalho. As principais bibliotecas utilizadas foram: numpy (NumPy, 2021), e pandas (Pandas, 2021), para a manusear os dados, matplotlib (Matplotlib, 2021), e seaborn (Seaborn, 2021), para a plotagem dos gráficos, e por fim a biblioteca sklearn (Scikit-Learn, 2021), que é a biblioteca responsável pelos algoritmos de aprendizado de máquina.

5.2 BASE DE DADOS

As bases de dados a serem utilizadas serão o SIM e SINASC, que são as duas principais fontes de informações sobre nascimentos e óbitos no Brasil. O SINASC é alimentado com base na Declaração de Nascido Vivo (Declaração de Nascido Vivo - DNV) (BELUZO et al.,2020b apud Oliveira et al., 2015). O SIM tem como objetivo principal apoiar a coleta, armazenamento e processo de gestão de registros de óbitos no Brasil (BELUZO et

25

al.,2020b apud Morais et al., 2017), e foi usado para rotular o óbito registrado no SIM, utilizando o campo DNV como chave de associação. O SIM será utilizado para rotular os registros do SINASC, pois o SIM coletas informações sobre mortalidade e é usado como base para o cálculo de estatísticas vitais, como a taxa de mortalidade neonatal e o SINASC reúne informações sobre dados demográficos e epidemiológicos do bebê, da mãe, do pré-natal e do parto (BELUZO et al.,2020b).

A Tabela 3 apresenta as variáveis da base de dados. Essa base tem variáveis que contêm valores quantitativos e categóricos. A coluna ?num_live_births? por exemplo contém o número de nascidos vivos anteriores da mãe e é composta por valores quantitativos contínuos, e a coluna ?tp_pregnancy? utiliza valores nominais categóricos que indicam o tipo de gravidez.

Tabela 3. Colunas da base de dados e suas descrições. (Fonte: Adaptado de BELUZO et al.,2020b).

Nome da coluna	Descrição	Domínio de dados
Variáveis demográficas e socioeconômicas		

maternal_age Idade da mãe Quantitativo Contínuo (inteiro)
tp_maternal_race Raça / cor da pele da
mãe
Nominal categórico (inteiro)
1 - branco; 2 - Preto; 3 -
amarelo; 4 - Pele morena; 5 -
Indígena.
tp_marital_status Estado civil da mãe Nominal categórico (inteiro)
1 - Único; 2 - Casado; 3 - Viúva;
4 - Separados / divorciados
judicialmente; 5 - Casamento
por união estável; 9 - Ignorado.
tp_maternal_schooling Anos de escolaridade da
mãe
Nominal categórico (inteiro)
1 - nenhum; 2 - de 1 a 3 anos; 3 -
de 4 a 7 anos; 4 - de 8 a 11 anos;
5 - 12 e mais; 9 - Ignorado.
Variáveis obstétricas maternas
num_live_births Número de nascidos
vivos
Quantitativo Contínuo (inteiro)
num_fetal_losses Número de perdas fetais Quantitativo Contínuo (inteiro)
num_previous_gestations Número de gestações Quantitativo Contínuo (inteiro)
26
anteriores
num_normal_labors Número de partos
normais (trabalhos de
parto)
Quantitativo Contínuo (inteiro)
num_cesarean_labors Número de partos
cesáreos (partos)
Quantitativo Contínuo (inteiro)
tp_pregnancy Tipo de gravidez Nominal categórico (inteiro)
1 - Singleton; 2 - Twin; 3 -
Trigêmeo ou mais; 9 - Ignorado.
Variáveis relacionadas a cuidados anteriores
tp_labor Tipo de parto infantil
(tipo de parto)
Categórico Nominal (inteiro)
1 - Vaginal; 2 - Cesariana;
num_prenatal_appointments Número de consultas de
pré-natal
Quantitativo Contínuo (inteiro)

tp_robson_group Classificação do grupo
Robson
Ordinal categórico (inteiro)
Variáveis relacionadas ao recém-nascido
tp_newborn_presentation Tipo de apresentação de
recém-nascido
Categórico Nominal (inteiro)
1 - Cefálico; 2 - Pélvico ou
culatra; 3 - Transversal; 9 -
Ignorado.
has_congenital_malformation Presença de
malformação congênita
Nominal categórico (inteiro)
1 - Sim; 2 - Não; 9 - Ignorado
newborn_weight Peso ao nascer em
gramas
Quantitativo Contínuo (inteiro)
cd_apgar1 Pontuação de Apgar de 1
minuto
Ordinal categórico (inteiro)
cd_apgar5 Pontuação de Apgar de 5
minuto
Ordinal categórico (inteiro)
gestaional_week Semana de gestação Quantitativo Contínuo (inteiro)
tp_childbirth_care Assistência ao parto Categórico Nominal (inteiro)
1 - Doutor; 2 - enfermeira ou
obstetra; 3 - Parteira; 4 - outros;
9 - Ignorado.
was_cesarean_before_labor Foi cesáreo antes do
parto.
27
was_labor_induced Foi induzido ao trabalho
de parto.
is_neonatal_death Morte antes de 28 dias
(rótulo)
Nominal categórico (inteiro)
0 - sobrevivente; 1 - morto.

Neste trabalho foi utilizado dados do período de 2014 à 2016 devido sua melhor qualidade. Este recorte possui 6.760.222 registros dos quais 99.4% são amostras de recém-nascidos vivos e 0.6% são amostras onde os recém-nascidos vieram a óbito conforme pode ser observado na Figura 4. Com essas informações consegue-se observar que a base de dados é desbalanceada.

Figura 4. Distribuição da base de dados considerando se o recém-nascido viveu ou morreu durante o período neonatal. (Fonte: Elaboração Própria).

5.3 PREPARAÇÃO DA BASE DE DADOS

Nessa etapa foi realizada a preparação da base de dados para o treinamento dos modelos que consiste na limpeza, transformação e preparação dos dados a serem utilizados na análise exploratória e na criação dos modelos preditivos. A base de dados pode conter informações desnecessárias para o modelo que pode acabar atrapalhando as predições,
28
algumas colunas podem ser retiradas ou seus valores podem ser modificados, por exemplo valores reais são modificados para inteiros.

5.4 DESENVOLVIMENTO DOS MODELOS DE APRENDIZADO DE MÁQUINA

Como já dito anteriormente na seção 4 os algoritmos de aprendizado de máquina escolhidos para esse trabalho foram os: Regressão Logística e Árvore de Decisão que utilizam a lógica mostrada na seção 4 , Fundamentação Teórica. Esses dois foram escolhidos por alguns motivos, ambos são algoritmos de classificação e supervisionados, atendendo os critérios necessários para a predição de mortalidade neonatal, ambos os algoritmos são bons para realizar predições, a Árvore de Decisão inclusive é um ótimo modelo para analisar as decisões que estão sendo feitas pelo modelo, afinal esse algoritmo tem a característica de ser um WhiteBox, ou seja conseguimos ver o que o modelo aprendeu e o que ele está decidindo. Esses algoritmos selecionados são algoritmos que se encaixam na necessidade deste trabalho e são muito utilizados na comunidade acadêmica.

Para o algoritmo de Regressão Logística, como foi dito anteriormente na seção ?Preparação da base de dados? a base de dados foi tratada e particionada, no particionamento foi utilizado 90% da base para o treino e 10% para os testes e então foi realizado o treinamento do modelo, também foi aplicada a função RFECV (Eliminação Recursiva de Feature com Validação Cruzada), esse função executa a RFE (Eliminação Recursiva de Feature), que tem como ideia selecionar features considerando recursivamente conjuntos cada vez menores de features, isso ocorre da seguinte forma. Primeiro, o estimador é treinado no conjunto inicial de features e a importância de cada feature é obtida por meio de um atributo "coef" ou por meio de um atributo "feature_importances". Em seguida, as features menos importantes são removidas do conjunto atual de features. Esse procedimento é repetido recursivamente no conjunto removido até que o número desejado de features a serem selecionados seja finalmente alcançado. Porém o diferencial da RFECV é que essa função executa a RFE em um loop de validação cruzada para encontrar o número ideal ou o melhor número de features.

Após essa seleção, foi realizado o treinamento do modelo usando todas as features e também treinamos usando as features que o RFECV selecionou, para compararmos os resultados no final. Também foi utilizado o método de K-folds cross-validation em um novo treinamento para conferirmos se o modelo estava sofrendo overfitting (sobreajuste). E por fim aplicamos um método chamado GridSearchCV, esse método permite que seja realizado testes
29
alterando algumas combinação de parâmetros nos nossos modelos, facilitando para achar a melhor combinação, esse método é parecido com o cross validation, o diferencial do Grid Search é que ele faz os cross validation de vários modelos com hiperparâmetros diferentes de uma só vez.

Para o algoritmo de Árvore de Decisão também foi utilizado o mesmo

particionamento de 90% para treino e 10% para teste, na criação do modelo foi colocado a entropy como critério de decisão e uma profundidade máxima de 4, pois com números maiores o modelo ficava muito complexo e ele errava mais as predições, também utilizamos um algoritmo para mostrar a importância de cada Feature para o modelo.
O algoritmo de Regressão Logística e análise exploratória foi baseado em códigos disponível na plataforma web Kaggle (MNASSRI, 2020), no exemplo do Kaggle o autor faz os códigos para predizer mortes no navio Titanic, para o trabalho de mortalidade neonatal os códigos foram alterados para realizar predição de mortes neonatais. Já no algoritmo de Árvore de Decisão foi baseado em uma vídeo aula do professor Diogo Cortiz (ÁRVORE..., 2020), nessa vídeo aula o professor explica sobre os algoritmos de Árvore de Decisão e depois implementa o um exemplo de código, o código usado neste trabalho usou o codigo do professor Diogo Cortiz, e foi alterado para realizar predições de mortalidade neonatal.

5.5 ANÁLISE DOS RESULTADOS

Para a análise de resultados foi utilizado dois métodos que é o de Matriz de confusão e a curva ROC esses métodos são explicados na seção 4, a Fundamentação Teórica. Com esses métodos pode-se realizar uma avaliação do modelo, e assim será possível analisar os valores de acurácia, precisão e recall do modelo, essas métricas são utilizadas avaliar o desempenho do modelo, com a acurácia é possível calcular a proximidade entre o valor obtido na predição dos modelos e os valores esperados, com a precisão é possível analisar as amostras que o modelo classificou como 0 eram realmente 0 na classe real e com o recall é possível analisar as amostras que o modelo conseguiu reconhecer como a classe correta.

5.6 ACESSO A BASE DE DADOS E CÓDIGOS

A base de dados utilizada neste projeto pode ser encontrada no link: ... . E os códigos podem ser encontrados no link: ... .

30

6 RESULTADOS

Os mesmos experimentos mostrados abaixo foram aplicados para uma base de dados que contém somente dados de São Paulo, essa base é um recorte da base do Brasil todo em contém cerca de 1.400.000 amostras, esse recorte da base também é bem desbalanceado. Os resultados obtidos nesse experimento usando o recorte de São Paulo foram bem parecidos com os experimentos da base do Brasil todo, e os códigos desse experimento podem ser visualizados no apêndice X.

6.1 ANÁLISE EXPLORATÓRIA

O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

6.1.1 Apresentação dos Dados

Como dito anteriormente na seção 5.2 ?Base de Dados? deste trabalho, a base de dados é formada pelas informações do SIM e do SINASC contém 6.760.222 amostras e 29 colunas, porém serão utilizadas somente 23 colunas sendo elas as colunas descritas na Tabela 3. Na seção 5.2 também tem a Figura 4 que mostra que a base dados é desbalanceada tendo 99.4% das amostras de recém-nascidos vivos e 0.6% das amostras onde os recém-nascidos vieram a óbito no período neonatal.

6.1.1.1 Caracteristicas demográficas e socioeconômicas maternas

O apêndice X contém uma tabela que possui as estatísticas sumarizadas das variáveis

demográficas e socioeconômicas maternas. Nesta tabela por exemplo podemos observar a coluna ?maternal_age? que contém a idade da mãe, e na tabela mostra que a média da idade das mães é de 26.4 anos, e os dados possuem uma variação de no mínimo 8 anos e no máximo 55 anos, e a mediana é 26 anos. Na tabela também tem as colunas: ?tp_maternal_schooling? que mostra a categoria de anos de escolaridade da mãe, ?tp_marital_status? que mostra o estado civil da mãe em dados categóricos e a coluna ?tp_maternal_race? que mostra a raça da mãe também em dados categóricos.
31
6.1.1.2 Variávies obstétricas maternas
No apêndice X pode-se observar uma tabela que possui as estatísticas sumarizadas das variáveis obstétricas maternas. Nesta tabela por exemplo temos a coluna ?num_live_births?, essa coluna mostra o número de nascidos vivos anteriores que a mãe obteve, a média desta coluna é 0.94 , a mediana é 1, o número mínimo é 0 e o máximo é 10 nascidos vivos. Na tabela podemos observar também a coluna, ?num_fetal_losses?, esta coluna mostra o número de perdas fetais anteriores da mãe, a média desta coluna é 0.21, a medina é 0, o número mínimo é 0 e o máximo é 5, esses valores mostram que poucas mães tiveram perdas fetais antes da gravidez atual. Também na tabela tem estatísticas da coluna ?num_previous_gestations? esta coluna mostra o número de gestações anteriores da mãe, esta coluna tem a média de 1.14, a mediana é 1, o número mínimo é 0 e o máximo 10. Essa tabela também contém as colunas: ?num_normal_labors? que é o número de partos normais da mãe, ?num_cesarean_labor? que mostra o número de partos cesáreos da mãe e por fim mostra a coluna ?tp_pregnancy? que é o tipo da gravidez.
6.1.1.3 Variávies de histórico de gravidez
No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis do histórico de gravidez. Nesta tabela temos as colunas ?num_prenatal_appointments? que mostra o número de consultas de pré-natal, a média dessa coluna é 7.8, a mediana é 8, a variação dos dados é de no mínimo 0 e no máximo 40. A tabela também contém as colunas: ?tp_labor? que mostra o tipo de parto, e também contém a coluna ?tp_robson_group? que mostra a classificação do grupo Robson.
6.1.1.4 Variávies relacionadas ao recém-nascido
No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis relacionadas ao recém-nascido. Nesta tabela por exemplo podemos observar a coluna ?newborn_weight? que armazena o peso ao nascer dos recém-nascidos em gramas, a média dos dados dessa coluna é 3187.3, a mediana é 3215, a variação dos dados é de no mínimo 0 e no máximo 6000 gramas. Também podemos observar a coluna ?gestacional_week? que mostra o número de semanas de gestação, a média é 38.4, a mediana é 39, a variação dos dados é de no mínimo 15 e no máximo 45 semanas. Além dessas colunas a tabela também
32
contém as colunas: ?cd_apgar1? que mostra a pontuação de Apgar de 1 minuto, ?cd_apgar5? que mostra a pontuação de Apgar de 5 minuto, ?has_congenital_malformation? que mostra se o recém-nascido contém a presença de alguma malformação, ?tp_newborn_presentation? que mostra o tipo de apresentação de recém-nascido, ?tp_labor? que mostra o tipo de parto, ?was_cesarean_before_labor? que mostra se o recém-nascido foi cesáreo antes do parto, ?was_labor_induced ? que mostra se o recém-nascido foi induzido ao trabalho de parto,

?tp_childbirth_care? que mostra se ouve assistência ao parto, e por fim temos a coluna ?is_neonatal_death? que mostra se o recém-nascido veio a óbito ou não.

6.1.2 Análise das variáveis

Nesta seção serão mostrados a parte de análise de variáveis com diferentes tipos de gráficos.

6.1.2.1 Variáveis contínuas

Na Figura 5 temos dois gráficos boxplot, no boxplot à esquerda temos a distribuição da variável peso ao nascer e nele é possível observar duas caixas onde a caixa azul são os vivos e o laranja são os mortos. Nas duas caixas mostradas no gráfico podemos observar pequenos círculos nas pontas, esses círculos são outliers (dados fora da curva), então para a caixa de vivos os outliers são recém-nascidos com pesos acima de 4500 gramas ou abaixo de 2000 gramas, e para a caixa de mortos os outliers são os recém-nascidos com mais de 5500 gramas. Para os vivos temos a mediana de aproximadamente 3200 gramas e para os mortos a mediana é 1300 gramas. Cada caixa representa 50% da base de dados, a caixa azul de vivos mostra que o peso de recém-nascidos que sobreviveram está aproximadamente entre 2700 a 3600 gramas, e para a caixa laranja de mortos o peso de recém-nascidos que vieram a óbito estão aproximadamente entre 700 a 2500 gramas. Então o gráfico mostra para nós que os recém-nascidos que têm maior risco de vir a óbito tem pesos menores que 2000 gramas.

33

Figura 5. BoxPlot das features Peso ao nascer e Idade da mãe. (Fonte: Elaboração Própria).

No gráfico à direita da Figura 5 temos um boxplot da distribuição da variável idade da mãe, como foi dito anteriormente a caixa azul são os vivos e o laranja são os mortos. Nesse boxplot então podemos ver que os outliers da caixa de vivos são mães com idade acima de 46 anos aproximadamente, para os mortos os outliers são mães com idade acima de 50 anos. Para os vivos temos a mediana de aproximadamente 26 anos e para os mortos a mediana é de 26 anos. Nos vivos a idade das mães está aproximadamente entre 21 a 32 anos, e para os mortos a idade das mães está aproximadamente entre 20 a 34 anos. É importante ressaltar que o gráfico mostra as duas caixas bem parecidas, então de acordo com os dados não contém diferença significativa entre vivos e mortos usando a idade da mãe para distribuir.

Já na Figura 6 temos um boxplot da distribuição da variável semanas de gestação.

Nesse boxplot então podemos ver que os outliers da caixa de vivos são gestações com mais de 44 semanas aproximadamente, e gestações com menos de 35 semanas. Para os vivos temos a mediana de aproximadamente 39 semanas e para os mortos a mediana é de 32 semanas. Nos vivos as semanas de gestação estão aproximadamente entre 36 a 40 semanas, e para os mortos a semanas de gestação estão aproximadamente entre 26 a 36 semanas. Nesse boxplot os dados mostraram que para as gestações que têm menos de 36 semanas os recém-nascidos têm maior risco de vir a óbito.

34

Figura 6. BoxPlot da feature semana de gestação. (Fonte: Elaboração Própria).

Na Figura 7 temos um histograma da distribuição de peso ao nascer por classe, podemos observar no gráfico que grande parte dos vivos (linha azul) estão distribuídos entre 2000 e 4000 gramas, a área entre esses valores tem muitos dados de vivos, já entre os valores 0 e 2000 gramas temos uma concentração dos dados de mortos.

Figura 7. Histograma de distribuição do peso entre as classes. (Fonte: Elaboração Própria)

6.1.2.2 Variáveis categoricas

Observando a Figura 8 podemos ver um gráfico de barras da distribuição da variável escolaridade da mãe, este gráfico mostra a contagem de registros por escolaridade da mãe, esse gráfico mostra para nós que a maior parte das mães estão na categoria 4 que representa de 8 a 11 anos de escolaridade.

35

Figura 8. Gráfico de barras da distribuição da variável Escolaridade da mãe. (Fonte: Elaboração Própria).

Observando a Figura 9 podemos ver um gráfico de barras da distribuição da variável número de vivos, este gráfico mostra a contagem de registros por número de vivos, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho.

Figura 9. Gráfico de barras da distribuição da variável Número de nascidos vivos. (Fonte: Elaboração Própria).

Observando a Figura 10 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 1 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 1 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem

36

apgar de 8 ou 9, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.

Figura 10. Gráfico de barras da distribuição da variável Pontuação Apgar de 1 minuto. (Fonte: Elaboração Própria).

Observando a Figura 11 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 5 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 5 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem apgar de 9 ou 10, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.

Figura 11. Gráfico de barras da distribuição da variável Pontuação Apgar de 5 minutos. (Fonte: Elaboração Própria).

37

Observando a Figura 12 podemos ver um gráfico de barras da distribuição da variável presença de malformação congênita, este gráfico mostra a contagem de registros por presença de malformação congênita, esse gráfico mostra para nós que a grande maioria dos dados estão na categoria 2 que mostra que o recém-nascido não possui malformação, esse grande volume para a categoria 2 é causada pelo desbalanceamento da base.

Figura 12. Gráfico de barras da distribuição da variável Presença de malformação congênita. (Fonte: Elaboração Própria).

Observando a Figura 13 podemos ver um gráfico de barras da distribuição da variável número de gestações, este gráfico mostra a contagem de registros por número de gestações, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho, da mesma forma mostrada na Figura 9.

38

Figura 13. Gráfico de barras da distribuição da variável Número de gestações. (Fonte: Elaboração Própria).

Observando a Figura 14 podemos ver um gráfico de barras da distribuição da variável assistência ao parto, este gráfico mostra a contagem de registros por assistência ao parto, esse gráfico mostra para nós que a grande maioria dos dados mostram que o parto teve assistência de uma Enfermeira ou obstetra ou essa informação foi ignorada na hora do resgistro.
Figura 14. Gráfico de barras da distribuição da variável Assistência ao parto. (Fonte: Elaboração Própria).
6.1.2.3 Analise bi-variada
A Figura 15 mostra para nós a importância da feature peso ao nascer com o gráfico mostrado é possível observar claramente a diferença de pesos entre vivos e mortos, recém-nascidos com peso acima de 2000 gramas sobreviveram, já os recém-nascidos com menos de 2000 gramas vieram a óbito.
39
Figura 15. Gráfico de densidade de peso entre as classes. (Fonte: Elaboração Própria).
Na Figura 16 podemos observar um gráfico de correlação entre algumas features de valores contínuos, algumas dessas features possuem correlações positivas, como pontuação apgar de 1 minuto e pontuação apgar de 5 minutos, essas colunas possuem uma correlação de 0.7, então nascidos que recebem um valor alto de apgar de 1 minuto tendem a receber um valor alto no apgar de 5 minutos. O gráfico também mostra outras features com correlações positivas como, número de partos normais e número de gestações anteriores que contém uma correlação de 0.81, número de gestações anteriores com idade da mãe que tem uma correção de 0.39, número de partos de cesáreos com idade da mãe que possui correlação de 0.24, número de partos normais com idade da mãe com a correlação de 0.26, semanas de gestação com peso ao nascer em gramas com correção de 0.52, e as apgar de 1 e 5 minutos com semanas de gestação. E grande parte das features, possuem correlações baixas então elas são inversamente correlacionadas pelo que o gráfico mostra, como por exemplo número de partos normais com número de consultas pré-natais possuem uma correção de -0.17, e número de partos cesáreos com número de partos normais que contém uma correlação de -0.15. Então a correlação dessas features estão bem baixas tirando as correlações acima de 0.7.
40
Figura 16. Gráfico de correlação. (Fonte: Elaboração Própria)
6.1.2.4 Distribuição por raça
Na Figura 17 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por raça, e podemos observar que a raça 5 (indigena), é o boxplot que contem a menor variação de idade, já as raças 1 (branco) e 3 (amarelo) são as caixas que possuem maior variação com a média em torno de 20 e 30 anos.
41
Figura 17. Distribuição da idade da mãe por raça. (Fonte: Elaboração Própria).
No boxplot mostrado na Figura 18 podemos observar a distribuição de peso ao nascer por raça, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre as raças para peso ao nascer.
Figura 18. Distribuição de peso ao nascer por raça. (Fonte: Elaboração Própria).
Na Figura 19 temos o boxplot da a distribuição de semanas de gestação por raça, e praticamente todas as caixas são iguais, somente a caixa da raça 1 (branco) possui menor variação que as outras.

42
Figura 19. Distribuição de semanas de gestação por raça. (Fonte: Elaboração Própria).
6.1.2.5 Distribuição por estado civil
A Figura 20 mostra o boxplot da a distribuição de peso ao nascer por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre os estados civis para peso ao nascer.
Figura 20. Distribuição de peso ao nascer por estado civil. (Fonte: Elaboração Própria).
A Figura 21 mostra o boxplot da a distribuição de semanas de gestação por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Porém os estados civis 2 (Casado) e 4 (Separados/divorciados judicialmente), contém variações menores.
43
Figura 21. Distribuição de semanas de gestação por estado civil. (Fonte: Elaboração Própria).
Na Figura 22 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por estado civil, e podemos observar que o estado civil 3 (Viúva), possui a maior variação, já o estado civil 4 (Separados/divorciados judicialmente) é a caixa que possuem menor variação
Figura 22. Distribuição da idade da mãe por estado civil. (Fonte: Elaboração Própria).
6.1.2.6 Distribuição por escolaridade da mãe
Na Figura 23 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por escolaridade da mãe, e podemos observar que a escolaridade 2 (1 a 3 anos) e 3 (4 a 7
44
anos), possui as maiores variações, já escolaridade 5 (12 ou mais anos) é a caixa que possuem menor variação
Figura 23. Distribuição da idade da mãe por escolaridade da mãe. (Fonte: Elaboração Própria).
A Figura 24 mostra o boxplot da a distribuição de peso ao nascer por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então, pouca variação é observada entre a escolaridade da mãe para peso ao nascer.
Figura 24. Distribuição de peso ao nascer por escolaridade da mãe. (Fonte: Elaboração Própria).
A Figura 25 mostra o boxplot da a distribuição de semanas de gestação por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas,
45
todas são praticamente iguais. Porém a escolaridade 5 (12 ou mais anos) contém variação menor que as outras.
Figura 25. Distribuição de semanas de gestação por escolaridade da mãe. (Fonte: Elaboração Própria).
6.2 ÁRVORE DE DECISÃO
A Figura 26 mostra um trecho do algoritmo utilizado, nesse trecho é realizado a divisão do DataFrame das features de entrada , e o DataFrame que contém o rótulo.Neste caso o rótulo será o DataFrame ?target?, esse Dataframe contém a feature

?is_neonatal_death?, essa feature informa se o recém nascido veio a óbito ou não, sendo o foco da predição que desejamos realizar essa feature entra como rótulo para o modelo, já no Dataframe ?nome_features? temos as features que vão servir de entrada para o modelo, é a partir dessas features que o modelo tentará encontrar padrões para predizer o risco de óbito do recém nascido. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

46

Figura 26. Separação dos DataFrames de entrada e rótulo. (Fonte: Elaboração Própria).

6.2.1 Modelo de Árvore de Decisão Utilizando Particionamento 90/10

O algoritmo de árvore de decisão foi treinado utilizando duas formas diferentes: usando as partições 90% para treino e 10% para os testes e utilizando o método K-folds cross-validation. E para os experimentos iniciais foi utilizado o particionamento 90/10.

Tabela 4. Resultados do modelo de árvore de decisão. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	671807
Morto(1)	0.71	0.29	0.41	4216

Analisando a Tabela 4 pode-se ver os seguintes resultados, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 71% que realmente pertenciam a classe 1, então cerca de 29% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 29% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem a classe 1 que seria dos recém nascidos que poderiam vir a óbito, já para a classe de vivos a

47

classe 0 o modelo está identificando muito bem e classificando de forma correta. A tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.

Esse modelo teve uma acurácia de 0.99 que é uma acurácia bem alta, porém somente pela acurácia não é possível dizer que o modelo está excelente pelo motivo da base que está sendo usada é altamente desbalanceada, então para a classe 0 que é a de vivos temos muito mais dados que para a classe de mortos, e o modelo está acertando bastante a classe de vivos logo a acurácia fica alta por esses acertos, enquanto para a classe de mortos o modelo não está acertando tanto.

A Figura 27 mostra a curva ROC do modelo, pode-se observar que a AUC da curva ROC ficou em 0.92, a AUC mostra que o modelo está acertando bastante as predições, pois quanto mais perto de 1 melhor está no nosso modelo, mas no caso desse modelo vimos acima que as predições para a classe de mortos(1) está ruim, o modelo está acertando poucas classificações como 1. Então a AUC assim como a acurácia está alta porque a base de dados utilizada é altamente desbalanceada tendo muito mais amostras de vivos(1), e como o modelo está bom para essa classificação e está acertando bastante os valores de AUC e acurácia ficam altos .

Figura 27. Curva ROC modelo de Árvore de decisão. (Fonte: Elaboração Própria).
A Figura 28 mostra a importância de cada feature para o modelo sendo assim a feature 10 - Peso ao nascer, a mais importante para o modelo, pois dela o modelo conseguiu tirar mais informações, seguido das features, 13 - Apgar dos 5 primeiros minutos, 11 - Semanas de 48
Gestação, 14 - Presença de malformação 12 - Apgar do 1 primeiro minutos. Então essas são as features que foram mais decisivas para a montagem da árvore e para as predições do modelo.
Figura 28. Gráfico de importância das features. (Fonte: Elaboração Própria).
Na Figura 29 temos uma imagem reduzida da árvore de decisão que foi montada com o treinamento utilizando 5 como profundidade máxima para a árvore, a imagem da árvore completa pode ser visualizada no apêndice A. É possível ver que as features marcadas como importante para o modelo apareceu diversas vezes na árvore, e a feature peso ao nascer foi escolhida para ser o nó raiz da árvore, de todas as features ela foi a que teve a melhor entropia para ser o nó raiz, e depois aparece mais vezes para outras decisões em outros níveis da árvore e isso faz com que essa feature se torne a mais importante para o modelo, também podemos ver que a feature ?Apgar do 1 primeiro minuto? mesmo sendo a menos importante para o modelo ela aparece somente no nó de número 42 isso mostra que a entropia e o ganho de informação dessa feature nas outras decisões não foram boas fazendo ela ser a feature com menos importância utilizada no modelo.
49
Figura 29. Árvore de decisão montada a partir do algoritmo. (Fonte: Elaboração Própria).
6.2.2 Modelo de Árvore de Decisão Utilizando K-folds cross-validation
Após realizar esse experimento com a partição 90/10, foi realizado outro experimento com esse mesmo algoritmo porém agora utilizando o método K-folds cross-validation para o treinamento. O K-folds cross-validation foi configurado para separar a base de dados em 10 partes iguais, e assim o algoritmo foi treinado novamente gerando um novo modelo. O método de K-folds cross-validation é utilizado para conferir se o modelo não está sofrendo problemas de overfitting, é importante garantir que o modelo está aprendendo com os dados e não está somente decorando.
Na Tabela 5 temos então o resultados desse modelo, pode-se observar que os resultados para a classificação de vivos(0) não teve alteração, o modelo continua tendo 100% na precisão e no recall, mas na classificação de mortos(1) o modelo teve uma diminuição na precisão que agora está em 65% e também teve uma diminuição no recall que agora está em 23%, porém essa diminuição é justificada porque para o teste deste modelo foi utilizado a base completa e não somente a partição de teste que contém 10% da base total. Logo os resultados não tiveram uma alteração drástica e tiveram um comportamento bem semelhante, incluindo a acurácia que se manteve em 0.99 e a AUC que teve uma diminuição pequena também e ficou com 0.89.
50
Tabela 5. Resultados do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	6719191

Morto(1) 0.65 0.23 0.34 41031
Na Tabela 6 é possível ver a matriz de confusão do modelo que mostra o número exato de erros e acertos do modelo, pode-se observar que para a classe de vivos o modelo classificou corretamente 6.714.117 amostras da base de dados completa, o modelo classificou como 0 as amostras que nos rótulos realmente eram 0, errou apenas 5074 amostras. Já para a classe de mortos o modelo acertou somente 9552 amostras, classificou como 1 as amostras que no rótulo eram 1 também, e errou 31479. Isso mostra que o modelo está muito desbalanceado, pois está errado muito mais do que acertando as predições de mortos, sendo isso um reflexo da base de dados desbalanceada também. E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.
Tabela 6. Matriz de confusão do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6714117 5074
Mortos (1) 31479 9552
6.3 REGRESSÃO LOGÍSTICA
Para o algoritmo de Regressão Logística também foi utilizado métodos diferentes para o treinamento, O algoritmo foi treinado utilizando três formas diferentes: usando as partições 90% para treino e 10% para os testes, utilizando o método K-folds cross-validation e foi treinada utilizando o método RFECV que seleciona as melhores features para o modelo, juntamente com o método GridSearchCV. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.
51
6.3.1 Modelo de Regressão Logística Utilizando Particionamento 90/10
Para o primeiro experimento do algoritmo de regressão logística foi utilizado o particionamento 90/10 para o treinamento do algoritmo, sendo 90% da base de dados para realizar o treinamento do modelo e 10% da base para realizar os testes.
Na Tabela 7 temos os resultados do modelo de regressão logística utilizando o método de particionamento 90/10, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 65% que realmente pertenciam a classe 1, então cerca de 35% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 24% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem as amostras da classe 1 que seria dos recém-nascidos que poderiam vir a óbito, já para a classe de vivos a classe 0 o modelo está identificando muito bem e classificando de forma correta.E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.

Tabela 7. Resultados do modelo de regressão logística usando particionamento 90/10.
(Fonte: Elaboração Própria)
Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 671885
Morto(1) 0.65 0.24 0.35 4138
A Figura 30 abaixo mostra a curva ROC do modelo de regressão logística usando o particionamento 90/10, analisando a curva ROC podemos ver que a AUC da curva ficou em 0.93 e a acurácia do modelo ficou 0.99, mostrando que o modelo está acertando bastante as predições, porém esses números para o nosso modelo não mostra que ele está ótimo, com a análise da Tabela 7 acima podemos observar que o modelo está acertando muitas predições da classe de vivos e poucas predições da classe de mortos, como a base de dados é
52
desbalanceada, como a maior quantidade de dados está na classe de vivos e o modelo está acerto quase todos dessa classe a acurácia e a AUC ficam altas por esses acertos.Logo é preciso analisar mais resultados além da acurácia e da AUC do modelo.
Figura 30. Curva ROC modelo regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria).
Na Tabela 8 podemos analisar os número exatos que o modelo acertou de cada classe, como é mostrado na matriz de confusão o modelo acertou 671.435 da classe de vivos e acertou apenas 915 da classe de mortos, e errou mais que o triplo da classe de mortos. Então as predições do modelo estão muito desbalanceadas, para a classe de vivos o modelo está predizendo bem, porém para as classes de mortos o modelo está errando muitas predições.
Tabela 8. Matriz de confusão do modelo de regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 671435 504
Mortos (1) 3169 915
6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation
Para o segundo experimento do algoritmo de regressão logística foi utilizado o método K-folds cross-validation para o treinamento do modelo com a intenção de conferir e confirmar
53
que o modelo não esteja tendo problemas com overfitting e realmente esteja aprendendo com os dados que está sendo usado para o treinamento. O método K-folds cross-validation foi configurado para realizar uma separação de 10 partes iguais.
Na Tabela 9 podemos observar os resultados do modelo, e os resultados foram parecidos com o experimento anterior usando o particionamento 90/10, única diferença nos resultado é que no experimento anterior a precisão da classe de mortos ficou em 65% e no caso desse experimento usando o K-folds cross-validation a precisão ficou em 64%, mas os valores de recall e F1-Score se mantiveram, assim como todos os resultados da classe de mortos se mantiveram em 100%. A execução desse experimento mostra que o modelo realmente está aprendendo com os dados e não está apenas decorando, retirando o risco do modelo estar com overfitting. Mesmo sem alterações no resultado é importante saber que o

modelo não está com overfitting e está conseguindo aprender e encontrar padrões nos dados.
Tabela 9. Resultados do modelo de regressão logística usando K-folds cross-validation.
(Fonte: Elaboração Própria)
Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 6719191
Morto(1) 0.64 0.24 0.35 41031
Já na Figura 31 temos a curva ROC do modelo, essa curva ROC mostra a AUC de
todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do
método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas
para todas as vezes que foi realizado os testes, a AUC se manteve em 0.93 e 0.94, sendo a
média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não
se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua
acertando muito a classe de vivos que é a classe predominante na base de dados. Já era
esperado os resultados não sofrerem alterações grandes, pois a base de dados não sofreu
nenhuma alteração, esse experimento foi somente para conferir se a modelo não estava
sofrendo overfitting, e pelos resultados aparentemente a base não está com problemas de
sobreajuste.
54
Figura 31. Curva ROC modelo regressão logística usando K-folds cross-validation. (Fonte:
Elaboração Própria).
Na Tabela 10 temos a matriz de confusão do modelo, onde mostra que o modelo
acertou 6.713.658 amostras da classe de vivos e errou apenas 5.533, já para a classe de mortos
o modelo acertou somente 9.847 e novamente errou mais que o triplo errando 31.182
amostras. Como os resultados mostrados anteriormente não tiveram alteração era de se
esperar que as predições corretas e incorretas do modelo também não sofressem alterações, o
modelo continua acertando muito a classe de vivos e errando muito a classe de mortos,
mantendo as predições bem desbalanceadas.
55
Tabela 10. Matriz de confusão do modelo de regressão logística usando K-folds
cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6713658 5533
Mortos (1) 31182 9847
6.3.3 Modelo de Regressão Logística Utilizando GridSearchCV e RFECV
Para o terceiro e último experimento de regressão logística foi utilizado técnicas
diferentes juntas para tentar melhorar as predições do modelo, para esse experimento usamos
o K-folds cross-validation para o particionamento da base de dados assim como foi utilizado
no experimento acima, e também foi utilizado duas técnicas que ainda não foram usadas nos
experimentos anteriores que são as funções RFECV e GridSearchCV. A função RFECV
seleciona a quantidade ideal de features para ser usadas no treinamento do modelo e também
mostra quais são as melhores features para essa predição, então essa função ela utiliza a
validação cruzada para ir treinando e testando N vezes o modelo, e sempre vai alterando a

quantidade e as features utilizadas para o teste, então cada vez que a função testa os modelos ela grava a pontuação dos acertos e assim depois mostra o gráfico da Figura 32 e assim é possível ver quais são as features ideais para o modelo.
56
Figura 32. Resultado da função RFECV Base Brasil. (Fonte: Elaboração Própria).
Então os resultados mostrados na Figura 32 apresenta que o número ideal para treinar o modelo de features é 6, e as features sao: 'tp_maternal_schooling', 'gestaional_week', 'cd_apgar1', 'cd_apgar5', 'has_congenital_malformation', 'tp_labor'. Essas são as features que tiveram o melhor desempenho no RFECV, então para o treinamento do modelo desse experimento as features que vão ser utilizadas serão somente essas 6.
Após a execução da função de RFECV foi executado a função de GridSearchCV, o GridSearchCV é utilizado para descobrir quais são os melhores parâmetros para utilizar no algoritmo de regressão logística e criar os modelos, foi utilizado também a validação cruzada para testar qual seria os melhores parâmetros para melhorar o desempenho do modelo. Na Figura 33 pode-se ver que os parâmetros escolhidos pela função foram:
Figura 33. Resultado da função GridSearchCV. (Fonte: Elaboração Própria).
Na Tabela 11 podemos observar os resultados do modelo, e os resultados foram parecidos com os experimentos anteriores, para a classe de mortos a precisão teve um aumento e foi para 69% e o recall diminuiu chegando em 20%, então utilizando as novas
57
funções o modelo ganhou precisão e perdeu recall. Para a classe de vivos o modelo se manteve em 100% e não teve alterações para os resultados de precisão e recall. Mesmo com a utilização das funções de RFECV e GridSearchCV o modelo não teve uma melhora significativa para as predições de mortos.
Tabela 11. Resultados do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	6719191
Morto(1)	0.69	0.20	0.31	41031

Na Figura 34 temos a curva ROC do modelo, essa curva ROC mostra a AUC de todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas para todas as vezes que foi realizado os testes, a AUC se manteve em 0.92 e 0.93, sendo a média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua acertando muito a classe de vivos que é a classe predominante na base de dados.
58
Figura 34. Curva ROC modelo regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria).
Na Tabela 12 temos a matriz de confusão do modelo, onde mostra que o modelo acertou 6.715.535 amostras da classe de vivos e errou apenas 3.656, comparando com os experimentos anteriores de regressão logística as predições de vivos teve uma pequena piora e teve uma pequena diminuição nos acertos. Já para a classe de mortos, o modelo acertou 8.296 e errou 32.735 amostras, logo as predições de vivos tiveram uma uma piora e acertou um

pouco menos do que os outros experimentos, e na classe de vivos teve uma melhora acertando mais predições, porém como o problema está nas predições de mortos e para essa classe teve uma piora pode-se dizer que a função de GridSearchCV não teve um ganho significativo nos resultados. Então como modelo final foi escolhido o modelo utilizando somente o K-folds cross-validation, pois foi o modelo que mostrou o melhor desempenho dos experimentos de regressão logística.

59

Tabela 12. Matriz de confusão do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

	Predito	
	Vivos (0)	Mortos (1)
Classe Real		
Vivos (0)	6715535	3656
Mortos (1)	32735	8296

60

7 CONCLUSÃO

O principal objetivo deste trabalho foi analisar os resultados das predições de mortalidade neonatal dos algoritmos de aprendizado de máquina usando uma base de dados com dados do Brasil inteiro. A partir destes resultados apresentados na secção anterior ?Resultados?, é possível concluir que analisar somente a AUC e a acurácia do modelo não é o suficiente para garantir que o modelo está excelente, sendo assim temos que ter mais atenção a precisão, recall e as predições mostradas na matriz de confusão.

Ambos os modelos ficaram desbalanceados nas predições, os modelos acertaram mais predições de vivos do que mortos, e as predições corretas de mortos foram mais baixas do que as incorretas, para os modelos ficarem melhor precisam passar por algoritmos que possam balancear essas predições, uma outra alternativa para melhorar o modelo é utilizar uma base de dados mais balanceada, pois essa base que foi utilizada é altamente desbalanceada.

Embora os dois modelos tenham tido resultados parecidos , o modelo de regressão logística utilizando somente o K-folds cross-validation se saiu melhor que os outros experimentos, o modelo criado na seção ?6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation? manteve a acurácia, AUC, precisão e recall dos demais modelos e acertou mais predições, talvez com algoritmos para melhorar as predições da classe de mortos, balanceando mais as predições, esse modelo fique com ótimas taxas preditivas.

Além disso podes-se afirmar que o peso ao nascer do recém nascido é um fator muito importante para as decisões dos modelos, ambos os modelos usaram muito essa informação para as predições, também podemos colocar, as colunas de presença de malformação e semana de gestação, como colunas que foram importantes para os modelos, essas causas podem ser evitadas se houver um acompanhamento médico com qualidade e eficiência.

61

REFERÊNCIAS

ÁRVORE de Decisão - Aula 3. Gravação de Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=ecYpXd4WREk&t=4089s. Acesso em: 1 jul. 2020.

BARRETO, J. O. M.; SOUZA, N. M.; CHAPMAN, E. Síntese de evidências para políticas de saúde: reduzindo a mortalidade perinatal. Ministério da Saúde, p. 1?44,2015.

BATISTA, André Filipe de Moraes; FILHO, Alexandre Dias Porto Chiavegatto. Machine Learning aplicado à Saúde. In: ZIVIANI, Artur; FERNANDES, Natalia Castro; SAADE, Débora Christina Muchaluat. SBCAS 2019: 19 Simpósio Brasileiro de Computação Aplicada à Saúde. [S. l.: s. n.], 2019. cap. Capítulo 1.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. Towards neonatal mortality risk classification: A data-driven approach using neonatal, maternal, and social factors, [s. l.], 23 out. 2020.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. SPNeoDeath: A demographic and epidemiological dataset having infant, mother, prenatal care and childbirth data related to births and neonatal deaths in São Paulo city Brazil ? 2012?2018, [s. l.], 19 jul. 2020.
BRESAN, Rodrigo. Um método para a detecção de ataques de apresentação em sistemas de reconhecimento facial através de propriedades intrínsecas e Deep Learning. Orientador: Me. Carlos Eduardo Beluzo. 2018. Trabalho de Conclusão de Curso (Superior de Tecnologia em Análise e Desenvolvimento de Sistemas) - Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas., [S. l.], 2018.
CABRAL, Cleidy Isolete Silva. Aplicação do Modelo de Regressão Logística num Estudo de Mercado. Orientador: João José Ferreira Gomes. 2013. Projeto Mestrado em Matemática Aplicada à Economia e à Gestão (Mestrado) - Universidade de Lisboa Faculdade de Ciências Departamento de Estatística e Investigação Operacional, [S. l.], 2013. Disponível em: https://repositorio.ul.pt/bitstream/10451/10671/1/ulfc106455_tm_Cleidy_Cabral.pdf. Acesso em: 1 maio 2021.
CAMPOS, Raphael. Árvores de Decisão: Então diga-me, como construo uma?. [S. l.], 28 nov. 2017. Disponível em: https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69. Acesso em: 23 jan. 2021.
Colab. 2021. Disponível em: https://colab.research.google.com/. Acesso em: 20 mar. 2021.
COX, D.R.;SNELL, E.J. Analysis of Binary Data.London: Chapman & Hall, 2ª Edição, 1989. HOSMER, D.W.; LEMESHOW, S.Applied Logistic Regression. New York: John Wiley, 2ª Edição, 2000.
E.França, S.Lansky. Mortalidade infantil neonatal no brasil: situação, tendências e perspectivas Rede Interagencial de Informações para Saúde - demografia e saúde: contribuição para análise de situação e tendências Série Informe de Situação e Tendências(2009), pp.83-112.
62
FREIRE, Sergio Miranda. Bioestatística Básica: Regressão Linear. [S. l.], 20 out. 2020. Disponível em: http://www.lampada.uerj.br/arquivosdb/_book/bioestatisticaBasica.html. Acesso em: 23 jan. 2021.
Jupyter. 2021. Disponível em: https://jupyter.org/. Acesso em: 20 jun. 2021.
GOODFELLOW, I. et al. Deep learning. [S.l.]: MIT press Cambridge, 2016.
Pandas. 2021. Disponível em: https://pandas.pydata.org/. Acesso em: 20 jun. 2021.
Python. 2021. Disponível em: https://www.python.org/. Acesso em: 20 jun. 2021.

KUHN, M.; JOHNSON, K. Applied predictive modeling. [S.l.]: Springer, 2013. v. 26.
LANSKY, S. et al. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante . Cadernos de Saúde Pública, Scielo, v. 30, p. S192 ? S207, 00 2014.
LANSKY, Sônia; FRICHE, Amélia Augusta de Lima; SILVA, Antônio Augusto Moura; CAMPOS, Deise; BITTENCOURT, Sonia Duarte de Azevedo; CARVALHO, Márcia Lazaro; FRIAS, Paulo Germano; CAVALCANTE, Rejane Silva; CUNHA, Antonio José Ledo Alves. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante e ao recém-nascido. [s. l.], Ago 2014. Disponivel em: https://www.scielosp.org/article/csp/2014.v30suppl1/S192-S207/pt/
Matplotlib. 2021. Disponível em: https://matplotlib.org/. Acesso em: 20 mar. 2021.
Maranhão AGK, Vasconcelos AMN, Trindade CM, Victora CG, Rabello Neto DL, Porto D,et al. Mortalidade infantil no Brasil: tendências, componentes e causas de morte no período de 2000 a 2010 . In: Departamento de Análise de Situação de Saúde, Secretaria de Vigilância em Saúde, Ministério da Saúde, organizador. Saúde Brasil 2011: uma análise da situação de saúde e a vigilância da saúde da mulher. v. 1. Brasília: Ministério da Saúde; 2012. p. 163-82.
MESQUITA, PAULO. UM MODELO DE REGRESSÃO LOGÍSTICA PARA AVALIAÇÃO DOS PROGRAMAS DE PÕS-GRADUAÇÃO NO BRASIL. Orientador: Prof. RODRIGO TAVARES NOGUEIRA. 2014. Dissertação (Mestre em Engenharia de Produção) - Universidade Estadual do Norte Fluminense, [S. l.], 2014.
MNASSRI , Baligh. Titanic: logistic regression with python. [S. l.], 1 ago. 2020. Disponível em: https://www.kaggle.com/mnassrib/titanic-logistic-regression-with-python. Acesso em: 14 jan. 2021.
Morais, R.M.d., Costa, A.L: Uma avaliação do sistema de informações sobre mortalidade. Saúde em Debate 41, 101?117 (2017). Disponível em: https://doi.org/10.1109/SIBGRAPI.2012.38.
MOTA, Caio Augusto de Souza; BELUZO , Carlos Eduardo; TRABUCO, Lavinia Pedrosa; SOUZA , Adriano; ALVES, Luciana; CARVALHO, Tiago. DESENVOLVIMENTO DE UMA PLATAFORMA WEB PARA CIÊNCIA DE DADOS APLICADA À SAÚDE MATERNO INFANTIL , [s. l.], 28 nov. 2019.
MULTIEDRO (Brasil). Conheça as aplicações do machine learning na área da saúde. [S. l.], 28 nov. 2019. Disponível em: https://blog.multiedro.com.br/machine-learning-na-area-da-saude/#:~:text=O%20machine%263
0learning%20na%20%C3%A1rea,diagn%C3%B3sticos%20e%20tratamentos%20mais%20precisos. Acesso em: 13 jun. 2021.
MUKHIYA,S.K.;AHMED,U. Hands-On Exploratory Data Analysis with Python: Perform EDA Techniques to Understand, Summarize, and Investigate Your Data. [S.l.]: PacktPublishingLtd,2020.ISBN978-1-78953-725-3.22,32
Murray CJ, Laakso T, Shibuya K, Hill K, Lopez AD. Can we achieve Millennium Development Goal 4? New analysis of country trends and forecasts of under-5 mortality to 2015. Lancet 2007; 370:1040-54.
NumPy. 2021. Disponível em:https://numpy.org/. Acesso em: 20 jun. 2021.
Oliveira, M.M.d., de Araújo, A.S.S.C., Santiago, D.G., Oliveira, J.a.C.G.d., Carvalho, M.D.,

de Lyra et al, R.N.D.: Evaluation of the national information system on live births in brazil , 2006-2010. Epidemiol. Serv. Saúde 24(4), 629?640 (2015).
PELISSARI, ANA CAROLINA CHEBEL. MORTALIDADE NEONATAL DA CIDADE DE SÃO PAULO: UMA ABORDAGEM UTILIZANDO APRENDIZADO DE MÁQUINA SUPERVISIONADO. 2021. Trabalho de Conclusão de Curso (Superior) - Instituto Federal de Educação, Ciência e Tecnologia Campus Campinas, [S. l.], 2021.
REGRESSÃO logística e binary cross entropy - Aula 7. Direção: Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=3J-LBtHVsm4. Acesso em: 21 jan. 2021.
Rmd Nascimento, AJM Leite, NMGSd Almeida, Pcd Almeida, Cfd Silva Determinantes da mortalidade neonatal: estudo caso-controle em fortaleza, ceará, brasil Cad Saúde Pública, 28(2012), pp.559 - 572.
SANTOS, Hellen Geremias. Machine learning e análise preditiva: considerações metodológicas: métodos lineares para classificação. In: SANTOS, Hellen Geremias. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina. 2018. Tese (Pós-Graduação em Epidemiologia) - Faculdade de Saúde Pública da Universidade de São Paulo, [S. l.], 2018.
Scikit-learn. 2021. Disponível em: https://scikit-learn.org/stable/. Acesso em: 20 jun. 2021.
Seaborn. 2021. Disponível em: https://seaborn.pydata.org/. Acesso em: 20 jun. 2021.
SILVA, Wesley; CORSO, Jansen; WELGACZ, Hanna; PEIXE, Julinês. AVALIAÇÃO DA ESCOLHA DE UM FORNECEDOR SOB CONDIÇÃO DE RISCOS A PARTIR DO MÉTODO DE ÁRVORE DE DECISÃO. ARTIGO ? MÉTODOS QUANTITATIVOS, [s. l.], 12 ago. 2008.
WHO, W. H. O.Women and health: today?s evidence, tomorrow?s agenda. [S.l.]: UNWorld Health Organization, 2009. ISBN 9789241563857.
64
APÊNDICE A

==========

Arquivo 1: monografia-V7.docx.pdf (9728 termos)
Arquivo 2: https://minerandodados.com.br/cross-validation-com-python (743 termos)
Termos comuns: 34
Similaridade: 0,32%
O texto abaixo é o conteúdo do documento monografia-V7.docx.pdf (9728 termos)
Os termos em vermelho foram encontrados no documento https://minerandodados.com.br/cross-validation-com-python (743 termos)

==========

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO
CÂMPUS CAMPINAS
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
CAMPINAS
2021

CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS

Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas.
Orientador: Prof. Me. Everton Josué da Silva.
CAMPINAS
2021

Dados Internacionais de Catalogação na Publicação

CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS

Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo Câmpus Campinas.
Aprovado pela banca examinadora em: _____ de _____ de _____.
BANCA EXAMINADORA

_____
Prof. Me. Everton Josué da Silva (orientador)

IFSP Câmpus Campinas

_____

Prof. Me. Carlos Eduardo Beluzo
IFSP Câmpus Campinas

_____

Prof. Dr. Ricardo Barz Sovat
IFSP Câmpus Campinas

Dedico este trabalho aos meus familiares,
colegas de classe, professores e servidores do Instituto
que colaboraram em minha jornada formativa.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pelo dom da vida e pela oportunidade de concluir mais uma etapa de minha experiência acadêmica.

Agradeço a todos os professores e servidores do IFSP Câmpus Campinas, que contribuíram direta e indiretamente para a conclusão deste trabalho.

Agradeço também à minha família, que deu todo o apoio necessário para que eu chegasse até aqui.

Agradeço ao meu orientador que me auxiliou a solucionar as dificuldades encontradas no caminho.

"Minha energia é o desafio, minha motivação é o impossível,
e é por isso que eu preciso
ser, à força e a esmo, inabalável.".
Augusto Branco

RESUMO

A mortalidade infantil pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível em uma população. O aprendizado de máquina é uma área da inteligência artificial que propõe métodos de análise de dados que automatizam a construção de modelos analíticos com base em reconhecimento de padrões, baseado no conceito de que sistemas podem aprender com dados, identificando padrões e tomando decisões com o mínimo de intervenção humana. O objetivo deste trabalho é testar e analisar dois tipos diferentes de algoritmos de aprendizado de máquina, utilizando a base de dados do SIM e SINASC do Brasil, do período de 2016 até 2018, para gerar modelos para predição de mortalidade neonatal. Os métodos utilizados foram Árvore de Decisão e de Regressão Logística e como métricas para avaliar os modelos resultantes foram utilizadas a AUC, Curva ROC e a Matriz de Confusão. Como resultado, apesar de terem alcançado valores de AUC 0.93, ambos modelos de predições acertaram muitas predições de vivos cerca de 671.000 e erraram muitas predições de mortos cerca de 2.600, principalmente devido ao fato de a base estar desbalanceada.

Palavras-chave: Mortalidade Neonatal, Predição, Aprendizado de Máquina.

ABSTRACT

Infant mortality can be used to analyze poverty and socioeconomic levels, as well as measure the quality of health and medical technology available in a population. Machine learning is an area of artificial intelligence that proposes data analysis methods that automate the

construction of analytical models based on pattern recognition, based on the concept that systems can learn from data, identifying patterns and making decisions with a minimum of human intervention. The objective of this work is to test and analyze two different types of machine learning algorithms, using the SIM and SINASC do Brasil database, from 2016 to 2018, to generate models for predicting neonatal mortality. The methods used were Decision Tree and Logistic Regression and as metrics to evaluate the resulting models, AUC, ROC Curve and Confusion Matrix were used. As a result, despite achieving values of AUC 0.93, both prediction models correct many live birth predictions around 671.000 and miss many stillbirth predictions around 2600, mainly due to the fact that the base is unbalanced.
Keywords: Neonatal Mortality, Prediction, Machine Learning.
LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS
SIM Sistema de Informação sobre Mortalidade
SINASC Sistema de Informações sobre Nascidos Vivos
TMI Taxa de Mortalidade Infantil
TMN Taxa de Mortalidade Neonatal
MN Mortalidade Neonatal
ML Machine Learning
DNV Declaração de Nascido Vivo
VN Verdadeiro Negativo
FN Falso Negativo
VP Verdadeiro Positivo
FP Falso Positivo
ROC Curva Característica de Operação do Receptor
NaN Não é um Número
RFE Eliminação Recursiva de Feature
RFECV Eliminação Recursiva de Feature com Validação Cruzada
SUMÁRIO

13

1 INTRODUÇÃO

A taxa de mortalidade infantil (TMI) é uma importante medida de saúde em uma população e é um grande problema no mundo todo. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível. Esta taxa é dividida em duas categorias: neonatal e pós-neonatal. É categorizada neonatal quando o óbito ocorre nos 28 primeiros dias de vida após o pós-parto, e o pós-neonatal é quando o óbito ocorre entre 29 e 364 dias de vida (BELUZO et al., 2020). Cerca de 46% das mortes no mundo acontecem entre os menores de cinco anos e grande parte está concentrada nos primeiros dias de vida (LANSKY, 2014). A diminuição dessas taxas é muito importante no mundo todo, refletindo nos indicadores de saúde pública e desenvolvimento do país.

Os fatores associados à mortalidade são profundamente influenciados pelas características biológicas maternas e neonatais, pelas condições sociais e pelos cuidados prestados pelos serviços de saúde (E. FRANÇA; S. LANSKY, 2009; R.M.D. NASCIMENTO, 2012). Em grande parte, o diagnóstico é altamente dependente da experiência adquirida pelo

profissional que o realiza. Apesar de indiscutivelmente necessário, não é perfeito, principalmente pelo fato de ser dependente de fatores humanos, por este motivo os profissionais sempre tiveram recursos tecnológicos para auxiliar nessa tarefa (BELUZO et al., 2020). Segundo estudos de 2010 no Brasil, calcula-se que aproximadamente 70% dos óbitos infantis ocorridos poderiam ter sido evitados por ações de saúde e que 60% dos óbitos neonatais ocorreram por situações que poderiam ser evitadas (BARRETO; SOUZA; CHAPMAN, 2015).

No presente trabalho foram utilizados dois algoritmos de aprendizado de máquina para criar modelos de predição de mortalidade neonatal. Além disso, foi realizada também uma análise exploratória da base de dados. Para este trabalho foram utilizadas duas fontes de dados: Sistema de Informação sobre Mortalidade (SIM) e Sistema de Informações sobre Nascidos Vivos (SINASC) do Brasil inteiro.

14

2 JUSTIFICATIVA

Com o avanço da tecnologia, podemos notar que ela está cada vez mais presente no nosso dia a dia e nos auxilia em diversas tarefas do cotidiano, e vem sendo aplicada em diversas áreas, dentre elas a saúde.

A mortalidade infantil é uma preocupação mundial na saúde pública, a ONU (Organizações das Nações Unidas) definiu a redução da mortalidade infantil como uma meta para o desenvolvimento global. A mortalidade neonatal é responsável por aproximadamente 60% da TMI nos países em desenvolvimento (A.K. SINGHA, 2016). Existem diversos fatores que podem contribuir com a redução de óbitos neonatais, como por exemplo acompanhamento médico à gestante no período de gravidez, acompanhamento médico ao recém-nascido nos primeiros dias de vida e a disponibilização deste serviço com qualidade e proficiência (WHO, 2009).

A diminuição dessa taxa é importante e de interesse para o mundo todo, e utilizar da tecnologia para auxiliar nessa diminuição é uma proposta funcional, e inovadora para a realidade brasileira (MOTA et al., 2019). Havendo disponibilidade de tecnologia para auxiliar nas decisões dos diagnósticos, os profissionais da saúde podem focar nos cuidados do recém-nascido com risco de vir a óbito, fornecendo tratamentos melhores.

Segundo o site Multiedro (2019), um grande problema que os médicos enfrentam ao realizar o diagnóstico, é analisar um grande volume de dados. Para a área médica obter diagnósticos rápidos e precisos pode salvar vidas, e a utilização de machine learning para realizar esses processos é importante, com a ML os dados podem ser processados em questões de segundos e resultar em um diagnóstico mais rápido, além disso utilizando bons modelos ML os diagnósticos serão mais precisos.

Diversos trabalhos vêm sendo desenvolvidos neste contexto. No trabalho realizado por BELUZO et al. (2020a), os autores utilizaram uma base de dados com informações da cidade de São Paulo para criação de modelos de aprendizado de máquina para predição de mortalidade neonatal. Este recorte foi utilizado também no presente trabalho para fins de exercício e definição de etapas da implementação de código. Na conclusão os autores discutem os fatores que tiveram mais influência nas predições, e na análise feita pelos autores, as consultas de pré-natal são muito importantes para a redução da mortalidade infantil, uma vez que algumas características muito importantes, como a malformação congênita, podem

ser detectadas ao longo das consultas de pré-natal.
Em um outro estudo realizado por BELUZO et al. (2020a), foi realizada uma análise exploratória da base de dados SPNeoDeath, onde e podemos observar detalhadamente cada
15
análise das colunas da tabela e diversos gráficos feito para um melhor entendimento dos dados.
Outro trabalho que foi realizado utilizando a base de dados com dados somente de São Paulo foi o de PELISSARI (2021), nesse trabalho é realizado uma análise exploratória na base de dados de São Paulo e é também analisado três algoritmos de aprendizado de máquina (Logistic Regression, Random Forest Classifier e XGBoost), nesse trabalho a autora conclui que os modelos de Logistic Regression e XGBoost foram os modelos que apresentaram os melhores desempenhos preditivos, e também mostra os problemas de estar utilizando uma base de dados desbalanceada.
O problema da mortalidade neonatal não é recente, e o mundo todo desenvolve iniciativas para lidar com ele. A ONU definiu como meta a redução da mortalidade infantil para o desenvolvimento global. Diversos fatores podem ajudar na diminuição da taxa de mortalidade neonatal como acompanhamento médico antes e após parto, tanto com a mãe quanto com o recém-nascido. Um serviço de qualidade e constante para os casos com maior risco de óbito pode salvar diversos recém-nascidos. Neste sentido, o desenvolvimento de ferramentas para a predição de mortalidade neonatal pode colaborar com esta iniciativa.

3 OBJETIVOS

3.1 OBJETIVO GERAL

O presente trabalho tem como objetivo testar e analisar os resultados da aplicação dos aprendizagem de máquina supervisionado Árvore de Decisão e Regressão Logística, utilizando dados do SIM e do SINASC, no período de 2016 até 2018, a fim de gerar modelos de predição de risco morte neonatal.

3.2 OBJETIVOS ESPECÍFICOS

? Realizar análise exploratória da base dados a ser utilizada na construção dos modelos;
? Implementar modelos de predição utilizando algoritmos Árvore de Decisão e Regressão Logística;
? Avaliar resultados e propor novas abordagens.
16

4 FUNDAMENTAÇÃO TEÓRICA

4.1 MORTALIDADE INFANTIL E NEONATAL

Como dito na monografia PELISSARI (2021), a TMI consiste no número de crianças que foram a óbito antes de concluir um ano de vida dividido por 1000 crianças nascidas vivas no tempo de um ano. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos e qualidade da saúde pública de um país (apud BELUZO et al., 2020).
Mencionado em PELISSARI (2021), a TMN é uma das duas categorias da TMI, é categorizado mortalidade neonatal quando o óbito ocorre do nascimento da criança até os 28 primeiros dias de vida. A mortalidade neonatal pode ser subdividida em mortalidade neonatal precoce que é quando o óbito ocorre entre 0 e 6 dias de vida, e o neonatal tardio quando o óbito ocorre entre 7 e 28 dias de vida (apud E. FRANÇA e S. LANSKY, 2009).
Segundo Lansky et al. (2014), a taxa de mortalidade neonatal é o principal

componente da TMI desde a década de 1990 no Brasil e vem se mantendo em níveis elevados, com taxa de 11,2 óbitos por mil nascidos vivos em 2010 (apud Maranhão et al., 2012). Também é dito em Lansky et al. (2014), que a TMI do Brasil em 2011 foi 15,3 por mil nascidos vivos, alcançando a meta 4 dos objetivos de desenvolvimento do milênio, compromisso dos governos integrantes das Nações Unidas de melhorar a saúde infantil e reduzir em 2/3 a mortalidade infantil entre 1990 e 2015. (apud Maranhão et al., 2012; apud Murray et al., 2015). Segundo Lansky et al. (2014) o principal componente da TMI em 2009 é a mortalidade neonatal precoce, e grande parte das mortes infantis acontece nas primeiras 24 horas (25%), indicando uma relação estreita com a atenção ao parto e nascimento (apud E. FRANÇA e S. LANSKY, 2009).

4.2 MÉTODOS DE APRENDIZADO DE MÁQUINA

A utilização de dados para a resolução de problemas, de modo a se identificar padrões ocultos e posteriormente auxiliar na tomada de decisões é definida como aprendizado de máquina (BRESAN, 2018 apud Brink et al. 2013).

O uso de técnicas de Aprendizado de Máquina se mostra altamente eficiente na resolução de tarefas que se apresentem difíceis de se resolver a partir de programas escritos por humanos (BRESAN, 2018 apud GOODFELLOW et al., 2016), bem como também possibilita uma maior compreensão sobre os funcionamentos dos princípios que compõem a inteligência humana.

17

Neste trabalho serão implementados modelos utilizando os algoritmos de Árvore de Decisão e Regressão Logística, os quais serão apresentados a seguir.

4.2.1 Árvore de Decisão

Segundo Batista e Filho (2019), os modelos preditivos baseados em árvores são geralmente utilizados para tarefas de classificação, embora também possam ser utilizados para tarefas de regressão. Considerado um dos mais populares algoritmos de predição, o algoritmo de árvore de decisão proporciona uma grande facilidade de interpretação.

As árvores de decisão utilizam a estratégia ?dividir-e-conquistar", na qual as árvores são construídas utilizando-se apenas alguns atributos. A árvores de decisão é uma das técnicas por meio da qual um problema complexo é decomposto em subproblemas mais simples. Recursivamente, a mesma estratégia é aplicada a cada subproblema (SILVA et al, 2008).

A árvore de decisões tem uma estrutura hierárquica onde cada nó da árvore representa uma decisão em uma das colunas do conjunto de dados, cada ramo da árvore representa uma tomada de decisão (BATISTA; FILHO, 2019).

O algoritmo segue algumas decisões para montar a árvore, o atributo mais importante é utilizado como nó raiz e será o primeiro nó, já os atributos menos importantes são mostrados nos ramos da árvore. Cada atributo é verificado para conferir se é possível gerar uma árvore menor e se é possível fazer uma classificação melhor, e assim é escolhido o nó que produz nós filhos (SILVA et al, 2008).

Existem diferentes tipos de algoritmo para a construção da árvore, como por exemplo, ID3 (Iterative Dichotomiser), C4.5 e CART (Classification and Regression Tree). O algoritmo ID3 escolhe os nós da árvore por meio da métrica do ganho de informação. Já o CART faz uso da equação de Gini (BATISTA; FILHO, 2019 apud KUHN; JOHNSON, 2013).

O algoritmo utilizado neste trabalho usará o cálculo da entropia e ganho de

informação, para decidir qual atributo será usado no nó, a entropia é uma medida da desorganização dos sistemas, maior é a incerteza do sistema, quanto maior a entropia maior está a desorganização, com a árvore de decisão a ideia é ir organizando o sistema e ir separando as decisões da melhor forma para que a entropia diminui e o ganho de informação aumente, para que a decisão do modelo fiquei mais assertiva. O cálculo da entropia utiliza a seguinte equação (ÁRVORE, 2020):

???????? =

?

?? ?? ???

2

??

18

Nesta equação o i representa possíveis classificações (rótulos), e o P é a probabilidade de cada uma. A ideia da árvore então é verificar qual é a entropia para cada uma das variáveis da base de dados, e verificar qual é a melhor organização de cada uma das variáveis. E isso é realizado através da técnica chamada de ganho de informação, essa técnica utiliza a equação a seguir (ÁRVORE, 2020):

????? = ????????(???) ? ? ????(??????) * ????????(??????)

Então o ganho de informação é calculado pela entropia do nó pai menos a soma do peso vezes a entropia dos nós filhos. Esse cálculo é realizado para todas as variáveis e a variável que tiver maior ganho de informação é utilizado para a decisão do nó. Esse processo é realizado recursivamente e a árvore vai sendo montada com base nesses cálculos (ÁRVORE, 2020).

Na Figura 1 temos a representação do início de uma árvore de decisão, no caso em questão o autor estava classificando exames positivos e negativos para o vírus SARS-CoV-2, ele utilizou uma base de dados que contém quase 6 mil exames realizados, contendo campos como o resultado do exame, idade do paciente, níveis de hemoglobina entre outras informações.

Figura 1. Exemplo de Árvore de Decisão. (Fonte: ÁRVORE 2020).

Pode-se observar que para o nó raiz foi utilizado a variável Leukocytes, logo foi o que apresentou a melhor organização para ser o nó raiz, e após ele temos os ramos da árvore. Cada retângulo da árvore é uma decisão do modelo e esses retângulos têm atributos importante como, a coluna da base de dados que será onde vai ser realizado a decisão, então pegando o nó raiz, pacientes que tiverem com os Leukocytes para menos que -0.43 seguiram o caminho

19

para a esquerda (true), já os pacientes que tiverem mais seguiram o caminho da direita (false). A entropia calculada do nó também é apresentada e a classificação do nó também, então se o resultado parar neste nó, a classificação que está nele será a classificação final, e temos o samples que é o número de amostras que se enquadram no na decisão. E esse processo de verificação do modelo vai se repetindo até não ter mais como dividir em subproblemas ou até chegar na profundidade máxima.

Uma característica importante da árvore de decisão é que ela é um modelo WhiteBox, ou seja, é possível visualizar a árvore montada e as decisões que o modelo terá que tomar na árvore. Na Figura 1 pode-se ver uma árvore de decisões montada, cada retângulo da árvore é

uma decisão que o modelo terá que tomar, conforme o modelo toma as decisões ele vai seguindo um caminho até chegar ao nó folha, nó folha é o nó que não contém nó filho, não tem mais ramos para baixo, chegando ao nó folha o modelo tem a classificação final.

4.2.2 Regressão Logística

O modelo de regressão logística estabelece uma relação entre a probabilidade de ocorrência da variável de interesse e as variáveis de entrado do modelo, sendo utilizada para tarefas de classificação, em que a variável de interesse é categórica, neste caso, a variável de interesse assume valores 0 ou 1, onde 1 é geralmente utilizado para indicar a ocorrência do evento de interesse (Batista e Filho, 2019).

De acordo com Mesquita (2014), a técnica de regressão logística, desenvolvida no século XIX, obteve maior visibilidade após 1950 ficando então mais conhecida. Ficou ainda mais difundida a partir dos trabalhos de Cox & Snell (1989) e Hosmer & Lemeshow (2000). Caracteriza-se por descrever a relação entre uma variável dependente qualitativa binária, associada a um conjunto de variáveis independentes qualitativas ou métricas.

Esta técnica inicialmente foi utilizada na área médica, porém a eficiência viabilizou sua implementação nas mais diversas áreas. Mesquita (2014) diz que o termo regressão logístico, tem sua origem na transformação usada com a variável dependente, que permite calcular diretamente a probabilidade da ocorrência do fenômeno em estudo.

Segundo Santos (2018), para estimar a probabilidade, é utilizado uma técnica chamada distribuição binomial para modelar a variável resposta do conjunto de treinamento, que tem como parâmetro a probabilidade de ocorrência de uma classe específica. Uma característica importante desse modelo é que a probabilidade estimada deve estar limitada ao intervalo de 0 a 1. O algoritmo de Regressão Logística gera um modelo de classificação binária (0 e 1). O modelo utiliza a Regressão Linear como base, a equação linear é (REGRESSÃO, 2020):

20

? = ? . ? + ?

O ?m? da equação é o coeficiente angular da reta, e o ?b? é o intercepto. Então, aplicando a equação linear obtemos um valor contínuo como resultado, após obter esse valor a Regressão Logística tem que realizar a classificação do resultado, então é aplicado uma função chamada sigmoide (REGRESSÃO, 2020):

? = 1

1+ ? ??

Essa função chamada de sigmoide, ou função logística, é responsável por ?achatar? o resultado da regressão linear, calculando a probabilidade de o resultado pertencer a classe 1, classificando assim o resultado da função linear em 0 ou 1.

Segundo Batista e Filho (2019), o modelo de regressão logística que tem como objetivo realizar uma classificação binária de uma variável de interesse irá predizer a probabilidade de pertencer à classe positiva. Tem-se, portanto, que é dado por:?

? = { 0, ?? ???? (? = 1 ??) < 0, 5

ou

? = { 1, ?? ???? (? = 1??) ? 0, 5

Então a decisão do modelo de regressão logística está relacionada à escolha de um ponto de corte para p(x). Caso o ponto de corte escolhido for p(x)=0,5, os resultados que forem p(x)>0,5 serão classificados como 1 e os resultados p(x)<5 serão classificados como 0

(SANTOS, 2018).
A Figura 2 ilustra um exemplo de como é o diagrama de Regressão Logística. Pelo diagrama então pode-se observar que o modelo possui uma representação gráfica em formato de ?S?, essa seria a curva logística. As predições do modelo sempre ficaram na linha 1 e 0 do eixo y, pois é o intervalo estabelecido pelo algoritmo, então a classificação sempre será nesse intervalo.
21
Figura 2. Exemplo de Diagrama de Regressão Linear. (Fonte: Batista e Filho 2019).
4.2.3 Treino, teste e validação do modelo de aprendizado de máquina
A criação de um modelo de aprendizado de máquina é realizada em duas etapas: treinamento do modelo, que é realizado a partir dos dados fornecidos para o algoritmo, e etapa de teste, onde os modelos recebem dados novos e sem o rótulo, para que seja avaliado a performance do modelo (PELISSARI, 2021).
Para o treinamento dos modelos foram utilizadas duas abordagens, a primeira onde a base foi particionada entre conjunto de dados para teste e para treino em uma porcentagem de 90% para treino e 10% para teste, porém dessa forma pode ocorrer problema de sobreajuste ou overfitting. Nesta situação, o que pode ocorrer é o modelo não aprender com os dados, e apenas ?decorar? os dados recebidos e quando é realizado o teste o modelo consegue predizer apenas situações parecidas, não sendo capaz de identificar padrões com diferenças mínimas. Para evitar este problema, foi utilizado uma técnica de validação cruzada chamada K-folds cross-validation. Essa técnica de validação cruzada divide a base de dados em K subconjuntos, e então são realizadas várias rodadas de treinamento e teste alternando os subconjuntos utilizados para teste e treino, e o resultado do modelo baseia-se na média da acurácia observada em cada rodada (PELISSARI, 2021).
Pode-se observar na Tabela 1 uma ilustração da técnica K-fold cross-validation, onde temos um exemplo onde a base é dividida em 5 subconjuntos e cada subconjunto tem a parte de teste diferente dos outros subconjuntos, assim o modelo vai receber valores novos de teste e treino todas as vezes que executar um novo subconjunto.
22
Tabela 1. Exemplo da separação K-fold cross-validation. (Fonte: Adaptado de PELISSARI, 2021).

Predição 1	Predição 2	Predição 3	Predição 4	Predição 5
Teste	Treino	Treino	Treino	Treino
Treino	Teste	Treino	Treino	Treino
Treino	Treino	Teste	Treino	Treino
Treino	Treino	Treino	Teste	Treino
Treino	Treino	Treino	Treino	Teste

4.2.4 Métricas para avaliação de modelos
Para fins de avaliação, neste trabalho foram utilizadas três métricas para avaliar o desempenho dos modelos: a Matriz de Confusão e a AUC (Area under Receiver Operating Characteristic Curve - ROC). Na matriz de confusão pode-se observar o comportamento do modelo em relação a cada uma das classes a serem preditas pelo modelo, utilizando variáveis binárias (positivo: 1; e negativo: 0), para apresentar uma matriz que faz uma comparação entre as predições do modelo e os valores corretos do conjunto de dados (PELISSARI, 2021).

Na Tabela 2 temos um exemplo de uma matriz de confusão, a qual consiste em duas colunas e duas linhas, e os valores são atribuídos nos quadrantes com os seguintes resultados (PELISSARI, 2021):

? quando o valor real do conjunto de dados é 0, e a predição do modelo também classificou como 0, então temos um Verdadeiro Negativo (VN);

? quando o valor real é 1, e o modelo classificado como 0, temos um Falso Negativo (FN);

? quando o valor real é 0, e o modelo classificado como 1, então o resultado se enquadra no quadrante de Falso Positivo (FP);

? e por fim o quadrante Verdadeiro Positivo (VP), que são os resultados que tem o valor real 1, e o modelo classificado como 1.

23

Tabela 2. Modelo de Matriz de confusão. (Fonte: Adaptado de PELISSARI, 2021).

Valor Predito
Negativo (0) Positivo (1)
Classe
Real
Negativo
(0)
Verdadeiro
Negativo
Falso
Positivo
Positivo
(1)
Falso
Negativo
Verdadeiro
Positivo

A segunda métrica de avaliação utilizada foi a Curva ROC que permite avaliar o desempenho de um modelo com relação às predições efetuadas. A curva ROC é um gráfico de Sensibilidade (taxa de verdadeiros positivos) versus taxa de falsos positivos, a representação da curva ROC permite evidenciar os valores para os quais existe otimização da Sensibilidade em função da Especificidade (CABRAL, 2013).

Na Figura 3 temos um exemplo de uma curva ROC. A linha traçada na cor preta é uma linha de referência, ela representa hipoteticamente a curva ROC de um classificador puramente aleatório. Caso a linha laranja, que representa o resultado do modelo que está sendo avaliado, ficasse igual a linha tracejada preta, indicaria que o modelo avaliado estaria predizendo aleatoriamente os resultados.

Figura 3: Exemplo de Curva ROC. (Fonte: Elaboração Própria).

Por uma análise de inspeção visual, quanto mais próximo a linha laranja estiver do valor 1 no eixo Y, melhor está a capacidade do modelo para diferenciar as classes. O valor da

24

AUC representa a capacidade do modelo de diferenciar as classes, quanto maior o valor do

AUC, melhor é a predição realizada pelo modelo, uma vez que, os valores resultantes iguais a 0 são realmente 0, e os iguais a 1 são de fato 1 (PELISSARI, 2021).

5 METODOLOGIA

O desenvolvimento deste trabalho foi realizado em três etapas: (1) Pré-processamento e Análise Exploratória do conjunto de dados, onde foram aplicadas técnicas comuns para preparar o conjunto de dados como tratamento de valores nulos e seleção de variáveis de interesse; (2) Implementação de modelos de aprendizado de máquina supervisionado para predição das mortes neonatais, utilizando os algoritmos de árvore de decisão e regressão logística; e (3) Análise de Resultados, que será realizada uma análise do desempenho do modelo.

5.1 TECNOLOGIAS E FERRAMENTAS

Neste trabalho vamos utilizar a linguagem de programação Python (PYTHON, 2021) pela sua simplicidade de codificação e alto poder de processamento. Foi utilizado também a plataforma Google Colab Research (Colab, 2021), para o desenvolvimento dos algoritmos e treinamento dos modelos. Além disso, foi utilizado também o Jupyter Notebook (Jupyter, 2021), instalado localmente em computador pessoal, pois o Google Colab Research possui limitação de recursos de memória e processamento, então as duas plataformas foram usadas para realização deste trabalho. As principais bibliotecas utilizadas foram: numpy (NumPy, 2021), e pandas (Pandas, 2021), para a manusear os dados, matplotlib (Matplotlib, 2021), e seaborn (Seaborn, 2021), para a plotagem dos gráficos, e por fim a biblioteca sklearn (Scikit-Learn, 2021), que é a biblioteca responsável pelos algoritmos de aprendizado de máquina.

5.2 BASE DE DADOS

As bases de dados a serem utilizadas serão o SIM e SINASC, que são as duas principais fontes de informações sobre nascimentos e óbitos no Brasil. O SINASC é alimentado com base na Declaração de Nascido Vivo (Declaração de Nascido Vivo - DNV) (BELUZO et al.,2020b apud Oliveira et al., 2015). O SIM tem como objetivo principal apoiar a coleta, armazenamento e processo de gestão de registros de óbitos no Brasil (BELUZO et 25
al.,2020b apud Morais et al., 2017), e foi usado para rotular o óbito registrado no SIM, utilizando o campo DNV como chave de associação. O SIM será utilizado para rotular os registros do SINASC, pois o SIM coletas informações sobre mortalidade e é usado como base para o cálculo de estatísticas vitais, como a taxa de mortalidade neonatal e o SINASC reúne informações sobre dados demográficos e epidemiológicos do bebê, da mãe, do pré-natal e do parto (BELUZO et al.,2020b).

A Tabela 3 apresenta as variáveis da base de dados. Essa base tem variáveis que contêm valores quantitativos e categóricos. A coluna ?num_live_births? por exemplo contém o número de nascidos vivos anteriores da mãe e é composta por valores quantitativos contínuos, e a coluna ?tp_pregnancy? utiliza valores nominais categóricos que indicam o tipo de gravidez.

Tabela 3. Colunas da base de dados e suas descrições. (Fonte: Adaptado de BELUZO et al.,2020b).

Nome da coluna	Descrição	Domínio de dados
Variáveis	demográficas e socioeconômicas	

maternal_age Idade da mãe Quantitativo Contínuo (inteiro)
tp_maternal_race Raça / cor da pele da
mãe
Nominal categórico (inteiro)
1 - branco; 2 - Preto; 3 -
amarelo; 4 - Pele morena; 5 -
Indígena.
tp_marital_status Estado civil da mãe Nominal categórico (inteiro)
1 - Único; 2 - Casado; 3 - Viúva;
4 - Separados / divorciados
judicialmente; 5 - Casamento
por união estável; 9 - Ignorado.
tp_maternal_schooling Anos de escolaridade da
mãe
Nominal categórico (inteiro)
1 - nenhum; 2 - de 1 a 3 anos; 3 -
de 4 a 7 anos; 4 - de 8 a 11 anos;
5 - 12 e mais; 9 - Ignorado.
Variáveis obstétricas maternas
num_live_births Número de nascidos
vivos
Quantitativo Contínuo (inteiro)
num_fetal_losses Número de perdas fetais Quantitativo Contínuo (inteiro)
num_previous_gestations Número de gestações Quantitativo Contínuo (inteiro)
26
anteriores
num_normal_labors Número de partos
normais (trabalhos de
parto)
Quantitativo Contínuo (inteiro)
num_cesarean_labors Número de partos
cesáreos (partos)
Quantitativo Contínuo (inteiro)
tp_pregnancy Tipo de gravidez Nominal categórico (inteiro)
1 - Singleton; 2 - Twin; 3 -
Trigêmeo ou mais; 9 - Ignorado.
Variáveis relacionadas a cuidados anteriores
tp_labor Tipo de parto infantil
(tipo de parto)
Categórico Nominal (inteiro)
1 - Vaginal; 2 - Cesariana;
num_prenatal_appointments Número de consultas de
pré-natal
Quantitativo Contínuo (inteiro)

tp_robson_group Classificação do grupo
Robson
Ordinal categórico (inteiro)
Variáveis relacionadas ao recém-nascido
tp_newborn_presentation Tipo de apresentação de
recém-nascido
Categórico Nominal (inteiro)
1 - Cefálico; 2 - Pélvico ou
culatra; 3 - Transversal; 9 -
Ignorado.
has_congenital_malformation Presença de
malformação congênita
Nominal categórico (inteiro)
1 - Sim; 2 - Não; 9 - Ignorado
newborn_weight Peso ao nascer em
gramas
Quantitativo Contínuo (inteiro)
cd_apgar1 Pontuação de Apgar de 1
minuto
Ordinal categórico (inteiro)
cd_apgar5 Pontuação de Apgar de 5
minuto
Ordinal categórico (inteiro)
gestaional_week Semana de gestação Quantitativo Contínuo (inteiro)
tp_childbirth_care Assistência ao parto Categórico Nominal (inteiro)
1 - Doutor; 2 - enfermeira ou
obstetra; 3 - Parteira; 4 - outros;
9 - Ignorado.
was_cesarean_before_labor Foi cesáreo antes do
parto.
27
was_labor_induced Foi induzido ao trabalho
de parto.
is_neonatal_death Morte antes de 28 dias
(rótulo)
Nominal categórico (inteiro)
0 - sobrevivente; 1 - morto.

Neste trabalho foi utilizado dados do período de 2014 à 2016 devido sua melhor qualidade. Este recorte possui 6.760.222 registros dos quais 99.4% são amostras de recém-nascidos vivos e 0.6% são amostras onde os recém-nascidos vieram a óbito conforme pode ser observado na Figura 4. Com essas informações consegue-se observar que a base de dados é desbalanceada.

Figura 4. Distribuição da base de dados considerando se o recém-nascido viveu ou morreu durante o período neonatal. (Fonte: Elaboração Própria).

5.3 PREPARAÇÃO DA BASE DE DADOS

Nessa etapa foi realizada a preparação da base de dados para o treinamento dos modelos que consiste na limpeza, transformação e preparação dos dados a serem utilizados na análise exploratória e na criação dos modelos preditivos. A base de dados pode conter informações desnecessárias para o modelo que pode acabar atrapalhando as predições,
28
algumas colunas podem ser retiradas ou seus valores podem ser modificados, por exemplo valores reais são modificados para inteiros.

5.4 DESENVOLVIMENTO DOS MODELOS DE APRENDIZADO DE MÁQUINA

Como já dito anteriormente na seção 4 os algoritmos de aprendizado de máquina escolhidos para esse trabalho foram os: Regressão Logística e Árvore de Decisão que utilizam a lógica mostrada na seção 4 , Fundamentação Teórica. Esses dois foram escolhidos por alguns motivos, ambos são algoritmos de classificação e supervisionados, atendendo os critérios necessários para a predição de mortalidade neonatal, ambos os algoritmos são bons para realizar predições, a Árvore de Decisão inclusive é um ótimo modelo para analisar as decisões que estão sendo feitas pelo modelo, afinal esse algoritmo tem a característica de ser um WhiteBox, ou seja conseguimos ver o que o modelo aprendeu e o que ele está decidindo. Esses algoritmos selecionados são algoritmos que se encaixam na necessidade deste trabalho e são muito utilizados na comunidade acadêmica.

Para o algoritmo de Regressão Logística, como foi dito anteriormente na seção ?Preparação da base de dados? a base de dados foi tratada e particionada, no particionamento foi utilizado 90% da base para o treino e 10% para os testes e então foi realizado o treinamento do modelo, também foi aplicada a função RFECV (Eliminação Recursiva de Feature com Validação Cruzada), esse função executa a RFE (Eliminação Recursiva de Feature), que tem como ideia selecionar features considerando recursivamente conjuntos cada vez menores de features, isso ocorre da seguinte forma. Primeiro, o estimador é treinado no conjunto inicial de features e a importância de cada feature é obtida por meio de um atributo "coef" ou por meio de um atributo "feature_importances". Em seguida, as features menos importantes são removidas do conjunto atual de features. Esse procedimento é repetido recursivamente no conjunto removido até que o número desejado de features a serem selecionados seja finalmente alcançado. Porém o diferencial da RFECV é que essa função executa a RFE em um loop de validação cruzada para encontrar o número ideal ou o melhor número de features.

Após essa seleção, foi realizado o treinamento do modelo usando todas as features e também treinamos usando as features que o RFECV selecionou, para compararmos os resultados no final. Também foi utilizado o método de K-folds cross-validation em um novo treinamento para conferirmos se o modelo estava sofrendo overfitting (sobreajuste). E por fim aplicamos um método chamado GridSearchCV, esse método permite que seja realizado testes
29
alterando algumas combinação de parâmetros nos nossos modelos, facilitando para achar a melhor combinação, esse método é parecido com o cross validation, o diferencial do Grid Search é que ele faz os cross validation de vários modelos com hiperparâmetros diferentes de uma só vez.

Para o algoritmo de Árvore de Decisão também foi utilizado o mesmo

particionamento de 90% para treino e 10% para teste, na criação do modelo foi colocado a entropy como critério de decisão e uma profundidade máxima de 4, pois com números maiores o modelo ficava muito complexo e ele errava mais as predições, também utilizamos um algoritmo para mostrar a importância de cada Feature para o modelo.
O algoritmo de Regressão Logística e análise exploratória foi baseado em códigos disponível na plataforma web Kaggle (MNASSRI, 2020), no exemplo do Kaggle o autor faz os códigos para predizer mortes no navio Titanic, para o trabalho de mortalidade neonatal os códigos foram alterados para realizar predição de mortes neonatais. Já no algoritmo de Árvore de Decisão foi baseado em uma vídeo aula do professor Diogo Cortiz (ÁRVORE..., 2020), nessa vídeo aula o professor explica sobre os algoritmos de Árvore de Decisão e depois implementa o um exemplo de código, o código usado neste trabalho usou o codigo do professor Diogo Cortiz, e foi alterado para realizar predições de mortalidade neonatal.

5.5 ANÁLISE DOS RESULTADOS

Para a análise de resultados foi utilizado dois métodos que é o de Matriz de confusão e a curva ROC esses métodos são explicados na seção 4, a Fundamentação Teórica. Com esses métodos pode-se realizar uma avaliação do modelo, e assim será possível analisar os valores de acurácia, precisão e recall do modelo, essas métricas são utilizadas avaliar o desempenho do modelo, com a acurácia é possível calcular a proximidade entre o valor obtido na predição dos modelos e os valores esperados, com a precisão é possível analisar as amostras que o modelo classificou como 0 eram realmente 0 na classe real e com o recall é possível analisar as amostras que o modelo conseguiu reconhecer como a classe correta.

5.6 ACESSO A BASE DE DADOS E CÓDIGOS

A base de dados utilizada neste projeto pode ser encontrada no link: ... . E os códigos podem ser encontrados no link: ... .

30

6 RESULTADOS

Os mesmos experimentos mostrados abaixo foram aplicados para uma base de dados que contém somente dados de São Paulo, essa base é um recorte da base do Brasil todo em contém cerca de 1.400.000 amostras, esse recorte da base também é bem desbalanceado. Os resultados obtidos nesse experimento usando o recorte de São Paulo foram bem parecidos com os experimentos da base do Brasil todo, e os códigos desse experimento podem ser visualizados no apêndice X.

6.1 ANÁLISE EXPLORATÓRIA

O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

6.1.1 Apresentação dos Dados

Como dito anteriormente na seção 5.2 ?Base de Dados? deste trabalho, a base de dados é formada pelas informações do SIM e do SINASC contém 6.760.222 amostras e 29 colunas, porém serão utilizadas somente 23 colunas sendo elas as colunas descritas na Tabela 3. Na seção 5.2 também tem a Figura 4 que mostra que a base dados é desbalanceada tendo 99.4% das amostras de recém-nascidos vivos e 0.6% das amostras onde os recém-nascidos vieram a óbito no período neonatal.

6.1.1.1 Caracteristicas demográficas e socioeconômicas maternas

O apêndice X contém uma tabela que possui as estatísticas sumarizadas das variáveis

demográficas e socioeconômicas maternas. Nesta tabela por exemplo podemos observar a coluna ?maternal_age? que contém a idade da mãe, e na tabela mostra que a média da idade das mães é de 26.4 anos, e os dados possuem uma variação de no mínimo 8 anos e no máximo 55 anos, e a mediana é 26 anos. Na tabela também tem as colunas: ?tp_maternal_schooling? que mostra a categoria de anos de escolaridade da mãe, ?tp_marital_status? que mostra o estado civil da mãe em dados categóricos e a coluna ?tp_maternal_race? que mostra a raça da mãe também em dados categóricos.
31
6.1.1.2 Variávies obstétricas maternas
No apêndice X pode-se observar uma tabela que possui as estatísticas sumarizadas das variáveis obstétricas maternas. Nesta tabela por exemplo temos a coluna ?num_live_births?, essa coluna mostra o número de nascidos vivos anteriores que a mãe obteve, a média desta coluna é 0.94 , a mediana é 1, o número mínimo é 0 e o máximo é 10 nascidos vivos. Na tabela podemos observar também a coluna, ?num_fetal_losses?, esta coluna mostra o número de perdas fetais anteriores da mãe, a média desta coluna é 0.21, a medina é 0, o número mínimo é 0 e o máximo é 5, esses valores mostram que poucas mães tiveram perdas fetais antes da gravidez atual. Também na tabela tem estatísticas da coluna ?num_previous_gestations? esta coluna mostra o número de gestações anteriores da mãe, esta coluna tem a média de 1.14, a mediana é 1, o número mínimo é 0 e o máximo 10. Essa tabela também contém as colunas: ?num_normal_labors? que é o número de partos normais da mãe, ?num_cesarean_labor? que mostra o número de partos cesáreos da mãe e por fim mostra a coluna ?tp_pregnancy? que é o tipo da gravidez.
6.1.1.3 Variávies de histórico de gravidez
No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis do histórico de gravidez. Nesta tabela temos as colunas ?num_prenatal_appointments? que mostra o número de consultas de pré-natal, a média dessa coluna é 7.8, a mediana é 8, a variação dos dados é de no mínimo 0 e no máximo 40. A tabela também contém as colunas: ?tp_labor? que mostra o tipo de parto, e também contém a coluna ?tp_robson_group? que mostra a classificação do grupo Robson.
6.1.1.4 Variávies relacionadas ao recém-nascido
No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis relacionadas ao recém-nascido. Nesta tabela por exemplo podemos observar a coluna ?newborn_weight? que armazena o peso ao nascer dos recém-nascidos em gramas, a média dos dados dessa coluna é 3187.3, a mediana é 3215, a variação dos dados é de no mínimo 0 e no máximo 6000 gramas. Também podemos observar a coluna ?gestacional_week? que mostra o número de semanas de gestação, a média é 38.4, a mediana é 39, a variação dos dados é de no mínimo 15 e no máximo 45 semanas. Além dessas colunas a tabela também
32
contém as colunas: ?cd_apgar1? que mostra a pontuação de Apgar de 1 minuto, ?cd_apgar5? que mostra a pontuação de Apgar de 5 minuto, ?has_congenital_malformation? que mostra se o recém-nascido contém a presença de alguma malformação, ?tp_newborn_presentation? que mostra o tipo de apresentação de recém-nascido, ?tp_labor? que mostra o tipo de parto, ?was_cesarean_before_labor? que mostra se o recém-nascido foi cesáreo antes do parto, ?was_labor_induced ? que mostra se o recém-nascido foi induzido ao trabalho de parto,

?tp_childbirth_care? que mostra se ouve assistência ao parto, e por fim temos a coluna ?is_neonatal_death? que mostra se o recém-nascido veio a óbito ou não.

6.1.2 Análise das variáveis

Nesta seção serão mostrados a parte de análise de variáveis com diferentes tipos de gráficos.

6.1.2.1 Variáveis contínuas

Na Figura 5 temos dois gráficos boxplot, no boxplot à esquerda temos a distribuição da variável peso ao nascer e nele é possível observar duas caixas onde a caixa azul são os vivos e o laranja são os mortos. Nas duas caixas mostradas no gráfico podemos observar pequenos círculos nas pontas, esses círculos são outliers (dados fora da curva), então para a caixa de vivos os outliers são recém-nascidos com pesos acima de 4500 gramas ou abaixo de 2000 gramas, e para a caixa de mortos os outliers são os recém-nascidos com mais de 5500 gramas. Para os vivos temos a mediana de aproximadamente 3200 gramas e para os mortos a mediana é 1300 gramas. Cada caixa representa 50% da base de dados, a caixa azul de vivos mostra que o peso de recém-nascidos que sobreviveram está aproximadamente entre 2700 a 3600 gramas, e para a caixa laranja de mortos o peso de recém-nascidos que vieram a óbito estão aproximadamente entre 700 a 2500 gramas. Então o gráfico mostra para nós que os recém-nascidos que têm maior risco de vir a óbito tem pesos menores que 2000 gramas.

33

Figura 5. BoxPlot das features Peso ao nascer e Idade da mãe. (Fonte: Elaboração Própria).

No gráfico à direita da Figura 5 temos um boxplot da distribuição da variável idade da mãe, como foi dito anteriormente a caixa azul são os vivos e o laranja são os mortos. Nesse boxplot então podemos ver que os outliers da caixa de vivos são mães com idade acima de 46 anos aproximadamente, para os mortos os outliers são mães com idade acima de 50 anos. Para os vivos temos a mediana de aproximadamente 26 anos e para os mortos a mediana é de 26 anos. Nos vivos a idade das mães está aproximadamente entre 21 a 32 anos, e para os mortos a idade das mães está aproximadamente entre 20 a 34 anos. É importante ressaltar que o gráfico mostra as duas caixas bem parecidas, então de acordo com os dados não contém diferença significativa entre vivos e mortos usando a idade da mãe para distribuir.

Já na Figura 6 temos um boxplot da distribuição da variável semanas de gestação.

Nesse boxplot então podemos ver que os outliers da caixa de vivos são gestações com mais de 44 semanas aproximadamente, e gestações com menos de 35 semanas. Para os vivos temos a mediana de aproximadamente 39 semanas e para os mortos a mediana é de 32 semanas. Nos vivos as semanas de gestação estão aproximadamente entre 36 a 40 semanas, e para os mortos a semanas de gestação estão aproximadamente entre 26 a 36 semanas. Nesse boxplot os dados mostraram que para as gestações que têm menos de 36 semanas os recém-nascidos têm maior risco de vir a óbito.

34

Figura 6. BoxPlot da feature semana de gestação. (Fonte: Elaboração Própria).

Na Figura 7 temos um histograma da distribuição de peso ao nascer por classe, podemos observar no gráfico que grande parte dos vivos (linha azul) estão distribuídos entre 2000 e 4000 gramas, a área entre esses valores tem muitos dados de vivos, já entre os valores 0 e 2000 gramas temos uma concentração dos dados de mortos.

Figura 7. Histograma de distribuição do peso entre as classes. (Fonte: Elaboração Própria)

6.1.2.2 Variáveis categoricas

Observando a Figura 8 podemos ver um gráfico de barras da distribuição da variável escolaridade da mãe, este gráfico mostra a contagem de registros por escolaridade da mãe, esse gráfico mostra para nós que a maior parte das mães estão na categoria 4 que representa de 8 a 11 anos de escolaridade.

35

Figura 8. Gráfico de barras da distribuição da variável Escolaridade da mãe. (Fonte: Elaboração Própria).

Observando a Figura 9 podemos ver um gráfico de barras da distribuição da variável número de vivos, este gráfico mostra a contagem de registros por número de vivos, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho.

Figura 9. Gráfico de barras da distribuição da variável Número de nascidos vivos. (Fonte: Elaboração Própria).

Observando a Figura 10 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 1 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 1 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem

36

apgar de 8 ou 9, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.

Figura 10. Gráfico de barras da distribuição da variável Pontuação Apgar de 1 minuto. (Fonte: Elaboração Própria).

Observando a Figura 11 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 5 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 5 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem apgar de 9 ou 10, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.

Figura 11. Gráfico de barras da distribuição da variável Pontuação Apgar de 5 minutos. (Fonte: Elaboração Própria).

37

Observando a Figura 12 podemos ver um gráfico de barras da distribuição da variável presença de malformação congênita, este gráfico mostra a contagem de registros por presença de malformação congênita, esse gráfico mostra para nós que a grande maioria dos dados estão na categoria 2 que mostra que o recém-nascido não possui malformação, esse grande volume para a categoria 2 é causada pelo desbalanceamento da base.

Figura 12. Gráfico de barras da distribuição da variável Presença de malformação congênita. (Fonte: Elaboração Própria).

Observando a Figura 13 podemos ver um gráfico de barras da distribuição da variável número de gestações, este gráfico mostra a contagem de registros por número de gestações, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho, da mesma forma mostrada na Figura 9.

38

Figura 13. Gráfico de barras da distribuição da variável Número de gestações. (Fonte: Elaboração Própria).

Observando a Figura 14 podemos ver um gráfico de barras da distribuição da variável assistência ao parto, este gráfico mostra a contagem de registros por assistência ao parto, esse gráfico mostra para nós que a grande maioria dos dados mostram que o parto teve assistência de uma Enfermeira ou obstetra ou essa informação foi ignorada na hora do resgistro.
Figura 14. Gráfico de barras da distribuição da variável Assistência ao parto. (Fonte: Elaboração Própria).
6.1.2.3 Analise bi-variada
A Figura 15 mostra para nós a importância da feature peso ao nascer com o gráfico mostrado é possível observar claramente a diferença de pesos entre vivos e mortos, recém-nascidos com peso acima de 2000 gramas sobreviveram, já os recém-nascidos com menos de 2000 gramas vieram a óbito.
39
Figura 15. Gráfico de densidade de peso entre as classes. (Fonte: Elaboração Própria).
Na Figura 16 podemos observar um gráfico de correlação entre algumas features de valores contínuos, algumas dessas features possuem correlações positivas, como pontuação apgar de 1 minuto e pontuação apgar de 5 minutos, essas colunas possuem uma correlação de 0.7, então nascidos que recebem um valor alto de apgar de 1 minuto tendem a receber um valor alto no apgar de 5 minutos. O gráfico também mostra outras features com correlações positivas como, número de partos normais e número de gestações anteriores que contém uma correlação de 0.81, número de gestações anteriores com idade da mãe que tem uma correção de 0.39, número de partos de cesáreos com idade da mãe que possui correlação de 0.24, número de partos normais com idade da mãe com a correlação de 0.26, semanas de gestação com peso ao nascer em gramas com correção de 0.52, e as apgar de 1 e 5 minutos com semanas de gestação. E grande parte das features, possuem correlações baixas então elas são inversamente correlacionadas pelo que o gráfico mostra, como por exemplo número de partos normais com número de consultas pré-natais possuem uma correção de -0.17, e número de partos cesáreos com número de partos normais que contém uma correlação de -0.15. Então a correlação dessas features estão bem baixas tirando as correlações acima de 0.7.
40
Figura 16. Gráfico de correlação. (Fonte: Elaboração Própria)
6.1.2.4 Distribuição por raça
Na Figura 17 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por raça, e podemos observar que a raça 5 (indigena), é o boxplot que contem a menor variação de idade, já as raças 1 (branco) e 3 (amarelo) são as caixas que possuem maior variação com a média em torno de 20 e 30 anos.
41
Figura 17. Distribuição da idade da mãe por raça. (Fonte: Elaboração Própria).
No boxplot mostrado na Figura 18 podemos observar a distribuição de peso ao nascer por raça, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre as raças para peso ao nascer.
Figura 18. Distribuição de peso ao nascer por raça. (Fonte: Elaboração Própria).
Na Figura 19 temos o boxplot da a distribuição de semanas de gestação por raça, e praticamente todas as caixas são iguais, somente a caixa da raça 1 (branco) possui menor variação que as outras.

42
Figura 19. Distribuição de semanas de gestação por raça. (Fonte: Elaboração Própria).
6.1.2.5 Distribuição por estado civil
A Figura 20 mostra o boxplot da a distribuição de peso ao nascer por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre os estados civis para peso ao nascer.
Figura 20. Distribuição de peso ao nascer por estado civil. (Fonte: Elaboração Própria).
A Figura 21 mostra o boxplot da a distribuição de semanas de gestação por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Porém os estados civis 2 (Casado) e 4 (Separados/divorciados judicialmente), contém variações menores.
43
Figura 21. Distribuição de semanas de gestação por estado civil. (Fonte: Elaboração Própria).
Na Figura 22 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por estado civil, e podemos observar que o estado civil 3 (Viúva), possui a maior variação, já o estado civil 4 (Separados/divorciados judicialmente) é a caixa que possuem menor variação
Figura 22. Distribuição da idade da mãe por estado civil. (Fonte: Elaboração Própria).
6.1.2.6 Distribuição por escolaridade da mãe
Na Figura 23 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por escolaridade da mãe, e podemos observar que a escolaridade 2 (1 a 3 anos) e 3 (4 a 7
44
anos), possui as maiores variações, já escolaridade 5 (12 ou mais anos) é a caixa que possuem menor variação
Figura 23. Distribuição da idade da mãe por escolaridade da mãe. (Fonte: Elaboração Própria).
A Figura 24 mostra o boxplot da a distribuição de peso ao nascer por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então, pouca variação é observada entre a escolaridade da mãe para peso ao nascer.
Figura 24. Distribuição de peso ao nascer por escolaridade da mãe. (Fonte: Elaboração Própria).
A Figura 25 mostra o boxplot da a distribuição de semanas de gestação por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas,
45
todas são praticamente iguais. Porém a escolaridade 5 (12 ou mais anos) contém variação menor que as outras.
Figura 25. Distribuição de semanas de gestação por escolaridade da mãe. (Fonte: Elaboração Própria).
6.2 ÁRVORE DE DECISÃO
A Figura 26 mostra um trecho do algoritmo utilizado, nesse trecho é realizado a divisão do DataFrame das features de entrada , e o DataFrame que contém o rótulo.Neste caso o rótulo será o DataFrame ?target?, esse Dataframe contém a feature

?is_neonatal_death?, essa feature informa se o recém nascido veio a óbito ou não, sendo o foco da predição que desejamos realizar essa feature entra como rótulo para o modelo, já no Dataframe ?nome_features? temos as features que vão servir de entrada para o modelo, é a partir dessas features que o modelo tentará encontrar padrões para predizer o risco de óbito do recém nascido. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

46

Figura 26. Separação dos DataFrames de entrada e rótulo. (Fonte: Elaboração Própria).

6.2.1 Modelo de Árvore de Decisão Utilizando Particionamento 90/10

O algoritmo de árvore de decisão foi treinado utilizando duas formas diferentes: usando as partições 90% para treino e 10% para os testes e utilizando o método K-folds cross-validation. E para os experimentos iniciais foi utilizado o particionamento 90/10.

Tabela 4. Resultados do modelo de árvore de decisão. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	671807
Morto(1)	0.71	0.29	0.41	4216

Analisando a Tabela 4 pode-se ver os seguintes resultados, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 71% que realmente pertenciam a classe 1, então cerca de 29% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 29% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem a classe 1 que seria dos recém nascidos que poderiam vir a óbito, já para a classe de vivos a

47

classe 0 o modelo está identificando muito bem e classificando de forma correta. A tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.

Esse modelo teve uma acurácia de 0.99 que é uma acurácia bem alta, porém somente pela acurácia não é possível dizer que o modelo está excelente pelo motivo da base que está sendo usada é altamente desbalanceada, então para a classe 0 que é a de vivos temos muito mais dados que para a classe de mortos, e o modelo está acertando bastante a classe de vivos logo a acurácia fica alta por esses acertos, enquanto para a classe de mortos o modelo não está acertando tanto.

A Figura 27 mostra a curva ROC do modelo, pode-se observar que a AUC da curva ROC ficou em 0.92, a AUC mostra que o modelo está acertando bastante as predições, pois quanto mais perto de 1 melhor está no nosso modelo, mas no caso desse modelo vimos acima que as predições para a classe de mortos(1) está ruim, o modelo está acertando poucas classificações como 1. Então a AUC assim como a acurácia está alta porque a base de dados utilizada é altamente desbalanceada tendo muito mais amostras de vivos(1), e como o modelo está bom para essa classificação e está acertando bastante os valores de AUC e acurácia ficam altos .

Figura 27. Curva ROC modelo de Árvore de decisão. (Fonte: Elaboração Própria).
A Figura 28 mostra a importância de cada feature para o modelo sendo assim a feature 10 - Peso ao nascer, a mais importante para o modelo, pois dela o modelo conseguiu tirar mais informações, seguido das features, 13 - Apgar dos 5 primeiros minutos, 11 - Semanas de 48
Gestação, 14 - Presença de malformação 12 - Apgar do 1 primeiro minutos. Então essas são as features que foram mais decisivas para a montagem da árvore e para as predições do modelo.
Figura 28. Gráfico de importância das features. (Fonte: Elaboração Própria).
Na Figura 29 temos uma imagem reduzida da árvore de decisão que foi montada com o treinamento utilizando 5 como profundidade máxima para a árvore, a imagem da árvore completa pode ser visualizada no apêndice A. É possível ver que as features marcadas como importante para o modelo apareceu diversas vezes na árvore, e a feature peso ao nascer foi escolhida para ser o nó raiz da árvore, de todas as features ela foi a que teve a melhor entropia para ser o nó raiz, e depois aparece mais vezes para outras decisões em outros níveis da árvore e isso faz com que essa feature se torne a mais importante para o modelo, também podemos ver que a feature ?Apgar do 1 primeiro minuto? mesmo sendo a menos importante para o modelo ela aparece somente no nó de número 42 isso mostra que a entropia e o ganho de informação dessa feature nas outras decisões não foram boas fazendo ela ser a feature com menos importância utilizada no modelo.
49
Figura 29. Árvore de decisão montada a partir do algoritmo. (Fonte: Elaboração Própria).
6.2.2 Modelo de Árvore de Decisão Utilizando K-folds cross-validation
Após realizar esse experimento com a partição 90/10, foi realizado outro experimento com esse mesmo algoritmo porém agora utilizando o método K-folds cross-validation para o treinamento. O K-folds cross-validation foi configurado para separar a base de dados em 10 partes iguais, e assim o algoritmo foi treinado novamente gerando um novo modelo. O método de K-folds cross-validation é utilizado para conferir se o modelo não está sofrendo problemas de overfitting, é importante garantir que o modelo está aprendendo com os dados e não está somente decorando.
Na Tabela 5 temos então o resultados desse modelo, pode-se observar que os resultados para a classificação de vivos(0) não teve alteração, o modelo continua tendo 100% na precisão e no recall, mas na classificação de mortos(1) o modelo teve uma diminuição na precisão que agora está em 65% e também teve uma diminuição no recall que agora está em 23%, porém essa diminuição é justificada porque para o teste deste modelo foi utilizado a base completa e não somente a partição de teste que contém 10% da base total. Logo os resultados não tiveram uma alteração drástica e tiveram um comportamento bem semelhante, incluindo a acurácia que se manteve em 0.99 e a AUC que teve uma diminuição pequena também e ficou com 0.89.
50
Tabela 5. Resultados do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	6719191

Morto(1) 0.65 0.23 0.34 41031
Na Tabela 6 é possível ver a matriz de confusão do modelo que mostra o número exato de erros e acertos do modelo, pode-se observar que para a classe de vivos o modelo classificou corretamente 6.714.117 amostras da base de dados completa, o modelo classificou como 0 as amostras que nos rótulos realmente eram 0, errou apenas 5074 amostras. Já para a classe de mortos o modelo acertou somente 9552 amostras, classificou como 1 as amostras que no rótulo eram 1 também, e errou 31479. Isso mostra que o modelo está muito desbalanceado, pois está errado muito mais do que acertando as predições de mortos, sendo isso um reflexo da base de dados desbalanceada também. E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.
Tabela 6. Matriz de confusão do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6714117 5074
Mortos (1) 31479 9552
6.3 REGRESSÃO LOGÍSTICA
Para o algoritmo de Regressão Logística também foi utilizado métodos diferentes para o treinamento, O algoritmo foi treinado utilizando três formas diferentes: usando as partições 90% para treino e 10% para os testes, utilizando o método K-folds cross-validation e foi treinada utilizando o método RFECV que seleciona as melhores features para o modelo, juntamente com o método GridSearchCV. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.
51
6.3.1 Modelo de Regressão Logística Utilizando Particionamento 90/10
Para o primeiro experimento do algoritmo de regressão logística foi utilizado o particionamento 90/10 para o treinamento do algoritmo, sendo 90% da base de dados para realizar o treinamento do modelo e 10% da base para realizar os testes.
Na Tabela 7 temos os resultados do modelo de regressão logística utilizando o método de particionamento 90/10, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 65% que realmente pertenciam a classe 1, então cerca de 35% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 24% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem as amostras da classe 1 que seria dos recém-nascidos que poderiam vir a óbito, já para a classe de vivos a classe 0 o modelo está identificando muito bem e classificando de forma correta.E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.

Tabela 7. Resultados do modelo de regressão logística usando particionamento 90/10.
(Fonte: Elaboração Própria)
Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 671885
Morto(1) 0.65 0.24 0.35 4138
A Figura 30 abaixo mostra a curva ROC do modelo de regressão logística usando o particionamento 90/10, analisando a curva ROC podemos ver que a AUC da curva ficou em 0.93 e a acurácia do modelo ficou 0.99, mostrando que o modelo está acertando bastante as predições, porém esses números para o nosso modelo não mostra que ele está ótimo, com a análise da Tabela 7 acima podemos observar que o modelo está acertando muitas predições da classe de vivos e poucas predições da classe de mortos, como a base de dados é
52
desbalanceada, como a maior quantidade de dados está na classe de vivos e o modelo está acerto quase todos dessa classe a acurácia e a AUC ficam altas por esses acertos.Logo é preciso analisar mais resultados além da acurácia e da AUC do modelo.
Figura 30. Curva ROC modelo regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria).
Na Tabela 8 podemos analisar os número exatos que o modelo acertou de cada classe, como é mostrado na matriz de confusão o modelo acertou 671.435 da classe de vivos e acertou apenas 915 da classe de mortos, e errou mais que o triplo da classe de mortos. Então as predições do modelo estão muito desbalanceadas, para a classe de vivos o modelo está predizendo bem, porém para as classes de mortos o modelo está errando muitas predições.
Tabela 8. Matriz de confusão do modelo de regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 671435 504
Mortos (1) 3169 915
6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation
Para o segundo experimento do algoritmo de regressão logística foi utilizado o método K-folds cross-validation para o treinamento do modelo com a intenção de conferir e confirmar
53
que o modelo não esteja tendo problemas com overfitting e realmente esteja aprendendo com os dados que está sendo usado para o treinamento. O método K-folds cross-validation foi configurado para realizar uma separação de 10 partes iguais.
Na Tabela 9 podemos observar os resultados do modelo, e os resultados foram parecidos com o experimento anterior usando o particionamento 90/10, única diferença nos resultado é que no experimento anterior a precisão da classe de mortos ficou em 65% e no caso desse experimento usando o K-folds cross-validation a precisão ficou em 64%, mas os valores de recall e F1-Score se mantiveram, assim como todos os resultados da classe de mortos se mantiveram em 100%. A execução desse experimento mostra que o modelo realmente está aprendendo com os dados e não está apenas decorando, retirando o risco do modelo estar com overfitting. Mesmo sem alterações no resultado é importante saber que o

modelo não está com overfitting e está conseguindo aprender e encontrar padrões nos dados.
Tabela 9. Resultados do modelo de regressão logística usando K-folds cross-validation.
(Fonte: Elaboração Própria)
Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 6719191
Morto(1) 0.64 0.24 0.35 41031
Já na Figura 31 temos a curva ROC do modelo, essa curva ROC mostra a AUC de
todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do
método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas
para todas as vezes que foi realizado os testes, a AUC se manteve em 0.93 e 0.94, sendo a
média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não
se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua
acertando muito a classe de vivos que é a classe predominante na base de dados. Já era
esperado os resultados não sofrerem alterações grandes, pois a base de dados não sofreu
nenhuma alteração, esse experimento foi somente para conferir se a modelo não estava
sofrendo overfitting, e pelos resultados aparentemente a base não está com problemas de
sobreajuste.
54
Figura 31. Curva ROC modelo regressão logística usando K-folds cross-validation. (Fonte:
Elaboração Própria).
Na Tabela 10 temos a matriz de confusão do modelo, onde mostra que o modelo
acertou 6.713.658 amostras da classe de vivos e errou apenas 5.533, já para a classe de mortos
o modelo acertou somente 9.847 e novamente errou mais que o triplo errando 31.182
amostras. Como os resultados mostrados anteriormente não tiveram alteração era de se
esperar que as predições corretas e incorretas do modelo também não sofressem alterações, o
modelo continua acertando muito a classe de vivos e errando muito a classe de mortos,
mantendo as predições bem desbalanceadas.
55
Tabela 10. Matriz de confusão do modelo de regressão logística usando K-folds
cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6713658 5533
Mortos (1) 31182 9847
6.3.3 Modelo de Regressão Logística Utilizando GridSearchCV e RFECV
Para o terceiro e último experimento de regressão logística foi utilizado técnicas
diferentes juntas para tentar melhorar as predições do modelo, para esse experimento usamos
o K-folds cross-validation para o particionamento da base de dados assim como foi utilizado
no experimento acima, e também foi utilizado duas técnicas que ainda não foram usadas nos
experimentos anteriores que são as funções RFECV e GridSearchCV. A função RFECV
seleciona a quantidade ideal de features para ser usadas no treinamento do modelo e também
mostra quais são as melhores features para essa predição, então essa função ela utiliza a
validação cruzada para ir treinando e testando N vezes o modelo, e sempre vai alterando a

quantidade e as features utilizadas para o teste, então cada vez que a função testa os modelos ela grava a pontuação dos acertos e assim depois mostra o gráfico da Figura 32 e assim é possível ver quais são as features ideais para o modelo.
56
Figura 32. Resultado da função RFECV Base Brasil. (Fonte: Elaboração Própria).
Então os resultados mostrados na Figura 32 apresenta que o número ideal para treinar o modelo de features é 6, e as features sao: 'tp_maternal_schooling', 'gestaional_week', 'cd_apgar1', 'cd_apgar5', 'has_congenital_malformation', 'tp_labor'. Essas são as features que tiveram o melhor desempenho no RFECV, então para o treinamento do modelo desse experimento as features que vão ser utilizadas serão somente essas 6.
Após a execução da função de RFECV foi executado a função de GridSearchCV, o GridSearchCV é utilizado para descobrir quais são os melhores parâmetros para utilizar no algoritmo de regressão logística e criar os modelos, foi utilizado também a validação cruzada para testar qual seria os melhores parâmetros para melhorar o desempenho do modelo. Na Figura 33 pode-se ver que os parâmetros escolhidos pela função foram:
Figura 33. Resultado da função GridSearchCV. (Fonte: Elaboração Própria).
Na Tabela 11 podemos observar os resultados do modelo, e os resultados foram parecidos com os experimentos anteriores, para a classe de mortos a precisão teve um aumento e foi para 69% e o recall diminuiu chegando em 20%, então utilizando as novas
57
funções o modelo ganhou precisão e perdeu recall. Para a classe de vivos o modelo se manteve em 100% e não teve alterações para os resultados de precisão e recall. Mesmo com a utilização das funções de RFECV e GridSearchCV o modelo não teve uma melhora significativa para as predições de mortos.
Tabela 11. Resultados do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	6719191
Morto(1)	0.69	0.20	0.31	41031

Na Figura 34 temos a curva ROC do modelo, essa curva ROC mostra a AUC de todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas para todas as vezes que foi realizado os testes, a AUC se manteve em 0.92 e 0.93, sendo a média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua acertando muito a classe de vivos que é a classe predominante na base de dados.
58
Figura 34. Curva ROC modelo regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria).
Na Tabela 12 temos a matriz de confusão do modelo, onde mostra que o modelo acertou 6.715.535 amostras da classe de vivos e errou apenas 3.656, comparando com os experimentos anteriores de regressão logística as predições de vivos teve uma pequena piora e teve uma pequena diminuição nos acertos. Já para a classe de mortos, o modelo acertou 8.296 e errou 32.735 amostras, logo as predições de vivos tiveram uma uma piora e acertou um

pouco menos do que os outros experimentos, e na classe de vivos teve uma melhora acertando mais predições, porém como o problema está nas predições de mortos e para essa classe teve uma piora pode-se dizer que a função de GridSearchCV não teve um ganho significativo nos resultados. Então como modelo final foi escolhido o modelo utilizando somente o K-folds cross-validation, pois foi o modelo que mostrou o melhor desempenho dos experimentos de regressão logística.

59

Tabela 12. Matriz de confusão do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

	Predito	
	Vivos (0)	Mortos (1)
Classe Real		
Vivos (0)	6715535	3656
Mortos (1)	32735	8296

60

7 CONCLUSÃO

O principal objetivo deste trabalho foi analisar os resultados das predições de mortalidade neonatal dos algoritmos de aprendizado de máquina usando uma base de dados com dados do Brasil inteiro. A partir destes resultados apresentados na secção anterior ?Resultados?, é possível concluir que analisar somente a AUC e a acurácia do modelo não é o suficiente para garantir que o modelo está excelente, sendo assim temos que ter mais atenção a precisão, recall e as predições mostradas na matriz de confusão.

Ambos os modelos ficaram desbalanceados nas predições, os modelos acertaram mais predições de vivos do que mortos, e as predições corretas de mortos foram mais baixas do que as incorretas, para os modelos ficarem melhor precisam passar por algoritmos que possam balancear essas predições, uma outra alternativa para melhorar o modelo é utilizar uma base de dados mais balanceada, pois essa base que foi utilizada é altamente desbalanceada.

Embora os dois modelos tenham tido resultados parecidos , o modelo de regressão logística utilizando somente o K-folds cross-validation se saiu melhor que os outros experimentos, o modelo criado na seção ?6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation? manteve a acurácia, AUC, precisão e recall dos demais modelos e acertou mais predições, talvez com algoritmos para melhorar as predições da classe de mortos, balanceando mais as predições, esse modelo fique com ótimas taxas preditivas.

Além disso podes-se afirmar que o peso ao nascer do recém nascido é um fator muito importante para as decisões dos modelos, ambos os modelos usaram muito essa informação para as predições, também podemos colocar, as colunas de presença de malformação e semana de gestação, como colunas que foram importantes para os modelos, essas causas podem ser evitadas se houver um acompanhamento médico com qualidade e eficiência.

61

REFERÊNCIAS

ÁRVORE de Decisão - Aula 3. Gravação de Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=ecYpXd4WREk&t=4089s. Acesso em: 1 jul. 2020.

BARRETO, J. O. M.; SOUZA, N. M.; CHAPMAN, E. Síntese de evidências para políticas de saúde: reduzindo a mortalidade perinatal. Ministério da Saúde, p. 1?44,2015.

BATISTA, André Filipe de Moraes; FILHO, Alexandre Dias Porto Chiavegatto. Machine Learning aplicado à Saúde. In: ZIVIANI, Artur; FERNANDES, Natalia Castro; SAADE, Débora Christina Muchaluat. SBCAS 2019: 19 Simpósio Brasileiro de Computação Aplicada à Saúde. [S. l.: s. n.], 2019. cap. Capítulo 1.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. Towards neonatal mortality risk classification: A data-driven approach using neonatal, maternal, and social factors, [s. l.], 23 out. 2020.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. SPNeoDeath: A demographic and epidemiological dataset having infant, mother, prenatal care and childbirth data related to births and neonatal deaths in São Paulo city Brazil ? 2012?2018, [s. l.], 19 jul. 2020.
BRESAN, Rodrigo. Um método para a detecção de ataques de apresentação em sistemas de reconhecimento facial através de propriedades intrínsecas e Deep Learning. Orientador: Me. Carlos Eduardo Beluzo. 2018. Trabalho de Conclusão de Curso (Superior de Tecnologia em Análise e Desenvolvimento de Sistemas) - Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas., [S. l.], 2018.
CABRAL, Cleidy Isolete Silva. Aplicação do Modelo de Regressão Logística num Estudo de Mercado. Orientador: João José Ferreira Gomes. 2013. Projeto Mestrado em Matemática Aplicada à Economia e à Gestão (Mestrado) - Universidade de Lisboa Faculdade de Ciências Departamento de Estatística e Investigação Operacional, [S. l.], 2013. Disponível em: https://repositorio.ul.pt/bitstream/10451/10671/1/ulfc106455_tm_Cleidy_Cabral.pdf. Acesso em: 1 maio 2021.
CAMPOS, Raphael. Árvores de Decisão: Então diga-me, como construo uma?. [S. l.], 28 nov. 2017. Disponível em: https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69. Acesso em: 23 jan. 2021.
Colab. 2021. Disponível em: https://colab.research.google.com/. Acesso em: 20 mar. 2021.
COX, D.R.;SNELL, E.J. Analysis of Binary Data.London: Chapman & Hall, 2ª Edição, 1989. HOSMER, D.W.; LEMESHOW, S.Applied Logistic Regression. New York: John Wiley, 2ª Edição, 2000.
E.França, S.Lansky. Mortalidade infantil neonatal no brasil: situação, tendências e perspectivas Rede Interagencial de Informações para Saúde - demografia e saúde: contribuição para análise de situação e tendências Série Informe de Situação e Tendências(2009), pp.83-112.
62
FREIRE, Sergio Miranda. Bioestatística Básica: Regressão Linear. [S. l.], 20 out. 2020. Disponível em: http://www.lampada.uerj.br/arquivosdb/_book/bioestatisticaBasica.html. Acesso em: 23 jan. 2021.
Jupyter. 2021. Disponível em: https://jupyter.org/. Acesso em: 20 jun. 2021.
GOODFELLOW, I. et al. Deep learning. [S.l.]: MIT press Cambridge, 2016.
Pandas. 2021. Disponível em: https://pandas.pydata.org/. Acesso em: 20 jun. 2021.
Python. 2021. Disponível em: https://www.python.org/. Acesso em: 20 jun. 2021.

KUHN, M.; JOHNSON, K. Applied predictive modeling. [S.l.]: Springer, 2013. v. 26.
LANSKY, S. et al. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante . Cadernos de Saúde Pública, Scielo, v. 30, p. S192 ? S207, 00 2014.
LANSKY, Sônia; FRICHE, Amélia Augusta de Lima; SILVA, Antônio Augusto Moura; CAMPOS, Deise; BITTENCOURT, Sonia Duarte de Azevedo; CARVALHO, Márcia Lazaro; FRIAS, Paulo Germano; CAVALCANTE, Rejane Silva; CUNHA, Antonio José Ledo Alves. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante e ao recém-nascido. [s. l.], Ago 2014. Disponivel em: https://www.scielosp.org/article/csp/2014.v30suppl1/S192-S207/pt/
Matplotlib. 2021. Disponível em: https://matplotlib.org/. Acesso em: 20 mar. 2021.
Maranhão AGK, Vasconcelos AMN, Trindade CM, Victora CG, Rabello Neto DL, Porto D,et al. Mortalidade infantil no Brasil: tendências, componentes e causas de morte no período de 2000 a 2010 . In: Departamento de Análise de Situação de Saúde, Secretaria de Vigilância em Saúde, Ministério da Saúde, organizador. Saúde Brasil 2011: uma análise da situação de saúde e a vigilância da saúde da mulher. v. 1. Brasília: Ministério da Saúde; 2012. p. 163-82.
MESQUITA, PAULO. UM MODELO DE REGRESSÃO LOGÍSTICA PARA AVALIAÇÃO DOS PROGRAMAS DE PÕS-GRADUAÇÃO NO BRASIL. Orientador: Prof. RODRIGO TAVARES NOGUEIRA. 2014. Dissertação (Mestre em Engenharia de Produção) - Universidade Estadual do Norte Fluminense, [S. l.], 2014.
MNASSRI , Baligh. Titanic: logistic regression with python. [S. l.], 1 ago. 2020. Disponível em: https://www.kaggle.com/mnassrib/titanic-logistic-regression-with-python. Acesso em: 14 jan. 2021.
Morais, R.M.d., Costa, A.L: Uma avaliação do sistema de informações sobre mortalidade. Saúde em Debate 41, 101?117 (2017). Disponível em: https://doi.org/10.1109/SIBGRAPI.2012.38.
MOTA, Caio Augusto de Souza; BELUZO , Carlos Eduardo; TRABUCO, Lavinia Pedrosa; SOUZA , Adriano; ALVES, Luciana; CARVALHO, Tiago. DESENVOLVIMENTO DE UMA PLATAFORMA WEB PARA CIÊNCIA DE DADOS APLICADA À SAÚDE MATERNO INFANTIL , [s. l.], 28 nov. 2019.
MULTIEDRO (Brasil). Conheça as aplicações do machine learning na área da saúde. [S. l.], 28 nov. 2019. Disponível em: https://blog.multiedro.com.br/machine-learning-na-area-da-saude/#:~:text=O%20machine%263
0learning%20na%20%C3%A1rea,diagn%C3%B3sticos%20e%20tratamentos%20mais%20precisos. Acesso em: 13 jun. 2021.
MUKHIYA,S.K.;AHMED,U. Hands-On Exploratory Data Analysis with Python: Perform EDA Techniques to Understand, Summarize, and Investigate Your Data. [S.l.]: PacktPublishingLtd,2020.ISBN978-1-78953-725-3.22,32
Murray CJ, Laakso T, Shibuya K, Hill K, Lopez AD. Can we achieve Millennium Development Goal 4? New analysis of country trends and forecasts of under-5 mortality to 2015. Lancet 2007; 370:1040-54.
NumPy. 2021. Disponível em:https://numpy.org/. Acesso em: 20 jun. 2021.
Oliveira, M.M.d., de Araújo, A.S.S.C., Santiago, D.G., Oliveira, J.a.C.G.d., Carvalho, M.D.,

de Lyra et al, R.N.D.: Evaluation of the national information system on live births in brazil , 2006-2010. Epidemiol. Serv. Saúde 24(4), 629?640 (2015).
PELISSARI, ANA CAROLINA CHEBEL. MORTALIDADE NEONATAL DA CIDADE DE SÃO PAULO: UMA ABORDAGEM UTILIZANDO APRENDIZADO DE MÁQUINA SUPERVISIONADO. 2021. Trabalho de Conclusão de Curso (Superior) - Instituto Federal de Educação, Ciência e Tecnologia Campus Campinas, [S. l.], 2021.
REGRESSÃO logística e binary cross entropy - Aula 7. Direção: Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=3J-LBtHVsm4. Acesso em: 21 jan. 2021.
Rmd Nascimento, AJM Leite, NMGSd Almeida, Pcd Almeida, Cfd Silva Determinantes da mortalidade neonatal: estudo caso-controle em fortaleza, ceará, brasil Cad Saúde Pública, 28(2012), pp.559 - 572.
SANTOS, Hellen Geremias. Machine learning e análise preditiva: considerações metodológicas: métodos lineares para classificação. In: SANTOS, Hellen Geremias. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina. 2018. Tese (Pós-Graduação em Epidemiologia) - Faculdade de Saúde Pública da Universidade de São Paulo, [S. l.], 2018.
Scikit-learn. 2021. Disponível em: https://scikit-learn.org/stable/. Acesso em: 20 jun. 2021.
Seaborn. 2021. Disponível em: https://seaborn.pydata.org/. Acesso em: 20 jun. 2021.
SILVA, Wesley; CORSO, Jansen; WELGACZ, Hanna; PEIXE, Julinês. AVALIAÇÃO DA ESCOLHA DE UM FORNECEDOR SOB CONDIÇÃO DE RISCOS A PARTIR DO MÉTODO DE ÁRVORE DE DECISÃO. ARTIGO ? MÉTODOS QUANTITATIVOS, [s. l.], 12 ago. 2008.
WHO, W. H. O.Women and health: today?s evidence, tomorrow?s agenda. [S.l.]: UNWorld Health Organization, 2009. ISBN 9789241563857.
64
APÊNDICE A

==========

Arquivo 1: monografia-V7.docx.pdf (9728 termos)
Arquivo 2: https://vitorborbarodrigues.medium.com/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-recall-quais-as-diferen%C3%A7as-c8f05e0a513c (700 termos)
Termos comuns: 33
Similaridade: 0,31%
O texto abaixo é o conteúdo do documento monografia-V7.docx.pdf (9728 termos)
Os termos em vermelho foram encontrados no documento
https://vitorborbarodrigues.medium.com/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-recall-quais-as-diferen%C3%A7as-c8f05e0a513c (700 termos)

==========

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO
CÂMPUS CAMPINAS
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
CAMPINAS
2021
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas.
Orientador: Prof. Me. Everton Josué da Silva.
CAMPINAS
2021
Dados Internacionais de Catalogação na Publicação
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo Câmpus Campinas.
Aprovado pela banca examinadora em: _____ de _____ de _____.
BANCA EXAMINADORA

Prof. Me. Everton Josué da Silva (orientador)
IFSP Câmpus Campinas

Prof. Me. Carlos Eduardo Beluzo
IFSP Câmpus Campinas

Prof. Dr. Ricardo Barz Sovat
IFSP Câmpus Campinas

Dedico este trabalho aos meus familiares,
colegas de classe, professores e servidores do Instituto
que colaboraram em minha jornada formativa.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pelo dom da vida e pela oportunidade de concluir mais uma etapa de minha experiência acadêmica.

Agradeço a todos os professores e servidores do IFSP Câmpus Campinas, que contribuíram direta e indiretamente para a conclusão deste trabalho.

Agradeço também à minha família, que deu todo o apoio necessário para que eu chegasse até aqui.

Agradeço ao meu orientador que me auxiliou a solucionar as dificuldades encontradas no caminho.

"Minha energia é o desafio, minha motivação é o impossível,
e é por isso que eu preciso
ser, à força e a esmo, inabalável.".
Augusto Branco

RESUMO

A mortalidade infantil pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível em uma população. O aprendizado de máquina é uma área da inteligência artificial que propõe métodos de análise de dados que automatizam a construção de modelos analíticos com base em reconhecimento de padrões, baseado no conceito de que sistemas podem aprender com dados, identificando padrões e tomando decisões com o mínimo de intervenção humana. O objetivo deste trabalho é testar e analisar dois tipos diferentes de algoritmos de aprendizado de máquina, utilizando a base de dados do SIM e SINASC do Brasil, do período de 2016 até 2018, para gerar modelos para predição de mortalidade neonatal. Os métodos utilizados foram Árvore de Decisão e de Regressão Logística e como métricas para avaliar os modelos resultantes foram utilizadas a AUC, Curva ROC e a Matriz de Confusão. Como resultado, apesar de terem alcançado valores de AUC 0.93, ambos modelos de predições acertaram muitas predições de vivos cerca de 671.000 e erraram muitas predições de mortos cerca de 2.600, principalmente devido ao fato de a base estar desbalanceada.

Palavras-chave: Mortalidade Neonatal, Predição, Aprendizado de Máquina.

ABSTRACT

Infant mortality can be used to analyze poverty and socioeconomic levels, as well as measure

the quality of health and medical technology available in a population. Machine learning is an area of artificial intelligence that proposes data analysis methods that automate the construction of analytical models based on pattern recognition, based on the concept that systems can learn from data, identifying patterns and making decisions with a minimum of human intervention. The objective of this work is to test and analyze two different types of machine learning algorithms, using the SIM and SINASC do Brasil database, from 2016 to 2018, to generate models for predicting neonatal mortality. The methods used were Decision Tree and Logistic Regression and as metrics to evaluate the resulting models, AUC, ROC Curve and Confusion Matrix were used. As a result, despite achieving values of AUC 0.93, both prediction models correct many live birth predictions around 671.000 and miss many stillbirth predictions around 2600, mainly due to the fact that the base is unbalanced.
Keywords: Neonatal Mortality, Prediction, Machine Learning.
LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS
SIM Sistema de Informação sobre Mortalidade
SINASC Sistema de Informações sobre Nascidos Vivos
TMI Taxa de Mortalidade Infantil
TMN Taxa de Mortalidade Neonatal
MN Mortalidade Neonatal
ML Machine Learning
DNV Declaração de Nascido Vivo
VN Verdadeiro Negativo
FN Falso Negativo
VP Verdadeiro Positivo
FP Falso Positivo
ROC Curva Característica de Operação do Receptor
NaN Não é um Número
RFE Eliminação Recursiva de Feature
RFECV Eliminação Recursiva de Feature com Validação Cruzada
SUMÁRIO

13
1 INTRODUÇÃO

A taxa de mortalidade infantil (TMI) é uma importante medida de saúde em uma população e é um grande problema no mundo todo. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível. Esta taxa é dividida em duas categorias: neonatal e pós-neonatal. É categorizada neonatal quando o óbito ocorre nos 28 primeiros dias de vida após o pós-parto, e o pós-neonatal é quando o óbito ocorre entre 29 e 364 dias de vida (BELUZO et al., 2020). Cerca de 46% das mortes no mundo acontecem entre os menores de cinco anos e grande parte está concentrada nos primeiros dias de vida (LANSKY, 2014). A diminuição dessas taxas é muito importante no mundo todo, refletindo nos indicadores de saúde pública e desenvolvimento do país.

Os fatores associados à mortalidade são profundamente influenciados pelas características biológicas maternas e neonatais, pelas condições sociais e pelos cuidados

prestados pelos serviços de saúde (E. FRANÇA; S. LANSKY, 2009; R.M.D. NASCIMENTO, 2012). Em grande parte, o diagnóstico é altamente dependente da experiência adquirida pelo profissional que o realiza. Apesar de indiscutivelmente necessário, não é perfeito, principalmente pelo fato de ser dependente de fatores humanos, por este motivo os profissionais sempre tiveram recursos tecnológicos para auxiliar nessa tarefa (BELUZO et al., 2020). Segundo estudos de 2010 no Brasil, calcula-se que aproximadamente 70% dos óbitos infantis ocorridos poderiam ter sido evitados por ações de saúde e que 60% dos óbitos neonatais ocorreram por situações que poderiam ser evitadas (BARRETO; SOUZA; CHAPMAN, 2015).

No presente trabalho foram utilizados dois algoritmos de aprendizado de máquina para criar modelos de predição de mortalidade neonatal. Além disso, foi realizada também uma análise exploratória da base de dados. Para este trabalho foram utilizadas duas fontes de dados: Sistema de Informação sobre Mortalidade (SIM) e Sistema de Informações sobre Nascidos Vivos (SINASC) do Brasil inteiro.

14

2 JUSTIFICATIVA

Com o avanço da tecnologia, podemos notar que ela está cada vez mais presente no nosso dia a dia e nos auxilia em diversas tarefas do cotidiano, e vem sendo aplicada em diversas áreas, dentre elas a saúde.

A mortalidade infantil é uma preocupação mundial na saúde pública, a ONU (Organizações das Nações Unidas) definiu a redução da mortalidade infantil como uma meta para o desenvolvimento global. A mortalidade neonatal é responsável por aproximadamente 60% da TMI nos países em desenvolvimento (A.K. SINGHA, 2016). Existem diversos fatores que podem contribuir com a redução de óbitos neonatais, como por exemplo acompanhamento médico à gestante no período de gravidez, acompanhamento médico ao recém-nascido nos primeiros dias de vida e a disponibilização deste serviço com qualidade e proficiência (WHO, 2009).

A diminuição dessa taxa é importante e de interesse para o mundo todo, e utilizar da tecnologia para auxiliar nessa diminuição é uma proposta funcional, e inovadora para a realidade brasileira (MOTA et al., 2019). Havendo disponibilidade de tecnologia para auxiliar nas decisões dos diagnósticos, os profissionais da saúde podem focar nos cuidados do recém-nascido com risco de vir a óbito, fornecendo tratamentos melhores.

Segundo o site Multiedro (2019), um grande problema que os médicos enfrentam ao realizar o diagnóstico, é analisar um grande volume de dados. Para a área médica obter diagnósticos rápidos e precisos pode salvar vidas, e a utilização de machine learning para realizar esses processos é importante, com a ML os dados podem ser processados em questões de segundos e resultar em um diagnóstico mais rápido, além disso utilizando bons modelos ML os diagnósticos serão mais precisos.

Diversos trabalhos vêm sendo desenvolvidos neste contexto. No trabalho realizado por BELUZO et al. (2020a), os autores utilizaram uma base de dados com informações da cidade de São Paulo para criação de modelos de aprendizado de máquina para predição de mortalidade neonatal. Este recorte foi utilizado também no presente trabalho para fins de exercício e definição de etapas da implementação de código. Na conclusão os autores discutem os fatores que tiveram mais influência nas predições, e na análise feita pelos autores,

as consultas de pré-natal são muito importantes para a redução da mortalidade infantil, uma vez que algumas características muito importantes, como a malformação congênita, podem ser detectadas ao longo das consultas de pré-natal.
Em um outro estudo realizado por BELUZO et al. (2020a), foi realizada uma análise exploratória da base de dados SPNeoDeath, onde e podemos observar detalhadamente cada
15
análise das colunas da tabela e diversos gráficos feito para um melhor entendimento dos dados.
Outro trabalho que foi realizado utilizando a base de dados com dados somente de São Paulo foi o de PELISSARI (2021), nesse trabalho é realizado uma análise exploratória na base de dados de São Paulo e é também analisado três algoritmos de aprendizado de máquina (Logistic Regression, Random Forest Classifier e XGBoost), nesse trabalho a autora conclui que os modelos de Logistic Regression e XGBoost foram os modelos que apresentaram os melhores desempenhos preditivos, e também mostra os problemas de estar utilizando uma base de dados desbalanceada.
O problema da mortalidade neonatal não é recente, e o mundo todo desenvolve iniciativas para lidar com ele. A ONU definiu como meta a redução da mortalidade infantil para o desenvolvimento global. Diversos fatores podem ajudar na diminuição da taxa de mortalidade neonatal como acompanhamento médico antes e após parto, tanto com a mãe quanto com o recém-nascido. Um serviço de qualidade e constante para os casos com maior risco de óbito pode salvar diversos recém-nascidos. Neste sentido, o desenvolvimento de ferramentas para a predição de mortalidade neonatal pode colaborar com esta iniciativa.

3 OBJETIVOS

3.1 OBJETIVO GERAL

O presente trabalho tem como objetivo testar e analisar os resultados da aplicação dos aprendizagem de máquina supervisionado Árvore de Decisão e Regressão Logística, utilizando dados do SIM e do SINASC, no período de 2016 até 2018, a fim de gerar modelos de predição de risco morte neonatal.

3.2 OBJETIVOS ESPECÍFICOS

? Realizar análise exploratória da base dados a ser utilizada na construção dos modelos;
? Implementar modelos de predição utilizando algoritmos Árvore de Decisão e Regressão Logística;
? Avaliar resultados e propor novas abordagens.
16

4 FUNDAMENTAÇÃO TEÓRICA

4.1 MORTALIDADE INFANTIL E NEONATAL

Como dito na monografia PELISSARI (2021), a TMI consiste no número de crianças que foram a óbito antes de concluir um ano de vida dividido por 1000 crianças nascidas vivas no tempo de um ano. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos e qualidade da saúde pública de um país (apud BELUZO et al., 2020).
Mencionado em PELISSARI (2021), a TMN é uma das duas categorias da TMI, é categorizado mortalidade neonatal quando o óbito ocorre do nascimento da criança até os 28 primeiros dias de vida. A mortalidade neonatal pode ser subdividida em mortalidade neonatal precoce que é quando o óbito ocorre entre 0 e 6 dias de vida, e o neonatal tardio quando o

óbito ocorre entre 7 e 28 dias de vida (apud E. FRANÇA e S. LANSKY, 2009).
Segundo Lansky et al. (2014), a taxa de mortalidade neonatal é o principal componente da TMI desde a década de 1990 no Brasil e vem se mantendo em níveis elevados, com taxa de 11,2 óbitos por mil nascidos vivos em 2010 (apud Maranhão et al., 2012).
Também é dito em Lansky et al. (2014), que a TMI do Brasil em 2011 foi 15,3 por mil nascidos vivos, alcançando a meta 4 dos objetivos de desenvolvimento do milênio, compromisso dos governos integrantes das Nações Unidas de melhorar a saúde infantil e reduzir em 2/3 a mortalidade infantil entre 1990 e 2015. (apud Maranhão et al., 2012; apud Murray et al., 2015). Segundo Lansky et al. (2014) o principal componente da TMI em 2009 é a mortalidade neonatal precoce, e grande parte das mortes infantis acontece nas primeiras 24 horas (25%), indicando uma relação estreita com a atenção ao parto e nascimento (apud E. FRANÇA e S. LANSKY, 2009).

4.2 MÉTODOS DE APRENDIZADO DE MÁQUINA

A utilização de dados para a resolução de problemas, de modo a se identificar padrões ocultos e posteriormente auxiliar na tomada de decisões é definida como aprendizado de máquina (BRESAN, 2018 apud Brink et al. 2013).
O uso de técnicas de Aprendizado de Máquina se mostra altamente eficiente na resolução de tarefas que se apresentem difíceis de se resolver a partir de programas escritos por humanos (BRESAN, 2018 apud GOODFELLOW et al., 2016), bem como também possibilita uma maior compreensão sobre os funcionamentos dos princípios que compõem a inteligência humana.
17
Neste trabalho serão implementados modelos utilizando os algoritmos de Árvore de Decisão e Regressão Logística, os quais serão apresentados a seguir.

4.2.1 Árvore de Decisão

Segundo Batista e Filho (2019), os modelos preditivos baseados em árvores são geralmente utilizados para tarefas de classificação, embora também possam ser utilizados para tarefas de regressão. Considerado um dos mais populares algoritmos de predição, o algoritmo de árvore de decisão proporciona uma grande facilidade de interpretação.
As árvores de decisão utilizam a estratégia ?dividir-e-conquistar", na qual as árvores são construídas utilizando-se apenas alguns atributos. A árvores de decisão é uma das técnicas por meio da qual um problema complexo é decomposto em subproblemas mais simples. Recursivamente, a mesma estratégia é aplicada a cada subproblema (SILVA et al, 2008).
A árvore de decisões tem uma estrutura hierárquica onde cada nó da árvore representa uma decisão em uma das colunas do conjunto de dados, cada ramo da árvore representa uma tomada de decisão (BATISTA; FILHO, 2019).
O algoritmo segue algumas decisões para montar a árvore, o atributo mais importante é utilizado como nó raiz e será o primeiro nó, já os atributos menos importantes são mostrados nos ramos da árvore. Cada atributo é verificado para conferir se é possível gerar uma árvore menor e se é possível fazer uma classificação melhor, e assim é escolhido o nó que produz nós filhos (SILVA et al, 2008).
Existem diferentes tipos de algoritmo para a construção da árvore, como por exemplo, ID3 (Iterative Dichotomiser), C4.5 e CART (Classification and Regression Tree). O algoritmo ID3 escolhe os nós da árvore por meio da métrica do ganho de informação. Já o CART faz

uso da equação de Gini (BATISTA; FILHO, 2019 apud KUHN; JOHNSON, 2013).
O algoritmo utilizado neste trabalho usará o cálculo da entropia e ganho de informação, para decidir qual atributo será usado no nó, a entropia é uma medida da desorganização dos sistemas, maior é a incerteza do sistema, quanto maior a entropia maior está a desorganização, com a árvore de decisão a ideia é ir organizando o sistema e ir separando as decisões da melhor forma para que a entropia diminui e o ganho de informação aumente, para que a decisão do modelo fiquei mais assertiva. O cálculo da entropia utiliza a seguinte equação (ÁRVORE, 2020):
???????? =
?
?? ?? ???
2
??
18
Nesta equação o i representa possíveis classificações (rótulos), e o P é a probabilidade de cada uma. A ideia da árvore então é verificar qual é a entropia para cada uma das variáveis da base de dados, e verificar qual é a melhor organização de cada uma das variáveis. E isso é realizado através da técnica chamada de ganho de informação, essa técnica utiliza a equação a seguir (ÁRVORE, 2020):
????? = ????????(???) ? ? ????(??????) * ????????(??????)
Então o ganho de informação é calculado pela entropia do nó pai menos a soma do peso vezes a entropia dos nós filhos. Esse cálculo é realizado para todas as variáveis e a variável que tiver maior ganho de informação é utilizado para a decisão do nó. Esse processo é realizado recursivamente e a árvore vai sendo montada com base nesses cálculos (ÁRVORE, 2020).
Na Figura 1 temos a representação do início de uma árvore de decisão, no caso em questão o autor estava classificando exames positivos e negativos para o vírus SARS-CoV-2, ele utilizou uma base de dados que contém quase 6 mil exames realizados, contendo campos como o resultado do exame, idade do paciente, níveis de hemoglobina entre outras informações.
Figura 1. Exemplo de Árvore de Decisão. (Fonte: ÁRVORE 2020).
Pode-se observar que para o nó raiz foi utilizado a variável Leukocytes, logo foi o que apresentou a melhor organização para ser o nó raiz, e após ele temos os ramos da árvore. Cada retângulo da árvore é uma decisão do modelo e esses retângulos têm atributos importante como, a coluna da base de dados que será onde vai ser realizado a decisão, então pegando o nó raiz, pacientes que tiverem com os Leukocytes para menos que -0.43 seguiram o caminho
19
para a esquerda (true), já os pacientes que tiverem mais seguiram o caminho da direita (false). A entropia calculada do nó também é apresentada e a classificação do nó também, então se o resultado parar neste nó, a classificação que está nele será a classificação final, e temos o samples que é o número de amostras que se enquadram no na decisão. E esse processo de verificação do modelo vai se repetindo até não ter mais como dividir em subproblemas ou até chegar na profundidade máxima.
Uma característica importante da árvore de decisão é que ela é um modelo WhiteBox,

ou seja, é possível visualizar a árvore montada e as decisões que o modelo terá que tomar na árvore. Na Figura 1 pode-se ver uma árvore de decisões montada, cada retângulo da árvore é uma decisão que o modelo terá que tomar, conforme o modelo toma as decisões ele vai seguindo um caminho até chegar ao nó folha, nó folha é o nó que não contém nó filho, não tem mais ramos para baixo, chegando ao nó folha o modelo tem a classificação final.

4.2.2 Regressão Logística

O modelo de regressão logística estabelece uma relação entre a probabilidade de ocorrência da variável de interesse e as variáveis de entrado do modelo, sendo utilizada para tarefas de classificação, em que a variável de interesse é categórica, neste caso, a variável de interesse assume valores 0 ou 1, onde 1 é geralmente utilizado para indicar a ocorrência do evento de interesse (Batista e Filho, 2019).

De acordo com Mesquita (2014), a técnica de regressão logística, desenvolvida no século XIX, obteve maior visibilidade após 1950 ficando então mais conhecida. Ficou ainda mais difundida a partir dos trabalhos de Cox & Snell (1989) e Hosmer & Lemeshow (2000). Caracteriza-se por descrever a relação entre uma variável dependente qualitativa binária, associada a um conjunto de variáveis independentes qualitativas ou métricas.

Esta técnica inicialmente foi utilizada na área médica, porém a eficiência viabilizou sua implementação nas mais diversas áreas. Mesquita (2014) diz que o termo regressão logístico, tem sua origem na transformação usada com a variável dependente, que permite calcular diretamente a probabilidade da ocorrência do fenômeno em estudo.

Segundo Santos (2018), para estimar a probabilidade, é utilizado uma técnica chamada distribuição binomial para modelar a variável resposta do conjunto de treinamento, que tem como parâmetro a probabilidade de ocorrência de uma classe específica. Uma característica importante desse modelo é que a probabilidade estimada deve estar limitada ao intervalo de 0 a 1. O algoritmo de Regressão Logística gera um modelo de classificação binária (0 e 1). O modelo utiliza a Regressão Linear como base, a equação linear é (REGRESSÃO, 2020):

20

? = ? . ? + ?

O ?m? da equação é o coeficiente angular da reta, e o ?b? é o intercepto. Então, aplicando a equação linear obtemos um valor contínuo como resultado, após obter esse valor a Regressão Logística tem que realizar a classificação do resultado, então é aplicado uma função chamada sigmoide (REGRESSÃO, 2020):

? = 1

1+ ? ??

Essa função chamada de sigmoide, ou função logística, é responsável por ?achatar? o resultado da regressão linear, calculando a probabilidade de o resultado pertencer a classe 1, classificando assim o resultado da função linear em 0 ou 1.

Segundo Batista e Filho (2019), o modelo de regressão logística que tem como objetivo realizar uma classificação binária de uma variável de interesse irá predizer a probabilidade de pertencer à classe positiva. Tem-se, portanto, que é dado por:?

? = { 0, ?? ???? (? = 1 ??) < 0, 5

ou

? = { 1, ?? ???? (? = 1??) ? 0, 5

Então a decisão do modelo de regressão logística está relacionada à escolha de um

ponto de corte para p(x). Caso o ponto de corte escolhido for p(x)=0,5, os resultados que forem p(x)>0,5 serão classificados como 1 e os resultados p(x)<5 serão classificados como 0 (SANTOS, 2018).

A Figura 2 ilustra um exemplo de como é o diagrama de Regressão Logística. Pelo diagrama então pode-se observar que o modelo possui uma representação gráfica em formato de ?S?, essa seria a curva logística. As predições do modelo sempre ficaram na linha 1 e 0 do eixo y, pois é o intervalo estabelecido pelo algoritmo, então a classificação sempre será nesse intervalo.

21

Figura 2. Exemplo de Diagrama de Regressão Linear. (Fonte: Batista e Filho 2019).

4.2.3 Treino, teste e validação do modelo de aprendizado de máquina

A criação de um modelo de aprendizado de máquina é realizada em duas etapas: treinamento do modelo, que é realizado a partir dos dados fornecidos para o algoritmo, e etapa de teste, onde os modelos recebem dados novos e sem o rótulo, para que seja avaliado a performance do modelo (PELISSARI, 2021).

Para o treinamento dos modelos foram utilizadas duas abordagens, a primeira onde a base foi particionada entre conjunto de dados para teste e para treino em uma porcentagem de 90% para treino e 10% para teste, porém dessa forma pode ocorrer problema de sobreajuste ou overfitting. Nesta situação, o que pode ocorrer é o modelo não aprender com os dados, e apenas ?decorar? os dados recebidos e quando é realizado o teste o modelo consegue predizer apenas situações parecidas, não sendo capaz de identificar padrões com diferenças mínimas.

Para evitar este problema, foi utilizado uma técnica de validação cruzada chamada K-folds cross-validation. Essa técnica de validação cruzada divide a base de dados em K subconjuntos, e então são realizadas várias rodadas de treinamento e teste alternando os subconjuntos utilizados para teste e treino, e o resultado do modelo baseia-se na média da acurácia observada em cada rodada (PELISSARI, 2021).

Pode-se observar na Tabela 1 uma ilustração da técnica K-fold cross-validation, onde temos um exemplo onde a base é dividida em 5 subconjuntos e cada subconjunto tem a parte de teste diferente dos outros subconjuntos, assim o modelo vai receber valores novos de teste e treino todas as vezes que executar um novo subconjunto.

22

Tabela 1. Exemplo da separação K-fold cross-validation. (Fonte: Adaptado de PELISSARI, 2021).

Predição 1	Predição 2	Predição 3	Predição 4	Predição 5
Teste	Treino	Treino	Treino	Treino
Treino	Teste	Treino	Treino	Treino
Treino	Treino	Teste	Treino	Treino
Treino	Treino	Treino	Teste	Treino
Treino	Treino	Treino	Treino	Teste

4.2.4 Métricas para avaliação de modelos

Para fins de avaliação, neste trabalho foram utilizadas três métricas para avaliar o desempenho dos modelos: a Matriz de Confusão e a AUC (Area under Receiver Operating Characteristic Curve - ROC). Na matriz de confusão pode-se observar o comportamento do modelo em relação a cada uma das classes a serem preditas pelo modelo, utilizando variáveis

binárias (positivo: 1; e negativo: 0), para apresentar uma matriz que faz uma comparação entre as predições do modelo e os valores corretos do conjunto de dados (PELISSARI, 2021).
Na Tabela 2 temos um exemplo de uma matriz de confusão, a qual consiste em duas colunas e duas linhas, e os valores são atribuídos nos quadrantes com os seguintes resultados (PELISSARI, 2021):
? quando o valor real do conjunto de dados é 0, e a predição do modelo também classificou como 0, então temos um Verdadeiro Negativo (VN);
? quando o valor real é 1, e o modelo classificado como 0, temos um Falso Negativo (FN);
? quando o valor real é 0, e o modelo classificado como 1, então o resultado se enquadra no quadrante de Falso Positivo (FP);
? e por fim o quadrante Verdadeiro Positivo (VP), que são os resultados que tem o valor real 1, e o modelo classificado como 1.
23
Tabela 2. Modelo de Matriz de confusão. (Fonte: Adaptado de PELISSARI, 2021).
Valor Predito
Negativo (0) Positivo (1)
Classe
Real
Negativo
(0)
Verdadeiro
Negativo
Falso
Positivo
Positivo
(1)
Falso
Negativo
Verdadeiro
Positivo
A segunda métrica de avaliação utilizada foi a Curva ROC que permite avaliar o desempenho de um modelo com relação às predições efetuadas. A curva ROC é um gráfico de Sensibilidade (taxa de verdadeiros positivos) versus taxa de falsos positivos, a representação da curva ROC permite evidenciar os valores para os quais existe otimização da Sensibilidade em função da Especificidade (CABRAL, 2013).
Na Figura 3 temos um exemplo de uma curva ROC. A linha traçada na cor preta é uma linha de referência, ela representa hipoteticamente a curva ROC de um classificador puramente aleatório. Caso a linha laranja, que representa o resultado do modelo que está sendo avaliado, ficasse igual a linha tracejada preta, indicaria que o modelo avaliado estaria predizendo aleatoriamente os resultados.
Figura 3: Exemplo de Curva ROC. (Fonte: Elaboração Própria).
Por uma análise de inspeção visual, quanto mais próximo a linha laranja estiver do valor 1 no eixo Y, melhor está a capacidade do modelo para diferenciar as classes. O valor da

24
AUC representa a capacidade do modelo de diferenciar as classes, quanto maior o valor do AUC, melhor é a predição realizada pelo modelo, uma vez que, os valores resultantes iguais a 0 são realmente 0, e os iguais a 1 são de fato 1 (PELISSARI, 2021).

5 METODOLOGIA

O desenvolvimento deste trabalho foi realizado em três etapas: (1) Pré-processamento e Análise Exploratória do conjunto de dados, onde foram aplicadas técnicas comuns para preparar o conjunto de dados como tratamento de valores nulos e seleção de variáveis de interesse; (2) Implementação de modelos de aprendizado de máquina supervisionado para predição das mortes neonatais, utilizando os algoritmos de árvore de decisão e regressão logística; e (3) Análise de Resultados, que será realizada uma análise do desempenho do modelo.

5.1 TECNOLOGIAS E FERRAMENTAS

Neste trabalho vamos utilizar a linguagem de programação Python (PYTHON, 2021) pela sua simplicidade de codificação e alto poder de processamento. Foi utilizado também a plataforma Google Colab Research (Colab, 2021), para o desenvolvimento dos algoritmos e treinamento dos modelos. Além disso, foi utilizado também o Jupyter Notebook (Jupyter, 2021), instalado localmente em computador pessoal, pois o Google Colab Research possui limitação de recursos de memória e processamento, então as duas plataformas foram usadas para realização deste trabalho. As principais bibliotecas utilizadas foram: numpy (NumPy, 2021), e pandas (Pandas, 2021), para a manusear os dados, matplotlib (Matplotlib, 2021), e seaborn (Seaborn, 2021), para a plotagem dos gráficos, e por fim a biblioteca sklearn (Scikit-Learn, 2021), que é a biblioteca responsável pelos algoritmos de aprendizado de máquina.

5.2 BASE DE DADOS

As bases de dados a serem utilizadas serão o SIM e SINASC, que são as duas principais fontes de informações sobre nascimentos e óbitos no Brasil. O SINASC é alimentado com base na Declaração de Nascido Vivo (Declaração de Nascido Vivo - DNV) (BELUZO et al.,2020b apud Oliveira et al., 2015). O SIM tem como objetivo principal apoiar a coleta, armazenamento e processo de gestão de registros de óbitos no Brasil (BELUZO et
25
al.,2020b apud Morais et al., 2017), e foi usado para rotular o óbito registrado no SIM, utilizando o campo DNV como chave de associação. O SIM será utilizado para rotular os registros do SINASC, pois o SIM coletas informações sobre mortalidade e é usado como base para o cálculo de estatísticas vitais, como a taxa de mortalidade neonatal e o SINASC reúne informações sobre dados demográficos e epidemiológicos do bebê, da mãe, do pré-natal e do parto (BELUZO et al.,2020b).

A Tabela 3 apresenta as variáveis da base de dados. Essa base tem variáveis que contêm valores quantitativos e categóricos. A coluna ?num_live_births? por exemplo contém o número de nascidos vivos anteriores da mãe e é composta por valores quantitativos contínuos, e a coluna ?tp_pregnancy? utiliza valores nominais categóricos que indicam o tipo de gravidez.

Tabela 3. Colunas da base de dados e suas descrições. (Fonte: Adaptado de BELUZO et al.,2020b).

Nome da coluna Descrição Domínio de dados
Variáveis demográficas e socioeconômicas
maternal_age Idade da mãe Quantitativo Contínuo (inteiro)
tp_maternal_race Raça / cor da pele da
mãe
Nominal categórico (inteiro)
1 - branco; 2 - Preto; 3 -
amarelo; 4 - Pele morena; 5 -
Indígena.
tp_marital_status Estado civil da mãe Nominal categórico (inteiro)
1 - Único; 2 - Casado; 3 - Viúva;
4 - Separados / divorciados
judicialmente; 5 - Casamento
por união estável; 9 - Ignorado.
tp_maternal_schooling Anos de escolaridade da
mãe
Nominal categórico (inteiro)
1 - nenhum; 2 - de 1 a 3 anos; 3 -
de 4 a 7 anos; 4 - de 8 a 11 anos;
5 - 12 e mais; 9 - Ignorado.
Variáveis obstétricas maternas
num_live_births Número de nascidos
vivos
Quantitativo Contínuo (inteiro)
num_fetal_losses Número de perdas fetais Quantitativo Contínuo (inteiro)
num_previous_gestations Número de gestações Quantitativo Contínuo (inteiro)
26
anteriores
num_normal_labors Número de partos
normais (trabalhos de
parto)
Quantitativo Contínuo (inteiro)
num_cesarean_labors Número de partos
cesáreos (partos)
Quantitativo Contínuo (inteiro)
tp_pregnancy Tipo de gravidez Nominal categórico (inteiro)
1 - Singleton; 2 - Twin; 3 -
Trigêmeo ou mais; 9 - Ignorado.
Variáveis relacionadas a cuidados anteriores
tp_labor Tipo de parto infantil
(tipo de parto)
Categórico Nominal (inteiro)
1 - Vaginal; 2 - Cesariana;
num_prenatal_appointments Número de consultas de

pré-natal
Quantitativo Contínuo (inteiro)
tp_robson_group Classificação do grupo
Robson
Ordinal categórico (inteiro)
Variáveis relacionadas ao recém-nascido
tp_newborn_presentation Tipo de apresentação de
recém-nascido
Categórico Nominal (inteiro)
1 - Cefálico; 2 - Pélvico ou
culatra; 3 - Transversal; 9 -
Ignorado.
has_congenital_malformation Presença de
malformação congênita
Nominal categórico (inteiro)
1 - Sim; 2 - Não; 9 - Ignorado
newborn_weight Peso ao nascer em
gramas
Quantitativo Contínuo (inteiro)
cd_apgar1 Pontuação de Apgar de 1
minuto
Ordinal categórico (inteiro)
cd_apgar5 Pontuação de Apgar de 5
minuto
Ordinal categórico (inteiro)
gestaional_week Semana de gestação Quantitativo Contínuo (inteiro)
tp_childbirth_care Assistência ao parto Categórico Nominal (inteiro)
1 - Doutor; 2 - enfermeira ou
obstetra; 3 - Parteira; 4 - outros;
9 - Ignorado.
was_cesarean_before_labor Foi cesáreo antes do
parto.
27
was_labor_induced Foi induzido ao trabalho
de parto.
is_neonatal_death Morte antes de 28 dias
(rótulo)
Nominal categórico (inteiro)
0 - sobrevivente; 1 - morto.

Neste trabalho foi utilizado dados do período de 2014 à 2016 devido sua melhor qualidade. Este recorte possui 6.760.222 registros dos quais 99.4% são amostras de recém-nascidos vivos e 0.6% são amostras onde os recém-nascidos vieram a óbito conforme pode ser observado na Figura 4. Com essas informações consegue-se observar que a base de dados é desbalanceada.

Figura 4. Distribuição da base de dados considerando se o recém-nascido viveu ou morreu durante o período neonatal. (Fonte: Elaboração Própria).

5.3 PREPARAÇÃO DA BASE DE DADOS

Nessa etapa foi realizada a preparação da base de dados para o treinamento dos modelos que consiste na limpeza, transformação e preparação dos dados a serem utilizados na análise exploratória e na criação dos modelos preditivos. A base de dados pode conter informações desnecessárias para o modelo que pode acabar atrapalhando as predições,

28

algumas colunas podem ser retiradas ou seus valores podem ser modificados, por exemplo valores reais são modificados para inteiros.

5.4 DESENVOLVIMENTO DOS MODELOS DE APRENDIZADO DE MÁQUINA

Como já dito anteriormente na seção 4 os algoritmos de aprendizado de máquina escolhidos para esse trabalho foram os: Regressão Logística e Árvore de Decisão que utilizam a lógica mostrada na seção 4 , Fundamentação Teórica. Esses dois foram escolhidos por alguns motivos, ambos são algoritmos de classificação e supervisionados, atendendo os critérios necessários para a predição de mortalidade neonatal, ambos os algoritmos são bons para realizar predições, a Árvore de Decisão inclusive é um ótimo modelo para analisar as decisões que estão sendo feitas pelo modelo, afinal esse algoritmo tem a característica de ser um WhiteBox, ou seja conseguimos ver o que o modelo aprendeu e o que ele está decidindo. Esses algoritmos selecionados são algoritmos que se encaixam na necessidade deste trabalho e são muito utilizados na comunidade acadêmica.

Para o algoritmo de Regressão Logística, como foi dito anteriormente na seção ?Preparação da base de dados? a base de dados foi tratada e particionada, no particionamento foi utilizado 90% da base para o treino e 10% para os testes e então foi realizado o treinamento do modelo, também foi aplicada a função RFECV (Eliminação Recursiva de Feature com Validação Cruzada), esse função executa a RFE (Eliminação Recursiva de Feature), que tem como ideia selecionar features considerando recursivamente conjuntos cada vez menores de features, isso ocorre da seguinte forma. Primeiro, o estimador é treinado no conjunto inicial de features e a importância de cada feature é obtida por meio de um atributo "coef" ou por meio de um atributo "feature_importances". Em seguida, as features menos importantes são removidas do conjunto atual de features. Esse procedimento é repetido recursivamente no conjunto removido até que o número desejado de features a serem selecionados seja finalmente alcançado. Porém o diferencial da RFECV é que essa função executa a RFE em um loop de validação cruzada para encontrar o número ideal ou o melhor número de features.

Após essa seleção, foi realizado o treinamento do modelo usando todas as features e também treinamos usando as features que o RFECV selecionou, para compararmos os resultados no final. Também foi utilizado o método de K-folds cross-validation em um novo treinamento para conferirmos se o modelo estava sofrendo overfitting (sobreajuste). E por fim aplicamos um método chamado GridSearchCV, esse método permite que seja realizado testes

29

alterando algumas combinação de parâmetros nos nossos modelos, facilitando para achar a melhor combinação, esse método é parecido com o cross validation, o diferencial do Grid Search é que ele faz os cross validation de vários modelos com hiperparâmetros diferentes de

uma só vez.
Para o algoritmo de Árvore de Decisão também foi utilizado o mesmo particionamento de 90% para treino e 10% para teste, na criação do modelo foi colocado a entropy como critério de decisão e uma profundidade máxima de 4, pois com números maiores o modelo ficava muito complexo e ele errava mais as predições, também utilizamos um algoritmo para mostrar a importância de cada Feature para o modelo.
O algoritmo de Regressão Logística e análise exploratória foi baseado em códigos disponível na plataforma web Kaggle (MNASSRI, 2020), no exemplo do Kaggle o autor faz os códigos para predizer mortes no navio Titanic, para o trabalho de mortalidade neonatal os códigos foram alterados para realizar predição de mortes neonatais. Já no algoritmo de Árvore de Decisão foi baseado em uma vídeo aula do professor Diogo Cortiz (ÁRVORE..., 2020), nessa vídeo aula o professor explica sobre os algoritmos de Árvore de Decisão e depois implementa o um exemplo de código, o código usado neste trabalho usou o codigo do professor Diogo Cortiz, e foi alterado para realizar predições de mortalidade neonatal.

5.5 ANÁLISE DOS RESULTADOS

Para a análise de resultados foi utilizado dois métodos que é o de Matriz de confusão e a curva ROC esses métodos são explicados na seção 4, a Fundamentação Teórica. Com esses métodos pode-se realizar uma avaliação do modelo, e assim será possível analisar os valores de acurácia, precisão e recall do modelo, essas métricas são utilizadas avaliar o desempenho do modelo, com a acurácia é possível calcular a proximidade entre o valor obtido na predição dos modelos e os valores esperados, com a precisão é possível analisar as amostras que o modelo classificou como 0 eram realmente 0 na classe real e com o recall é possível analisar as amostras que o modelo conseguiu reconhecer como a classe correta.

5.6 ACESSO A BASE DE DADOS E CÓDIGOS

A base de dados utilizada neste projeto pode ser encontrada no link: ... . E os códigos podem ser encontrados no link: ... .

30

6 RESULTADOS

Os mesmos experimentos mostrados abaixo foram aplicados para uma base de dados que contém somente dados de São Paulo, essa base é um recorte da base do Brasil todo em contém cerca de 1.400.000 amostras, esse recorte da base também é bem desbalanceado. Os resultados obtidos nesse experimento usando o recorte de São Paulo foram bem parecidos com os experimentos da base do Brasil todo, e os códigos desse experimento podem ser visualizados no apêndice X.

6.1 ANÁLISE EXPLORATÓRIA

O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

6.1.1 Apresentação dos Dados

Como dito anteriormente na seção 5.2 ?Base de Dados? deste trabalho, a base de dados é formada pelas informações do SIM e do SINASC contém 6.760.222 amostras e 29 colunas, porém serão utilizadas somente 23 colunas sendo elas as colunas descritas na Tabela 3. Na seção 5.2 também tem a Figura 4 que mostra que a base dados é desbalanceada tendo 99.4% das amostras de recém-nascidos vivos e 0.6% das amostras onde os recém-nascidos vieram a óbito no período neonatal.

6.1.1.1 Caracteristicas demográficas e socioeconômicas maternas
O apêndice X contém uma tabela que possui as estatísticas sumarizadas das variáveis demográficas e socioeconômicas maternas. Nesta tabela por exemplo podemos observar a coluna ?maternal_age? que contém a idade da mãe, e na tabela mostra que a média da idade das mães é de 26.4 anos, e os dados possuem uma variação de no mínimo 8 anos e no máximo 55 anos, e a mediana é 26 anos. Na tabela também tem as colunas: ?tp_maternal_schooling? que mostra a categoria de anos de escolaridade da mãe, ?tp_marital_status? que mostra o estado civil da mãe em dados categóricos e a coluna ?tp_maternal_race? que mostra a raça da mãe também em dados categóricos.
31
6.1.1.2 Variávies obstétricas maternas
No apêndice X pode-se observar uma tabela que possui as estatísticas sumarizadas das variáveis obstétricas maternas. Nesta tabela por exemplo temos a coluna ?num_live_births?, essa coluna mostra o número de nascidos vivos anteriores que a mãe obteve, a média desta coluna é 0.94 , a mediana é 1, o número mínimo é 0 e o máximo é 10 nascidos vivos. Na tabela podemos observar também a coluna, ?num_fetal_losses?, esta coluna mostra o número de perdas fetais anteriores da mãe, a média desta coluna é 0.21, a medina é 0, o número mínimo é 0 e o máximo é 5, esses valores mostram que poucas mães tiveram perdas fetais antes da gravidez atual. Também na tabela tem estatísticas da coluna ?num_previous_gestations? esta coluna mostra o número de gestações anteriores da mãe, esta coluna tem a média de 1.14, a mediana é 1, o número mínimo é 0 e o máximo 10. Essa tabela também contém as colunas: ?num_normal_labors? que é o número de partos normais da mãe, ?num_cesarean_labor? que mostra o número de partos cesáreos da mãe e por fim mostra a coluna ?tp_pregnancy? que é o tipo da gravidez.
6.1.1.3 Variávies de histórico de gravidez
No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis do histórico de gravidez. Nesta tabela temos as colunas ?num_prenatal_appointments? que mostra o número de consultas de pré-natal, a média dessa coluna é 7.8, a mediana é 8, a variação dos dados é de no mínimo 0 e no máximo 40. A tabela também contém as colunas: ?tp_labor? que mostra o tipo de parto, e também contém a coluna ?tp_robson_group? que mostra a classificação do grupo Robson.
6.1.1.4 Variávies relacionadas ao recém-nascido
No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis relacionadas ao recém-nascido. Nesta tabela por exemplo podemos observar a coluna ?newborn_weight? que armazena o peso ao nascer dos recém-nascidos em gramas, a média dos dados dessa coluna é 3187.3, a mediana é 3215, a variação dos dados é de no mínimo 0 e no máximo 6000 gramas. Também podemos observar a coluna ?gestacional_week? que mostra o número de semanas de gestação, a média é 38.4, a mediana é 39, a variação dos dados é de no mínimo 15 e no máximo 45 semanas. Além dessas colunas a tabela também
32
contém as colunas: ?cd_apgar1? que mostra a pontuação de Apgar de 1 minuto, ?cd_apgar5? que mostra a pontuação de Apgar de 5 minuto, ?has_congenital_malformation? que mostra se o recém-nascido contém a presença de alguma malformação, ?tp_newborn_presentation? que mostra o tipo de apresentação de recém-nascido, ?tp_labor? que mostra o tipo de parto,

?was_cesarean_before_labor? que mostra se o recém-nascido foi cesáreo antes do parto, ?was_labor_induced ? que mostra se o recém-nascido foi induzido ao trabalho de parto, ?tp_childbirth_care? que mostra se ouve assistência ao parto, e por fim temos a coluna ?is_neonatal_death? que mostra se o recém-nascido veio a óbito ou não.

6.1.2 Análise das variáveis

Nesta seção serão mostrados a parte de análise de variáveis com diferentes tipos de gráficos.

6.1.2.1 Variáveis contínuas

Na Figura 5 temos dois gráficos boxplot, no boxplot à esquerda temos a distribuição da variável peso ao nascer e nele é possível observar duas caixas onde a caixa azul são os vivos e o laranja são os mortos. Nas duas caixas mostradas no gráfico podemos observar pequenos círculos nas pontas, esses círculos são outliers (dados fora da curva), então para a caixa de vivos os outliers são recém-nascidos com pesos acima de 4500 gramas ou abaixo de 2000 gramas, e para a caixa de mortos os outliers são os recém-nascidos com mais de 5500 gramas. Para os vivos temos a mediana de aproximadamente 3200 gramas e para os mortos a mediana é 1300 gramas. Cada caixa representa 50% da base de dados, a caixa azul de vivos mostra que o peso de recém-nascidos que sobreviveram está aproximadamente entre 2700 a 3600 gramas, e para a caixa laranja de mortos o peso de recém-nascidos que vieram a óbito estão aproximadamente entre 700 a 2500 gramas. Então o gráfico mostra para nós que os recém-nascidos que têm maior risco de vir a óbito tem pesos menores que 2000 gramas.

33

Figura 5. BoxPlot das features Peso ao nascer e Idade da mãe. (Fonte: Elaboração Própria).

No gráfico à direita da Figura 5 temos um boxplot da distribuição da variável idade da mãe, como foi dito anteriormente a caixa azul são os vivos e o laranja são os mortos. Nesse boxplot então podemos ver que os outliers da caixa de vivos são mães com idade acima de 46 anos aproximadamente, para os mortos os outliers são mães com idade acima de 50 anos. Para os vivos temos a mediana de aproximadamente 26 anos e para os mortos a mediana é de 26 anos. Nos vivos a idade das mães está aproximadamente entre 21 a 32 anos, e para os mortos a idade das mães está aproximadamente entre 20 a 34 anos. É importante ressaltar que o gráfico mostra as duas caixas bem parecidas, então de acordo com os dados não contém diferença significativa entre vivos e mortos usando a idade da mãe para distribuir.

Já na Figura 6 temos um boxplot da distribuição da variável semanas de gestação.

Nesse boxplot então podemos ver que os outliers da caixa de vivos são gestações com mais de 44 semanas aproximadamente, e gestações com menos de 35 semanas. Para os vivos temos a mediana de aproximadamente 39 semanas e para os mortos a mediana é de 32 semanas. Nos vivos as semanas de gestação estão aproximadamente entre 36 a 40 semanas, e para os mortos a semanas de gestação estão aproximadamente entre 26 a 36 semanas. Nesse boxplot os dados mostraram que para as gestações que têm menos de 36 semanas os recém-nascidos têm maior risco de vir a óbito.

34

Figura 6. BoxPlot da feature semana de gestação. (Fonte: Elaboração Própria).

Na Figura 7 temos um histograma da distribuição de peso ao nascer por classe, podemos observar no gráfico que grande parte dos vivos (linha azul) estão distribuídos entre 2000 e 4000 gramas, a área entre esses valores tem muitos dados de vivos, já entre os valores

0 e 2000 gramas temos uma concentração dos dados de mortos.
Figura 7. Histograma de distribuição do peso entre as classes. (Fonte: Elaboração Própria)
6.1.2.2 Variáveis categoricas
Observando a Figura 8 podemos ver um gráfico de barras da distribuição da variável escolaridade da mãe, este gráfico mostra a contagem de registros por escolaridade da mãe, esse gráfico mostra para nós que a maior parte das mães estão na categoria 4 que representa de 8 a 11 anos de escolaridade.
35
Figura 8. Gráfico de barras da distribuição da variável Escolaridade da mãe. (Fonte: Elaboração Própria).
Observando a Figura 9 podemos ver um gráfico de barras da distribuição da variável número de vivos, este gráfico mostra a contagem de registros por número de vivos, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho.
Figura 9. Gráfico de barras da distribuição da variável Número de nascidos vivos. (Fonte: Elaboração Própria).
Observando a Figura 10 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 1 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 1 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem
36
apgar de 8 ou 9, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.
Figura 10. Gráfico de barras da distribuição da variável Pontuação Apgar de 1 minuto. (Fonte: Elaboração Própria).
Observando a Figura 11 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 5 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 5 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem apgar de 9 ou 10, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.
Figura 11. Gráfico de barras da distribuição da variável Pontuação Apgar de 5 minutos. (Fonte: Elaboração Própria).
37
Observando a Figura 12 podemos ver um gráfico de barras da distribuição da variável presença de malformação congênita, este gráfico mostra a contagem de registros por presença de malformação congênita, esse gráfico mostra para nós que a grande maioria dos dados estão na categoria 2 que mostra que o recém-nascido não possui malformação, esse grande volume para a categoria 2 é causada pelo desbalanceamento da base.
Figura 12. Gráfico de barras da distribuição da variável Presença de malformação congênita. (Fonte: Elaboração Própria).
Observando a Figura 13 podemos ver um gráfico de barras da distribuição da variável número de gestações, este gráfico mostra a contagem de registros por número de gestações, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho, da mesma forma mostrada na Figura 9.
38

Figura 13. Gráfico de barras da distribuição da variável Número de gestações. (Fonte: Elaboração Própria).

Observando a Figura 14 podemos ver um gráfico de barras da distribuição da variável assistência ao parto, este gráfico mostra a contagem de registros por assistência ao parto, esse gráfico mostra para nós que a grande maioria dos dados mostram que o parto teve assistência de uma Enfermeira ou obstetra ou essa informação foi ignorada na hora do resgistro.

Figura 14. Gráfico de barras da distribuição da variável Assistência ao parto. (Fonte: Elaboração Própria).

6.1.2.3 Analise bi-variada

A Figura 15 mostra para nós a importância da feature peso ao nascer com o gráfico mostrado é possível observar claramente a diferença de pesos entre vivos e mortos, recém-nascidos com peso acima de 2000 gramas sobreviveram, já os recém-nascidos com menos de 2000 gramas vieram a óbito.

39

Figura 15. Gráfico de densidade de peso entre as classes. (Fonte: Elaboração Própria).

Na Figura 16 podemos observar um gráfico de correlação entre algumas features de valores contínuos, algumas dessas features possuem correlações positivas, como pontuação apgar de 1 minuto e pontuação apgar de 5 minutos, essas colunas possuem uma correlação de 0.7, então nascidos que recebem um valor alto de apgar de 1 minuto tendem a receber um valor alto no apgar de 5 minutos. O gráfico também mostra outras features com correlações positivas como, número de partos normais e número de gestações anteriores que contém uma correlação de 0.81, número de gestações anteriores com idade da mãe que tem uma correção de 0.39, número de partos de cesáreos com idade da mãe que possui correlação de 0.24, número de partos normais com idade da mãe com a correlação de 0.26, semanas de gestação com peso ao nascer em gramas com correção de 0.52, e as apgar de 1 e 5 minutos com semanas de gestação. E grande parte das features, possuem correlações baixas então elas são inversamente correlacionadas pelo que o gráfico mostra, como por exemplo número de partos normais com número de consultas pré-natais possuem uma correção de -0.17, e número de partos cesáreos com número de partos normais que contém uma correlação de -0.15. Então a correlação dessas features estão bem baixas tirando as correlações acima de 0.7.

40

Figura 16. Gráfico de correlação. (Fonte: Elaboração Própria)

6.1.2.4 Distribuição por raça

Na Figura 17 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por raça, e podemos observar que a raça 5 (indigena), é o boxplot que contem a menor variação de idade, já as raças 1 (branco) e 3 (amarelo) são as caixas que possuem maior variação com a média em torno de 20 e 30 anos.

41

Figura 17. Distribuição da idade da mãe por raça. (Fonte: Elaboração Própria).

No boxplot mostrado na Figura 18 podemos observar a distribuição de peso ao nascer por raça, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre as raças para peso ao nascer.

Figura 18. Distribuição de peso ao nascer por raça. (Fonte: Elaboração Própria).

Na Figura 19 temos o boxplot da a distribuição de semanas de gestação por raça, e

praticamente todas as caixas são iguais, somente a caixa da raça 1 (branco) possui menor variação que as outras.

42

Figura 19. Distribuição de semanas de gestação por raça. (Fonte: Elaboração Própria).

6.1.2.5 Distribuição por estado civil

A Figura 20 mostra o boxplot da a distribuição de peso ao nascer por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre os estados civis para peso ao nascer.

Figura 20. Distribuição de peso ao nascer por estado civil. (Fonte: Elaboração Própria).

A Figura 21 mostra o boxplot da a distribuição de semanas de gestação por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Porém os estados civis 2 (Casado) e 4 (Separados/divorciados judicialmente), contém variações menores.

43

Figura 21. Distribuição de semanas de gestação por estado civil. (Fonte: Elaboração Própria).

Na Figura 22 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por estado civil, e podemos observar que o estado civil 3 (Viúva), possui a maior variação, já o estado civil 4 (Separados/divorciados judicialmente) é a caixa que possuem menor variação

Figura 22. Distribuição da idade da mãe por estado civil. (Fonte: Elaboração Própria).

6.1.2.6 Distribuição por escolaridade da mãe

Na Figura 23 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por escolaridade da mãe, e podemos observar que a escolaridade 2 (1 a 3 anos) e 3 (4 a 7

44

anos), possui as maiores variações, já escolaridade 5 (12 ou mais anos) é a caixa que possuem menor variação

Figura 23. Distribuição da idade da mãe por escolaridade da mãe. (Fonte: Elaboração Própria).

A Figura 24 mostra o boxplot da a distribuição de peso ao nascer por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então, pouca variação é observada entre a escolaridade da mãe para peso ao nascer.

Figura 24. Distribuição de peso ao nascer por escolaridade da mãe. (Fonte: Elaboração Própria).

A Figura 25 mostra o boxplot da a distribuição de semanas de gestação por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas,

45

todas são praticamente iguais. Porém a escolaridade 5 (12 ou mais anos) contém variação menor que as outras.

Figura 25. Distribuição de semanas de gestação por escolaridade da mãe. (Fonte: Elaboração Própria).

6.2 ÁRVORE DE DECISÃO

A Figura 26 mostra um trecho do algoritmo utilizado, nesse trecho é realizado a

divisão do DataFrame das features de entrada , e o DataFrame que contém o rótulo.Neste caso o rótulo será o DataFrame ?target?, esse Dataframe contém a feature ?is_neonatal_death?, essa feature informa se o recém nascido veio a óbito ou não, sendo o foco da predição que desejamos realizar essa feature entra como rótulo para o modelo, já no Dataframe ?nome_features? temos as features que vão servir de entrada para o modelo, é a partir dessas features que o modelo tentará encontrar padrões para predizer o risco de óbito do recém nascido. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

46

Figura 26. Separação dos DataFrames de entrada e rótulo. (Fonte: Elaboração Própria).

6.2.1 Modelo de Árvore de Decisão Utilizando Particionamento 90/10

O algoritmo de árvore de decisão foi treinado utilizando duas formas diferentes: usando as partições 90% para treino e 10% para os testes e utilizando o método K-folds cross-validation. E para os experimentos iniciais foi utilizado o particionamento 90/10.

Tabela 4. Resultados do modelo de árvore de decisão. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	671807
Morto(1)	0.71	0.29	0.41	4216

Analisando a Tabela 4 pode-se ver os seguintes resultados, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 71% que realmente pertenciam a classe 1, então cerca de 29% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 29% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem a classe 1 que seria dos recém nascidos que poderiam vir a óbito, já para a classe de vivos a

47

classe 0 o modelo está identificando muito bem e classificando de forma correta. A tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.

Esse modelo teve uma acurácia de 0.99 que é uma acurácia bem alta, porém somente pela acurácia não é possível dizer que o modelo está excelente pelo motivo da base que está sendo usada é altamente desbalanceada, então para a classe 0 que é a de vivos temos muito mais dados que para a classe de mortos, e o modelo está acertando bastante a classe de vivos logo a acurácia fica alta por esses acertos, enquanto para a classe de mortos o modelo não está acertando tanto.

A Figura 27 mostra a curva ROC do modelo, pode-se observar que a AUC da curva ROC ficou em 0.92, a AUC mostra que o modelo está acertando bastante as predições, pois quanto mais perto de 1 melhor está no nosso modelo, mas no caso desse modelo vimos acima que as predições para a classe de mortos(1) está ruim, o modelo está acertando poucas classificações como 1. Então a AUC assim como a acurácia está alta porque a base de dados utilizada é altamente desbalanceada tendo muito mais amostras de vivos(1), e como o modelo

está bom para essa classificação e está acertando bastante os valores de AUC e acurácia ficam altos .

Figura 27. Curva ROC modelo de Árvore de decisão. (Fonte: Elaboração Própria).

A Figura 28 mostra a importância de cada feature para o modelo sendo assim a feature 10 - Peso ao nascer, a mais importante para o modelo, pois dela o modelo conseguiu tirar mais informações, seguido das features, 13 - Apgar dos 5 primeiros minutos, 11 - Semanas de 48
Gestação, 14 - Presença de malformação 12 - Apgar do 1 primeiro minutos. Então essas são as features que foram mais decisivas para a montagem da árvore e para as predições do modelo.

Figura 28. Gráfico de importância das features. (Fonte: Elaboração Própria).

Na Figura 29 temos uma imagem reduzida da árvore de decisão que foi montada com o treinamento utilizando 5 como profundidade máxima para a árvore, a imagem da árvore completa pode ser visualizada no apêndice A. É possível ver que as features marcadas como importante para o modelo apareceu diversas vezes na árvore, e a feature peso ao nascer foi escolhida para ser o nó raiz da árvore, de todas as features ela foi a que teve a melhor entropia para ser o nó raiz, e depois aparece mais vezes para outras decisões em outros níveis da árvore e isso faz com que essa feature se torne a mais importante para o modelo, também podemos ver que a feature ?Apgar do 1 primeiro minuto? mesmo sendo a menos importante para o modelo ela aparece somente no nó de número 42 isso mostra que a entropia e o ganho de informação dessa feature nas outras decisões não foram boas fazendo ela ser a feature com menos importância utilizada no modelo.

49

Figura 29. Árvore de decisão montada a partir do algoritmo. (Fonte: Elaboração Própria).

6.2.2 Modelo de Árvore de Decisão Utilizando K-folds cross-validation

Após realizar esse experimento com a partição 90/10, foi realizado outro experimento com esse mesmo algoritmo porém agora utilizando o método K-folds cross-validation para o treinamento. O K-folds cross-validation foi configurado para separar a base de dados em 10 partes iguais, e assim o algoritmo foi treinado novamente gerando um novo modelo. O método de K-folds cross-validation é utilizado para conferir se o modelo não está sofrendo problemas de overfitting, é importante garantir que o modelo está aprendendo com os dados e não está somente decorando.

Na Tabela 5 temos então o resultados desse modelo, pode-se observar que os resultados para a classificação de vivos(0) não teve alteração, o modelo continua tendo 100% na precisão e no recall, mas na classificação de mortos(1) o modelo teve uma diminuição na precisão que agora está em 65% e também teve uma diminuição no recall que agora está em 23%, porém essa diminuição é justificada porque para o teste deste modelo foi utilizado a base completa e não somente a partição de teste que contém 10% da base total. Logo os resultados não tiveram uma alteração drástica e tiveram um comportamento bem semelhante, incluindo a acurácia que se manteve em 0.99 e a AUC que teve uma diminuição pequena também e ficou com 0.89.

50

Tabela 5. Resultados do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)

Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 6719191
Morto(1) 0.65 0.23 0.34 41031
Na Tabela 6 é possível ver a matriz de confusão do modelo que mostra o número exato de erros e acertos do modelo, pode-se observar que para a classe de vivos o modelo classificou corretamente 6.714.117 amostras da base de dados completa, o modelo classificou como 0 as amostras que nos rótulos realmente eram 0, errou apenas 5074 amostras. Já para a classe de mortos o modelo acertou somente 9552 amostras, classificou como 1 as amostras que no rótulo eram 1 também, e errou 31479. Isso mostra que o modelo está muito desbalanceado, pois está errado muito mais do que acertando as predições de mortos, sendo isso um reflexo da base de dados desbalanceada também. E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.
Tabela 6. Matriz de confusão do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6714117 5074
Mortos (1) 31479 9552
6.3 REGRESSÃO LOGÍSTICA
Para o algoritmo de Regressão Logística também foi utilizado métodos diferentes para o treinamento, O algoritmo foi treinado utilizando três formas diferentes: usando as partições 90% para treino e 10% para os testes, utilizando o método K-folds cross-validation e foi treinada utilizando o método RFECV que seleciona as melhores features para o modelo, juntamente com o método GridSearchCV. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.
51
6.3.1 Modelo de Regressão Logística Utilizando Particionamento 90/10
Para o primeiro experimento do algoritmo de regressão logística foi utilizado o particionamento 90/10 para o treinamento do algoritmo, sendo 90% da base de dados para realizar o treinamento do modelo e 10% da base para realizar os testes.
Na Tabela 7 temos os resultados do modelo de regressão logística utilizando o método de particionamento 90/10, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 65% que realmente pertenciam a classe 1, então cerca de 35% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 24% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem as amostras da classe 1 que seria dos recém-nascidos que poderiam vir a óbito, já para a classe de vivos a classe 0 o modelo está identificando muito bem e classificando de forma correta.E na tabela também mostra os

valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.

Tabela 7. Resultados do modelo de regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	671885
Morto(1)	0.65	0.24	0.35	4138

A Figura 30 abaixo mostra a curva ROC do modelo de regressão logística usando o particionamento 90/10, analisando a curva ROC podemos ver que a AUC da curva ficou em 0.93 e a acurácia do modelo ficou 0.99, mostrando que o modelo está acertando bastante as predições, porém esses números para o nosso modelo não mostra que ele está ótimo, com a análise da Tabela 7 acima podemos observar que o modelo está acertando muitas predições da classe de vivos e poucas predições da classe de mortos, como a base de dados é
52
desbalanceada, como a maior quantidade de dados está na classe de vivos e o modelo está acerto quase todos dessa classe a acurácia e a AUC ficam altas por esses acertos.Logo é preciso analisar mais resultados além da acurácia e da AUC do modelo.

Figura 30. Curva ROC modelo regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria).

Na Tabela 8 podemos analisar os número exatos que o modelo acertou de cada classe, como é mostrado na matriz de confusão o modelo acertou 671.435 da classe de vivos e acertou apenas 915 da classe de mortos, e errou mais que o triplo da classe de mortos. Então as predições do modelo estão muito desbalanceadas, para a classe de vivos o modelo está predizendo bem, porém para as classes de mortos o modelo está errando muitas predições.

Tabela 8. Matriz de confusão do modelo de regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria)

	Predito	
	Vivos (0)	Mortos (1)
Classe Real		
Vivos (0)	671435	504
Mortos (1)	3169	915

6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation

Para o segundo experimento do algoritmo de regressão logística foi utilizado o método K-folds cross-validation para o treinamento do modelo com a intenção de conferir e confirmar
53
que o modelo não esteja tendo problemas com overfitting e realmente esteja aprendendo com os dados que está sendo usado para o treinamento. O método K-folds cross-validation foi configurado para realizar uma separação de 10 partes iguais.

Na Tabela 9 podemos observar os resultados do modelo, e os resultados foram parecidos com o experimento anterior usando o particionamento 90/10, única diferença nos resultado é que no experimento anterior a precisão da classe de mortos ficou em 65% e no caso desse experimento usando o K-folds cross-validation a precisão ficou em 64%, mas os valores de recall e F1-Score se mantiveram, assim como todos os resultados da classe de mortos se mantiveram em 100%. A execução desse experimento mostra que o modelo

realmente está aprendendo com os dados e não está apenas decorando, retirando o risco do modelo estar com overfitting. Mesmo sem alterações no resultado é importante saber que o modelo não está com overfitting e está conseguindo aprender e encontrar padrões nos dados.

Tabela 9. Resultados do modelo de regressão logística usando K-folds cross-validation. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	6719191
Morto(1)	0.64	0.24	0.35	41031

Já na Figura 31 temos a curva ROC do modelo, essa curva ROC mostra a AUC de todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas para todas as vezes que foi realizado os testes, a AUC se manteve em 0.93 e 0.94, sendo a média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua acertando muito a classe de vivos que é a classe predominante na base de dados. Já era esperado os resultados não sofrerem alterações grandes, pois a base de dados não sofreu nenhuma alteração, esse experimento foi somente para conferir se a modelo não estava sofrendo overfitting, e pelos resultados aparentemente a base não está com problemas de sobreajuste.

54

Figura 31. Curva ROC modelo regressão logística usando K-folds cross-validation. (Fonte: Elaboração Própria).

Na Tabela 10 temos a matriz de confusão do modelo, onde mostra que o modelo acertou 6.713.658 amostras da classe de vivos e errou apenas 5.533, já para a classe de mortos o modelo acertou somente 9.847 e novamente errou mais que o triplo errando 31.182 amostras. Como os resultados mostrados anteriormente não tiveram alteração era de se esperar que as predições corretas e incorretas do modelo também não sofressem alterações, o modelo continua acertando muito a classe de vivos e errando muito a classe de mortos, mantendo as predições bem desbalanceadas.

55

Tabela 10. Matriz de confusão do modelo de regressão logística usando K-folds cross-validation. (Fonte: Elaboração Própria)

	Predito	
	Vivos (0)	Mortos (1)
Classe Real		
Vivos (0)	6713658	5533
Mortos (1)	31182	9847

6.3.3 Modelo de Regressão Logística Utilizando GridSearchCV e RFECV

Para o terceiro e último experimento de regressão logística foi utilizado técnicas diferentes juntas para tentar melhorar as predições do modelo, para esse experimento usamos o K-folds cross-validation para o particionamento da base de dados assim como foi utilizado no experimento acima, e também foi utilizado duas técnicas que ainda não foram usadas nos experimentos anteriores que são as funções RFECV e GridSearchCV. A função RFECV seleciona a quantidade ideal de features para ser usadas no treinamento do modelo e também

mostra quais são as melhores features para essa predição, então essa função ela utiliza a validação cruzada para ir treinando e testando N vezes o modelo, e sempre vai alterando a quantidade e as features utilizadas para o teste, então cada vez que a função testa os modelos ela grava a pontuação dos acertos e assim depois mostra o gráfico da Figura 32 e assim é possível ver quais são as features ideais para o modelo.
56
Figura 32. Resultado da função RFECV Base Brasil. (Fonte: Elaboração Própria).
Então os resultados mostrados na Figura 32 apresenta que o número ideal para treinar o modelo de features é 6, e as features sao: 'tp_maternal_schooling', 'gestaional_week', 'cd_apgar1', 'cd_apgar5', 'has_congenital_malformation', 'tp_labor'. Essas são as features que tiveram o melhor desempenho no RFECV, então para o treinamento do modelo desse experimento as features que vão ser utilizadas serão somente essas 6.
Após a execução da função de RFECV foi executado a função de GridSearchCV, o GridSearchCV é utilizado para descobrir quais são os melhores parâmetros para utilizar no algoritmo de regressão logística e criar os modelos, foi utilizado também a validação cruzada para testar qual seria os melhores parâmetros para melhorar o desempenho do modelo. Na Figura 33 pode-se ver que os parâmetros escolhidos pela função foram:
Figura 33. Resultado da função GridSearchCV. (Fonte: Elaboração Própria).
Na Tabela 11 podemos observar os resultados do modelo, e os resultados foram parecidos com os experimentos anteriores, para a classe de mortos a precisão teve um aumento e foi para 69% e o recall diminuiu chegando em 20%, então utilizando as novas
57
funções o modelo ganhou precisão e perdeu recall. Para a classe de vivos o modelo se manteve em 100% e não teve alterações para os resultados de precisão e recall. Mesmo com a utilização das funções de RFECV e GridSearchCV o modelo não teve uma melhora significativa para as predições de mortos.
Tabela 11. Resultados do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	6719191
Morto(1)	0.69	0.20	0.31	41031

Na Figura 34 temos a curva ROC do modelo, essa curva ROC mostra a AUC de todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas para todas as vezes que foi realizado os testes, a AUC se manteve em 0.92 e 0.93, sendo a média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua acertando muito a classe de vivos que é a classe predominante na base de dados.
58
Figura 34. Curva ROC modelo regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria).
Na Tabela 12 temos a matriz de confusão do modelo, onde mostra que o modelo acertou 6.715.535 amostras da classe de vivos e errou apenas 3.656, comparando com os experimentos anteriores de regressão logística as predições de vivos teve uma pequena piora e

teve uma pequena diminuição nos acertos. Já para a classe de mortos, o modelo acertou 8.296 e errou 32.735 amostras, logo as predições de vivos tiveram uma uma piora e acertou um pouco menos do que os outros experimentos, e na classe de vivos teve uma melhora acertando mais predições, porém como o problema está nas predições de mortos e para essa classe teve uma piora pode-se dizer que a função de GridSearchCV não teve um ganho significativo nos resultados. Então como modelo final foi escolhido o modelo utilizando somente o K-folds cross-validation, pois foi o modelo que mostrou o melhor desempenho dos experimentos de regressão logística.

59

Tabela 12. Matriz de confusão do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

Predito

Vivos (0) Mortos (1)

Classe Real

Vivos (0) 6715535 3656

Mortos (1) 32735 8296

60

7 CONCLUSÃO

O principal objetivo deste trabalho foi analisar os resultados das predições de mortalidade neonatal dos algoritmos de aprendizado de máquina usando uma base de dados com dados do Brasil inteiro. A partir destes resultados apresentados na secção anterior ?Resultados?, é possível concluir que analisar somente a AUC e a acurácia do modelo não é o suficiente para garantir que o modelo está excelente, sendo assim temos que ter mais atenção a precisão, recall e as predições mostradas na matriz de confusão.

Ambos os modelos ficaram desbalanceados nas predições, os modelos acertaram mais predições de vivos do que mortos, e as predições corretas de mortos foram mais baixas do que as incorretas, para os modelos ficarem melhor precisam passar por algoritmos que possam balancear essas predições, uma outra alternativa para melhorar o modelo é utilizar uma base de dados mais balanceada, pois essa base que foi utilizada é altamente desbalanceada.

Embora os dois modelos tenham tido resultados parecidos , o modelo de regressão logística utilizando somente o K-folds cross-validation se saiu melhor que os outros experimentos, o modelo criado na seção ?6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation? manteve a acurácia, AUC, precisão e recall dos demais modelos e acertou mais predições, talvez com algoritmos para melhorar as predições da classe de mortos, balanceando mais as predições, esse modelo fique com ótimas taxas preditivas.

Além disso podes-se afirmar que o peso ao nascer do recém nascido é um fator muito importante para as decisões dos modelos, ambos os modelos usaram muito essa informação para as predições, também podemos colocar, as colunas de presença de malformação e semana de gestação, como colunas que foram importantes para os modelos, essas causas podem ser evitadas se houver um acompanhamento médico com qualidade e eficiência.

61

REFERÊNCIAS

ÁRVORE de Decisão - Aula 3. Gravação de Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=ecYpXd4WREk&t=4089s. Acesso em: 1 jul. 2020.

BARRETO, J. O. M.; SOUZA, N. M.; CHAPMAN, E. Síntese de evidências para políticas de saúde: reduzindo a mortalidade perinatal. Ministério da Saúde, p. 1?44,2015.
BATISTA, André Filipe de Moraes; FILHO, Alexandre Dias Porto Chiavegatto. Machine Learning aplicado à Saúde. In: ZIVIANI, Artur; FERNANDES, Natalia Castro; SAADE, Débora Christina Muchaluat. SBCAS 2019: 19 Simpósio Brasileiro de Computação Aplicada à Saúde. [S. l.: s. n.], 2019. cap. Capítulo 1.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. Towards neonatal mortality risk classification: A data-driven approach using neonatal, maternal, and social factors, [s. l.], 23 out. 2020.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. SPNeoDeath: A demographic and epidemiological dataset having infant, mother, prenatal care and childbirth data related to births and neonatal deaths in São Paulo city Brazil ? 2012?2018, [s. l.], 19 jul. 2020.
BRESAN, Rodrigo. Um método para a detecção de ataques de apresentação em sistemas de reconhecimento facial através de propriedades intrínsecas e Deep Learning. Orientador: Me. Carlos Eduardo Beluzo. 2018. Trabalho de Conclusão de Curso (Superior de Tecnologia em Análise e Desenvolvimento de Sistemas) - Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas., [S. l.], 2018.
CABRAL, Cleidy Isolete Silva. Aplicação do Modelo de Regressão Logística num Estudo de Mercado. Orientador: João José Ferreira Gomes. 2013. Projeto Mestrado em Matemática Aplicada à Economia e à Gestão (Mestrado) - Universidade de Lisboa Faculdade de Ciências Departamento de Estatística e Investigação Operacional, [S. l.], 2013. Disponível em: https://repositorio.ul.pt/bitstream/10451/10671/1/ulfc106455_tm_Cleidy_Cabral.pdf. Acesso em: 1 maio 2021.
CAMPOS, Raphael. Árvores de Decisão: Então diga-me, como construo uma?. [S. l.], 28 nov. 2017. Disponível em: https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69. Acesso em: 23 jan. 2021.
Colab. 2021. Disponível em: https://colab.research.google.com/. Acesso em: 20 mar. 2021.
COX, D.R.;SNELL, E.J. Analysis of Binary Data.London: Chapman & Hall, 2ª Edição, 1989. HOSMER, D.W.; LEMESHOW, S.Applied Logistic Regression. New York: John Wiley, 2ª Edição, 2000.
E.França, S.Lansky. Mortalidade infantil neonatal no brasil: situação, tendências e perspectivas Rede Interagencial de Informações para Saúde - demografia e saúde: contribuição para análise de situação e tendências Série Informe de Situação e Tendências(2009), pp.83-112.
62
FREIRE, Sergio Miranda. Bioestatística Básica: Regressão Linear. [S. l.], 20 out. 2020. Disponível em: http://www.lampada.uerj.br/arquivosdb/_book/bioestatisticaBasica.html. Acesso em: 23 jan. 2021.
Jupyter. 2021. Disponível em: https://jupyter.org/. Acesso em: 20 jun. 2021.
GOODFELLOW, I. et al. Deep learning. [S.l.]: MIT press Cambridge, 2016.

Pandas. 2021. Disponível em: https://pandas.pydata.org/. Acesso em: 20 jun. 2021.
Python. 2021. Disponível em: https://www.python.org/. Acesso em: 20 jun. 2021.
KUHN, M.; JOHNSON, K. Applied predictive modeling. [S.l.]: Springer, 2013. v. 26.
LANSKY, S. et al. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante . Cadernos de Saúde Pública, Scielo, v. 30, p. S192 ? S207, 00 2014.
LANSKY, Sônia; FRICHE, Amélia Augusta de Lima; SILVA, Antônio Augusto Moura; CAMPOS, Deise; BITTENCOURT, Sonia Duarte de Azevedo; CARVALHO, Márcia Lazaro; FRIAS, Paulo Germano; CAVALCANTE, Rejane Silva; CUNHA, Antonio José Ledo Alves. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante e ao recém-nascido. [s. l.], Ago 2014. Disponivel em: https://www.scielosp.org/article/csp/2014.v30suppl1/S192-S207/pt/
Matplotlib. 2021. Disponível em: https://matplotlib.org/. Acesso em: 20 mar. 2021.
Maranhão AGK, Vasconcelos AMN, Trindade CM, Victora CG, Rabello Neto DL, Porto D,et al. Mortalidade infantil no Brasil: tendências, componentes e causas de morte no período de 2000 a 2010 . In: Departamento de Análise de Situação de Saúde, Secretaria de Vigilância em Saúde, Ministério da Saúde, organizador. Saúde Brasil 2011: uma análise da situação de saúde e a vigilância da saúde da mulher. v. 1. Brasília: Ministério da Saúde; 2012. p. 163-82.
MESQUITA, PAULO. UM MODELO DE REGRESSÃO LOGÍSTICA PARA AVALIAÇÃO DOS PROGRAMAS DE PÕS-GRADUAÇÃO NO BRASIL. Orientador: Prof. RODRIGO TAVARES NOGUEIRA. 2014. Dissertação (Mestre em Engenharia de Produção) - Universidade Estadual do Norte Fluminense, [S. l.], 2014.
MNASSRI , Baligh. Titanic: logistic regression with python. [S. l.], 1 ago. 2020. Disponível em: https://www.kaggle.com/mnassrib/titanic-logistic-regression-with-python. Acesso em: 14 jan. 2021.
Morais, R.M.d., Costa, A.L: Uma avaliação do sistema de informações sobre mortalidade. Saúde em Debate 41, 101?117 (2017). Disponível em: https://doi.org/10.1109/SIBGRAPI.2012.38.
MOTA, Caio Augusto de Souza; BELUZO , Carlos Eduardo; TRABUCO, Lavinia Pedrosa; SOUZA , Adriano; ALVES, Luciana; CARVALHO, Tiago. DESENVOLVIMENTO DE UMA PLATAFORMA WEB PARA CIÊNCIA DE DADOS APLICADA À SAÚDE MATERNO INFANTIL , [s. l.], 28 nov. 2019.
MULTIEDRO (Brasil). Conheça as aplicações do machine learning na área da saúde. [S. l.], 28 nov. 2019. Disponível em: https://blog.multiedro.com.br/machine-learning-na-area-da-saude/#:~:text=O%20machine%263
0learning%20na%20%C3%A1rea,diagn%C3%B3sticos%20e%20tratamentos%20mais%20precisos. Acesso em: 13 jun. 2021.
MUKHIYA,S.K.;AHMED,U. Hands-On Exploratory Data Analysis with Python: Perform EDA Techniques to Understand, Summarize, and Investigate Your Data. [S.l.]: PacktPublishingLtd,2020.ISBN978-1-78953-725-3.22,32
Murray CJ, Laakso T, Shibuya K, Hill K, Lopez AD. Can we achieve Millennium Development Goal 4? New analysis of country trends and forecasts of under-5 mortality to 2015. Lancet 2007; 370:1040-54.

NumPy. 2021. Disponível em:https://numpy.org/. Acesso em: 20 jun. 2021.
Oliveira, M.M.d., de Araújo, A.S.S.C., Santiago, D.G., Oliveira, J.a.C.G.d., Carvalho, M.D., de Lyra et al, R.N.D.: Evaluation of the national information system on live births in brazil , 2006-2010. Epidemiol. Serv. Saúde 24(4), 629?640 (2015).
PELISSARI, ANA CAROLINA CHEBEL. MORTALIDADE NEONATAL DA CIDADE DE SÃO PAULO: UMA ABORDAGEM UTILIZANDO APRENDIZADO DE MÁQUINA SUPERVISIONADO. 2021. Trabalho de Conclusão de Curso (Superior) - Instituto Federal de Educação, Ciência e Tecnologia Campus Campinas, [S. l.], 2021.
REGRESSÃO logística e binary cross entropy - Aula 7. Direção: Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=3J-LBtHVsm4. Acesso em: 21 jan. 2021.
Rmd Nascimento, AJM Leite, NMGSd Almeida, Pcd Almeida, Cfd Silva Determinantes da mortalidade neonatal: estudo caso-controle em fortaleza, ceará, brasil Cad Saúde Pública, 28(2012), pp.559 - 572.
SANTOS, Hellen Geremias. Machine learning e análise preditiva: considerações metodológicas: métodos lineares para classificação. In: SANTOS, Hellen Geremias. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina. 2018. Tese (Pós-Graduação em Epidemiologia) - Faculdade de Saúde Pública da Universidade de São Paulo, [S. l.], 2018.
Scikit-learn. 2021. Disponível em: https://scikit-learn.org/stable/. Acesso em: 20 jun. 2021.
Seaborn. 2021. Disponível em: https://seaborn.pydata.org/. Acesso em: 20 jun. 2021.
SILVA, Wesley; CORSO, Jansen; WELGACZ, Hanna; PEIXE, Julinês. AVALIAÇÃO DA ESCOLHA DE UM FORNECEDOR SOB CONDIÇÃO DE RISCOS A PARTIR DO MÉTODO DE ÁRVORE DE DECISÃO. ARTIGO ? MÉTODOS QUANTITATIVOS, [s. l.], 12 ago. 2008.
WHO, W. H. O.Women and health: today?s evidence, tomorrow?s agenda. [S.l.]: UNWorld Health Organization, 2009. ISBN 9789241563857.
64
APÊNDICE A

==========

Arquivo 1: monografia-V7.docx.pdf (9728 termos)
Arquivo 2: https://medium.com/data-hackers/crossvalidation-de-maneira-did%C3%A1tica-79c9b080a6ec (761 termos)
Termos comuns: 24
Similaridade: 0,22%
O texto abaixo é o conteúdo do documento monografia-V7.docx.pdf (9728 termos)
Os termos em vermelho foram encontrados no documento https://medium.com/data-hackers/crossvalidation-de-maneira-did%C3%A1tica-79c9b080a6ec (761 termos)

==========

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO
CÂMPUS CAMPINAS
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
CAMPINAS
2021
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas.
Orientador: Prof. Me. Everton Josué da Silva.
CAMPINAS
2021
Dados Internacionais de Catalogação na Publicação
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo Câmpus Campinas.
Aprovado pela banca examinadora em: _____ de _____ de _____.
BANCA EXAMINADORA

_____

Prof. Me. Everton Josué da Silva (orientador)
IFSP Câmpus Campinas

_____

Prof. Me. Carlos Eduardo Beluzo
IFSP Câmpus Campinas

_____

Prof. Dr. Ricardo Barz Sovat
IFSP Câmpus Campinas

Dedico este trabalho aos meus familiares,
colegas de classe, professores e servidores do Instituto
que colaboraram em minha jornada formativa.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pelo dom da vida e pela oportunidade de concluir mais uma etapa de minha experiência acadêmica.

Agradeço a todos os professores e servidores do IFSP Câmpus Campinas, que contribuíram direta e indiretamente para a conclusão deste trabalho.

Agradeço também à minha família, que deu todo o apoio necessário para que eu chegasse até aqui.

Agradeço ao meu orientador que me auxiliou a solucionar as dificuldades encontradas no caminho.

"Minha energia é o desafio, minha motivação é o impossível,
e é por isso que eu preciso
ser, à força e a esmo, inabalável.".
Augusto Branco

RESUMO

A mortalidade infantil pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível em uma população. O aprendizado de máquina é uma área da inteligência artificial que propõe métodos de análise de dados que automatizam a construção de modelos analíticos com base em reconhecimento de padrões, baseado no conceito de que sistemas podem aprender com dados, identificando padrões e tomando decisões com o mínimo de intervenção humana. O objetivo deste trabalho é testar e analisar dois tipos diferentes de algoritmos de aprendizado de máquina, utilizando a base de dados do SIM e SINASC do Brasil, do período de 2016 até 2018, para gerar modelos para predição de mortalidade neonatal. Os métodos utilizados foram Árvore de Decisão e de Regressão Logística e como métricas para avaliar os modelos resultantes foram utilizadas a AUC, Curva ROC e a Matriz de Confusão. Como resultado, apesar de terem alcançado valores de AUC 0.93, ambos modelos de predições acertaram muitas predições de vivos cerca de 671.000 e erraram muitas predições de mortos cerca de 2.600, principalmente devido ao fato de a base estar desbalanceada.

Palavras-chave: Mortalidade Neonatal, Predição, Aprendizado de Máquina.

ABSTRACT

Infant mortality can be used to analyze poverty and socioeconomic levels, as well as measure the quality of health and medical technology available in a population. Machine learning is an

area of artificial intelligence that proposes data analysis methods that automate the construction of analytical models based on pattern recognition, based on the concept that systems can learn from data, identifying patterns and making decisions with a minimum of human intervention. The objective of this work is to test and analyze two different types of machine learning algorithms, using the SIM and SINASC do Brasil database, from 2016 to 2018, to generate models for predicting neonatal mortality. The methods used were Decision Tree and Logistic Regression and as metrics to evaluate the resulting models, AUC, ROC Curve and Confusion Matrix were used. As a result, despite achieving values of AUC 0.93, both prediction models correct many live birth predictions around 671.000 and miss many stillbirth predictions around 2600, mainly due to the fact that the base is unbalanced.
Keywords: Neonatal Mortality, Prediction, Machine Learning.
LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

SIM Sistema de Informação sobre Mortalidade
SINASC Sistema de Informações sobre Nascidos Vivos
TMI Taxa de Mortalidade Infantil
TMN Taxa de Mortalidade Neonatal
MN Mortalidade Neonatal
ML Machine Learning
DNV Declaração de Nascido Vivo
VN Verdadeiro Negativo
FN Falso Negativo
VP Verdadeiro Positivo
FP Falso Positivo
ROC Curva Característica de Operação do Receptor
NaN Não é um Número
RFE Eliminação Recursiva de Feature
RFECV Eliminação Recursiva de Feature com Validação Cruzada

SUMÁRIO

13
1 INTRODUÇÃO

A taxa de mortalidade infantil (TMI) é uma importante medida de saúde em uma população e é um grande problema no mundo todo. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível. Esta taxa é dividida em duas categorias: neonatal e pós-neonatal. É categorizada neonatal quando o óbito ocorre nos 28 primeiros dias de vida após o pós-parto, e o pós-neonatal é quando o óbito ocorre entre 29 e 364 dias de vida (BELUZO et al., 2020). Cerca de 46% das mortes no mundo acontecem entre os menores de cinco anos e grande parte está concentrada nos primeiros dias de vida (LANSKY, 2014). A diminuição dessas taxas é muito importante no mundo todo, refletindo nos indicadores de saúde pública e desenvolvimento do país.

Os fatores associados à mortalidade são profundamente influenciados pelas características biológicas maternas e neonatais, pelas condições sociais e pelos cuidados prestados pelos serviços de saúde (E. FRANÇA; S. LANSKY, 2009; R.M.D. NASCIMENTO,

2012). Em grande parte, o diagnóstico é altamente dependente da experiência adquirida pelo profissional que o realiza. Apesar de indiscutivelmente necessário, não é perfeito, principalmente pelo fato de ser dependente de fatores humanos, por este motivo os profissionais sempre tiveram recursos tecnológicos para auxiliar nessa tarefa (BELUZO et al., 2020). Segundo estudos de 2010 no Brasil, calcula-se que aproximadamente 70% dos óbitos infantis ocorridos poderiam ter sido evitados por ações de saúde e que 60% dos óbitos neonatais ocorreram por situações que poderiam ser evitadas (BARRETO; SOUZA; CHAPMAN, 2015).

No presente trabalho foram utilizados dois algoritmos de aprendizado de máquina para criar modelos de predição de mortalidade neonatal. Além disso, foi realizada também uma análise exploratória da base de dados. Para este trabalho foram utilizadas duas fontes de dados: Sistema de Informação sobre Mortalidade (SIM) e Sistema de Informações sobre Nascidos Vivos (SINASC) do Brasil inteiro.

14

2 JUSTIFICATIVA

Com o avanço da tecnologia, podemos notar que ela está cada vez mais presente no nosso dia a dia e nos auxilia em diversas tarefas do cotidiano, e vem sendo aplicada em diversas áreas, dentre elas a saúde.

A mortalidade infantil é uma preocupação mundial na saúde pública, a ONU (Organizações das Nações Unidas) definiu a redução da mortalidade infantil como uma meta para o desenvolvimento global. A mortalidade neonatal é responsável por aproximadamente 60% da TMI nos países em desenvolvimento (A.K. SINGHA, 2016). Existem diversos fatores que podem contribuir com a redução de óbitos neonatais, como por exemplo acompanhamento médico à gestante no período de gravidez, acompanhamento médico ao recém-nascido nos primeiros dias de vida e a disponibilização deste serviço com qualidade e proficiência (WHO, 2009).

A diminuição dessa taxa é importante e de interesse para o mundo todo, e utilizar da tecnologia para auxiliar nessa diminuição é uma proposta funcional, e inovadora para a realidade brasileira (MOTA et al., 2019). Havendo disponibilidade de tecnologia para auxiliar nas decisões dos diagnósticos, os profissionais da saúde podem focar nos cuidados do recém-nascido com risco de vir a óbito, fornecendo tratamentos melhores.

Segundo o site Multiedro (2019), um grande problema que os médicos enfrentam ao realizar o diagnóstico, é analisar um grande volume de dados. Para a área médica obter diagnósticos rápidos e precisos pode salvar vidas, e a utilização de machine learning para realizar esses processos é importante, com a ML os dados podem ser processados em questões de segundos e resultar em um diagnóstico mais rápido, além disso utilizando bons modelos ML os diagnósticos serão mais precisos.

Diversos trabalhos vêm sendo desenvolvidos neste contexto. No trabalho realizado por BELUZO et al. (2020a), os autores utilizaram uma base de dados com informações da cidade de São Paulo para criação de modelos de aprendizado de máquina para predição de mortalidade neonatal. Este recorte foi utilizado também no presente trabalho para fins de exercício e definição de etapas da implementação de código. Na conclusão os autores discutem os fatores que tiveram mais influência nas predições, e na análise feita pelos autores, as consultas de pré-natal são muito importantes para a redução da mortalidade infantil, uma

vez que algumas características muito importantes, como a malformação congênita, podem ser detectadas ao longo das consultas de pré-natal.

Em um outro estudo realizado por BELUZO et al. (2020a), foi realizada uma análise exploratória da base de dados SPNeoDeath, onde e podemos observar detalhadamente cada 15
análise das colunas da tabela e diversos gráficos feito para um melhor entendimento dos dados.

Outro trabalho que foi realizado utilizando a base de dados com dados somente de São Paulo foi o de PELISSARI (2021), nesse trabalho é realizado uma análise exploratória na base de dados de São Paulo e é também analisado três algoritmos de aprendizado de máquina (Logistic Regression, Random Forest Classifier e XGBoost), nesse trabalho a autora conclui que os modelos de Logistic Regression e XGBoost foram os modelos que apresentaram os melhores desempenhos preditivos, e também mostra os problemas de estar utilizando uma base de dados desbalanceada.

O problema da mortalidade neonatal não é recente, e o mundo todo desenvolve iniciativas para lidar com ele. A ONU definiu como meta a redução da mortalidade infantil para o desenvolvimento global. Diversos fatores podem ajudar na diminuição da taxa de mortalidade neonatal como acompanhamento médico antes e após parto, tanto com a mãe quanto com o recém-nascido. Um serviço de qualidade e constante para os casos com maior risco de óbito pode salvar diversos recém-nascidos. Neste sentido, o desenvolvimento de ferramentas para a predição de mortalidade neonatal pode colaborar com esta iniciativa.

3 OBJETIVOS

3.1 OBJETIVO GERAL

O presente trabalho tem como objetivo testar e analisar os resultados da aplicação dos aprendizagem de máquina supervisionado Árvore de Decisão e Regressão Logística, utilizando dados do SIM e do SINASC, no período de 2016 até 2018, a fim de gerar modelos de predição de risco morte neonatal.

3.2 OBJETIVOS ESPECÍFICOS

? Realizar análise exploratória da base dados a ser utilizada na construção dos modelos;
? Implementar modelos de predição utilizando algoritmos Árvore de Decisão e Regressão Logística;
? Avaliar resultados e propor novas abordagens.

16

4 FUNDAMENTAÇÃO TEÓRICA

4.1 MORTALIDADE INFANTIL E NEONATAL

Como dito na monografia PELISSARI (2021), a TMI consiste no número de crianças que foram a óbito antes de concluir um ano de vida dividido por 1000 crianças nascidas vivas no tempo de um ano. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos e qualidade da saúde pública de um país (apud BELUZO et al., 2020).

Mencionado em PELISSARI (2021), a TMN é uma das duas categorias da TMI, é categorizado mortalidade neonatal quando o óbito ocorre do nascimento da criança até os 28 primeiros dias de vida. A mortalidade neonatal pode ser subdividida em mortalidade neonatal precoce que é quando o óbito ocorre entre 0 e 6 dias de vida, e o neonatal tardio quando o óbito ocorre entre 7 e 28 dias de vida (apud E. FRANÇA e S. LANSKY, 2009).

Segundo Lansky et al. (2014), a taxa de mortalidade neonatal é o principal componente da TMI desde a década de 1990 no Brasil e vem se mantendo em níveis elevados, com taxa de 11,2 óbitos por mil nascidos vivos em 2010 (apud Maranhão et al., 2012). Também é dito em Lansky et al. (2014), que a TMI do Brasil em 2011 foi 15,3 por mil nascidos vivos, alcançando a meta 4 dos objetivos de desenvolvimento do milênio, compromisso dos governos integrantes das Nações Unidas de melhorar a saúde infantil e reduzir em 2/3 a mortalidade infantil entre 1990 e 2015. (apud Maranhão et al., 2012; apud Murray et al., 2015). Segundo Lansky et al. (2014) o principal componente da TMI em 2009 é a mortalidade neonatal precoce, e grande parte das mortes infantis acontece nas primeiras 24 horas (25%), indicando uma relação estreita com a atenção ao parto e nascimento (apud E. FRANÇA e S. LANSKY, 2009).

4.2 MÉTODOS DE APRENDIZADO DE MÁQUINA

A utilização de dados para a resolução de problemas, de modo a se identificar padrões ocultos e posteriormente auxiliar na tomada de decisões é definida como aprendizado de máquina (BRESAN, 2018 apud Brink et al. 2013).

O uso de técnicas de Aprendizado de Máquina se mostra altamente eficiente na resolução de tarefas que se apresentem difíceis de se resolver a partir de programas escritos por humanos (BRESAN, 2018 apud GOODFELLOW et al., 2016), bem como também possibilita uma maior compreensão sobre os funcionamentos dos princípios que compõem a inteligência humana.

17

Neste trabalho serão implementados modelos utilizando os algoritmos de Árvore de Decisão e Regressão Logística, os quais serão apresentados a seguir.

4.2.1 Árvore de Decisão

Segundo Batista e Filho (2019), os modelos preditivos baseados em árvores são geralmente utilizados para tarefas de classificação, embora também possam ser utilizados para tarefas de regressão. Considerado um dos mais populares algoritmos de predição, o algoritmo de árvore de decisão proporciona uma grande facilidade de interpretação.

As árvores de decisão utilizam a estratégia ?dividir-e-conquistar", na qual as árvores são construídas utilizando-se apenas alguns atributos. A árvores de decisão é uma das técnicas por meio da qual um problema complexo é decomposto em subproblemas mais simples. Recursivamente, a mesma estratégia é aplicada a cada subproblema (SILVA et al, 2008).

A árvore de decisões tem uma estrutura hierárquica onde cada nó da árvore representa uma decisão em uma das colunas do conjunto de dados, cada ramo da árvore representa uma tomada de decisão (BATISTA; FILHO, 2019).

O algoritmo segue algumas decisões para montar a árvore, o atributo mais importante é utilizado como nó raiz e será o primeiro nó, já os atributos menos importantes são mostrados nos ramos da árvore. Cada atributo é verificado para conferir se é possível gerar uma árvore menor e se é possível fazer uma classificação melhor, e assim é escolhido o nó que produz nós filhos (SILVA et al, 2008).

Existem diferentes tipos de algoritmo para a construção da árvore, como por exemplo, ID3 (Iterative Dichotomiser), C4.5 e CART (Classification and Regression Tree). O algoritmo ID3 escolhe os nós da árvore por meio da métrica do ganho de informação. Já o CART faz uso da equação de Gini (BATISTA; FILHO, 2019 apud KUHN; JOHNSON, 2013).

O algoritmo utilizado neste trabalho usará o cálculo da entropia e ganho de informação, para decidir qual atributo será usado no nó, a entropia é uma medida da desorganização dos sistemas, maior é a incerteza do sistema, quanto maior a entropia maior está a desorganização, com a árvore de decisão a ideia é ir organizando o sistema e ir separando as decisões da melhor forma para que a entropia diminui e o ganho de informação aumente, para que a decisão do modelo fiquei mais assertiva. O cálculo da entropia utiliza a seguinte equação (ÁRVORE, 2020):

???????? =
?
?? ?? ???
2
??
18

Nesta equação o i representa possíveis classificações (rótulos), e o P é a probabilidade de cada uma. A ideia da árvore então é verificar qual é a entropia para cada uma das variáveis da base de dados, e verificar qual é a melhor organização de cada uma das variáveis. E isso é realizado através da técnica chamada de ganho de informação, essa técnica utiliza a equação a seguir (ÁRVORE, 2020):

????? = ????????(???) ? ? ????(??????) * ????????(??????)

Então o ganho de informação é calculado pela entropia do nó pai menos a soma do peso vezes a entropia dos nós filhos. Esse cálculo é realizado para todas as variáveis e a variável que tiver maior ganho de informação é utilizado para a decisão do nó. Esse processo é realizado recursivamente e a árvore vai sendo montada com base nesses cálculos (ÁRVORE, 2020).

Na Figura 1 temos a representação do início de uma árvore de decisão, no caso em questão o autor estava classificando exames positivos e negativos para o vírus SARS-CoV-2, ele utilizou uma base de dados que contém quase 6 mil exames realizados, contendo campos como o resultado do exame, idade do paciente, níveis de hemoglobina entre outras informações.

Figura 1. Exemplo de Árvore de Decisão. (Fonte: ÁRVORE 2020).

Pode-se observar que para o nó raiz foi utilizado a variável Leukocytes, logo foi o que apresentou a melhor organização para ser o nó raiz, e após ele temos os ramos da árvore. Cada retângulo da árvore é uma decisão do modelo e esses retângulos têm atributos importante como, a coluna da base de dados que será onde vai ser realizado a decisão, então pegando o nó raiz, pacientes que tiverem com os Leukocytes para menos que -0.43 seguiram o caminho
19
para a esquerda (true), já os pacientes que tiverem mais seguiram o caminho da direita (false). A entropia calculada do nó também é apresentada e a classificação do nó também, então se o resultado parar neste nó, a classificação que está nele será a classificação final, e temos o samples que é o número de amostras que se enquadram no na decisão. E esse processo de verificação do modelo vai se repetindo até não ter mais como dividir em subproblemas ou até chegar na profundidade máxima.

Uma característica importante da árvore de decisão é que ela é um modelo WhiteBox, ou seja, é possível visualizar a árvore montada e as decisões que o modelo terá que tomar na

árvore. Na Figura 1 pode-se ver uma árvore de decisões montada, cada retângulo da árvore é uma decisão que o modelo terá que tomar, conforme o modelo toma as decisões ele vai seguindo um caminho até chegar ao nó folha, nó folha é o nó que não contém nó filho, não tem mais ramos para baixo, chegando ao nó folha o modelo tem a classificação final.

4.2.2 Regressão Logística

O modelo de regressão logística estabelece uma relação entre a probabilidade de ocorrência da variável de interesse e as variáveis de entrado do modelo, sendo utilizada para tarefas de classificação, em que a variável de interesse é categórica, neste caso, a variável de interesse assume valores 0 ou 1, onde 1 é geralmente utilizado para indicar a ocorrência do evento de interesse (Batista e Filho, 2019).

De acordo com Mesquita (2014), a técnica de regressão logística, desenvolvida no século XIX, obteve maior visibilidade após 1950 ficando então mais conhecida. Ficou ainda mais difundida a partir dos trabalhos de Cox & Snell (1989) e Hosmer & Lemeshow (2000). Caracteriza-se por descrever a relação entre uma variável dependente qualitativa binária, associada a um conjunto de variáveis independentes qualitativas ou métricas.

Esta técnica inicialmente foi utilizada na área médica, porém a eficiência viabilizou sua implementação nas mais diversas áreas. Mesquita (2014) diz que o termo regressão logístico, tem sua origem na transformação usada com a variável dependente, que permite calcular diretamente a probabilidade da ocorrência do fenômeno em estudo.

Segundo Santos (2018), para estimar a probabilidade, é utilizado uma técnica chamada distribuição binomial para modelar a variável resposta do conjunto de treinamento, que tem como parâmetro a probabilidade de ocorrência de uma classe específica. Uma característica importante desse modelo é que a probabilidade estimada deve estar limitada ao intervalo de 0 a 1. O algoritmo de Regressão Logística gera um modelo de classificação binária (0 e 1). O modelo utiliza a Regressão Linear como base, a equação linear é (REGRESSÃO, 2020):

20

? = ? . ? + ?

O ?m? da equação é o coeficiente angular da reta, e o ?b? é o intercepto. Então, aplicando a equação linear obtemos um valor contínuo como resultado, após obter esse valor a Regressão Logística tem que realizar a classificação do resultado, então é aplicado uma função chamada sigmoide (REGRESSÃO, 2020):

? = 1

1+ ? ??

Essa função chamada de sigmoide, ou função logística, é responsável por ?achatar? o resultado da regressão linear, calculando a probabilidade de o resultado pertencer a classe 1, classificando assim o resultado da função linear em 0 ou 1.

Segundo Batista e Filho (2019), o modelo de regressão logística que tem como objetivo realizar uma classificação binária de uma variável de interesse irá predizer a probabilidade de pertencer à classe positiva. Tem-se, portanto, que é dado por:?

? = { 0, ?? ???? (? = 1 ??) < 0, 5

ou

? = { 1, ?? ???? (? = 1??) ? 0, 5

Então a decisão do modelo de regressão logística está relacionada à escolha de um ponto de corte para p(x). Caso o ponto de corte escolhido for p(x)=0,5, os resultados que

forem p(x)>0,5 serão classificados como 1 e os resultados p(x)<5 serão classificados como 0 (SANTOS, 2018).

A Figura 2 ilustra um exemplo de como é o diagrama de Regressão Logística. Pelo diagrama então pode-se observar que o modelo possui uma representação gráfica em formato de ?S?, essa seria a curva logística. As predições do modelo sempre ficaram na linha 1 e 0 do eixo y, pois é o intervalo estabelecido pelo algoritmo, então a classificação sempre será nesse intervalo.

21

Figura 2. Exemplo de Diagrama de Regressão Linear. (Fonte: Batista e Filho 2019).

4.2.3 Treino, teste e validação do modelo de aprendizado de máquina

A criação de um modelo de aprendizado de máquina é realizada em duas etapas: treinamento do modelo, que é realizado a partir dos dados fornecidos para o algoritmo, e etapa de teste, onde os modelos recebem dados novos e sem o rótulo, para que seja avaliado a performance do modelo (PELISSARI, 2021).

Para o treinamento dos modelos foram utilizadas duas abordagens, a primeira onde a base foi particionada entre conjunto de dados para teste e para treino em uma porcentagem de 90% para treino e 10% para teste, porém dessa forma pode ocorrer problema de sobreajuste ou overfitting. Nesta situação, o que pode ocorrer é o modelo não aprender com os dados, e apenas ?decorar? os dados recebidos e quando é realizado o teste o modelo consegue predizer apenas situações parecidas, não sendo capaz de identificar padrões com diferenças mínimas.

Para evitar este problema, foi utilizado uma técnica de validação cruzada chamada K-folds cross-validation. Essa técnica de validação cruzada divide a base de dados em K subconjuntos, e então são realizadas várias rodadas de treinamento e teste alternando os subconjuntos utilizados para teste e treino, e o resultado do modelo baseia-se na média da acurácia observada em cada rodada (PELISSARI, 2021).

Pode-se observar na Tabela 1 uma ilustração da técnica K-fold cross-validation, onde temos um exemplo onde a base é dividida em 5 subconjuntos e cada subconjunto tem a parte de teste diferente dos outros subconjuntos, assim o modelo vai receber valores novos de teste e treino todas as vezes que executar um novo subconjunto.

22

Tabela 1. Exemplo da separação K-fold cross-validation. (Fonte: Adaptado de PELISSARI, 2021).

Predição 1	Predição 2	Predição 3	Predição 4	Predição 5
Teste	Treino	Treino	Treino	Treino
Treino	Teste	Treino	Treino	Treino
Treino	Treino	Teste	Treino	Treino
Treino	Treino	Treino	Teste	Treino
Treino	Treino	Treino	Treino	Teste

4.2.4 Métricas para avaliação de modelos

Para fins de avaliação, neste trabalho foram utilizadas três métricas para avaliar o desempenho dos modelos: a Matriz de Confusão e a AUC (Area under Receiver Operating Characteristic Curve - ROC). Na matriz de confusão pode-se observar o comportamento do modelo em relação a cada uma das classes a serem preditas pelo modelo, utilizando variáveis binárias (positivo: 1; e negativo: 0), para apresentar uma matriz que faz uma comparação

entre as predições do modelo e os valores corretos do conjunto de dados (PELISSARI, 2021).
Na Tabela 2 temos um exemplo de uma matriz de confusão, a qual consiste em duas colunas e duas linhas, e os valores são atribuídos nos quadrantes com os seguintes resultados (PELISSARI, 2021):

? quando o valor real do conjunto de dados é 0, e a predição do modelo também classificou como 0, então temos um Verdadeiro Negativo (VN);
? quando o valor real é 1, e o modelo classificado como 0, temos um Falso Negativo (FN);
? quando o valor real é 0, e o modelo classificado como 1, então o resultado se enquadra no quadrante de Falso Positivo (FP);
? e por fim o quadrante Verdadeiro Positivo (VP), que são os resultados que tem o valor real 1, e o modelo classificado como 1.

23

Tabela 2. Modelo de Matriz de confusão. (Fonte: Adaptado de PELISSARI, 2021).

Valor Predito
Negativo (0) Positivo (1)
Classe
Real
Negativo
(0)
Verdadeiro
Negativo
Falso
Positivo
Positivo
(1)
Falso
Negativo
Verdadeiro
Positivo

A segunda métrica de avaliação utilizada foi a Curva ROC que permite avaliar o desempenho de um modelo com relação às predições efetuadas. A curva ROC é um gráfico de Sensibilidade (taxa de verdadeiros positivos) versus taxa de falsos positivos, a representação da curva ROC permite evidenciar os valores para os quais existe otimização da Sensibilidade em função da Especificidade (CABRAL, 2013).
Na Figura 3 temos um exemplo de uma curva ROC. A linha traçada na cor preta é uma linha de referência, ela representa hipoteticamente a curva ROC de um classificador puramente aleatório. Caso a linha laranja, que representa o resultado do modelo que está sendo avaliado, ficasse igual a linha tracejada preta, indicaria que o modelo avaliado estaria predizendo aleatoriamente os resultados.

Figura 3: Exemplo de Curva ROC. (Fonte: Elaboração Própria).

Por uma análise de inspeção visual, quanto mais próximo a linha laranja estiver do valor 1 no eixo Y, melhor está a capacidade do modelo para diferenciar as classes. O valor da

24

AUC representa a capacidade do modelo de diferenciar as classes, quanto maior o valor do AUC, melhor é a predição realizada pelo modelo, uma vez que, os valores resultantes iguais a 0 são realmente 0, e os iguais a 1 são de fato 1 (PELISSARI, 2021).

5 METODOLOGIA

O desenvolvimento deste trabalho foi realizado em três etapas: (1) Pré-processamento e Análise Exploratória do conjunto de dados, onde foram aplicadas técnicas comuns para preparar o conjunto de dados como tratamento de valores nulos e seleção de variáveis de interesse; (2) Implementação de modelos de aprendizado de máquina supervisionado para predição das mortes neonatais, utilizando os algoritmos de árvore de decisão e regressão logística; e (3) Análise de Resultados, que será realizada uma análise do desempenho do modelo.

5.1 TECNOLOGIAS E FERRAMENTAS

Neste trabalho vamos utilizar a linguagem de programação Python (PYTHON, 2021) pela sua simplicidade de codificação e alto poder de processamento. Foi utilizado também a plataforma Google Colab Research (Colab, 2021), para o desenvolvimento dos algoritmos e treinamento dos modelos. Além disso, foi utilizado também o Jupyter Notebook (Jupyter, 2021), instalado localmente em computador pessoal, pois o Google Colab Research possui limitação de recursos de memória e processamento, então as duas plataformas foram usadas para realização deste trabalho. As principais bibliotecas utilizadas foram: numpy (NumPy, 2021), e pandas (Pandas, 2021), para a manusear os dados, matplotlib (Matplotlib, 2021), e seaborn (Seaborn, 2021), para a plotagem dos gráficos, e por fim a biblioteca sklearn (Scikit-Learn, 2021), que é a biblioteca responsável pelos algoritmos de aprendizado de máquina.

5.2 BASE DE DADOS

As bases de dados a serem utilizadas serão o SIM e SINASC, que são as duas principais fontes de informações sobre nascimentos e óbitos no Brasil. O SINASC é alimentado com base na Declaração de Nascido Vivo (Declaração de Nascido Vivo - DNV) (BELUZO et al.,2020b apud Oliveira et al., 2015). O SIM tem como objetivo principal apoiar a coleta, armazenamento e processo de gestão de registros de óbitos no Brasil (BELUZO et

25

al.,2020b apud Morais et al., 2017), e foi usado para rotular o óbito registrado no SIM, utilizando o campo DNV como chave de associação. O SIM será utilizado para rotular os registros do SINASC, pois o SIM coletas informações sobre mortalidade e é usado como base para o cálculo de estatísticas vitais, como a taxa de mortalidade neonatal e o SINASC reúne informações sobre dados demográficos e epidemiológicos do bebê, da mãe, do pré-natal e do parto (BELUZO et al.,2020b).

A Tabela 3 apresenta as variáveis da base de dados. Essa base tem variáveis que contêm valores quantitativos e categóricos. A coluna ?num_live_births? por exemplo contém o número de nascidos vivos anteriores da mãe e é composta por valores quantitativos contínuos, e a coluna ?tp_pregnancy? utiliza valores nominais categóricos que indicam o tipo de gravidez.

Tabela 3. Colunas da base de dados e suas descrições. (Fonte: Adaptado de BELUZO et al.,2020b).

Nome da coluna	Descrição	Domínio de dados

Variáveis demográficas e socioeconômicas
maternal_age Idade da mãe Quantitativo Contínuo (inteiro)
tp_maternal_race Raça / cor da pele da
mãe
Nominal categórico (inteiro)
1 - branco; 2 - Preto; 3 -
amarelo; 4 - Pele morena; 5 -
Indígena.
tp_marital_status Estado civil da mãe Nominal categórico (inteiro)
1 - Único; 2 - Casado; 3 - Viúva;
4 - Separados / divorciados
judicialmente; 5 - Casamento
por união estável; 9 - Ignorado.
tp_maternal_schooling Anos de escolaridade da
mãe
Nominal categórico (inteiro)
1 - nenhum; 2 - de 1 a 3 anos; 3 -
de 4 a 7 anos; 4 - de 8 a 11 anos;
5 - 12 e mais; 9 - Ignorado.
Variáveis obstétricas maternas
num_live_births Número de nascidos
vivos
Quantitativo Contínuo (inteiro)
num_fetal_losses Número de perdas fetais Quantitativo Contínuo (inteiro)
num_previous_gestations Número de gestações Quantitativo Contínuo (inteiro)
26
anteriores
num_normal_labors Número de partos
normais (trabalhos de
parto)
Quantitativo Contínuo (inteiro)
num_cesarean_labors Número de partos
cesáreos (partos)
Quantitativo Contínuo (inteiro)
tp_pregnancy Tipo de gravidez Nominal categórico (inteiro)
1 - Singleton; 2 - Twin; 3 -
Trigêmeo ou mais; 9 - Ignorado.
Variáveis relacionadas a cuidados anteriores
tp_labor Tipo de parto infantil
(tipo de parto)
Categórico Nominal (inteiro)
1 - Vaginal; 2 - Cesariana;
num_prenatal_appointments Número de consultas de
pré-natal

Quantitativo Contínuo (inteiro)
tp_robson_group Classificação do grupo
Robson
Ordinal categórico (inteiro)
Variáveis relacionadas ao recém-nascido
tp_newborn_presentation Tipo de apresentação de
recém-nascido
Categórico Nominal (inteiro)
1 - Cefálico; 2 - Pélvico ou
culatra; 3 - Transversal; 9 -
Ignorado.
has_congenital_malformation Presença de
malformação congênita
Nominal categórico (inteiro)
1 - Sim; 2 - Não; 9 - Ignorado
newborn_weight Peso ao nascer em
gramas
Quantitativo Contínuo (inteiro)
cd_apgar1 Pontuação de Apgar de 1
minuto
Ordinal categórico (inteiro)
cd_apgar5 Pontuação de Apgar de 5
minuto
Ordinal categórico (inteiro)
gestaional_week Semana de gestação Quantitativo Contínuo (inteiro)
tp_childbirth_care Assistência ao parto Categórico Nominal (inteiro)
1 - Doutor; 2 - enfermeira ou
obstetra; 3 - Parteira; 4 - outros;
9 - Ignorado.
was_cesarean_before_labor Foi cesáreo antes do
parto.
27
was_labor_induced Foi induzido ao trabalho
de parto.
is_neonatal_death Morte antes de 28 dias
(rótulo)
Nominal categórico (inteiro)
0 - sobrevivente; 1 - morto.

Neste trabalho foi utilizado dados do período de 2014 à 2016 devido sua melhor qualidade. Este recorte possui 6.760.222 registros dos quais 99.4% são amostras de recém-nascidos vivos e 0.6% são amostras onde os recém-nascidos vieram a óbito conforme pode ser observado na Figura 4. Com essas informações consegue-se observar que a base de dados é desbalanceada.

Figura 4. Distribuição da base de dados considerando se o recém-nascido viveu ou

morreu durante o período neonatal. (Fonte: Elaboração Própria).

5.3 PREPARAÇÃO DA BASE DE DADOS

Nessa etapa foi realizada a preparação da base de dados para o treinamento dos modelos que consiste na limpeza, transformação e preparação dos dados a serem utilizados na análise exploratória e na criação dos modelos preditivos. A base de dados pode conter informações desnecessárias para o modelo que pode acabar atrapalhando as predições,

28

algumas colunas podem ser retiradas ou seus valores podem ser modificados, por exemplo valores reais são modificados para inteiros.

5.4 DESENVOLVIMENTO DOS MODELOS DE APRENDIZADO DE MÁQUINA

Como já dito anteriormente na seção 4 os algoritmos de aprendizado de máquina escolhidos para esse trabalho foram os: Regressão Logística e Árvore de Decisão que utilizam a lógica mostrada na seção 4 , Fundamentação Teórica. Esses dois foram escolhidos por alguns motivos, ambos são algoritmos de classificação e supervisionados, atendendo os critérios necessários para a predição de mortalidade neonatal, ambos os algoritmos são bons para realizar predições, a Árvore de Decisão inclusive é um ótimo modelo para analisar as decisões que estão sendo feitas pelo modelo, afinal esse algoritmo tem a característica de ser um WhiteBox, ou seja conseguimos ver o que o modelo aprendeu e o que ele está decidindo. Esses algoritmos selecionados são algoritmos que se encaixam na necessidade deste trabalho e são muito utilizados na comunidade acadêmica.

Para o algoritmo de Regressão Logística, como foi dito anteriormente na seção ?Preparação da base de dados? a base de dados foi tratada e particionada, no particionamento foi utilizado 90% da base para o treino e 10% para os testes e então foi realizado o treinamento do modelo, também foi aplicada a função RFECV (Eliminação Recursiva de Feature com Validação Cruzada), esse função executa a RFE (Eliminação Recursiva de Feature), que tem como ideia selecionar features considerando recursivamente conjuntos cada vez menores de features, isso ocorre da seguinte forma. Primeiro, o estimador é treinado no conjunto inicial de features e a importância de cada feature é obtida por meio de um atributo "coef" ou por meio de um atributo "feature_importances". Em seguida, as features menos importantes são removidas do conjunto atual de features. Esse procedimento é repetido recursivamente no conjunto removido até que o número desejado de features a serem selecionados seja finalmente alcançado. Porém o diferencial da RFECV é que essa função executa a RFE em um loop de validação cruzada para encontrar o número ideal ou o melhor número de features.

Após essa seleção, foi realizado o treinamento do modelo usando todas as features e também treinamos usando as features que o RFECV selecionou, para compararmos os resultados no final. Também foi utilizado o método de K-folds cross-validation em um novo treinamento para conferirmos se o modelo estava sofrendo overfitting (sobreajuste). E por fim aplicamos um método chamado GridSearchCV, esse método permite que seja realizado testes

29

alterando algumas combinação de parâmetros nos nossos modelos, facilitando para achar a melhor combinação, esse método é parecido com o cross validation, o diferencial do Grid Search é que ele faz os cross validation de vários modelos com hiperparâmetros diferentes de uma só vez.

Para o algoritmo de Árvore de Decisão também foi utilizado o mesmo particionamento de 90% para treino e 10% para teste, na criação do modelo foi colocado a entropy como critério de decisão e uma profundidade máxima de 4, pois com números maiores o modelo ficava muito complexo e ele errava mais as predições, também utilizamos um algoritmo para mostrar a importância de cada Feature para o modelo.
O algoritmo de Regressão Logística e análise exploratória foi baseado em códigos disponível na plataforma web Kaggle (MNASSRI, 2020), no exemplo do Kaggle o autor faz os códigos para predizer mortes no navio Titanic, para o trabalho de mortalidade neonatal os códigos foram alterados para realizar predição de mortes neonatais. Já no algoritmo de Árvore de Decisão foi baseado em uma vídeo aula do professor Diogo Cortiz (ÁRVORE..., 2020), nessa vídeo aula o professor explica sobre os algoritmos de Árvore de Decisão e depois implementa o um exemplo de código, o código usado neste trabalho usou o codigo do professor Diogo Cortiz, e foi alterado para realizar predições de mortalidade neonatal.

5.5 ANÁLISE DOS RESULTADOS

Para a análise de resultados foi utilizado dois métodos que é o de Matriz de confusão e a curva ROC esses métodos são explicados na seção 4, a Fundamentação Teórica. Com esses métodos pode-se realizar uma avaliação do modelo, e assim será possível analisar os valores de acurácia, precisão e recall do modelo, essas métricas são utilizadas avaliar o desempenho do modelo, com a acurácia é possível calcular a proximidade entre o valor obtido na predição dos modelos e os valores esperados, com a precisão é possível analisar as amostras que o modelo classificou como 0 eram realmente 0 na classe real e com o recall é possível analisar as amostras que o modelo conseguiu reconhecer como a classe correta.

5.6 ACESSO A BASE DE DADOS E CÓDIGOS

A base de dados utilizada neste projeto pode ser encontrada no link: ... . E os códigos podem ser encontrados no link: ... .

30

6 RESULTADOS

Os mesmos experimentos mostrados abaixo foram aplicados para uma base de dados que contém somente dados de São Paulo, essa base é um recorte da base do Brasil todo em contém cerca de 1.400.000 amostras, esse recorte da base também é bem desbalanceado. Os resultados obtidos nesse experimento usando o recorte de São Paulo foram bem parecidos com os experimentos da base do Brasil todo, e os códigos desse experimento podem ser visualizados no apêndice X.

6.1 ANÁLISE EXPLORATÓRIA

O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

6.1.1 Apresentação dos Dados

Como dito anteriormente na seção 5.2 ?Base de Dados? deste trabalho, a base de dados é formada pelas informações do SIM e do SINASC contém 6.760.222 amostras e 29 colunas, porém serão utilizadas somente 23 colunas sendo elas as colunas descritas na Tabela 3. Na seção 5.2 também tem a Figura 4 que mostra que a base dados é desbalanceada tendo 99.4% das amostras de recém-nascidos vivos e 0.6% das amostras onde os recém-nascidos vieram a óbito no período neonatal.

6.1.1.1 Caracteristicas demográficas e socioeconômicas maternas

O apêndice X contém uma tabela que possui as estatísticas sumarizadas das variáveis demográficas e socioeconômicas maternas. Nesta tabela por exemplo podemos observar a coluna ?maternal_age? que contém a idade da mãe, e na tabela mostra que a média da idade das mães é de 26.4 anos, e os dados possuem uma variação de no mínimo 8 anos e no máximo 55 anos, e a mediana é 26 anos. Na tabela também tem as colunas: ?tp_maternal_schooling? que mostra a categoria de anos de escolaridade da mãe, ?tp_marital_status? que mostra o estado civil da mãe em dados categóricos e a coluna ?tp_maternal_race? que mostra a raça da mãe também em dados categóricos.

31

6.1.1.2 Variávies obstétricas maternas

No apêndice X pode-se observar uma tabela que possui as estatísticas sumarizadas das variáveis obstétricas maternas. Nesta tabela por exemplo temos a coluna ?num_live_births?, essa coluna mostra o número de nascidos vivos anteriores que a mãe obteve, a média desta coluna é 0.94 , a mediana é 1, o número mínimo é 0 e o máximo é 10 nascidos vivos. Na tabela podemos observar também a coluna, ?num_fetal_losses?, esta coluna mostra o número de perdas fetais anteriores da mãe, a média desta coluna é 0.21, a medina é 0, o número mínimo é 0 e o máximo é 5, esses valores mostram que poucas mães tiveram perdas fetais antes da gravidez atual. Também na tabela tem estatísticas da coluna ?num_previous_gestations? esta coluna mostra o número de gestações anteriores da mãe, esta coluna tem a média de 1.14, a mediana é 1, o número mínimo é 0 e o máximo 10. Essa tabela também contém as colunas: ?num_normal_labors? que é o número de partos normais da mãe, ?num_cesarean_labor? que mostra o número de partos cesáreos da mãe e por fim mostra a coluna ?tp_pregnancy? que é o tipo da gravidez.

6.1.1.3 Variávies de histórico de gravidez

No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis do histórico de gravidez. Nesta tabela temos as colunas ?num_prenatal_appointments? que mostra o número de consultas de pré-natal, a média dessa coluna é 7.8, a mediana é 8, a variação dos dados é de no mínimo 0 e no máximo 40. A tabela também contém as colunas: ?tp_labor? que mostra o tipo de parto, e também contém a coluna ?tp_robson_group? que mostra a classificação do grupo Robson.

6.1.1.4 Variávies relacionadas ao recém-nascido

No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis relacionadas ao recém-nascido. Nesta tabela por exemplo podemos observar a coluna ?newborn_weight? que armazena o peso ao nascer dos recém-nascidos em gramas, a média dos dados dessa coluna é 3187.3, a mediana é 3215, a variação dos dados é de no mínimo 0 e no máximo 6000 gramas. Também podemos observar a coluna ?gestacional_week? que mostra o número de semanas de gestação, a média é 38.4, a mediana é 39, a variação dos dados é de no mínimo 15 e no máximo 45 semanas. Além dessas colunas a tabela também

32

contém as colunas: ?cd_apgar1? que mostra a pontuação de Apgar de 1 minuto, ?cd_apgar5? que mostra a pontuação de Apgar de 5 minuto, ?has_congenital_malformation? que mostra se o recém-nascido contém a presença de alguma malformação, ?tp_newborn_presentation? que mostra o tipo de apresentação de recém-nascido, ?tp_labor? que mostra o tipo de parto, ?was_cesarean_before_labor? que mostra se o recém-nascido foi cesáreo antes do parto,

?was_labor_induced ? que mostra se o recém-nascido foi induzido ao trabalho de parto, ?tp_childbirth_care? que mostra se ouve assistência ao parto, e por fim temos a coluna ?is_neonatal_death? que mostra se o recém-nascido veio a óbito ou não.

6.1.2 Análise das variáveis

Nesta seção serão mostrados a parte de análise de variáveis com diferentes tipos de gráficos.

6.1.2.1 Variáveis contínuas

Na Figura 5 temos dois gráficos boxplot, no boxplot à esquerda temos a distribuição da variável peso ao nascer e nele é possível observar duas caixas onde a caixa azul são os vivos e o laranja são os mortos. Nas duas caixas mostradas no gráfico podemos observar pequenos círculos nas pontas, esses círculos são outliers (dados fora da curva), então para a caixa de vivos os outliers são recém-nascidos com pesos acima de 4500 gramas ou abaixo de 2000 gramas, e para a caixa de mortos os outliers são os recém-nascidos com mais de 5500 gramas. Para os vivos temos a mediana de aproximadamente 3200 gramas e para os mortos a mediana é 1300 gramas. Cada caixa representa 50% da base de dados, a caixa azul de vivos mostra que o peso de recém-nascidos que sobreviveram está aproximadamente entre 2700 a 3600 gramas, e para a caixa laranja de mortos o peso de recém-nascidos que vieram a óbito estão aproximadamente entre 700 a 2500 gramas. Então o gráfico mostra para nós que os recém-nascidos que têm maior risco de vir a óbito tem pesos menores que 2000 gramas.

33

Figura 5. BoxPlot das features Peso ao nascer e Idade da mãe. (Fonte: Elaboração Própria).

No gráfico à direita da Figura 5 temos um boxplot da distribuição da variável idade da mãe, como foi dito anteriormente a caixa azul são os vivos e o laranja são os mortos. Nesse boxplot então podemos ver que os outliers da caixa de vivos são mães com idade acima de 46 anos aproximadamente, para os mortos os outliers são mães com idade acima de 50 anos. Para os vivos temos a mediana de aproximadamente 26 anos e para os mortos a mediana é de 26 anos. Nos vivos a idade das mães está aproximadamente entre 21 a 32 anos, e para os mortos a idade das mães está aproximadamente entre 20 a 34 anos. É importante ressaltar que o gráfico mostra as duas caixas bem parecidas, então de acordo com os dados não contém diferença significativa entre vivos e mortos usando a idade da mãe para distribuir.

Já na Figura 6 temos um boxplot da distribuição da variável semanas de gestação. Nesse boxplot então podemos ver que os outliers da caixa de vivos são gestações com mais de 44 semanas aproximadamente, e gestações com menos de 35 semanas. Para os vivos temos a mediana de aproximadamente 39 semanas e para os mortos a mediana é de 32 semanas. Nos vivos as semanas de gestação estão aproximadamente entre 36 a 40 semanas, e para os mortos a semanas de gestação estão aproximadamente entre 26 a 36 semanas. Nesse boxplot os dados mostraram que para as gestações que têm menos de 36 semanas os recém-nascidos têm maior risco de vir a óbito.

34

Figura 6. BoxPlot da feature semana de gestação. (Fonte: Elaboração Própria).

Na Figura 7 temos um histograma da distribuição de peso ao nascer por classe, podemos observar no gráfico que grande parte dos vivos (linha azul) estão distribuídos entre 2000 e 4000 gramas, a área entre esses valores tem muitos dados de vivos, já entre os valores 0 e 2000 gramas temos uma concentração dos dados de mortos.

Figura 7. Histograma de distribuição do peso entre as classes. (Fonte: Elaboração Própria)

6.1.2.2 Variáveis categoricas

Observando a Figura 8 podemos ver um gráfico de barras da distribuição da variável escolaridade da mãe, este gráfico mostra a contagem de registros por escolaridade da mãe, esse gráfico mostra para nós que a maior parte das mães estão na categoria 4 que representa de 8 a 11 anos de escolaridade.

35

Figura 8. Gráfico de barras da distribuição da variável Escolaridade da mãe. (Fonte: Elaboração Própria).

Observando a Figura 9 podemos ver um gráfico de barras da distribuição da variável número de vivos, este gráfico mostra a contagem de registros por número de vivos, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho.

Figura 9. Gráfico de barras da distribuição da variável Número de nascidos vivos. (Fonte: Elaboração Própria).

Observando a Figura 10 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 1 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 1 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem

36

apgar de 8 ou 9, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.

Figura 10. Gráfico de barras da distribuição da variável Pontuação Apgar de 1 minuto. (Fonte: Elaboração Própria).

Observando a Figura 11 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 5 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 5 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem apgar de 9 ou 10, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.

Figura 11. Gráfico de barras da distribuição da variável Pontuação Apgar de 5 minutos. (Fonte: Elaboração Própria).

37

Observando a Figura 12 podemos ver um gráfico de barras da distribuição da variável presença de malformação congênita, este gráfico mostra a contagem de registros por presença de malformação congênita, esse gráfico mostra para nós que a grande maioria dos dados estão na categoria 2 que mostra que o recém-nascido não possui malformação, esse grande volume para a categoria 2 é causada pelo desbalanceamento da base.

Figura 12. Gráfico de barras da distribuição da variável Presença de malformação congênita. (Fonte: Elaboração Própria).

Observando a Figura 13 podemos ver um gráfico de barras da distribuição da variável número de gestações, este gráfico mostra a contagem de registros por número de gestações, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho, da mesma forma mostrada na Figura 9.

38

Figura 13. Gráfico de barras da distribuição da variável Número de gestações. (Fonte:

Elaboração Própria).
Observando a Figura 14 podemos ver um gráfico de barras da distribuição da variável assistência ao parto, este gráfico mostra a contagem de registros por assistência ao parto, esse gráfico mostra para nós que a grande maioria dos dados mostram que o parto teve assistência de uma Enfermeira ou obstetra ou essa informação foi ignorada na hora do resgistro.
Figura 14. Gráfico de barras da distribuição da variável Assistência ao parto. (Fonte: Elaboração Própria).
6.1.2.3 Analise bi-variada
A Figura 15 mostra para nós a importância da feature peso ao nascer com o gráfico mostrado é possível observar claramente a diferença de pesos entre vivos e mortos, recém-nascidos com peso acima de 2000 gramas sobreviveram, já os recém-nascidos com menos de 2000 gramas vieram a óbito.
39
Figura 15. Gráfico de densidade de peso entre as classes. (Fonte: Elaboração Própria).
Na Figura 16 podemos observar um gráfico de correlação entre algumas features de valores contínuos, algumas dessas features possuem correlações positivas, como pontuação apgar de 1 minuto e pontuação apgar de 5 minutos, essas colunas possuem uma correlação de 0.7, então nascidos que recebem um valor alto de apgar de 1 minuto tendem a receber um valor alto no apgar de 5 minutos. O gráfico também mostra outras features com correlações positivas como, número de partos normais e número de gestações anteriores que contém uma correlação de 0.81, número de gestações anteriores com idade da mãe que tem uma correção de 0.39, número de partos de cesáreos com idade da mãe que possui correlação de 0.24, número de partos normais com idade da mãe com a correlação de 0.26, semanas de gestação com peso ao nascer em gramas com correção de 0.52, e as apgar de 1 e 5 minutos com semanas de gestação. E grande parte das features, possuem correlações baixas então elas são inversamente correlacionadas pelo que o gráfico mostra, como por exemplo número de partos normais com número de consultas pré-natais possuem uma correção de -0.17, e número de partos cesáreos com número de partos normais que contém uma correlação de -0.15. Então a correlação dessas features estão bem baixas tirando as correlações acima de 0.7.
40
Figura 16. Gráfico de correlação. (Fonte: Elaboração Própria)
6.1.2.4 Distribuição por raça
Na Figura 17 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por raça, e podemos observar que a raça 5 (indigena), é o boxplot que contem a menor variação de idade, já as raças 1 (branco) e 3 (amarelo) são as caixas que possuem maior variação com a média em torno de 20 e 30 anos.
41
Figura 17. Distribuição da idade da mãe por raça. (Fonte: Elaboração Própria).
No boxplot mostrado na Figura 18 podemos observar a distribuição de peso ao nascer por raça, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre as raças para peso ao nascer.
Figura 18. Distribuição de peso ao nascer por raça. (Fonte: Elaboração Própria).
Na Figura 19 temos o boxplot da a distribuição de semanas de gestação por raça, e praticamente todas as caixas são iguais, somente a caixa da raça 1 (branco) possui menor

variação que as outras.
42
Figura 19. Distribuição de semanas de gestação por raça. (Fonte: Elaboração Própria).
6.1.2.5 Distribuição por estado civil
A Figura 20 mostra o boxplot da a distribuição de peso ao nascer por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre os estados civis para peso ao nascer.
Figura 20. Distribuição de peso ao nascer por estado civil. (Fonte: Elaboração Própria).
A Figura 21 mostra o boxplot da a distribuição de semanas de gestação por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Porém os estados civis 2 (Casado) e 4 (Separados/divorciados judicialmente), contém variações menores.
43
Figura 21. Distribuição de semanas de gestação por estado civil. (Fonte: Elaboração Própria).
Na Figura 22 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por estado civil, e podemos observar que o estado civil 3 (Viúva), possui a maior variação, já o estado civil 4 (Separados/divorciados judicialmente) é a caixa que possuem menor variação
Figura 22. Distribuição da idade da mãe por estado civil. (Fonte: Elaboração Própria).
6.1.2.6 Distribuição por escolaridade da mãe
Na Figura 23 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por escolaridade da mãe, e podemos observar que a escolaridade 2 (1 a 3 anos) e 3 (4 a 7
44
anos), possui as maiores variações, já escolaridade 5 (12 ou mais anos) é a caixa que possuem menor variação
Figura 23. Distribuição da idade da mãe por escolaridade da mãe. (Fonte: Elaboração Própria).
A Figura 24 mostra o boxplot da a distribuição de peso ao nascer por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então, pouca variação é observada entre a escolaridade da mãe para peso ao nascer.
Figura 24. Distribuição de peso ao nascer por escolaridade da mãe. (Fonte: Elaboração Própria).
A Figura 25 mostra o boxplot da a distribuição de semanas de gestação por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas,
45
todas são praticamente iguais. Porém a escolaridade 5 (12 ou mais anos) contém variação menor que as outras.
Figura 25. Distribuição de semanas de gestação por escolaridade da mãe. (Fonte: Elaboração Própria).
6.2 ÁRVORE DE DECISÃO
A Figura 26 mostra um trecho do algoritmo utilizado, nesse trecho é realizado a divisão do DataFrame das features de entrada , e o DataFrame que contém o rótulo.Neste

caso o rótulo será o DataFrame ?target?, esse Dataframe contém a feature ?is_neonatal_death?, essa feature informa se o recém nascido veio a óbito ou não, sendo o foco da predição que desejamos realizar essa feature entra como rótulo para o modelo, já no Dataframe ?nome_features? temos as features que vão servir de entrada para o modelo, é a partir dessas features que o modelo tentará encontrar padrões para predizer o risco de óbito do recém nascido. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

46

Figura 26. Separação dos DataFrames de entrada e rótulo. (Fonte: Elaboração Própria).

6.2.1 Modelo de Árvore de Decisão Utilizando Particionamento 90/10

O algoritmo de árvore de decisão foi treinado utilizando duas formas diferentes: usando as partições 90% para treino e 10% para os testes e utilizando o método K-folds cross-validation. E para os experimentos iniciais foi utilizado o particionamento 90/10.

Tabela 4. Resultados do modelo de árvore de decisão. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	671807
Morto(1)	0.71	0.29	0.41	4216

Analisando a Tabela 4 pode-se ver os seguintes resultados, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 71% que realmente pertenciam a classe 1, então cerca de 29% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 29% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem a classe 1 que seria dos recém nascidos que poderiam vir a óbito, já para a classe de vivos a

47

classe 0 o modelo está identificando muito bem e classificando de forma correta. A tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.

Esse modelo teve uma acurácia de 0.99 que é uma acurácia bem alta, porém somente pela acurácia não é possível dizer que o modelo está excelente pelo motivo da base que está sendo usada é altamente desbalanceada, então para a classe 0 que é a de vivos temos muito mais dados que para a classe de mortos, e o modelo está acertando bastante a classe de vivos logo a acurácia fica alta por esses acertos, enquanto para a classe de mortos o modelo não está acertando tanto.

A Figura 27 mostra a curva ROC do modelo, pode-se observar que a AUC da curva ROC ficou em 0.92, a AUC mostra que o modelo está acertando bastante as predições, pois quanto mais perto de 1 melhor está no nosso modelo, mas no caso desse modelo vimos acima que as predições para a classe de mortos(1) está ruim, o modelo está acertando poucas classificações como 1. Então a AUC assim como a acurácia está alta porque a base de dados utilizada é altamente desbalanceada tendo muito mais amostras de vivos(1), e como o modelo está bom para essa classificação e está acertando bastante os valores de AUC e acurácia ficam

altos .
Figura 27. Curva ROC modelo de Árvore de decisão. (Fonte: Elaboração Própria).
A Figura 28 mostra a importância de cada feature para o modelo sendo assim a feature 10 - Peso ao nascer, a mais importante para o modelo, pois dela o modelo conseguiu tirar mais informações, seguido das features, 13 - Apgar dos 5 primeiros minutos, 11 - Semanas de
48
Gestação, 14 - Presença de malformação 12 - Apgar do 1 primeiro minutos. Então essas são as features que foram mais decisivas para a montagem da árvore e para as predições do modelo.
Figura 28. Gráfico de importância das features. (Fonte: Elaboração Própria).
Na Figura 29 temos uma imagem reduzida da árvore de decisão que foi montada com o treinamento utilizando 5 como profundidade máxima para a árvore, a imagem da árvore completa pode ser visualizada no apêndice A. É possível ver que as features marcadas como importante para o modelo apareceu diversas vezes na árvore, e a feature peso ao nascer foi escolhida para ser o nó raiz da árvore, de todas as features ela foi a que teve a melhor entropia para ser o nó raiz, e depois aparece mais vezes para outras decisões em outros níveis da árvore e isso faz com que essa feature se torne a mais importante para o modelo, também podemos ver que a feature ?Apgar do 1 primeiro minuto? mesmo sendo a menos importante para o modelo ela aparece somente no nó de número 42 isso mostra que a entropia e o ganho de informação dessa feature nas outras decisões não foram boas fazendo ela ser a feature com menos importância utilizada no modelo.
49
Figura 29. Árvore de decisão montada a partir do algoritmo. (Fonte: Elaboração Própria).
6.2.2 Modelo de Árvore de Decisão Utilizando K-folds cross-validation
Após realizar esse experimento com a partição 90/10, foi realizado outro experimento com esse mesmo algoritmo porém agora utilizando o método K-folds cross-validation para o treinamento. O K-folds cross-validation foi configurado para separar a base de dados em 10 partes iguais, e assim o algoritmo foi treinado novamente gerando um novo modelo. O método de K-folds cross-validation é utilizado para conferir se o modelo não está sofrendo problemas de overfitting, é importante garantir que o modelo está aprendendo com os dados e não está somente decorando.
Na Tabela 5 temos então o resultados desse modelo, pode-se observar que os resultados para a classificação de vivos(0) não teve alteração, o modelo continua tendo 100% na precisão e no recall, mas na classificação de mortos(1) o modelo teve uma diminuição na precisão que agora está em 65% e também teve uma diminuição no recall que agora está em 23%, porém essa diminuição é justificada porque para o teste deste modelo foi utilizado a base completa e não somente a partição de teste que contém 10% da base total. Logo os resultados não tiveram uma alteração drástica e tiveram um comportamento bem semelhante, incluindo a acurácia que se manteve em 0.99 e a AUC que teve uma diminuição pequena também e ficou com 0.89.
50
Tabela 5. Resultados do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)
Precision Recall F1-Score Support

Vivo(0) 1.00 1.00 1.00 6719191
Morto(1) 0.65 0.23 0.34 41031
Na Tabela 6 é possível ver a matriz de confusão do modelo que mostra o número exato de erros e acertos do modelo, pode-se observar que para a classe de vivos o modelo classificou corretamente 6.714.117 amostras da base de dados completa, o modelo classificou como 0 as amostras que nos rótulos realmente eram 0, errou apenas 5074 amostras. Já para a classe de mortos o modelo acertou somente 9552 amostras, classificou como 1 as amostras que no rótulo eram 1 também, e errou 31479. Isso mostra que o modelo está muito desbalanceado, pois está errado muito mais do que acertando as predições de mortos, sendo isso um reflexo da base de dados desbalanceada também. E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.
Tabela 6. Matriz de confusão do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6714117 5074
Mortos (1) 31479 9552
6.3 REGRESSÃO LOGÍSTICA
Para o algoritmo de Regressão Logística também foi utilizado métodos diferentes para o treinamento, O algoritmo foi treinado utilizando três formas diferentes: usando as partições 90% para treino e 10% para os testes, utilizando o método K-folds cross-validation e foi treinada utilizando o método RFECV que seleciona as melhores features para o modelo, juntamente com o método GridSearchCV. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.
51
6.3.1 Modelo de Regressão Logística Utilizando Particionamento 90/10
Para o primeiro experimento do algoritmo de regressão logística foi utilizado o particionamento 90/10 para o treinamento do algoritmo, sendo 90% da base de dados para realizar o treinamento do modelo e 10% da base para realizar os testes.
Na Tabela 7 temos os resultados do modelo de regressão logística utilizando o método de particionamento 90/10, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 65% que realmente pertenciam a classe 1, então cerca de 35% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 24% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem as amostras da classe 1 que seria dos recém-nascidos que poderiam vir a óbito, já para a classe de vivos a classe 0 o modelo está identificando muito bem e classificando de forma correta.E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do

Recall.
Tabela 7. Resultados do modelo de regressão logística usando particionamento 90/10.
(Fonte: Elaboração Própria)
Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 671885
Morto(1) 0.65 0.24 0.35 4138
A Figura 30 abaixo mostra a curva ROC do modelo de regressão logística usando o particionamento 90/10, analisando a curva ROC podemos ver que a AUC da curva ficou em 0.93 e a acurácia do modelo ficou 0.99, mostrando que o modelo está acertando bastante as predições, porém esses números para o nosso modelo não mostra que ele está ótimo, com a análise da Tabela 7 acima podemos observar que o modelo está acertando muitas predições da classe de vivos e poucas predições da classe de mortos, como a base de dados é
52
desbalanceada, como a maior quantidade de dados está na classe de vivos e o modelo está acerto quase todos dessa classe a acurácia e a AUC ficam altas por esses acertos.Logo é preciso analisar mais resultados além da acurácia e da AUC do modelo.
Figura 30. Curva ROC modelo regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria).
Na Tabela 8 podemos analisar os número exatos que o modelo acertou de cada classe, como é mostrado na matriz de confusão o modelo acertou 671.435 da classe de vivos e acertou apenas 915 da classe de mortos, e errou mais que o triplo da classe de mortos. Então as predições do modelo estão muito desbalanceadas, para a classe de vivos o modelo está predizendo bem, porém para as classes de mortos o modelo está errando muitas predições.
Tabela 8. Matriz de confusão do modelo de regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 671435 504
Mortos (1) 3169 915
6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation
Para o segundo experimento do algoritmo de regressão logística foi utilizado o método K-folds cross-validation para o treinamento do modelo com a intenção de conferir e confirmar
53
que o modelo não esteja tendo problemas com overfitting e realmente esteja aprendendo com os dados que está sendo usado para o treinamento. O método K-folds cross-validation foi configurado para realizar uma separação de 10 partes iguais.
Na Tabela 9 podemos observar os resultados do modelo, e os resultados foram parecidos com o experimento anterior usando o particionamento 90/10, única diferença nos resultado é que no experimento anterior a precisão da classe de mortos ficou em 65% e no caso desse experimento usando o K-folds cross-validation a precisão ficou em 64%, mas os valores de recall e F1-Score se mantiveram, assim como todos os resultados da classe de mortos se mantiveram em 100%. A execução desse experimento mostra que o modelo realmente está aprendendo com os dados e não está apenas decorando, retirando o risco do

modelo estar com overfitting. Mesmo sem alterações no resultado é importante saber que o modelo não está com overfitting e está conseguindo aprender e encontrar padrões nos dados.
Tabela 9. Resultados do modelo de regressão logística usando K-folds cross-validation. (Fonte: Elaboração Própria)
Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 6719191
Morto(1) 0.64 0.24 0.35 41031
Já na Figura 31 temos a curva ROC do modelo, essa curva ROC mostra a AUC de todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas para todas as vezes que foi realizado os testes, a AUC se manteve em 0.93 e 0.94, sendo a média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua acertando muito a classe de vivos que é a classe predominante na base de dados. Já era esperado os resultados não sofrerem alterações grandes, pois a base de dados não sofreu nenhuma alteração, esse experimento foi somente para conferir se a modelo não estava sofrendo overfitting, e pelos resultados aparentemente a base não está com problemas de sobreajuste.
54
Figura 31. Curva ROC modelo regressão logística usando K-folds cross-validation. (Fonte: Elaboração Própria).
Na Tabela 10 temos a matriz de confusão do modelo, onde mostra que o modelo acertou 6.713.658 amostras da classe de vivos e errou apenas 5.533, já para a classe de mortos o modelo acertou somente 9.847 e novamente errou mais que o triplo errando 31.182 amostras. Como os resultados mostrados anteriormente não tiveram alteração era de se esperar que as predições corretas e incorretas do modelo também não sofressem alterações, o modelo continua acertando muito a classe de vivos e errando muito a classe de mortos, mantendo as predições bem desbalanceadas.
55
Tabela 10. Matriz de confusão do modelo de regressão logística usando K-folds cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6713658 5533
Mortos (1) 31182 9847
6.3.3 Modelo de Regressão Logística Utilizando GridSearchCV e RFECV
Para o terceiro e último experimento de regressão logística foi utilizado técnicas diferentes juntas para tentar melhorar as predições do modelo, para esse experimento usamos o K-folds cross-validation para o particionamento da base de dados assim como foi utilizado no experimento acima, e também foi utilizado duas técnicas que ainda não foram usadas nos experimentos anteriores que são as funções RFECV e GridSearchCV. A função RFECV seleciona a quantidade ideal de features para ser usadas no treinamento do modelo e também mostra quais são as melhores features para essa predição, então essa função ela utiliza a

validação cruzada para ir treinando e testando N vezes o modelo, e sempre vai alterando a quantidade e as features utilizadas para o teste, então cada vez que a função testa os modelos ela grava a pontuação dos acertos e assim depois mostra o gráfico da Figura 32 e assim é possível ver quais são as features ideais para o modelo.
56
Figura 32. Resultado da função RFECV Base Brasil. (Fonte: Elaboração Própria).
Então os resultados mostrados na Figura 32 apresenta que o número ideal para treinar o modelo de features é 6, e as features sao: 'tp_maternal_schooling', 'gestaional_week', 'cd_apgar1', 'cd_apgar5', 'has_congenital_malformation', 'tp_labor'. Essas são as features que tiveram o melhor desempenho no RFECV, então para o treinamento do modelo desse experimento as features que vão ser utilizadas serão somente essas 6.
Após a execução da função de RFECV foi executado a função de GridSearchCV, o GridSearchCV é utilizado para descobrir quais são os melhores parâmetros para utilizar no algoritmo de regressão logística e criar os modelos, foi utilizado também a validação cruzada para testar qual seria os melhores parâmetros para melhorar o desempenho do modelo. Na Figura 33 pode-se ver que os parâmetros escolhidos pela função foram:
Figura 33. Resultado da função GridSearchCV. (Fonte: Elaboração Própria).
Na Tabela 11 podemos observar os resultados do modelo, e os resultados foram parecidos com os experimentos anteriores, para a classe de mortos a precisão teve um aumento e foi para 69% e o recall diminuiu chegando em 20%, então utilizando as novas
57
funções o modelo ganhou precisão e perdeu recall. Para a classe de vivos o modelo se manteve em 100% e não teve alterações para os resultados de precisão e recall. Mesmo com a utilização das funções de RFECV e GridSearchCV o modelo não teve uma melhora significativa para as predições de mortos.
Tabela 11. Resultados do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	6719191
Morto(1)	0.69	0.20	0.31	41031

Na Figura 34 temos a curva ROC do modelo, essa curva ROC mostra a AUC de todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas para todas as vezes que foi realizado os testes, a AUC se manteve em 0.92 e 0.93, sendo a média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua acertando muito a classe de vivos que é a classe predominante na base de dados.
58
Figura 34. Curva ROC modelo regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria).
Na Tabela 12 temos a matriz de confusão do modelo, onde mostra que o modelo acertou 6.715.535 amostras da classe de vivos e errou apenas 3.656, comparando com os experimentos anteriores de regressão logística as predições de vivos teve uma pequena piora e teve uma pequena diminuição nos acertos. Já para a classe de mortos, o modelo acertou 8.296

e errou 32.735 amostras, logo as predições de vivos tiveram uma uma piora e acertou um pouco menos do que os outros experimentos, e na classe de vivos teve uma melhora acertando mais predições, porém como o problema está nas predições de mortos e para essa classe teve uma piora pode-se dizer que a função de GridSearchCV não teve um ganho significativo nos resultados. Então como modelo final foi escolhido o modelo utilizando somente o K-folds cross-validation, pois foi o modelo que mostrou o melhor desempenho dos experimentos de regressão logística.

59

Tabela 12. Matriz de confusão do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

	Predito	
	Vivos (0)	Mortos (1)
Classe Real		
Vivos (0)	6715535	3656
Mortos (1)	32735	8296

60

7 CONCLUSÃO

O principal objetivo deste trabalho foi analisar os resultados das predições de mortalidade neonatal dos algoritmos de aprendizado de máquina usando uma base de dados com dados do Brasil inteiro. A partir destes resultados apresentados na secção anterior ?Resultados?, é possível concluir que analisar somente a AUC e a acurácia do modelo não é o suficiente para garantir que o modelo está excelente, sendo assim temos que ter mais atenção a precisão, recall e as predições mostradas na matriz de confusão.

Ambos os modelos ficaram desbalanceados nas predições, os modelos acertaram mais predições de vivos do que mortos, e as predições corretas de mortos foram mais baixas do que as incorretas, para os modelos ficarem melhor precisam passar por algoritmos que possam balancear essas predições, uma outra alternativa para melhorar o modelo é utilizar uma base de dados mais balanceada, pois essa base que foi utilizada é altamente desbalanceada.

Embora os dois modelos tenham tido resultados parecidos , o modelo de regressão logística utilizando somente o K-folds cross-validation se saiu melhor que os outros experimentos, o modelo criado na seção ?6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation? manteve a acurácia, AUC, precisão e recall dos demais modelos e acertou mais predições, talvez com algoritmos para melhorar as predições da classe de mortos, balanceando mais as predições, esse modelo fique com ótimas taxas preditivas.

Além disso podes-se afirmar que o peso ao nascer do recém nascido é um fator muito importante para as decisões dos modelos, ambos os modelos usaram muito essa informação para as predições, também podemos colocar, as colunas de presença de malformação e semana de gestação, como colunas que foram importantes para os modelos, essas causas podem ser evitadas se houver um acompanhamento médico com qualidade e eficiência.

61

REFERÊNCIAS

ÁRVORE de Decisão - Aula 3. Gravação de Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=ecYpXd4WREk&t=4089s. Acesso em: 1 jul. 2020.

BARRETO, J. O. M.; SOUZA, N. M.; CHAPMAN, E. Síntese de evidências para políticas

de saúde: reduzindo a mortalidade perinatal. Ministério da Saúde, p. 1?44,2015.
BATISTA, André Filipe de Moraes; FILHO, Alexandre Dias Porto Chiavegatto. Machine Learning aplicado à Saúde. In: ZIVIANI, Artur; FERNANDES, Natalia Castro; SAADE, Débora Christina Muchaluat. SBCAS 2019: 19 Simpósio Brasileiro de Computação Aplicada à Saúde. [S. l.: s. n.], 2019. cap. Capítulo 1.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. Towards neonatal mortality risk classification: A data-driven approach using neonatal, maternal, and social factors, [s. l.], 23 out. 2020.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. SPNeoDeath: A demographic and epidemiological dataset having infant, mother, prenatal care and childbirth data related to births and neonatal deaths in São Paulo city Brazil ? 2012?2018, [s. l.], 19 jul. 2020.
BRESAN, Rodrigo. Um método para a detecção de ataques de apresentação em sistemas de reconhecimento facial através de propriedades intrínsecas e Deep Learning. Orientador: Me. Carlos Eduardo Beluzo. 2018. Trabalho de Conclusão de Curso (Superior de Tecnologia em Análise e Desenvolvimento de Sistemas) - Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas., [S. l.], 2018.
CABRAL, Cleidy Isolete Silva. Aplicação do Modelo de Regressão Logística num Estudo de Mercado. Orientador: João José Ferreira Gomes. 2013. Projeto Mestrado em Matemática Aplicada à Economia e à Gestão (Mestrado) - Universidade de Lisboa Faculdade de Ciências Departamento de Estatística e Investigação Operacional, [S. l.], 2013. Disponível em: https://repositorio.ul.pt/bitstream/10451/10671/1/ulfc106455_tm_Cleidy_Cabral.pdf. Acesso em: 1 maio 2021.
CAMPOS, Raphael. Árvores de Decisão: Então diga-me, como construo uma?. [S. l.], 28 nov. 2017. Disponível em: https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69. Acesso em: 23 jan. 2021.
Colab. 2021. Disponível em: https://colab.research.google.com/. Acesso em: 20 mar. 2021.
COX, D.R.;SNELL, E.J. Analysis of Binary Data.London: Chapman & Hall, 2ª Edição, 1989. HOSMER, D.W.; LEMESHOW, S.Applied Logistic Regression. New York: John Wiley, 2ª Edição, 2000.
E.França, S.Lansky. Mortalidade infantil neonatal no brasil: situação, tendências e perspectivas Rede Interagencial de Informações para Saúde - demografia e saúde: contribuição para análise de situação e tendências Série Informe de Situação e Tendências(2009), pp.83-112.
62
FREIRE, Sergio Miranda. Bioestatística Básica: Regressão Linear. [S. l.], 20 out. 2020. Disponível em: http://www.lampada.uerj.br/arquivosdb/_book/bioestatisticaBasica.html. Acesso em: 23 jan. 2021.
Jupyter. 2021. Disponível em: https://jupyter.org/. Acesso em: 20 jun. 2021.
GOODFELLOW, I. et al. Deep learning. [S.l.]: MIT press Cambridge, 2016.
Pandas. 2021. Disponível em: https://pandas.pydata.org/. Acesso em: 20 jun. 2021.

Python. 2021. Disponível em: https://www.python.org/. Acesso em: 20 jun. 2021.
KUHN, M.; JOHNSON, K. Applied predictive modeling. [S.l.]: Springer, 2013. v. 26.
LANSKY, S. et al. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante . Cadernos de Saúde Pública, Scielo, v. 30, p. S192 ? S207, 00 2014.
LANSKY, Sônia; FRICHE, Amélia Augusta de Lima; SILVA, Antônio Augusto Moura; CAMPOS, Deise; BITTENCOURT, Sonia Duarte de Azevedo; CARVALHO, Márcia Lazaro; FRIAS, Paulo Germano; CAVALCANTE, Rejane Silva; CUNHA, Antonio José Ledo Alves. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante e ao recém-nascido. [s. l.], Ago 2014. Disponivel em: https://www.scielosp.org/article/csp/2014.v30suppl1/S192-S207/pt/
Matplotlib. 2021. Disponível em: https://matplotlib.org/. Acesso em: 20 mar. 2021.
Maranhão AGK, Vasconcelos AMN, Trindade CM, Victora CG, Rabello Neto DL, Porto D,et al. Mortalidade infantil no Brasil: tendências, componentes e causas de morte no período de 2000 a 2010 . In: Departamento de Análise de Situação de Saúde, Secretaria de Vigilância em Saúde, Ministério da Saúde, organizador. Saúde Brasil 2011: uma análise da situação de saúde e a vigilância da saúde da mulher. v. 1. Brasília: Ministério da Saúde; 2012. p. 163-82.
MESQUITA, PAULO. UM MODELO DE REGRESSÃO LOGÍSTICA PARA AVALIAÇÃO DOS PROGRAMAS DE PÕS-GRADUAÇÃO NO BRASIL. Orientador: Prof. RODRIGO TAVARES NOGUEIRA. 2014. Dissertação (Mestre em Engenharia de Produção) - Universidade Estadual do Norte Fluminense, [S. l.], 2014.
MNASSRI , Baligh. Titanic: logistic regression with python. [S. l.], 1 ago. 2020. Disponível em: https://www.kaggle.com/mnassrib/titanic-logistic-regression-with-python. Acesso em: 14 jan. 2021.
Morais, R.M.d., Costa, A.L: Uma avaliação do sistema de informações sobre mortalidade. Saúde em Debate 41, 101?117 (2017). Disponível em: https://doi.org/10.1109/SIBGRAPI.2012.38.
MOTA, Caio Augusto de Souza; BELUZO , Carlos Eduardo; TRABUCO, Lavinia Pedrosa; SOUZA , Adriano; ALVES, Luciana; CARVALHO, Tiago. DESENVOLVIMENTO DE UMA PLATAFORMA WEB PARA CIÊNCIA DE DADOS APLICADA À SAÚDE MATERNO INFANTIL , [s. l.], 28 nov. 2019.
MULTIEDRO (Brasil). Conheça as aplicações do machine learning na área da saúde. [S. l.], 28 nov. 2019. Disponível em: https://blog.multiedro.com.br/machine-learning-na-area-da-saude/#:~:text=O%20machine%263
0learning%20na%20%C3%A1rea,diagn%C3%B3sticos%20e%20tratamentos%20mais%20precisos. Acesso em: 13 jun. 2021.
MUKHIYA,S.K.;AHMED,U. Hands-On Exploratory Data Analysis with Python: Perform EDA Techniques to Understand, Summarize, and Investigate Your Data. [S.l.]: PacktPublishingLtd,2020.ISBN978-1-78953-725-3.22,32
Murray CJ, Laakso T, Shibuya K, Hill K, Lopez AD. Can we achieve Millennium Development Goal 4? New analysis of country trends and forecasts of under-5 mortality to 2015. Lancet 2007; 370:1040-54.
NumPy. 2021. Disponível em:https://numpy.org/. Acesso em: 20 jun. 2021.

Oliveira, M.M.d., de Araújo, A.S.S.C., Santiago, D.G., Oliveira, J.a.C.G.d., Carvalho, M.D., de Lyra et al, R.N.D.: Evaluation of the national information system on live births in brazil , 2006-2010. Epidemiol. Serv. Saúde 24(4), 629?640 (2015).
PELISSARI, ANA CAROLINA CHEBEL. MORTALIDADE NEONATAL DA CIDADE DE SÃO PAULO: UMA ABORDAGEM UTILIZANDO APRENDIZADO DE MÁQUINA SUPERVISIONADO. 2021. Trabalho de Conclusão de Curso (Superior) - Instituto Federal de Educação, Ciência e Tecnologia Campus Campinas, [S. l.], 2021.
REGRESSÃO logística e binary cross entropy - Aula 7. Direção: Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=3J-LBtHVsm4. Acesso em: 21 jan. 2021.
Rmd Nascimento, AJM Leite, NMGSd Almeida, Pcd Almeida, Cfd Silva Determinantes da mortalidade neonatal: estudo caso-controle em fortaleza, ceará, brasil Cad Saúde Pública, 28(2012), pp.559 - 572.
SANTOS, Hellen Geremias. Machine learning e análise preditiva: considerações metodológicas: métodos lineares para classificação. In: SANTOS, Hellen Geremias. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina. 2018. Tese (Pós-Graduação em Epidemiologia) - Faculdade de Saúde Pública da Universidade de São Paulo, [S. l.], 2018.
Scikit-learn. 2021. Disponível em: https://scikit-learn.org/stable/. Acesso em: 20 jun. 2021.
Seaborn. 2021. Disponível em: https://seaborn.pydata.org/. Acesso em: 20 jun. 2021.
SILVA, Wesley; CORSO, Jansen; WELGACZ, Hanna; PEIXE, Julinês. AVALIAÇÃO DA ESCOLHA DE UM FORNECEDOR SOB CONDIÇÃO DE RISCOS A PARTIR DO MÉTODO DE ÁRVORE DE DECISÃO. ARTIGO ? MÉTODOS QUANTITATIVOS, [s. l.], 12 ago. 2008.
WHO, W. H. O.Women and health: today?s evidence, tomorrow?s agenda. [S.l.]: UNWorld Health Organization, 2009. ISBN 9789241563857.
64
APÊNDICE A

==========

Arquivo 1: monografia-V7.docx.pdf (9728 termos)
Arquivo 2: http://www.cpdm.ufpr.br/testediaginostico.html (421 termos)
Termos comuns: 20
Similaridade: 0,19%
O texto abaixo é o conteúdo do documento monografia-V7.docx.pdf (9728 termos)
Os termos em vermelho foram encontrados no documento
http://www.cpdm.ufpr.br/testediaginostico.html (421 termos)

==========

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO
CÂMPUS CAMPINAS
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
CAMPINAS
2021
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas.
Orientador: Prof. Me. Everton Josué da Silva.
CAMPINAS
2021
Dados Internacionais de Catalogação na Publicação
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo Câmpus Campinas.
Aprovado pela banca examinadora em: _____ de _____ de _____.
BANCA EXAMINADORA

_____
Prof. Me. Everton Josué da Silva (orientador)

IFSP Câmpus Campinas

_____

Prof. Me. Carlos Eduardo Beluzo
IFSP Câmpus Campinas

_____

Prof. Dr. Ricardo Barz Sovat
IFSP Câmpus Campinas

Dedico este trabalho aos meus familiares,
colegas de classe, professores e servidores do Instituto
que colaboraram em minha jornada formativa.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pelo dom da vida e pela oportunidade de concluir mais uma etapa de minha experiência acadêmica.

Agradeço a todos os professores e servidores do IFSP Câmpus Campinas, que contribuíram direta e indiretamente para a conclusão deste trabalho.

Agradeço também à minha família, que deu todo o apoio necessário para que eu chegasse até aqui.

Agradeço ao meu orientador que me auxiliou a solucionar as dificuldades encontradas no caminho.

"Minha energia é o desafio, minha motivação é o impossível,
e é por isso que eu preciso
ser, à força e a esmo, inabalável.".
Augusto Branco

RESUMO

A mortalidade infantil pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível em uma população. O aprendizado de máquina é uma área da inteligência artificial que propõe métodos de análise de dados que automatizam a construção de modelos analíticos com base em reconhecimento de padrões, baseado no conceito de que sistemas podem aprender com dados, identificando padrões e tomando decisões com o mínimo de intervenção humana. O objetivo deste trabalho é testar e analisar dois tipos diferentes de algoritmos de aprendizado de máquina, utilizando a base de dados do SIM e SINASC do Brasil, do período de 2016 até 2018, para gerar modelos para predição de mortalidade neonatal. Os métodos utilizados foram Árvore de Decisão e de Regressão Logística e como métricas para avaliar os modelos resultantes foram utilizadas a AUC, Curva ROC e a Matriz de Confusão. Como resultado, apesar de terem alcançado valores de AUC 0.93, ambos modelos de predições acertaram muitas predições de vivos cerca de 671.000 e erraram muitas predições de mortos cerca de 2.600, principalmente devido ao fato de a base estar desbalanceada.

Palavras-chave: Mortalidade Neonatal, Predição, Aprendizado de Máquina.

ABSTRACT

Infant mortality can be used to analyze poverty and socioeconomic levels, as well as measure the quality of health and medical technology available in a population. Machine learning is an area of artificial intelligence that proposes data analysis methods that automate the

construction of analytical models based on pattern recognition, based on the concept that systems can learn from data, identifying patterns and making decisions with a minimum of human intervention. The objective of this work is to test and analyze two different types of machine learning algorithms, using the SIM and SINASC do Brasil database, from 2016 to 2018, to generate models for predicting neonatal mortality. The methods used were Decision Tree and Logistic Regression and as metrics to evaluate the resulting models, AUC, ROC Curve and Confusion Matrix were used. As a result, despite achieving values of AUC 0.93, both prediction models correct many live birth predictions around 671.000 and miss many stillbirth predictions around 2600, mainly due to the fact that the base is unbalanced.
Keywords: Neonatal Mortality, Prediction, Machine Learning.
LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS
SIM Sistema de Informação sobre Mortalidade
SINASC Sistema de Informações sobre Nascidos Vivos
TMI Taxa de Mortalidade Infantil
TMN Taxa de Mortalidade Neonatal
MN Mortalidade Neonatal
ML Machine Learning
DNV Declaração de Nascido Vivo
VN Verdadeiro Negativo
FN Falso Negativo
VP Verdadeiro Positivo
FP Falso Positivo
ROC Curva Característica de Operação do Receptor
NaN Não é um Número
RFE Eliminação Recursiva de Feature
RFECV Eliminação Recursiva de Feature com Validação Cruzada
SUMÁRIO

13
1 INTRODUÇÃO

A taxa de mortalidade infantil (TMI) é uma importante medida de saúde em uma população e é um grande problema no mundo todo. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível. Esta taxa é dividida em duas categorias: neonatal e pós-neonatal. É categorizada neonatal quando o óbito ocorre nos 28 primeiros dias de vida após o pós-parto, e o pós-neonatal é quando o óbito ocorre entre 29 e 364 dias de vida (BELUZO et al., 2020). Cerca de 46% das mortes no mundo acontecem entre os menores de cinco anos e grande parte está concentrada nos primeiros dias de vida (LANSKY, 2014). A diminuição dessas taxas é muito importante no mundo todo, refletindo nos indicadores de saúde pública e desenvolvimento do país.

Os fatores associados à mortalidade são profundamente influenciados pelas características biológicas maternas e neonatais, pelas condições sociais e pelos cuidados prestados pelos serviços de saúde (E. FRANÇA; S. LANSKY, 2009; R.M.D. NASCIMENTO, 2012). Em grande parte, o diagnóstico é altamente dependente da experiência adquirida pelo

profissional que o realiza. Apesar de indiscutivelmente necessário, não é perfeito, principalmente pelo fato de ser dependente de fatores humanos, por este motivo os profissionais sempre tiveram recursos tecnológicos para auxiliar nessa tarefa (BELUZO et al., 2020). Segundo estudos de 2010 no Brasil, calcula-se que aproximadamente 70% dos óbitos infantis ocorridos poderiam ter sido evitados por ações de saúde e que 60% dos óbitos neonatais ocorreram por situações que poderiam ser evitadas (BARRETO; SOUZA; CHAPMAN, 2015).

No presente trabalho foram utilizados dois algoritmos de aprendizado de máquina para criar modelos de predição de mortalidade neonatal. Além disso, foi realizada também uma análise exploratória da base de dados. Para este trabalho foram utilizadas duas fontes de dados: Sistema de Informação sobre Mortalidade (SIM) e Sistema de Informações sobre Nascidos Vivos (SINASC) do Brasil inteiro.

14

2 JUSTIFICATIVA

Com o avanço da tecnologia, podemos notar que ela está cada vez mais presente no nosso dia a dia e nos auxilia em diversas tarefas do cotidiano, e vem sendo aplicada em diversas áreas, dentre elas a saúde.

A mortalidade infantil é uma preocupação mundial na saúde pública, a ONU (Organizações das Nações Unidas) definiu a redução da mortalidade infantil como uma meta para o desenvolvimento global. A mortalidade neonatal é responsável por aproximadamente 60% da TMI nos países em desenvolvimento (A.K. SINGHA, 2016). Existem diversos fatores que podem contribuir com a redução de óbitos neonatais, como por exemplo acompanhamento médico à gestante no período de gravidez, acompanhamento médico ao recém-nascido nos primeiros dias de vida e a disponibilização deste serviço com qualidade e proficiência (WHO, 2009).

A diminuição dessa taxa é importante e de interesse para o mundo todo, e utilizar da tecnologia para auxiliar nessa diminuição é uma proposta funcional, e inovadora para a realidade brasileira (MOTA et al., 2019). Havendo disponibilidade de tecnologia para auxiliar nas decisões dos diagnósticos, os profissionais da saúde podem focar nos cuidados do recém-nascido com risco de vir a óbito, fornecendo tratamentos melhores.

Segundo o site Multiedro (2019), um grande problema que os médicos enfrentam ao realizar o diagnóstico, é analisar um grande volume de dados. Para a área médica obter diagnósticos rápidos e precisos pode salvar vidas, e a utilização de machine learning para realizar esses processos é importante, com a ML os dados podem ser processados em questões de segundos e resultar em um diagnóstico mais rápido, além disso utilizando bons modelos ML os diagnósticos serão mais precisos.

Diversos trabalhos vêm sendo desenvolvidos neste contexto. No trabalho realizado por BELUZO et al. (2020a), os autores utilizaram uma base de dados com informações da cidade de São Paulo para criação de modelos de aprendizado de máquina para predição de mortalidade neonatal. Este recorte foi utilizado também no presente trabalho para fins de exercício e definição de etapas da implementação de código. Na conclusão os autores discutem os fatores que tiveram mais influência nas predições, e na análise feita pelos autores, as consultas de pré-natal são muito importantes para a redução da mortalidade infantil, uma vez que algumas características muito importantes, como a malformação congênita, podem

ser detectadas ao longo das consultas de pré-natal.
Em um outro estudo realizado por BELUZO et al. (2020a), foi realizada uma análise exploratória da base de dados SPNeoDeath, onde e podemos observar detalhadamente cada
15
análise das colunas da tabela e diversos gráficos feito para um melhor entendimento dos dados.
Outro trabalho que foi realizado utilizando a base de dados com dados somente de São Paulo foi o de PELISSARI (2021), nesse trabalho é realizado uma análise exploratória na base de dados de São Paulo e é também analisado três algoritmos de aprendizado de máquina (Logistic Regression, Random Forest Classifier e XGBoost), nesse trabalho a autora conclui que os modelos de Logistic Regression e XGBoost foram os modelos que apresentaram os melhores desempenhos preditivos, e também mostra os problemas de estar utilizando uma base de dados desbalanceada.
O problema da mortalidade neonatal não é recente, e o mundo todo desenvolve iniciativas para lidar com ele. A ONU definiu como meta a redução da mortalidade infantil para o desenvolvimento global. Diversos fatores podem ajudar na diminuição da taxa de mortalidade neonatal como acompanhamento médico antes e após parto, tanto com a mãe quanto com o recém-nascido. Um serviço de qualidade e constante para os casos com maior risco de óbito pode salvar diversos recém-nascidos. Neste sentido, o desenvolvimento de ferramentas para a predição de mortalidade neonatal pode colaborar com esta iniciativa.

3 OBJETIVOS

3.1 OBJETIVO GERAL

O presente trabalho tem como objetivo testar e analisar os resultados da aplicação dos aprendizagem de máquina supervisionado Árvore de Decisão e Regressão Logística, utilizando dados do SIM e do SINASC, no período de 2016 até 2018, a fim de gerar modelos de predição de risco morte neonatal.

3.2 OBJETIVOS ESPECÍFICOS

? Realizar análise exploratória da base dados a ser utilizada na construção dos modelos;
? Implementar modelos de predição utilizando algoritmos Árvore de Decisão e Regressão Logística;
? Avaliar resultados e propor novas abordagens.
16

4 FUNDAMENTAÇÃO TEÓRICA

4.1 MORTALIDADE INFANTIL E NEONATAL

Como dito na monografia PELISSARI (2021), a TMI consiste no número de crianças que foram a óbito antes de concluir um ano de vida dividido por 1000 crianças nascidas vivas no tempo de um ano. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos e qualidade da saúde pública de um país (apud BELUZO et al., 2020).
Mencionado em PELISSARI (2021), a TMN é uma das duas categorias da TMI, é categorizado mortalidade neonatal quando o óbito ocorre do nascimento da criança até os 28 primeiros dias de vida. A mortalidade neonatal pode ser subdividida em mortalidade neonatal precoce que é quando o óbito ocorre entre 0 e 6 dias de vida, e o neonatal tardio quando o óbito ocorre entre 7 e 28 dias de vida (apud E. FRANÇA e S. LANSKY, 2009).
Segundo Lansky et al. (2014), a taxa de mortalidade neonatal é o principal

componente da TMI desde a década de 1990 no Brasil e vem se mantendo em níveis elevados, com taxa de 11,2 óbitos por mil nascidos vivos em 2010 (apud Maranhão et al., 2012). Também é dito em Lansky et al. (2014), que a TMI do Brasil em 2011 foi 15,3 por mil nascidos vivos, alcançando a meta 4 dos objetivos de desenvolvimento do milênio, compromisso dos governos integrantes das Nações Unidas de melhorar a saúde infantil e reduzir em 2/3 a mortalidade infantil entre 1990 e 2015. (apud Maranhão et al., 2012; apud Murray et al., 2015). Segundo Lansky et al. (2014) o principal componente da TMI em 2009 é a mortalidade neonatal precoce, e grande parte das mortes infantis acontece nas primeiras 24 horas (25%), indicando uma relação estreita com a atenção ao parto e nascimento (apud E. FRANÇA e S. LANSKY, 2009).

4.2 MÉTODOS DE APRENDIZADO DE MÁQUINA

A utilização de dados para a resolução de problemas, de modo a se identificar padrões ocultos e posteriormente auxiliar na tomada de decisões é definida como aprendizado de máquina (BRESAN, 2018 apud Brink et al. 2013).

O uso de técnicas de Aprendizado de Máquina se mostra altamente eficiente na resolução de tarefas que se apresentem difíceis de se resolver a partir de programas escritos por humanos (BRESAN, 2018 apud GOODFELLOW et al., 2016), bem como também possibilita uma maior compreensão sobre os funcionamentos dos princípios que compõem a inteligência humana.

17

Neste trabalho serão implementados modelos utilizando os algoritmos de Árvore de Decisão e Regressão Logística, os quais serão apresentados a seguir.

4.2.1 Árvore de Decisão

Segundo Batista e Filho (2019), os modelos preditivos baseados em árvores são geralmente utilizados para tarefas de classificação, embora também possam ser utilizados para tarefas de regressão. Considerado um dos mais populares algoritmos de predição, o algoritmo de árvore de decisão proporciona uma grande facilidade de interpretação.

As árvores de decisão utilizam a estratégia ?dividir-e-conquistar", na qual as árvores são construídas utilizando-se apenas alguns atributos. A árvores de decisão é uma das técnicas por meio da qual um problema complexo é decomposto em subproblemas mais simples. Recursivamente, a mesma estratégia é aplicada a cada subproblema (SILVA et al, 2008).

A árvore de decisões tem uma estrutura hierárquica onde cada nó da árvore representa uma decisão em uma das colunas do conjunto de dados, cada ramo da árvore representa uma tomada de decisão (BATISTA; FILHO, 2019).

O algoritmo segue algumas decisões para montar a árvore, o atributo mais importante é utilizado como nó raiz e será o primeiro nó, já os atributos menos importantes são mostrados nos ramos da árvore. Cada atributo é verificado para conferir se é possível gerar uma árvore menor e se é possível fazer uma classificação melhor, e assim é escolhido o nó que produz nós filhos (SILVA et al, 2008).

Existem diferentes tipos de algoritmo para a construção da árvore, como por exemplo, ID3 (Iterative Dichotomiser), C4.5 e CART (Classification and Regression Tree). O algoritmo ID3 escolhe os nós da árvore por meio da métrica do ganho de informação. Já o CART faz uso da equação de Gini (BATISTA; FILHO, 2019 apud KUHN; JOHNSON, 2013).

O algoritmo utilizado neste trabalho usará o cálculo da entropia e ganho de

informação, para decidir qual atributo será usado no nó, a entropia é uma medida da desorganização dos sistemas, maior é a incerteza do sistema, quanto maior a entropia maior está a desorganização, com a árvore de decisão a ideia é ir organizando o sistema e ir separando as decisões da melhor forma para que a entropia diminui e o ganho de informação aumente, para que a decisão do modelo fiquei mais assertiva. O cálculo da entropia utiliza a seguinte equação (ÁRVORE, 2020):

???????? =

?

?? ?? ???

2

??

18

Nesta equação o i representa possíveis classificações (rótulos), e o P é a probabilidade de cada uma. A ideia da árvore então é verificar qual é a entropia para cada uma das variáveis da base de dados, e verificar qual é a melhor organização de cada uma das variáveis. E isso é realizado através da técnica chamada de ganho de informação, essa técnica utiliza a equação a seguir (ÁRVORE, 2020):

????? = ????????(???) ? ? ????(??????) * ????????(??????)

Então o ganho de informação é calculado pela entropia do nó pai menos a soma do peso vezes a entropia dos nós filhos. Esse cálculo é realizado para todas as variáveis e a variável que tiver maior ganho de informação é utilizado para a decisão do nó. Esse processo é realizado recursivamente e a árvore vai sendo montada com base nesses cálculos (ÁRVORE, 2020).

Na Figura 1 temos a representação do início de uma árvore de decisão, no caso em questão o autor estava classificando exames positivos e negativos para o vírus SARS-CoV-2, ele utilizou uma base de dados que contém quase 6 mil exames realizados, contendo campos como o resultado do exame, idade do paciente, níveis de hemoglobina entre outras informações.

Figura 1. Exemplo de Árvore de Decisão. (Fonte: ÁRVORE 2020).

Pode-se observar que para o nó raiz foi utilizado a variável Leukocytes, logo foi o que apresentou a melhor organização para ser o nó raiz, e após ele temos os ramos da árvore. Cada retângulo da árvore é uma decisão do modelo e esses retângulos têm atributos importante como, a coluna da base de dados que será onde vai ser realizado a decisão, então pegando o nó raiz, pacientes que tiverem com os Leukocytes para menos que -0.43 seguiram o caminho

19

para a esquerda (true), já os pacientes que tiverem mais seguiram o caminho da direita (false). A entropia calculada do nó também é apresentada e a classificação do nó também, então se o resultado parar neste nó, a classificação que está nele será a classificação final, e temos o samples que é o número de amostras que se enquadram no na decisão. E esse processo de verificação do modelo vai se repetindo até não ter mais como dividir em subproblemas ou até chegar na profundidade máxima.

Uma característica importante da árvore de decisão é que ela é um modelo WhiteBox, ou seja, é possível visualizar a árvore montada e as decisões que o modelo terá que tomar na árvore. Na Figura 1 pode-se ver uma árvore de decisões montada, cada retângulo da árvore é

uma decisão que o modelo terá que tomar, conforme o modelo toma as decisões ele vai seguindo um caminho até chegar ao nó folha, nó folha é o nó que não contém nó filho, não tem mais ramos para baixo, chegando ao nó folha o modelo tem a classificação final.

4.2.2 Regressão Logística

O modelo de regressão logística estabelece uma relação entre a probabilidade de ocorrência da variável de interesse e as variáveis de entrado do modelo, sendo utilizada para tarefas de classificação, em que a variável de interesse é categórica, neste caso, a variável de interesse assume valores 0 ou 1, onde 1 é geralmente utilizado para indicar a ocorrência do evento de interesse (Batista e Filho, 2019).

De acordo com Mesquita (2014), a técnica de regressão logística, desenvolvida no século XIX, obteve maior visibilidade após 1950 ficando então mais conhecida. Ficou ainda mais difundida a partir dos trabalhos de Cox & Snell (1989) e Hosmer & Lemeshow (2000). Caracteriza-se por descrever a relação entre uma variável dependente qualitativa binária, associada a um conjunto de variáveis independentes qualitativas ou métricas.

Esta técnica inicialmente foi utilizada na área médica, porém a eficiência viabilizou sua implementação nas mais diversas áreas. Mesquita (2014) diz que o termo regressão logístico, tem sua origem na transformação usada com a variável dependente, que permite calcular diretamente a probabilidade da ocorrência do fenômeno em estudo.

Segundo Santos (2018), para estimar a probabilidade, é utilizado uma técnica chamada distribuição binomial para modelar a variável resposta do conjunto de treinamento, que tem como parâmetro a probabilidade de ocorrência de uma classe específica. Uma característica importante desse modelo é que a probabilidade estimada deve estar limitada ao intervalo de 0 a 1. O algoritmo de Regressão Logística gera um modelo de classificação binária (0 e 1). O modelo utiliza a Regressão Linear como base, a equação linear é (REGRESSÃO, 2020):

20

? = ? . ? + ?

O ?m? da equação é o coeficiente angular da reta, e o ?b? é o intercepto. Então, aplicando a equação linear obtemos um valor contínuo como resultado, após obter esse valor a Regressão Logística tem que realizar a classificação do resultado, então é aplicado uma função chamada sigmoide (REGRESSÃO, 2020):

? = 1

1+ ? ??

Essa função chamada de sigmoide, ou função logística, é responsável por ?achatar? o resultado da regressão linear, calculando a probabilidade de o resultado pertencer a classe 1, classificando assim o resultado da função linear em 0 ou 1.

Segundo Batista e Filho (2019), o modelo de regressão logística que tem como objetivo realizar uma classificação binária de uma variável de interesse irá predizer a probabilidade de pertencer à classe positiva. Tem-se, portanto, que é dado por:?

? = { 0, ?? ???? (? = 1 ??) < 0, 5

ou

? = { 1, ?? ???? (? = 1??) ? 0, 5

Então a decisão do modelo de regressão logística está relacionada à escolha de um ponto de corte para p(x). Caso o ponto de corte escolhido for p(x)=0,5, os resultados que forem p(x)>0,5 serão classificados como 1 e os resultados p(x)<5 serão classificados como 0

(SANTOS, 2018).
A Figura 2 ilustra um exemplo de como é o diagrama de Regressão Logística. Pelo diagrama então pode-se observar que o modelo possui uma representação gráfica em formato de ?S?, essa seria a curva logística. As predições do modelo sempre ficaram na linha 1 e 0 do eixo y, pois é o intervalo estabelecido pelo algoritmo, então a classificação sempre será nesse intervalo.
21
Figura 2. Exemplo de Diagrama de Regressão Linear. (Fonte: Batista e Filho 2019).
4.2.3 Treino, teste e validação do modelo de aprendizado de máquina
A criação de um modelo de aprendizado de máquina é realizada em duas etapas: treinamento do modelo, que é realizado a partir dos dados fornecidos para o algoritmo, e etapa de teste, onde os modelos recebem dados novos e sem o rótulo, para que seja avaliado a performance do modelo (PELISSARI, 2021).
Para o treinamento dos modelos foram utilizadas duas abordagens, a primeira onde a base foi particionada entre conjunto de dados para teste e para treino em uma porcentagem de 90% para treino e 10% para teste, porém dessa forma pode ocorrer problema de sobreajuste ou overfitting. Nesta situação, o que pode ocorrer é o modelo não aprender com os dados, e apenas ?decorar? os dados recebidos e quando é realizado o teste o modelo consegue predizer apenas situações parecidas, não sendo capaz de identificar padrões com diferenças mínimas. Para evitar este problema, foi utilizado uma técnica de validação cruzada chamada K-folds cross-validation. Essa técnica de validação cruzada divide a base de dados em K subconjuntos, e então são realizadas várias rodadas de treinamento e teste alternando os subconjuntos utilizados para teste e treino, e o resultado do modelo baseia-se na média da acurácia observada em cada rodada (PELISSARI, 2021).
Pode-se observar na Tabela 1 uma ilustração da técnica K-fold cross-validation, onde temos um exemplo onde a base é dividida em 5 subconjuntos e cada subconjunto tem a parte de teste diferente dos outros subconjuntos, assim o modelo vai receber valores novos de teste e treino todas as vezes que executar um novo subconjunto.
22
Tabela 1. Exemplo da separação K-fold cross-validation. (Fonte: Adaptado de PELISSARI, 2021).

Predição 1	Predição 2	Predição 3	Predição 4	Predição 5
Teste	Treino	Treino	Treino	Treino
Treino	Teste	Treino	Treino	Treino
Treino	Treino	Teste	Treino	Treino
Treino	Treino	Treino	Teste	Treino
Treino	Treino	Treino	Treino	Teste

4.2.4 Métricas para avaliação de modelos
Para fins de avaliação, neste trabalho foram utilizadas três métricas para avaliar o desempenho dos modelos: a Matriz de Confusão e a AUC (Area under Receiver Operating Characteristic Curve - ROC). Na matriz de confusão pode-se observar o comportamento do modelo em relação a cada uma das classes a serem preditas pelo modelo, utilizando variáveis binárias (positivo: 1; e negativo: 0), para apresentar uma matriz que faz uma comparação entre as predições do modelo e os valores corretos do conjunto de dados (PELISSARI, 2021).

Na Tabela 2 temos um exemplo de uma matriz de confusão, a qual consiste em duas colunas e duas linhas, e os valores são atribuídos nos quadrantes com os seguintes resultados (PELISSARI, 2021):

? quando o valor real do conjunto de dados é 0, e a predição do modelo também classificou como 0, então temos um Verdadeiro Negativo (VN);

? quando o valor real é 1, e o modelo classificado como 0, temos um Falso Negativo (FN);

? quando o valor real é 0, e o modelo classificado como 1, então o resultado se enquadra no quadrante de Falso Positivo (FP);

? e por fim o quadrante Verdadeiro Positivo (VP), que são os resultados que tem o valor real 1, e o modelo classificado como 1.

23

Tabela 2. Modelo de Matriz de confusão. (Fonte: Adaptado de PELISSARI, 2021).

Valor Predito
Negativo (0) Positivo (1)
Classe
Real
Negativo
(0)
Verdadeiro
Negativo
Falso
Positivo
Positivo
(1)
Falso
Negativo
Verdadeiro
Positivo

A segunda métrica de avaliação utilizada foi a Curva ROC que permite avaliar o desempenho de um modelo com relação às predições efetuadas. A curva ROC é um gráfico de Sensibilidade (taxa de verdadeiros positivos) versus taxa de falsos positivos, a representação da curva ROC permite evidenciar os valores para os quais existe otimização da Sensibilidade em função da Especificidade (CABRAL, 2013).

Na Figura 3 temos um exemplo de uma curva ROC. A linha traçada na cor preta é uma linha de referência, ela representa hipoteticamente a curva ROC de um classificador puramente aleatório. Caso a linha laranja, que representa o resultado do modelo que está sendo avaliado, ficasse igual a linha tracejada preta, indicaria que o modelo avaliado estaria predizendo aleatoriamente os resultados.

Figura 3: Exemplo de Curva ROC. (Fonte: Elaboração Própria).

Por uma análise de inspeção visual, quanto mais próximo a linha laranja estiver do valor 1 no eixo Y, melhor está a capacidade do modelo para diferenciar as classes. O valor da

24

AUC representa a capacidade do modelo de diferenciar as classes, quanto maior o valor do

AUC, melhor é a predição realizada pelo modelo, uma vez que, os valores resultantes iguais a 0 são realmente 0, e os iguais a 1 são de fato 1 (PELISSARI, 2021).

5 METODOLOGIA

O desenvolvimento deste trabalho foi realizado em três etapas: (1) Pré-processamento e Análise Exploratória do conjunto de dados, onde foram aplicadas técnicas comuns para preparar o conjunto de dados como tratamento de valores nulos e seleção de variáveis de interesse; (2) Implementação de modelos de aprendizado de máquina supervisionado para predição das mortes neonatais, utilizando os algoritmos de árvore de decisão e regressão logística; e (3) Análise de Resultados, que será realizada uma análise do desempenho do modelo.

5.1 TECNOLOGIAS E FERRAMENTAS

Neste trabalho vamos utilizar a linguagem de programação Python (PYTHON, 2021) pela sua simplicidade de codificação e alto poder de processamento. Foi utilizado também a plataforma Google Colab Research (Colab, 2021), para o desenvolvimento dos algoritmos e treinamento dos modelos. Além disso, foi utilizado também o Jupyter Notebook (Jupyter, 2021), instalado localmente em computador pessoal, pois o Google Colab Research possui limitação de recursos de memória e processamento, então as duas plataformas foram usadas para realização deste trabalho. As principais bibliotecas utilizadas foram: numpy (NumPy, 2021), e pandas (Pandas, 2021), para a manusear os dados, matplotlib (Matplotlib, 2021), e seaborn (Seaborn, 2021), para a plotagem dos gráficos, e por fim a biblioteca sklearn (Scikit-Learn, 2021), que é a biblioteca responsável pelos algoritmos de aprendizado de máquina.

5.2 BASE DE DADOS

As bases de dados a serem utilizadas serão o SIM e SINASC, que são as duas principais fontes de informações sobre nascimentos e óbitos no Brasil. O SINASC é alimentado com base na Declaração de Nascido Vivo (Declaração de Nascido Vivo - DNV) (BELUZO et al.,2020b apud Oliveira et al., 2015). O SIM tem como objetivo principal apoiar a coleta, armazenamento e processo de gestão de registros de óbitos no Brasil (BELUZO et

25

al.,2020b apud Morais et al., 2017), e foi usado para rotular o óbito registrado no SIM, utilizando o campo DNV como chave de associação. O SIM será utilizado para rotular os registros do SINASC, pois o SIM coletas informações sobre mortalidade e é usado como base para o cálculo de estatísticas vitais, como a taxa de mortalidade neonatal e o SINASC reúne informações sobre dados demográficos e epidemiológicos do bebê, da mãe, do pré-natal e do parto (BELUZO et al.,2020b).

A Tabela 3 apresenta as variáveis da base de dados. Essa base tem variáveis que contêm valores quantitativos e categóricos. A coluna ?num_live_births? por exemplo contém o número de nascidos vivos anteriores da mãe e é composta por valores quantitativos contínuos, e a coluna ?tp_pregnancy? utiliza valores nominais categóricos que indicam o tipo de gravidez.

Tabela 3. Colunas da base de dados e suas descrições. (Fonte: Adaptado de BELUZO et al.,2020b).

Nome da coluna	Descrição	Domínio de dados
Variáveis	demográficas e socioeconômicas	

maternal_age Idade da mãe Quantitativo Contínuo (inteiro)
tp_maternal_race Raça / cor da pele da
mãe
Nominal categórico (inteiro)
1 - branco; 2 - Preto; 3 -
amarelo; 4 - Pele morena; 5 -
Indígena.
tp_marital_status Estado civil da mãe Nominal categórico (inteiro)
1 - Único; 2 - Casado; 3 - Viúva;
4 - Separados / divorciados
judicialmente; 5 - Casamento
por união estável; 9 - Ignorado.
tp_maternal_schooling Anos de escolaridade da
mãe
Nominal categórico (inteiro)
1 - nenhum; 2 - de 1 a 3 anos; 3 -
de 4 a 7 anos; 4 - de 8 a 11 anos;
5 - 12 e mais; 9 - Ignorado.
Variáveis obstétricas maternas
num_live_births Número de nascidos
vivos
Quantitativo Contínuo (inteiro)
num_fetal_losses Número de perdas fetais Quantitativo Contínuo (inteiro)
num_previous_gestations Número de gestações Quantitativo Contínuo (inteiro)
26
anteriores
num_normal_labors Número de partos
normais (trabalhos de
parto)
Quantitativo Contínuo (inteiro)
num_cesarean_labors Número de partos
cesáreos (partos)
Quantitativo Contínuo (inteiro)
tp_pregnancy Tipo de gravidez Nominal categórico (inteiro)
1 - Singleton; 2 - Twin; 3 -
Trigêmeo ou mais; 9 - Ignorado.
Variáveis relacionadas a cuidados anteriores
tp_labor Tipo de parto infantil
(tipo de parto)
Categórico Nominal (inteiro)
1 - Vaginal; 2 - Cesariana;
num_prenatal_appointments Número de consultas de
pré-natal
Quantitativo Contínuo (inteiro)

tp_robson_group Classificação do grupo
Robson
Ordinal categórico (inteiro)
Variáveis relacionadas ao recém-nascido
tp_newborn_presentation Tipo de apresentação de
recém-nascido
Categórico Nominal (inteiro)
1 - Cefálico; 2 - Pélvico ou
culatra; 3 - Transversal; 9 -
Ignorado.
has_congenital_malformation Presença de
malformação congênita
Nominal categórico (inteiro)
1 - Sim; 2 - Não; 9 - Ignorado
newborn_weight Peso ao nascer em
gramas
Quantitativo Contínuo (inteiro)
cd_apgar1 Pontuação de Apgar de 1
minuto
Ordinal categórico (inteiro)
cd_apgar5 Pontuação de Apgar de 5
minuto
Ordinal categórico (inteiro)
gestaional_week Semana de gestação Quantitativo Contínuo (inteiro)
tp_childbirth_care Assistência ao parto Categórico Nominal (inteiro)
1 - Doutor; 2 - enfermeira ou
obstetra; 3 - Parteira; 4 - outros;
9 - Ignorado.
was_cesarean_before_labor Foi cesáreo antes do
parto.
27
was_labor_induced Foi induzido ao trabalho
de parto.
is_neonatal_death Morte antes de 28 dias
(rótulo)
Nominal categórico (inteiro)
0 - sobrevivente; 1 - morto.

Neste trabalho foi utilizado dados do período de 2014 à 2016 devido sua melhor qualidade. Este recorte possui 6.760.222 registros dos quais 99.4% são amostras de recém-nascidos vivos e 0.6% são amostras onde os recém-nascidos vieram a óbito conforme pode ser observado na Figura 4. Com essas informações consegue-se observar que a base de dados é desbalanceada.

Figura 4. Distribuição da base de dados considerando se o recém-nascido viveu ou morreu durante o período neonatal. (Fonte: Elaboração Própria).

5.3 PREPARAÇÃO DA BASE DE DADOS

Nessa etapa foi realizada a preparação da base de dados para o treinamento dos modelos que consiste na limpeza, transformação e preparação dos dados a serem utilizados na análise exploratória e na criação dos modelos preditivos. A base de dados pode conter informações desnecessárias para o modelo que pode acabar atrapalhando as predições,
28
algumas colunas podem ser retiradas ou seus valores podem ser modificados, por exemplo valores reais são modificados para inteiros.

5.4 DESENVOLVIMENTO DOS MODELOS DE APRENDIZADO DE MÁQUINA

Como já dito anteriormente na seção 4 os algoritmos de aprendizado de máquina escolhidos para esse trabalho foram os: Regressão Logística e Árvore de Decisão que utilizam a lógica mostrada na seção 4 , Fundamentação Teórica. Esses dois foram escolhidos por alguns motivos, ambos são algoritmos de classificação e supervisionados, atendendo os critérios necessários para a predição de mortalidade neonatal, ambos os algoritmos são bons para realizar predições, a Árvore de Decisão inclusive é um ótimo modelo para analisar as decisões que estão sendo feitas pelo modelo, afinal esse algoritmo tem a característica de ser um WhiteBox, ou seja conseguimos ver o que o modelo aprendeu e o que ele está decidindo. Esses algoritmos selecionados são algoritmos que se encaixam na necessidade deste trabalho e são muito utilizados na comunidade acadêmica.

Para o algoritmo de Regressão Logística, como foi dito anteriormente na seção ?Preparação da base de dados? a base de dados foi tratada e particionada, no particionamento foi utilizado 90% da base para o treino e 10% para os testes e então foi realizado o treinamento do modelo, também foi aplicada a função RFECV (Eliminação Recursiva de Feature com Validação Cruzada), esse função executa a RFE (Eliminação Recursiva de Feature), que tem como ideia selecionar features considerando recursivamente conjuntos cada vez menores de features, isso ocorre da seguinte forma. Primeiro, o estimador é treinado no conjunto inicial de features e a importância de cada feature é obtida por meio de um atributo "coef" ou por meio de um atributo "feature_importances". Em seguida, as features menos importantes são removidas do conjunto atual de features. Esse procedimento é repetido recursivamente no conjunto removido até que o número desejado de features a serem selecionados seja finalmente alcançado. Porém o diferencial da RFECV é que essa função executa a RFE em um loop de validação cruzada para encontrar o número ideal ou o melhor número de features.

Após essa seleção, foi realizado o treinamento do modelo usando todas as features e também treinamos usando as features que o RFECV selecionou, para compararmos os resultados no final. Também foi utilizado o método de K-folds cross-validation em um novo treinamento para conferirmos se o modelo estava sofrendo overfitting (sobreajuste). E por fim aplicamos um método chamado GridSearchCV, esse método permite que seja realizado testes
29
alterando algumas combinação de parâmetros nos nossos modelos, facilitando para achar a melhor combinação, esse método é parecido com o cross validation, o diferencial do Grid Search é que ele faz os cross validation de vários modelos com hiperparâmetros diferentes de uma só vez.

Para o algoritmo de Árvore de Decisão também foi utilizado o mesmo

particionamento de 90% para treino e 10% para teste, na criação do modelo foi colocado a entropy como critério de decisão e uma profundidade máxima de 4, pois com números maiores o modelo ficava muito complexo e ele errava mais as predições, também utilizamos um algoritmo para mostrar a importância de cada Feature para o modelo.
O algoritmo de Regressão Logística e análise exploratória foi baseado em códigos disponível na plataforma web Kaggle (MNASSRI, 2020), no exemplo do Kaggle o autor faz os códigos para predizer mortes no navio Titanic, para o trabalho de mortalidade neonatal os códigos foram alterados para realizar predição de mortes neonatais. Já no algoritmo de Árvore de Decisão foi baseado em uma vídeo aula do professor Diogo Cortiz (ÁRVORE..., 2020), nessa vídeo aula o professor explica sobre os algoritmos de Árvore de Decisão e depois implementa o um exemplo de código, o código usado neste trabalho usou o codigo do professor Diogo Cortiz, e foi alterado para realizar predições de mortalidade neonatal.

5.5 ANÁLISE DOS RESULTADOS

Para a análise de resultados foi utilizado dois métodos que é o de Matriz de confusão e a curva ROC esses métodos são explicados na seção 4, a Fundamentação Teórica. Com esses métodos pode-se realizar uma avaliação do modelo, e assim será possível analisar os valores de acurácia, precisão e recall do modelo, essas métricas são utilizadas avaliar o desempenho do modelo, com a acurácia é possível calcular a proximidade entre o valor obtido na predição dos modelos e os valores esperados, com a precisão é possível analisar as amostras que o modelo classificou como 0 eram realmente 0 na classe real e com o recall é possível analisar as amostras que o modelo conseguiu reconhecer como a classe correta.

5.6 ACESSO A BASE DE DADOS E CÓDIGOS

A base de dados utilizada neste projeto pode ser encontrada no link: ... . E os códigos podem ser encontrados no link: ... .

30

6 RESULTADOS

Os mesmos experimentos mostrados abaixo foram aplicados para uma base de dados que contém somente dados de São Paulo, essa base é um recorte da base do Brasil todo em contém cerca de 1.400.000 amostras, esse recorte da base também é bem desbalanceado. Os resultados obtidos nesse experimento usando o recorte de São Paulo foram bem parecidos com os experimentos da base do Brasil todo, e os códigos desse experimento podem ser visualizados no apêndice X.

6.1 ANÁLISE EXPLORATÓRIA

O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

6.1.1 Apresentação dos Dados

Como dito anteriormente na seção 5.2 ?Base de Dados? deste trabalho, a base de dados é formada pelas informações do SIM e do SINASC contém 6.760.222 amostras e 29 colunas, porém serão utilizadas somente 23 colunas sendo elas as colunas descritas na Tabela 3. Na seção 5.2 também tem a Figura 4 que mostra que a base dados é desbalanceada tendo 99.4% das amostras de recém-nascidos vivos e 0.6% das amostras onde os recém-nascidos vieram a óbito no período neonatal.

6.1.1.1 Caracteristicas demográficas e socioeconômicas maternas

O apêndice X contém uma tabela que possui as estatísticas sumarizadas das variáveis

demográficas e socioeconômicas maternas. Nesta tabela por exemplo podemos observar a coluna ?maternal_age? que contém a idade da mãe, e na tabela mostra que a média da idade das mães é de 26.4 anos, e os dados possuem uma variação de no mínimo 8 anos e no máximo 55 anos, e a mediana é 26 anos. Na tabela também tem as colunas: ?tp_maternal_schooling? que mostra a categoria de anos de escolaridade da mãe, ?tp_marital_status? que mostra o estado civil da mãe em dados categóricos e a coluna ?tp_maternal_race? que mostra a raça da mãe também em dados categóricos.
31
6.1.1.2 Variávies obstétricas maternas
No apêndice X pode-se observar uma tabela que possui as estatísticas sumarizadas das variáveis obstétricas maternas. Nesta tabela por exemplo temos a coluna ?num_live_births?, essa coluna mostra o número de nascidos vivos anteriores que a mãe obteve, a média desta coluna é 0.94 , a mediana é 1, o número mínimo é 0 e o máximo é 10 nascidos vivos. Na tabela podemos observar também a coluna, ?num_fetal_losses?, esta coluna mostra o número de perdas fetais anteriores da mãe, a média desta coluna é 0.21, a medina é 0, o número mínimo é 0 e o máximo é 5, esses valores mostram que poucas mães tiveram perdas fetais antes da gravidez atual. Também na tabela tem estatísticas da coluna ?num_previous_gestations? esta coluna mostra o número de gestações anteriores da mãe, esta coluna tem a média de 1.14, a mediana é 1, o número mínimo é 0 e o máximo 10. Essa tabela também contém as colunas: ?num_normal_labors? que é o número de partos normais da mãe, ?num_cesarean_labor? que mostra o número de partos cesáreos da mãe e por fim mostra a coluna ?tp_pregnancy? que é o tipo da gravidez.
6.1.1.3 Variávies de histórico de gravidez
No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis do histórico de gravidez. Nesta tabela temos as colunas ?num_prenatal_appointments? que mostra o número de consultas de pré-natal, a média dessa coluna é 7.8, a mediana é 8, a variação dos dados é de no mínimo 0 e no máximo 40. A tabela também contém as colunas: ?tp_labor? que mostra o tipo de parto, e também contém a coluna ?tp_robson_group? que mostra a classificação do grupo Robson.
6.1.1.4 Variávies relacionadas ao recém-nascido
No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis relacionadas ao recém-nascido. Nesta tabela por exemplo podemos observar a coluna ?newborn_weight? que armazena o peso ao nascer dos recém-nascidos em gramas, a média dos dados dessa coluna é 3187.3, a mediana é 3215, a variação dos dados é de no mínimo 0 e no máximo 6000 gramas. Também podemos observar a coluna ?gestacional_week? que mostra o número de semanas de gestação, a média é 38.4, a mediana é 39, a variação dos dados é de no mínimo 15 e no máximo 45 semanas. Além dessas colunas a tabela também
32
contém as colunas: ?cd_apgar1? que mostra a pontuação de Apgar de 1 minuto, ?cd_apgar5? que mostra a pontuação de Apgar de 5 minuto, ?has_congenital_malformation? que mostra se o recém-nascido contém a presença de alguma malformação, ?tp_newborn_presentation? que mostra o tipo de apresentação de recém-nascido, ?tp_labor? que mostra o tipo de parto, ?was_cesarean_before_labor? que mostra se o recém-nascido foi cesáreo antes do parto, ?was_labor_induced ? que mostra se o recém-nascido foi induzido ao trabalho de parto,

?tp_childbirth_care? que mostra se ouve assistência ao parto, e por fim temos a coluna ?is_neonatal_death? que mostra se o recém-nascido veio a óbito ou não.

6.1.2 Análise das variáveis

Nesta seção serão mostrados a parte de análise de variáveis com diferentes tipos de gráficos.

6.1.2.1 Variáveis contínuas

Na Figura 5 temos dois gráficos boxplot, no boxplot à esquerda temos a distribuição da variável peso ao nascer e nele é possível observar duas caixas onde a caixa azul são os vivos e o laranja são os mortos. Nas duas caixas mostradas no gráfico podemos observar pequenos círculos nas pontas, esses círculos são outliers (dados fora da curva), então para a caixa de vivos os outliers são recém-nascidos com pesos acima de 4500 gramas ou abaixo de 2000 gramas, e para a caixa de mortos os outliers são os recém-nascidos com mais de 5500 gramas. Para os vivos temos a mediana de aproximadamente 3200 gramas e para os mortos a mediana é 1300 gramas. Cada caixa representa 50% da base de dados, a caixa azul de vivos mostra que o peso de recém-nascidos que sobreviveram está aproximadamente entre 2700 a 3600 gramas, e para a caixa laranja de mortos o peso de recém-nascidos que vieram a óbito estão aproximadamente entre 700 a 2500 gramas. Então o gráfico mostra para nós que os recém-nascidos que têm maior risco de vir a óbito tem pesos menores que 2000 gramas.

33

Figura 5. BoxPlot das features Peso ao nascer e Idade da mãe. (Fonte: Elaboração Própria).

No gráfico à direita da Figura 5 temos um boxplot da distribuição da variável idade da mãe, como foi dito anteriormente a caixa azul são os vivos e o laranja são os mortos. Nesse boxplot então podemos ver que os outliers da caixa de vivos são mães com idade acima de 46 anos aproximadamente, para os mortos os outliers são mães com idade acima de 50 anos. Para os vivos temos a mediana de aproximadamente 26 anos e para os mortos a mediana é de 26 anos. Nos vivos a idade das mães está aproximadamente entre 21 a 32 anos, e para os mortos a idade das mães está aproximadamente entre 20 a 34 anos. É importante ressaltar que o gráfico mostra as duas caixas bem parecidas, então de acordo com os dados não contém diferença significativa entre vivos e mortos usando a idade da mãe para distribuir.

Já na Figura 6 temos um boxplot da distribuição da variável semanas de gestação.

Nesse boxplot então podemos ver que os outliers da caixa de vivos são gestações com mais de 44 semanas aproximadamente, e gestações com menos de 35 semanas. Para os vivos temos a mediana de aproximadamente 39 semanas e para os mortos a mediana é de 32 semanas. Nos vivos as semanas de gestação estão aproximadamente entre 36 a 40 semanas, e para os mortos a semanas de gestação estão aproximadamente entre 26 a 36 semanas. Nesse boxplot os dados mostraram que para as gestações que têm menos de 36 semanas os recém-nascidos têm maior risco de vir a óbito.

34

Figura 6. BoxPlot da feature semana de gestação. (Fonte: Elaboração Própria).

Na Figura 7 temos um histograma da distribuição de peso ao nascer por classe, podemos observar no gráfico que grande parte dos vivos (linha azul) estão distribuídos entre 2000 e 4000 gramas, a área entre esses valores tem muitos dados de vivos, já entre os valores 0 e 2000 gramas temos uma concentração dos dados de mortos.

Figura 7. Histograma de distribuição do peso entre as classes. (Fonte: Elaboração Própria)

6.1.2.2 Variáveis categoricas

Observando a Figura 8 podemos ver um gráfico de barras da distribuição da variável escolaridade da mãe, este gráfico mostra a contagem de registros por escolaridade da mãe, esse gráfico mostra para nós que a maior parte das mães estão na categoria 4 que representa de 8 a 11 anos de escolaridade.

35

Figura 8. Gráfico de barras da distribuição da variável Escolaridade da mãe. (Fonte: Elaboração Própria).

Observando a Figura 9 podemos ver um gráfico de barras da distribuição da variável número de vivos, este gráfico mostra a contagem de registros por número de vivos, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho.

Figura 9. Gráfico de barras da distribuição da variável Número de nascidos vivos. (Fonte: Elaboração Própria).

Observando a Figura 10 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 1 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 1 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem

36

apgar de 8 ou 9, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.

Figura 10. Gráfico de barras da distribuição da variável Pontuação Apgar de 1 minuto. (Fonte: Elaboração Própria).

Observando a Figura 11 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 5 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 5 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem apgar de 9 ou 10, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.

Figura 11. Gráfico de barras da distribuição da variável Pontuação Apgar de 5 minutos. (Fonte: Elaboração Própria).

37

Observando a Figura 12 podemos ver um gráfico de barras da distribuição da variável presença de malformação congênita, este gráfico mostra a contagem de registros por presença de malformação congênita, esse gráfico mostra para nós que a grande maioria dos dados estão na categoria 2 que mostra que o recém-nascido não possui malformação, esse grande volume para a categoria 2 é causada pelo desbalanceamento da base.

Figura 12. Gráfico de barras da distribuição da variável Presença de malformação congênita. (Fonte: Elaboração Própria).

Observando a Figura 13 podemos ver um gráfico de barras da distribuição da variável número de gestações, este gráfico mostra a contagem de registros por número de gestações, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho, da mesma forma mostrada na Figura 9.

38

Figura 13. Gráfico de barras da distribuição da variável Número de gestações. (Fonte: Elaboração Própria).

Observando a Figura 14 podemos ver um gráfico de barras da distribuição da variável assistência ao parto, este gráfico mostra a contagem de registros por assistência ao parto, esse gráfico mostra para nós que a grande maioria dos dados mostram que o parto teve assistência de uma Enfermeira ou obstetra ou essa informação foi ignorada na hora do resgistro.
Figura 14. Gráfico de barras da distribuição da variável Assistência ao parto. (Fonte: Elaboração Própria).
6.1.2.3 Analise bi-variada
A Figura 15 mostra para nós a importância da feature peso ao nascer com o gráfico mostrado é possível observar claramente a diferença de pesos entre vivos e mortos, recém-nascidos com peso acima de 2000 gramas sobreviveram, já os recém-nascidos com menos de 2000 gramas vieram a óbito.
39
Figura 15. Gráfico de densidade de peso entre as classes. (Fonte: Elaboração Própria).
Na Figura 16 podemos observar um gráfico de correlação entre algumas features de valores contínuos, algumas dessas features possuem correlações positivas, como pontuação apgar de 1 minuto e pontuação apgar de 5 minutos, essas colunas possuem uma correlação de 0.7, então nascidos que recebem um valor alto de apgar de 1 minuto tendem a receber um valor alto no apgar de 5 minutos. O gráfico também mostra outras features com correlações positivas como, número de partos normais e número de gestações anteriores que contém uma correlação de 0.81, número de gestações anteriores com idade da mãe que tem uma correção de 0.39, número de partos de cesáreos com idade da mãe que possui correlação de 0.24, número de partos normais com idade da mãe com a correlação de 0.26, semanas de gestação com peso ao nascer em gramas com correção de 0.52, e as apgar de 1 e 5 minutos com semanas de gestação. E grande parte das features, possuem correlações baixas então elas são inversamente correlacionadas pelo que o gráfico mostra, como por exemplo número de partos normais com número de consultas pré-natais possuem uma correção de -0.17, e número de partos cesáreos com número de partos normais que contém uma correlação de -0.15. Então a correlação dessas features estão bem baixas tirando as correlações acima de 0.7.
40
Figura 16. Gráfico de correlação. (Fonte: Elaboração Própria)
6.1.2.4 Distribuição por raça
Na Figura 17 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por raça, e podemos observar que a raça 5 (indigena), é o boxplot que contem a menor variação de idade, já as raças 1 (branco) e 3 (amarelo) são as caixas que possuem maior variação com a média em torno de 20 e 30 anos.
41
Figura 17. Distribuição da idade da mãe por raça. (Fonte: Elaboração Própria).
No boxplot mostrado na Figura 18 podemos observar a distribuição de peso ao nascer por raça, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre as raças para peso ao nascer.
Figura 18. Distribuição de peso ao nascer por raça. (Fonte: Elaboração Própria).
Na Figura 19 temos o boxplot da a distribuição de semanas de gestação por raça, e praticamente todas as caixas são iguais, somente a caixa da raça 1 (branco) possui menor variação que as outras.

42
Figura 19. Distribuição de semanas de gestação por raça. (Fonte: Elaboração Própria).
6.1.2.5 Distribuição por estado civil
A Figura 20 mostra o boxplot da a distribuição de peso ao nascer por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre os estados civis para peso ao nascer.
Figura 20. Distribuição de peso ao nascer por estado civil. (Fonte: Elaboração Própria).
A Figura 21 mostra o boxplot da a distribuição de semanas de gestação por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Porém os estados civis 2 (Casado) e 4 (Separados/divorciados judicialmente), contém variações menores.
43
Figura 21. Distribuição de semanas de gestação por estado civil. (Fonte: Elaboração Própria).
Na Figura 22 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por estado civil, e podemos observar que o estado civil 3 (Viúva), possui a maior variação, já o estado civil 4 (Separados/divorciados judicialmente) é a caixa que possuem menor variação
Figura 22. Distribuição da idade da mãe por estado civil. (Fonte: Elaboração Própria).
6.1.2.6 Distribuição por escolaridade da mãe
Na Figura 23 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por escolaridade da mãe, e podemos observar que a escolaridade 2 (1 a 3 anos) e 3 (4 a 7
44
anos), possui as maiores variações, já escolaridade 5 (12 ou mais anos) é a caixa que possuem menor variação
Figura 23. Distribuição da idade da mãe por escolaridade da mãe. (Fonte: Elaboração Própria).
A Figura 24 mostra o boxplot da a distribuição de peso ao nascer por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então, pouca variação é observada entre a escolaridade da mãe para peso ao nascer.
Figura 24. Distribuição de peso ao nascer por escolaridade da mãe. (Fonte: Elaboração Própria).
A Figura 25 mostra o boxplot da a distribuição de semanas de gestação por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas,
45
todas são praticamente iguais. Porém a escolaridade 5 (12 ou mais anos) contém variação menor que as outras.
Figura 25. Distribuição de semanas de gestação por escolaridade da mãe. (Fonte: Elaboração Própria).
6.2 ÁRVORE DE DECISÃO
A Figura 26 mostra um trecho do algoritmo utilizado, nesse trecho é realizado a divisão do DataFrame das features de entrada , e o DataFrame que contém o rótulo.Neste caso o rótulo será o DataFrame ?target?, esse Dataframe contém a feature

?is_neonatal_death?, essa feature informa se o recém nascido veio a óbito ou não, sendo o foco da predição que desejamos realizar essa feature entra como rótulo para o modelo, já no Dataframe ?nome_features? temos as features que vão servir de entrada para o modelo, é a partir dessas features que o modelo tentará encontrar padrões para predizer o risco de óbito do recém nascido. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

46

Figura 26. Separação dos DataFrames de entrada e rótulo. (Fonte: Elaboração Própria).

6.2.1 Modelo de Árvore de Decisão Utilizando Particionamento 90/10

O algoritmo de árvore de decisão foi treinado utilizando duas formas diferentes: usando as partições 90% para treino e 10% para os testes e utilizando o método K-folds cross-validation. E para os experimentos iniciais foi utilizado o particionamento 90/10.

Tabela 4. Resultados do modelo de árvore de decisão. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	671807
Morto(1)	0.71	0.29	0.41	4216

Analisando a Tabela 4 pode-se ver os seguintes resultados, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 71% que realmente pertenciam a classe 1, então cerca de 29% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 29% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem a classe 1 que seria dos recém nascidos que poderiam vir a óbito, já para a classe de vivos a

47

classe 0 o modelo está identificando muito bem e classificando de forma correta. A tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.

Esse modelo teve uma acurácia de 0.99 que é uma acurácia bem alta, porém somente pela acurácia não é possível dizer que o modelo está excelente pelo motivo da base que está sendo usada é altamente desbalanceada, então para a classe 0 que é a de vivos temos muito mais dados que para a classe de mortos, e o modelo está acertando bastante a classe de vivos logo a acurácia fica alta por esses acertos, enquanto para a classe de mortos o modelo não está acertando tanto.

A Figura 27 mostra a curva ROC do modelo, pode-se observar que a AUC da curva ROC ficou em 0.92, a AUC mostra que o modelo está acertando bastante as predições, pois quanto mais perto de 1 melhor está no nosso modelo, mas no caso desse modelo vimos acima que as predições para a classe de mortos(1) está ruim, o modelo está acertando poucas classificações como 1. Então a AUC assim como a acurácia está alta porque a base de dados utilizada é altamente desbalanceada tendo muito mais amostras de vivos(1), e como o modelo está bom para essa classificação e está acertando bastante os valores de AUC e acurácia ficam altos .

Figura 27. Curva ROC modelo de Árvore de decisão. (Fonte: Elaboração Própria).
A Figura 28 mostra a importância de cada feature para o modelo sendo assim a feature 10 - Peso ao nascer, a mais importante para o modelo, pois dela o modelo conseguiu tirar mais informações, seguido das features, 13 - Apgar dos 5 primeiros minutos, 11 - Semanas de 48
Gestação, 14 - Presença de malformação 12 - Apgar do 1 primeiro minutos. Então essas são as features que foram mais decisivas para a montagem da árvore e para as predições do modelo.
Figura 28. Gráfico de importância das features. (Fonte: Elaboração Própria).
Na Figura 29 temos uma imagem reduzida da árvore de decisão que foi montada com o treinamento utilizando 5 como profundidade máxima para a árvore, a imagem da árvore completa pode ser visualizada no apêndice A. É possível ver que as features marcadas como importante para o modelo apareceu diversas vezes na árvore, e a feature peso ao nascer foi escolhida para ser o nó raiz da árvore, de todas as features ela foi a que teve a melhor entropia para ser o nó raiz, e depois aparece mais vezes para outras decisões em outros níveis da árvore e isso faz com que essa feature se torne a mais importante para o modelo, também podemos ver que a feature ?Apgar do 1 primeiro minuto? mesmo sendo a menos importante para o modelo ela aparece somente no nó de número 42 isso mostra que a entropia e o ganho de informação dessa feature nas outras decisões não foram boas fazendo ela ser a feature com menos importância utilizada no modelo.
49
Figura 29. Árvore de decisão montada a partir do algoritmo. (Fonte: Elaboração Própria).
6.2.2 Modelo de Árvore de Decisão Utilizando K-folds cross-validation
Após realizar esse experimento com a partição 90/10, foi realizado outro experimento com esse mesmo algoritmo porém agora utilizando o método K-folds cross-validation para o treinamento. O K-folds cross-validation foi configurado para separar a base de dados em 10 partes iguais, e assim o algoritmo foi treinado novamente gerando um novo modelo. O método de K-folds cross-validation é utilizado para conferir se o modelo não está sofrendo problemas de overfitting, é importante garantir que o modelo está aprendendo com os dados e não está somente decorando.
Na Tabela 5 temos então o resultados desse modelo, pode-se observar que os resultados para a classificação de vivos(0) não teve alteração, o modelo continua tendo 100% na precisão e no recall, mas na classificação de mortos(1) o modelo teve uma diminuição na precisão que agora está em 65% e também teve uma diminuição no recall que agora está em 23%, porém essa diminuição é justificada porque para o teste deste modelo foi utilizado a base completa e não somente a partição de teste que contém 10% da base total. Logo os resultados não tiveram uma alteração drástica e tiveram um comportamento bem semelhante, incluindo a acurácia que se manteve em 0.99 e a AUC que teve uma diminuição pequena também e ficou com 0.89.
50
Tabela 5. Resultados do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	6719191

Morto(1) 0.65 0.23 0.34 41031
Na Tabela 6 é possível ver a matriz de confusão do modelo que mostra o número exato de erros e acertos do modelo, pode-se observar que para a classe de vivos o modelo classificou corretamente 6.714.117 amostras da base de dados completa, o modelo classificou como 0 as amostras que nos rótulos realmente eram 0, errou apenas 5074 amostras. Já para a classe de mortos o modelo acertou somente 9552 amostras, classificou como 1 as amostras que no rótulo eram 1 também, e errou 31479. Isso mostra que o modelo está muito desbalanceado, pois está errado muito mais do que acertando as predições de mortos, sendo isso um reflexo da base de dados desbalanceada também. E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.
Tabela 6. Matriz de confusão do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6714117 5074
Mortos (1) 31479 9552
6.3 REGRESSÃO LOGÍSTICA
Para o algoritmo de Regressão Logística também foi utilizado métodos diferentes para o treinamento, O algoritmo foi treinado utilizando três formas diferentes: usando as partições 90% para treino e 10% para os testes, utilizando o método K-folds cross-validation e foi treinada utilizando o método RFECV que seleciona as melhores features para o modelo, juntamente com o método GridSearchCV. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.
51
6.3.1 Modelo de Regressão Logística Utilizando Particionamento 90/10
Para o primeiro experimento do algoritmo de regressão logística foi utilizado o particionamento 90/10 para o treinamento do algoritmo, sendo 90% da base de dados para realizar o treinamento do modelo e 10% da base para realizar os testes.
Na Tabela 7 temos os resultados do modelo de regressão logística utilizando o método de particionamento 90/10, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 65% que realmente pertenciam a classe 1, então cerca de 35% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 24% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem as amostras da classe 1 que seria dos recém-nascidos que poderiam vir a óbito, já para a classe de vivos a classe 0 o modelo está identificando muito bem e classificando de forma correta.E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.

Tabela 7. Resultados do modelo de regressão logística usando particionamento 90/10.
(Fonte: Elaboração Própria)
Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 671885
Morto(1) 0.65 0.24 0.35 4138
A Figura 30 abaixo mostra a curva ROC do modelo de regressão logística usando o particionamento 90/10, analisando a curva ROC podemos ver que a AUC da curva ficou em 0.93 e a acurácia do modelo ficou 0.99, mostrando que o modelo está acertando bastante as predições, porém esses números para o nosso modelo não mostra que ele está ótimo, com a análise da Tabela 7 acima podemos observar que o modelo está acertando muitas predições da classe de vivos e poucas predições da classe de mortos, como a base de dados é
52
desbalanceada, como a maior quantidade de dados está na classe de vivos e o modelo está acerto quase todos dessa classe a acurácia e a AUC ficam altas por esses acertos.Logo é preciso analisar mais resultados além da acurácia e da AUC do modelo.
Figura 30. Curva ROC modelo regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria).
Na Tabela 8 podemos analisar os número exatos que o modelo acertou de cada classe, como é mostrado na matriz de confusão o modelo acertou 671.435 da classe de vivos e acertou apenas 915 da classe de mortos, e errou mais que o triplo da classe de mortos. Então as predições do modelo estão muito desbalanceadas, para a classe de vivos o modelo está predizendo bem, porém para as classes de mortos o modelo está errando muitas predições.
Tabela 8. Matriz de confusão do modelo de regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 671435 504
Mortos (1) 3169 915
6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation
Para o segundo experimento do algoritmo de regressão logística foi utilizado o método K-folds cross-validation para o treinamento do modelo com a intenção de conferir e confirmar
53
que o modelo não esteja tendo problemas com overfitting e realmente esteja aprendendo com os dados que está sendo usado para o treinamento. O método K-folds cross-validation foi configurado para realizar uma separação de 10 partes iguais.
Na Tabela 9 podemos observar os resultados do modelo, e os resultados foram parecidos com o experimento anterior usando o particionamento 90/10, única diferença nos resultado é que no experimento anterior a precisão da classe de mortos ficou em 65% e no caso desse experimento usando o K-folds cross-validation a precisão ficou em 64%, mas os valores de recall e F1-Score se mantiveram, assim como todos os resultados da classe de mortos se mantiveram em 100%. A execução desse experimento mostra que o modelo realmente está aprendendo com os dados e não está apenas decorando, retirando o risco do modelo estar com overfitting. Mesmo sem alterações no resultado é importante saber que o

modelo não está com overfitting e está conseguindo aprender e encontrar padrões nos dados.
Tabela 9. Resultados do modelo de regressão logística usando K-folds cross-validation.
(Fonte: Elaboração Própria)
Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 6719191
Morto(1) 0.64 0.24 0.35 41031
Já na Figura 31 temos a curva ROC do modelo, essa curva ROC mostra a AUC de
todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do
método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas
para todas as vezes que foi realizado os testes, a AUC se manteve em 0.93 e 0.94, sendo a
média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não
se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua
acertando muito a classe de vivos que é a classe predominante na base de dados. Já era
esperado os resultados não sofrerem alterações grandes, pois a base de dados não sofreu
nenhuma alteração, esse experimento foi somente para conferir se a modelo não estava
sofrendo overfitting, e pelos resultados aparentemente a base não está com problemas de
sobreajuste.
54
Figura 31. Curva ROC modelo regressão logística usando K-folds cross-validation. (Fonte:
Elaboração Própria).
Na Tabela 10 temos a matriz de confusão do modelo, onde mostra que o modelo
acertou 6.713.658 amostras da classe de vivos e errou apenas 5.533, já para a classe de mortos
o modelo acertou somente 9.847 e novamente errou mais que o triplo errando 31.182
amostras. Como os resultados mostrados anteriormente não tiveram alteração era de se
esperar que as predições corretas e incorretas do modelo também não sofressem alterações, o
modelo continua acertando muito a classe de vivos e errando muito a classe de mortos,
mantendo as predições bem desbalanceadas.
55
Tabela 10. Matriz de confusão do modelo de regressão logística usando K-folds
cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6713658 5533
Mortos (1) 31182 9847
6.3.3 Modelo de Regressão Logística Utilizando GridSearchCV e RFECV
Para o terceiro e último experimento de regressão logística foi utilizado técnicas
diferentes juntas para tentar melhorar as predições do modelo, para esse experimento usamos
o K-folds cross-validation para o particionamento da base de dados assim como foi utilizado
no experimento acima, e também foi utilizado duas técnicas que ainda não foram usadas nos
experimentos anteriores que são as funções RFECV e GridSearchCV. A função RFECV
seleciona a quantidade ideal de features para ser usadas no treinamento do modelo e também
mostra quais são as melhores features para essa predição, então essa função ela utiliza a
validação cruzada para ir treinando e testando N vezes o modelo, e sempre vai alterando a

quantidade e as features utilizadas para o teste, então cada vez que a função testa os modelos ela grava a pontuação dos acertos e assim depois mostra o gráfico da Figura 32 e assim é possível ver quais são as features ideais para o modelo.
56
Figura 32. Resultado da função RFECV Base Brasil. (Fonte: Elaboração Própria).
Então os resultados mostrados na Figura 32 apresenta que o número ideal para treinar o modelo de features é 6, e as features sao: 'tp_maternal_schooling', 'gestaional_week', 'cd_apgar1', 'cd_apgar5', 'has_congenital_malformation', 'tp_labor'. Essas são as features que tiveram o melhor desempenho no RFECV, então para o treinamento do modelo desse experimento as features que vão ser utilizadas serão somente essas 6.
Após a execução da função de RFECV foi executado a função de GridSearchCV, o GridSearchCV é utilizado para descobrir quais são os melhores parâmetros para utilizar no algoritmo de regressão logística e criar os modelos, foi utilizado também a validação cruzada para testar qual seria os melhores parâmetros para melhorar o desempenho do modelo. Na Figura 33 pode-se ver que os parâmetros escolhidos pela função foram:
Figura 33. Resultado da função GridSearchCV. (Fonte: Elaboração Própria).
Na Tabela 11 podemos observar os resultados do modelo, e os resultados foram parecidos com os experimentos anteriores, para a classe de mortos a precisão teve um aumento e foi para 69% e o recall diminuiu chegando em 20%, então utilizando as novas
57
funções o modelo ganhou precisão e perdeu recall. Para a classe de vivos o modelo se manteve em 100% e não teve alterações para os resultados de precisão e recall. Mesmo com a utilização das funções de RFECV e GridSearchCV o modelo não teve uma melhora significativa para as predições de mortos.
Tabela 11. Resultados do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	6719191
Morto(1)	0.69	0.20	0.31	41031

Na Figura 34 temos a curva ROC do modelo, essa curva ROC mostra a AUC de todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas para todas as vezes que foi realizado os testes, a AUC se manteve em 0.92 e 0.93, sendo a média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua acertando muito a classe de vivos que é a classe predominante na base de dados.
58
Figura 34. Curva ROC modelo regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria).
Na Tabela 12 temos a matriz de confusão do modelo, onde mostra que o modelo acertou 6.715.535 amostras da classe de vivos e errou apenas 3.656, comparando com os experimentos anteriores de regressão logística as predições de vivos teve uma pequena piora e teve uma pequena diminuição nos acertos. Já para a classe de mortos, o modelo acertou 8.296 e errou 32.735 amostras, logo as predições de vivos tiveram uma uma piora e acertou um

pouco menos do que os outros experimentos, e na classe de vivos teve uma melhora acertando mais predições, porém como o problema está nas predições de mortos e para essa classe teve uma piora pode-se dizer que a função de GridSearchCV não teve um ganho significativo nos resultados. Então como modelo final foi escolhido o modelo utilizando somente o K-folds cross-validation, pois foi o modelo que mostrou o melhor desempenho dos experimentos de regressão logística.

59

Tabela 12. Matriz de confusão do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

Predito

Vivos (0) Mortos (1)

Classe Real

Vivos (0) 6715535 3656

Mortos (1) 32735 8296

60

7 CONCLUSÃO

O principal objetivo deste trabalho foi analisar os resultados das predições de mortalidade neonatal dos algoritmos de aprendizado de máquina usando uma base de dados com dados do Brasil inteiro. A partir destes resultados apresentados na secção anterior ?Resultados?, é possível concluir que analisar somente a AUC e a acurácia do modelo não é o suficiente para garantir que o modelo está excelente, sendo assim temos que ter mais atenção a precisão, recall e as predições mostradas na matriz de confusão.

Ambos os modelos ficaram desbalanceados nas predições, os modelos acertaram mais predições de vivos do que mortos, e as predições corretas de mortos foram mais baixas do que as incorretas, para os modelos ficarem melhor precisam passar por algoritmos que possam balancear essas predições, uma outra alternativa para melhorar o modelo é utilizar uma base de dados mais balanceada, pois essa base que foi utilizada é altamente desbalanceada.

Embora os dois modelos tenham tido resultados parecidos , o modelo de regressão logística utilizando somente o K-folds cross-validation se saiu melhor que os outros experimentos, o modelo criado na seção ?6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation? manteve a acurácia, AUC, precisão e recall dos demais modelos e acertou mais predições, talvez com algoritmos para melhorar as predições da classe de mortos, balanceando mais as predições, esse modelo fique com ótimas taxas preditivas.

Além disso podes-se afirmar que o peso ao nascer do recém nascido é um fator muito importante para as decisões dos modelos, ambos os modelos usaram muito essa informação para as predições, também podemos colocar, as colunas de presença de malformação e semana de gestação, como colunas que foram importantes para os modelos, essas causas podem ser evitadas se houver um acompanhamento médico com qualidade e eficiência.

61

REFERÊNCIAS

ÁRVORE de Decisão - Aula 3. Gravação de Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=ecYpXd4WREk&t=4089s. Acesso em: 1 jul. 2020.

BARRETO, J. O. M.; SOUZA, N. M.; CHAPMAN, E. Síntese de evidências para políticas de saúde: reduzindo a mortalidade perinatal. Ministério da Saúde, p. 1?44,2015.

BATISTA, André Filipe de Moraes; FILHO, Alexandre Dias Porto Chiavegatto. Machine Learning aplicado à Saúde. In: ZIVIANI, Artur; FERNANDES, Natalia Castro; SAADE, Débora Christina Muchaluat. SBCAS 2019: 19 Simpósio Brasileiro de Computação Aplicada à Saúde. [S. l.: s. n.], 2019. cap. Capítulo 1.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. Towards neonatal mortality risk classification: A data-driven approach using neonatal, maternal, and social factors, [s. l.], 23 out. 2020.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. SPNeoDeath: A demographic and epidemiological dataset having infant, mother, prenatal care and childbirth data related to births and neonatal deaths in São Paulo city Brazil ? 2012?2018, [s. l.], 19 jul. 2020.
BRESAN, Rodrigo. Um método para a detecção de ataques de apresentação em sistemas de reconhecimento facial através de propriedades intrínsecas e Deep Learning. Orientador: Me. Carlos Eduardo Beluzo. 2018. Trabalho de Conclusão de Curso (Superior de Tecnologia em Análise e Desenvolvimento de Sistemas) - Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas., [S. l.], 2018.
CABRAL, Cleidy Isolete Silva. Aplicação do Modelo de Regressão Logística num Estudo de Mercado. Orientador: João José Ferreira Gomes. 2013. Projeto Mestrado em Matemática Aplicada à Economia e à Gestão (Mestrado) - Universidade de Lisboa Faculdade de Ciências Departamento de Estatística e Investigação Operacional, [S. l.], 2013. Disponível em: https://repositorio.ul.pt/bitstream/10451/10671/1/ulfc106455_tm_Cleidy_Cabral.pdf. Acesso em: 1 maio 2021.
CAMPOS, Raphael. Árvores de Decisão: Então diga-me, como construo uma?. [S. l.], 28 nov. 2017. Disponível em: https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69. Acesso em: 23 jan. 2021.
Colab. 2021. Disponível em: https://colab.research.google.com/. Acesso em: 20 mar. 2021.
COX, D.R.;SNELL, E.J. Analysis of Binary Data.London: Chapman & Hall, 2ª Edição, 1989. HOSMER, D.W.; LEMESHOW, S.Applied Logistic Regression. New York: John Wiley, 2ª Edição, 2000.
E.França, S.Lansky. Mortalidade infantil neonatal no brasil: situação, tendências e perspectivas Rede Interagencial de Informações para Saúde - demografia e saúde: contribuição para análise de situação e tendências Série Informe de Situação e Tendências(2009), pp.83-112.
62
FREIRE, Sergio Miranda. Bioestatística Básica: Regressão Linear. [S. l.], 20 out. 2020. Disponível em: http://www.lampada.uerj.br/arquivosdb/_book/bioestatisticaBasica.html. Acesso em: 23 jan. 2021.
Jupyter. 2021. Disponível em: https://jupyter.org/. Acesso em: 20 jun. 2021.
GOODFELLOW, I. et al. Deep learning. [S.l.]: MIT press Cambridge, 2016.
Pandas. 2021. Disponível em: https://pandas.pydata.org/. Acesso em: 20 jun. 2021.
Python. 2021. Disponível em: https://www.python.org/. Acesso em: 20 jun. 2021.

KUHN, M.; JOHNSON, K. Applied predictive modeling. [S.l.]: Springer, 2013. v. 26.
LANSKY, S. et al. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante . Cadernos de Saúde Pública, Scielo, v. 30, p. S192 ? S207, 00 2014.
LANSKY, Sônia; FRICHE, Amélia Augusta de Lima; SILVA, Antônio Augusto Moura; CAMPOS, Deise; BITTENCOURT, Sonia Duarte de Azevedo; CARVALHO, Márcia Lazaro; FRIAS, Paulo Germano; CAVALCANTE, Rejane Silva; CUNHA, Antonio José Ledo Alves. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante e ao recém-nascido. [s. l.], Ago 2014. Disponivel em: https://www.scielosp.org/article/csp/2014.v30suppl1/S192-S207/pt/
Matplotlib. 2021. Disponível em: https://matplotlib.org/. Acesso em: 20 mar. 2021.
Maranhão AGK, Vasconcelos AMN, Trindade CM, Victora CG, Rabello Neto DL, Porto D,et al. Mortalidade infantil no Brasil: tendências, componentes e causas de morte no período de 2000 a 2010 . In: Departamento de Análise de Situação de Saúde, Secretaria de Vigilância em Saúde, Ministério da Saúde, organizador. Saúde Brasil 2011: uma análise da situação de saúde e a vigilância da saúde da mulher. v. 1. Brasília: Ministério da Saúde; 2012. p. 163-82.
MESQUITA, PAULO. UM MODELO DE REGRESSÃO LOGÍSTICA PARA AVALIAÇÃO DOS PROGRAMAS DE PÕS-GRADUAÇÃO NO BRASIL. Orientador: Prof. RODRIGO TAVARES NOGUEIRA. 2014. Dissertação (Mestre em Engenharia de Produção) - Universidade Estadual do Norte Fluminense, [S. l.], 2014.
MNASSRI , Baligh. Titanic: logistic regression with python. [S. l.], 1 ago. 2020. Disponível em: https://www.kaggle.com/mnassrib/titanic-logistic-regression-with-python. Acesso em: 14 jan. 2021.
Morais, R.M.d., Costa, A.L: Uma avaliação do sistema de informações sobre mortalidade. Saúde em Debate 41, 101?117 (2017). Disponível em: https://doi.org/10.1109/SIBGRAPI.2012.38.
MOTA, Caio Augusto de Souza; BELUZO , Carlos Eduardo; TRABUCO, Lavinia Pedrosa; SOUZA , Adriano; ALVES, Luciana; CARVALHO, Tiago. DESENVOLVIMENTO DE UMA PLATAFORMA WEB PARA CIÊNCIA DE DADOS APLICADA À SAÚDE MATERNO INFANTIL , [s. l.], 28 nov. 2019.
MULTIEDRO (Brasil). Conheça as aplicações do machine learning na área da saúde. [S. l.], 28 nov. 2019. Disponível em: https://blog.multiedro.com.br/machine-learning-na-area-da-saude/#:~:text=O%20machine%263
0learning%20na%20%C3%A1rea,diagn%C3%B3sticos%20e%20tratamentos%20mais%20precisos. Acesso em: 13 jun. 2021.
MUKHIYA,S.K.;AHMED,U. Hands-On Exploratory Data Analysis with Python: Perform EDA Techniques to Understand, Summarize, and Investigate Your Data. [S.l.]: PacktPublishingLtd,2020.ISBN978-1-78953-725-3.22,32
Murray CJ, Laakso T, Shibuya K, Hill K, Lopez AD. Can we achieve Millennium Development Goal 4? New analysis of country trends and forecasts of under-5 mortality to 2015. Lancet 2007; 370:1040-54.
NumPy. 2021. Disponível em:https://numpy.org/. Acesso em: 20 jun. 2021.
Oliveira, M.M.d., de Araújo, A.S.S.C., Santiago, D.G., Oliveira, J.a.C.G.d., Carvalho, M.D.,

de Lyra et al, R.N.D.: Evaluation of the national information system on live births in brazil , 2006-2010. Epidemiol. Serv. Saúde 24(4), 629?640 (2015).
PELISSARI, ANA CAROLINA CHEBEL. MORTALIDADE NEONATAL DA CIDADE DE SÃO PAULO: UMA ABORDAGEM UTILIZANDO APRENDIZADO DE MÁQUINA SUPERVISIONADO. 2021. Trabalho de Conclusão de Curso (Superior) - Instituto Federal de Educação, Ciência e Tecnologia Campus Campinas, [S. l.], 2021.
REGRESSÃO logística e binary cross entropy - Aula 7. Direção: Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=3J-LBtHVsm4. Acesso em: 21 jan. 2021.
Rmd Nascimento, AJM Leite, NMGSd Almeida, Pcd Almeida, Cfd Silva Determinantes da mortalidade neonatal: estudo caso-controle em fortaleza, ceará, brasil Cad Saúde Pública, 28(2012), pp.559 - 572.
SANTOS, Hellen Geremias. Machine learning e análise preditiva: considerações metodológicas: métodos lineares para classificação. In: SANTOS, Hellen Geremias. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina. 2018. Tese (Pós-Graduação em Epidemiologia) - Faculdade de Saúde Pública da Universidade de São Paulo, [S. l.], 2018.
Scikit-learn. 2021. Disponível em: https://scikit-learn.org/stable/. Acesso em: 20 jun. 2021.
Seaborn. 2021. Disponível em: https://seaborn.pydata.org/. Acesso em: 20 jun. 2021.
SILVA, Wesley; CORSO, Jansen; WELGACZ, Hanna; PEIXE, Julinês. AVALIAÇÃO DA ESCOLHA DE UM FORNECEDOR SOB CONDIÇÃO DE RISCOS A PARTIR DO MÉTODO DE ÁRVORE DE DECISÃO. ARTIGO ? MÉTODOS QUANTITATIVOS, [s. l.], 12 ago. 2008.
WHO, W. H. O.Women and health: today?s evidence, tomorrow?s agenda. [S.l.]: UNWorld Health Organization, 2009. ISBN 9789241563857.
64
APÊNDICE A

==========

Arquivo 1: monografia-V7.docx.pdf (9728 termos)
Arquivo 2: https://www.lume.ufrgs.br/bitstream/handle/10183/46711/Poster_11760.pdf?sequence=2 (525 termos)
Termos comuns: 14
Similaridade: 0,13%
O texto abaixo é o conteúdo do documento monografia-V7.docx.pdf (9728 termos)
Os termos em vermelho foram encontrados no documento
https://www.lume.ufrgs.br/bitstream/handle/10183/46711/Poster_11760.pdf?sequence=2 (525 termos)

==========

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO
CÂMPUS CAMPINAS
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
CAMPINAS
2021
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas.
Orientador: Prof. Me. Everton Josué da Silva.
CAMPINAS
2021
Dados Internacionais de Catalogação na Publicação
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo Câmpus Campinas.
Aprovado pela banca examinadora em: _____ de _____ de _____.
BANCA EXAMINADORA

_____

Prof. Me. Everton Josué da Silva (orientador)
IFSP Câmpus Campinas

_____

Prof. Me. Carlos Eduardo Beluzo
IFSP Câmpus Campinas

_____

Prof. Dr. Ricardo Barz Sovat
IFSP Câmpus Campinas

Dedico este trabalho aos meus familiares,
colegas de classe, professores e servidores do Instituto
que colaboraram em minha jornada formativa.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pelo dom da vida e pela oportunidade de concluir mais uma etapa de minha experiência acadêmica.

Agradeço a todos os professores e servidores do IFSP Câmpus Campinas, que contribuíram direta e indiretamente para a conclusão deste trabalho.

Agradeço também à minha família, que deu todo o apoio necessário para que eu chegasse até aqui.

Agradeço ao meu orientador que me auxiliou a solucionar as dificuldades encontradas no caminho.

"Minha energia é o desafio, minha motivação é o impossível,
e é por isso que eu preciso
ser, à força e a esmo, inabalável.".
Augusto Branco

RESUMO

A mortalidade infantil pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível em uma população. O aprendizado de máquina é uma área da inteligência artificial que propõe métodos de análise de dados que automatizam a construção de modelos analíticos com base em reconhecimento de padrões, baseado no conceito de que sistemas podem aprender com dados, identificando padrões e tomando decisões com o mínimo de intervenção humana. O objetivo deste trabalho é testar e analisar dois tipos diferentes de algoritmos de aprendizado de máquina, utilizando a base de dados do SIM e SINASC do Brasil, do período de 2016 até 2018, para gerar modelos para predição de mortalidade neonatal. Os métodos utilizados foram Árvore de Decisão e de Regressão Logística e como métricas para avaliar os modelos resultantes foram utilizadas a AUC, Curva ROC e a Matriz de Confusão. Como resultado, apesar de terem alcançado valores de AUC 0.93, ambos modelos de predições acertaram muitas predições de vivos cerca de 671.000 e erraram muitas predições de mortos cerca de 2.600, principalmente devido ao fato de a base estar desbalanceada.

Palavras-chave: Mortalidade Neonatal, Predição, Aprendizado de Máquina.

ABSTRACT

Infant mortality can be used to analyze poverty and socioeconomic levels, as well as measure the quality of health and medical technology available in a population. Machine learning is an

area of artificial intelligence that proposes data analysis methods that automate the construction of analytical models based on pattern recognition, based on the concept that systems can learn from data, identifying patterns and making decisions with a minimum of human intervention. The objective of this work is to test and analyze two different types of machine learning algorithms, using the SIM and SINASC do Brasil database, from 2016 to 2018, to generate models for predicting neonatal mortality. The methods used were Decision Tree and Logistic Regression and as metrics to evaluate the resulting models, AUC, ROC Curve and Confusion Matrix were used. As a result, despite achieving values of AUC 0.93, both prediction models correct many live birth predictions around 671.000 and miss many stillbirth predictions around 2600, mainly due to the fact that the base is unbalanced.
Keywords: Neonatal Mortality, Prediction, Machine Learning.
LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

SIM Sistema de Informação sobre Mortalidade

SINASC Sistema de Informações sobre Nascidos Vivos

TMI Taxa de Mortalidade Infantil

TMN Taxa de Mortalidade Neonatal

MN Mortalidade Neonatal

ML Machine Learning

DNV Declaração de Nascido Vivo

VN Verdadeiro Negativo

FN Falso Negativo

VP Verdadeiro Positivo

FP Falso Positivo

ROC Curva Característica de Operação do Receptor

NaN Não é um Número

RFE Eliminação Recursiva de Feature

RFECV Eliminação Recursiva de Feature com Validação Cruzada

SUMÁRIO

13

1 INTRODUÇÃO

A taxa de mortalidade infantil (TMI) é uma importante medida de saúde em uma população e é um grande problema no mundo todo. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível. Esta taxa é dividida em duas categorias: neonatal e pós-neonatal. É categorizada neonatal quando o óbito ocorre nos 28 primeiros dias de vida após o pós-parto, e o pós-neonatal é quando o óbito ocorre entre 29 e 364 dias de vida (BELUZO et al., 2020). Cerca de 46% das mortes no mundo acontecem entre os menores de cinco anos e grande parte está concentrada nos primeiros dias de vida (LANSKY, 2014). A diminuição dessas taxas é muito importante no mundo todo, refletindo nos indicadores de saúde pública e desenvolvimento do país.

Os fatores associados à mortalidade são profundamente influenciados pelas características biológicas maternas e neonatais, pelas condições sociais e pelos cuidados prestados pelos serviços de saúde (E. FRANÇA; S. LANSKY, 2009; R.M.D. NASCIMENTO,

2012). Em grande parte, o diagnóstico é altamente dependente da experiência adquirida pelo profissional que o realiza. Apesar de indiscutivelmente necessário, não é perfeito, principalmente pelo fato de ser dependente de fatores humanos, por este motivo os profissionais sempre tiveram recursos tecnológicos para auxiliar nessa tarefa (BELUZO et al., 2020). Segundo estudos de 2010 no Brasil, calcula-se que aproximadamente 70% dos óbitos infantis ocorridos poderiam ter sido evitados por ações de saúde e que 60% dos óbitos neonatais ocorreram por situações que poderiam ser evitadas (BARRETO; SOUZA; CHAPMAN, 2015).

No presente trabalho foram utilizados dois algoritmos de aprendizado de máquina para criar modelos de predição de mortalidade neonatal. Além disso, foi realizada também uma análise exploratória da base de dados. Para este trabalho foram utilizadas duas fontes de dados: Sistema de Informação sobre Mortalidade (SIM) e Sistema de Informações sobre Nascidos Vivos (SINASC) do Brasil inteiro.

14

2 JUSTIFICATIVA

Com o avanço da tecnologia, podemos notar que ela está cada vez mais presente no nosso dia a dia e nos auxilia em diversas tarefas do cotidiano, e vem sendo aplicada em diversas áreas, dentre elas a saúde.

A mortalidade infantil é uma preocupação mundial na saúde pública, a ONU (Organizações das Nações Unidas) definiu a redução da mortalidade infantil como uma meta para o desenvolvimento global. A mortalidade neonatal é responsável por aproximadamente 60% da TMI nos países em desenvolvimento (A.K. SINGHA, 2016). Existem diversos fatores que podem contribuir com a redução de óbitos neonatais, como por exemplo acompanhamento médico à gestante no período de gravidez, acompanhamento médico ao recém-nascido nos primeiros dias de vida e a disponibilização deste serviço com qualidade e proficiência (WHO, 2009).

A diminuição dessa taxa é importante e de interesse para o mundo todo, e utilizar da tecnologia para auxiliar nessa diminuição é uma proposta funcional, e inovadora para a realidade brasileira (MOTA et al., 2019). Havendo disponibilidade de tecnologia para auxiliar nas decisões dos diagnósticos, os profissionais da saúde podem focar nos cuidados do recém-nascido com risco de vir a óbito, fornecendo tratamentos melhores.

Segundo o site Multiedro (2019), um grande problema que os médicos enfrentam ao realizar o diagnóstico, é analisar um grande volume de dados. Para a área médica obter diagnósticos rápidos e precisos pode salvar vidas, e a utilização de machine learning para realizar esses processos é importante, com a ML os dados podem ser processados em questões de segundos e resultar em um diagnóstico mais rápido, além disso utilizando bons modelos ML os diagnósticos serão mais precisos.

Diversos trabalhos vêm sendo desenvolvidos neste contexto. No trabalho realizado por BELUZO et al. (2020a), os autores utilizaram uma base de dados com informações da cidade de São Paulo para criação de modelos de aprendizado de máquina para predição de mortalidade neonatal. Este recorte foi utilizado também no presente trabalho para fins de exercício e definição de etapas da implementação de código. Na conclusão os autores discutem os fatores que tiveram mais influência nas predições, e na análise feita pelos autores, as consultas de pré-natal são muito importantes para a redução da mortalidade infantil, uma

vez que algumas características muito importantes, como a malformação congênita, podem ser detectadas ao longo das consultas de pré-natal.
Em um outro estudo realizado por BELUZO et al. (2020a), foi realizada uma análise exploratória da base de dados SPNeoDeath, onde e podemos observar detalhadamente cada
15
análise das colunas da tabela e diversos gráficos feito para um melhor entendimento dos dados.
Outro trabalho que foi realizado utilizando a base de dados com dados somente de São Paulo foi o de PELISSARI (2021), nesse trabalho é realizado uma análise exploratória na base de dados de São Paulo e é também analisado três algoritmos de aprendizado de máquina (Logistic Regression, Random Forest Classifier e XGBoost), nesse trabalho a autora conclui que os modelos de Logistic Regression e XGBoost foram os modelos que apresentaram os melhores desempenhos preditivos, e também mostra os problemas de estar utilizando uma base de dados desbalanceada.
O problema da mortalidade neonatal não é recente, e o mundo todo desenvolve iniciativas para lidar com ele. A ONU definiu como meta a redução da mortalidade infantil para o desenvolvimento global. Diversos fatores podem ajudar na diminuição da taxa de mortalidade neonatal como acompanhamento médico antes e após parto, tanto com a mãe quanto com o recém-nascido. Um serviço de qualidade e constante para os casos com maior risco de óbito pode salvar diversos recém-nascidos. Neste sentido, o desenvolvimento de ferramentas para a predição de mortalidade neonatal pode colaborar com esta iniciativa.

3 OBJETIVOS

3.1 OBJETIVO GERAL

O presente trabalho tem como objetivo testar e analisar os resultados da aplicação dos aprendizagem de máquina supervisionado Árvore de Decisão e Regressão Logística, utilizando dados do SIM e do SINASC, no período de 2016 até 2018, a fim de gerar modelos de predição de risco morte neonatal.

3.2 OBJETIVOS ESPECÍFICOS

? Realizar análise exploratória da base dados a ser utilizada na construção dos modelos;
? Implementar modelos de predição utilizando algoritmos Árvore de Decisão e Regressão Logística;
? Avaliar resultados e propor novas abordagens.
16

4 FUNDAMENTAÇÃO TEÓRICA

4.1 MORTALIDADE INFANTIL E NEONATAL

Como dito na monografia PELISSARI (2021), a TMI consiste no número de crianças que foram a óbito antes de concluir um ano de vida dividido por 1000 crianças nascidas vivas no tempo de um ano. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos e qualidade da saúde pública de um país (apud BELUZO et al., 2020).
Mencionado em PELISSARI (2021), a TMN é uma das duas categorias da TMI, é categorizado mortalidade neonatal quando o óbito ocorre do nascimento da criança até os 28 primeiros dias de vida. A mortalidade neonatal pode ser subdividida em mortalidade neonatal precoce que é quando o óbito ocorre entre 0 e 6 dias de vida, e o neonatal tardio quando o óbito ocorre entre 7 e 28 dias de vida (apud E. FRANÇA e S. LANSKY, 2009).

Segundo Lansky et al. (2014), a taxa de mortalidade neonatal é o principal componente da TMI desde a década de 1990 no Brasil e vem se mantendo em níveis elevados, com taxa de 11,2 óbitos por mil nascidos vivos em 2010 (apud Maranhão et al., 2012). Também é dito em Lansky et al. (2014), que a TMI do Brasil em 2011 foi 15,3 por mil nascidos vivos, alcançando a meta 4 dos objetivos de desenvolvimento do milênio, compromisso dos governos integrantes das Nações Unidas de melhorar a saúde infantil e reduzir em 2/3 a mortalidade infantil entre 1990 e 2015. (apud Maranhão et al., 2012; apud Murray et al., 2015). Segundo Lansky et al. (2014) o principal componente da TMI em 2009 é a mortalidade neonatal precoce, e grande parte das mortes infantis acontece nas primeiras 24 horas (25%), indicando uma relação estreita com a atenção ao parto e nascimento (apud E. FRANÇA e S. LANSKY, 2009).

4.2 MÉTODOS DE APRENDIZADO DE MÁQUINA

A utilização de dados para a resolução de problemas, de modo a se identificar padrões ocultos e posteriormente auxiliar na tomada de decisões é definida como aprendizado de máquina (BRESAN, 2018 apud Brink et al. 2013).

O uso de técnicas de Aprendizado de Máquina se mostra altamente eficiente na resolução de tarefas que se apresentem difíceis de se resolver a partir de programas escritos por humanos (BRESAN, 2018 apud GOODFELLOW et al., 2016), bem como também possibilita uma maior compreensão sobre os funcionamentos dos princípios que compõem a inteligência humana.

17

Neste trabalho serão implementados modelos utilizando os algoritmos de Árvore de Decisão e Regressão Logística, os quais serão apresentados a seguir.

4.2.1 Árvore de Decisão

Segundo Batista e Filho (2019), os modelos preditivos baseados em árvores são geralmente utilizados para tarefas de classificação, embora também possam ser utilizados para tarefas de regressão. Considerado um dos mais populares algoritmos de predição, o algoritmo de árvore de decisão proporciona uma grande facilidade de interpretação.

As árvores de decisão utilizam a estratégia ?dividir-e-conquistar", na qual as árvores são construídas utilizando-se apenas alguns atributos. A árvores de decisão é uma das técnicas por meio da qual um problema complexo é decomposto em subproblemas mais simples. Recursivamente, a mesma estratégia é aplicada a cada subproblema (SILVA et al, 2008).

A árvore de decisões tem uma estrutura hierárquica onde cada nó da árvore representa uma decisão em uma das colunas do conjunto de dados, cada ramo da árvore representa uma tomada de decisão (BATISTA; FILHO, 2019).

O algoritmo segue algumas decisões para montar a árvore, o atributo mais importante é utilizado como nó raiz e será o primeiro nó, já os atributos menos importantes são mostrados nos ramos da árvore. Cada atributo é verificado para conferir se é possível gerar uma árvore menor e se é possível fazer uma classificação melhor, e assim é escolhido o nó que produz nós filhos (SILVA et al, 2008).

Existem diferentes tipos de algoritmo para a construção da árvore, como por exemplo, ID3 (Iterative Dichotomiser), C4.5 e CART (Classification and Regression Tree). O algoritmo ID3 escolhe os nós da árvore por meio da métrica do ganho de informação. Já o CART faz uso da equação de Gini (BATISTA; FILHO, 2019 apud KUHN; JOHNSON, 2013).

O algoritmo utilizado neste trabalho usará o cálculo da entropia e ganho de informação, para decidir qual atributo será usado no nó, a entropia é uma medida da desorganização dos sistemas, maior é a incerteza do sistema, quanto maior a entropia maior está a desorganização, com a árvore de decisão a ideia é ir organizando o sistema e ir separando as decisões da melhor forma para que a entropia diminui e o ganho de informação aumente, para que a decisão do modelo fiquei mais assertiva. O cálculo da entropia utiliza a seguinte equação (ÁRVORE, 2020):

???????? =
?
?? ?? ???
2
??
18

Nesta equação o i representa possíveis classificações (rótulos), e o P é a probabilidade de cada uma. A ideia da árvore então é verificar qual é a entropia para cada uma das variáveis da base de dados, e verificar qual é a melhor organização de cada uma das variáveis. E isso é realizado através da técnica chamada de ganho de informação, essa técnica utiliza a equação a seguir (ÁRVORE, 2020):

????? = ????????(???) ? ? ????(??????) * ????????(??????)

Então o ganho de informação é calculado pela entropia do nó pai menos a soma do peso vezes a entropia dos nós filhos. Esse cálculo é realizado para todas as variáveis e a variável que tiver maior ganho de informação é utilizado para a decisão do nó. Esse processo é realizado recursivamente e a árvore vai sendo montada com base nesses cálculos (ÁRVORE, 2020).

Na Figura 1 temos a representação do início de uma árvore de decisão, no caso em questão o autor estava classificando exames positivos e negativos para o vírus SARS-CoV-2, ele utilizou uma base de dados que contém quase 6 mil exames realizados, contendo campos como o resultado do exame, idade do paciente, níveis de hemoglobina entre outras informações.

Figura 1. Exemplo de Árvore de Decisão. (Fonte: ÁRVORE 2020).

Pode-se observar que para o nó raiz foi utilizado a variável Leukocytes, logo foi o que apresentou a melhor organização para ser o nó raiz, e após ele temos os ramos da árvore. Cada retângulo da árvore é uma decisão do modelo e esses retângulos têm atributos importante como, a coluna da base de dados que será onde vai ser realizado a decisão, então pegando o nó raiz, pacientes que tiverem com os Leukocytes para menos que -0.43 seguiram o caminho
19
para a esquerda (true), já os pacientes que tiverem mais seguiram o caminho da direita (false). A entropia calculada do nó também é apresentada e a classificação do nó também, então se o resultado parar neste nó, a classificação que está nele será a classificação final, e temos o samples que é o número de amostras que se enquadram no na decisão. E esse processo de verificação do modelo vai se repetindo até não ter mais como dividir em subproblemas ou até chegar na profundidade máxima.

Uma característica importante da árvore de decisão é que ela é um modelo WhiteBox, ou seja, é possível visualizar a árvore montada e as decisões que o modelo terá que tomar na

árvore. Na Figura 1 pode-se ver uma árvore de decisões montada, cada retângulo da árvore é uma decisão que o modelo terá que tomar, conforme o modelo toma as decisões ele vai seguindo um caminho até chegar ao nó folha, nó folha é o nó que não contém nó filho, não tem mais ramos para baixo, chegando ao nó folha o modelo tem a classificação final.

4.2.2 Regressão Logística

O modelo de regressão logística estabelece uma relação entre a probabilidade de ocorrência da variável de interesse e as variáveis de entrado do modelo, sendo utilizada para tarefas de classificação, em que a variável de interesse é categórica, neste caso, a variável de interesse assume valores 0 ou 1, onde 1 é geralmente utilizado para indicar a ocorrência do evento de interesse (Batista e Filho, 2019).

De acordo com Mesquita (2014), a técnica de regressão logística, desenvolvida no século XIX, obteve maior visibilidade após 1950 ficando então mais conhecida. Ficou ainda mais difundida a partir dos trabalhos de Cox & Snell (1989) e Hosmer & Lemeshow (2000). Caracteriza-se por descrever a relação entre uma variável dependente qualitativa binária, associada a um conjunto de variáveis independentes qualitativas ou métricas.

Esta técnica inicialmente foi utilizada na área médica, porém a eficiência viabilizou sua implementação nas mais diversas áreas. Mesquita (2014) diz que o termo regressão logístico, tem sua origem na transformação usada com a variável dependente, que permite calcular diretamente a probabilidade da ocorrência do fenômeno em estudo.

Segundo Santos (2018), para estimar a probabilidade, é utilizado uma técnica chamada distribuição binomial para modelar a variável resposta do conjunto de treinamento, que tem como parâmetro a probabilidade de ocorrência de uma classe específica. Uma característica importante desse modelo é que a probabilidade estimada deve estar limitada ao intervalo de 0 a 1. O algoritmo de Regressão Logística gera um modelo de classificação binária (0 e 1). O modelo utiliza a Regressão Linear como base, a equação linear é (REGRESSÃO, 2020):

20

? = ? . ? + ?

O ?m? da equação é o coeficiente angular da reta, e o ?b? é o intercepto. Então, aplicando a equação linear obtemos um valor contínuo como resultado, após obter esse valor a Regressão Logística tem que realizar a classificação do resultado, então é aplicado uma função chamada sigmoide (REGRESSÃO, 2020):

? = 1

1+ ? ??

Essa função chamada de sigmoide, ou função logística, é responsável por ?achatar? o resultado da regressão linear, calculando a probabilidade de o resultado pertencer a classe 1, classificando assim o resultado da função linear em 0 ou 1.

Segundo Batista e Filho (2019), o modelo de regressão logística que tem como objetivo realizar uma classificação binária de uma variável de interesse irá predizer a probabilidade de pertencer à classe positiva. Tem-se, portanto, que é dado por:?

? = { 0, ?? ???? (? = 1 ??) < 0, 5

ou

? = { 1, ?? ???? (? = 1??) ? 0, 5

Então a decisão do modelo de regressão logística está relacionada à escolha de um ponto de corte para p(x). Caso o ponto de corte escolhido for p(x)=0,5, os resultados que

forem p(x)>0,5 serão classificados como 1 e os resultados p(x)<5 serão classificados como 0 (SANTOS, 2018).

A Figura 2 ilustra um exemplo de como é o diagrama de Regressão Logística. Pelo diagrama então pode-se observar que o modelo possui uma representação gráfica em formato de ?S?, essa seria a curva logística. As predições do modelo sempre ficaram na linha 1 e 0 do eixo y, pois é o intervalo estabelecido pelo algoritmo, então a classificação sempre será nesse intervalo.

21

Figura 2. Exemplo de Diagrama de Regressão Linear. (Fonte: Batista e Filho 2019).

4.2.3 Treino, teste e validação do modelo de aprendizado de máquina

A criação de um modelo de aprendizado de máquina é realizada em duas etapas: treinamento do modelo, que é realizado a partir dos dados fornecidos para o algoritmo, e etapa de teste, onde os modelos recebem dados novos e sem o rótulo, para que seja avaliado a performance do modelo (PELISSARI, 2021).

Para o treinamento dos modelos foram utilizadas duas abordagens, a primeira onde a base foi particionada entre conjunto de dados para teste e para treino em uma porcentagem de 90% para treino e 10% para teste, porém dessa forma pode ocorrer problema de sobreajuste ou overfitting. Nesta situação, o que pode ocorrer é o modelo não aprender com os dados, e apenas ?decorar? os dados recebidos e quando é realizado o teste o modelo consegue predizer apenas situações parecidas, não sendo capaz de identificar padrões com diferenças mínimas.

Para evitar este problema, foi utilizado uma técnica de validação cruzada chamada K-folds cross-validation. Essa técnica de validação cruzada divide a base de dados em K subconjuntos, e então são realizadas várias rodadas de treinamento e teste alternando os subconjuntos utilizados para teste e treino, e o resultado do modelo baseia-se na média da acurácia observada em cada rodada (PELISSARI, 2021).

Pode-se observar na Tabela 1 uma ilustração da técnica K-fold cross-validation, onde temos um exemplo onde a base é dividida em 5 subconjuntos e cada subconjunto tem a parte de teste diferente dos outros subconjuntos, assim o modelo vai receber valores novos de teste e treino todas as vezes que executar um novo subconjunto.

22

Tabela 1. Exemplo da separação K-fold cross-validation. (Fonte: Adaptado de PELISSARI, 2021).

Predição 1	Predição 2	Predição 3	Predição 4	Predição 5
Teste	Treino	Treino	Treino	Treino
Treino	Teste	Treino	Treino	Treino
Treino	Treino	Teste	Treino	Treino
Treino	Treino	Treino	Teste	Treino
Treino	Treino	Treino	Treino	Teste

4.2.4 Métricas para avaliação de modelos

Para fins de avaliação, neste trabalho foram utilizadas três métricas para avaliar o desempenho dos modelos: a Matriz de Confusão e a AUC (Area under Receiver Operating Characteristic Curve - ROC). Na matriz de confusão pode-se observar o comportamento do modelo em relação a cada uma das classes a serem preditas pelo modelo, utilizando variáveis binárias (positivo: 1; e negativo: 0), para apresentar uma matriz que faz uma comparação

entre as predições do modelo e os valores corretos do conjunto de dados (PELISSARI, 2021).
Na Tabela 2 temos um exemplo de uma matriz de confusão, a qual consiste em duas colunas e duas linhas, e os valores são atribuídos nos quadrantes com os seguintes resultados (PELISSARI, 2021):
? quando o valor real do conjunto de dados é 0, e a predição do modelo também classificou como 0, então temos um Verdadeiro Negativo (VN);
? quando o valor real é 1, e o modelo classificado como 0, temos um Falso Negativo (FN);
? quando o valor real é 0, e o modelo classificado como 1, então o resultado se enquadra no quadrante de Falso Positivo (FP);
? e por fim o quadrante Verdadeiro Positivo (VP), que são os resultados que tem o valor real 1, e o modelo classificado como 1.
23
Tabela 2. Modelo de Matriz de confusão. (Fonte: Adaptado de PELISSARI, 2021).
Valor Predito
Negativo (0) Positivo (1)
Classe
Real
Negativo
(0)
Verdadeiro
Negativo
Falso
Positivo
Positivo
(1)
Falso
Negativo
Verdadeiro
Positivo
A segunda métrica de avaliação utilizada foi a Curva ROC que permite avaliar o desempenho de um modelo com relação às predições efetuadas. A curva ROC é um gráfico de Sensibilidade (taxa de verdadeiros positivos) versus taxa de falsos positivos, a representação da curva ROC permite evidenciar os valores para os quais existe otimização da Sensibilidade em função da Especificidade (CABRAL, 2013).
Na Figura 3 temos um exemplo de uma curva ROC. A linha traçada na cor preta é uma linha de referência, ela representa hipoteticamente a curva ROC de um classificador puramente aleatório. Caso a linha laranja, que representa o resultado do modelo que está sendo avaliado, ficasse igual a linha tracejada preta, indicaria que o modelo avaliado estaria predizendo aleatoriamente os resultados.
Figura 3: Exemplo de Curva ROC. (Fonte: Elaboração Própria).
Por uma análise de inspeção visual, quanto mais próximo a linha laranja estiver do valor 1 no eixo Y, melhor está a capacidade do modelo para diferenciar as classes. O valor da
24

AUC representa a capacidade do modelo de diferenciar as classes, quanto maior o valor do AUC, melhor é a predição realizada pelo modelo, uma vez que, os valores resultantes iguais a 0 são realmente 0, e os iguais a 1 são de fato 1 (PELISSARI, 2021).

5 METODOLOGIA

O desenvolvimento deste trabalho foi realizado em três etapas: (1) Pré-processamento e Análise Exploratória do conjunto de dados, onde foram aplicadas técnicas comuns para preparar o conjunto de dados como tratamento de valores nulos e seleção de variáveis de interesse; (2) Implementação de modelos de aprendizado de máquina supervisionado para predição das mortes neonatais, utilizando os algoritmos de árvore de decisão e regressão logística; e (3) Análise de Resultados, que será realizada uma análise do desempenho do modelo.

5.1 TECNOLOGIAS E FERRAMENTAS

Neste trabalho vamos utilizar a linguagem de programação Python (PYTHON, 2021) pela sua simplicidade de codificação e alto poder de processamento. Foi utilizado também a plataforma Google Colab Research (Colab, 2021), para o desenvolvimento dos algoritmos e treinamento dos modelos. Além disso, foi utilizado também o Jupyter Notebook (Jupyter, 2021), instalado localmente em computador pessoal, pois o Google Colab Research possui limitação de recursos de memória e processamento, então as duas plataformas foram usadas para realização deste trabalho. As principais bibliotecas utilizadas foram: numpy (NumPy, 2021), e pandas (Pandas, 2021), para a manusear os dados, matplotlib (Matplotlib, 2021), e seaborn (Seaborn, 2021), para a plotagem dos gráficos, e por fim a biblioteca sklearn (Scikit-Learn, 2021), que é a biblioteca responsável pelos algoritmos de aprendizado de máquina.

5.2 BASE DE DADOS

As bases de dados a serem utilizadas serão o SIM e SINASC, que são as duas principais fontes de informações sobre nascimentos e óbitos no Brasil. O SINASC é alimentado com base na Declaração de Nascido Vivo (Declaração de Nascido Vivo - DNV) (BELUZO et al.,2020b apud Oliveira et al., 2015). O SIM tem como objetivo principal apoiar a coleta, armazenamento e processo de gestão de registros de óbitos no Brasil (BELUZO et

25

al.,2020b apud Morais et al., 2017), e foi usado para rotular o óbito registrado no SIM, utilizando o campo DNV como chave de associação. O SIM será utilizado para rotular os registros do SINASC, pois o SIM coletas informações sobre mortalidade e é usado como base para o cálculo de estatísticas vitais, como a taxa de mortalidade neonatal e o SINASC reúne informações sobre dados demográficos e epidemiológicos do bebê, da mãe, do pré-natal e do parto (BELUZO et al.,2020b).

A Tabela 3 apresenta as variáveis da base de dados. Essa base tem variáveis que contêm valores quantitativos e categóricos. A coluna ?num_live_births? por exemplo contém o número de nascidos vivos anteriores da mãe e é composta por valores quantitativos contínuos, e a coluna ?tp_pregnancy? utiliza valores nominais categóricos que indicam o tipo de gravidez.

Tabela 3. Colunas da base de dados e suas descrições. (Fonte: Adaptado de BELUZO et al.,2020b).

Nome da coluna	Descrição	Domínio de dados

Variáveis demográficas e socioeconômicas
maternal_age Idade da mãe Quantitativo Contínuo (inteiro)
tp_maternal_race Raça / cor da pele da
mãe
Nominal categórico (inteiro)
1 - branco; 2 - Preto; 3 -
amarelo; 4 - Pele morena; 5 -
Indígena.
tp_marital_status Estado civil da mãe Nominal categórico (inteiro)
1 - Único; 2 - Casado; 3 - Viúva;
4 - Separados / divorciados
judicialmente; 5 - Casamento
por união estável; 9 - Ignorado.
tp_maternal_schooling Anos de escolaridade da
mãe
Nominal categórico (inteiro)
1 - nenhum; 2 - de 1 a 3 anos; 3 -
de 4 a 7 anos; 4 - de 8 a 11 anos;
5 - 12 e mais; 9 - Ignorado.
Variáveis obstétricas maternas
num_live_births Número de nascidos
vivos
Quantitativo Contínuo (inteiro)
num_fetal_losses Número de perdas fetais Quantitativo Contínuo (inteiro)
num_previous_gestations Número de gestações Quantitativo Contínuo (inteiro)
26
anteriores
num_normal_labors Número de partos
normais (trabalhos de
parto)
Quantitativo Contínuo (inteiro)
num_cesarean_labors Número de partos
cesáreos (partos)
Quantitativo Contínuo (inteiro)
tp_pregnancy Tipo de gravidez Nominal categórico (inteiro)
1 - Singleton; 2 - Twin; 3 -
Trigêmeo ou mais; 9 - Ignorado.
Variáveis relacionadas a cuidados anteriores
tp_labor Tipo de parto infantil
(tipo de parto)
Categórico Nominal (inteiro)
1 - Vaginal; 2 - Cesariana;
num_prenatal_appointments Número de consultas de
pré-natal

Quantitativo Contínuo (inteiro)
tp_robson_group Classificação do grupo
Robson
Ordinal categórico (inteiro)
Variáveis relacionadas ao recém-nascido
tp_newborn_presentation Tipo de apresentação de
recém-nascido
Categórico Nominal (inteiro)
1 - Cefálico; 2 - Pélvico ou
culatra; 3 - Transversal; 9 -
Ignorado.
has_congenital_malformation Presença de
malformação congênita
Nominal categórico (inteiro)
1 - Sim; 2 - Não; 9 - Ignorado
newborn_weight Peso ao nascer em
gramas
Quantitativo Contínuo (inteiro)
cd_apgar1 Pontuação de Apgar de 1
minuto
Ordinal categórico (inteiro)
cd_apgar5 Pontuação de Apgar de 5
minuto
Ordinal categórico (inteiro)
gestaional_week Semana de gestação Quantitativo Contínuo (inteiro)
tp_childbirth_care Assistência ao parto Categórico Nominal (inteiro)
1 - Doutor; 2 - enfermeira ou
obstetra; 3 - Parteira; 4 - outros;
9 - Ignorado.
was_cesarean_before_labor Foi cesáreo antes do
parto.
27
was_labor_induced Foi induzido ao trabalho
de parto.
is_neonatal_death Morte antes de 28 dias
(rótulo)
Nominal categórico (inteiro)
0 - sobrevivente; 1 - morto.

Neste trabalho foi utilizado dados do período de 2014 à 2016 devido sua melhor qualidade. Este recorte possui 6.760.222 registros dos quais 99.4% são amostras de recém-nascidos vivos e 0.6% são amostras onde os recém-nascidos vieram a óbito conforme pode ser observado na Figura 4. Com essas informações consegue-se observar que a base de dados é desbalanceada.

Figura 4. Distribuição da base de dados considerando se o recém-nascido viveu ou

morreu durante o período neonatal. (Fonte: Elaboração Própria).

5.3 PREPARAÇÃO DA BASE DE DADOS

Nessa etapa foi realizada a preparação da base de dados para o treinamento dos modelos que consiste na limpeza, transformação e preparação dos dados a serem utilizados na análise exploratória e na criação dos modelos preditivos. A base de dados pode conter informações desnecessárias para o modelo que pode acabar atrapalhando as predições,

28

algumas colunas podem ser retiradas ou seus valores podem ser modificados, por exemplo valores reais são modificados para inteiros.

5.4 DESENVOLVIMENTO DOS MODELOS DE APRENDIZADO DE MÁQUINA

Como já dito anteriormente na seção 4 os algoritmos de aprendizado de máquina escolhidos para esse trabalho foram os: Regressão Logística e Árvore de Decisão que utilizam a lógica mostrada na seção 4 , Fundamentação Teórica. Esses dois foram escolhidos por alguns motivos, ambos são algoritmos de classificação e supervisionados, atendendo os critérios necessários para a predição de mortalidade neonatal, ambos os algoritmos são bons para realizar predições, a Árvore de Decisão inclusive é um ótimo modelo para analisar as decisões que estão sendo feitas pelo modelo, afinal esse algoritmo tem a característica de ser um WhiteBox, ou seja conseguimos ver o que o modelo aprendeu e o que ele está decidindo. Esses algoritmos selecionados são algoritmos que se encaixam na necessidade deste trabalho e são muito utilizados na comunidade acadêmica.

Para o algoritmo de Regressão Logística, como foi dito anteriormente na seção ?Preparação da base de dados? a base de dados foi tratada e particionada, no particionamento foi utilizado 90% da base para o treino e 10% para os testes e então foi realizado o treinamento do modelo, também foi aplicada a função RFECV (Eliminação Recursiva de Feature com Validação Cruzada), esse função executa a RFE (Eliminação Recursiva de Feature), que tem como ideia selecionar features considerando recursivamente conjuntos cada vez menores de features, isso ocorre da seguinte forma. Primeiro, o estimador é treinado no conjunto inicial de features e a importância de cada feature é obtida por meio de um atributo "coef" ou por meio de um atributo "feature_importances". Em seguida, as features menos importantes são removidas do conjunto atual de features. Esse procedimento é repetido recursivamente no conjunto removido até que o número desejado de features a serem selecionados seja finalmente alcançado. Porém o diferencial da RFECV é que essa função executa a RFE em um loop de validação cruzada para encontrar o número ideal ou o melhor número de features.

Após essa seleção, foi realizado o treinamento do modelo usando todas as features e também treinamos usando as features que o RFECV selecionou, para compararmos os resultados no final. Também foi utilizado o método de K-folds cross-validation em um novo treinamento para conferirmos se o modelo estava sofrendo overfitting (sobreajuste). E por fim aplicamos um método chamado GridSearchCV, esse método permite que seja realizado testes

29

alterando algumas combinação de parâmetros nos nossos modelos, facilitando para achar a melhor combinação, esse método é parecido com o cross validation, o diferencial do Grid Search é que ele faz os cross validation de vários modelos com hiperparâmetros diferentes de uma só vez.

Para o algoritmo de Árvore de Decisão também foi utilizado o mesmo particionamento de 90% para treino e 10% para teste, na criação do modelo foi colocado a entropy como critério de decisão e uma profundidade máxima de 4, pois com números maiores o modelo ficava muito complexo e ele errava mais as predições, também utilizamos um algoritmo para mostrar a importância de cada Feature para o modelo.
O algoritmo de Regressão Logística e análise exploratória foi baseado em códigos disponível na plataforma web Kaggle (MNASSRI, 2020), no exemplo do Kaggle o autor faz os códigos para predizer mortes no navio Titanic, para o trabalho de mortalidade neonatal os códigos foram alterados para realizar predição de mortes neonatais. Já no algoritmo de Árvore de Decisão foi baseado em uma vídeo aula do professor Diogo Cortiz (ÁRVORE..., 2020), nessa vídeo aula o professor explica sobre os algoritmos de Árvore de Decisão e depois implementa o um exemplo de código, o código usado neste trabalho usou o codigo do professor Diogo Cortiz, e foi alterado para realizar predições de mortalidade neonatal.

5.5 ANÁLISE DOS RESULTADOS

Para a análise de resultados foi utilizado dois métodos que é o de Matriz de confusão e a curva ROC esses métodos são explicados na seção 4, a Fundamentação Teórica. Com esses métodos pode-se realizar uma avaliação do modelo, e assim será possível analisar os valores de acurácia, precisão e recall do modelo, essas métricas são utilizadas avaliar o desempenho do modelo, com a acurácia é possível calcular a proximidade entre o valor obtido na predição dos modelos e os valores esperados, com a precisão é possível analisar as amostras que o modelo classificou como 0 eram realmente 0 na classe real e com o recall é possível analisar as amostras que o modelo conseguiu reconhecer como a classe correta.

5.6 ACESSO A BASE DE DADOS E CÓDIGOS

A base de dados utilizada neste projeto pode ser encontrada no link: ... . E os códigos podem ser encontrados no link: ... .

30

6 RESULTADOS

Os mesmos experimentos mostrados abaixo foram aplicados para uma base de dados que contém somente dados de São Paulo, essa base é um recorte da base do Brasil todo em contém cerca de 1.400.000 amostras, esse recorte da base também é bem desbalanceado. Os resultados obtidos nesse experimento usando o recorte de São Paulo foram bem parecidos com os experimentos da base do Brasil todo, e os códigos desse experimento podem ser visualizados no apêndice X.

6.1 ANÁLISE EXPLORATÓRIA

O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

6.1.1 Apresentação dos Dados

Como dito anteriormente na seção 5.2 ?Base de Dados? deste trabalho, a base de dados é formada pelas informações do SIM e do SINASC contém 6.760.222 amostras e 29 colunas, porém serão utilizadas somente 23 colunas sendo elas as colunas descritas na Tabela 3. Na seção 5.2 também tem a Figura 4 que mostra que a base dados é desbalanceada tendo 99.4% das amostras de recém-nascidos vivos e 0.6% das amostras onde os recém-nascidos vieram a óbito no período neonatal.

6.1.1.1 Caracteristicas demográficas e socioeconômicas maternas

O apêndice X contém uma tabela que possui as estatísticas sumarizadas das variáveis demográficas e socioeconômicas maternas. Nesta tabela por exemplo podemos observar a coluna ?maternal_age? que contém a idade da mãe, e na tabela mostra que a média da idade das mães é de 26.4 anos, e os dados possuem uma variação de no mínimo 8 anos e no máximo 55 anos, e a mediana é 26 anos. Na tabela também tem as colunas: ?tp_maternal_schooling? que mostra a categoria de anos de escolaridade da mãe, ?tp_marital_status? que mostra o estado civil da mãe em dados categóricos e a coluna ?tp_maternal_race? que mostra a raça da mãe também em dados categóricos.

31

6.1.1.2 Variávies obstétricas maternas

No apêndice X pode-se observar uma tabela que possui as estatísticas sumarizadas das variáveis obstétricas maternas. Nesta tabela por exemplo temos a coluna ?num_live_births?, essa coluna mostra o número de nascidos vivos anteriores que a mãe obteve, a média desta coluna é 0.94 , a mediana é 1, o número mínimo é 0 e o máximo é 10 nascidos vivos. Na tabela podemos observar também a coluna, ?num_fetal_losses?, esta coluna mostra o número de perdas fetais anteriores da mãe, a média desta coluna é 0.21, a medina é 0, o número mínimo é 0 e o máximo é 5, esses valores mostram que poucas mães tiveram perdas fetais antes da gravidez atual. Também na tabela tem estatísticas da coluna ?num_previous_gestations? esta coluna mostra o número de gestações anteriores da mãe, esta coluna tem a média de 1.14, a mediana é 1, o número mínimo é 0 e o máximo 10. Essa tabela também contém as colunas: ?num_normal_labors? que é o número de partos normais da mãe, ?num_cesarean_labor? que mostra o número de partos cesáreos da mãe e por fim mostra a coluna ?tp_pregnancy? que é o tipo da gravidez.

6.1.1.3 Variávies de histórico de gravidez

No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis do histórico de gravidez. Nesta tabela temos as colunas ?num_prenatal_appointments? que mostra o número de consultas de pré-natal, a média dessa coluna é 7.8, a mediana é 8, a variação dos dados é de no mínimo 0 e no máximo 40. A tabela também contém as colunas: ?tp_labor? que mostra o tipo de parto, e também contém a coluna ?tp_robson_group? que mostra a classificação do grupo Robson.

6.1.1.4 Variávies relacionadas ao recém-nascido

No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis relacionadas ao recém-nascido. Nesta tabela por exemplo podemos observar a coluna ?newborn_weight? que armazena o peso ao nascer dos recém-nascidos em gramas, a média dos dados dessa coluna é 3187.3, a mediana é 3215, a variação dos dados é de no mínimo 0 e no máximo 6000 gramas. Também podemos observar a coluna ?gestacional_week? que mostra o número de semanas de gestação, a média é 38.4, a mediana é 39, a variação dos dados é de no mínimo 15 e no máximo 45 semanas. Além dessas colunas a tabela também

32

contém as colunas: ?cd_apgar1? que mostra a pontuação de Apgar de 1 minuto, ?cd_apgar5? que mostra a pontuação de Apgar de 5 minuto, ?has_congenital_malformation? que mostra se o recém-nascido contém a presença de alguma malformação, ?tp_newborn_presentation? que mostra o tipo de apresentação de recém-nascido, ?tp_labor? que mostra o tipo de parto, ?was_cesarean_before_labor? que mostra se o recém-nascido foi cesáreo antes do parto,

?was_labor_induced ? que mostra se o recém-nascido foi induzido ao trabalho de parto, ?tp_childbirth_care? que mostra se ouve assistência ao parto, e por fim temos a coluna ?is_neonatal_death? que mostra se o recém-nascido veio a óbito ou não.

6.1.2 Análise das variáveis

Nesta seção serão mostrados a parte de análise de variáveis com diferentes tipos de gráficos.

6.1.2.1 Variáveis contínuas

Na Figura 5 temos dois gráficos boxplot, no boxplot à esquerda temos a distribuição da variável peso ao nascer e nele é possível observar duas caixas onde a caixa azul são os vivos e o laranja são os mortos. Nas duas caixas mostradas no gráfico podemos observar pequenos círculos nas pontas, esses círculos são outliers (dados fora da curva), então para a caixa de vivos os outliers são recém-nascidos com pesos acima de 4500 gramas ou abaixo de 2000 gramas, e para a caixa de mortos os outliers são os recém-nascidos com mais de 5500 gramas. Para os vivos temos a mediana de aproximadamente 3200 gramas e para os mortos a mediana é 1300 gramas. Cada caixa representa 50% da base de dados, a caixa azul de vivos mostra que o peso de recém-nascidos que sobreviveram está aproximadamente entre 2700 a 3600 gramas, e para a caixa laranja de mortos o peso de recém-nascidos que vieram a óbito estão aproximadamente entre 700 a 2500 gramas. Então o gráfico mostra para nós que os recém-nascidos que têm maior risco de vir a óbito tem pesos menores que 2000 gramas.

33

Figura 5. BoxPlot das features Peso ao nascer e Idade da mãe. (Fonte: Elaboração Própria).

No gráfico à direita da Figura 5 temos um boxplot da distribuição da variável idade da mãe, como foi dito anteriormente a caixa azul são os vivos e o laranja são os mortos. Nesse boxplot então podemos ver que os outliers da caixa de vivos são mães com idade acima de 46 anos aproximadamente, para os mortos os outliers são mães com idade acima de 50 anos. Para os vivos temos a mediana de aproximadamente 26 anos e para os mortos a mediana é de 26 anos. Nos vivos a idade das mães está aproximadamente entre 21 a 32 anos, e para os mortos a idade das mães está aproximadamente entre 20 a 34 anos. É importante ressaltar que o gráfico mostra as duas caixas bem parecidas, então de acordo com os dados não contém diferença significativa entre vivos e mortos usando a idade da mãe para distribuir.

Já na Figura 6 temos um boxplot da distribuição da variável semanas de gestação. Nesse boxplot então podemos ver que os outliers da caixa de vivos são gestações com mais de 44 semanas aproximadamente, e gestações com menos de 35 semanas. Para os vivos temos a mediana de aproximadamente 39 semanas e para os mortos a mediana é de 32 semanas. Nos vivos as semanas de gestação estão aproximadamente entre 36 a 40 semanas, e para os mortos a semanas de gestação estão aproximadamente entre 26 a 36 semanas. Nesse boxplot os dados mostraram que para as gestações que têm menos de 36 semanas os recém-nascidos têm maior risco de vir a óbito.

34

Figura 6. BoxPlot da feature semana de gestação. (Fonte: Elaboração Própria).

Na Figura 7 temos um histograma da distribuição de peso ao nascer por classe, podemos observar no gráfico que grande parte dos vivos (linha azul) estão distribuídos entre 2000 e 4000 gramas, a área entre esses valores tem muitos dados de vivos, já entre os valores 0 e 2000 gramas temos uma concentração dos dados de mortos.

Figura 7. Histograma de distribuição do peso entre as classes. (Fonte: Elaboração Própria)

6.1.2.2 Variáveis categoricas

Observando a Figura 8 podemos ver um gráfico de barras da distribuição da variável escolaridade da mãe, este gráfico mostra a contagem de registros por escolaridade da mãe, esse gráfico mostra para nós que a maior parte das mães estão na categoria 4 que representa de 8 a 11 anos de escolaridade.

35

Figura 8. Gráfico de barras da distribuição da variável Escolaridade da mãe. (Fonte: Elaboração Própria).

Observando a Figura 9 podemos ver um gráfico de barras da distribuição da variável número de vivos, este gráfico mostra a contagem de registros por número de vivos, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho.

Figura 9. Gráfico de barras da distribuição da variável Número de nascidos vivos. (Fonte: Elaboração Própria).

Observando a Figura 10 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 1 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 1 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem

36

apgar de 8 ou 9, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.

Figura 10. Gráfico de barras da distribuição da variável Pontuação Apgar de 1 minuto. (Fonte: Elaboração Própria).

Observando a Figura 11 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 5 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 5 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem apgar de 9 ou 10, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.

Figura 11. Gráfico de barras da distribuição da variável Pontuação Apgar de 5 minutos. (Fonte: Elaboração Própria).

37

Observando a Figura 12 podemos ver um gráfico de barras da distribuição da variável presença de malformação congênita, este gráfico mostra a contagem de registros por presença de malformação congênita, esse gráfico mostra para nós que a grande maioria dos dados estão na categoria 2 que mostra que o recém-nascido não possui malformação, esse grande volume para a categoria 2 é causada pelo desbalanceamento da base.

Figura 12. Gráfico de barras da distribuição da variável Presença de malformação congênita. (Fonte: Elaboração Própria).

Observando a Figura 13 podemos ver um gráfico de barras da distribuição da variável número de gestações, este gráfico mostra a contagem de registros por número de gestações, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho, da mesma forma mostrada na Figura 9.

38

Figura 13. Gráfico de barras da distribuição da variável Número de gestações. (Fonte:

Elaboração Própria).
Observando a Figura 14 podemos ver um gráfico de barras da distribuição da variável assistência ao parto, este gráfico mostra a contagem de registros por assistência ao parto, esse gráfico mostra para nós que a grande maioria dos dados mostram que o parto teve assistência de uma Enfermeira ou obstetra ou essa informação foi ignorada na hora do resgistro.
Figura 14. Gráfico de barras da distribuição da variável Assistência ao parto. (Fonte: Elaboração Própria).
6.1.2.3 Analise bi-variada
A Figura 15 mostra para nós a importância da feature peso ao nascer com o gráfico mostrado é possível observar claramente a diferença de pesos entre vivos e mortos, recém-nascidos com peso acima de 2000 gramas sobreviveram, já os recém-nascidos com menos de 2000 gramas vieram a óbito.
39
Figura 15. Gráfico de densidade de peso entre as classes. (Fonte: Elaboração Própria).
Na Figura 16 podemos observar um gráfico de correlação entre algumas features de valores contínuos, algumas dessas features possuem correlações positivas, como pontuação apgar de 1 minuto e pontuação apgar de 5 minutos, essas colunas possuem uma correlação de 0.7, então nascidos que recebem um valor alto de apgar de 1 minuto tendem a receber um valor alto no apgar de 5 minutos. O gráfico também mostra outras features com correlações positivas como, número de partos normais e número de gestações anteriores que contém uma correlação de 0.81, número de gestações anteriores com idade da mãe que tem uma correção de 0.39, número de partos de cesáreos com idade da mãe que possui correlação de 0.24, número de partos normais com idade da mãe com a correlação de 0.26, semanas de gestação com peso ao nascer em gramas com correção de 0.52, e as apgar de 1 e 5 minutos com semanas de gestação. E grande parte das features, possuem correlações baixas então elas são inversamente correlacionadas pelo que o gráfico mostra, como por exemplo número de partos normais com número de consultas pré-natais possuem uma correção de -0.17, e número de partos cesáreos com número de partos normais que contém uma correlação de -0.15. Então a correlação dessas features estão bem baixas tirando as correlações acima de 0.7.
40
Figura 16. Gráfico de correlação. (Fonte: Elaboração Própria)
6.1.2.4 Distribuição por raça
Na Figura 17 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por raça, e podemos observar que a raça 5 (indigena), é o boxplot que contem a menor variação de idade, já as raças 1 (branco) e 3 (amarelo) são as caixas que possuem maior variação com a média em torno de 20 e 30 anos.
41
Figura 17. Distribuição da idade da mãe por raça. (Fonte: Elaboração Própria).
No boxplot mostrado na Figura 18 podemos observar a distribuição de peso ao nascer por raça, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre as raças para peso ao nascer.
Figura 18. Distribuição de peso ao nascer por raça. (Fonte: Elaboração Própria).
Na Figura 19 temos o boxplot da a distribuição de semanas de gestação por raça, e praticamente todas as caixas são iguais, somente a caixa da raça 1 (branco) possui menor

variação que as outras.
42
Figura 19. Distribuição de semanas de gestação por raça. (Fonte: Elaboração Própria).
6.1.2.5 Distribuição por estado civil
A Figura 20 mostra o boxplot da a distribuição de peso ao nascer por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre os estados civis para peso ao nascer.
Figura 20. Distribuição de peso ao nascer por estado civil. (Fonte: Elaboração Própria).
A Figura 21 mostra o boxplot da a distribuição de semanas de gestação por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Porém os estados civis 2 (Casado) e 4 (Separados/divorciados judicialmente), contém variações menores.
43
Figura 21. Distribuição de semanas de gestação por estado civil. (Fonte: Elaboração Própria).
Na Figura 22 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por estado civil, e podemos observar que o estado civil 3 (Viúva), possui a maior variação, já o estado civil 4 (Separados/divorciados judicialmente) é a caixa que possuem menor variação
Figura 22. Distribuição da idade da mãe por estado civil. (Fonte: Elaboração Própria).
6.1.2.6 Distribuição por escolaridade da mãe
Na Figura 23 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por escolaridade da mãe, e podemos observar que a escolaridade 2 (1 a 3 anos) e 3 (4 a 7
44
anos), possui as maiores variações, já escolaridade 5 (12 ou mais anos) é a caixa que possuem menor variação
Figura 23. Distribuição da idade da mãe por escolaridade da mãe. (Fonte: Elaboração Própria).
A Figura 24 mostra o boxplot da a distribuição de peso ao nascer por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então, pouca variação é observada entre a escolaridade da mãe para peso ao nascer.
Figura 24. Distribuição de peso ao nascer por escolaridade da mãe. (Fonte: Elaboração Própria).
A Figura 25 mostra o boxplot da a distribuição de semanas de gestação por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas,
45
todas são praticamente iguais. Porém a escolaridade 5 (12 ou mais anos) contém variação menor que as outras.
Figura 25. Distribuição de semanas de gestação por escolaridade da mãe. (Fonte: Elaboração Própria).
6.2 ÁRVORE DE DECISÃO
A Figura 26 mostra um trecho do algoritmo utilizado, nesse trecho é realizado a divisão do DataFrame das features de entrada , e o DataFrame que contém o rótulo.Neste

caso o rótulo será o DataFrame ?target?, esse Dataframe contém a feature ?is_neonatal_death?, essa feature informa se o recém nascido veio a óbito ou não, sendo o foco da predição que desejamos realizar essa feature entra como rótulo para o modelo, já no Dataframe ?nome_features? temos as features que vão servir de entrada para o modelo, é a partir dessas features que o modelo tentará encontrar padrões para predizer o risco de óbito do recém nascido. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

46

Figura 26. Separação dos DataFrames de entrada e rótulo. (Fonte: Elaboração Própria).

6.2.1 Modelo de Árvore de Decisão Utilizando Particionamento 90/10

O algoritmo de árvore de decisão foi treinado utilizando duas formas diferentes: usando as partições 90% para treino e 10% para os testes e utilizando o método K-folds cross-validation. E para os experimentos iniciais foi utilizado o particionamento 90/10.

Tabela 4. Resultados do modelo de árvore de decisão. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	671807
Morto(1)	0.71	0.29	0.41	4216

Analisando a Tabela 4 pode-se ver os seguintes resultados, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 71% que realmente pertenciam a classe 1, então cerca de 29% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 29% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem a classe 1 que seria dos recém nascidos que poderiam vir a óbito, já para a classe de vivos a

47

classe 0 o modelo está identificando muito bem e classificando de forma correta. A tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.

Esse modelo teve uma acurácia de 0.99 que é uma acurácia bem alta, porém somente pela acurácia não é possível dizer que o modelo está excelente pelo motivo da base que está sendo usada é altamente desbalanceada, então para a classe 0 que é a de vivos temos muito mais dados que para a classe de mortos, e o modelo está acertando bastante a classe de vivos logo a acurácia fica alta por esses acertos, enquanto para a classe de mortos o modelo não está acertando tanto.

A Figura 27 mostra a curva ROC do modelo, pode-se observar que a AUC da curva ROC ficou em 0.92, a AUC mostra que o modelo está acertando bastante as predições, pois quanto mais perto de 1 melhor está no nosso modelo, mas no caso desse modelo vimos acima que as predições para a classe de mortos(1) está ruim, o modelo está acertando poucas classificações como 1. Então a AUC assim como a acurácia está alta porque a base de dados utilizada é altamente desbalanceada tendo muito mais amostras de vivos(1), e como o modelo está bom para essa classificação e está acertando bastante os valores de AUC e acurácia ficam

altos .
Figura 27. Curva ROC modelo de Árvore de decisão. (Fonte: Elaboração Própria).
A Figura 28 mostra a importância de cada feature para o modelo sendo assim a feature 10 - Peso ao nascer, a mais importante para o modelo, pois dela o modelo conseguiu tirar mais informações, seguido das features, 13 - Apgar dos 5 primeiros minutos, 11 - Semanas de
48
Gestação, 14 - Presença de malformação 12 - Apgar do 1 primeiro minutos. Então essas são as features que foram mais decisivas para a montagem da árvore e para as predições do modelo.
Figura 28. Gráfico de importância das features. (Fonte: Elaboração Própria).
Na Figura 29 temos uma imagem reduzida da árvore de decisão que foi montada com o treinamento utilizando 5 como profundidade máxima para a árvore, a imagem da árvore completa pode ser visualizada no apêndice A. É possível ver que as features marcadas como importante para o modelo apareceu diversas vezes na árvore, e a feature peso ao nascer foi escolhida para ser o nó raiz da árvore, de todas as features ela foi a que teve a melhor entropia para ser o nó raiz, e depois aparece mais vezes para outras decisões em outros níveis da árvore e isso faz com que essa feature se torne a mais importante para o modelo, também podemos ver que a feature ?Apgar do 1 primeiro minuto? mesmo sendo a menos importante para o modelo ela aparece somente no nó de número 42 isso mostra que a entropia e o ganho de informação dessa feature nas outras decisões não foram boas fazendo ela ser a feature com menos importância utilizada no modelo.
49
Figura 29. Árvore de decisão montada a partir do algoritmo. (Fonte: Elaboração Própria).
6.2.2 Modelo de Árvore de Decisão Utilizando K-folds cross-validation
Após realizar esse experimento com a partição 90/10, foi realizado outro experimento com esse mesmo algoritmo porém agora utilizando o método K-folds cross-validation para o treinamento. O K-folds cross-validation foi configurado para separar a base de dados em 10 partes iguais, e assim o algoritmo foi treinado novamente gerando um novo modelo. O método de K-folds cross-validation é utilizado para conferir se o modelo não está sofrendo problemas de overfitting, é importante garantir que o modelo está aprendendo com os dados e não está somente decorando.
Na Tabela 5 temos então o resultados desse modelo, pode-se observar que os resultados para a classificação de vivos(0) não teve alteração, o modelo continua tendo 100% na precisão e no recall, mas na classificação de mortos(1) o modelo teve uma diminuição na precisão que agora está em 65% e também teve uma diminuição no recall que agora está em 23%, porém essa diminuição é justificada porque para o teste deste modelo foi utilizado a base completa e não somente a partição de teste que contém 10% da base total. Logo os resultados não tiveram uma alteração drástica e tiveram um comportamento bem semelhante, incluindo a acurácia que se manteve em 0.99 e a AUC que teve uma diminuição pequena também e ficou com 0.89.
50
Tabela 5. Resultados do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)
Precision Recall F1-Score Support

Vivo(0) 1.00 1.00 1.00 6719191
Morto(1) 0.65 0.23 0.34 41031
Na Tabela 6 é possível ver a matriz de confusão do modelo que mostra o número exato de erros e acertos do modelo, pode-se observar que para a classe de vivos o modelo classificou corretamente 6.714.117 amostras da base de dados completa, o modelo classificou como 0 as amostras que nos rótulos realmente eram 0, errou apenas 5074 amostras. Já para a classe de mortos o modelo acertou somente 9552 amostras, classificou como 1 as amostras que no rótulo eram 1 também, e errou 31479. Isso mostra que o modelo está muito desbalanceado, pois está errado muito mais do que acertando as predições de mortos, sendo isso um reflexo da base de dados desbalanceada também. E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.
Tabela 6. Matriz de confusão do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6714117 5074
Mortos (1) 31479 9552
6.3 REGRESSÃO LOGÍSTICA
Para o algoritmo de Regressão Logística também foi utilizado métodos diferentes para o treinamento, O algoritmo foi treinado utilizando três formas diferentes: usando as partições 90% para treino e 10% para os testes, utilizando o método K-folds cross-validation e foi treinada utilizando o método RFECV que seleciona as melhores features para o modelo, juntamente com o método GridSearchCV. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.
51
6.3.1 Modelo de Regressão Logística Utilizando Particionamento 90/10
Para o primeiro experimento do algoritmo de regressão logística foi utilizado o particionamento 90/10 para o treinamento do algoritmo, sendo 90% da base de dados para realizar o treinamento do modelo e 10% da base para realizar os testes.
Na Tabela 7 temos os resultados do modelo de regressão logística utilizando o método de particionamento 90/10, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 65% que realmente pertenciam a classe 1, então cerca de 35% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 24% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem as amostras da classe 1 que seria dos recém-nascidos que poderiam vir a óbito, já para a classe de vivos a classe 0 o modelo está identificando muito bem e classificando de forma correta.E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do

Recall.
Tabela 7. Resultados do modelo de regressão logística usando particionamento 90/10.
(Fonte: Elaboração Própria)
Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 671885
Morto(1) 0.65 0.24 0.35 4138
A Figura 30 abaixo mostra a curva ROC do modelo de regressão logística usando o particionamento 90/10, analisando a curva ROC podemos ver que a AUC da curva ficou em 0.93 e a acurácia do modelo ficou 0.99, mostrando que o modelo está acertando bastante as predições, porém esses números para o nosso modelo não mostra que ele está ótimo, com a análise da Tabela 7 acima podemos observar que o modelo está acertando muitas predições da classe de vivos e poucas predições da classe de mortos, como a base de dados é
52
desbalanceada, como a maior quantidade de dados está na classe de vivos e o modelo está acerto quase todos dessa classe a acurácia e a AUC ficam altas por esses acertos.Logo é preciso analisar mais resultados além da acurácia e da AUC do modelo.
Figura 30. Curva ROC modelo regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria).
Na Tabela 8 podemos analisar os número exatos que o modelo acertou de cada classe, como é mostrado na matriz de confusão o modelo acertou 671.435 da classe de vivos e acertou apenas 915 da classe de mortos, e errou mais que o triplo da classe de mortos. Então as predições do modelo estão muito desbalanceadas, para a classe de vivos o modelo está predizendo bem, porém para as classes de mortos o modelo está errando muitas predições.
Tabela 8. Matriz de confusão do modelo de regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 671435 504
Mortos (1) 3169 915
6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation
Para o segundo experimento do algoritmo de regressão logística foi utilizado o método K-folds cross-validation para o treinamento do modelo com a intenção de conferir e confirmar
53
que o modelo não esteja tendo problemas com overfitting e realmente esteja aprendendo com os dados que está sendo usado para o treinamento. O método K-folds cross-validation foi configurado para realizar uma separação de 10 partes iguais.
Na Tabela 9 podemos observar os resultados do modelo, e os resultados foram parecidos com o experimento anterior usando o particionamento 90/10, única diferença nos resultado é que no experimento anterior a precisão da classe de mortos ficou em 65% e no caso desse experimento usando o K-folds cross-validation a precisão ficou em 64%, mas os valores de recall e F1-Score se mantiveram, assim como todos os resultados da classe de mortos se mantiveram em 100%. A execução desse experimento mostra que o modelo realmente está aprendendo com os dados e não está apenas decorando, retirando o risco do

modelo estar com overfitting. Mesmo sem alterações no resultado é importante saber que o modelo não está com overfitting e está conseguindo aprender e encontrar padrões nos dados.
Tabela 9. Resultados do modelo de regressão logística usando K-folds cross-validation. (Fonte: Elaboração Própria)
Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 6719191
Morto(1) 0.64 0.24 0.35 41031
Já na Figura 31 temos a curva ROC do modelo, essa curva ROC mostra a AUC de todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas para todas as vezes que foi realizado os testes, a AUC se manteve em 0.93 e 0.94, sendo a média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua acertando muito a classe de vivos que é a classe predominante na base de dados. Já era esperado os resultados não sofrerem alterações grandes, pois a base de dados não sofreu nenhuma alteração, esse experimento foi somente para conferir se a modelo não estava sofrendo overfitting, e pelos resultados aparentemente a base não está com problemas de sobreajuste.
54
Figura 31. Curva ROC modelo regressão logística usando K-folds cross-validation. (Fonte: Elaboração Própria).
Na Tabela 10 temos a matriz de confusão do modelo, onde mostra que o modelo acertou 6.713.658 amostras da classe de vivos e errou apenas 5.533, já para a classe de mortos o modelo acertou somente 9.847 e novamente errou mais que o triplo errando 31.182 amostras. Como os resultados mostrados anteriormente não tiveram alteração era de se esperar que as predições corretas e incorretas do modelo também não sofressem alterações, o modelo continua acertando muito a classe de vivos e errando muito a classe de mortos, mantendo as predições bem desbalanceadas.
55
Tabela 10. Matriz de confusão do modelo de regressão logística usando K-folds cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6713658 5533
Mortos (1) 31182 9847

6.3.3 Modelo de Regressão Logística Utilizando GridSearchCV e RFECV

Para o terceiro e último experimento de regressão logística foi utilizado técnicas diferentes juntas para tentar melhorar as predições do modelo, para esse experimento usamos o K-folds cross-validation para o particionamento da base de dados assim como foi utilizado no experimento acima, e também foi utilizado duas técnicas que ainda não foram usadas nos experimentos anteriores que são as funções RFECV e GridSearchCV. A função RFECV seleciona a quantidade ideal de features para ser usadas no treinamento do modelo e também mostra quais são as melhores features para essa predição, então essa função ela utiliza a

validação cruzada para ir treinando e testando N vezes o modelo, e sempre vai alterando a quantidade e as features utilizadas para o teste, então cada vez que a função testa os modelos ela grava a pontuação dos acertos e assim depois mostra o gráfico da Figura 32 e assim é possível ver quais são as features ideais para o modelo.

56

Figura 32. Resultado da função RFECV Base Brasil. (Fonte: Elaboração Própria).

Então os resultados mostrados na Figura 32 apresenta que o número ideal para treinar o modelo de features é 6, e as features sao: 'tp_maternal_schooling', 'gestaional_week', 'cd_apgar1', 'cd_apgar5', 'has_congenital_malformation', 'tp_labor'. Essas são as features que tiveram o melhor desempenho no RFECV, então para o treinamento do modelo desse experimento as features que vão ser utilizadas serão somente essas 6.

Após a execução da função de RFECV foi executado a função de GridSearchCV, o GridSearchCV é utilizado para descobrir quais são os melhores parâmetros para utilizar no algoritmo de regressão logística e criar os modelos, foi utilizado também a validação cruzada para testar qual seria os melhores parâmetros para melhorar o desempenho do modelo. Na Figura 33 pode-se ver que os parâmetros escolhidos pela função foram:

Figura 33. Resultado da função GridSearchCV. (Fonte: Elaboração Própria).

Na Tabela 11 podemos observar os resultados do modelo, e os resultados foram parecidos com os experimentos anteriores, para a classe de mortos a precisão teve um aumento e foi para 69% e o recall diminuiu chegando em 20%, então utilizando as novas

57

funções o modelo ganhou precisão e perdeu recall. Para a classe de vivos o modelo se manteve em 100% e não teve alterações para os resultados de precisão e recall. Mesmo com a utilização das funções de RFECV e GridSearchCV o modelo não teve uma melhora significativa para as predições de mortos.

Tabela 11. Resultados do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	6719191
Morto(1)	0.69	0.20	0.31	41031

Na Figura 34 temos a curva ROC do modelo, essa curva ROC mostra a AUC de todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas para todas as vezes que foi realizado os testes, a AUC se manteve em 0.92 e 0.93, sendo a média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua acertando muito a classe de vivos que é a classe predominante na base de dados.

58

Figura 34. Curva ROC modelo regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria).

Na Tabela 12 temos a matriz de confusão do modelo, onde mostra que o modelo acertou 6.715.535 amostras da classe de vivos e errou apenas 3.656, comparando com os experimentos anteriores de regressão logística as predições de vivos teve uma pequena piora e teve uma pequena diminuição nos acertos. Já para a classe de mortos, o modelo acertou 8.296

e errou 32.735 amostras, logo as predições de vivos tiveram uma uma piora e acertou um pouco menos do que os outros experimentos, e na classe de vivos teve uma melhora acertando mais predições, porém como o problema está nas predições de mortos e para essa classe teve uma piora pode-se dizer que a função de GridSearchCV não teve um ganho significativo nos resultados. Então como modelo final foi escolhido o modelo utilizando somente o K-folds cross-validation, pois foi o modelo que mostrou o melhor desempenho dos experimentos de regressão logística.

59

Tabela 12. Matriz de confusão do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

	Predito	
	Vivos (0)	Mortos (1)
Classe Real		
Vivos (0)	6715535	3656
Mortos (1)	32735	8296

60

7 CONCLUSÃO

O principal objetivo deste trabalho foi analisar os resultados das predições de mortalidade neonatal dos algoritmos de aprendizado de máquina usando uma base de dados com dados do Brasil inteiro. A partir destes resultados apresentados na secção anterior ?Resultados?, é possível concluir que analisar somente a AUC e a acurácia do modelo não é o suficiente para garantir que o modelo está excelente, sendo assim temos que ter mais atenção a precisão, recall e as predições mostradas na matriz de confusão.

Ambos os modelos ficaram desbalanceados nas predições, os modelos acertaram mais predições de vivos do que mortos, e as predições corretas de mortos foram mais baixas do que as incorretas, para os modelos ficarem melhor precisam passar por algoritmos que possam balancear essas predições, uma outra alternativa para melhorar o modelo é utilizar uma base de dados mais balanceada, pois essa base que foi utilizada é altamente desbalanceada.

Embora os dois modelos tenham tido resultados parecidos , o modelo de regressão logística utilizando somente o K-folds cross-validation se saiu melhor que os outros experimentos, o modelo criado na seção ?6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation? manteve a acurácia, AUC, precisão e recall dos demais modelos e acertou mais predições, talvez com algoritmos para melhorar as predições da classe de mortos, balanceando mais as predições, esse modelo fique com ótimas taxas preditivas.

Além disso podes-se afirmar que o peso ao nascer do recém nascido é um fator muito importante para as decisões dos modelos, ambos os modelos usaram muito essa informação para as predições, também podemos colocar, as colunas de presença de malformação e semana de gestação, como colunas que foram importantes para os modelos, essas causas podem ser evitadas se houver um acompanhamento médico com qualidade e eficiência.

61

REFERÊNCIAS

ÁRVORE de Decisão - Aula 3. Gravação de Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=ecYpXd4WREk&t=4089s. Acesso em: 1 jul. 2020.

BARRETO, J. O. M.; SOUZA, N. M.; CHAPMAN, E. Síntese de evidências para políticas

de saúde: reduzindo a mortalidade perinatal. Ministério da Saúde, p. 1?44,2015.
BATISTA, André Filipe de Moraes; FILHO, Alexandre Dias Porto Chiavegatto. Machine Learning aplicado à Saúde. In: ZIVIANI, Artur; FERNANDES, Natalia Castro; SAADE, Débora Christina Muchaluat. SBCAS 2019: 19 Simpósio Brasileiro de Computação Aplicada à Saúde. [S. l.: s. n.], 2019. cap. Capítulo 1.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. Towards neonatal mortality risk classification: A data-driven approach using neonatal, maternal, and social factors, [s. l.], 23 out. 2020.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. SPNeoDeath: A demographic and epidemiological dataset having infant, mother, prenatal care and childbirth data related to births and neonatal deaths in São Paulo city Brazil ? 2012?2018, [s. l.], 19 jul. 2020.
BRESAN, Rodrigo. Um método para a detecção de ataques de apresentação em sistemas de reconhecimento facial através de propriedades intrínsecas e Deep Learning. Orientador: Me. Carlos Eduardo Beluzo. 2018. Trabalho de Conclusão de Curso (Superior de Tecnologia em Análise e Desenvolvimento de Sistemas) - Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas., [S. l.], 2018.
CABRAL, Cleidy Isolete Silva. Aplicação do Modelo de Regressão Logística num Estudo de Mercado. Orientador: João José Ferreira Gomes. 2013. Projeto Mestrado em Matemática Aplicada à Economia e à Gestão (Mestrado) - Universidade de Lisboa Faculdade de Ciências Departamento de Estatística e Investigação Operacional, [S. l.], 2013. Disponível em: https://repositorio.ul.pt/bitstream/10451/10671/1/ulfc106455_tm_Cleidy_Cabral.pdf. Acesso em: 1 maio 2021.
CAMPOS, Raphael. Árvores de Decisão: Então diga-me, como construo uma?. [S. l.], 28 nov. 2017. Disponível em: https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69. Acesso em: 23 jan. 2021.
Colab. 2021. Disponível em: https://colab.research.google.com/. Acesso em: 20 mar. 2021.
COX, D.R.;SNELL, E.J. Analysis of Binary Data.London: Chapman & Hall, 2ª Edição, 1989. HOSMER, D.W.; LEMESHOW, S.Applied Logistic Regression. New York: John Wiley, 2ª Edição, 2000.
E.França, S.Lansky. Mortalidade infantil neonatal no brasil: situação, tendências e perspectivas Rede Interagencial de Informações para Saúde - demografia e saúde: contribuição para análise de situação e tendências Série Informe de Situação e Tendências(2009), pp.83-112.
62
FREIRE, Sergio Miranda. Bioestatística Básica: Regressão Linear. [S. l.], 20 out. 2020. Disponível em: http://www.lampada.uerj.br/arquivosdb/_book/bioestatisticaBasica.html. Acesso em: 23 jan. 2021.
Jupyter. 2021. Disponível em: https://jupyter.org/. Acesso em: 20 jun. 2021.
GOODFELLOW, I. et al. Deep learning. [S.l.]: MIT press Cambridge, 2016.
Pandas. 2021. Disponível em: https://pandas.pydata.org/. Acesso em: 20 jun. 2021.

Python. 2021. Disponível em: https://www.python.org/. Acesso em: 20 jun. 2021.
KUHN, M.; JOHNSON, K. Applied predictive modeling. [S.l.]: Springer, 2013. v. 26.
LANSKY, S. et al. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante . Cadernos de Saúde Pública, Scielo, v. 30, p. S192 ? S207, 00 2014.
LANSKY, Sônia; FRICHE, Amélia Augusta de Lima; SILVA, Antônio Augusto Moura; CAMPOS, Deise; BITTENCOURT, Sonia Duarte de Azevedo; CARVALHO, Márcia Lazaro; FRIAS, Paulo Germano; CAVALCANTE, Rejane Silva; CUNHA, Antonio José Ledo Alves. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante e ao recém-nascido. [s. l.], Ago 2014. Disponivel em: https://www.scielosp.org/article/csp/2014.v30suppl1/S192-S207/pt/
Matplotlib. 2021. Disponível em: https://matplotlib.org/. Acesso em: 20 mar. 2021.
Maranhão AGK, Vasconcelos AMN, Trindade CM, Victora CG, Rabello Neto DL, Porto D,et al. Mortalidade infantil no Brasil: tendências, componentes e causas de morte no período de 2000 a 2010 . In: Departamento de Análise de Situação de Saúde, Secretaria de Vigilância em Saúde, Ministério da Saúde, organizador. Saúde Brasil 2011: uma análise da situação de saúde e a vigilância da saúde da mulher. v. 1. Brasília: Ministério da Saúde; 2012. p. 163-82.
MESQUITA, PAULO. UM MODELO DE REGRESSÃO LOGÍSTICA PARA AVALIAÇÃO DOS PROGRAMAS DE PÕS-GRADUAÇÃO NO BRASIL. Orientador: Prof. RODRIGO TAVARES NOGUEIRA. 2014. Dissertação (Mestre em Engenharia de Produção) - Universidade Estadual do Norte Fluminense, [S. l.], 2014.
MNASSRI , Baligh. Titanic: logistic regression with python. [S. l.], 1 ago. 2020. Disponível em: https://www.kaggle.com/mnassrib/titanic-logistic-regression-with-python. Acesso em: 14 jan. 2021.
Morais, R.M.d., Costa, A.L: Uma avaliação do sistema de informações sobre mortalidade. Saúde em Debate 41, 101?117 (2017). Disponível em: https://doi.org/10.1109/SIBGRAPI.2012.38.
MOTA, Caio Augusto de Souza; BELUZO , Carlos Eduardo; TRABUCO, Lavinia Pedrosa; SOUZA , Adriano; ALVES, Luciana; CARVALHO, Tiago. DESENVOLVIMENTO DE UMA PLATAFORMA WEB PARA CIÊNCIA DE DADOS APLICADA À SAÚDE MATERNO INFANTIL , [s. l.], 28 nov. 2019.
MULTIEDRO (Brasil). Conheça as aplicações do machine learning na área da saúde. [S. l.], 28 nov. 2019. Disponível em: https://blog.multiedro.com.br/machine-learning-na-area-da-saude/#:~:text=O%20machine%263
0learning%20na%20%C3%A1rea,diagn%C3%B3sticos%20e%20tratamentos%20mais%20precisos. Acesso em: 13 jun. 2021.
MUKHIYA,S.K.;AHMED,U. Hands-On Exploratory Data Analysis with Python: Perform EDA Techniques to Understand, Summarize, and Investigate Your Data. [S.l.]: PacktPublishingLtd,2020.ISBN978-1-78953-725-3.22,32
Murray CJ, Laakso T, Shibuya K, Hill K, Lopez AD. Can we achieve Millennium Development Goal 4? New analysis of country trends and forecasts of under-5 mortality to 2015. Lancet 2007; 370:1040-54.
NumPy. 2021. Disponível em:https://numpy.org/. Acesso em: 20 jun. 2021.

Oliveira, M.M.d., de Araújo, A.S.S.C., Santiago, D.G., Oliveira, J.a.C.G.d., Carvalho, M.D., de Lyra et al, R.N.D.: Evaluation of the national information system on live births in brazil , 2006-2010. Epidemiol. Serv. Saúde 24(4), 629?640 (2015).
PELISSARI, ANA CAROLINA CHEBEL. MORTALIDADE NEONATAL DA CIDADE DE SÃO PAULO: UMA ABORDAGEM UTILIZANDO APRENDIZADO DE MÁQUINA SUPERVISIONADO. 2021. Trabalho de Conclusão de Curso (Superior) - Instituto Federal de Educação, Ciência e Tecnologia Campus Campinas, [S. l.], 2021.
REGRESSÃO logística e binary cross entropy - Aula 7. Direção: Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=3J-LBtHVsm4. Acesso em: 21 jan. 2021.
Rmd Nascimento, AJM Leite, NMGSd Almeida, Pcd Almeida, Cfd Silva Determinantes da mortalidade neonatal: estudo caso-controle em fortaleza, ceará, brasil Cad Saúde Pública, 28(2012), pp.559 - 572.
SANTOS, Hellen Geremias. Machine learning e análise preditiva: considerações metodológicas: métodos lineares para classificação. In: SANTOS, Hellen Geremias. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina. 2018. Tese (Pós-Graduação em Epidemiologia) - Faculdade de Saúde Pública da Universidade de São Paulo, [S. l.], 2018.
Scikit-learn. 2021. Disponível em: https://scikit-learn.org/stable/. Acesso em: 20 jun. 2021.
Seaborn. 2021. Disponível em: https://seaborn.pydata.org/. Acesso em: 20 jun. 2021.
SILVA, Wesley; CORSO, Jansen; WELGACZ, Hanna; PEIXE, Julinês. AVALIAÇÃO DA ESCOLHA DE UM FORNECEDOR SOB CONDIÇÃO DE RISCOS A PARTIR DO MÉTODO DE ÁRVORE DE DECISÃO. ARTIGO ? MÉTODOS QUANTITATIVOS, [s. l.], 12 ago. 2008.
WHO, W. H. O.Women and health: today?s evidence, tomorrow?s agenda. [S.l.]: UNWorld Health Organization, 2009. ISBN 9789241563857.
64
APÊNDICE A

==========

Arquivo 1: monografia-V7.docx.pdf (9728 termos)
Arquivo 2: https://studylibpt.com/doc/4536178/hist%C3%B3ria---figure-b (248 termos)
Termos comuns: 2
Similaridade: 0,02%
O texto abaixo é o conteúdo do documento monografia-V7.docx.pdf (9728 termos)
Os termos em vermelho foram encontrados no documento
https://studylibpt.com/doc/4536178/hist%C3%B3ria---figure-b (248 termos)

==========

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO
CÂMPUS CAMPINAS
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
CAMPINAS
2021
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
Trabalho de Conclusão de Curso apresentado como
exigência parcial para obtenção do diploma do
Curso de Tecnologia em Análise e
Desenvolvimento de Sistemas do Instituto Federal
de Educação, Ciência e Tecnologia Câmpus
Campinas.
Orientador: Prof. Me. Everton Josué da Silva.
CAMPINAS
2021
Dados Internacionais de Catalogação na Publicação
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
Trabalho de Conclusão de Curso apresentado
como exigência parcial para obtenção do diploma
do Curso de Tecnologia em Análise e
Desenvolvimento de Sistemas do Instituto
Federal de Educação, Ciência e Tecnologia de
São Paulo Câmpus Campinas.
Aprovado pela banca examinadora em: _____ de _____ de _____.
BANCA EXAMINADORA

_____
Prof. Me. Everton Josué da Silva (orientador)

IFSP Câmpus Campinas

_____

Prof. Me. Carlos Eduardo Beluzo
IFSP Câmpus Campinas

_____

Prof. Dr. Ricardo Barz Sovat
IFSP Câmpus Campinas

Dedico este trabalho aos meus familiares,
colegas de classe, professores e servidores do Instituto
que colaboraram em minha jornada formativa.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pelo dom da vida e pela oportunidade de concluir mais uma etapa de minha experiência acadêmica.

Agradeço a todos os professores e servidores do IFSP Câmpus Campinas, que contribuíram direta e indiretamente para a conclusão deste trabalho.

Agradeço também à minha família, que deu todo o apoio necessário para que eu chegasse até aqui.

Agradeço ao meu orientador que me auxiliou a solucionar as dificuldades encontradas no caminho.

"Minha energia é o desafio, minha motivação é o impossível,
e é por isso que eu preciso
ser, à força e a esmo, inabalável.".
Augusto Branco

RESUMO

A mortalidade infantil pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível em uma população. O aprendizado de máquina é uma área da inteligência artificial que propõe métodos de análise de dados que automatizam a construção de modelos analíticos com base em reconhecimento de padrões, baseado no conceito de que sistemas podem aprender com dados, identificando padrões e tomando decisões com o mínimo de intervenção humana. O objetivo deste trabalho é testar e analisar dois tipos diferentes de algoritmos de aprendizado de máquina, utilizando a base de dados do SIM e SINASC do Brasil, do período de 2016 até 2018, para gerar modelos para predição de mortalidade neonatal. Os métodos utilizados foram Árvore de Decisão e de Regressão Logística e como métricas para avaliar os modelos resultantes foram utilizadas a AUC, Curva ROC e a Matriz de Confusão. Como resultado, apesar de terem alcançado valores de AUC 0.93, ambos modelos de predições acertaram muitas predições de vivos cerca de 671.000 e erraram muitas predições de mortos cerca de 2.600, principalmente devido ao fato de a base estar desbalanceada.

Palavras-chave: Mortalidade Neonatal, Predição, Aprendizado de Máquina.

ABSTRACT

Infant mortality can be used to analyze poverty and socioeconomic levels, as well as measure the quality of health and medical technology available in a population. Machine learning is an area of artificial intelligence that proposes data analysis methods that automate the

construction of analytical models based on pattern recognition, based on the concept that systems can learn from data, identifying patterns and making decisions with a minimum of human intervention. The objective of this work is to test and analyze two different types of machine learning algorithms, using the SIM and SINASC do Brasil database, from 2016 to 2018, to generate models for predicting neonatal mortality. The methods used were Decision Tree and Logistic Regression and as metrics to evaluate the resulting models, AUC, ROC Curve and Confusion Matrix were used. As a result, despite achieving values of AUC 0.93, both prediction models correct many live birth predictions around 671.000 and miss many stillbirth predictions around 2600, mainly due to the fact that the base is unbalanced.
Keywords: Neonatal Mortality, Prediction, Machine Learning.
LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

SIM Sistema de Informação sobre Mortalidade
SINASC Sistema de Informações sobre Nascidos Vivos
TMI Taxa de Mortalidade Infantil
TMN Taxa de Mortalidade Neonatal
MN Mortalidade Neonatal
ML Machine Learning
DNV Declaração de Nascido Vivo
VN Verdadeiro Negativo
FN Falso Negativo
VP Verdadeiro Positivo
FP Falso Positivo
ROC Curva Característica de Operação do Receptor
NaN Não é um Número
RFE Eliminação Recursiva de Feature
RFECV Eliminação Recursiva de Feature com Validação Cruzada

SUMÁRIO

13

1 INTRODUÇÃO

A taxa de mortalidade infantil (TMI) é uma importante medida de saúde em uma população e é um grande problema no mundo todo. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível. Esta taxa é dividida em duas categorias: neonatal e pós-neonatal. É categorizada neonatal quando o óbito ocorre nos 28 primeiros dias de vida após o pós-parto, e o pós-neonatal é quando o óbito ocorre entre 29 e 364 dias de vida (BELUZO et al., 2020). Cerca de 46% das mortes no mundo acontecem entre os menores de cinco anos e grande parte está concentrada nos primeiros dias de vida (LANSKY, 2014). A diminuição dessas taxas é muito importante no mundo todo, refletindo nos indicadores de saúde pública e desenvolvimento do país.

Os fatores associados à mortalidade são profundamente influenciados pelas características biológicas maternas e neonatais, pelas condições sociais e pelos cuidados prestados pelos serviços de saúde (E. FRANÇA; S. LANSKY, 2009; R.M.D. NASCIMENTO, 2012). Em grande parte, o diagnóstico é altamente dependente da experiência adquirida pelo

profissional que o realiza. Apesar de indiscutivelmente necessário, não é perfeito, principalmente pelo fato de ser dependente de fatores humanos, por este motivo os profissionais sempre tiveram recursos tecnológicos para auxiliar nessa tarefa (BELUZO et al., 2020). Segundo estudos de 2010 no Brasil, calcula-se que aproximadamente 70% dos óbitos infantis ocorridos poderiam ter sido evitados por ações de saúde e que 60% dos óbitos neonatais ocorreram por situações que poderiam ser evitadas (BARRETO; SOUZA; CHAPMAN, 2015).

No presente trabalho foram utilizados dois algoritmos de aprendizado de máquina para criar modelos de predição de mortalidade neonatal. Além disso, foi realizada também uma análise exploratória da base de dados. Para este trabalho foram utilizadas duas fontes de dados: Sistema de Informação sobre Mortalidade (SIM) e Sistema de Informações sobre Nascidos Vivos (SINASC) do Brasil inteiro.

14

2 JUSTIFICATIVA

Com o avanço da tecnologia, podemos notar que ela está cada vez mais presente no nosso dia a dia e nos auxilia em diversas tarefas do cotidiano, e vem sendo aplicada em diversas áreas, dentre elas a saúde.

A mortalidade infantil é uma preocupação mundial na saúde pública, a ONU (Organizações das Nações Unidas) definiu a redução da mortalidade infantil como uma meta para o desenvolvimento global. A mortalidade neonatal é responsável por aproximadamente 60% da TMI nos países em desenvolvimento (A.K. SINGHA, 2016). Existem diversos fatores que podem contribuir com a redução de óbitos neonatais, como por exemplo acompanhamento médico à gestante no período de gravidez, acompanhamento médico ao recém-nascido nos primeiros dias de vida e a disponibilização deste serviço com qualidade e proficiência (WHO, 2009).

A diminuição dessa taxa é importante e de interesse para o mundo todo, e utilizar da tecnologia para auxiliar nessa diminuição é uma proposta funcional, e inovadora para a realidade brasileira (MOTA et al., 2019). Havendo disponibilidade de tecnologia para auxiliar nas decisões dos diagnósticos, os profissionais da saúde podem focar nos cuidados do recém-nascido com risco de vir a óbito, fornecendo tratamentos melhores.

Segundo o site Multiedro (2019), um grande problema que os médicos enfrentam ao realizar o diagnóstico, é analisar um grande volume de dados. Para a área médica obter diagnósticos rápidos e precisos pode salvar vidas, e a utilização de machine learning para realizar esses processos é importante, com a ML os dados podem ser processados em questões de segundos e resultar em um diagnóstico mais rápido, além disso utilizando bons modelos ML os diagnósticos serão mais precisos.

Diversos trabalhos vêm sendo desenvolvidos neste contexto. No trabalho realizado por BELUZO et al. (2020a), os autores utilizaram uma base de dados com informações da cidade de São Paulo para criação de modelos de aprendizado de máquina para predição de mortalidade neonatal. Este recorte foi utilizado também no presente trabalho para fins de exercício e definição de etapas da implementação de código. Na conclusão os autores discutem os fatores que tiveram mais influência nas predições, e na análise feita pelos autores, as consultas de pré-natal são muito importantes para a redução da mortalidade infantil, uma vez que algumas características muito importantes, como a malformação congênita, podem

ser detectadas ao longo das consultas de pré-natal.
Em um outro estudo realizado por BELUZO et al. (2020a), foi realizada uma análise exploratória da base de dados SPNeoDeath, onde e podemos observar detalhadamente cada
15
análise das colunas da tabela e diversos gráficos feito para um melhor entendimento dos dados.
Outro trabalho que foi realizado utilizando a base de dados com dados somente de São Paulo foi o de PELISSARI (2021), nesse trabalho é realizado uma análise exploratória na base de dados de São Paulo e é também analisado três algoritmos de aprendizado de máquina (Logistic Regression, Random Forest Classifier e XGBoost), nesse trabalho a autora conclui que os modelos de Logistic Regression e XGBoost foram os modelos que apresentaram os melhores desempenhos preditivos, e também mostra os problemas de estar utilizando uma base de dados desbalanceada.
O problema da mortalidade neonatal não é recente, e o mundo todo desenvolve iniciativas para lidar com ele. A ONU definiu como meta a redução da mortalidade infantil para o desenvolvimento global. Diversos fatores podem ajudar na diminuição da taxa de mortalidade neonatal como acompanhamento médico antes e após parto, tanto com a mãe quanto com o recém-nascido. Um serviço de qualidade e constante para os casos com maior risco de óbito pode salvar diversos recém-nascidos. Neste sentido, o desenvolvimento de ferramentas para a predição de mortalidade neonatal pode colaborar com esta iniciativa.

3 OBJETIVOS

3.1 OBJETIVO GERAL

O presente trabalho tem como objetivo testar e analisar os resultados da aplicação dos aprendizagem de máquina supervisionado Árvore de Decisão e Regressão Logística, utilizando dados do SIM e do SINASC, no período de 2016 até 2018, a fim de gerar modelos de predição de risco morte neonatal.

3.2 OBJETIVOS ESPECÍFICOS

? Realizar análise exploratória da base dados a ser utilizada na construção dos modelos;
? Implementar modelos de predição utilizando algoritmos Árvore de Decisão e Regressão Logística;
? Avaliar resultados e propor novas abordagens.
16

4 FUNDAMENTAÇÃO TEÓRICA

4.1 MORTALIDADE INFANTIL E NEONATAL

Como dito na monografia PELISSARI (2021), a TMI consiste no número de crianças que foram a óbito antes de concluir um ano de vida dividido por 1000 crianças nascidas vivas no tempo de um ano. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos e qualidade da saúde pública de um país (apud BELUZO et al., 2020).
Mencionado em PELISSARI (2021), a TMN é uma das duas categorias da TMI, é categorizado mortalidade neonatal quando o óbito ocorre do nascimento da criança até os 28 primeiros dias de vida. A mortalidade neonatal pode ser subdividida em mortalidade neonatal precoce que é quando o óbito ocorre entre 0 e 6 dias de vida, e o neonatal tardio quando o óbito ocorre entre 7 e 28 dias de vida (apud E. FRANÇA e S. LANSKY, 2009).
Segundo Lansky et al. (2014), a taxa de mortalidade neonatal é o principal

componente da TMI desde a década de 1990 no Brasil e vem se mantendo em níveis elevados, com taxa de 11,2 óbitos por mil nascidos vivos em 2010 (apud Maranhão et al., 2012). Também é dito em Lansky et al. (2014), que a TMI do Brasil em 2011 foi 15,3 por mil nascidos vivos, alcançando a meta 4 dos objetivos de desenvolvimento do milênio, compromisso dos governos integrantes das Nações Unidas de melhorar a saúde infantil e reduzir em 2/3 a mortalidade infantil entre 1990 e 2015. (apud Maranhão et al., 2012; apud Murray et al., 2015). Segundo Lansky et al. (2014) o principal componente da TMI em 2009 é a mortalidade neonatal precoce, e grande parte das mortes infantis acontece nas primeiras 24 horas (25%), indicando uma relação estreita com a atenção ao parto e nascimento (apud E. FRANÇA e S. LANSKY, 2009).

4.2 MÉTODOS DE APRENDIZADO DE MÁQUINA

A utilização de dados para a resolução de problemas, de modo a se identificar padrões ocultos e posteriormente auxiliar na tomada de decisões é definida como aprendizado de máquina (BRESAN, 2018 apud Brink et al. 2013).

O uso de técnicas de Aprendizado de Máquina se mostra altamente eficiente na resolução de tarefas que se apresentem difíceis de se resolver a partir de programas escritos por humanos (BRESAN, 2018 apud GOODFELLOW et al., 2016), bem como também possibilita uma maior compreensão sobre os funcionamentos dos princípios que compõem a inteligência humana.

17

Neste trabalho serão implementados modelos utilizando os algoritmos de Árvore de Decisão e Regressão Logística, os quais serão apresentados a seguir.

4.2.1 Árvore de Decisão

Segundo Batista e Filho (2019), os modelos preditivos baseados em árvores são geralmente utilizados para tarefas de classificação, embora também possam ser utilizados para tarefas de regressão. Considerado um dos mais populares algoritmos de predição, o algoritmo de árvore de decisão proporciona uma grande facilidade de interpretação.

As árvores de decisão utilizam a estratégia ?dividir-e-conquistar", na qual as árvores são construídas utilizando-se apenas alguns atributos. A árvores de decisão é uma das técnicas por meio da qual um problema complexo é decomposto em subproblemas mais simples. Recursivamente, a mesma estratégia é aplicada a cada subproblema (SILVA et al, 2008).

A árvore de decisões tem uma estrutura hierárquica onde cada nó da árvore representa uma decisão em uma das colunas do conjunto de dados, cada ramo da árvore representa uma tomada de decisão (BATISTA; FILHO, 2019).

O algoritmo segue algumas decisões para montar a árvore, o atributo mais importante é utilizado como nó raiz e será o primeiro nó, já os atributos menos importantes são mostrados nos ramos da árvore. Cada atributo é verificado para conferir se é possível gerar uma árvore menor e se é possível fazer uma classificação melhor, e assim é escolhido o nó que produz nós filhos (SILVA et al, 2008).

Existem diferentes tipos de algoritmo para a construção da árvore, como por exemplo, ID3 (Iterative Dichotomiser), C4.5 e CART (Classification and Regression Tree). O algoritmo ID3 escolhe os nós da árvore por meio da métrica do ganho de informação. Já o CART faz uso da equação de Gini (BATISTA; FILHO, 2019 apud KUHN; JOHNSON, 2013).

O algoritmo utilizado neste trabalho usará o cálculo da entropia e ganho de

informação, para decidir qual atributo será usado no nó, a entropia é uma medida da desorganização dos sistemas, maior é a incerteza do sistema, quanto maior a entropia maior está a desorganização, com a árvore de decisão a ideia é ir organizando o sistema e ir separando as decisões da melhor forma para que a entropia diminui e o ganho de informação aumente, para que a decisão do modelo fiquei mais assertiva. O cálculo da entropia utiliza a seguinte equação (ÁRVORE, 2020):

???????? =
?
?? ?? ???
2
??
18

Nesta equação o i representa possíveis classificações (rótulos), e o P é a probabilidade de cada uma. A ideia da árvore então é verificar qual é a entropia para cada uma das variáveis da base de dados, e verificar qual é a melhor organização de cada uma das variáveis. E isso é realizado através da técnica chamada de ganho de informação, essa técnica utiliza a equação a seguir (ÁRVORE, 2020):

????? = ????????(???) ? ? ????(??????) * ????????(??????)

Então o ganho de informação é calculado pela entropia do nó pai menos a soma do peso vezes a entropia dos nós filhos. Esse cálculo é realizado para todas as variáveis e a variável que tiver maior ganho de informação é utilizado para a decisão do nó. Esse processo é realizado recursivamente e a árvore vai sendo montada com base nesses cálculos (ÁRVORE, 2020).

Na Figura 1 temos a representação do início de uma árvore de decisão, no caso em questão o autor estava classificando exames positivos e negativos para o vírus SARS-CoV-2, ele utilizou uma base de dados que contém quase 6 mil exames realizados, contendo campos como o resultado do exame, idade do paciente, níveis de hemoglobina entre outras informações.

Figura 1. Exemplo de Árvore de Decisão. (Fonte: ÁRVORE 2020).

Pode-se observar que para o nó raiz foi utilizado a variável Leukocytes, logo foi o que apresentou a melhor organização para ser o nó raiz, e após ele temos os ramos da árvore. Cada retângulo da árvore é uma decisão do modelo e esses retângulos têm atributos importante como, a coluna da base de dados que será onde vai ser realizado a decisão, então pegando o nó raiz, pacientes que tiverem com os Leukocytes para menos que -0.43 seguiram o caminho
19
para a esquerda (true), já os pacientes que tiverem mais seguiram o caminho da direita (false). A entropia calculada do nó também é apresentada e a classificação do nó também, então se o resultado parar neste nó, a classificação que está nele será a classificação final, e temos o samples que é o número de amostras que se enquadram no na decisão. E esse processo de verificação do modelo vai se repetindo até não ter mais como dividir em subproblemas ou até chegar na profundidade máxima.

Uma característica importante da árvore de decisão é que ela é um modelo WhiteBox, ou seja, é possível visualizar a árvore montada e as decisões que o modelo terá que tomar na árvore. Na Figura 1 pode-se ver uma árvore de decisões montada, cada retângulo da árvore é

uma decisão que o modelo terá que tomar, conforme o modelo toma as decisões ele vai seguindo um caminho até chegar ao nó folha, nó folha é o nó que não contém nó filho, não tem mais ramos para baixo, chegando ao nó folha o modelo tem a classificação final.

4.2.2 Regressão Logística

O modelo de regressão logística estabelece uma relação entre a probabilidade de ocorrência da variável de interesse e as variáveis de entrado do modelo, sendo utilizada para tarefas de classificação, em que a variável de interesse é categórica, neste caso, a variável de interesse assume valores 0 ou 1, onde 1 é geralmente utilizado para indicar a ocorrência do evento de interesse (Batista e Filho, 2019).

De acordo com Mesquita (2014), a técnica de regressão logística, desenvolvida no século XIX, obteve maior visibilidade após 1950 ficando então mais conhecida. Ficou ainda mais difundida a partir dos trabalhos de Cox & Snell (1989) e Hosmer & Lemeshow (2000). Caracteriza-se por descrever a relação entre uma variável dependente qualitativa binária, associada a um conjunto de variáveis independentes qualitativas ou métricas.

Esta técnica inicialmente foi utilizada na área médica, porém a eficiência viabilizou sua implementação nas mais diversas áreas. Mesquita (2014) diz que o termo regressão logístico, tem sua origem na transformação usada com a variável dependente, que permite calcular diretamente a probabilidade da ocorrência do fenômeno em estudo.

Segundo Santos (2018), para estimar a probabilidade, é utilizado uma técnica chamada distribuição binomial para modelar a variável resposta do conjunto de treinamento, que tem como parâmetro a probabilidade de ocorrência de uma classe específica. Uma característica importante desse modelo é que a probabilidade estimada deve estar limitada ao intervalo de 0 a 1. O algoritmo de Regressão Logística gera um modelo de classificação binária (0 e 1). O modelo utiliza a Regressão Linear como base, a equação linear é (REGRESSÃO, 2020):

20

? = ? . ? + ?

O ?m? da equação é o coeficiente angular da reta, e o ?b? é o intercepto. Então, aplicando a equação linear obtemos um valor contínuo como resultado, após obter esse valor a Regressão Logística tem que realizar a classificação do resultado, então é aplicado uma função chamada sigmoide (REGRESSÃO, 2020):

? = 1

1+ ? ??

Essa função chamada de sigmoide, ou função logística, é responsável por ?achatar? o resultado da regressão linear, calculando a probabilidade de o resultado pertencer a classe 1, classificando assim o resultado da função linear em 0 ou 1.

Segundo Batista e Filho (2019), o modelo de regressão logística que tem como objetivo realizar uma classificação binária de uma variável de interesse irá predizer a probabilidade de pertencer à classe positiva. Tem-se, portanto, que é dado por:?

? = { 0, ?? ???? (? = 1 ??) < 0, 5

ou

? = { 1, ?? ???? (? = 1??) ? 0, 5

Então a decisão do modelo de regressão logística está relacionada à escolha de um ponto de corte para p(x). Caso o ponto de corte escolhido for p(x)=0,5, os resultados que forem p(x)>0,5 serão classificados como 1 e os resultados p(x)<5 serão classificados como 0

(SANTOS, 2018).
A Figura 2 ilustra um exemplo de como é o diagrama de Regressão Logística. Pelo diagrama então pode-se observar que o modelo possui uma representação gráfica em formato de ?S?, essa seria a curva logística. As predições do modelo sempre ficaram na linha 1 e 0 do eixo y, pois é o intervalo estabelecido pelo algoritmo, então a classificação sempre será nesse intervalo.
21
Figura 2. Exemplo de Diagrama de Regressão Linear. (Fonte: Batista e Filho 2019).
4.2.3 Treino, teste e validação do modelo de aprendizado de máquina
A criação de um modelo de aprendizado de máquina é realizada em duas etapas: treinamento do modelo, que é realizado a partir dos dados fornecidos para o algoritmo, e etapa de teste, onde os modelos recebem dados novos e sem o rótulo, para que seja avaliado a performance do modelo (PELISSARI, 2021).
Para o treinamento dos modelos foram utilizadas duas abordagens, a primeira onde a base foi particionada entre conjunto de dados para teste e para treino em uma porcentagem de 90% para treino e 10% para teste, porém dessa forma pode ocorrer problema de sobreajuste ou overfitting. Nesta situação, o que pode ocorrer é o modelo não aprender com os dados, e apenas ?decorar? os dados recebidos e quando é realizado o teste o modelo consegue predizer apenas situações parecidas, não sendo capaz de identificar padrões com diferenças mínimas.
Para evitar este problema, foi utilizado uma técnica de validação cruzada chamada K-folds cross-validation. Essa técnica de validação cruzada divide a base de dados em K subconjuntos, e então são realizadas várias rodadas de treinamento e teste alternando os subconjuntos utilizados para teste e treino, e o resultado do modelo baseia-se na média da acurácia observada em cada rodada (PELISSARI, 2021).
Pode-se observar na Tabela 1 uma ilustração da técnica K-fold cross-validation, onde temos um exemplo onde a base é dividida em 5 subconjuntos e cada subconjunto tem a parte de teste diferente dos outros subconjuntos, assim o modelo vai receber valores novos de teste e treino todas as vezes que executar um novo subconjunto.
22
Tabela 1. Exemplo da separação K-fold cross-validation. (Fonte: Adaptado de PELISSARI, 2021).

Predição 1	Predição 2	Predição 3	Predição 4	Predição 5
Teste	Treino	Treino	Treino	Treino
Treino	Teste	Treino	Treino	Treino
Treino	Treino	Teste	Treino	Treino
Treino	Treino	Treino	Teste	Treino
Treino	Treino	Treino	Treino	Teste

4.2.4 Métricas para avaliação de modelos
Para fins de avaliação, neste trabalho foram utilizadas três métricas para avaliar o desempenho dos modelos: a Matriz de Confusão e a AUC (Area under Receiver Operating Characteristic Curve - ROC). Na matriz de confusão pode-se observar o comportamento do modelo em relação a cada uma das classes a serem preditas pelo modelo, utilizando variáveis binárias (positivo: 1; e negativo: 0), para apresentar uma matriz que faz uma comparação entre as predições do modelo e os valores corretos do conjunto de dados (PELISSARI, 2021).

Na Tabela 2 temos um exemplo de uma matriz de confusão, a qual consiste em duas colunas e duas linhas, e os valores são atribuídos nos quadrantes com os seguintes resultados (PELISSARI, 2021):

? quando o valor real do conjunto de dados é 0, e a predição do modelo também classificou como 0, então temos um Verdadeiro Negativo (VN);

? quando o valor real é 1, e o modelo classificado como 0, temos um Falso Negativo (FN);

? quando o valor real é 0, e o modelo classificado como 1, então o resultado se enquadra no quadrante de Falso Positivo (FP);

? e por fim o quadrante Verdadeiro Positivo (VP), que são os resultados que tem o valor real 1, e o modelo classificado como 1.

23

Tabela 2. Modelo de Matriz de confusão. (Fonte: Adaptado de PELISSARI, 2021).

Valor Predito
Negativo (0) Positivo (1)
Classe
Real
Negativo
(0)
Verdadeiro
Negativo
Falso
Positivo
Positivo
(1)
Falso
Negativo
Verdadeiro
Positivo

A segunda métrica de avaliação utilizada foi a Curva ROC que permite avaliar o desempenho de um modelo com relação às predições efetuadas. A curva ROC é um gráfico de Sensibilidade (taxa de verdadeiros positivos) versus taxa de falsos positivos, a representação da curva ROC permite evidenciar os valores para os quais existe otimização da Sensibilidade em função da Especificidade (CABRAL, 2013).

Na Figura 3 temos um exemplo de uma curva ROC. A linha traçada na cor preta é uma linha de referência, ela representa hipoteticamente a curva ROC de um classificador puramente aleatório. Caso a linha laranja, que representa o resultado do modelo que está sendo avaliado, ficasse igual a linha tracejada preta, indicaria que o modelo avaliado estaria predizendo aleatoriamente os resultados.

Figura 3: Exemplo de Curva ROC. (Fonte: Elaboração Própria).

Por uma análise de inspeção visual, quanto mais próximo a linha laranja estiver do valor 1 no eixo Y, melhor está a capacidade do modelo para diferenciar as classes. O valor da

24

AUC representa a capacidade do modelo de diferenciar as classes, quanto maior o valor do

AUC, melhor é a predição realizada pelo modelo, uma vez que, os valores resultantes iguais a 0 são realmente 0, e os iguais a 1 são de fato 1 (PELISSARI, 2021).

5 METODOLOGIA

O desenvolvimento deste trabalho foi realizado em três etapas: (1) Pré-processamento e Análise Exploratória do conjunto de dados, onde foram aplicadas técnicas comuns para preparar o conjunto de dados como tratamento de valores nulos e seleção de variáveis de interesse; (2) Implementação de modelos de aprendizado de máquina supervisionado para predição das mortes neonatais, utilizando os algoritmos de árvore de decisão e regressão logística; e (3) Análise de Resultados, que será realizada uma análise do desempenho do modelo.

5.1 TECNOLOGIAS E FERRAMENTAS

Neste trabalho vamos utilizar a linguagem de programação Python (PYTHON, 2021) pela sua simplicidade de codificação e alto poder de processamento. Foi utilizado também a plataforma Google Colab Research (Colab, 2021), para o desenvolvimento dos algoritmos e treinamento dos modelos. Além disso, foi utilizado também o Jupyter Notebook (Jupyter, 2021), instalado localmente em computador pessoal, pois o Google Colab Research possui limitação de recursos de memória e processamento, então as duas plataformas foram usadas para realização deste trabalho. As principais bibliotecas utilizadas foram: numpy (NumPy, 2021), e pandas (Pandas, 2021), para a manusear os dados, matplotlib (Matplotlib, 2021), e seaborn (Seaborn, 2021), para a plotagem dos gráficos, e por fim a biblioteca sklearn (Scikit-Learn, 2021), que é a biblioteca responsável pelos algoritmos de aprendizado de máquina.

5.2 BASE DE DADOS

As bases de dados a serem utilizadas serão o SIM e SINASC, que são as duas principais fontes de informações sobre nascimentos e óbitos no Brasil. O SINASC é alimentado com base na Declaração de Nascido Vivo (Declaração de Nascido Vivo - DNV) (BELUZO et al.,2020b apud Oliveira et al., 2015). O SIM tem como objetivo principal apoiar a coleta, armazenamento e processo de gestão de registros de óbitos no Brasil (BELUZO et 25
al.,2020b apud Morais et al., 2017), e foi usado para rotular o óbito registrado no SIM, utilizando o campo DNV como chave de associação. O SIM será utilizado para rotular os registros do SINASC, pois o SIM coletas informações sobre mortalidade e é usado como base para o cálculo de estatísticas vitais, como a taxa de mortalidade neonatal e o SINASC reúne informações sobre dados demográficos e epidemiológicos do bebê, da mãe, do pré-natal e do parto (BELUZO et al.,2020b).

A Tabela 3 apresenta as variáveis da base de dados. Essa base tem variáveis que contêm valores quantitativos e categóricos. A coluna ?num_live_births? por exemplo contém o número de nascidos vivos anteriores da mãe e é composta por valores quantitativos contínuos, e a coluna ?tp_pregnancy? utiliza valores nominais categóricos que indicam o tipo de gravidez.

Tabela 3. Colunas da base de dados e suas descrições. (Fonte: Adaptado de BELUZO et al.,2020b).

Nome da coluna	Descrição	Domínio de dados
Variáveis	demográficas e socioeconômicas	

maternal_age Idade da mãe Quantitativo Contínuo (inteiro)
tp_maternal_race Raça / cor da pele da
mãe
Nominal categórico (inteiro)
1 - branco; 2 - Preto; 3 -
amarelo; 4 - Pele morena; 5 -
Indígena.
tp_marital_status Estado civil da mãe Nominal categórico (inteiro)
1 - Único; 2 - Casado; 3 - Viúva;
4 - Separados / divorciados
judicialmente; 5 - Casamento
por união estável; 9 - Ignorado.
tp_maternal_schooling Anos de escolaridade da
mãe
Nominal categórico (inteiro)
1 - nenhum; 2 - de 1 a 3 anos; 3 -
de 4 a 7 anos; 4 - de 8 a 11 anos;
5 - 12 e mais; 9 - Ignorado.
Variáveis obstétricas maternas
num_live_births Número de nascidos
vivos
Quantitativo Contínuo (inteiro)
num_fetal_losses Número de perdas fetais Quantitativo Contínuo (inteiro)
num_previous_gestations Número de gestações Quantitativo Contínuo (inteiro)
26
anteriores
num_normal_labors Número de partos
normais (trabalhos de
parto)
Quantitativo Contínuo (inteiro)
num_cesarean_labors Número de partos
cesáreos (partos)
Quantitativo Contínuo (inteiro)
tp_pregnancy Tipo de gravidez Nominal categórico (inteiro)
1 - Singleton; 2 - Twin; 3 -
Trigêmeo ou mais; 9 - Ignorado.
Variáveis relacionadas a cuidados anteriores
tp_labor Tipo de parto infantil
(tipo de parto)
Categórico Nominal (inteiro)
1 - Vaginal; 2 - Cesariana;
num_prenatal_appointments Número de consultas de
pré-natal
Quantitativo Contínuo (inteiro)

tp_robson_group Classificação do grupo
Robson
Ordinal categórico (inteiro)
Variáveis relacionadas ao recém-nascido
tp_newborn_presentation Tipo de apresentação de
recém-nascido
Categórico Nominal (inteiro)
1 - Cefálico; 2 - Pélvico ou
culatra; 3 - Transversal; 9 -
Ignorado.
has_congenital_malformation Presença de
malformação congênita
Nominal categórico (inteiro)
1 - Sim; 2 - Não; 9 - Ignorado
newborn_weight Peso ao nascer em
gramas
Quantitativo Contínuo (inteiro)
cd_apgar1 Pontuação de Apgar de 1
minuto
Ordinal categórico (inteiro)
cd_apgar5 Pontuação de Apgar de 5
minuto
Ordinal categórico (inteiro)
gestaional_week Semana de gestação Quantitativo Contínuo (inteiro)
tp_childbirth_care Assistência ao parto Categórico Nominal (inteiro)
1 - Doutor; 2 - enfermeira ou
obstetra; 3 - Parteira; 4 - outros;
9 - Ignorado.
was_cesarean_before_labor Foi cesáreo antes do
parto.
27
was_labor_induced Foi induzido ao trabalho
de parto.
is_neonatal_death Morte antes de 28 dias
(rótulo)
Nominal categórico (inteiro)
0 - sobrevivente; 1 - morto.

Neste trabalho foi utilizado dados do período de 2014 à 2016 devido sua melhor qualidade. Este recorte possui 6.760.222 registros dos quais 99.4% são amostras de recém-nascidos vivos e 0.6% são amostras onde os recém-nascidos vieram a óbito conforme pode ser observado na Figura 4. Com essas informações consegue-se observar que a base de dados é desbalanceada.

Figura 4. Distribuição da base de dados considerando se o recém-nascido viveu ou morreu durante o período neonatal. (Fonte: Elaboração Própria).

5.3 PREPARAÇÃO DA BASE DE DADOS

Nessa etapa foi realizada a preparação da base de dados para o treinamento dos modelos que consiste na limpeza, transformação e preparação dos dados a serem utilizados na análise exploratória e na criação dos modelos preditivos. A base de dados pode conter informações desnecessárias para o modelo que pode acabar atrapalhando as predições,
28
algumas colunas podem ser retiradas ou seus valores podem ser modificados, por exemplo valores reais são modificados para inteiros.

5.4 DESENVOLVIMENTO DOS MODELOS DE APRENDIZADO DE MÁQUINA

Como já dito anteriormente na seção 4 os algoritmos de aprendizado de máquina escolhidos para esse trabalho foram os: Regressão Logística e Árvore de Decisão que utilizam a lógica mostrada na seção 4 , Fundamentação Teórica. Esses dois foram escolhidos por alguns motivos, ambos são algoritmos de classificação e supervisionados, atendendo os critérios necessários para a predição de mortalidade neonatal, ambos os algoritmos são bons para realizar predições, a Árvore de Decisão inclusive é um ótimo modelo para analisar as decisões que estão sendo feitas pelo modelo, afinal esse algoritmo tem a característica de ser um WhiteBox, ou seja conseguimos ver o que o modelo aprendeu e o que ele está decidindo. Esses algoritmos selecionados são algoritmos que se encaixam na necessidade deste trabalho e são muito utilizados na comunidade acadêmica.

Para o algoritmo de Regressão Logística, como foi dito anteriormente na seção ?Preparação da base de dados? a base de dados foi tratada e particionada, no particionamento foi utilizado 90% da base para o treino e 10% para os testes e então foi realizado o treinamento do modelo, também foi aplicada a função RFECV (Eliminação Recursiva de Feature com Validação Cruzada), esse função executa a RFE (Eliminação Recursiva de Feature), que tem como ideia selecionar features considerando recursivamente conjuntos cada vez menores de features, isso ocorre da seguinte forma. Primeiro, o estimador é treinado no conjunto inicial de features e a importância de cada feature é obtida por meio de um atributo "coef" ou por meio de um atributo "feature_importances". Em seguida, as features menos importantes são removidas do conjunto atual de features. Esse procedimento é repetido recursivamente no conjunto removido até que o número desejado de features a serem selecionados seja finalmente alcançado. Porém o diferencial da RFECV é que essa função executa a RFE em um loop de validação cruzada para encontrar o número ideal ou o melhor número de features.

Após essa seleção, foi realizado o treinamento do modelo usando todas as features e também treinamos usando as features que o RFECV selecionou, para compararmos os resultados no final. Também foi utilizado o método de K-folds cross-validation em um novo treinamento para conferirmos se o modelo estava sofrendo overfitting (sobreajuste). E por fim aplicamos um método chamado GridSearchCV, esse método permite que seja realizado testes
29
alterando algumas combinação de parâmetros nos nossos modelos, facilitando para achar a melhor combinação, esse método é parecido com o cross validation, o diferencial do Grid Search é que ele faz os cross validation de vários modelos com hiperparâmetros diferentes de uma só vez.

Para o algoritmo de Árvore de Decisão também foi utilizado o mesmo

particionamento de 90% para treino e 10% para teste, na criação do modelo foi colocado a entropy como critério de decisão e uma profundidade máxima de 4, pois com números maiores o modelo ficava muito complexo e ele errava mais as predições, também utilizamos um algoritmo para mostrar a importância de cada Feature para o modelo.
O algoritmo de Regressão Logística e análise exploratória foi baseado em códigos disponível na plataforma web Kaggle (MNASSRI, 2020), no exemplo do Kaggle o autor faz os códigos para predizer mortes no navio Titanic, para o trabalho de mortalidade neonatal os códigos foram alterados para realizar predição de mortes neonatais. Já no algoritmo de Árvore de Decisão foi baseado em uma vídeo aula do professor Diogo Cortiz (ÁRVORE..., 2020), nessa vídeo aula o professor explica sobre os algoritmos de Árvore de Decisão e depois implementa o um exemplo de código, o código usado neste trabalho usou o codigo do professor Diogo Cortiz, e foi alterado para realizar predições de mortalidade neonatal.

5.5 ANÁLISE DOS RESULTADOS

Para a análise de resultados foi utilizado dois métodos que é o de Matriz de confusão e a curva ROC esses métodos são explicados na seção 4, a Fundamentação Teórica. Com esses métodos pode-se realizar uma avaliação do modelo, e assim será possível analisar os valores de acurácia, precisão e recall do modelo, essas métricas são utilizadas avaliar o desempenho do modelo, com a acurácia é possível calcular a proximidade entre o valor obtido na predição dos modelos e os valores esperados, com a precisão é possível analisar as amostras que o modelo classificou como 0 eram realmente 0 na classe real e com o recall é possível analisar as amostras que o modelo conseguiu reconhecer como a classe correta.

5.6 ACESSO A BASE DE DADOS E CÓDIGOS

A base de dados utilizada neste projeto pode ser encontrada no link: ... . E os códigos podem ser encontrados no link: ... .

30

6 RESULTADOS

Os mesmos experimentos mostrados abaixo foram aplicados para uma base de dados que contém somente dados de São Paulo, essa base é um recorte da base do Brasil todo em contém cerca de 1.400.000 amostras, esse recorte da base também é bem desbalanceado. Os resultados obtidos nesse experimento usando o recorte de São Paulo foram bem parecidos com os experimentos da base do Brasil todo, e os códigos desse experimento podem ser visualizados no apêndice X.

6.1 ANÁLISE EXPLORATÓRIA

O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

6.1.1 Apresentação dos Dados

Como dito anteriormente na seção 5.2 ?Base de Dados? deste trabalho, a base de dados é formada pelas informações do SIM e do SINASC contém 6.760.222 amostras e 29 colunas, porém serão utilizadas somente 23 colunas sendo elas as colunas descritas na Tabela 3. Na seção 5.2 também tem a Figura 4 que mostra que a base dados é desbalanceada tendo 99.4% das amostras de recém-nascidos vivos e 0.6% das amostras onde os recém-nascidos vieram a óbito no período neonatal.

6.1.1.1 Caracteristicas demográficas e socioeconômicas maternas

O apêndice X contém uma tabela que possui as estatísticas sumarizadas das variáveis

demográficas e socioeconômicas maternas. Nesta tabela por exemplo podemos observar a coluna ?maternal_age? que contém a idade da mãe, e na tabela mostra que a média da idade das mães é de 26.4 anos, e os dados possuem uma variação de no mínimo 8 anos e no máximo 55 anos, e a mediana é 26 anos. Na tabela também tem as colunas: ?tp_maternal_schooling? que mostra a categoria de anos de escolaridade da mãe, ?tp_marital_status? que mostra o estado civil da mãe em dados categóricos e a coluna ?tp_maternal_race? que mostra a raça da mãe também em dados categóricos.
31
6.1.1.2 Variávies obstétricas maternas
No apêndice X pode-se observar uma tabela que possui as estatísticas sumarizadas das variáveis obstétricas maternas. Nesta tabela por exemplo temos a coluna ?num_live_births?, essa coluna mostra o número de nascidos vivos anteriores que a mãe obteve, a média desta coluna é 0.94 , a mediana é 1, o número mínimo é 0 e o máximo é 10 nascidos vivos. Na tabela podemos observar também a coluna, ?num_fetal_losses?, esta coluna mostra o número de perdas fetais anteriores da mãe, a média desta coluna é 0.21, a medina é 0, o número mínimo é 0 e o máximo é 5, esses valores mostram que poucas mães tiveram perdas fetais antes da gravidez atual. Também na tabela tem estatísticas da coluna ?num_previous_gestations? esta coluna mostra o número de gestações anteriores da mãe, esta coluna tem a média de 1.14, a mediana é 1, o número mínimo é 0 e o máximo 10. Essa tabela também contém as colunas: ?num_normal_labors? que é o número de partos normais da mãe, ?num_cesarean_labor? que mostra o número de partos cesáreos da mãe e por fim mostra a coluna ?tp_pregnancy? que é o tipo da gravidez.
6.1.1.3 Variávies de histórico de gravidez
No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis do histórico de gravidez. Nesta tabela temos as colunas ?num_prenatal_appointments? que mostra o número de consultas de pré-natal, a média dessa coluna é 7.8, a mediana é 8, a variação dos dados é de no mínimo 0 e no máximo 40. A tabela também contém as colunas: ?tp_labor? que mostra o tipo de parto, e também contém a coluna ?tp_robson_group? que mostra a classificação do grupo Robson.
6.1.1.4 Variávies relacionadas ao recém-nascido
No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis relacionadas ao recém-nascido. Nesta tabela por exemplo podemos observar a coluna ?newborn_weight? que armazena o peso ao nascer dos recém-nascidos em gramas, a média dos dados dessa coluna é 3187.3, a mediana é 3215, a variação dos dados é de no mínimo 0 e no máximo 6000 gramas. Também podemos observar a coluna ?gestacional_week? que mostra o número de semanas de gestação, a média é 38.4, a mediana é 39, a variação dos dados é de no mínimo 15 e no máximo 45 semanas. Além dessas colunas a tabela também
32
contém as colunas: ?cd_apgar1? que mostra a pontuação de Apgar de 1 minuto, ?cd_apgar5? que mostra a pontuação de Apgar de 5 minuto, ?has_congenital_malformation? que mostra se o recém-nascido contém a presença de alguma malformação, ?tp_newborn_presentation? que mostra o tipo de apresentação de recém-nascido, ?tp_labor? que mostra o tipo de parto, ?was_cesarean_before_labor? que mostra se o recém-nascido foi cesáreo antes do parto, ?was_labor_induced ? que mostra se o recém-nascido foi induzido ao trabalho de parto,

?tp_childbirth_care? que mostra se ouve assistência ao parto, e por fim temos a coluna ?is_neonatal_death? que mostra se o recém-nascido veio a óbito ou não.

6.1.2 Análise das variáveis

Nesta seção serão mostrados a parte de análise de variáveis com diferentes tipos de gráficos.

6.1.2.1 Variáveis contínuas

Na Figura 5 temos dois gráficos boxplot, no boxplot à esquerda temos a distribuição da variável peso ao nascer e nele é possível observar duas caixas onde a caixa azul são os vivos e o laranja são os mortos. Nas duas caixas mostradas no gráfico podemos observar pequenos círculos nas pontas, esses círculos são outliers (dados fora da curva), então para a caixa de vivos os outliers são recém-nascidos com pesos acima de 4500 gramas ou abaixo de 2000 gramas, e para a caixa de mortos os outliers são os recém-nascidos com mais de 5500 gramas. Para os vivos temos a mediana de aproximadamente 3200 gramas e para os mortos a mediana é 1300 gramas. Cada caixa representa 50% da base de dados, a caixa azul de vivos mostra que o peso de recém-nascidos que sobreviveram está aproximadamente entre 2700 a 3600 gramas, e para a caixa laranja de mortos o peso de recém-nascidos que vieram a óbito estão aproximadamente entre 700 a 2500 gramas. Então o gráfico mostra para nós que os recém-nascidos que têm maior risco de vir a óbito tem pesos menores que 2000 gramas.

33

Figura 5. BoxPlot das features Peso ao nascer e Idade da mãe. (Fonte: Elaboração Própria).

No gráfico à direita da Figura 5 temos um boxplot da distribuição da variável idade da mãe, como foi dito anteriormente a caixa azul são os vivos e o laranja são os mortos. Nesse boxplot então podemos ver que os outliers da caixa de vivos são mães com idade acima de 46 anos aproximadamente, para os mortos os outliers são mães com idade acima de 50 anos. Para os vivos temos a mediana de aproximadamente 26 anos e para os mortos a mediana é de 26 anos. Nos vivos a idade das mães está aproximadamente entre 21 a 32 anos, e para os mortos a idade das mães está aproximadamente entre 20 a 34 anos. É importante ressaltar que o gráfico mostra as duas caixas bem parecidas, então de acordo com os dados não contém diferença significativa entre vivos e mortos usando a idade da mãe para distribuir.

Já na Figura 6 temos um boxplot da distribuição da variável semanas de gestação.

Nesse boxplot então podemos ver que os outliers da caixa de vivos são gestações com mais de 44 semanas aproximadamente, e gestações com menos de 35 semanas. Para os vivos temos a mediana de aproximadamente 39 semanas e para os mortos a mediana é de 32 semanas. Nos vivos as semanas de gestação estão aproximadamente entre 36 a 40 semanas, e para os mortos a semanas de gestação estão aproximadamente entre 26 a 36 semanas. Nesse boxplot os dados mostraram que para as gestações que têm menos de 36 semanas os recém-nascidos têm maior risco de vir a óbito.

34

Figura 6. BoxPlot da feature semana de gestação. (Fonte: Elaboração Própria).

Na Figura 7 temos um histograma da distribuição de peso ao nascer por classe, podemos observar no gráfico que grande parte dos vivos (linha azul) estão distribuídos entre 2000 e 4000 gramas, a área entre esses valores tem muitos dados de vivos, já entre os valores 0 e 2000 gramas temos uma concentração dos dados de mortos.

Figura 7. Histograma de distribuição do peso entre as classes. (Fonte: Elaboração Própria)

6.1.2.2 Variáveis categoricas

Observando a Figura 8 podemos ver um gráfico de barras da distribuição da variável escolaridade da mãe, este gráfico mostra a contagem de registros por escolaridade da mãe, esse gráfico mostra para nós que a maior parte das mães estão na categoria 4 que representa de 8 a 11 anos de escolaridade.

35

Figura 8. Gráfico de barras da distribuição da variável Escolaridade da mãe. (Fonte: Elaboração Própria).

Observando a Figura 9 podemos ver um gráfico de barras da distribuição da variável número de vivos, este gráfico mostra a contagem de registros por número de vivos, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho.

Figura 9. Gráfico de barras da distribuição da variável Número de nascidos vivos. (Fonte: Elaboração Própria).

Observando a Figura 10 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 1 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 1 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem

36

apgar de 8 ou 9, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.

Figura 10. Gráfico de barras da distribuição da variável Pontuação Apgar de 1 minuto. (Fonte: Elaboração Própria).

Observando a Figura 11 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 5 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 5 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem apgar de 9 ou 10, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.

Figura 11. Gráfico de barras da distribuição da variável Pontuação Apgar de 5 minutos. (Fonte: Elaboração Própria).

37

Observando a Figura 12 podemos ver um gráfico de barras da distribuição da variável presença de malformação congênita, este gráfico mostra a contagem de registros por presença de malformação congênita, esse gráfico mostra para nós que a grande maioria dos dados estão na categoria 2 que mostra que o recém-nascido não possui malformação, esse grande volume para a categoria 2 é causada pelo desbalanceamento da base.

Figura 12. Gráfico de barras da distribuição da variável Presença de malformação congênita. (Fonte: Elaboração Própria).

Observando a Figura 13 podemos ver um gráfico de barras da distribuição da variável número de gestações, este gráfico mostra a contagem de registros por número de gestações, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho, da mesma forma mostrada na Figura 9.

38

Figura 13. Gráfico de barras da distribuição da variável Número de gestações. (Fonte: Elaboração Própria).

Observando a Figura 14 podemos ver um gráfico de barras da distribuição da variável assistência ao parto, este gráfico mostra a contagem de registros por assistência ao parto, esse gráfico mostra para nós que a grande maioria dos dados mostram que o parto teve assistência de uma Enfermeira ou obstetra ou essa informação foi ignorada na hora do resgistro.
Figura 14. Gráfico de barras da distribuição da variável Assistência ao parto. (Fonte: Elaboração Própria).
6.1.2.3 Analise bi-variada
A Figura 15 mostra para nós a importância da feature peso ao nascer com o gráfico mostrado é possível observar claramente a diferença de pesos entre vivos e mortos, recém-nascidos com peso acima de 2000 gramas sobreviveram, já os recém-nascidos com menos de 2000 gramas vieram a óbito.
39
Figura 15. Gráfico de densidade de peso entre as classes. (Fonte: Elaboração Própria).
Na Figura 16 podemos observar um gráfico de correlação entre algumas features de valores contínuos, algumas dessas features possuem correlações positivas, como pontuação apgar de 1 minuto e pontuação apgar de 5 minutos, essas colunas possuem uma correlação de 0.7, então nascidos que recebem um valor alto de apgar de 1 minuto tendem a receber um valor alto no apgar de 5 minutos. O gráfico também mostra outras features com correlações positivas como, número de partos normais e número de gestações anteriores que contém uma correlação de 0.81, número de gestações anteriores com idade da mãe que tem uma correção de 0.39, número de partos de cesáreos com idade da mãe que possui correlação de 0.24, número de partos normais com idade da mãe com a correlação de 0.26, semanas de gestação com peso ao nascer em gramas com correção de 0.52, e as apgar de 1 e 5 minutos com semanas de gestação. E grande parte das features, possuem correlações baixas então elas são inversamente correlacionadas pelo que o gráfico mostra, como por exemplo número de partos normais com número de consultas pré-natais possuem uma correção de -0.17, e número de partos cesáreos com número de partos normais que contém uma correlação de -0.15. Então a correlação dessas features estão bem baixas tirando as correlações acima de 0.7.
40
Figura 16. Gráfico de correlação. (Fonte: Elaboração Própria)
6.1.2.4 Distribuição por raça
Na Figura 17 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por raça, e podemos observar que a raça 5 (indigena), é o boxplot que contem a menor variação de idade, já as raças 1 (branco) e 3 (amarelo) são as caixas que possuem maior variação com a média em torno de 20 e 30 anos.
41
Figura 17. Distribuição da idade da mãe por raça. (Fonte: Elaboração Própria).
No boxplot mostrado na Figura 18 podemos observar a distribuição de peso ao nascer por raça, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre as raças para peso ao nascer.
Figura 18. Distribuição de peso ao nascer por raça. (Fonte: Elaboração Própria).
Na Figura 19 temos o boxplot da a distribuição de semanas de gestação por raça, e praticamente todas as caixas são iguais, somente a caixa da raça 1 (branco) possui menor variação que as outras.

42
Figura 19. Distribuição de semanas de gestação por raça. (Fonte: Elaboração Própria).
6.1.2.5 Distribuição por estado civil
A Figura 20 mostra o boxplot da a distribuição de peso ao nascer por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre os estados civis para peso ao nascer.
Figura 20. Distribuição de peso ao nascer por estado civil. (Fonte: Elaboração Própria).
A Figura 21 mostra o boxplot da a distribuição de semanas de gestação por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Porém os estados civis 2 (Casado) e 4 (Separados/divorciados judicialmente), contém variações menores.
43
Figura 21. Distribuição de semanas de gestação por estado civil. (Fonte: Elaboração Própria).
Na Figura 22 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por estado civil, e podemos observar que o estado civil 3 (Viúva), possui a maior variação, já o estado civil 4 (Separados/divorciados judicialmente) é a caixa que possuem menor variação
Figura 22. Distribuição da idade da mãe por estado civil. (Fonte: Elaboração Própria).
6.1.2.6 Distribuição por escolaridade da mãe
Na Figura 23 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por escolaridade da mãe, e podemos observar que a escolaridade 2 (1 a 3 anos) e 3 (4 a 7
44
anos), possui as maiores variações, já escolaridade 5 (12 ou mais anos) é a caixa que possuem menor variação
Figura 23. Distribuição da idade da mãe por escolaridade da mãe. (Fonte: Elaboração Própria).
A Figura 24 mostra o boxplot da a distribuição de peso ao nascer por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então, pouca variação é observada entre a escolaridade da mãe para peso ao nascer.
Figura 24. Distribuição de peso ao nascer por escolaridade da mãe. (Fonte: Elaboração Própria).
A Figura 25 mostra o boxplot da a distribuição de semanas de gestação por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas,
45
todas são praticamente iguais. Porém a escolaridade 5 (12 ou mais anos) contém variação menor que as outras.
Figura 25. Distribuição de semanas de gestação por escolaridade da mãe. (Fonte: Elaboração Própria).
6.2 ÁRVORE DE DECISÃO
A Figura 26 mostra um trecho do algoritmo utilizado, nesse trecho é realizado a divisão do DataFrame das features de entrada , e o DataFrame que contém o rótulo.Neste caso o rótulo será o DataFrame ?target?, esse Dataframe contém a feature

?is_neonatal_death?, essa feature informa se o recém nascido veio a óbito ou não, sendo o foco da predição que desejamos realizar essa feature entra como rótulo para o modelo, já no Dataframe ?nome_features? temos as features que vão servir de entrada para o modelo, é a partir dessas features que o modelo tentará encontrar padrões para predizer o risco de óbito do recém nascido. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

46

Figura 26. Separação dos DataFrames de entrada e rótulo. (Fonte: Elaboração Própria).

6.2.1 Modelo de Árvore de Decisão Utilizando Particionamento 90/10

O algoritmo de árvore de decisão foi treinado utilizando duas formas diferentes: usando as partições 90% para treino e 10% para os testes e utilizando o método K-folds cross-validation. E para os experimentos iniciais foi utilizado o particionamento 90/10.

Tabela 4. Resultados do modelo de árvore de decisão. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	671807
Morto(1)	0.71	0.29	0.41	4216

Analisando a Tabela 4 pode-se ver os seguintes resultados, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 71% que realmente pertenciam a classe 1, então cerca de 29% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 29% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem a classe 1 que seria dos recém nascidos que poderiam vir a óbito, já para a classe de vivos a

47

classe 0 o modelo está identificando muito bem e classificando de forma correta. A tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.

Esse modelo teve uma acurácia de 0.99 que é uma acurácia bem alta, porém somente pela acurácia não é possível dizer que o modelo está excelente pelo motivo da base que está sendo usada é altamente desbalanceada, então para a classe 0 que é a de vivos temos muito mais dados que para a classe de mortos, e o modelo está acertando bastante a classe de vivos logo a acurácia fica alta por esses acertos, enquanto para a classe de mortos o modelo não está acertando tanto.

A Figura 27 mostra a curva ROC do modelo, pode-se observar que a AUC da curva ROC ficou em 0.92, a AUC mostra que o modelo está acertando bastante as predições, pois quanto mais perto de 1 melhor está no nosso modelo, mas no caso desse modelo vimos acima que as predições para a classe de mortos(1) está ruim, o modelo está acertando poucas classificações como 1. Então a AUC assim como a acurácia está alta porque a base de dados utilizada é altamente desbalanceada tendo muito mais amostras de vivos(1), e como o modelo está bom para essa classificação e está acertando bastante os valores de AUC e acurácia ficam altos .

Figura 27. Curva ROC modelo de Árvore de decisão. (Fonte: Elaboração Própria).
A Figura 28 mostra a importância de cada feature para o modelo sendo assim a feature 10 - Peso ao nascer, a mais importante para o modelo, pois dela o modelo conseguiu tirar mais informações, seguido das features, 13 - Apgar dos 5 primeiros minutos, 11 - Semanas de 48
Gestação, 14 - Presença de malformação 12 - Apgar do 1 primeiro minutos. Então essas são as features que foram mais decisivas para a montagem da árvore e para as predições do modelo.
Figura 28. Gráfico de importância das features. (Fonte: Elaboração Própria).
Na Figura 29 temos uma imagem reduzida da árvore de decisão que foi montada com o treinamento utilizando 5 como profundidade máxima para a árvore, a imagem da árvore completa pode ser visualizada no apêndice A. É possível ver que as features marcadas como importante para o modelo apareceu diversas vezes na árvore, e a feature peso ao nascer foi escolhida para ser o nó raiz da árvore, de todas as features ela foi a que teve a melhor entropia para ser o nó raiz, e depois aparece mais vezes para outras decisões em outros níveis da árvore e isso faz com que essa feature se torne a mais importante para o modelo, também podemos ver que a feature ?Apgar do 1 primeiro minuto? mesmo sendo a menos importante para o modelo ela aparece somente no nó de número 42 isso mostra que a entropia e o ganho de informação dessa feature nas outras decisões não foram boas fazendo ela ser a feature com menos importância utilizada no modelo.
49
Figura 29. Árvore de decisão montada a partir do algoritmo. (Fonte: Elaboração Própria).
6.2.2 Modelo de Árvore de Decisão Utilizando K-folds cross-validation
Após realizar esse experimento com a partição 90/10, foi realizado outro experimento com esse mesmo algoritmo porém agora utilizando o método K-folds cross-validation para o treinamento. O K-folds cross-validation foi configurado para separar a base de dados em 10 partes iguais, e assim o algoritmo foi treinado novamente gerando um novo modelo. O método de K-folds cross-validation é utilizado para conferir se o modelo não está sofrendo problemas de overfitting, é importante garantir que o modelo está aprendendo com os dados e não está somente decorando.
Na Tabela 5 temos então o resultados desse modelo, pode-se observar que os resultados para a classificação de vivos(0) não teve alteração, o modelo continua tendo 100% na precisão e no recall, mas na classificação de mortos(1) o modelo teve uma diminuição na precisão que agora está em 65% e também teve uma diminuição no recall que agora está em 23%, porém essa diminuição é justificada porque para o teste deste modelo foi utilizado a base completa e não somente a partição de teste que contém 10% da base total. Logo os resultados não tiveram uma alteração drástica e tiveram um comportamento bem semelhante, incluindo a acurácia que se manteve em 0.99 e a AUC que teve uma diminuição pequena também e ficou com 0.89.
50
Tabela 5. Resultados do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	6719191

Morto(1) 0.65 0.23 0.34 41031
Na Tabela 6 é possível ver a matriz de confusão do modelo que mostra o número exato de erros e acertos do modelo, pode-se observar que para a classe de vivos o modelo classificou corretamente 6.714.117 amostras da base de dados completa, o modelo classificou como 0 as amostras que nos rótulos realmente eram 0, errou apenas 5074 amostras. Já para a classe de mortos o modelo acertou somente 9552 amostras, classificou como 1 as amostras que no rótulo eram 1 também, e errou 31479. Isso mostra que o modelo está muito desbalanceado, pois está errado muito mais do que acertando as predições de mortos, sendo isso um reflexo da base de dados desbalanceada também. E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.
Tabela 6. Matriz de confusão do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6714117 5074
Mortos (1) 31479 9552
6.3 REGRESSÃO LOGÍSTICA
Para o algoritmo de Regressão Logística também foi utilizado métodos diferentes para o treinamento, O algoritmo foi treinado utilizando três formas diferentes: usando as partições 90% para treino e 10% para os testes, utilizando o método K-folds cross-validation e foi treinada utilizando o método RFECV que seleciona as melhores features para o modelo, juntamente com o método GridSearchCV. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.
51
6.3.1 Modelo de Regressão Logística Utilizando Particionamento 90/10
Para o primeiro experimento do algoritmo de regressão logística foi utilizado o particionamento 90/10 para o treinamento do algoritmo, sendo 90% da base de dados para realizar o treinamento do modelo e 10% da base para realizar os testes.
Na Tabela 7 temos os resultados do modelo de regressão logística utilizando o método de particionamento 90/10, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 65% que realmente pertenciam a classe 1, então cerca de 35% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 24% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem as amostras da classe 1 que seria dos recém-nascidos que poderiam vir a óbito, já para a classe de vivos a classe 0 o modelo está identificando muito bem e classificando de forma correta.E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.

Tabela 7. Resultados do modelo de regressão logística usando particionamento 90/10.
(Fonte: Elaboração Própria)
Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 671885
Morto(1) 0.65 0.24 0.35 4138
A Figura 30 abaixo mostra a curva ROC do modelo de regressão logística usando o particionamento 90/10, analisando a curva ROC podemos ver que a AUC da curva ficou em 0.93 e a acurácia do modelo ficou 0.99, mostrando que o modelo está acertando bastante as predições, porém esses números para o nosso modelo não mostra que ele está ótimo, com a análise da Tabela 7 acima podemos observar que o modelo está acertando muitas predições da classe de vivos e poucas predições da classe de mortos, como a base de dados é
52
desbalanceada, como a maior quantidade de dados está na classe de vivos e o modelo está acerto quase todos dessa classe a acurácia e a AUC ficam altas por esses acertos.Logo é preciso analisar mais resultados além da acurácia e da AUC do modelo.
Figura 30. Curva ROC modelo regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria).
Na Tabela 8 podemos analisar os número exatos que o modelo acertou de cada classe, como é mostrado na matriz de confusão o modelo acertou 671.435 da classe de vivos e acertou apenas 915 da classe de mortos, e errou mais que o triplo da classe de mortos. Então as predições do modelo estão muito desbalanceadas, para a classe de vivos o modelo está predizendo bem, porém para as classes de mortos o modelo está errando muitas predições.
Tabela 8. Matriz de confusão do modelo de regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 671435 504
Mortos (1) 3169 915
6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation
Para o segundo experimento do algoritmo de regressão logística foi utilizado o método K-folds cross-validation para o treinamento do modelo com a intenção de conferir e confirmar
53
que o modelo não esteja tendo problemas com overfitting e realmente esteja aprendendo com os dados que está sendo usado para o treinamento. O método K-folds cross-validation foi configurado para realizar uma separação de 10 partes iguais.
Na Tabela 9 podemos observar os resultados do modelo, e os resultados foram parecidos com o experimento anterior usando o particionamento 90/10, única diferença nos resultado é que no experimento anterior a precisão da classe de mortos ficou em 65% e no caso desse experimento usando o K-folds cross-validation a precisão ficou em 64%, mas os valores de recall e F1-Score se mantiveram, assim como todos os resultados da classe de mortos se mantiveram em 100%. A execução desse experimento mostra que o modelo realmente está aprendendo com os dados e não está apenas decorando, retirando o risco do modelo estar com overfitting. Mesmo sem alterações no resultado é importante saber que o

modelo não está com overfitting e está conseguindo aprender e encontrar padrões nos dados.
Tabela 9. Resultados do modelo de regressão logística usando K-folds cross-validation.
(Fonte: Elaboração Própria)
Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 6719191
Morto(1) 0.64 0.24 0.35 41031
Já na Figura 31 temos a curva ROC do modelo, essa curva ROC mostra a AUC de
todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do
método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas
para todas as vezes que foi realizado os testes, a AUC se manteve em 0.93 e 0.94, sendo a
média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não
se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua
acertando muito a classe de vivos que é a classe predominante na base de dados. Já era
esperado os resultados não sofrerem alterações grandes, pois a base de dados não sofreu
nenhuma alteração, esse experimento foi somente para conferir se a modelo não estava
sofrendo overfitting, e pelos resultados aparentemente a base não está com problemas de
sobreajuste.
54
Figura 31. Curva ROC modelo regressão logística usando K-folds cross-validation. (Fonte:
Elaboração Própria).
Na Tabela 10 temos a matriz de confusão do modelo, onde mostra que o modelo
acertou 6.713.658 amostras da classe de vivos e errou apenas 5.533, já para a classe de mortos
o modelo acertou somente 9.847 e novamente errou mais que o triplo errando 31.182
amostras. Como os resultados mostrados anteriormente não tiveram alteração era de se
esperar que as predições corretas e incorretas do modelo também não sofressem alterações, o
modelo continua acertando muito a classe de vivos e errando muito a classe de mortos,
mantendo as predições bem desbalanceadas.
55
Tabela 10. Matriz de confusão do modelo de regressão logística usando K-folds
cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6713658 5533
Mortos (1) 31182 9847
6.3.3 Modelo de Regressão Logística Utilizando GridSearchCV e RFECV
Para o terceiro e último experimento de regressão logística foi utilizado técnicas
diferentes juntas para tentar melhorar as predições do modelo, para esse experimento usamos
o K-folds cross-validation para o particionamento da base de dados assim como foi utilizado
no experimento acima, e também foi utilizado duas técnicas que ainda não foram usadas nos
experimentos anteriores que são as funções RFECV e GridSearchCV. A função RFECV
seleciona a quantidade ideal de features para ser usadas no treinamento do modelo e também
mostra quais são as melhores features para essa predição, então essa função ela utiliza a
validação cruzada para ir treinando e testando N vezes o modelo, e sempre vai alterando a

quantidade e as features utilizadas para o teste, então cada vez que a função testa os modelos ela grava a pontuação dos acertos e assim depois mostra o gráfico da Figura 32 e assim é possível ver quais são as features ideais para o modelo.
56
Figura 32. Resultado da função RFECV Base Brasil. (Fonte: Elaboração Própria).
Então os resultados mostrados na Figura 32 apresenta que o número ideal para treinar o modelo de features é 6, e as features sao: 'tp_maternal_schooling', 'gestaional_week', 'cd_apgar1', 'cd_apgar5', 'has_congenital_malformation', 'tp_labor'. Essas são as features que tiveram o melhor desempenho no RFECV, então para o treinamento do modelo desse experimento as features que vão ser utilizadas serão somente essas 6.
Após a execução da função de RFECV foi executado a função de GridSearchCV, o GridSearchCV é utilizado para descobrir quais são os melhores parâmetros para utilizar no algoritmo de regressão logística e criar os modelos, foi utilizado também a validação cruzada para testar qual seria os melhores parâmetros para melhorar o desempenho do modelo. Na Figura 33 pode-se ver que os parâmetros escolhidos pela função foram:
Figura 33. Resultado da função GridSearchCV. (Fonte: Elaboração Própria).
Na Tabela 11 podemos observar os resultados do modelo, e os resultados foram parecidos com os experimentos anteriores, para a classe de mortos a precisão teve um aumento e foi para 69% e o recall diminuiu chegando em 20%, então utilizando as novas
57
funções o modelo ganhou precisão e perdeu recall. Para a classe de vivos o modelo se manteve em 100% e não teve alterações para os resultados de precisão e recall. Mesmo com a utilização das funções de RFECV e GridSearchCV o modelo não teve uma melhora significativa para as predições de mortos.
Tabela 11. Resultados do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	6719191
Morto(1)	0.69	0.20	0.31	41031

Na Figura 34 temos a curva ROC do modelo, essa curva ROC mostra a AUC de todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas para todas as vezes que foi realizado os testes, a AUC se manteve em 0.92 e 0.93, sendo a média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua acertando muito a classe de vivos que é a classe predominante na base de dados.
58
Figura 34. Curva ROC modelo regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria).
Na Tabela 12 temos a matriz de confusão do modelo, onde mostra que o modelo acertou 6.715.535 amostras da classe de vivos e errou apenas 3.656, comparando com os experimentos anteriores de regressão logística as predições de vivos teve uma pequena piora e teve uma pequena diminuição nos acertos. Já para a classe de mortos, o modelo acertou 8.296 e errou 32.735 amostras, logo as predições de vivos tiveram uma uma piora e acertou um

pouco menos do que os outros experimentos, e na classe de vivos teve uma melhora acertando mais predições, porém como o problema está nas predições de mortos e para essa classe teve uma piora pode-se dizer que a função de GridSearchCV não teve um ganho significativo nos resultados. Então como modelo final foi escolhido o modelo utilizando somente o K-folds cross-validation, pois foi o modelo que mostrou o melhor desempenho dos experimentos de regressão logística.

59

Tabela 12. Matriz de confusão do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

Predito

Vivos (0) Mortos (1)

Classe Real

Vivos (0) 6715535 3656

Mortos (1) 32735 8296

60

7 CONCLUSÃO

O principal objetivo deste trabalho foi analisar os resultados das predições de mortalidade neonatal dos algoritmos de aprendizado de máquina usando uma base de dados com dados do Brasil inteiro. A partir destes resultados apresentados na secção anterior ?Resultados?, é possível concluir que analisar somente a AUC e a acurácia do modelo não é o suficiente para garantir que o modelo está excelente, sendo assim temos que ter mais atenção a precisão, recall e as predições mostradas na matriz de confusão.

Ambos os modelos ficaram desbalanceados nas predições, os modelos acertaram mais predições de vivos do que mortos, e as predições corretas de mortos foram mais baixas do que as incorretas, para os modelos ficarem melhor precisam passar por algoritmos que possam balancear essas predições, uma outra alternativa para melhorar o modelo é utilizar uma base de dados mais balanceada, pois essa base que foi utilizada é altamente desbalanceada.

Embora os dois modelos tenham tido resultados parecidos , o modelo de regressão logística utilizando somente o K-folds cross-validation se saiu melhor que os outros experimentos, o modelo criado na seção ?6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation? manteve a acurácia, AUC, precisão e recall dos demais modelos e acertou mais predições, talvez com algoritmos para melhorar as predições da classe de mortos, balanceando mais as predições, esse modelo fique com ótimas taxas preditivas.

Além disso podes-se afirmar que o peso ao nascer do recém nascido é um fator muito importante para as decisões dos modelos, ambos os modelos usaram muito essa informação para as predições, também podemos colocar, as colunas de presença de malformação e semana de gestação, como colunas que foram importantes para os modelos, essas causas podem ser evitadas se houver um acompanhamento médico com qualidade e eficiência.

61

REFERÊNCIAS

ÁRVORE de Decisão - Aula 3. Gravação de Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=ecYpXd4WREk&t=4089s. Acesso em: 1 jul. 2020.

BARRETO, J. O. M.; SOUZA, N. M.; CHAPMAN, E. Síntese de evidências para políticas de saúde: reduzindo a mortalidade perinatal. Ministério da Saúde, p. 1?44,2015.

BATISTA, André Filipe de Moraes; FILHO, Alexandre Dias Porto Chiavegatto. Machine Learning aplicado à Saúde. In: ZIVIANI, Artur; FERNANDES, Natalia Castro; SAADE, Débora Christina Muchaluat. SBCAS 2019: 19 Simpósio Brasileiro de Computação Aplicada à Saúde. [S. l.: s. n.], 2019. cap. Capítulo 1.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. Towards neonatal mortality risk classification: A data-driven approach using neonatal, maternal, and social factors, [s. l.], 23 out. 2020.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. SPNeoDeath: A demographic and epidemiological dataset having infant, mother, prenatal care and childbirth data related to births and neonatal deaths in São Paulo city Brazil ? 2012?2018, [s. l.], 19 jul. 2020.
BRESAN, Rodrigo. Um método para a detecção de ataques de apresentação em sistemas de reconhecimento facial através de propriedades intrínsecas e Deep Learning. Orientador: Me. Carlos Eduardo Beluzo. 2018. Trabalho de Conclusão de Curso (Superior de Tecnologia em Análise e Desenvolvimento de Sistemas) - Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas., [S. l.], 2018.
CABRAL, Cleidy Isolete Silva. Aplicação do Modelo de Regressão Logística num Estudo de Mercado. Orientador: João José Ferreira Gomes. 2013. Projeto Mestrado em Matemática Aplicada à Economia e à Gestão (Mestrado) - Universidade de Lisboa Faculdade de Ciências Departamento de Estatística e Investigação Operacional, [S. l.], 2013. Disponível em: https://repositorio.ul.pt/bitstream/10451/10671/1/ulfc106455_tm_Cleidy_Cabral.pdf. Acesso em: 1 maio 2021.
CAMPOS, Raphael. Árvores de Decisão: Então diga-me, como construo uma?. [S. l.], 28 nov. 2017. Disponível em: https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69. Acesso em: 23 jan. 2021.
Colab. 2021. Disponível em: https://colab.research.google.com/. Acesso em: 20 mar. 2021.
COX, D.R.;SNELL, E.J. Analysis of Binary Data.London: Chapman & Hall, 2ª Edição, 1989. HOSMER, D.W.; LEMESHOW, S.Applied Logistic Regression. New York: John Wiley, 2ª Edição, 2000.
E.França, S.Lansky. Mortalidade infantil neonatal no brasil: situação, tendências e perspectivas Rede Interagencial de Informações para Saúde - demografia e saúde: contribuição para análise de situação e tendências Série Informe de Situação e Tendências(2009), pp.83-112.
62
FREIRE, Sergio Miranda. Bioestatística Básica: Regressão Linear. [S. l.], 20 out. 2020. Disponível em: http://www.lampada.uerj.br/arquivosdb/_book/bioestatisticaBasica.html. Acesso em: 23 jan. 2021.
Jupyter. 2021. Disponível em: https://jupyter.org/. Acesso em: 20 jun. 2021.
GOODFELLOW, I. et al. Deep learning. [S.l.]: MIT press Cambridge, 2016.
Pandas. 2021. Disponível em: https://pandas.pydata.org/. Acesso em: 20 jun. 2021.
Python. 2021. Disponível em: https://www.python.org/. Acesso em: 20 jun. 2021.

KUHN, M.; JOHNSON, K. Applied predictive modeling. [S.l.]: Springer, 2013. v. 26.
LANSKY, S. et al. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante . Cadernos de Saúde Pública, Scielo, v. 30, p. S192 ? S207, 00 2014.
LANSKY, Sônia; FRICHE, Amélia Augusta de Lima; SILVA, Antônio Augusto Moura; CAMPOS, Deise; BITTENCOURT, Sonia Duarte de Azevedo; CARVALHO, Márcia Lazaro; FRIAS, Paulo Germano; CAVALCANTE, Rejane Silva; CUNHA, Antonio José Ledo Alves. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante e ao recém-nascido. [s. l.], Ago 2014. Disponivel em: https://www.scielosp.org/article/csp/2014.v30suppl1/S192-S207/pt/
Matplotlib. 2021. Disponível em: https://matplotlib.org/. Acesso em: 20 mar. 2021.
Maranhão AGK, Vasconcelos AMN, Trindade CM, Victora CG, Rabello Neto DL, Porto D,et al. Mortalidade infantil no Brasil: tendências, componentes e causas de morte no período de 2000 a 2010 . In: Departamento de Análise de Situação de Saúde, Secretaria de Vigilância em Saúde, Ministério da Saúde, organizador. Saúde Brasil 2011: uma análise da situação de saúde e a vigilância da saúde da mulher. v. 1. Brasília: Ministério da Saúde; 2012. p. 163-82.
MESQUITA, PAULO. UM MODELO DE REGRESSÃO LOGÍSTICA PARA AVALIAÇÃO DOS PROGRAMAS DE PÕS-GRADUAÇÃO NO BRASIL. Orientador: Prof. RODRIGO TAVARES NOGUEIRA. 2014. Dissertação (Mestre em Engenharia de Produção) - Universidade Estadual do Norte Fluminense, [S. l.], 2014.
MNASSRI , Baligh. Titanic: logistic regression with python. [S. l.], 1 ago. 2020. Disponível em: https://www.kaggle.com/mnassrib/titanic-logistic-regression-with-python. Acesso em: 14 jan. 2021.
Morais, R.M.d., Costa, A.L: Uma avaliação do sistema de informações sobre mortalidade. Saúde em Debate 41, 101?117 (2017). Disponível em: https://doi.org/10.1109/SIBGRAPI.2012.38.
MOTA, Caio Augusto de Souza; BELUZO , Carlos Eduardo; TRABUCO, Lavinia Pedrosa; SOUZA , Adriano; ALVES, Luciana; CARVALHO, Tiago. DESENVOLVIMENTO DE UMA PLATAFORMA WEB PARA CIÊNCIA DE DADOS APLICADA À SAÚDE MATERNO INFANTIL , [s. l.], 28 nov. 2019.
MULTIEDRO (Brasil). Conheça as aplicações do machine learning na área da saúde. [S. l.], 28 nov. 2019. Disponível em: https://blog.multiedro.com.br/machine-learning-na-area-da-saude/#:~:text=O%20machine%263
0learning%20na%20%C3%A1rea,diagn%C3%B3sticos%20e%20tratamentos%20mais%20precisos. Acesso em: 13 jun. 2021.
MUKHIYA,S.K.;AHMED,U. Hands-On Exploratory Data Analysis with Python: Perform EDA Techniques to Understand, Summarize, and Investigate Your Data. [S.l.]: PacktPublishingLtd,2020.ISBN978-1-78953-725-3.22,32
Murray CJ, Laakso T, Shibuya K, Hill K, Lopez AD. Can we achieve Millennium Development Goal 4? New analysis of country trends and forecasts of under-5 mortality to 2015. Lancet 2007; 370:1040-54.
NumPy. 2021. Disponível em:https://numpy.org/. Acesso em: 20 jun. 2021.
Oliveira, M.M.d., de Araújo, A.S.S.C., Santiago, D.G., Oliveira, J.a.C.G.d., Carvalho, M.D.,

de Lyra et al, R.N.D.: Evaluation of the national information system on live births in brazil , 2006-2010. Epidemiol. Serv. Saúde 24(4), 629?640 (2015).
PELISSARI, ANA CAROLINA CHEBEL. MORTALIDADE NEONATAL DA CIDADE DE SÃO PAULO: UMA ABORDAGEM UTILIZANDO APRENDIZADO DE MÁQUINA SUPERVISIONADO. 2021. Trabalho de Conclusão de Curso (Superior) - Instituto Federal de Educação, Ciência e Tecnologia Campus Campinas, [S. l.], 2021.
REGRESSÃO logística e binary cross entropy - Aula 7. Direção: Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=3J-LBtHVsm4. Acesso em: 21 jan. 2021.
Rmd Nascimento, AJM Leite, NMGSd Almeida, Pcd Almeida, Cfd Silva Determinantes da mortalidade neonatal: estudo caso-controle em fortaleza, ceará, brasil Cad Saúde Pública, 28(2012), pp.559 - 572.
SANTOS, Hellen Geremias. Machine learning e análise preditiva: considerações metodológicas: métodos lineares para classificação. In: SANTOS, Hellen Geremias. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina. 2018. Tese (Pós-Graduação em Epidemiologia) - Faculdade de Saúde Pública da Universidade de São Paulo, [S. l.], 2018.
Scikit-learn. 2021. Disponível em: https://scikit-learn.org/stable/. Acesso em: 20 jun. 2021.
Seaborn. 2021. Disponível em: https://seaborn.pydata.org/. Acesso em: 20 jun. 2021.
SILVA, Wesley; CORSO, Jansen; WELGACZ, Hanna; PEIXE, Julinês. AVALIAÇÃO DA ESCOLHA DE UM FORNECEDOR SOB CONDIÇÃO DE RISCOS A PARTIR DO MÉTODO DE ÁRVORE DE DECISÃO. ARTIGO ? MÉTODOS QUANTITATIVOS, [s. l.], 12 ago. 2008.
WHO, W. H. O.Women and health: today?s evidence, tomorrow?s agenda. [S.l.]: UNWorld Health Organization, 2009. ISBN 9789241563857.
64
APÊNDICE A

==========

Arquivo 1: monografia-V7.docx.pdf (9728 termos)
Arquivo 2: https://www.youtube.com/watch?v=Jv6wdM7HXOQ (27 termos)
Termos comuns: 0
Similaridade: 0,00%
O texto abaixo é o conteúdo do documento monografia-V7.docx.pdf (9728 termos)
Os termos em vermelho foram encontrados no documento
https://www.youtube.com/watch?v=Jv6wdM7HXOQ (27 termos)

==========

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO
CÂMPUS CAMPINAS
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
CAMPINAS
2021
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
Trabalho de Conclusão de Curso apresentado como
exigência parcial para obtenção do diploma do
Curso de Tecnologia em Análise e
Desenvolvimento de Sistemas do Instituto Federal
de Educação, Ciência e Tecnologia Câmpus
Campinas.
Orientador: Prof. Me. Everton Josué da Silva.
CAMPINAS
2021
Dados Internacionais de Catalogação na Publicação
CAIO AUGUSTO DE SOUZA MOTA
AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTALIDADE NEONATAL UTILIZANDO DADOS DO DATASUS
Trabalho de Conclusão de Curso apresentado
como exigência parcial para obtenção do diploma
do Curso de Tecnologia em Análise e
Desenvolvimento de Sistemas do Instituto
Federal de Educação, Ciência e Tecnologia de
São Paulo Câmpus Campinas.
Aprovado pela banca examinadora em: _____ de _____ de _____.
BANCA EXAMINADORA

_____
Prof. Me. Everton Josué da Silva (orientador)

IFSP Câmpus Campinas

_____

Prof. Me. Carlos Eduardo Beluzo
IFSP Câmpus Campinas

_____

Prof. Dr. Ricardo Barz Sovat
IFSP Câmpus Campinas

Dedico este trabalho aos meus familiares,
colegas de classe, professores e servidores do Instituto
que colaboraram em minha jornada formativa.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pelo dom da vida e pela oportunidade de concluir mais uma etapa de minha experiência acadêmica.

Agradeço a todos os professores e servidores do IFSP Câmpus Campinas, que contribuíram direta e indiretamente para a conclusão deste trabalho.

Agradeço também à minha família, que deu todo o apoio necessário para que eu chegasse até aqui.

Agradeço ao meu orientador que me auxiliou a solucionar as dificuldades encontradas no caminho.

"Minha energia é o desafio, minha motivação é o impossível,
e é por isso que eu preciso
ser, à força e a esmo, inabalável.".
Augusto Branco

RESUMO

A mortalidade infantil pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível em uma população. O aprendizado de máquina é uma área da inteligência artificial que propõe métodos de análise de dados que automatizam a construção de modelos analíticos com base em reconhecimento de padrões, baseado no conceito de que sistemas podem aprender com dados, identificando padrões e tomando decisões com o mínimo de intervenção humana. O objetivo deste trabalho é testar e analisar dois tipos diferentes de algoritmos de aprendizado de máquina, utilizando a base de dados do SIM e SINASC do Brasil, do período de 2016 até 2018, para gerar modelos para predição de mortalidade neonatal. Os métodos utilizados foram Árvore de Decisão e de Regressão Logística e como métricas para avaliar os modelos resultantes foram utilizadas a AUC, Curva ROC e a Matriz de Confusão. Como resultado, apesar de terem alcançado valores de AUC 0.93, ambos modelos de predições acertaram muitas predições de vivos cerca de 671.000 e erraram muitas predições de mortos cerca de 2.600, principalmente devido ao fato de a base estar desbalanceada.

Palavras-chave: Mortalidade Neonatal, Predição, Aprendizado de Máquina.

ABSTRACT

Infant mortality can be used to analyze poverty and socioeconomic levels, as well as measure the quality of health and medical technology available in a population. Machine learning is an area of artificial intelligence that proposes data analysis methods that automate the

construction of analytical models based on pattern recognition, based on the concept that systems can learn from data, identifying patterns and making decisions with a minimum of human intervention. The objective of this work is to test and analyze two different types of machine learning algorithms, using the SIM and SINASC do Brasil database, from 2016 to 2018, to generate models for predicting neonatal mortality. The methods used were Decision Tree and Logistic Regression and as metrics to evaluate the resulting models, AUC, ROC Curve and Confusion Matrix were used. As a result, despite achieving values of AUC 0.93, both prediction models correct many live birth predictions around 671.000 and miss many stillbirth predictions around 2600, mainly due to the fact that the base is unbalanced.
Keywords: Neonatal Mortality, Prediction, Machine Learning.
LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

SIM Sistema de Informação sobre Mortalidade
SINASC Sistema de Informações sobre Nascidos Vivos
TMI Taxa de Mortalidade Infantil
TMN Taxa de Mortalidade Neonatal
MN Mortalidade Neonatal
ML Machine Learning
DNV Declaração de Nascido Vivo
VN Verdadeiro Negativo
FN Falso Negativo
VP Verdadeiro Positivo
FP Falso Positivo
ROC Curva Característica de Operação do Receptor
NaN Não é um Número
RFE Eliminação Recursiva de Feature
RFECV Eliminação Recursiva de Feature com Validação Cruzada

SUMÁRIO

13
1 INTRODUÇÃO

A taxa de mortalidade infantil (TMI) é uma importante medida de saúde em uma população e é um grande problema no mundo todo. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos, assim como medir qualidade de saúde e tecnologia médica disponível. Esta taxa é dividida em duas categorias: neonatal e pós-neonatal. É categorizada neonatal quando o óbito ocorre nos 28 primeiros dias de vida após o pós-parto, e o pós-neonatal é quando o óbito ocorre entre 29 e 364 dias de vida (BELUZO et al., 2020). Cerca de 46% das mortes no mundo acontecem entre os menores de cinco anos e grande parte está concentrada nos primeiros dias de vida (LANSKY, 2014). A diminuição dessas taxas é muito importante no mundo todo, refletindo nos indicadores de saúde pública e desenvolvimento do país.

Os fatores associados à mortalidade são profundamente influenciados pelas características biológicas maternas e neonatais, pelas condições sociais e pelos cuidados prestados pelos serviços de saúde (E. FRANÇA; S. LANSKY, 2009; R.M.D. NASCIMENTO, 2012). Em grande parte, o diagnóstico é altamente dependente da experiência adquirida pelo

profissional que o realiza. Apesar de indiscutivelmente necessário, não é perfeito, principalmente pelo fato de ser dependente de fatores humanos, por este motivo os profissionais sempre tiveram recursos tecnológicos para auxiliar nessa tarefa (BELUZO et al., 2020). Segundo estudos de 2010 no Brasil, calcula-se que aproximadamente 70% dos óbitos infantis ocorridos poderiam ter sido evitados por ações de saúde e que 60% dos óbitos neonatais ocorreram por situações que poderiam ser evitadas (BARRETO; SOUZA; CHAPMAN, 2015).

No presente trabalho foram utilizados dois algoritmos de aprendizado de máquina para criar modelos de predição de mortalidade neonatal. Além disso, foi realizada também uma análise exploratória da base de dados. Para este trabalho foram utilizadas duas fontes de dados: Sistema de Informação sobre Mortalidade (SIM) e Sistema de Informações sobre Nascidos Vivos (SINASC) do Brasil inteiro.

14

2 JUSTIFICATIVA

Com o avanço da tecnologia, podemos notar que ela está cada vez mais presente no nosso dia a dia e nos auxilia em diversas tarefas do cotidiano, e vem sendo aplicada em diversas áreas, dentre elas a saúde.

A mortalidade infantil é uma preocupação mundial na saúde pública, a ONU (Organizações das Nações Unidas) definiu a redução da mortalidade infantil como uma meta para o desenvolvimento global. A mortalidade neonatal é responsável por aproximadamente 60% da TMI nos países em desenvolvimento (A.K. SINGHA, 2016). Existem diversos fatores que podem contribuir com a redução de óbitos neonatais, como por exemplo acompanhamento médico à gestante no período de gravidez, acompanhamento médico ao recém-nascido nos primeiros dias de vida e a disponibilização deste serviço com qualidade e proficiência (WHO, 2009).

A diminuição dessa taxa é importante e de interesse para o mundo todo, e utilizar da tecnologia para auxiliar nessa diminuição é uma proposta funcional, e inovadora para a realidade brasileira (MOTA et al., 2019). Havendo disponibilidade de tecnologia para auxiliar nas decisões dos diagnósticos, os profissionais da saúde podem focar nos cuidados do recém-nascido com risco de vir a óbito, fornecendo tratamentos melhores.

Segundo o site Multiedro (2019), um grande problema que os médicos enfrentam ao realizar o diagnóstico, é analisar um grande volume de dados. Para a área médica obter diagnósticos rápidos e precisos pode salvar vidas, e a utilização de machine learning para realizar esses processos é importante, com a ML os dados podem ser processados em questões de segundos e resultar em um diagnóstico mais rápido, além disso utilizando bons modelos ML os diagnósticos serão mais precisos.

Diversos trabalhos vêm sendo desenvolvidos neste contexto. No trabalho realizado por BELUZO et al. (2020a), os autores utilizaram uma base de dados com informações da cidade de São Paulo para criação de modelos de aprendizado de máquina para predição de mortalidade neonatal. Este recorte foi utilizado também no presente trabalho para fins de exercício e definição de etapas da implementação de código. Na conclusão os autores discutem os fatores que tiveram mais influência nas predições, e na análise feita pelos autores, as consultas de pré-natal são muito importantes para a redução da mortalidade infantil, uma vez que algumas características muito importantes, como a malformação congênita, podem

ser detectadas ao longo das consultas de pré-natal.
Em um outro estudo realizado por BELUZO et al. (2020a), foi realizada uma análise exploratória da base de dados SPNeoDeath, onde e podemos observar detalhadamente cada
15
análise das colunas da tabela e diversos gráficos feito para um melhor entendimento dos dados.
Outro trabalho que foi realizado utilizando a base de dados com dados somente de São Paulo foi o de PELISSARI (2021), nesse trabalho é realizado uma análise exploratória na base de dados de São Paulo e é também analisado três algoritmos de aprendizado de máquina (Logistic Regression, Random Forest Classifier e XGBoost), nesse trabalho a autora conclui que os modelos de Logistic Regression e XGBoost foram os modelos que apresentaram os melhores desempenhos preditivos, e também mostra os problemas de estar utilizando uma base de dados desbalanceada.
O problema da mortalidade neonatal não é recente, e o mundo todo desenvolve iniciativas para lidar com ele. A ONU definiu como meta a redução da mortalidade infantil para o desenvolvimento global. Diversos fatores podem ajudar na diminuição da taxa de mortalidade neonatal como acompanhamento médico antes e após parto, tanto com a mãe quanto com o recém-nascido. Um serviço de qualidade e constante para os casos com maior risco de óbito pode salvar diversos recém-nascidos. Neste sentido, o desenvolvimento de ferramentas para a predição de mortalidade neonatal pode colaborar com esta iniciativa.

3 OBJETIVOS

3.1 OBJETIVO GERAL

O presente trabalho tem como objetivo testar e analisar os resultados da aplicação dos aprendizagem de máquina supervisionado Árvore de Decisão e Regressão Logística, utilizando dados do SIM e do SINASC, no período de 2016 até 2018, a fim de gerar modelos de predição de risco morte neonatal.

3.2 OBJETIVOS ESPECÍFICOS

? Realizar análise exploratória da base dados a ser utilizada na construção dos modelos;
? Implementar modelos de predição utilizando algoritmos Árvore de Decisão e Regressão Logística;
? Avaliar resultados e propor novas abordagens.
16

4 FUNDAMENTAÇÃO TEÓRICA

4.1 MORTALIDADE INFANTIL E NEONATAL

Como dito na monografia PELISSARI (2021), a TMI consiste no número de crianças que foram a óbito antes de concluir um ano de vida dividido por 1000 crianças nascidas vivas no tempo de um ano. A TMI pode ser usada para analisar níveis de pobreza e socioeconômicos e qualidade da saúde pública de um país (apud BELUZO et al., 2020).
Mencionado em PELISSARI (2021), a TMN é uma das duas categorias da TMI, é categorizado mortalidade neonatal quando o óbito ocorre do nascimento da criança até os 28 primeiros dias de vida. A mortalidade neonatal pode ser subdividida em mortalidade neonatal precoce que é quando o óbito ocorre entre 0 e 6 dias de vida, e o neonatal tardio quando o óbito ocorre entre 7 e 28 dias de vida (apud E. FRANÇA e S. LANSKY, 2009).
Segundo Lansky et al. (2014), a taxa de mortalidade neonatal é o principal

componente da TMI desde a década de 1990 no Brasil e vem se mantendo em níveis elevados, com taxa de 11,2 óbitos por mil nascidos vivos em 2010 (apud Maranhão et al., 2012). Também é dito em Lansky et al. (2014), que a TMI do Brasil em 2011 foi 15,3 por mil nascidos vivos, alcançando a meta 4 dos objetivos de desenvolvimento do milênio, compromisso dos governos integrantes das Nações Unidas de melhorar a saúde infantil e reduzir em 2/3 a mortalidade infantil entre 1990 e 2015. (apud Maranhão et al., 2012; apud Murray et al., 2015). Segundo Lansky et al. (2014) o principal componente da TMI em 2009 é a mortalidade neonatal precoce, e grande parte das mortes infantis acontece nas primeiras 24 horas (25%), indicando uma relação estreita com a atenção ao parto e nascimento (apud E. FRANÇA e S. LANSKY, 2009).

4.2 MÉTODOS DE APRENDIZADO DE MÁQUINA

A utilização de dados para a resolução de problemas, de modo a se identificar padrões ocultos e posteriormente auxiliar na tomada de decisões é definida como aprendizado de máquina (BRESAN, 2018 apud Brink et al. 2013).

O uso de técnicas de Aprendizado de Máquina se mostra altamente eficiente na resolução de tarefas que se apresentem difíceis de se resolver a partir de programas escritos por humanos (BRESAN, 2018 apud GOODFELLOW et al., 2016), bem como também possibilita uma maior compreensão sobre os funcionamentos dos princípios que compõem a inteligência humana.

17

Neste trabalho serão implementados modelos utilizando os algoritmos de Árvore de Decisão e Regressão Logística, os quais serão apresentados a seguir.

4.2.1 Árvore de Decisão

Segundo Batista e Filho (2019), os modelos preditivos baseados em árvores são geralmente utilizados para tarefas de classificação, embora também possam ser utilizados para tarefas de regressão. Considerado um dos mais populares algoritmos de predição, o algoritmo de árvore de decisão proporciona uma grande facilidade de interpretação.

As árvores de decisão utilizam a estratégia ?dividir-e-conquistar", na qual as árvores são construídas utilizando-se apenas alguns atributos. A árvores de decisão é uma das técnicas por meio da qual um problema complexo é decomposto em subproblemas mais simples. Recursivamente, a mesma estratégia é aplicada a cada subproblema (SILVA et al, 2008).

A árvore de decisões tem uma estrutura hierárquica onde cada nó da árvore representa uma decisão em uma das colunas do conjunto de dados, cada ramo da árvore representa uma tomada de decisão (BATISTA; FILHO, 2019).

O algoritmo segue algumas decisões para montar a árvore, o atributo mais importante é utilizado como nó raiz e será o primeiro nó, já os atributos menos importantes são mostrados nos ramos da árvore. Cada atributo é verificado para conferir se é possível gerar uma árvore menor e se é possível fazer uma classificação melhor, e assim é escolhido o nó que produz nós filhos (SILVA et al, 2008).

Existem diferentes tipos de algoritmo para a construção da árvore, como por exemplo, ID3 (Iterative Dichotomiser), C4.5 e CART (Classification and Regression Tree). O algoritmo ID3 escolhe os nós da árvore por meio da métrica do ganho de informação. Já o CART faz uso da equação de Gini (BATISTA; FILHO, 2019 apud KUHN; JOHNSON, 2013).

O algoritmo utilizado neste trabalho usará o cálculo da entropia e ganho de

informação, para decidir qual atributo será usado no nó, a entropia é uma medida da desorganização dos sistemas, maior é a incerteza do sistema, quanto maior a entropia maior está a desorganização, com a árvore de decisão a ideia é ir organizando o sistema e ir separando as decisões da melhor forma para que a entropia diminui e o ganho de informação aumente, para que a decisão do modelo fiquei mais assertiva. O cálculo da entropia utiliza a seguinte equação (ÁRVORE, 2020):

???????? =

?

?? ?? ???

2

??

18

Nesta equação o i representa possíveis classificações (rótulos), e o P é a probabilidade de cada uma. A ideia da árvore então é verificar qual é a entropia para cada uma das variáveis da base de dados, e verificar qual é a melhor organização de cada uma das variáveis. E isso é realizado através da técnica chamada de ganho de informação, essa técnica utiliza a equação a seguir (ÁRVORE, 2020):

????? = ????????(???) ? ? ????(??????) * ????????(??????)

Então o ganho de informação é calculado pela entropia do nó pai menos a soma do peso vezes a entropia dos nós filhos. Esse cálculo é realizado para todas as variáveis e a variável que tiver maior ganho de informação é utilizado para a decisão do nó. Esse processo é realizado recursivamente e a árvore vai sendo montada com base nesses cálculos (ÁRVORE, 2020).

Na Figura 1 temos a representação do início de uma árvore de decisão, no caso em questão o autor estava classificando exames positivos e negativos para o vírus SARS-CoV-2, ele utilizou uma base de dados que contém quase 6 mil exames realizados, contendo campos como o resultado do exame, idade do paciente, níveis de hemoglobina entre outras informações.

Figura 1. Exemplo de Árvore de Decisão. (Fonte: ÁRVORE 2020).

Pode-se observar que para o nó raiz foi utilizado a variável Leukocytes, logo foi o que apresentou a melhor organização para ser o nó raiz, e após ele temos os ramos da árvore. Cada retângulo da árvore é uma decisão do modelo e esses retângulos têm atributos importante como, a coluna da base de dados que será onde vai ser realizado a decisão, então pegando o nó raiz, pacientes que tiverem com os Leukocytes para menos que -0.43 seguiram o caminho

19

para a esquerda (true), já os pacientes que tiverem mais seguiram o caminho da direita (false). A entropia calculada do nó também é apresentada e a classificação do nó também, então se o resultado parar neste nó, a classificação que está nele será a classificação final, e temos o samples que é o número de amostras que se enquadram no na decisão. E esse processo de verificação do modelo vai se repetindo até não ter mais como dividir em subproblemas ou até chegar na profundidade máxima.

Uma característica importante da árvore de decisão é que ela é um modelo WhiteBox, ou seja, é possível visualizar a árvore montada e as decisões que o modelo terá que tomar na árvore. Na Figura 1 pode-se ver uma árvore de decisões montada, cada retângulo da árvore é

uma decisão que o modelo terá que tomar, conforme o modelo toma as decisões ele vai seguindo um caminho até chegar ao nó folha, nó folha é o nó que não contém nó filho, não tem mais ramos para baixo, chegando ao nó folha o modelo tem a classificação final.

4.2.2 Regressão Logística

O modelo de regressão logística estabelece uma relação entre a probabilidade de ocorrência da variável de interesse e as variáveis de entrado do modelo, sendo utilizada para tarefas de classificação, em que a variável de interesse é categórica, neste caso, a variável de interesse assume valores 0 ou 1, onde 1 é geralmente utilizado para indicar a ocorrência do evento de interesse (Batista e Filho, 2019).

De acordo com Mesquita (2014), a técnica de regressão logística, desenvolvida no século XIX, obteve maior visibilidade após 1950 ficando então mais conhecida. Ficou ainda mais difundida a partir dos trabalhos de Cox & Snell (1989) e Hosmer & Lemeshow (2000). Caracteriza-se por descrever a relação entre uma variável dependente qualitativa binária, associada a um conjunto de variáveis independentes qualitativas ou métricas.

Esta técnica inicialmente foi utilizada na área médica, porém a eficiência viabilizou sua implementação nas mais diversas áreas. Mesquita (2014) diz que o termo regressão logístico, tem sua origem na transformação usada com a variável dependente, que permite calcular diretamente a probabilidade da ocorrência do fenômeno em estudo.

Segundo Santos (2018), para estimar a probabilidade, é utilizado uma técnica chamada distribuição binomial para modelar a variável resposta do conjunto de treinamento, que tem como parâmetro a probabilidade de ocorrência de uma classe específica. Uma característica importante desse modelo é que a probabilidade estimada deve estar limitada ao intervalo de 0 a 1. O algoritmo de Regressão Logística gera um modelo de classificação binária (0 e 1). O modelo utiliza a Regressão Linear como base, a equação linear é (REGRESSÃO, 2020):

20

? = ? . ? + ?

O ?m? da equação é o coeficiente angular da reta, e o ?b? é o intercepto. Então, aplicando a equação linear obtemos um valor contínuo como resultado, após obter esse valor a Regressão Logística tem que realizar a classificação do resultado, então é aplicado uma função chamada sigmoide (REGRESSÃO, 2020):

? = 1

1+ ? ??

Essa função chamada de sigmoide, ou função logística, é responsável por ?achatar? o resultado da regressão linear, calculando a probabilidade de o resultado pertencer a classe 1, classificando assim o resultado da função linear em 0 ou 1.

Segundo Batista e Filho (2019), o modelo de regressão logística que tem como objetivo realizar uma classificação binária de uma variável de interesse irá predizer a probabilidade de pertencer à classe positiva. Tem-se, portanto, que é dado por:?

? = { 0, ?? ???? (? = 1 ??) < 0, 5

ou

? = { 1, ?? ???? (? = 1??) ? 0, 5

Então a decisão do modelo de regressão logística está relacionada à escolha de um ponto de corte para p(x). Caso o ponto de corte escolhido for p(x)=0,5, os resultados que forem p(x)>0,5 serão classificados como 1 e os resultados p(x)<5 serão classificados como 0

(SANTOS, 2018).
A Figura 2 ilustra um exemplo de como é o diagrama de Regressão Logística. Pelo diagrama então pode-se observar que o modelo possui uma representação gráfica em formato de ?S?, essa seria a curva logística. As predições do modelo sempre ficaram na linha 1 e 0 do eixo y, pois é o intervalo estabelecido pelo algoritmo, então a classificação sempre será nesse intervalo.
21
Figura 2. Exemplo de Diagrama de Regressão Linear. (Fonte: Batista e Filho 2019).
4.2.3 Treino, teste e validação do modelo de aprendizado de máquina
A criação de um modelo de aprendizado de máquina é realizada em duas etapas: treinamento do modelo, que é realizado a partir dos dados fornecidos para o algoritmo, e etapa de teste, onde os modelos recebem dados novos e sem o rótulo, para que seja avaliado a performance do modelo (PELISSARI, 2021).
Para o treinamento dos modelos foram utilizadas duas abordagens, a primeira onde a base foi particionada entre conjunto de dados para teste e para treino em uma porcentagem de 90% para treino e 10% para teste, porém dessa forma pode ocorrer problema de sobreajuste ou overfitting. Nesta situação, o que pode ocorrer é o modelo não aprender com os dados, e apenas ?decorar? os dados recebidos e quando é realizado o teste o modelo consegue predizer apenas situações parecidas, não sendo capaz de identificar padrões com diferenças mínimas.
Para evitar este problema, foi utilizado uma técnica de validação cruzada chamada K-folds cross-validation. Essa técnica de validação cruzada divide a base de dados em K subconjuntos, e então são realizadas várias rodadas de treinamento e teste alternando os subconjuntos utilizados para teste e treino, e o resultado do modelo baseia-se na média da acurácia observada em cada rodada (PELISSARI, 2021).
Pode-se observar na Tabela 1 uma ilustração da técnica K-fold cross-validation, onde temos um exemplo onde a base é dividida em 5 subconjuntos e cada subconjunto tem a parte de teste diferente dos outros subconjuntos, assim o modelo vai receber valores novos de teste e treino todas as vezes que executar um novo subconjunto.
22
Tabela 1. Exemplo da separação K-fold cross-validation. (Fonte: Adaptado de PELISSARI, 2021).

Predição 1	Predição 2	Predição 3	Predição 4	Predição 5
Teste	Treino	Treino	Treino	Treino
Treino	Teste	Treino	Treino	Treino
Treino	Treino	Teste	Treino	Treino
Treino	Treino	Treino	Teste	Treino
Treino	Treino	Treino	Treino	Teste

4.2.4 Métricas para avaliação de modelos
Para fins de avaliação, neste trabalho foram utilizadas três métricas para avaliar o desempenho dos modelos: a Matriz de Confusão e a AUC (Area under Receiver Operating Characteristic Curve - ROC). Na matriz de confusão pode-se observar o comportamento do modelo em relação a cada uma das classes a serem preditas pelo modelo, utilizando variáveis binárias (positivo: 1; e negativo: 0), para apresentar uma matriz que faz uma comparação entre as predições do modelo e os valores corretos do conjunto de dados (PELISSARI, 2021).

Na Tabela 2 temos um exemplo de uma matriz de confusão, a qual consiste em duas colunas e duas linhas, e os valores são atribuídos nos quadrantes com os seguintes resultados (PELISSARI, 2021):

? quando o valor real do conjunto de dados é 0, e a predição do modelo também classificou como 0, então temos um Verdadeiro Negativo (VN);

? quando o valor real é 1, e o modelo classificado como 0, temos um Falso Negativo (FN);

? quando o valor real é 0, e o modelo classificado como 1, então o resultado se enquadra no quadrante de Falso Positivo (FP);

? e por fim o quadrante Verdadeiro Positivo (VP), que são os resultados que tem o valor real 1, e o modelo classificado como 1.

23

Tabela 2. Modelo de Matriz de confusão. (Fonte: Adaptado de PELISSARI, 2021).

Valor Predito
Negativo (0) Positivo (1)
Classe
Real
Negativo
(0)
Verdadeiro
Negativo
Falso
Positivo
Positivo
(1)
Falso
Negativo
Verdadeiro
Positivo

A segunda métrica de avaliação utilizada foi a Curva ROC que permite avaliar o desempenho de um modelo com relação às predições efetuadas. A curva ROC é um gráfico de Sensibilidade (taxa de verdadeiros positivos) versus taxa de falsos positivos, a representação da curva ROC permite evidenciar os valores para os quais existe otimização da Sensibilidade em função da Especificidade (CABRAL, 2013).

Na Figura 3 temos um exemplo de uma curva ROC. A linha traçada na cor preta é uma linha de referência, ela representa hipoteticamente a curva ROC de um classificador puramente aleatório. Caso a linha laranja, que representa o resultado do modelo que está sendo avaliado, ficasse igual a linha tracejada preta, indicaria que o modelo avaliado estaria predizendo aleatoriamente os resultados.

Figura 3: Exemplo de Curva ROC. (Fonte: Elaboração Própria).

Por uma análise de inspeção visual, quanto mais próximo a linha laranja estiver do valor 1 no eixo Y, melhor está a capacidade do modelo para diferenciar as classes. O valor da

24

AUC representa a capacidade do modelo de diferenciar as classes, quanto maior o valor do

AUC, melhor é a predição realizada pelo modelo, uma vez que, os valores resultantes iguais a 0 são realmente 0, e os iguais a 1 são de fato 1 (PELISSARI, 2021).

5 METODOLOGIA

O desenvolvimento deste trabalho foi realizado em três etapas: (1) Pré-processamento e Análise Exploratória do conjunto de dados, onde foram aplicadas técnicas comuns para preparar o conjunto de dados como tratamento de valores nulos e seleção de variáveis de interesse; (2) Implementação de modelos de aprendizado de máquina supervisionado para predição das mortes neonatais, utilizando os algoritmos de árvore de decisão e regressão logística; e (3) Análise de Resultados, que será realizada uma análise do desempenho do modelo.

5.1 TECNOLOGIAS E FERRAMENTAS

Neste trabalho vamos utilizar a linguagem de programação Python (PYTHON, 2021) pela sua simplicidade de codificação e alto poder de processamento. Foi utilizado também a plataforma Google Colab Research (Colab, 2021), para o desenvolvimento dos algoritmos e treinamento dos modelos. Além disso, foi utilizado também o Jupyter Notebook (Jupyter, 2021), instalado localmente em computador pessoal, pois o Google Colab Research possui limitação de recursos de memória e processamento, então as duas plataformas foram usadas para realização deste trabalho. As principais bibliotecas utilizadas foram: numpy (NumPy, 2021), e pandas (Pandas, 2021), para a manusear os dados, matplotlib (Matplotlib, 2021), e seaborn (Seaborn, 2021), para a plotagem dos gráficos, e por fim a biblioteca sklearn (Scikit-Learn, 2021), que é a biblioteca responsável pelos algoritmos de aprendizado de máquina.

5.2 BASE DE DADOS

As bases de dados a serem utilizadas serão o SIM e SINASC, que são as duas principais fontes de informações sobre nascimentos e óbitos no Brasil. O SINASC é alimentado com base na Declaração de Nascido Vivo (Declaração de Nascido Vivo - DNV) (BELUZO et al.,2020b apud Oliveira et al., 2015). O SIM tem como objetivo principal apoiar a coleta, armazenamento e processo de gestão de registros de óbitos no Brasil (BELUZO et 25
al.,2020b apud Morais et al., 2017), e foi usado para rotular o óbito registrado no SIM, utilizando o campo DNV como chave de associação. O SIM será utilizado para rotular os registros do SINASC, pois o SIM coletas informações sobre mortalidade e é usado como base para o cálculo de estatísticas vitais, como a taxa de mortalidade neonatal e o SINASC reúne informações sobre dados demográficos e epidemiológicos do bebê, da mãe, do pré-natal e do parto (BELUZO et al.,2020b).

A Tabela 3 apresenta as variáveis da base de dados. Essa base tem variáveis que contêm valores quantitativos e categóricos. A coluna ?num_live_births? por exemplo contém o número de nascidos vivos anteriores da mãe e é composta por valores quantitativos contínuos, e a coluna ?tp_pregnancy? utiliza valores nominais categóricos que indicam o tipo de gravidez.

Tabela 3. Colunas da base de dados e suas descrições. (Fonte: Adaptado de BELUZO et al.,2020b).

Nome da coluna	Descrição	Domínio de dados
Variáveis	demográficas e socioeconômicas	

maternal_age Idade da mãe Quantitativo Contínuo (inteiro)
tp_maternal_race Raça / cor da pele da
mãe
Nominal categórico (inteiro)
1 - branco; 2 - Preto; 3 -
amarelo; 4 - Pele morena; 5 -
Indígena.
tp_marital_status Estado civil da mãe Nominal categórico (inteiro)
1 - Único; 2 - Casado; 3 - Viúva;
4 - Separados / divorciados
judicialmente; 5 - Casamento
por união estável; 9 - Ignorado.
tp_maternal_schooling Anos de escolaridade da
mãe
Nominal categórico (inteiro)
1 - nenhum; 2 - de 1 a 3 anos; 3 -
de 4 a 7 anos; 4 - de 8 a 11 anos;
5 - 12 e mais; 9 - Ignorado.
Variáveis obstétricas maternas
num_live_births Número de nascidos
vivos
Quantitativo Contínuo (inteiro)
num_fetal_losses Número de perdas fetais Quantitativo Contínuo (inteiro)
num_previous_gestations Número de gestações Quantitativo Contínuo (inteiro)
26
anteriores
num_normal_labors Número de partos
normais (trabalhos de
parto)
Quantitativo Contínuo (inteiro)
num_cesarean_labors Número de partos
cesáreos (partos)
Quantitativo Contínuo (inteiro)
tp_pregnancy Tipo de gravidez Nominal categórico (inteiro)
1 - Singleton; 2 - Twin; 3 -
Trigêmeo ou mais; 9 - Ignorado.
Variáveis relacionadas a cuidados anteriores
tp_labor Tipo de parto infantil
(tipo de parto)
Categórico Nominal (inteiro)
1 - Vaginal; 2 - Cesariana;
num_prenatal_appointments Número de consultas de
pré-natal
Quantitativo Contínuo (inteiro)

tp_robson_group Classificação do grupo
Robson
Ordinal categórico (inteiro)
Variáveis relacionadas ao recém-nascido
tp_newborn_presentation Tipo de apresentação de
recém-nascido
Categórico Nominal (inteiro)
1 - Cefálico; 2 - Pélvico ou
culatra; 3 - Transversal; 9 -
Ignorado.
has_congenital_malformation Presença de
malformação congênita
Nominal categórico (inteiro)
1 - Sim; 2 - Não; 9 - Ignorado
newborn_weight Peso ao nascer em
gramas
Quantitativo Contínuo (inteiro)
cd_apgar1 Pontuação de Apgar de 1
minuto
Ordinal categórico (inteiro)
cd_apgar5 Pontuação de Apgar de 5
minuto
Ordinal categórico (inteiro)
gestaional_week Semana de gestação Quantitativo Contínuo (inteiro)
tp_childbirth_care Assistência ao parto Categórico Nominal (inteiro)
1 - Doutor; 2 - enfermeira ou
obstetra; 3 - Parteira; 4 - outros;
9 - Ignorado.
was_cesarean_before_labor Foi cesáreo antes do
parto.
27
was_labor_induced Foi induzido ao trabalho
de parto.
is_neonatal_death Morte antes de 28 dias
(rótulo)
Nominal categórico (inteiro)
0 - sobrevivente; 1 - morto.

Neste trabalho foi utilizado dados do período de 2014 à 2016 devido sua melhor qualidade. Este recorte possui 6.760.222 registros dos quais 99.4% são amostras de recém-nascidos vivos e 0.6% são amostras onde os recém-nascidos vieram a óbito conforme pode ser observado na Figura 4. Com essas informações consegue-se observar que a base de dados é desbalanceada.

Figura 4. Distribuição da base de dados considerando se o recém-nascido viveu ou morreu durante o período neonatal. (Fonte: Elaboração Própria).

5.3 PREPARAÇÃO DA BASE DE DADOS

Nessa etapa foi realizada a preparação da base de dados para o treinamento dos modelos que consiste na limpeza, transformação e preparação dos dados a serem utilizados na análise exploratória e na criação dos modelos preditivos. A base de dados pode conter informações desnecessárias para o modelo que pode acabar atrapalhando as predições,
28
algumas colunas podem ser retiradas ou seus valores podem ser modificados, por exemplo valores reais são modificados para inteiros.

5.4 DESENVOLVIMENTO DOS MODELOS DE APRENDIZADO DE MÁQUINA

Como já dito anteriormente na seção 4 os algoritmos de aprendizado de máquina escolhidos para esse trabalho foram os: Regressão Logística e Árvore de Decisão que utilizam a lógica mostrada na seção 4 , Fundamentação Teórica. Esses dois foram escolhidos por alguns motivos, ambos são algoritmos de classificação e supervisionados, atendendo os critérios necessários para a predição de mortalidade neonatal, ambos os algoritmos são bons para realizar predições, a Árvore de Decisão inclusive é um ótimo modelo para analisar as decisões que estão sendo feitas pelo modelo, afinal esse algoritmo tem a característica de ser um WhiteBox, ou seja conseguimos ver o que o modelo aprendeu e o que ele está decidindo. Esses algoritmos selecionados são algoritmos que se encaixam na necessidade deste trabalho e são muito utilizados na comunidade acadêmica.

Para o algoritmo de Regressão Logística, como foi dito anteriormente na seção ?Preparação da base de dados? a base de dados foi tratada e particionada, no particionamento foi utilizado 90% da base para o treino e 10% para os testes e então foi realizado o treinamento do modelo, também foi aplicada a função RFECV (Eliminação Recursiva de Feature com Validação Cruzada), esse função executa a RFE (Eliminação Recursiva de Feature), que tem como ideia selecionar features considerando recursivamente conjuntos cada vez menores de features, isso ocorre da seguinte forma. Primeiro, o estimador é treinado no conjunto inicial de features e a importância de cada feature é obtida por meio de um atributo "coef" ou por meio de um atributo "feature_importances". Em seguida, as features menos importantes são removidas do conjunto atual de features. Esse procedimento é repetido recursivamente no conjunto removido até que o número desejado de features a serem selecionados seja finalmente alcançado. Porém o diferencial da RFECV é que essa função executa a RFE em um loop de validação cruzada para encontrar o número ideal ou o melhor número de features.

Após essa seleção, foi realizado o treinamento do modelo usando todas as features e também treinamos usando as features que o RFECV selecionou, para compararmos os resultados no final. Também foi utilizado o método de K-folds cross-validation em um novo treinamento para conferirmos se o modelo estava sofrendo overfitting (sobreajuste). E por fim aplicamos um método chamado GridSearchCV, esse método permite que seja realizado testes
29
alterando algumas combinação de parâmetros nos nossos modelos, facilitando para achar a melhor combinação, esse método é parecido com o cross validation, o diferencial do Grid Search é que ele faz os cross validation de vários modelos com hiperparâmetros diferentes de uma só vez.

Para o algoritmo de Árvore de Decisão também foi utilizado o mesmo

particionamento de 90% para treino e 10% para teste, na criação do modelo foi colocado a entropy como critério de decisão e uma profundidade máxima de 4, pois com números maiores o modelo ficava muito complexo e ele errava mais as predições, também utilizamos um algoritmo para mostrar a importância de cada Feature para o modelo.
O algoritmo de Regressão Logística e análise exploratória foi baseado em códigos disponível na plataforma web Kaggle (MNASSRI, 2020), no exemplo do Kaggle o autor faz os códigos para predizer mortes no navio Titanic, para o trabalho de mortalidade neonatal os códigos foram alterados para realizar predição de mortes neonatais. Já no algoritmo de Árvore de Decisão foi baseado em uma vídeo aula do professor Diogo Cortiz (ÁRVORE..., 2020), nessa vídeo aula o professor explica sobre os algoritmos de Árvore de Decisão e depois implementa o um exemplo de código, o código usado neste trabalho usou o codigo do professor Diogo Cortiz, e foi alterado para realizar predições de mortalidade neonatal.

5.5 ANÁLISE DOS RESULTADOS

Para a análise de resultados foi utilizado dois métodos que é o de Matriz de confusão e a curva ROC esses métodos são explicados na seção 4, a Fundamentação Teórica. Com esses métodos pode-se realizar uma avaliação do modelo, e assim será possível analisar os valores de acurácia, precisão e recall do modelo, essas métricas são utilizadas avaliar o desempenho do modelo, com a acurácia é possível calcular a proximidade entre o valor obtido na predição dos modelos e os valores esperados, com a precisão é possível analisar as amostras que o modelo classificou como 0 eram realmente 0 na classe real e com o recall é possível analisar as amostras que o modelo conseguiu reconhecer como a classe correta.

5.6 ACESSO A BASE DE DADOS E CÓDIGOS

A base de dados utilizada neste projeto pode ser encontrada no link: ... . E os códigos podem ser encontrados no link: ... .

30

6 RESULTADOS

Os mesmos experimentos mostrados abaixo foram aplicados para uma base de dados que contém somente dados de São Paulo, essa base é um recorte da base do Brasil todo em contém cerca de 1.400.000 amostras, esse recorte da base também é bem desbalanceado. Os resultados obtidos nesse experimento usando o recorte de São Paulo foram bem parecidos com os experimentos da base do Brasil todo, e os códigos desse experimento podem ser visualizados no apêndice X.

6.1 ANÁLISE EXPLORATÓRIA

O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

6.1.1 Apresentação dos Dados

Como dito anteriormente na seção 5.2 ?Base de Dados? deste trabalho, a base de dados é formada pelas informações do SIM e do SINASC contém 6.760.222 amostras e 29 colunas, porém serão utilizadas somente 23 colunas sendo elas as colunas descritas na Tabela 3. Na seção 5.2 também tem a Figura 4 que mostra que a base dados é desbalanceada tendo 99.4% das amostras de recém-nascidos vivos e 0.6% das amostras onde os recém-nascidos vieram a óbito no período neonatal.

6.1.1.1 Caracteristicas demográficas e socioeconômicas maternas

O apêndice X contém uma tabela que possui as estatísticas sumarizadas das variáveis

demográficas e socioeconômicas maternas. Nesta tabela por exemplo podemos observar a coluna ?maternal_age? que contém a idade da mãe, e na tabela mostra que a média da idade das mães é de 26.4 anos, e os dados possuem uma variação de no mínimo 8 anos e no máximo 55 anos, e a mediana é 26 anos. Na tabela também tem as colunas: ?tp_maternal_schooling? que mostra a categoria de anos de escolaridade da mãe, ?tp_marital_status? que mostra o estado civil da mãe em dados categóricos e a coluna ?tp_maternal_race? que mostra a raça da mãe também em dados categóricos.
31
6.1.1.2 Variávies obstétricas maternas
No apêndice X pode-se observar uma tabela que possui as estatísticas sumarizadas das variáveis obstétricas maternas. Nesta tabela por exemplo temos a coluna ?num_live_births?, essa coluna mostra o número de nascidos vivos anteriores que a mãe obteve, a média desta coluna é 0.94 , a mediana é 1, o número mínimo é 0 e o máximo é 10 nascidos vivos. Na tabela podemos observar também a coluna, ?num_fetal_losses?, esta coluna mostra o número de perdas fetais anteriores da mãe, a média desta coluna é 0.21, a medina é 0, o número mínimo é 0 e o máximo é 5, esses valores mostram que poucas mães tiveram perdas fetais antes da gravidez atual. Também na tabela tem estatísticas da coluna ?num_previous_gestations? esta coluna mostra o número de gestações anteriores da mãe, esta coluna tem a média de 1.14, a mediana é 1, o número mínimo é 0 e o máximo 10. Essa tabela também contém as colunas: ?num_normal_labors? que é o número de partos normais da mãe, ?num_cesarean_labor? que mostra o número de partos cesáreos da mãe e por fim mostra a coluna ?tp_pregnancy? que é o tipo da gravidez.
6.1.1.3 Variávies de histórico de gravidez
No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis do histórico de gravidez. Nesta tabela temos as colunas ?num_prenatal_appointments? que mostra o número de consultas de pré-natal, a média dessa coluna é 7.8, a mediana é 8, a variação dos dados é de no mínimo 0 e no máximo 40. A tabela também contém as colunas: ?tp_labor? que mostra o tipo de parto, e também contém a coluna ?tp_robson_group? que mostra a classificação do grupo Robson.
6.1.1.4 Variávies relacionadas ao recém-nascido
No apêndice X temos uma tabela que possui as estatísticas sumarizadas das variáveis relacionadas ao recém-nascido. Nesta tabela por exemplo podemos observar a coluna ?newborn_weight? que armazena o peso ao nascer dos recém-nascidos em gramas, a média dos dados dessa coluna é 3187.3, a mediana é 3215, a variação dos dados é de no mínimo 0 e no máximo 6000 gramas. Também podemos observar a coluna ?gestacional_week? que mostra o número de semanas de gestação, a média é 38.4, a mediana é 39, a variação dos dados é de no mínimo 15 e no máximo 45 semanas. Além dessas colunas a tabela também
32
contém as colunas: ?cd_apgar1? que mostra a pontuação de Apgar de 1 minuto, ?cd_apgar5? que mostra a pontuação de Apgar de 5 minuto, ?has_congenital_malformation? que mostra se o recém-nascido contém a presença de alguma malformação, ?tp_newborn_presentation? que mostra o tipo de apresentação de recém-nascido, ?tp_labor? que mostra o tipo de parto, ?was_cesarean_before_labor? que mostra se o recém-nascido foi cesáreo antes do parto, ?was_labor_induced ? que mostra se o recém-nascido foi induzido ao trabalho de parto,

?tp_childbirth_care? que mostra se ouve assistência ao parto, e por fim temos a coluna ?is_neonatal_death? que mostra se o recém-nascido veio a óbito ou não.

6.1.2 Análise das variáveis

Nesta seção serão mostrados a parte de análise de variáveis com diferentes tipos de gráficos.

6.1.2.1 Variáveis contínuas

Na Figura 5 temos dois gráficos boxplot, no boxplot à esquerda temos a distribuição da variável peso ao nascer e nele é possível observar duas caixas onde a caixa azul são os vivos e o laranja são os mortos. Nas duas caixas mostradas no gráfico podemos observar pequenos círculos nas pontas, esses círculos são outliers (dados fora da curva), então para a caixa de vivos os outliers são recém-nascidos com pesos acima de 4500 gramas ou abaixo de 2000 gramas, e para a caixa de mortos os outliers são os recém-nascidos com mais de 5500 gramas. Para os vivos temos a mediana de aproximadamente 3200 gramas e para os mortos a mediana é 1300 gramas. Cada caixa representa 50% da base de dados, a caixa azul de vivos mostra que o peso de recém-nascidos que sobreviveram está aproximadamente entre 2700 a 3600 gramas, e para a caixa laranja de mortos o peso de recém-nascidos que vieram a óbito estão aproximadamente entre 700 a 2500 gramas. Então o gráfico mostra para nós que os recém-nascidos que têm maior risco de vir a óbito tem pesos menores que 2000 gramas.

33

Figura 5. BoxPlot das features Peso ao nascer e Idade da mãe. (Fonte: Elaboração Própria).

No gráfico à direita da Figura 5 temos um boxplot da distribuição da variável idade da mãe, como foi dito anteriormente a caixa azul são os vivos e o laranja são os mortos. Nesse boxplot então podemos ver que os outliers da caixa de vivos são mães com idade acima de 46 anos aproximadamente, para os mortos os outliers são mães com idade acima de 50 anos. Para os vivos temos a mediana de aproximadamente 26 anos e para os mortos a mediana é de 26 anos. Nos vivos a idade das mães está aproximadamente entre 21 a 32 anos, e para os mortos a idade das mães está aproximadamente entre 20 a 34 anos. É importante ressaltar que o gráfico mostra as duas caixas bem parecidas, então de acordo com os dados não contém diferença significativa entre vivos e mortos usando a idade da mãe para distribuir.

Já na Figura 6 temos um boxplot da distribuição da variável semanas de gestação.

Nesse boxplot então podemos ver que os outliers da caixa de vivos são gestações com mais de 44 semanas aproximadamente, e gestações com menos de 35 semanas. Para os vivos temos a mediana de aproximadamente 39 semanas e para os mortos a mediana é de 32 semanas. Nos vivos as semanas de gestação estão aproximadamente entre 36 a 40 semanas, e para os mortos a semanas de gestação estão aproximadamente entre 26 a 36 semanas. Nesse boxplot os dados mostraram que para as gestações que têm menos de 36 semanas os recém-nascidos têm maior risco de vir a óbito.

34

Figura 6. BoxPlot da feature semana de gestação. (Fonte: Elaboração Própria).

Na Figura 7 temos um histograma da distribuição de peso ao nascer por classe, podemos observar no gráfico que grande parte dos vivos (linha azul) estão distribuídos entre 2000 e 4000 gramas, a área entre esses valores tem muitos dados de vivos, já entre os valores 0 e 2000 gramas temos uma concentração dos dados de mortos.

Figura 7. Histograma de distribuição do peso entre as classes. (Fonte: Elaboração Própria)

6.1.2.2 Variáveis categoricas

Observando a Figura 8 podemos ver um gráfico de barras da distribuição da variável escolaridade da mãe, este gráfico mostra a contagem de registros por escolaridade da mãe, esse gráfico mostra para nós que a maior parte das mães estão na categoria 4 que representa de 8 a 11 anos de escolaridade.

35

Figura 8. Gráfico de barras da distribuição da variável Escolaridade da mãe. (Fonte: Elaboração Própria).

Observando a Figura 9 podemos ver um gráfico de barras da distribuição da variável número de vivos, este gráfico mostra a contagem de registros por número de vivos, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho.

Figura 9. Gráfico de barras da distribuição da variável Número de nascidos vivos. (Fonte: Elaboração Própria).

Observando a Figura 10 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 1 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 1 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem

36

apgar de 8 ou 9, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.

Figura 10. Gráfico de barras da distribuição da variável Pontuação Apgar de 1 minuto. (Fonte: Elaboração Própria).

Observando a Figura 11 podemos ver um gráfico de barras da distribuição da variável pontuação Apgar de 5 minuto, este gráfico mostra a contagem de registros por pontuação Apgar de 5 minuto, esse gráfico mostra para nós que a grande maioria dos dados possuem apgar de 9 ou 10, esse alto valor de apgar é por conta da base ser desbalanceada e a maior parte dos dados são de recém-nascidos que sobreviveram.

Figura 11. Gráfico de barras da distribuição da variável Pontuação Apgar de 5 minutos. (Fonte: Elaboração Própria).

37

Observando a Figura 12 podemos ver um gráfico de barras da distribuição da variável presença de malformação congênita, este gráfico mostra a contagem de registros por presença de malformação congênita, esse gráfico mostra para nós que a grande maioria dos dados estão na categoria 2 que mostra que o recém-nascido não possui malformação, esse grande volume para a categoria 2 é causada pelo desbalanceamento da base.

Figura 12. Gráfico de barras da distribuição da variável Presença de malformação congênita. (Fonte: Elaboração Própria).

Observando a Figura 13 podemos ver um gráfico de barras da distribuição da variável número de gestações, este gráfico mostra a contagem de registros por número de gestações, esse gráfico mostra para nós que a grande maioria das mães está tendo o primeiro filho ou o segundo filho, da mesma forma mostrada na Figura 9.

38

Figura 13. Gráfico de barras da distribuição da variável Número de gestações. (Fonte: Elaboração Própria).

Observando a Figura 14 podemos ver um gráfico de barras da distribuição da variável assistência ao parto, este gráfico mostra a contagem de registros por assistência ao parto, esse gráfico mostra para nós que a grande maioria dos dados mostram que o parto teve assistência de uma Enfermeira ou obstetra ou essa informação foi ignorada na hora do resgistro.
Figura 14. Gráfico de barras da distribuição da variável Assistência ao parto. (Fonte: Elaboração Própria).
6.1.2.3 Analise bi-variada
A Figura 15 mostra para nós a importância da feature peso ao nascer com o gráfico mostrado é possível observar claramente a diferença de pesos entre vivos e mortos, recém-nascidos com peso acima de 2000 gramas sobreviveram, já os recém-nascidos com menos de 2000 gramas vieram a óbito.
39
Figura 15. Gráfico de densidade de peso entre as classes. (Fonte: Elaboração Própria).
Na Figura 16 podemos observar um gráfico de correlação entre algumas features de valores contínuos, algumas dessas features possuem correlações positivas, como pontuação apgar de 1 minuto e pontuação apgar de 5 minutos, essas colunas possuem uma correlação de 0.7, então nascidos que recebem um valor alto de apgar de 1 minuto tendem a receber um valor alto no apgar de 5 minutos. O gráfico também mostra outras features com correlações positivas como, número de partos normais e número de gestações anteriores que contém uma correlação de 0.81, número de gestações anteriores com idade da mãe que tem uma correção de 0.39, número de partos de cesáreos com idade da mãe que possui correlação de 0.24, número de partos normais com idade da mãe com a correlação de 0.26, semanas de gestação com peso ao nascer em gramas com correção de 0.52, e as apgar de 1 e 5 minutos com semanas de gestação. E grande parte das features, possuem correlações baixas então elas são inversamente correlacionadas pelo que o gráfico mostra, como por exemplo número de partos normais com número de consultas pré-natais possuem uma correção de -0.17, e número de partos cesáreos com número de partos normais que contém uma correlação de -0.15. Então a correlação dessas features estão bem baixas tirando as correlações acima de 0.7.
40
Figura 16. Gráfico de correlação. (Fonte: Elaboração Própria)
6.1.2.4 Distribuição por raça
Na Figura 17 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por raça, e podemos observar que a raça 5 (indigena), é o boxplot que contem a menor variação de idade, já as raças 1 (branco) e 3 (amarelo) são as caixas que possuem maior variação com a média em torno de 20 e 30 anos.
41
Figura 17. Distribuição da idade da mãe por raça. (Fonte: Elaboração Própria).
No boxplot mostrado na Figura 18 podemos observar a distribuição de peso ao nascer por raça, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre as raças para peso ao nascer.
Figura 18. Distribuição de peso ao nascer por raça. (Fonte: Elaboração Própria).
Na Figura 19 temos o boxplot da a distribuição de semanas de gestação por raça, e praticamente todas as caixas são iguais, somente a caixa da raça 1 (branco) possui menor variação que as outras.

42
Figura 19. Distribuição de semanas de gestação por raça. (Fonte: Elaboração Própria).
6.1.2.5 Distribuição por estado civil
A Figura 20 mostra o boxplot da a distribuição de peso ao nascer por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então pouca variação é observada entre os estados civis para peso ao nascer.
Figura 20. Distribuição de peso ao nascer por estado civil. (Fonte: Elaboração Própria).
A Figura 21 mostra o boxplot da a distribuição de semanas de gestação por estado civil, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Porém os estados civis 2 (Casado) e 4 (Separados/divorciados judicialmente), contém variações menores.
43
Figura 21. Distribuição de semanas de gestação por estado civil. (Fonte: Elaboração Própria).
Na Figura 22 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por estado civil, e podemos observar que o estado civil 3 (Viúva), possui a maior variação, já o estado civil 4 (Separados/divorciados judicialmente) é a caixa que possuem menor variação
Figura 22. Distribuição da idade da mãe por estado civil. (Fonte: Elaboração Própria).
6.1.2.6 Distribuição por escolaridade da mãe
Na Figura 23 podemos ver um gráfico de boxplot mostrando a distribuição da idade da mãe por escolaridade da mãe, e podemos observar que a escolaridade 2 (1 a 3 anos) e 3 (4 a 7
44
anos), possui as maiores variações, já escolaridade 5 (12 ou mais anos) é a caixa que possuem menor variação
Figura 23. Distribuição da idade da mãe por escolaridade da mãe. (Fonte: Elaboração Própria).
A Figura 24 mostra o boxplot da a distribuição de peso ao nascer por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas, todas são praticamente iguais. Então, pouca variação é observada entre a escolaridade da mãe para peso ao nascer.
Figura 24. Distribuição de peso ao nascer por escolaridade da mãe. (Fonte: Elaboração Própria).
A Figura 25 mostra o boxplot da a distribuição de semanas de gestação por escolaridade da mãe, e podemos ver que que não tem diferença significativa entre as caixas,
45
todas são praticamente iguais. Porém a escolaridade 5 (12 ou mais anos) contém variação menor que as outras.
Figura 25. Distribuição de semanas de gestação por escolaridade da mãe. (Fonte: Elaboração Própria).
6.2 ÁRVORE DE DECISÃO
A Figura 26 mostra um trecho do algoritmo utilizado, nesse trecho é realizado a divisão do DataFrame das features de entrada , e o DataFrame que contém o rótulo.Neste caso o rótulo será o DataFrame ?target?, esse Dataframe contém a feature

?is_neonatal_death?, essa feature informa se o recém nascido veio a óbito ou não, sendo o foco da predição que desejamos realizar essa feature entra como rótulo para o modelo, já no Dataframe ?nome_features? temos as features que vão servir de entrada para o modelo, é a partir dessas features que o modelo tentará encontrar padrões para predizer o risco de óbito do recém nascido. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.

46

Figura 26. Separação dos DataFrames de entrada e rótulo. (Fonte: Elaboração Própria).

6.2.1 Modelo de Árvore de Decisão Utilizando Particionamento 90/10

O algoritmo de árvore de decisão foi treinado utilizando duas formas diferentes: usando as partições 90% para treino e 10% para os testes e utilizando o método K-folds cross-validation. E para os experimentos iniciais foi utilizado o particionamento 90/10.

Tabela 4. Resultados do modelo de árvore de decisão. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	671807
Morto(1)	0.71	0.29	0.41	4216

Analisando a Tabela 4 pode-se ver os seguintes resultados, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 71% que realmente pertenciam a classe 1, então cerca de 29% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 29% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem a classe 1 que seria dos recém nascidos que poderiam vir a óbito, já para a classe de vivos a

47

classe 0 o modelo está identificando muito bem e classificando de forma correta. A tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.

Esse modelo teve uma acurácia de 0.99 que é uma acurácia bem alta, porém somente pela acurácia não é possível dizer que o modelo está excelente pelo motivo da base que está sendo usada é altamente desbalanceada, então para a classe 0 que é a de vivos temos muito mais dados que para a classe de mortos, e o modelo está acertando bastante a classe de vivos logo a acurácia fica alta por esses acertos, enquanto para a classe de mortos o modelo não está acertando tanto.

A Figura 27 mostra a curva ROC do modelo, pode-se observar que a AUC da curva ROC ficou em 0.92, a AUC mostra que o modelo está acertando bastante as predições, pois quanto mais perto de 1 melhor está no nosso modelo, mas no caso desse modelo vimos acima que as predições para a classe de mortos(1) está ruim, o modelo está acertando poucas classificações como 1. Então a AUC assim como a acurácia está alta porque a base de dados utilizada é altamente desbalanceada tendo muito mais amostras de vivos(1), e como o modelo está bom para essa classificação e está acertando bastante os valores de AUC e acurácia ficam altos .

Figura 27. Curva ROC modelo de Árvore de decisão. (Fonte: Elaboração Própria).
A Figura 28 mostra a importância de cada feature para o modelo sendo assim a feature 10 - Peso ao nascer, a mais importante para o modelo, pois dela o modelo conseguiu tirar mais informações, seguido das features, 13 - Apgar dos 5 primeiros minutos, 11 - Semanas de 48
Gestação, 14 - Presença de malformação 12 - Apgar do 1 primeiro minutos. Então essas são as features que foram mais decisivas para a montagem da árvore e para as predições do modelo.
Figura 28. Gráfico de importância das features. (Fonte: Elaboração Própria).
Na Figura 29 temos uma imagem reduzida da árvore de decisão que foi montada com o treinamento utilizando 5 como profundidade máxima para a árvore, a imagem da árvore completa pode ser visualizada no apêndice A. É possível ver que as features marcadas como importante para o modelo apareceu diversas vezes na árvore, e a feature peso ao nascer foi escolhida para ser o nó raiz da árvore, de todas as features ela foi a que teve a melhor entropia para ser o nó raiz, e depois aparece mais vezes para outras decisões em outros níveis da árvore e isso faz com que essa feature se torne a mais importante para o modelo, também podemos ver que a feature ?Apgar do 1 primeiro minuto? mesmo sendo a menos importante para o modelo ela aparece somente no nó de número 42 isso mostra que a entropia e o ganho de informação dessa feature nas outras decisões não foram boas fazendo ela ser a feature com menos importância utilizada no modelo.
49
Figura 29. Árvore de decisão montada a partir do algoritmo. (Fonte: Elaboração Própria).
6.2.2 Modelo de Árvore de Decisão Utilizando K-folds cross-validation
Após realizar esse experimento com a partição 90/10, foi realizado outro experimento com esse mesmo algoritmo porém agora utilizando o método K-folds cross-validation para o treinamento. O K-folds cross-validation foi configurado para separar a base de dados em 10 partes iguais, e assim o algoritmo foi treinado novamente gerando um novo modelo. O método de K-folds cross-validation é utilizado para conferir se o modelo não está sofrendo problemas de overfitting, é importante garantir que o modelo está aprendendo com os dados e não está somente decorando.
Na Tabela 5 temos então o resultados desse modelo, pode-se observar que os resultados para a classificação de vivos(0) não teve alteração, o modelo continua tendo 100% na precisão e no recall, mas na classificação de mortos(1) o modelo teve uma diminuição na precisão que agora está em 65% e também teve uma diminuição no recall que agora está em 23%, porém essa diminuição é justificada porque para o teste deste modelo foi utilizado a base completa e não somente a partição de teste que contém 10% da base total. Logo os resultados não tiveram uma alteração drástica e tiveram um comportamento bem semelhante, incluindo a acurácia que se manteve em 0.99 e a AUC que teve uma diminuição pequena também e ficou com 0.89.
50
Tabela 5. Resultados do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	6719191

Morto(1) 0.65 0.23 0.34 41031
Na Tabela 6 é possível ver a matriz de confusão do modelo que mostra o número exato de erros e acertos do modelo, pode-se observar que para a classe de vivos o modelo classificou corretamente 6.714.117 amostras da base de dados completa, o modelo classificou como 0 as amostras que nos rótulos realmente eram 0, errou apenas 5074 amostras. Já para a classe de mortos o modelo acertou somente 9552 amostras, classificou como 1 as amostras que no rótulo eram 1 também, e errou 31479. Isso mostra que o modelo está muito desbalanceado, pois está errado muito mais do que acertando as predições de mortos, sendo isso um reflexo da base de dados desbalanceada também. E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.
Tabela 6. Matriz de confusão do modelo de árvore de decisão utilizando K-folds cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6714117 5074
Mortos (1) 31479 9552
6.3 REGRESSÃO LOGÍSTICA
Para o algoritmo de Regressão Logística também foi utilizado métodos diferentes para o treinamento, O algoritmo foi treinado utilizando três formas diferentes: usando as partições 90% para treino e 10% para os testes, utilizando o método K-folds cross-validation e foi treinada utilizando o método RFECV que seleciona as melhores features para o modelo, juntamente com o método GridSearchCV. O código completo deste experimento pode ser visualizado no apêndice X, ou também no link do github exibido na seção X.
51
6.3.1 Modelo de Regressão Logística Utilizando Particionamento 90/10
Para o primeiro experimento do algoritmo de regressão logística foi utilizado o particionamento 90/10 para o treinamento do algoritmo, sendo 90% da base de dados para realizar o treinamento do modelo e 10% da base para realizar os testes.
Na Tabela 7 temos os resultados do modelo de regressão logística utilizando o método de particionamento 90/10, analisando a precision vemos que para a classe 0 a precisão é de 1.00, ou seja, 100% então todos as amostras que o modelo classificou como 0 eram realmente 0 na classe real, já para a classe 1 o modelo classificou 65% que realmente pertenciam a classe 1, então cerca de 35% das classificações que o modelo colocou como 1 estão incorretas. Analisando o recall é possível ver que o modelo conseguiu identificar 100% das classificações 0, o modelo conseguiu reconhecer muito bem as amostras classificadas como 0, já para as amostras classificadas como 1 o modelo reconheceu apenas 24% das amostras , o modelo deixou passar muitas amostras da classe 1 e não classificou como 1. Então o modelo treinado não está identificando muito bem as amostras da classe 1 que seria dos recém-nascidos que poderiam vir a óbito, já para a classe de vivos a classe 0 o modelo está identificando muito bem e classificando de forma correta.E na tabela também mostra os valores de F1-Score para cada classe, esse resultado é formado pela soma da Precision e do Recall.

Tabela 7. Resultados do modelo de regressão logística usando particionamento 90/10.
(Fonte: Elaboração Própria)
Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 671885
Morto(1) 0.65 0.24 0.35 4138
A Figura 30 abaixo mostra a curva ROC do modelo de regressão logística usando o particionamento 90/10, analisando a curva ROC podemos ver que a AUC da curva ficou em 0.93 e a acurácia do modelo ficou 0.99, mostrando que o modelo está acertando bastante as predições, porém esses números para o nosso modelo não mostra que ele está ótimo, com a análise da Tabela 7 acima podemos observar que o modelo está acertando muitas predições da classe de vivos e poucas predições da classe de mortos, como a base de dados é
52
desbalanceada, como a maior quantidade de dados está na classe de vivos e o modelo está acerto quase todos dessa classe a acurácia e a AUC ficam altas por esses acertos.Logo é preciso analisar mais resultados além da acurácia e da AUC do modelo.
Figura 30. Curva ROC modelo regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria).
Na Tabela 8 podemos analisar os número exatos que o modelo acertou de cada classe, como é mostrado na matriz de confusão o modelo acertou 671.435 da classe de vivos e acertou apenas 915 da classe de mortos, e errou mais que o triplo da classe de mortos. Então as predições do modelo estão muito desbalanceadas, para a classe de vivos o modelo está predizendo bem, porém para as classes de mortos o modelo está errando muitas predições.
Tabela 8. Matriz de confusão do modelo de regressão logística usando particionamento 90/10. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 671435 504
Mortos (1) 3169 915
6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation
Para o segundo experimento do algoritmo de regressão logística foi utilizado o método K-folds cross-validation para o treinamento do modelo com a intenção de conferir e confirmar
53
que o modelo não esteja tendo problemas com overfitting e realmente esteja aprendendo com os dados que está sendo usado para o treinamento. O método K-folds cross-validation foi configurado para realizar uma separação de 10 partes iguais.
Na Tabela 9 podemos observar os resultados do modelo, e os resultados foram parecidos com o experimento anterior usando o particionamento 90/10, única diferença nos resultado é que no experimento anterior a precisão da classe de mortos ficou em 65% e no caso desse experimento usando o K-folds cross-validation a precisão ficou em 64%, mas os valores de recall e F1-Score se mantiveram, assim como todos os resultados da classe de mortos se mantiveram em 100%. A execução desse experimento mostra que o modelo realmente está aprendendo com os dados e não está apenas decorando, retirando o risco do modelo estar com overfitting. Mesmo sem alterações no resultado é importante saber que o

modelo não está com overfitting e está conseguindo aprender e encontrar padrões nos dados.
Tabela 9. Resultados do modelo de regressão logística usando K-folds cross-validation.
(Fonte: Elaboração Própria)
Precision Recall F1-Score Support
Vivo(0) 1.00 1.00 1.00 6719191
Morto(1) 0.64 0.24 0.35 41031
Já na Figura 31 temos a curva ROC do modelo, essa curva ROC mostra a AUC de
todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do
método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas
para todas as vezes que foi realizado os testes, a AUC se manteve em 0.93 e 0.94, sendo a
média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não
se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua
acertando muito a classe de vivos que é a classe predominante na base de dados. Já era
esperado os resultados não sofrerem alterações grandes, pois a base de dados não sofreu
nenhuma alteração, esse experimento foi somente para conferir se a modelo não estava
sofrendo overfitting, e pelos resultados aparentemente a base não está com problemas de
sobreajuste.
54
Figura 31. Curva ROC modelo regressão logística usando K-folds cross-validation. (Fonte:
Elaboração Própria).
Na Tabela 10 temos a matriz de confusão do modelo, onde mostra que o modelo
acertou 6.713.658 amostras da classe de vivos e errou apenas 5.533, já para a classe de mortos
o modelo acertou somente 9.847 e novamente errou mais que o triplo errando 31.182
amostras. Como os resultados mostrados anteriormente não tiveram alteração era de se
esperar que as predições corretas e incorretas do modelo também não sofressem alterações, o
modelo continua acertando muito a classe de vivos e errando muito a classe de mortos,
mantendo as predições bem desbalanceadas.
55
Tabela 10. Matriz de confusão do modelo de regressão logística usando K-folds
cross-validation. (Fonte: Elaboração Própria)
Predito
Vivos (0) Mortos (1)
Classe Real
Vivos (0) 6713658 5533
Mortos (1) 31182 9847
6.3.3 Modelo de Regressão Logística Utilizando GridSearchCV e RFECV
Para o terceiro e último experimento de regressão logística foi utilizado técnicas
diferentes juntas para tentar melhorar as predições do modelo, para esse experimento usamos
o K-folds cross-validation para o particionamento da base de dados assim como foi utilizado
no experimento acima, e também foi utilizado duas técnicas que ainda não foram usadas nos
experimentos anteriores que são as funções RFECV e GridSearchCV. A função RFECV
seleciona a quantidade ideal de features para ser usadas no treinamento do modelo e também
mostra quais são as melhores features para essa predição, então essa função ela utiliza a
validação cruzada para ir treinando e testando N vezes o modelo, e sempre vai alterando a

quantidade e as features utilizadas para o teste, então cada vez que a função testa os modelos ela grava a pontuação dos acertos e assim depois mostra o gráfico da Figura 32 e assim é possível ver quais são as features ideais para o modelo.
56
Figura 32. Resultado da função RFECV Base Brasil. (Fonte: Elaboração Própria).
Então os resultados mostrados na Figura 32 apresenta que o número ideal para treinar o modelo de features é 6, e as features sao: 'tp_maternal_schooling', 'gestaional_week', 'cd_apgar1', 'cd_apgar5', 'has_congenital_malformation', 'tp_labor'. Essas são as features que tiveram o melhor desempenho no RFECV, então para o treinamento do modelo desse experimento as features que vão ser utilizadas serão somente essas 6.
Após a execução da função de RFECV foi executado a função de GridSearchCV, o GridSearchCV é utilizado para descobrir quais são os melhores parâmetros para utilizar no algoritmo de regressão logística e criar os modelos, foi utilizado também a validação cruzada para testar qual seria os melhores parâmetros para melhorar o desempenho do modelo. Na Figura 33 pode-se ver que os parâmetros escolhidos pela função foram:
Figura 33. Resultado da função GridSearchCV. (Fonte: Elaboração Própria).
Na Tabela 11 podemos observar os resultados do modelo, e os resultados foram parecidos com os experimentos anteriores, para a classe de mortos a precisão teve um aumento e foi para 69% e o recall diminuiu chegando em 20%, então utilizando as novas
57
funções o modelo ganhou precisão e perdeu recall. Para a classe de vivos o modelo se manteve em 100% e não teve alterações para os resultados de precisão e recall. Mesmo com a utilização das funções de RFECV e GridSearchCV o modelo não teve uma melhora significativa para as predições de mortos.
Tabela 11. Resultados do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

	Precision	Recall	F1-Score	Support
Vivo(0)	1.00	1.00	1.00	6719191
Morto(1)	0.69	0.20	0.31	41031

Na Figura 34 temos a curva ROC do modelo, essa curva ROC mostra a AUC de todas as 10 vezes que o modelo foi treinado e testado usando as 10 partes diferentes do método de K-folds cross-validation, e como pode-se observar as AUCs foram bem parecidas para todas as vezes que foi realizado os testes, a AUC se manteve em 0.92 e 0.93, sendo a média de AUC 0.93 a mesma AUC do experimento anterior então esse resultado também não se alterou. A acurácia do modelo também se manteve em 0.99, afinal o modelo continua acertando muito a classe de vivos que é a classe predominante na base de dados.
58
Figura 34. Curva ROC modelo regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria).
Na Tabela 12 temos a matriz de confusão do modelo, onde mostra que o modelo acertou 6.715.535 amostras da classe de vivos e errou apenas 3.656, comparando com os experimentos anteriores de regressão logística as predições de vivos teve uma pequena piora e teve uma pequena diminuição nos acertos. Já para a classe de mortos, o modelo acertou 8.296 e errou 32.735 amostras, logo as predições de vivos tiveram uma uma piora e acertou um

pouco menos do que os outros experimentos, e na classe de vivos teve uma melhora acertando mais predições, porém como o problema está nas predições de mortos e para essa classe teve uma piora pode-se dizer que a função de GridSearchCV não teve um ganho significativo nos resultados. Então como modelo final foi escolhido o modelo utilizando somente o K-folds cross-validation, pois foi o modelo que mostrou o melhor desempenho dos experimentos de regressão logística.

59

Tabela 12. Matriz de confusão do modelo de regressão logística usando GridSearchCV e RFECV. (Fonte: Elaboração Própria)

Predito

Vivos (0) Mortos (1)

Classe Real

Vivos (0) 6715535 3656

Mortos (1) 32735 8296

60

7 CONCLUSÃO

O principal objetivo deste trabalho foi analisar os resultados das predições de mortalidade neonatal dos algoritmos de aprendizado de máquina usando uma base de dados com dados do Brasil inteiro. A partir destes resultados apresentados na secção anterior ?Resultados?, é possível concluir que analisar somente a AUC e a acurácia do modelo não é o suficiente para garantir que o modelo está excelente, sendo assim temos que ter mais atenção a precisão, recall e as predições mostradas na matriz de confusão.

Ambos os modelos ficaram desbalanceados nas predições, os modelos acertaram mais predições de vivos do que mortos, e as predições corretas de mortos foram mais baixas do que as incorretas, para os modelos ficarem melhor precisam passar por algoritmos que possam balancear essas predições, uma outra alternativa para melhorar o modelo é utilizar uma base de dados mais balanceada, pois essa base que foi utilizada é altamente desbalanceada.

Embora os dois modelos tenham tido resultados parecidos , o modelo de regressão logística utilizando somente o K-folds cross-validation se saiu melhor que os outros experimentos, o modelo criado na seção ?6.3.2 Modelo de Regressão Logística Utilizando K-folds cross-validation? manteve a acurácia, AUC, precisão e recall dos demais modelos e acertou mais predições, talvez com algoritmos para melhorar as predições da classe de mortos, balanceando mais as predições, esse modelo fique com ótimas taxas preditivas.

Além disso podes-se afirmar que o peso ao nascer do recém nascido é um fator muito importante para as decisões dos modelos, ambos os modelos usaram muito essa informação para as predições, também podemos colocar, as colunas de presença de malformação e semana de gestação, como colunas que foram importantes para os modelos, essas causas podem ser evitadas se houver um acompanhamento médico com qualidade e eficiência.

61

REFERÊNCIAS

ÁRVORE de Decisão - Aula 3. Gravação de Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=ecYpXd4WREk&t=4089s. Acesso em: 1 jul. 2020.

BARRETO, J. O. M.; SOUZA, N. M.; CHAPMAN, E. Síntese de evidências para políticas de saúde: reduzindo a mortalidade perinatal. Ministério da Saúde, p. 1?44,2015.

BATISTA, André Filipe de Moraes; FILHO, Alexandre Dias Porto Chiavegatto. Machine Learning aplicado à Saúde. In: ZIVIANI, Artur; FERNANDES, Natalia Castro; SAADE, Débora Christina Muchaluat. SBCAS 2019: 19 Simpósio Brasileiro de Computação Aplicada à Saúde. [S. l.: s. n.], 2019. cap. Capítulo 1.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. Towards neonatal mortality risk classification: A data-driven approach using neonatal, maternal, and social factors, [s. l.], 23 out. 2020.
BELUZO, Carlos Eduardo; SILVA, Everton; ALVES, Luciana Correia; BRESAN, Rodrigo Campos; ARRUDA, Natália Martins; SOVAT, Ricardo; CARVALHO, Tiago. SPNeoDeath: A demographic and epidemiological dataset having infant, mother, prenatal care and childbirth data related to births and neonatal deaths in São Paulo city Brazil ? 2012?2018, [s. l.], 19 jul. 2020.
BRESAN, Rodrigo. Um método para a detecção de ataques de apresentação em sistemas de reconhecimento facial através de propriedades intrínsecas e Deep Learning. Orientador: Me. Carlos Eduardo Beluzo. 2018. Trabalho de Conclusão de Curso (Superior de Tecnologia em Análise e Desenvolvimento de Sistemas) - Instituto Federal de Educação, Ciência e Tecnologia Câmpus Campinas., [S. l.], 2018.
CABRAL, Cleidy Isolete Silva. Aplicação do Modelo de Regressão Logística num Estudo de Mercado. Orientador: João José Ferreira Gomes. 2013. Projeto Mestrado em Matemática Aplicada à Economia e à Gestão (Mestrado) - Universidade de Lisboa Faculdade de Ciências Departamento de Estatística e Investigação Operacional, [S. l.], 2013. Disponível em: https://repositorio.ul.pt/bitstream/10451/10671/1/ulfc106455_tm_Cleidy_Cabral.pdf. Acesso em: 1 maio 2021.
CAMPOS, Raphael. Árvores de Decisão: Então diga-me, como construo uma?. [S. l.], 28 nov. 2017. Disponível em: https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69. Acesso em: 23 jan. 2021.
Colab. 2021. Disponível em: https://colab.research.google.com/. Acesso em: 20 mar. 2021.
COX, D.R.;SNELL, E.J. Analysis of Binary Data.London: Chapman & Hall, 2ª Edição, 1989. HOSMER, D.W.; LEMESHOW, S.Applied Logistic Regression. New York: John Wiley, 2ª Edição, 2000.
E.França, S.Lansky. Mortalidade infantil neonatal no brasil: situação, tendências e perspectivas Rede Interagencial de Informações para Saúde - demografia e saúde: contribuição para análise de situação e tendências Série Informe de Situação e Tendências(2009), pp.83-112.
62
FREIRE, Sergio Miranda. Bioestatística Básica: Regressão Linear. [S. l.], 20 out. 2020. Disponível em: http://www.lampada.uerj.br/arquivosdb/_book/bioestatisticaBasica.html. Acesso em: 23 jan. 2021.
Jupyter. 2021. Disponível em: https://jupyter.org/. Acesso em: 20 jun. 2021.
GOODFELLOW, I. et al. Deep learning. [S.l.]: MIT press Cambridge, 2016.
Pandas. 2021. Disponível em: https://pandas.pydata.org/. Acesso em: 20 jun. 2021.
Python. 2021. Disponível em: https://www.python.org/. Acesso em: 20 jun. 2021.

KUHN, M.; JOHNSON, K. Applied predictive modeling. [S.l.]: Springer, 2013. v. 26.
LANSKY, S. et al. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante . Cadernos de Saúde Pública, Scielo, v. 30, p. S192 ? S207, 00 2014.
LANSKY, Sônia; FRICHE, Amélia Augusta de Lima; SILVA, Antônio Augusto Moura; CAMPOS, Deise; BITTENCOURT, Sonia Duarte de Azevedo; CARVALHO, Márcia Lazaro; FRIAS, Paulo Germano; CAVALCANTE, Rejane Silva; CUNHA, Antonio José Ledo Alves. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante e ao recém-nascido. [s. l.], Ago 2014. Disponivel em: https://www.scielosp.org/article/csp/2014.v30suppl1/S192-S207/pt/
Matplotlib. 2021. Disponível em: https://matplotlib.org/. Acesso em: 20 mar. 2021.
Maranhão AGK, Vasconcelos AMN, Trindade CM, Victora CG, Rabello Neto DL, Porto D,et al. Mortalidade infantil no Brasil: tendências, componentes e causas de morte no período de 2000 a 2010 . In: Departamento de Análise de Situação de Saúde, Secretaria de Vigilância em Saúde, Ministério da Saúde, organizador. Saúde Brasil 2011: uma análise da situação de saúde e a vigilância da saúde da mulher. v. 1. Brasília: Ministério da Saúde; 2012. p. 163-82.
MESQUITA, PAULO. UM MODELO DE REGRESSÃO LOGÍSTICA PARA AVALIAÇÃO DOS PROGRAMAS DE PÕS-GRADUAÇÃO NO BRASIL. Orientador: Prof. RODRIGO TAVARES NOGUEIRA. 2014. Dissertação (Mestre em Engenharia de Produção) - Universidade Estadual do Norte Fluminense, [S. l.], 2014.
MNASSRI , Baligh. Titanic: logistic regression with python. [S. l.], 1 ago. 2020. Disponível em: https://www.kaggle.com/mnassrib/titanic-logistic-regression-with-python. Acesso em: 14 jan. 2021.
Morais, R.M.d., Costa, A.L: Uma avaliação do sistema de informações sobre mortalidade. Saúde em Debate 41, 101?117 (2017). Disponível em: https://doi.org/10.1109/SIBGRAPI.2012.38.
MOTA, Caio Augusto de Souza; BELUZO , Carlos Eduardo; TRABUCO, Lavinia Pedrosa; SOUZA , Adriano; ALVES, Luciana; CARVALHO, Tiago. DESENVOLVIMENTO DE UMA PLATAFORMA WEB PARA CIÊNCIA DE DADOS APLICADA À SAÚDE MATERNO INFANTIL , [s. l.], 28 nov. 2019.
MULTIEDRO (Brasil). Conheça as aplicações do machine learning na área da saúde. [S. l.], 28 nov. 2019. Disponível em: https://blog.multiedro.com.br/machine-learning-na-area-da-saude/#:~:text=O%20machine%263
0learning%20na%20%C3%A1rea,diagn%C3%B3sticos%20e%20tratamentos%20mais%20precisos. Acesso em: 13 jun. 2021.
MUKHIYA,S.K.;AHMED,U. Hands-On Exploratory Data Analysis with Python: Perform EDA Techniques to Understand, Summarize, and Investigate Your Data. [S.l.]: PacktPublishingLtd,2020.ISBN978-1-78953-725-3.22,32
Murray CJ, Laakso T, Shibuya K, Hill K, Lopez AD. Can we achieve Millennium Development Goal 4? New analysis of country trends and forecasts of under-5 mortality to 2015. Lancet 2007; 370:1040-54.
NumPy. 2021. Disponível em:https://numpy.org/. Acesso em: 20 jun. 2021.
Oliveira, M.M.d., de Araújo, A.S.S.C., Santiago, D.G., Oliveira, J.a.C.G.d., Carvalho, M.D.,

de Lyra et al, R.N.D.: Evaluation of the national information system on live births in brazil , 2006-2010. Epidemiol. Serv. Saúde 24(4), 629?640 (2015).
PELISSARI, ANA CAROLINA CHEBEL. MORTALIDADE NEONATAL DA CIDADE DE SÃO PAULO: UMA ABORDAGEM UTILIZANDO APRENDIZADO DE MÁQUINA SUPERVISIONADO. 2021. Trabalho de Conclusão de Curso (Superior) - Instituto Federal de Educação, Ciência e Tecnologia Campus Campinas, [S. l.], 2021.
REGRESSÃO logística e binary cross entropy - Aula 7. Direção: Diogo Cortiz. [S. l.]: YouTube, 2020. Disponível em: https://www.youtube.com/watch?v=3J-LBtHVsm4. Acesso em: 21 jan. 2021.
Rmd Nascimento, AJM Leite, NMGSd Almeida, Pcd Almeida, Cfd Silva Determinantes da mortalidade neonatal: estudo caso-controle em fortaleza, ceará, brasil Cad Saúde Pública, 28(2012), pp.559 - 572.
SANTOS, Hellen Geremias. Machine learning e análise preditiva: considerações metodológicas: métodos lineares para classificação. In: SANTOS, Hellen Geremias. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina. 2018. Tese (Pós-Graduação em Epidemiologia) - Faculdade de Saúde Pública da Universidade de São Paulo, [S. l.], 2018.
Scikit-learn. 2021. Disponível em: https://scikit-learn.org/stable/. Acesso em: 20 jun. 2021.
Seaborn. 2021. Disponível em: https://seaborn.pydata.org/. Acesso em: 20 jun. 2021.
SILVA, Wesley; CORSO, Jansen; WELGACZ, Hanna; PEIXE, Julinês. AVALIAÇÃO DA ESCOLHA DE UM FORNECEDOR SOB CONDIÇÃO DE RISCOS A PARTIR DO MÉTODO DE ÁRVORE DE DECISÃO. ARTIGO ? MÉTODOS QUANTITATIVOS, [s. l.], 12 ago. 2008.
WHO, W. H. O.Women and health: today?s evidence, tomorrow?s agenda. [S.l.]: UNWorld Health Organization, 2009. ISBN 9789241563857.
64
APÊNDICE A