

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO
CAMPUS CAMPINAS

ANA CAROLINA CHEBEL PELISSARI

**MORTALIDADE NEONATAL DA CIDADE DE SÃO PAULO: UMA
ABORDAGEM UTILIZANDO APRENDIZADO DE MÁQUINA
SUPERVISIONADO**

CAMPINAS

2021

ANA CAROLINA CHEBEL PELISSARI

**MORTALIDADE NEONATAL DA CIDADE DE SÃO PAULO: UMA
ABORDAGEM UTILIZANDO APRENDIZADO DE MÁQUINA
SUPERVISIONADO**

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 Campus Campinas.

Orientador: Prof. Me. Everton Josué da Silva

CAMPINAS

2021

Ficha catalográfica
Instituto Federal de São Paulo – Câmpus Campinas
Biblioteca
Rosana Gomes – CRB 8/8733

P384m Pelissari, Ana Carolina Chebel
Mortalidade neonatal da cidade de São Paulo: uma abordagem utilizando
aprendizado de máquina supervisionado / Ana Carolina Chebel Pelissari. –
Campinas, SP: [s.n.], 2020.
50 f. il. :

Orientador: Everton Josué da Silva.
Trabalho de Conclusão de Curso (graduação) – Instituto Federal de
Educação, Ciência e Tecnologia de São Paulo Câmpus Campinas. Curso de
Tecnologia em Análise e Desenvolvimento de Sistemas, 2020.

1. Mortalidade neonatal – São Paulo (SP). 2. Aprendizado de máquina. 3.
Estatística. I. Instituto Federal de Educação, Ciência e Tecnologia de São Paulo
Câmpus Campinas. Curso de Tecnologia em Análise e Desenvolvimento de
Sistemas. II. Título.

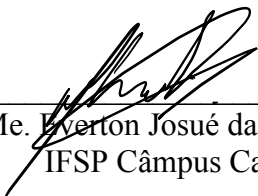
ANA CAROLINA CHEBEL PELISSARI

**MORTALIDADE NEONATAL DA CIDADE DE SÃO PAULO:
UMA ABORDAGEM UTILIZANDO APRENDIZADO DE MÁQUINA**

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: 07 de janeiro de 2021

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

AGRADECIMENTOS

Agradeço à Deus, à minha família que sempre me apoiou em tudo.

Agradeço a todos os professores e servidores do IFSP

Campus Campinas, que contribuíram direta e

indiretamente para a conclusão deste trabalho.

Agradeço ao meu orientador pela ajuda para a conclusão deste trabalho.

*"Se não puder voar, corra. Se não puder correr, ande.
Se não puder andar, rasteje, mas continue em frente de qualquer jeito."
Martin Luther King Jr.*

Resumo

Este projeto realizou uma análise exploratória de dados da mortalidade neonatal da cidade de São Paulo entre os anos 2012 e 2018, além da criação de modelos de aprendizado de máquina capazes de prever o risco de morte de recém-nascidos durante o período neonatal e estudar os fatores que mais influenciaram no resultado e as características mais importantes utilizadas pelos modelos na previsão. Foram utilizados três diferentes tipos de algoritmos de aprendizado de máquina, *XGBoost*, *Random Forest* e *Logistic Regression*, e dois métodos para treinamento, teste e validação dos modelos, o *K-fold cross validation* e *Hold Out*. Todos os modelos criados neste projeto obtiveram um bom desempenho preditivo, se considerarmos a acurácia e o valor AUC. Porém, uma vez que a base de dados utilizada é desbalanceada, não é viável analisar modelos criados utilizando apenas pela acurácia e o valor AUC, e sim realizar esta análise considerando métricas como a precisão e a cobertura. Os resultados das predições dos modelos criados utilizando a base de dados desbalanceada apresentaram uma precisão e cobertura da classe majoritária alta, entretanto, no caso da classe minoritária, o valor da precisão foi alto e o valor da cobertura foi baixo. Ao atribuir pesos à classe minoritária, a cobertura da classe minoritária aumentou.

Palavras-chave: Mortalidade Neonatal, Análise Exploratória, Aprendizado de Máquina.

Abstract

This project performed an exploratory data analysis of Sao Paulo's neonatal mortality within 2012 and 2018, besides the creation of machine learning models to predict the death risk of a new born during the neonatal period and study the factors and features that influenced the result of the prediction. It was used three different machine learning algorithms, *XGBoost*, *Random Forest* and *Logistic Regression*, and two training, test and validation methods, *K-fold cross validation* and *Hold Out*. All the models created obtained a good predictive performance, if only considered the accuracy and the AUC value, but since the dataset used in this project is unbalanced, it is not viable to analyze the models based on the accuracy and AUC value. In these cases, it is recommended to perform this analysis considering metrics such as precision and recall. The prediction results from the models created using the unbalanced dataset shows that the majority class precision and recall values are high while the minority class precision values are high and the recall values are low. By assigning weights to the minority class, their recall values increased.

Keywords: Neonatal Mortality, Exploratory Analysis, Machine Learning.

Lista de Figuras

1	Exemplo de árvore de decisão. (Fonte: Elaboração Própria)	20
2	Exemplo do resultado do algoritmo Kmeans. (Fonte: Adaptado de ML K-means++ Algorithm - Khosla, 2020)	21
3	Curva ROC. (Fonte: Elaboração Própria)	25
4	Curva PR. (Fonte: Elaboração Própria)	25
5	Etapas da Metodologia. (Fonte: Elaboração Própria)	28
6	Distribuição da base de dados considerando se o recém-nascido viveu ou morreu durante os período neonatal. (Fonte: Elaboração Própria) . .	29
7	Gráfico que apresenta a porcentagem dos valores NaN presentes em algumas colunas. (Fonte: Elaboração Própria)	31
8	Gráfico da relação entre a idade e escolaridade da gestante. (Fonte: Elaboração Própria)	32
9	Tipo de gravidez com relação a quantidade de recém-nascidos vivos e mortos. (Fonte: Elaboração Própria)	33
10	Estado civil da gestante com relação a quantidade de recém-nascidos vivos e mortos. (Fonte: Adaptado de (BELUZO et al., 2020b))	34
11	Peso com relação a quantidade de recém-nascidos que vieram a óbito. (Fonte: Elaboração Própria)	35
12	Semanas de gestação com relação a quantidade de recém-nascidos vivos e mortos. (Fonte: Adaptado de (BELUZO et al., 2020b))	35
13	Curva ROC dos modelos utilizando <i>K-fold cross validation</i> . (Fonte: El- aboração Própria)	41
14	Importância das variáveis na predição. (Fonte: Elaboração Própria) . .	46

Sumário

1	INTRODUÇÃO	15
2	JUSTIFICATIVA	17
3	OBJETIVOS	18
3.1	Objetivo Geral	18
3.2	Objetivos Específicos	18
4	FUNDAMENTAÇÃO TEÓRICA	19
4.1	Mortalidade Infantil e Neonatal	19
4.2	Aprendizado de Máquina	20
4.2.1	<i>Treinamento, teste e validação</i>	22
4.2.2	<i>Métricas de avaliação de modelos de classificação</i>	22
5	TRABALHOS RELACIONADOS	26
6	METODOLOGIA	28
6.1	Tecnologias e Ferramentas	28
6.2	Base de dados	28
6.3	Preparação da base de dados	30
6.4	Análise Exploratória dos Dados	32
6.5	Criação dos Modelos de Aprendizado de Máquina	36
6.6	Avaliação dos Resultados	37
6.7	Acesso a base de dados e código	38
7	RESULTADOS	39
7.1	Criação e análise dos modelos preditivos	39
7.2	Interpretações dos modelos preditivos	46
8	CONCLUSÃO	47
	REFERÊNCIAS	48

1 INTRODUÇÃO

A taxa de mortalidade infantil é um dos mais importantes indicadores de qualidade de vida e saúde de um país e é segmentada em duas categorias: neonatal e pós-neonatal. A taxa neonatal está relacionada ao número de óbitos ocorridos nos primeiros 28 dias de vida, quanto a taxa pós-neonatal, é o número de óbitos ocorridos entre 29 dias e 1 ano. Em ambas as taxas são considerados 1.000 nascidos vivos (BELUZO et al., 2020b).

A mortalidade infantil é um problema de escala mundial e sua redução é crucial. Aproximadamente 46% das mortes no mundo acontecem até os cinco anos de idade, sendo que, a maior parte desses óbitos ocorrem na primeira semana de vida do bebê. Os primeiros dias de vida de uma criança representam um período vulnerável e requer muita atenção e cuidados medicinais tanto com a criança quanto com a mãe, e a falta desse cuidado é um dos principais fatores que levam o recém-nascido a óbito (SILVA et al., 2016). Em 2010, no Brasil, estima-se que 70% dos óbitos infantis ocorridos poderiam ter sido evitados por ações de saúde e que 60% dos óbitos neonatais ocorreram por causas evitáveis (BARRETO; SOUZA; CHAPMAN, 2015). Com base nisso, conclui-se que é necessário o uso de tecnologias especializadas, como aprendizado de máquina, para substanciar o poder dos estudos, permitindo a visualização e manipulação de dados oriundos de grandes partes da população, tornando possível a elaboração de indicadores de acompanhamento (FRIAS et al., 2017). Segundo (RAJ-KOMAR; DEAN; KOHANE, 2019), a criação de modelos de Aprendizado de Máquina (*Machine Learning*) utilizando dados da mortalidade neonatal podem apresentar resultados significativos, uma vez que estes modelos têm a capacidade de aprender padrões presentes nos dados, que humanos, muitas vezes, não conseguem encontrar, contribuindo para com diagnósticos ou precauções que deverão ser tomadas.

A partir disso, alguns trabalhos foram desenvolvidos sobre a mortalidade neonatal e, também sobre o uso de aprendizado de máquina na área da saúde. Podemos citar o trabalho desenvolvido por (DEMITTO et al., 2017), que teve como objetivo identificar os fatores associados a mortalidade neonatal em um hospital no estado do Paraná, no Brasil. Podemos citar também, (AGUIAR, 2019) que propõe o uso de algoritmos de aprendizado de máquina para gerar modelos preditivos da taxa de Mortalidade Infantil no estado do Ceará, entre os anos 2013 e 2017, tal projeto gerou relevantes

resultados, que auxiliam a gestão pública e o combate da mortalidade infantil.

Deste modo, este projeto tem como principal contribuição realizar uma exploratória sobre as bases de dados Sistema de Informação sobre Mortalidade (SIM) e Sistema de Informações sobre Nascidos Vivos (SINASC) da cidade de São Paulo, entre os anos 2012 e 2018, assim como criar modelos de aprendizado de máquina para prever o risco de morte durante o período neonatal e avaliar os resultados. Tais modelos podem ser usados como apoio a criação de políticas públicas para a prevenção de morte neonatal.

Este trabalho é constituído por mais sete seções. Na seção 2 são apresentadas as justificativas para o desenvolvimento deste trabalho. A seção 3 apresenta os objetivos que o trabalho se propôs a alcançar. A seção 4 apresenta os conceitos necessários para um melhor entendimento dessa monografia. Na seção 5 são apresentados trabalhos relacionados ao assunto abordado neste trabalho. A seção 6 apresenta o método utilizado na execução deste trabalho e as tecnologias e ferramentas utilizadas. Os resultados obtidos e as conclusões são apresentadas nas seções 7 e 8, respectivamente.

2 JUSTIFICATIVA

Na contemporaneidade, a presença da tecnologia está alcançando diversas áreas da nossa vida facilitando-a e muitas vezes, melhorando-a. Além do uso cotidiano, a tecnologia pode ser usada para pesquisas e análises de dados de áreas como a saúde, podendo garantir autonomia e soberania para definição de políticas nacionais que podem garantir os direitos e o bem-estar da população (GUIMARÃES et al., 2019).

Considerando a saúde e bem-estar da mãe e o recém-nascido, temos que a taxa de mortalidade é o número de óbitos de crianças nos primeiros 28 dias de vida a cada 1.000 nascidos vivos. No mundo, em 2018, o número de mortalidades neonatais foi de 2.5 milhões, que corresponde a 47% do total de mortes de crianças com menos de 5 anos (UNIGME, 2019). São vários os fatores que podem contribuir para com a redução de óbitos neonatais, como acompanhamento médico à gestante no período de gravidez e do recém-nascido nos primeiros dias de vida, e a disponibilização deste serviço com qualidade e proficiência (WHO, 2009).

Neste contexto, este trabalho apresenta uma análise exploratória de dados da mortalidade neonatal para a cidade de São Paulo, assim como a criação de modelos de Aprendizado de Máquina para prever o risco de morte durante o período neonatal. Esses recursos podem auxiliar e contribuir com a criação de políticas públicas para diminuir a taxa de mortalidade neonatal.

3 OBJETIVOS

3.1 Objetivo Geral

O objetivo deste trabalho é realizar uma análise exploratória de dados da mortalidade neonatal da cidade de São Paulo e criar modelos de aprendizado de máquina capazes de prever o risco de morte durante o período neonatal.

3.2 Objetivos Específicos

- Realizar uma Análise Exploratória de Dados (AED);
- Criar modelos de predição de risco de morte neonatal utilizando diferentes técnicas de aprendizado de máquina;
- Analisar e avaliar os resultados e o desempenho preditivo dos modelos gerados;

4 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta os conceitos relacionados a mortalidade neonatal e ao aprendizado de máquina com intuito de contextualizar a temática deste trabalho.

4.1 Mortalidade Infantil e Neonatal

A mortalidade infantil é um indicador importante da qualidade da saúde pública de um país e sua redução é plausível, uma vez que estas mortes precoces podem ser, em sua maioria, evitáveis. A Taxa de Mortalidade Infantil (TMI) consiste no número de crianças que morreram antes de completar um ano de vida dividido por 1000 crianças nascidas vivas no período de um ano. A TMI pode ser dividida em 2 segmentos: a mortalidade neonatal, óbitos ocorridos entre 0 e 28 dias de vida, e a pós-neonatal, óbitos ocorridos entre 29 dias e 1 ano de vida. Já a Mortalidade Neonatal (MN), é subdividida em mortalidade neonatal precoce, óbito entre 0 e 6 dias de vida, e neonatal tardio, óbito entre 7 e 28 dias de vida. Tais taxas são afetadas por uma combinação de fatores biológicos, sociais, culturais e do sistema de saúde, sendo necessárias intervenções e políticas públicas relacionadas às condições de vida da população (FRANÇA; LANSKY, 2008). No mundo, entre 2000 e 2017 houve uma redução de 41% na taxa de mortalidade neonatal, porém, em 2017, ocorreram 2.5 milhões de mortes de crianças com menos de 1 mês de idade, sendo um número elevado considerando o avanço da saúde no passar dos anos (WHO, 2019). A mortalidade neonatal, no Brasil, corresponde a mais de 70% da mortalidade infantil, das quais 25% se encontram na subdivisão de mortalidade neonatal precoce, muitas vezes acontecendo nas primeiras 24 horas de vida (GAIVA; FUJIMORI; SATO, 2016), e por causas que, majoritariamente, poderiam ser evitadas.

Assim, a redução da mortalidade neonatal, no Brasil, é de extrema importância, sendo que, é o período mais vulnerável para a sobrevivência da criança e é fortemente influenciado por condições desfavoráveis de vida e da atenção à saúde. Algumas providências que podem ser tomadas para evitar a ocorrência de óbitos neonatais, são : a prevenção de gravidez de alto risco, a qualidade da assistência dada a gestante no período gestacional, no parto e no período pós-parto, assim como, o cuidado com o recém-nascido, nos seus primeiros momentos de vida e, também, na sua estadia no hospital (KASSAR et al., 2013).

4.2 Aprendizado de Máquina

O aprendizado de máquina é um método de análise de dados que, por meio de algoritmos, consegue encontrar e aprender padrões presentes nos dados recebidos que, muitas vezes, humanos não conseguiriam sozinhos (CHEN; ASCH, 2017), assim como prever valores a partir dos padrões aprendidos. A partir dos algoritmos de aprendizado de máquina e da base de dados, são criados os modelos de acordo com o objetivo da análise. Entre as muitas abordagens de aprendizado de máquina, temos a supervisionada e a não supervisionada. Na abordagem supervisionada o treinamento, especificado na seção 4.3.1, do modelo de aprendizado de máquina é realizado a partir dos resultados da predição esperados, chamados de "labels" (GERON, 2019) e sendo assim, seu objetivo é prever um valor.

Existem várias técnicas de aprendizagem de máquina supervisionado, entre elas, temos a de classificação na qual o resultado do modelo é uma classe, ou seja, o algoritmo tem como objetivo realizar a predição de valores categóricos, pré-definidos (MULLER; GUIDO, 2016). Um exemplo de um algoritmo de classificação é o *Random Forest* (RF) que tem como base o conceito de árvore de decisão, no qual é o aprendizado através das estruturas que envolvem expressões com operadores matemáticos, condicionais e lógicos. Na árvore de decisão as estruturas condicionais se encontram nos nós da árvore e os últimos nós correspondem ao resultado da predição. Na Figura abaixo temos uma árvore de decisão que utiliza a expressão condicional se/senão.

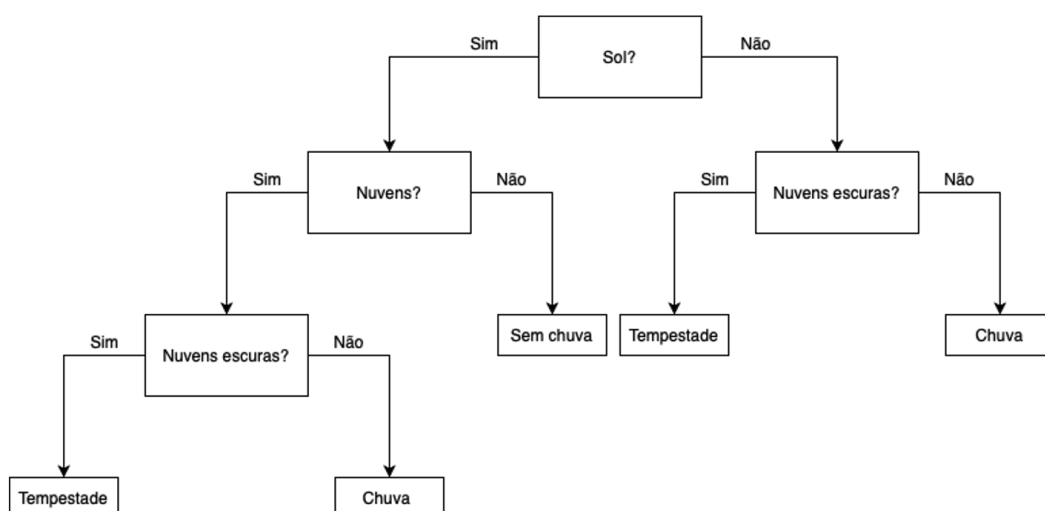


Figura 1: Exemplo de árvore de decisão. (Fonte: Elaboração Própria)

Como podemos observar na Figura 1, a resposta de cada nó é o que guia a predição, ou seja, as perguntas são feitas de acordo com as respostas dos nós anteriores. Sendo assim, a decisão é tomada a partir do caminhar do nó inicial até o resultado da predição (MULLER; GUIDO, 2016). Também podemos observar, que é possível ter 2 ou mais classes (Sem Chuva, Chuva, Tempestade) como resultado da classificação, sendo chamados de multiclasse.

Podemos citar também, outra técnica de algoritmo de aprendizado de máquina supervisionado, que é utilizada em problemas de previsão. Diferente dos algoritmos de classificação, o resultado da predição é composto por um valor real, tal como o preço de um carro (GERON, 2019) ou o valor de uma ação na bolsa de valores num determinado período de tempo futuro. Um exemplo de algoritmo de máquina supervisionada de previsão é o *Linear Regression* que a partir de uma função linear com as variáveis recebidas, prevê o resultado.

Outra abordagem de aprendizado de máquina é a não supervisionada, que ao contrário da supervisionada, no treinamento não são utilizadas as "*labels*", ou seja, não existe um resultado esperado, o modelo irá descobrir as semelhanças entre as variáveis afim de agrupar em classes ou detectar anomalias nos dados (GERON, 2019). Há várias técnicas de aprendizado de máquina não supervisionado, o mais utilizado é o algoritmo de "*Cluster*" ou agrupamento, que tem como objetivo agrupar os dados em diferentes classes. Um exemplo de um algoritmo de *cluster* é o *K-means*, que utiliza o conceito matemático de centróide para realizar o agrupamento dos dados.

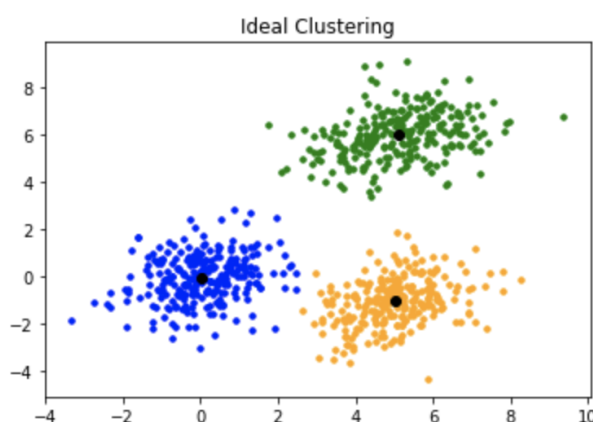


Figura 2: Exemplo do resultado do algoritmo Kmeans. (Fonte: Adaptado de ML | K-means++ Algorithm - Khosla, 2020)

Como podemos observar na Figura 2, existem 3 pontos da cor preta no centro de

cada grupo, esses pontos são os centróides e a partir deles, o agrupamento acontece.

4.2.1 **Treinamento, teste e validação**

O treinamento do modelo de aprendizado de máquina é a etapa na qual ocorre a criação do modelo a partir dos dados recebidos. Já na etapa de teste, temos a validação e o teste do modelo, para que sua performance seja avaliada (MUKHIYA; AHMED, 2020).

A fim de obter uma validação mais robusta dos resultados obtidos, no treinamento do modelo, utiliza-se métodos como o *K-fold cross-validation*, que divide a base de dados em K partes iguais e são realizadas várias rodadas de teste alternando as partes utilizadas para teste e treinamento. Tal processo é repetido K vezes, até que todas as partes sejam utilizadas para o teste (FONTAINE, 2018). Com este método, é possível avaliar os algoritmos e comparar os erros de predição, para que possam ser feitas mudanças para melhorá-lo ou para que se ache um algoritmo que mais se adeque a necessidade (CROWN, 2015). Podemos observar na Tabela 1, uma ilustração que exemplifica o método *K-fold cross-validation*, utilizando 5 validações. A taxa de erro média é calculada a partir da média das taxas de cada rodada.

Predição 1	Teste	Treinamento	Treinamento	Treinamento	Treinamento
Predição 2	Treinamento	Teste	Treinamento	Treinamento	Treinamento
Predição 3	Treinamento	Treinamento	Teste	Treinamento	Treinamento
Predição 4	Treinamento	Treinamento	Treinamento	Teste	Treinamento
Predição 5	Treinamento	Treinamento	Treinamento	Treinamento	Teste

Tabela 1: Exemplo de *K-fold cross-validation*. (Fonte: Elaboração Própria)

Outro método utilizado na validação dos modelos de aprendizado de máquina é o *Hold Out*, que consiste na divisão da base de dados em duas partes distintas, os dados utilizados para o treinamento e os dados utilizados para realizar o teste dos modelos (FONTAINE, 2018). A divisão da base de dados frequentemente utilizada é a qual temos 70% para a etapa do treinamento e 30% para a etapa de teste do modelo.

4.2.2 **Métricas de avaliação de modelos de classificação**

Uma das melhores formas de avaliar a performance do modelo é através da matriz de confusão (*confusion matrix*), pois considera o comportamento do modelo em relação a cada classe. Utilizando as variáveis binárias positive (1) e false (0), a matriz de

confusão apresenta uma comparação em forma de matriz entre a predição do modelo e os valores corretos correspondentes (SHUKLA, 2018).

		Previsto	
		Negativo (0)	Positivo (1)
Atual	Negativo (0)	VN	FP
	Positivo (1)	FN	VP

Tabela 2: Matriz de confusão. (Fonte: Elaboração Própria)

Como pode ser visto na Tabela 2, a matriz de confusão consiste em duas linhas, duas colunas e, conseqüentemente, 4 quadrantes, sendo que, os valores são atribuídos aos quadrantes de acordo com as seguintes situações: quando o valor da predição é 0 e corresponde ao valor correto, denominamo-lo Verdadeiro Negativo (VN). Já o quadrante Falso Negativo (FN) representa os valores das predições iguais a 0, que pelo contrário, correspondem ao valor 1. Denominamos Verdadeiro Positivo (VP) os resultados das predições iguais a 1 que de fato estão corretos, e Falso Positivo (FP), quando os valores resultantes da predição são iguais a 1 e não condizem com os valores corretos, que nesse caso seria 0.

Apesar destas denominações serem importantes e úteis individualmente, podemos utilizar métricas mais explicativas e relevantes. Como a acurácia, em inglês *accuracy*, que é utilizada para calcular a proximidade entre o valor obtido na predição dos modelos e os valores esperados, ou seja, entre todos os valores oriundos da predição, quantos estavam de fato corretos $((VP+VN)/(VP+VN+FP+FN))$. Já o índice da precisão, em inglês *precision*, é uma métrica que apresenta o cálculo de quanto o modelo realizou a predição correta da classe positiva, considerando o número total de predições da mesma classe, estando esta correta ou incorreta $(VP/(VP+FP))$. A cobertura, em inglês *recall*, entretanto, é outra métrica que representa o cálculo da quantidade de predições corretas da classe positiva em relação ao número total de valores que são, de fato, da classe positiva $(VP/(VP+FN))$ (SHUKLA, 2018). Na Tabela 3, temos um exemplo prático da matriz de confusão.

		Previsto	
		Negativo (0)	Positivo (1)
Atual	Negativo (0)	250	43
	Positivo (1)	21	369

Tabela 3: Exemplo Matriz de Confusão. (Fonte: Elaboração Própria)

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} = \frac{369 + 250}{369 + 250 + 43 + 21} = 0,90629575$$

$$\text{Precisão} = \frac{VP}{VP + FP} = \frac{369}{369 + 43} = 0,89563107$$

$$\text{Cobertura} = \frac{VP}{VP + FN} = \frac{369}{369 + 21} = 0,94615385$$

Esse exemplo apresenta a matriz de confusão oriunda de um modelo de aprendizado de máquina que tinha como objetivo prever se o email é spam (1) ou normal (0). Ao analisar a matriz, podemos afirmar que das 683 predições realizadas, 619 tiveram resultado correto, tendo um valor de acurácia alto. Assim como as métricas precisão e cobertura, que também são altos e apresentam um equilíbrio.

Outra forma de analisar a performance de um modelo de aprendizado de máquina é através de gráficos, como a Curva Característica de Operação do Receptor, ou, do inglês, *Receiver Operating Characteristic* (ROC) curve e a Curva de Precisão e Cobertura, ou, do inglês, *Precision Recall* (PR) curve. O ROC é um gráfico que analisa e caracteriza o comportamento do modelo de aprendizado de máquina com base nos valores da Taxa de Verdadeiros Positivos (Sensibilidade), ou, do inglês, *True Positive Rate* (TPR), que corresponde ao resultado da expressão matemática $VP/(VP+FN)$, e da Taxa de Falsos Positivos (Especificidade), ou *False Positive Rate* (FPR), que consiste o resultado da expressão matemática $FP/(FP+VN)$ (FLACH, 2003). Com base na ROC, é possível calcular a Área abaixo da curva, do inglês, *Area Under the Curve* (AUC), que apresenta a área da forma originada abaixo da linha do ROC curve, como pode ser visto na Figura 3. O uso do gráfico ROC para realizar a análise do modelo de aprendizagem de máquina é apropriado quando se tem uma base de dados balanceada, já o gráfico PR é utilizado quando se tem uma base de dados desbalanceada, uma vez que utiliza as métricas precisão e cobertura.

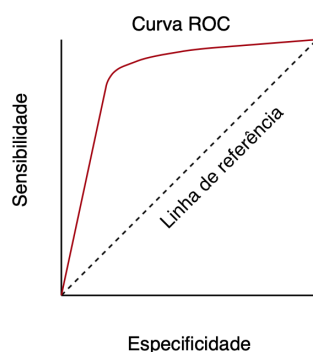


Figura 3: Curva ROC. (Fonte: Elaboração Própria)

O valor do 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. Já o gráfico PR, como podemos ver na Figura 4, apresenta uma relação entre os valores precisão e cobertura. Para se ter um resultado bom é desejável que ambos os valores sejam altos. Assim como no gráfico ROC, a partir do gráfico PR é possível calcular um índice que chama Precisão Média, ou do inglês, *Average Precision* (AP), este índice apresenta a precisão média do modelo em relação à cobertura.

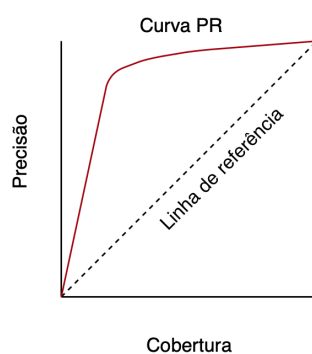


Figura 4: Curva PR. (Fonte: Elaboração Própria)

5 TRABALHOS RELACIONADOS

A alta taxa de mortalidade neonatal não é um problema recente e ocorre, muitas vezes, por causas evitáveis e relacionadas à qualidade do sistema de saúde, assim como do tratamento que o bebê e a mãe recebem antes do parto e nos primeiros dias de vida do recém-nascido. Entre as possíveis soluções para este problema, temos a melhora do sistema de saúde, acompanhamento qualificado do período gestacional e do período pós gestação (WHO, 2019).

Na área da medicina, sempre foi exigido que médicos saibam como manipular e analisar grande quantidade de dados, oriundas de fisiologias, imagens, estudos laboratoriais, entre outros. Conforme a complexidade das condições dos pacientes aumentam, algoritmos de aprendizado de máquina podem se tornar uma ferramenta indispensável e auxiliadora para estes médicos (OBERMEYER; EMANUEL, 2016). E, com o aumento da quantidade de dados, se torna viável integrar diversas variáveis, com técnicas computacionais, para entender e encontrar os determinantes do status da saúde de um paciente, para que também, no futuro, doenças e óbitos possam ser evitados (SONG et al., 2004).

Sendo assim, foram desenvolvidos vários estudos e projetos que abordam a taxa de mortalidade neonatal e o uso da tecnologia para apoiar e auxiliar esta causa. Entre eles, podemos citar o artigo escrito por (DEMITTO et al., 2017) que propõe um estudo sobre a mortalidade neonatal no estado do Paraná. Ao concluir o artigo, (DEMITTO et al., 2017) apresenta que os principais fatores de risco são, muitas vezes, relacionados à qualidade do sistema de saúde e a qualificação dos processos assistenciais no período pré-natal da gestante, parto e do recém-nascido. Além de recomendar a criação de políticas e programas com foco na melhora do atendimento as gestantes e ao bebê tanto no período pré-natal quanto após o parto.

Podemos citar também, o estudo realizado por Aguiar (AGUIAR, 2019), que desenvolveu modelos preditivos com base no conceito de Redes Neurais, uma vertente da inteligência artificial, utilizando os dados da mortalidade infantil no estado do Ceará. Uma das conclusões apresentadas por (AGUIAR, 2019), é o destaque a importância do grau de instrução da gestante, uma vez que 50% dos óbitos infantis ocorreram entre as mães de até 25 anos que não concluíram o Ensino Fundamental I. Além do grau de instrução da gestante, também foi apresentadas como possíveis causas a falta de

acesso à informação, aos cuidados necessários e a higiene.

Em (BELUZO et al., 2020b), os autores utilizaram a base de dados da mortalidade neonatal na cidade de São Paulo para realizar um estudo e, também, criar modelos de aprendizado de máquina supervisionado, com o intuito de auxiliar a criação de políticas públicas para diminuir a taxa de mortalidade neonatal no Brasil. Na conclusão (BELUZO et al., 2020b) discute os fatores que tiveram mais influência nas previsões realizadas, entre eles, podemos citar o peso do recém-nascido, a nota Apgar do primeiro e quinto minuto, as semanas gestacionais e a malformação do bebê.

6 METODOLOGIA

Esta seção apresenta a metodologia utilizada no desenvolvimento deste projeto, a qual é ilustrada pela Figura 5. As etapas são descritas nas subseções presentes nessa seção.

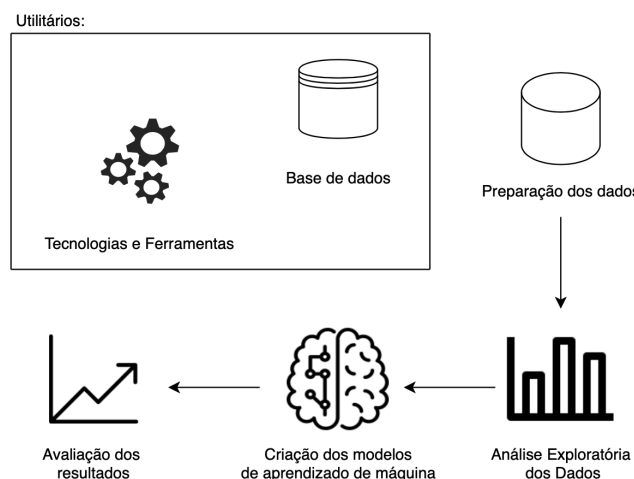


Figura 5: Etapas da Metodologia. (Fonte: Elaboração Própria)

6.1 Tecnologias e Ferramentas

Para desenvolver este trabalho foi utilizada a linguagem de programação *Python*, devido ao seu alto poder de processamento e simples codificação. As bibliotecas utilizadas foram: *pandas* e *numpy* para tratar e manusear os dados, *matplotlib* e *seaborn* para criar os gráficos, e *sklearn*, *xgboost* e *SHAP* para a criação e análise de modelos preditivos. A plataforma utilizada para a codificação e desenvolvimento foi o *Jupyter Notebook*, escolhida por suas vantagens, como a possibilidade de executar os scripts em partes.

6.2 Base de dados

Para realizar este trabalho, foram utilizadas as bases de dados SIM e SINASC da cidade de São Paulo dos anos 2012-2018, providenciada pelo Departamento de Informática do SUS (DATASUS). Estes dados contém as informações tanto dos bebês que estão vivos quanto dos bebês que vieram a óbito, tais como o peso, local de nascimento, escolaridade da mãe e outros. A base de dados utilizada foi coletada e

processada por (BELUZO et al., 2020a) e em seu artigo, são apresentadas as etapas e as explicações de como foi realizado tal processo. Na Figura 6 é apresentada a distribuição dos dados com base na ocorrência de óbito ou durante o período neonatal. Dos 1.425.834 registros, 1.427.906 (99,4%) representam recém-nascido vivos, enquanto 7928 (0,6%) são registros de recém-nascidos que vieram a óbito. Esses números mostram que a base de dados é altamente desbalanceada.

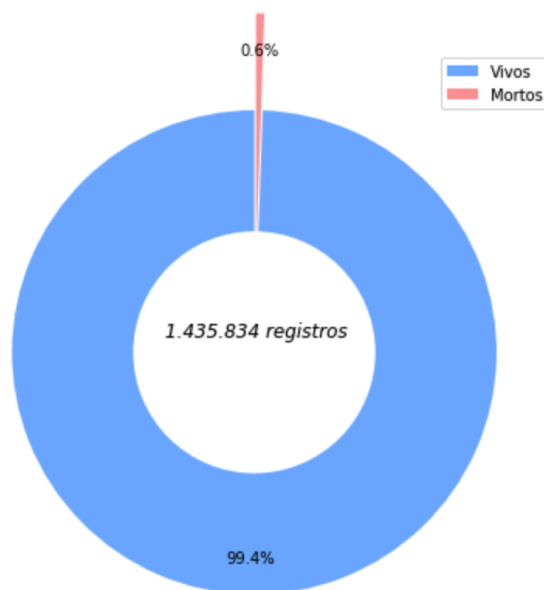


Figura 6: Distribuição da base de dados considerando se o recém-nascido viveu ou morreu durante os período neonatal. (Fonte: Elaboração Própria)

Variável	Descrição
n_nu_peso	Peso do recém-nascido, em gramas.
n_st_malformacao	Presença de anomalia.
n_nu_semana_gestacao	Número de semanas de gestação.
n_nu_apgar1	Nota Apgar no primeiro minuto.
n_nu_apgar5	Nota Apgar no quinto minuto.
n_tp_escolaridade	Escolaridade em anos de estudo concluídos.
n_tp_prenatal	Número de consultas no período pré-natal.
n_nu_idade	Idade da gestante.
n_tp_grupo_robson	Classificação Robson.
n_tp_raca_cor_mae	Raça/cor da gestante.
n_qt_parto_normal	Número de partos normais antecedentes.
n_tp_parto	Tipo de parto.
n_tp_gestacao	Semana de gestação.
n_tp_estado_civil	Situação conjugal da gestante.
n_tp_gravidez	Tipo de gravidez.
n_qt_gestacao_anterior	Número de gestações anteriores.
n_tp_funcao_responsavel	Tipo de função do responsável pelo preenchimento.
n_qt_nascidos_vivos	Número de nascidos vivos nas gestações antecedentes.
n_qt_nascidos_mortos	Número de nascidos mortos nas gestações antecedentes.
n_tp_ocorrendia	Local de nascimento.
n_tp_nascimento_assistido	Função de quem assistiu o parto.
n_tp_apresentacao	Tipo de apresentação do recém-nascido.
n_qt_parto_cesarea	Número de partos cesárea antecedentes.

Tabela 4: Exemplo de *K-fold cross-validation*. (Fonte: Elaboração Própria)

A base de dados é composta por variáveis que contêm valores reais ou categóricas, assim como é mostrado na Tabela 4. A coluna 'n_nu_peso', por exemplo, apresenta o peso do recém-nascido e é composta por valores reais, enquanto a coluna 'n_tp_estado_civil', é populada com valores categóricos de 1 a 5, que indicam o estado civil da gestante que deu a luz ao recém-nascido, podendo ser solteira, casada, viúva, divorciada ou com união estável.

6.3 Preparação da base de dados

Nesta etapa, foi realizado um pré-processamento dos dados, que consiste na limpeza, transformação e preparação dos dados que serão utilizados na análise exploratória e na criação dos modelos preditivos. Para realizar tanto este processamento foram utilizadas as bibliotecas *pandas* e *numpy* que auxiliam o tratamento de dados.

Foi realizada a conversão de algumas colunas para o tipo de valor inteiro e os registros que continham valores que não eram válidos ou valores que não correspondem a nenhuma categoria, receberam a moda ou a média da coluna (variável) correspondente. Além disso, foram geradas colunas novas, a partir das já presentes, que seriam necessárias na etapa da Análise Exploratória dos dados, por exemplo a extração do

ano a partir da coluna que contém a data (dia, mês e ano) na qual o bebê veio a óbito e a coluna que armazena números binários (0 e 1) indicando se ocorreu mortalidade neonatal.

Variável	Valores Originais	Valores Finais
n_tp_estado_civil	['2' '1' '5' '4' ' ' '3' '5.0' '2.0' '1.0' '4.0' '3.0' '9.0']	[1 2 3 4 5]
n_tp_escolaridade	['4' '3' '5' '2' '1' ' ' '0' '4.0' '3.0' '5.0' '2.0' '1.0' '9.0' '0.0']	[0 1 2 3 4 5]
n_tp_gestacao	['5' '4' ' ' '6' '2' '3' '1' '5.0' '6.0' '4.0' '3.0' '2.0' '1.0']	[1 2 3 4 5 6]
n_tp_gravidez	['1' '2' '3' ' ' '1.0' '2.0' '3.0']	[1 2 3]
n_tp_parto	['2' '1' '2.0' '1.0' ' ']	[1 2]
n_st_malformacao	['2.0' '1.0' '2' '1' ' ' '9.0']	[1 2]
n_tp_raca_cor_mae	['1' '4' '2' ' ' '3' '5']	[1 2 3 4 5]
n_tp_apresentacao	['1' '2' ' ' '3']	[1 2 3]
n_tp_nascimento_assistido	['1' '2' ' ' '4' '3']	[1 2 3 4]
n_tp_funcao_responsavel	['5' '2' ' ' '1' '3' '4']	[1 2 3 4 5]
n_tp_grupo_robson	['5' '7' '4' '3' '2' ' ' '10' '11' '1' '6' '8' '9']	[1 2 3 4 5 6 7 8 9 10 11]

Tabela 5: Colunas que foram padronizadas, seus valores originais e finais. (Fonte: Elaboração Própria)

Na Tabela 5, temos as colunas que passaram pelo processo de limpeza e seus valores antes e depois da mesma. Em todas as colunas, foi feita a conversão do tipo de dados, do tipo *Object* para inteiro, assim como a substituição de valores que não se encontravam na amostra das variáveis categóricas, pela respectiva moda das mesmas.

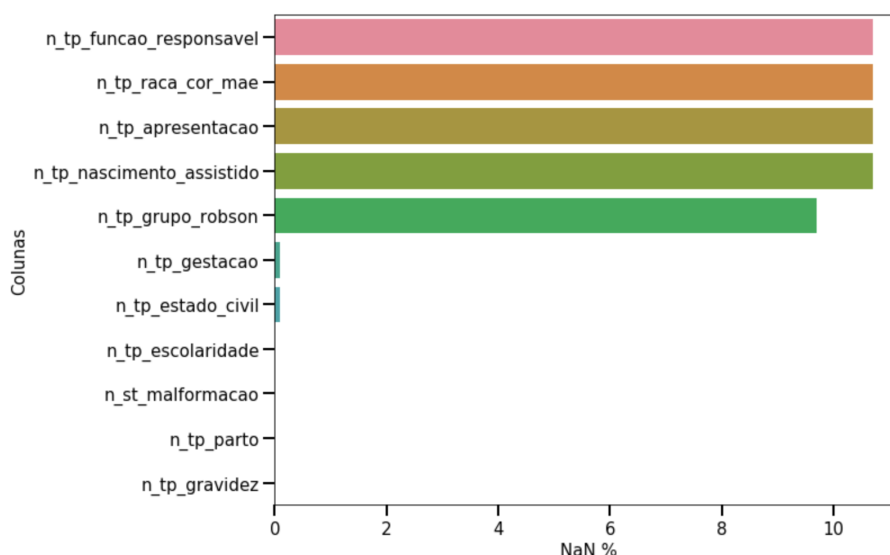


Figura 7: Gráfico que apresenta a porcentagem dos valores NaN presentes em algumas colunas. (Fonte: Elaboração Própria)

A sigla "NaN", em inglês, *Not a Number*, que significa "não é um número", é utilizada para substituir os campos que não possuem valor ou que possuem valores, do tipo *string*, vazios. Na Figura 7, há um gráfico que apresenta a porcenta-

gem de valores "NaN" em cada coluna, assim como sua porcentagem correspondente. A figura mostra as 5 colunas que contém um maior número de NaN, são elas: 'n_tp_função_responsavel', 'n_tp_raca_cor_mae', 'n_tp_apresentacao', 'n_tp_nascimento_assistido', 'n_tp_grupo_robson'.

6.4 Análise Exploratória dos Dados

A Análise Exploratória de Dados tem como objetivo realizar um estudo sobre os dados de tal forma que se possa extrair informações, descobrir padrões, visualizar e entender os dados (MUKHIYA; AHMED, 2020). Na etapa da Análise Exploratória dos Dados, foi realizado, primeiramente, um estudo de quais gráficos apresentam informações relevantes e, também, de fácil e simples interpretação. Os tipos de gráficos escolhidos foram: o gráfico de barra, linha, *boxplot* e circular.

Os gráficos foram criados com o uso de uma, duas ou mais variáveis da base de dados, que, por sua vez, foi filtrada e apenas os registros de óbitos foram utilizados na criação dos gráficos. Ao utilizarmos uma variável apenas na criação dos gráficos, foi possível analisar o comportamento desta sozinha, sem a interferência das outras. Já os gráficos que foram criados com duas ou mais variáveis, tiveram o propósito de analisar a relação entre as variáveis, o comportamento destas quando analisadas em conjunto.

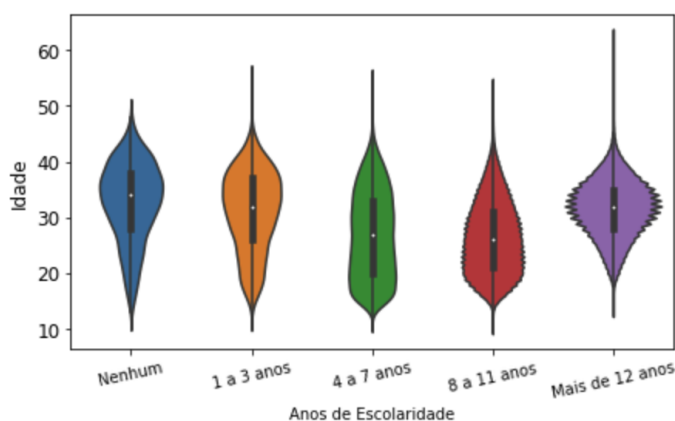


Figura 8: Gráfico da relação entre a idade e escolaridade da gestante. (Fonte: Elaboração Própria)

A Figura 8 apresenta a relação entre a escolaridade e a idade da gestante no ano o qual o recém-nascido veio a óbito. A maior parte das gestantes que não tinham escolaridade tinham entre 30 e 40 anos, enquanto as gestantes com 1 a 3 anos de

escolaridade tinham entre 20 e 40 anos. Já a maioria das gestantes com escolaridade entre 4 e 7 anos, tinham entre 18 e 35 anos, assim como as gestantes com 8 a 11 anos de escolaridade. As gestantes com mais de 12 anos de escolaridade, tinham, em sua maior parte, idades entre 25 e 35 anos. A partir disso, é possível analisar e concluir que a maior parte das gestantes tinham mais de 25 anos e que não há uma diferença significativa na quantidade de ocorrências, se analisarmos os anos de escolaridade.

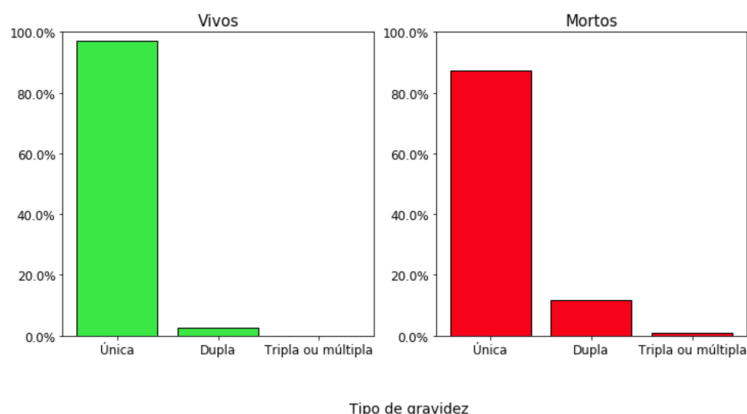


Figura 9: Tipo de gravidez com relação a quantidade de recém-nascidos vivos e mortos. (Fonte: Elaboração Própria)

O gráfico representado pela Figura 9, apresenta a relação do tipo de gravidez e a quantidade de recém-nascidos vivos e mortos, sendo que, os tipos de gravidez, são: única (um bebê), dupla (dois bebês) e tripla ou múltipla (três ou mais bebês). No caso de recém-nascidos vivos, a porcentagem do tipo de gravidez única corresponde a mais de 95%, enquanto a dupla corresponde a menos de 5% e a tripla ou múltipla tem um valor muito perto de 0%. Já no caso de recém-nascidos que vieram a óbito, a porcentagem correspondente ao tipo de gravidez única, está entre 80% e 90%, a porcentagem do tipo de gravidez dupla é de aproximadamente 10% e a tripla tem seu valor menor que 5%. Ao observar os gráficos, podemos concluir que nos dois casos, o tipo de gravidez que predominou foi a única.

Na Figura 10, temos dois gráficos criados por (BELUZO et al., 2020b) que apresentam a relação entre o estado civil da gestante e sua porcentagem correspondente. No primeiro gráfico temos a relação dos recém-nascidos vivos e no segundo temos dos que vieram a óbito. No eixo X, temos as seguintes categorias: solteira, casada, viúva, divorciada e com uma relação estável. O gráfico dos recém-nascidos vivos, apresenta que aproximadamente 42% das gestantes eram solteiras, 40% eram casa-

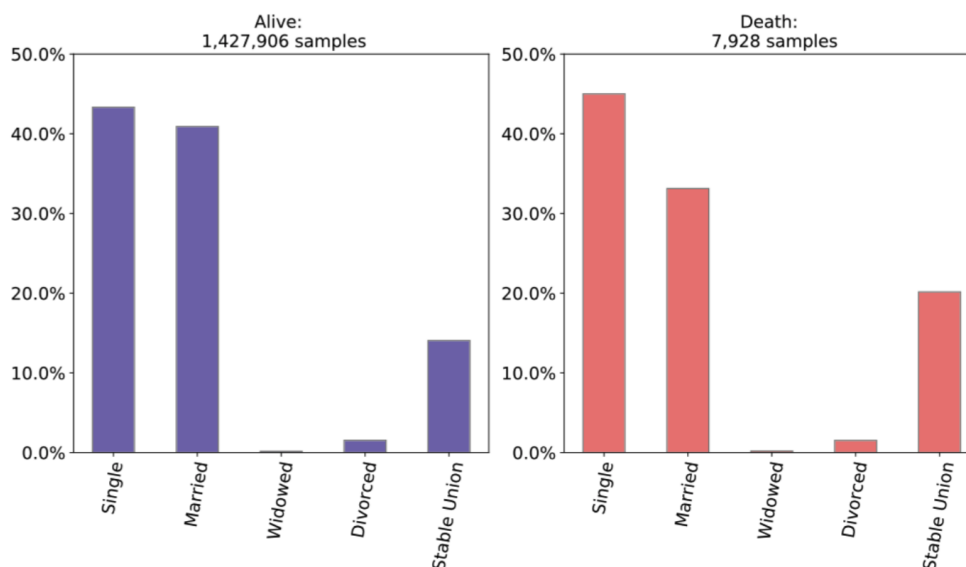


Figura 10: Estado civil da gestante com relação a quantidade de recém-nascidos vivos e mortos. (Fonte: Adaptado de (BELUZO et al., 2020b))

das. Enquanto menos de 5% eram viúvas ou divorciadas e aproximadamente 15% tinham uma relação estável. Já no gráfico dos recém-nascidos que vieram a óbito, aproximadamente 45% das gestantes eram solteiras, em torno de 32% eram casadas e menos de 5% eram viúvas ou divorciadas, além disso, aproximadamente 20% das gestantes tinham um relacionamento estável. A partir disso, é possível afirmar que em ambos os casos dos recém-nascidos vivos e mortos, a diferença entre o percentual de solteiras e casadas é pequena. E também, é importante citar que o percentual de gestantes com união estável, compõe uma parte significativa do total, tanto de casos de mortos quanto de vivos.

O gráfico da Figura 11 apresenta a distribuição dos pesos, em gramas, dos recém-nascidos que vieram a óbito. É possível observar que os pesos variam entre aproximadamente 200 e 5000 gramas, tendo uma concentração maior de ocorrências entre aproximadamente 400 e 1500 gramas, com suas respectivas porcentagens entre 6%, e 15,1%. Já as ocorrências de recém-nascidos com peso maior que 1500 gramas, apresentam porcentagens constantes.

Já na Figura 12, temos outros gráficos criados por (BELUZO et al., 2020b) que apresentam a relação entre o número de semanas de gestação e sua porcentagem correspondente. No primeiro gráfico temos a relação dos recém-nascidos vivos e no segundo temos dos que vieram a óbito. No eixo X, temos os números de semanas de gestação. O gráfico dos recém-nascidos vivos apresenta que há uma concentração

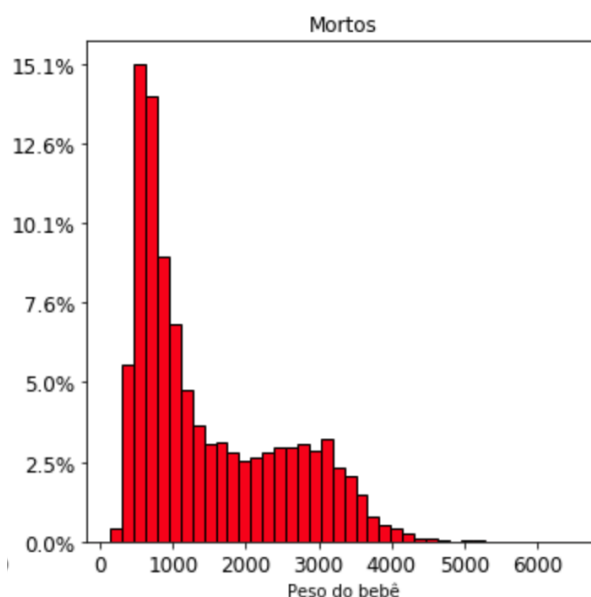


Figura 11: Peso com relação a quantidade de recém-nascidos que vieram a óbito. (Fonte: Elaboração Própria)

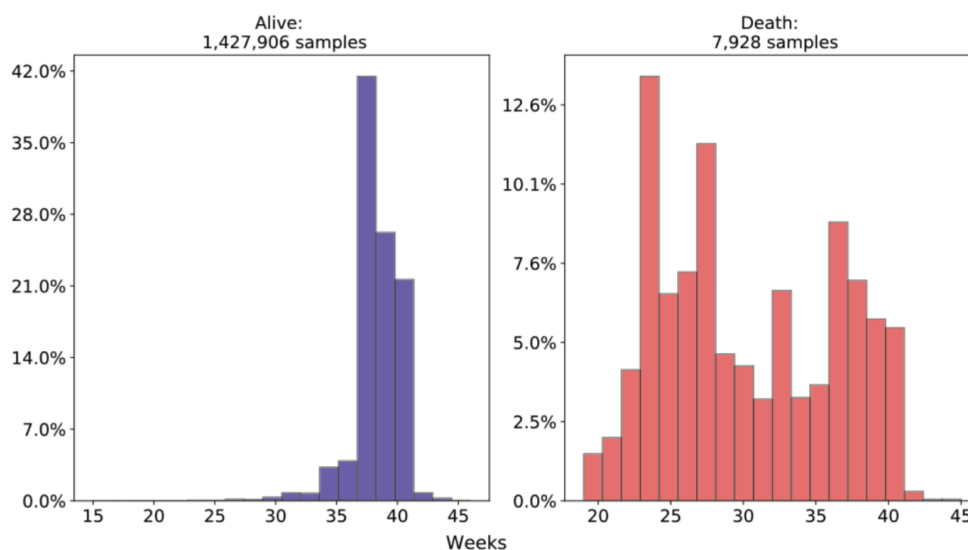


Figura 12: Semanas de gestação com relação a quantidade de recém-nascidos vivos e mortos. (Fonte: Adaptado de (BELUZO et al., 2020b))

de ocorrências com número de semanas de gestação entre 35 e 42, sendo que as porcentagens variam entre 42% e 21%, sendo que a variação se encontra entre 27 e 45 semanas. Já no gráfico dos recém-nascidos que vieram a óbito, a variação no eixo X se encontra entre 20 e 45 semanas, e a quantidade de ocorrências são mais distribuídas. O número de semanas que apresentam uma maior porcentagem é o de aproximadamente 25 semanas, por volta de 13% e o de aproximadamente 28 semanas, com cerca de 11.5%.

6.5 Criação dos Modelos de Aprendizado de Máquina

A escolha dos algoritmos de aprendizado de máquina foi baseada em um estudo das vantagens e desvantagens de cada um e também no objetivo deste projeto, que é prever o risco da mortalidade neonatal utilizando a base de dados descrita na seção 4. Os algoritmos de aprendizado de máquina utilizados neste projeto são da abordagem supervisionada e do tipo classificação, sendo eles: *Logistic Regression*, *Random Forest Classifier* e *XGBoost*. Foram escolhidos por serem algoritmos que têm alta performance e mais utilizados na comunidade acadêmica.

Como citado na seção da Fundamentação Teórica, muitos algoritmos de aprendizado de máquina supervisionada do tipo classificação utiliza o conceito de árvore de decisão, porém uma das desvantagens destes algoritmos, é o *overfitting*, que acontece quando o algoritmo adquire um bom resultado quando realiza o processamento com os dados de treinamento, porém quando utiliza dados que nunca processou, não se tem um bom resultado, este problema ocorre, uma vez que, o modelo aprende casos/regras muito específicas da base de dados de treinamento. Tal problema pode ser resolvido ao utilizar o modelo *Random Forest*, que realiza a predição através da média das predições de várias árvores de decisão aleatórias, por utilizarem diferentes variáveis nas estruturas condicionais de cada nó (MULLER; GUIDO, 2016).

O algoritmo *XGBoost*, também surgiu como uma solução do problema de *overfitting* das árvores de decisão, porém sua abordagem é diferente. Assim como o *Random Forest*, são utilizadas várias árvores para se criar um modelo mais poderoso, mas o objetivo deste algoritmo é criar estas árvores de forma sequencial, de certa forma que, cada árvore de decisão tem sua performance melhorada ao corrigir os erros da anterior. É considerado um dos melhores modelos para se utilizar, apesar de precisar de mais cuidado ao escolher os valores dos parâmetros (MULLER; GUIDO, 2016).

Já o algoritmo *Logistic Regression* foi implementado a partir dos modelos lineares, os quais realizam a predição a partir de uma função linear com as variáveis recebidas, apesar de ser baseado em modelos de regressão, o algoritmo prevê valores categóricos e utiliza a função linear para gerar uma linha, um plano ou um hiperplano com o intuito de separar as 2 ou mais classes (categorias) (MULLER; GUIDO, 2016).

Em relação aos parâmetros que foram utilizados na criação dos modelos, podemos destacar o uso do *"scale_pos_weight"* e *"class_weight"*, ambos têm o mesmo

propósito, que é dar um peso para uma variável, de acordo com a sua classe. Tal parâmetro foi de extrema importância neste projeto, uma vez que a base de dados utilizada era desbalanceada, ou seja, havia uma grande diferença entre o número dos valores da classe que representa se o bebê veio a óbito ou não, portanto a quantidade de bebês que vieram a óbito era muito menor do que a dos que sobreviveram. Como o objetivo do algoritmo de aprendizado de máquina é realizar o máximo de predições corretas, é esperado que o modelo simplesmente aprenda que a maioria dos valores é de certa classe, e arrisque predizer que quase todos os registros são desta classe, tendo-se assim uma acurácia alta (GERON, 2019). Por conta disso, devemos analisar, também, outros cálculos relevantes, como precisão e cobertura, e adequar o resultado esperado de acordo com o objetivo da criação do modelo. Neste projeto, a classe que representa os óbitos era minoria e os pesos aplicados nesta classe foram: 1, 10, 100, 200, 400, 1000.

Para realizar treinamento, teste e validação dos modelos, primeiramente foi utilizado o método *K-fold cross-validation*. Para os outros experimentos, foi-se utilizado o método *Hold Out* no qual a base de dados é dividida em 2 partes, treinamento e teste. Neste projeto, a distribuição dos dados foi feita utilizando 70% da base para o treinamento e 30% para teste.

6.6 Avaliação dos Resultados

Para realizar a análise e avaliação dos resultados das predições foram utilizadas as métricas explicadas na seção 4, a Fundamentação Teórica, como precisão, cobertura e acurácia, assim como os gráficos que apresentam a curva ROC e a curva PR.

Além disso, para realizar a interpretação do modelo, foi utilizado a biblioteca *SHapley Additive exPlanations* (SHAP) presente na linguagem *Python*, que mede a importância e influência de cada classe em relação ao resultado da predição e quais fatores contribuem a favor e contra cada classe (RATHI, 2019). O uso do SHAP é recomendável em modelos de complexidade alta e que utilizam um grande número de dados, nos quais, muitas vezes, nem pessoas especializadas na área conseguem interpretar e realizar as mesmas análises realizadas por tal biblioteca, apresentando informações que são tão cruciais quanto a acurácia da predição (LUNDBERG; LEE, 2017).

6.7 Acesso a base de dados e código

A base de dados utilizada neste projeto pode ser encontrada no link a seguir: <https://doi.org/10.7303/syn22240254>. Já os códigos desenvolvidos neste projeto podem ser encontrados neste link: <https://github.com/anacchp/TCC>.

7 RESULTADOS

7.1 Criação e análise dos modelos preditivos

Os modelos preditivos foram criados com o intuito de realizar uma predição que informa se o recém-nascido vai a óbito ou não (Vivo ou Morto). Para a criação, teste e validação dos primeiros modelos foi utilizado o método *K-fold cross validation*. Na Tabela 6, temos os resultados dos modelos criados com os algoritmos *XGBoost*, *Random Forest* e *Logistic Regression*. O modelo criado a partir do algoritmo *XGBoost* teve como acurácia aproximadamente 0.99, sendo o valor de precisão e cobertura da classe Vivo, 0.99 e 0.99, e os da classe Morto, 0.64 e 0.35. Já o modelo criado com o algoritmo *Random Forest*, teve a acurácia por volta de 0.99, precisão da classe Vivo de 0.99 e da classe Morto de 0.66, e cobertura da classe Vivo de 0.99 e 0.32 da classe Morto. A partir do algoritmo *Logistic Regression*, foi criado um modelo com cerca de 0.99 de acurácia, 0.99 e 0.99 de precisão e cobertura da classe Vivo e 0.66 e 0.32 de precisão e cobertura da classe Morto. Podemos observar que a acurácia dos 3 modelos é alta, porém os valores correspondentes a precisão e a cobertura para a classe Morto, tem uma grande diferença, mostrando que os modelos tiveram um baixo índice de acerto desta classe. No gráfico da Figura 13, o valor da AUC para o modelo *XGBoost* é o maior de todos, sendo assim, o que teve a melhor performance preditiva. É possível observar também que os modelos acertam quase todos os exemplos da classe vivo, pois sua tendência é ter uma cobertura maior da classe majoritária.

- XGBoost:

Acurácia = 0.9953490445274314

Matriz de confusão:

Atual	Previsto	
	Vivo (0)	Morto (1)
	Vivo (0)	Morto (1)
	Vivo (0)	1426365
	Morto (1)	5137

Relatório da Classificação:

	Precisão	Cobertura
Vivo (0)	0.99	0.99
Morto (1)	0.64	0.35

- Random Forest:

Acurácia = 0.9953455622307315

Matriz de confusão:

Atual	Previsto	
	Vivo (0)	Morto (1)
	Vivo (0)	Morto (1)
	Vivo (0)	1426605
	Morto (1)	5382

Relatório da Classificação:

	Precisão	Cobertura
Vivo (0)	0.99	0.99
Morto (1)	0.66	0.32

- Logistic Regression:

Acurácia = 0.9948907742817067

Matriz de confusão:

Atual	Previsto	
	Vivo (0)	Morto (1)
	Vivo (0)	Morto (1)
	Vivo (0)	1426719
	Morto (1)	6149

Relatório da Classificação:

	Precisão	Cobertura
Vivo (0)	0.99	0.99
Morto (1)	0.60	0.22

Tabela 6: Resultados das predições com os algoritmos *XGBoost*, *Random Forest* e *Logistic Regression* utilizando o *K-fold cross validation*. (Fonte: Elaboração Própria)

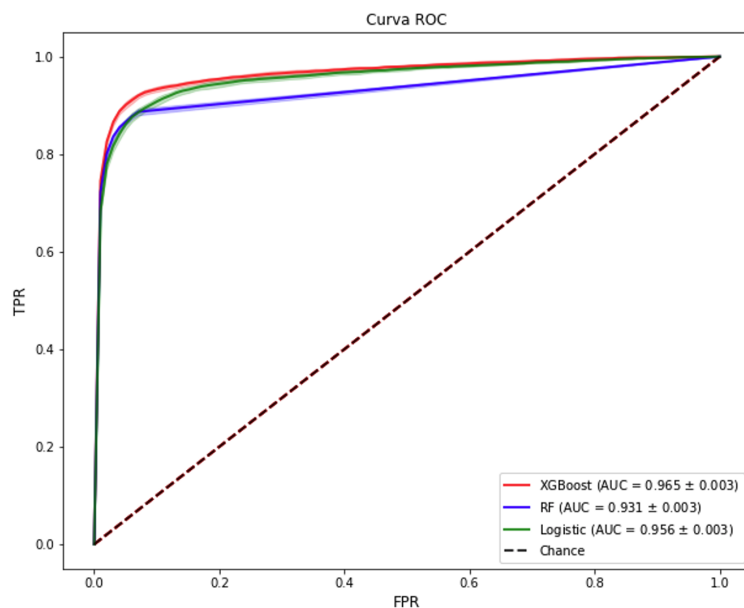


Figura 13: Curva ROC dos modelos utilizando *K-fold cross validation*. (Fonte: Elaboração Própria)

Os próximos experimentos foram realizados utilizando 70% da base para treinamento e 30% para teste. Para cada algoritmo, *XGBoost*, *Random Forest* e *Logistic Regression*, foi passado os seguintes valores como argumento para os parâmetros "*scale_pos_weight*" e "*class_weight*": 1, 10, 100, 200, 400, 1000.

Peso 1

Previsto

	Vivo (0)	Morto (1)
Vivo (0)	427889	453
Morto (1)	1582	827

Atual

Vivo (0)	427889	453
Morto (1)	1582	827

Acurácia: 0.9952756929177181

Precisão

Cobertura

Vivo (0)	0.99	0.99
Morto (1)	0.65	0.34

Peso 10

Previsto

	Vivo (0)	Morto (1)
Vivo (0)	425212	3130
Morto (1)	741	1668

Atual

Vivo (0)	425212	3130
Morto (1)	741	1668

Acurácia: 0.9910133696729665

Precisão

Cobertura

Vivo (0)	0.99	0.99
Morto (1)	0.35	0.69

Peso 100

Previsto

	Vivo (0)	Morto (1)
Vivo (0)	417476	10866
Morto (1)	421	1988

Atual

Vivo (0)	417476	10866
Morto (1)	421	1988

Acurácia: 0.9737969267627934

Precisão

Cobertura

Vivo (0)	0.99	0.97
Morto (1)	0.15	0.83

Peso 200

Previsto

	Vivo (0)	Morto (1)
Vivo (0)	413407	14935
Morto (1)	362	2047

Atual

Vivo (0)	413407	14935
Morto (1)	362	2047

Acurácia: 0.9644876042075352

Precisão

Cobertura

Vivo (0)	0.99	0.97
Morto (1)	0.12	0.85

Peso 400

Previsto

	Vivo (0)	Morto (1)
Vivo (0)	408759	19583
Morto (1)	348	2061

Atual

Vivo (0)	408759	19583
Morto (1)	348	2061

Acurácia: 0.9537296489154987

Precisão

Cobertura

Vivo (0)	0.99	0.95
Morto (1)	0.10	0.86

Peso 1000

Previsto

	Vivo (0)	Morto (1)
Vivo (0)	393703	34639
Morto (1)	300	2109

Atual

Vivo (0)	393703	34639
Morto (1)	300	2109

Acurácia: 0.9188881743745226

Precisão

Cobertura

Vivo (0)	0.99	0.92
Morto (1)	0.06	0.88

peso 1000, a acurácia é de 0.91, a cobertura da classe Vivo é de 0.92 e a precisão e cobertura da classe Morto é de 0.06 e 0.88. Ao analisar os resultados, é visível que, conforme o valor atribuído ao peso aumenta, a acurácia do modelo diminui, a precisão da classe Morto diminui e a cobertura aumenta. Vale ressaltar que neste problema, o resultado esperado seria prever corretamente os casos de morte, uma vez que são os casos que precisam de maior cuidado e atenção, os quais são representados pela cobertura da classe Morto.

Os resultados das previsões dos modelos criados a partir do algoritmo *Random Forest*, na Tabela 8, apresentam um comportamento diferente, se compararmos com os modelos preditivos do *XGBoost*. Quando atribuído um peso de 1, a acurácia do modelo é de aproximadamente 0.99, o valor da precisão e cobertura da classe Vivo são 0.99 e 0.99 e o valor da precisão e cobertura da classe Morto são 0.63 e 0.39. Já com o peso 10 a acurácia continua a mesma, assim como a precisão e a cobertura da classe Vivo, porém a precisão e a cobertura da classe Morto correspondem a 0.64 e 0.25. Quando o peso atribuído é 100, a acurácia do modelo é aproximadamente 0.99, sendo que a cobertura da classe Vivo diminui para 0.97 e a precisão e a cobertura da classe Morto é de 0.64 e 0.23. Com o peso 200, a acurácia continua com o valor aproximado de 0.99, assim como o valor da precisão e cobertura da classe Morto e Morto. Ao atribuir um peso de 400 na classe Morto, a acurácia continua a mesma, já a cobertura da classe Vivo diminui para 0.95, e há uma alteração na precisão da classe Morto, com valor de 0.63. Já com o peso 1000, a cobertura da classe Vivo é de 0.92 e a precisão e cobertura da classe Morto é de 0.65 e 0.23. Assim, conforme o peso aplicado na classe Morto aumenta, a acurácia dos modelos diminui um pouco e, a precisão e a cobertura permanecem constantes. Neste caso, o uso de pesos para balancear a base de dados não influenciou e nem melhorou a predição, se levarmos em conta o foco em prever as ocorrências de morte.

Peso 1				Peso 10				Peso 100			
Atual	Previsto			Atual	Previsto			Atual	Previsto		
		Vivo (0)	Morto (1)			Vivo (0)	Morto (1)			Vivo (0)	Morto (1)
	Vivo (0)	427932	410		Vivo (0)	428004	338		Vivo (0)	428029	313
	Morto (1)	1706	703		Morto (1)	1803	606		Morto (1)	1859	550
Acurácia: 0.9950876492451556				Acurácia: 0.9950296110746115				Acurácia: 0.994957643743137			
		Precisão	Cobertura			Precisão	Cobertura			Precisão	Cobertura
Vivo (0)		0.99	0.99	Vivo (0)		0.99	0.99	Vivo (0)		0.99	0.97
Morto (1)		0.63	0.29	Morto (1)		0.64	0.25	Morto (1)		0.64	0.23

Peso 200				Peso 400				Peso 1000			
Atual	Previsto			Atual	Previsto			Atual	Previsto		
		Vivo (0)	Morto (1)			Vivo (0)	Morto (1)			Vivo (0)	Morto (1)
	Vivo (0)	428023	319		Vivo (0)	428012	330		Vivo (0)	428034	308
	Morto (1)	1850	559		Morto (1)	1849	560		Morto (1)	1848	561
Acurácia: 0.9949646083236022				Acurácia: 0.9949413930553846				Acurácia: 0.9949947881722851			
		Precisão	Cobertura			Precisão	Cobertura			Precisão	Cobertura
Vivo (0)		0.99	0.97	Vivo (0)		0.99	0.95	Vivo (0)		0.99	0.92
Morto (1)		0.64	0.23	Morto (1)		0.63	0.23	Morto (1)		0.65	0.23

Tabela 8: Resultados das predições com o algoritmo *Random Forest* utilizando o *Hold Out*. Cada um dos 6 modelos criados possuem um peso diferente para a classe morto. (Fonte: Elaboração Própria)

Já no caso dos modelos preditivos criados a partir do algoritmo *Logistic Regression*, podemos observar, na Tabela 9, que o resultado é muito similar ao caso dos modelos com algoritmo *XGBoost*. Quando atribuído um peso de 1, a acurácia do modelo é de aproximadamente 0.99, o valor da precisão e cobertura da classe Vivo são 0.99 e 0.99 e o valor da precisão e cobertura da classe Morto são 0.61 e 0.24. Já com o peso 10 a acurácia continua a mesma, assim como a precisão e a cobertura da classe Vivo, porém a precisão e a cobertura da classe Morto correspondem a 0.33 e 0.68. Quando o peso atribuído é 100, a acurácia do modelo é aproximadamente 0.96, sendo que a cobertura da classe Vivo diminui para 0.96 e a precisão e a cobertura da classe Morto é de 0.12 e 0.68. Com o peso 200, a acurácia diminui para 0.94, assim como a cobertura da classe Vivo e a precisão da classe Morto, 0.08, porém a cobertura da classe Morto aumenta para 0.91. Ao atribuir um peso de 400 na classe Morto, a acurácia, mais uma vez, diminui e tem um valor de cerca de 0.91, assim como a cobertura da classe Vivo, para 0.91 e há uma alteração na precisão e cobertura da

classe Morto, com valores de 0.06 e 0.93. Enfim, com o peso 1000, a acurácia é de 0.82, a cobertura da classe Vivo é de 0.83 e a precisão e cobertura da classe Morto é de 0.03 e 0.95. Enfim, conforme o valor do peso foi aumentando, tanto a acurácia do modelo quanto a precisão da classe Morto diminuíram enquanto a cobertura da classe Morto aumentou.

Peso 1				Peso 10				Peso 100			
		Previsto				Previsto				Previsto	
		Vivo (0)	Morto (1)			Vivo (0)	Morto (1)			Vivo (0)	Morto (1)
Atual	Vivo (0)	427974	368	Atual	Vivo (0)	425063	3279	Atual	Vivo (0)	412081	16261
	Morto (1)	1827	582		Morto (1)	760	1649		Morto (1)	282	2127
Acurácia: 0.9949042486262365				Acurácia: 0.9906233531669109				Acurácia: 0.961594981787622			
		Precisão	Cobertura			Precisão	Cobertura			Precisão	Cobertura
		Vivo (0)	0.99			Vivo (0)	0.99			Vivo (0)	0.99
		Morto (1)	0.61			Morto (1)	0.33			Morto (1)	0.12
			0.24				0.68				0.88
Peso 200				Peso 400				Peso 1000			
		Previsto				Previsto				Previsto	
		Vivo (0)	Morto (1)			Vivo (0)	Morto (1)			Vivo (0)	Morto (1)
Atual	Vivo (0)	403319	25023	Atual	Vivo (0)	390392	37950	Atual	Vivo (0)	355204	73138
	Morto (1)	213	2196		Morto (1)	171	2238		Morto (1)	120	2289
Acurácia: 0.9414139491260612				Acurácia: 0.9115010760276819				Acurácia: 0.829929588091496			
		Precisão	Cobertura			Precisão	Cobertura			Precisão	Cobertura
		Vivo (0)	0.99			Vivo (0)	0.99			Vivo (0)	0.99
		Morto (1)	0.08			Morto (1)	0.06			Morto (1)	0.03
			0.91				0.93				0.83
											0.95

Tabela 9: Resultados das predições com o algoritmo *Logistic Regression* utilizando o *Hold Out*. Cada um dos 6 modelos criados possuem um peso diferente para a classe morto. (Fonte: Elaboração Própria)

Ao analisar tais modelos preditivos, podemos constatar que o valor da cobertura da classe Vivo, na maioria das vezes, diminuiu enquanto o valor da cobertura da classe Morto aumentou. Além disso, podemos constatar que os modelos preditivos criados a partir do algoritmo *Random Forest* não tiveram resultados significantes quando atribuídos pesos à classe Morto. Enfim, o uso do peso para dar importância a assertividade da classe Morto, fez com que a cobertura também aumentasse e isso é muito importante para o problema levantado neste projeto. Uma maior cobertura da classe Morto significa que o modelo acerta muitos casos de óbito, mas em contrapartida tem uma baixa precisão. Mesmo assim, se o modelo tiver o resultado de uma predição como

Morto e a classe real for Vivo, poderia apenas ocorrer a alocação de recurso para acompanhamento à gestante e o bebê mesmo sem risco algum.

7.2 Interpretações dos modelos preditivos

A partir de um modelo preditivo que utiliza o algoritmo *XGboost*, o gráfico da Figura 14 apresenta as variáveis que mais influenciaram os resultados da predição. É possível concluir que as características mais importantes são: "n_nu_peso", que armazena o peso do recém-nascido, "n_st_malformacao", que indica se houve ou não má formação do feto, "n_nu_semana_gestacao", que apresenta qual semana da gestação o bebê nasceu, "n_nu_apgar1", que contém a nota apgar do bebê no primeiro minuto, "n_nu_apgar5", que contém a nota apgar do bebê no quinto minuto entre outros. As variáveis presentes neste gráfico estão descritas na seção 6.

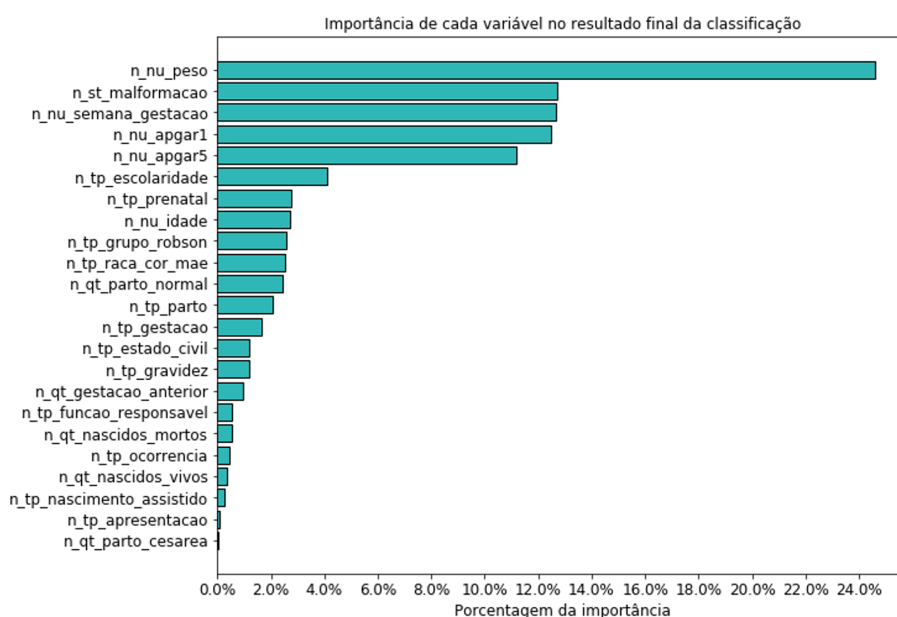


Figura 14: Importância das variáveis na predição. (Fonte: Elaboração Própria)

8 CONCLUSÃO

O principal objetivo deste projeto foi realizar uma análise exploratória dos dados da mortalidade neonatal da cidade de São Paulo e criar modelos de aprendizado de máquina para prever o risco de morte durante o período neonatal. A partir dos resultados dos modelos preditivos é possível concluir que não podemos avaliar o desempenho preditivo dos modelos apenas pela AUC ou acurácia, pois a base de dados é altamente desbalanceada. Sendo assim, temos também que dar mais atenção às métricas precisão e cobertura. Outra questão é o objetivo da criação dos modelos preditivos, que foram criados com a finalidade de prever se o recém-nascido virá a óbito ou não. Dessa forma, os modelos nos quais as predições apresentaram uma cobertura maior na classe Morto devem ser considerados como os mais adequados, uma vez que uma maior cobertura desta classe indica que estes modelos acertaram muitos casos de óbitos.

Em relação aos tipos de algoritmos utilizados na criação dos modelos preditivos e seus desempenhos, os criados a partir dos algoritmos *XGBoost* e *Logistic Regression* foram os que apresentaram melhor desempenho preditivo, quando atribuído um peso de valor 1000 na classe Morto, uma vez que de 2409 ocorrências de óbito, as predições erradas contabilizaram 300 e 120, ou seja, 0.88 e 0.95 de cobertura, respectivamente.

Os resultados das predições citados, nos quais a cobertura da classe Morto é maior, podem ser utilizados como base para a alocação de recursos relacionados a saúde, com o objetivo de diminuir os casos de óbito durante o período neonatal. Além disso, é possível afirmar que o peso do recém-nascido foi um dos fatores que mais influenciaram os resultados das predições, assim como a presença de malformação, sendo estas causas que podem ser evitadas se houver acompanhamento médico e a qualidade deste serviço.

REFERÊNCIAS

- AGUIAR, W. S. Desenvolvimento de modelos preditivos de mortalidade infantil com base em inteligência artificial no estado do ceará. Faculdade de Medicina, Universidade Federal do Ceará, 2019. 15, 26
- 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. 15
- BELUZO, C. E. et al. 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. *Informatics in Medicine Unlocked*, v. 20, 2020. 29
- BELUZO, C. E. et al. Towards neonatal mortality risk classification: A data-driven approach using neonatal, maternal, and social factors. *Informatics in Medicine Unlocked*, Elsevier, v. 20, 2020. , 15, 27, 33, 34, 35
- CHEN, J. H.; ASCH, S. M. Machine learning and prediction in medicine — beyond the peak of inflated expectations. *The New England Journal of Medicine*, The New England Journal of Medicine, v. 376, n. 26, p. 2507–2509, 2017. 20
- CROWN, W. H. Potential application of machine learning in health outcomes research and some statistical cautions. *Value in Health*, Elsevier Inc, v. 18, p. 137–140, 2015. 22
- DEMITTO, M. d. O. et al. Gestação de alto risco e fatores associados ao óbito neonatal. *Revista da Escola de Enfermagem*, Universidade de São Paulo, v. 51, 2017. 15, 26
- FLACH, P. A. The geometry of roc space: Understanding machine learning metrics through roc isometrics. International Conference on Machine Learning(ICML), 2003. 24
- FONTAINE, A. *Mastering Predictive Analytics with scikit-learn and TensorFlow: Implement machine learning techniques to build advanced predictive models using Python*. [S.l.]: Packt Publishing Ltd, 2018. ISBN 978-1-78961-774-0. 22
- FRANÇA, E.; LANSKY, S. Mortalidade infantil neonatal no Brasil: situação, tendências e perspectivas. 2008. 19
- FRIAS, P. G. d. et al. Utilização das informações vitais para a estimação de indicadores de mortalidade no brasil: da busca ativa de eventos ao desenvolvimento de métodos. *Cadernos de Saúde Pública*, Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, v. 33, n. 3, p. 1–13, 2017. 15
- GAIVA, M. A. M.; FUJIMORI, E.; SATO, A. P. S. Fatores de risco maternos e infantis associados à mortalidade neonatal. *Texto & Contexto Enfermagem*, Universidade Federal de Santa Catarina, Programa de Pós Graduação em Enfermagem, v. 25, n. 4, p. 1–9, 2016. 19

GERON, A. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2. ed. [S.l.]: O'Reilly Media Inc, 2019. ISBN 978-1-492-03264-9. 20, 21, 37

GUIMARÃES, R. et al. Política de ciência, tecnologia e inovação em saúde. *Ciência & Saúde Coletiva*, ABRASCO - Associação Brasileira de Saúde Coletiva, v. 24, n. 3, p. 881–886, 2019. ISSN 1413-8123. 17

KASSAR, S. B. et al. Fatores de risco para mortalidade neonatal, com especial atenção aos fatores assistenciais relacionados com os cuidados durante o período pré-natal, parto e história reprodutiva materna. *Jornal de Pediatria*, Sociedade Brasileira de Pediatria, v. 89, n. 3, p. 269–277, 2013. 19

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30*, Neural Information Processing Systems Conference, p. 4768–4777, 2017. 37

MUKHIYA, S. K.; AHMED, U. *Hands-On Exploratory Data Analysis with Python: Perform EDA Techniques to Understand, Summarize, and Investigate Your Data*. [S.l.]: Packt Publishing Ltd, 2020. ISBN 978-1-78953-725-3. 22, 32

MULLER, A. C.; GUIDO, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. [S.l.]: O'Reilly Media Inc, 2016. ISBN 978-1-449-36941-5. 20, 21, 36

OBERMEYER, Z.; EMANUEL, E. J. Predicting the future — big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, The New England Journal of Medicine, v. 375, n. 13, p. 1216–1219, 2016. 26

RAJKOMAR, A.; DEAN, J.; KOHANE, I. Machine learning in medicine. *The New England Journal of Medicine*, Massachusetts Medical Society, v. 380, n. 14, p. 1347–1358, 2019. 15

RATHI, S. Generating counterfactual and contrastive explanations using shap. 2019. 37

SHUKLA, N. *Machine Learning with TensorFlow*. [S.l.]: Manning Publications, 2018. ISBN 978-1-61729-387-0. 23

SILVA, A. L. A. d. et al. Assistência ao parto no brasil: uma situação crítica ainda não superada 1999- 2013. *Revista Brasileira de Saúde Materno Infantil*, Instituto de Medicina Integral Prof. Fernando Figueira, v. 16, n. 2, p. 139–148, 2016. 15

SONG, X. et al. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. *Studies in health technology and informatics*, IOS Press, v. 107, n. 1, p. 736–740, 2004. 26

UNIGME, U. N. I.-a. G. f. C. M. E. Levels & trends in child mortality: Report 2019, estimates developed by the united nations inter-agency group for child mortality estimation. United Nations Children's Fund, 2019. 17

WHO, W. H. O. *Women and health: today's evidence, tomorrow's agenda*. [S.l.]: UN World Health Organization, 2009. ISBN 9789241563857. 17

WHO, W. H. O. World health statistics 2019: monitoring health for the sdgs, sustainable development goals. Geneva World Health Organization, p. 16–22, 2019. 19, 26