Universidade Federal de Goiás
Instituto de Informática
Sistemas de Informação

Pedro Gustavo D.R. de Andrade

Criando um Modelo de Classificação de Óbito para Sars-CoV-2

Goiânia
2021

Universidade Federal de Goiás
Instituto de Informática
Sistemas de Informação

Autorização para Publicação de Trabalho de Conclusão de Curso em Formato Eletrônico

Na qualidade de titular dos direitos de autor, **AUTORIZO** o Instituto de Informática da Universidade Federal de Goiás – UFG a reproduzir, inclusive em outro formato ou mídia e através de armazenamento permanente ou temporário, bem como a publicar na rede mundial de computadores (*Internet*) e na biblioteca virtual da UFG, entendendo-se os termos "reproduzir" e "publicar" conforme definições dos incisos VI e I, respectivamente, do artigo 5º da Lei nº 9610/98 de 10/02/1998, a obra abaixo especificada, sem que me seja devido pagamento a título de direitos autorais, desde que a reprodução e/ou publicação tenham a finalidade exclusiva de uso por quem a consulta, e a título de divulgação da produção acadêmica gerada pela Universidade, a partir desta data.

Título: Criando um Modelo de Classificação de Óbito para Sars-CoV-2

Autor(a): Pedro Gustavo D.R. de Andrade

Goiânia, 06 de Outubro de 2021.

Pedro Gustavo D.R. de Andrade – Autor

Nádia Félix F. da Silva – Orientador

Pedro Gustavo D.R. de Andrade

Criando um Modelo de Classificação de Óbito para Sars-CoV-2

Trabalho de Conclusão apresentado à Coordenação do Curso de Sistemas de Informação do Instituto de Informática da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Área de concentração: Aprendizado de Máquina.

Orientador: Prof. Nádia Félix F. da Silva

Goiânia
2021

Pedro Gustavo D.R. de Andrade

Iniciou a graduação na Universidade Federal de Goiás no segundo semestre de 2012 no curso de Sistemas de Informação. Durante o andamento do curso foi estagiário da Escola de Agronomia - UFG por dois ano (2014-2016) atuando na parte de suporte técnico de toda infraestrutura dos centros de aula e salas de professores.

Dedico este trabalho a minha esposa Fernanda, meus pais e amigos.

Agradecimentos

Primeiramente agradeço aos meus pais, por sempre me apoiarem durante essa trajetória e estarem presentes me fortalecendo nos momentos bons e ruins. Tenho muito orgulho de quem são e do que representam para mim.

Agradeço a minha esposa Fernanda Martins da Paixão, por me incentivar sempre que estava desmotivado e desencorajado, você é muito importante para mim e agradeço por estar sempre ao meu lado.

Agradeço aos meus amigos em especial ao Lucas Lima e Akemy Nogueira, por fazerem parte ativamente das minhas conquistas desde o início dessa jornada.

Agradeço ao Instituto de Informática(INF) e aos professores da Universidade Federal de Goiás - UFG, por me oferecerem todo o suporte em ensino, pesquisa, educação, infraestrutura e amadurecimento pessoal.

Em especial, agradeço à minha orientadora Nádia Félix pela ótima pessoa que é. Agradeço por ser quem despertou vontade de estudar cada vez mais a área que escolhi dentro da universidade, sua passagem na minha vida me fez criar amor e carinho pela informática.

A emoção mais antiga e mais forte da humanidade é o medo, e o mais antigo e mais forte de todos os medos é o medo do desconhecido.

H.P. Lovecraft,
.

Resumo

D.R. de Andrade, Pedro Gustavo. **Criando um Modelo de Classificação de Óbito para Sars-CoV-2**. Goiânia, 2021. 40p. Relatório de Graduação. Sistemas de Informação, Instituto de Informática, Universidade Federal de Goiás.

O Sars-CoV-2 é um vírus da família dos coronavirus, vírus responsável por causar doenças respiratórias, em 2020 esse vírus foi classificado como pandêmico e em pouco tempo levando a morte de milhares de pessoas. A criação de um modelo de classificação pode auxiliar a identificar quais os principais fatores que levam ao óbito de pessoas infectadas pelo Sars-Cov-2, esses fatores identificados podem levar a diferentes abordagens durante o tratamento da doença.
O uso de modelos de aprendizado de máquina tem crescido cada vez mais devido ao auxílio que traz nas diversas áreas em que é aplicado. É comum o uso de técnicas de validação que ocasiona uma interpretação equivocada do seu desempenho. Esse trabalho tem como objetivo desenvolver um modelo de classificação utilizando o classificador Random Forest seguindo todos os passos necessários tais como, pre-processamento, visualização, normalização e balanceamento dos dados, utilizando diferentes abordagens de balanceamento exemplo, Random Oversampling, Random Undersampling e SMOTEENN, identificando as dificuldades no desenvolvimento.

Palavras–chave

Aprendizado de Máquina.

Abstract

D.R. de Andrade, Pedro Gustavo. **Creating a classification model of death for Sars-CoV-2**. Goiânia, 2021. 40p. Relatório de Graduação. Sistemas de Informação, Instituto de Informática, Universidade Federal de Goiás.

Sars-CoV-2 is a virus of the coronavirus family, virus responsible for causing respiratory diseases, in 2020 this virus was classified as pandemic and in a short time leading to the death of thousands of people. The creation of a model of classification can help identify what are the main factors that can lead to the death of people infected by the virus Sars-CoV-2, these factors identified can help to new approaches during the disease treatment.

The use of machine learning has grown constantly due to its efficiency in areas which it is applied. It is common to use validation techniques that lead to a misinterpretation of its performance. This study aims to develop a classification model using the Random Forest classifier following all the necessary steps such as, Data pre-processing, data visualization, data normalization and data balancing, using different balancing approaches such as, Random oversampling, Random Undersampling and SMOTEENN, and identifying the difficulties in development.

Keywords

Machine Learning.

Conteúdo

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

Em Dezembro de 2019, a China comunicava ao mundo o surgimento do *SARS-CoV-2*, um vírus de acometimento respiratório e de alta infectividade, conhecido popularmente como COVID19 (HASOKSUZ; KILIç; SARAC, 2020). É transmitido pelo ar, principalmente por gotículas de saliva que se espalham por meio de tosse e espirro de portadores da doença, e por contato com superfícies contaminadas. No ano de 2020 A Organização Mundial da Saúde (OMS) declarou o vírus como pandêmico, e no Brasil neste mesmo ano causou a morte de mais de 194 mil pessoas (ORGANIZATION, 2021).

O COVID19 pertence a família dos coronavirus (HASOKSUZ; KILIç; SARAC, 2020) que já são conhecidos por causarem doenças respiratórias em humanos com raros casos de infecção severa, diferente do SARS-CoV-2 que tem uma taxa de mortalidade que varia de 2 a 4% além de não possuir um tratamento específico.

Modelos de Classificação são usados frequentemente e tem como objetivo predizer resultados através de dados coletados, essa coleção de dados possui ou não relacionamento direto com o que se quer prever, geralmente em classificações binárias prever se ocorre ou se não ocorre determinado acontecimento. Podem ser usados na área médica (RESMINI AURA CONCI, 2012), identificação de fraudes (VARMEDJA et al., 2019), corridas eleitorais (TSAI et al., 2019) e identificação de perfis de clientes (TALóN-BALLESTERO et al., 2018).

Geralmente a coleta dos dados é feita através de pesquisas, preenchimento de formulários, imagens, áudios e vídeos(MARRELLI, 2007). Para este trabalho, os dados foram recolhidos pelos pontos de atendimento no momento do cadastro de pessoas suspeitas de estarem infectadas pelo vírus do COVID19, e disponibilizados no site openDataSUS (SAUDE, 2021) no formato *comma separated values* (CSV), acompanhado de um dicionário dos dados, identificando cada campo e os possíveis valores.

Desenvolver modelos de previsão demanda um grande esforço dos envolvidos, por se tratar de um processo longo e trabalhoso, na grande maioria das vezes os

dados obtidos são incompletos, não estruturados e podem não possuir uma descrição clara do seu significado, ficando na responsabilidade dos envolvidos identificar e resolver os problemas e em alguns casos, se constatado, descartando os dados se forem de má qualidade.

Outro detalhe importante que se deve atentar é a distribuição dos valores que se deseja prever, é necessário que as classes estejam distribuídas de forma igual ou o mais próximo da igualdade possível. Em casos onde as classes não possuam uma distribuição semelhante, havendo maior ocorrência de um determinado fenômeno em relação ao outro, o nosso modelo pode sofrer uma influência da classe majoritária ocasionando má performance na predição da classe minoritária.

Para avaliar o desempenho de um modelo de aprendizado utilizamos as métricas obtidas como, acurácia, precisão, *recall* e média geométrica, essas métricas devem ser usadas com cuidado, pois em modelos desbalanceados o uso de uma métrica incoerente pode passar uma falsa percepção de bom desempenho (GU; ZHU; CAI, 2009). Em um modelo onde 90% dos casos uma pessoa não vem ao óbito se usarmos somente a acurácia o modelo pode indicar assertividade 90% das vezes, porém, o modelo pode estar errando todas as predições para casos de óbito.

1.1 Motivação e Problemática

Uma das maneiras de se combater doenças é conhecer quais são os principais grupos de risco, os seus estágios e qualquer fator que influencie no seu agravamento. Através desses dados, governos e profissionais da saúde são capazes de tomarem decisões com maior exatidão, criando medidas sociais de controle e elaborando um plano de tratamento eficiente para pacientes de uma determinada doença.

Apesar de serem conhecidas as formas como o COVID19 é transmitido, sabe-se muito pouco a respeito de medidas eficientes de tratamento da doença em pessoas infectadas. Tendo em vista as grandes incertezas a respeito desse vírus, nesse trabalho foi proposto o desenvolvimento de um modelo de Aprendizado de Máquina que tenha como objetivo classificar o óbito ou não óbito para o Sars-CoV2. Dessa forma identificando quais são os fatores que influênciam nos possíveis resultados.

Para desenvolver esse trabalho, foi utilizada a coleção de dados disponibilizada no site openDataSuS (SAUDE, 2021), essa coleção possui mais de um milhão de instâncias, cada instancia é referente a uma pessoa diferente e contem informações relacionadas a pessoa e ao seu estado de saúde.

Esse trabalho não tem o objetivo de desenvolver planos técnicos para o tratamento de pessoas infectadas pelo vírus do COVID19, mas sim desenvolver um

modelo de Aprendizado de Máquina do começo ao fim e tentar descobrir quais são os principais fatores que levam uma pessoa infectada pelo vírus a óbito.

1.2 Objetivo do Trabalho

O uso de modelos de Aprendizado de Máquina tem se tornado cada vez mais comum em diversas áreas diferentes. As instituições utilizam esses modelos para gerarem algum tipo de valor ao seu negócio, seja encontrando soluções para determinados problemas, prevendo determinados riscos ao negócio e identificando fatores que influenciam na tomada de decisões.

O processo de criação de um modelo de Aprendizado de Máquina as vezes é tratado como um processo simples e que sempre irá se comportar da mesma maneira, é comum essa percepção devido ao uso de coleções de dados criadas para o uso didático em que os dados presentes foram coletados de forma controlada ou até mesmo criados sinteticamente. Em aplicações reais, esses dados podem vir de diferentes fontes ou coletados por diferentes pessoas o que pode levar a dados inconsistente, faltantes e com valores errados.

Mencionado os desafios, o foco desse trabalho é desenvolver um modelo de aprendizado de máquina (ML, do inglês *Machine Learning*) para classificar e prever casos de óbito a partir dos dados coletados no openDataSUS, seguindo os passos que devem ser seguidos para garantir que o seu modelo seja de alta confiabilidade e então validar as técnicas utilizadas a partir dos seus resultados, comparando as métricas obtidas em cada abordagem.

1.3 Organização do Trabalho

Este artigo está organizado da forma seguinte:

- O Capítulo 2 apresenta a fundamentação teórica, que visa elucidar alguns conceitos importantes que serão a base para o desenvolvimento deste trabalho.
- O Capítulo 3 apresenta trabalhos relacionados, que tem como objetivo descrever trabalhos de outros autores que vem ao encontro com o exposto por este trabalho;
- O Capítulo 4 apresenta a metodologia e implementação;
- O Capítulo 5 apresenta o estudo de caso e resultados obtidos, que explica as ferramentas que foram utilizadas para realizar a predição e explicação dos resultados obtidos em cada método de predição;
- O Capítulo 6 apresenta as conclusão e trabalhos futuros, levando em conta todo o exposto neste trabalho;

CAPÍTULO 2

Fundamentação teórica

O processo de desenvolvimento de um sistema de aprendizado de máquina é usado frequentemente por pesquisadores e empresas com o objetivo de se criar um modelo para auxiliar no entendimento dos dados coletados, tomada de decisões e prevenção de fraudes. Para garantir que o modelo tenha um alto nível de confiança é necessário que todas as etapas da criação do modelo de ML sejam executados seguindo uma metodologia. (LONES, 2021) apresenta em seu trabalho uma série de abordagens que devem ser consideradas no desenvolvimento de um modelo de ML para assegurar que os resultados obtidos sejam de qualidade.

2.1 Aprendizado de Máquina

Aprendizado de Máquina é o processo onde um computador através do uso de algoritmos alcança um determinado objetivo, para atingir esse objetivo ele utiliza dados como entrada e então produz um determinado resultado de saída. Utilizando amostras de dado, chamadas de amostras de treino, o algoritmo através da repetição se torna melhor em atingir o objetivo desejado, essa etapa é chamada de etapa de treino (NAQA; MURPHY, 2015) . Após o fim dessa etapa, o algoritmo está configurado para prever resultados a partir do uso de novos dados. O Aprendizado de Máquina pode ser Supervisionado e Não supervisionado.

2.2 Aprendizado Supervisionado

Aprendizado de Máquina supervisionado é a capacidade do computador realizar previsões a partir do uso de dados rotulados, criando um mapeamento desses dados na fase de treino e então aplicando esse mapeamento para gerar previsões para dados antes não vistos (CUNNINGHAM; CORD; DELANY, 2008). A criação de um Modelo de Aprendizado de Máquina supervisonado é dividido em algumas etapas e ele pode ser um modelo de regressão ou classificação:

- Primeiro os dados são obtidos através de alguma fonte ja disponível ou coletadas;
- Esses dados devem ser tratados para que somente dados de qualidade sejam utilizados para a criação do modelo;
- Pelo uso dos dados tratados será realizado o treinamento do modelo de Aprendizado de Máquina;
- A eficiência do modelo deve ser testada a partir do uso de métricas;
- Aplicação do modelo em dados novos.

2.2.1 Regressão

Um modelo de regressão tem o objetivo de prever como resultado um valor contínuo, um valor contínuo pode ser o preço de uma casa, o valor de um determinado produto, o tempo necessário pra uma determinada ação, ou a altura ou o peso. Para esse tipo de problema, os dados devem ser rotulados e geralmente são diferentes um do outro e a partir desses dados conhecidos é realizada a previsão de novos valores pela entrada de novos dados.

2.2.2 Classificação

Modelos de classificação também fazem uso de dados rotulados, porém, diferente dos modelos de regressão, um modelo de classificação tem o objetivo de identificar a qual grupo ou classe uma determinada sequencia de dados pertence, encontrar um valor não-contínuo. Podemos citar valores não-contínuos como grupos de idade, nível de satisfação, nota para filmes e produtos.

2.3 Árvores de Decisão

O algoritmo de Árvores de Decisão é um algoritmo muito utilizado para criar tanto modelos de classificação quanto regressão. Uma Árvore de decisão é composta por folhas(chamados também de nós), em cada nó é realizada uma pergunta a um determinado atributo do conjunto de dados, os caminhos(arestas) é o caminho percorrido a partir da resposta obtida em um nó podendo ser verdadeiro ou falso, uma árvore de decisão inicia na sua raiz e percorre até o nó final respondendo as perguntas encontradas em cada nó anterior, o nó final contém a resposta do rótulo dos dados (QUINLAN, 1986).

A figura 2.1 mostra um exemplo de uma árvore de decisão relacionada a coleção de dados referente a indentificação de óbito para o COVID19. Na raiz da árvore é perguntado sobre a existencia de suport de ventilador respiratório, se for

menor ou igual a 1.5, o caminho escolhido será o da esquerda, caso contrário o caminho percorrido será o da direita, o novo nó terá uma nova regra de decisão e dessa forma percorrerá até o nó final correspondente a uma classificação.

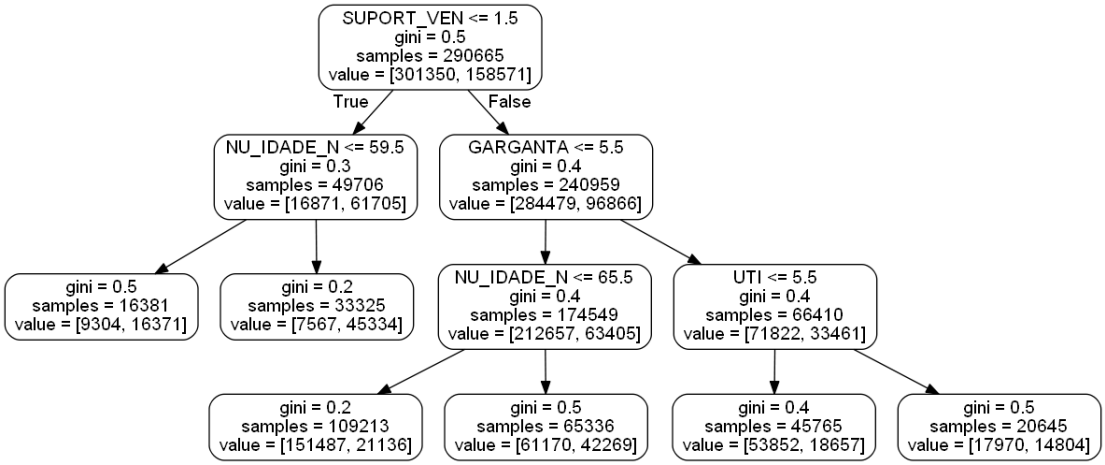


Figura 2.1: *Árvore de decisão.*

2.4 Floresta Aleatória

Florestas aleatórias fazem o uso de diversas árvores de decisão para criar um modelo, ao invés de um único modelo ser treinado, em uma floresta de decisão é escolhido a quantidade de árvores que serão criadas e cada árvore é criada a partir de uma amostra aleatória com reposição de toda a coleção de dados usada para treino. Uma amostra com reposição é um tipo de amostragem onde o elemento sorteado retorna para a coleção de dados para que possa ser sorteado novamente. Esse tipo de modelo é chamado de conjunto de modelos ou ensemble, a decisão tomada para a previsão do rótulo dos dados é decidida por uma votação do resultado de todas as árvores criadas (BREIMAN, 2001).

Vale ressaltar que só é possível criar um conjunto de modelos utilizando diferentes algoritmos de decisão ou gerando várias amostras diferentes dos dados a partir da coleção de dados original, que é a forma como uma Floresta Aleatória funciona. A Figura 2.2 mostra como funciona um modelo de floresta aleatória.

2.5 Aprendizado Não Supervisionado

Modelos de Aprendizado Não Supervisionado não necessariamente precisam de dados rotulados, diferente de modelos supervisionados onde o objetivo é criar um modelo capaz de generalizar previsões para novos dados, modelos não-supervisionados tem o objetivo de descobrir padrões existentes entre os dados for-

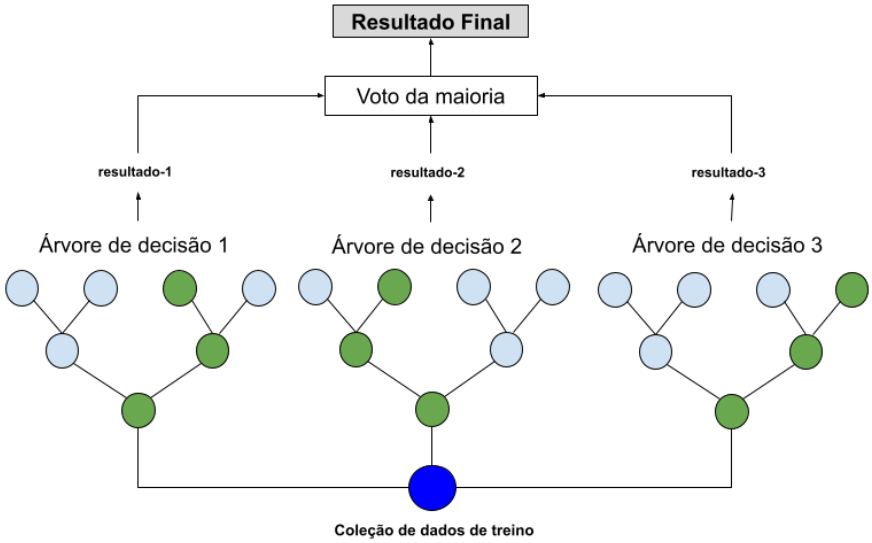


Figura 2.2: *Floresta aleatória.*

necidos a ele, esses padrões na grande maioria das vezes são usados para identificar grupos, que podem ser usados para a tomada de decisões, identificar fraudes, gerar novos dados para serem usados em Modelos Supervisionados ou definir sistemas de recomendação.

A figura 2.3 mostra a identificação de grupos de clientes de um mercado, utilizado os dados do ganho anual e o score de gasto de cada cliente naquele mercado durante o ano.

2.6 Coleta dos Dados

A obtenção dos dados é a etapa do desenvolvimento de um modelo de ML onde os dados etiquetados serão fornecidos para a análise e possível treinamento de um modelo, os dados podem ser classificados em dados primários ou secundários. Dados primários são obtidos para um problema específico, modelado de acordo para o problema, os dados secundários são aqueles que após obtidos para um determinado problema são disponibilizados para a comunidade em geral e poderão ser usados para outras finalidades (HOX; BOEIJE, 2005).

Os dados coletados podem ser classificados como ativos, onde são obtidos através de pesquisas, entrevistas, formulários preenchidos no momento de algum

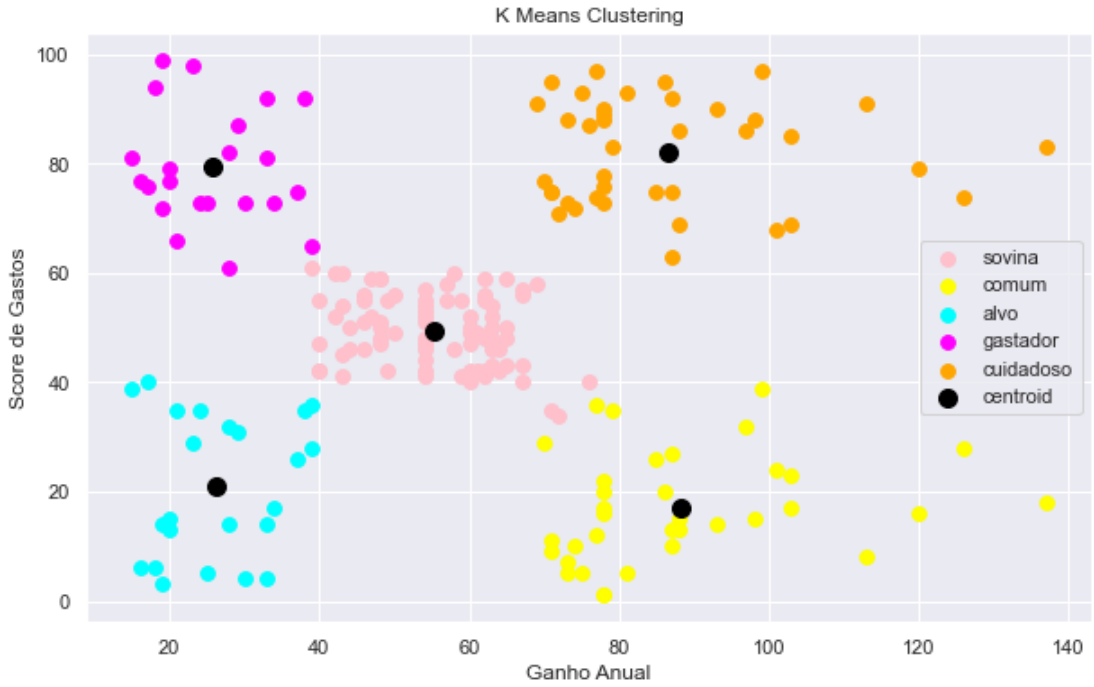


Figura 2.3: *Grupos de clientes de um mercado.*

cadastro, ou seja, interação direta com outras pessoas e passivos, aqueles obtidos sem participação direta de uma pessoa, dados de geolocalização, sensores ou metadados (MAHER et al., 2019).

2.7 Pré-Processamento dos Dados

Na grande maioria das vezes os dados coletados precisam ser tratados para que possam ser usados no modelo de ML, é comum os dados colhidos possuírem aspectos que pode influenciar negativamente o processo de aprendizado, essas anomalias podem ser a presença de outliers, instâncias com valores errados ou a ausencia de valores.

Técnicas de tratamento de valores ausentes são necessárias para a criação de modelos de ML quando presentes. A substituição pela média consiste no pre-enchimento de valores ausentes pela média dos valores, outra técnica utilizada é a criação de um valor que representa a ausência dos valores (ACOCK, 2005).

2.7.1 Outliers

Outliers são valores que destoam em relação a grande maioria dos valores de um determinado atributo de uma coleção de dados, podem ser valores muito altos ou muito baixos e também categorias que aconteçam com uma certa raridade (BEN-GAL, 2005).

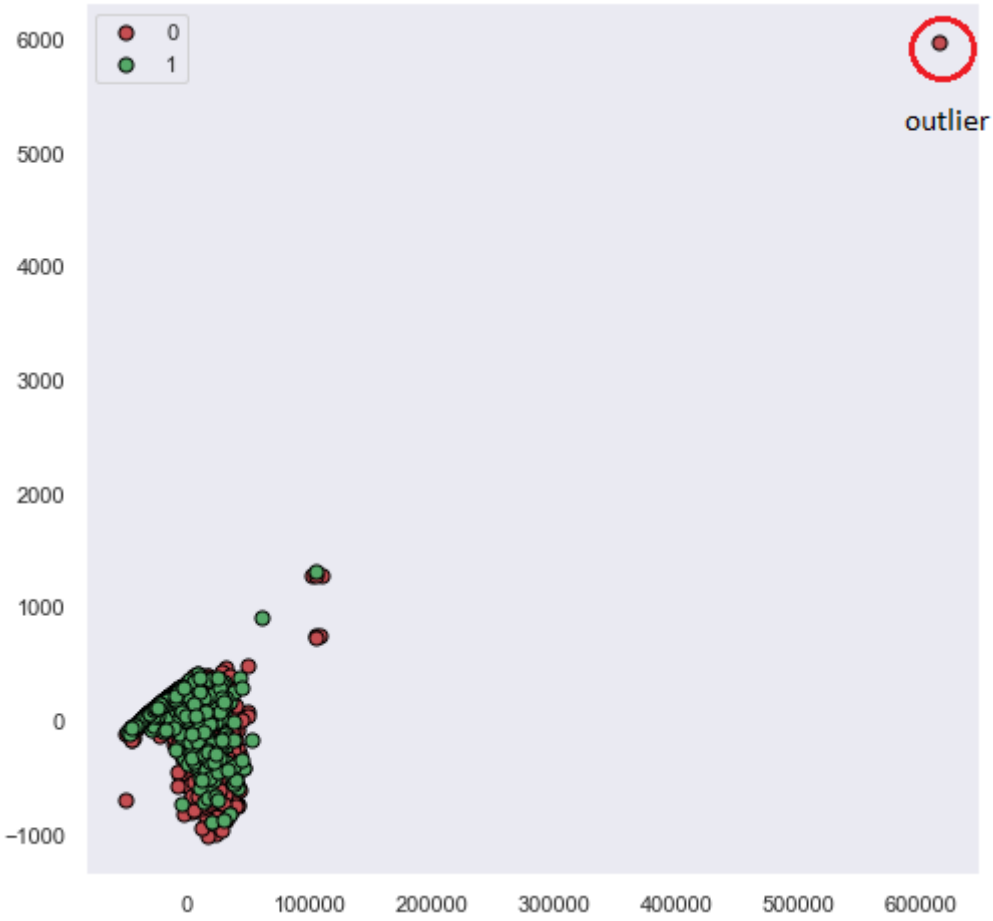


Figura 2.4: *Exemplo de um outlier.*

É comum outliers influênciarem negativamente no aprendizado de um modelo de classificação ou regressão, porém, podem existir casos onde os outliers sejam os dados mais importantes para a proposta de um modelo de previsão, como em casos de detecção de fraude.

Na figura 2.4 mostra um exemplo de um ponto muito diferente dos demais da distribuição, esse ponto é um outlier.

2.8 Técnicas de Balanceamento

Dados desbalanceados é um problema comum onde os rótulos de classe estão distribuídos em proporções diferentes. Vários autores apresentam diferentes abordagens para resolver esse problema, onde o objetivo principal é igualar ou atingir uma proporção próxima entre as classes (LIU; GHOSH; MARTIN, 2007).

Random oversampling consiste na repetição de dados de forma randômica da classe minoritária da Coleção de Dados, até que ambas as classes estejam uniformes ou em uma proporção próxima desejada. Como esse modelo reproduz de forma randômica os dados da classificação minoritária, irá aumentar a probabilidade

de acontecer *over-fitting*[1]. Outro problema levantado é o custo de performance que é ocasionado quando aplicado em uma Coleção de Dados de grande volume (MOHAMMED; RAWASHDEH; ABDULLAH, 2020).

Diferente do *Oversampling* onde a classe minoritária é o alvo, em *Random Undersampling* a classe majoritária é diminuída pela eliminação randômica dos dados até atingir uma proporção simétrica em relação a classe minoritária, como a eliminação dos dados é feita de forma aleatória, informações que podem ter relevância para o modelo de aprendizado podem ser eliminadas, prejudicando o modelo de aprendizado (COHEN et al., 2006).

Synthetic Minority Oversampling e Edited Nearest Neighbours (SMOTE-ENN) é a combinação do *oversampling*, *Synthetic Minority oversampling Technique* (SMOTE) e *Undersampling Edited Nearest Neighbors* (ENN). SMOTE seleciona os k vizinhos próximos pelo método *K-Nearest Neighbors* (KNN), os conecta e gera novas amostras entre esse espaço. ENN faz a limpeza da coleção de dados gerada após o *Oversampling* realizado, eliminando os ruídos[2] (MANJU; NAIR, 2019). A criação de amostras por SMOTE se dá por:

$$X_{new} = X_i + (X_i^{'} - X_i) * a$$

Onde, X_{new} = é o dado sintético, X_i = são exemplos da classe minoritária, $X_i^{'}$ = um dos k vizinhos próximos de X_i, e a = um valor randômico de 0 a 1.

2.9 Métricas

No desenvolvimento dos modelos de classificação, as métricas são utilizadas para validar a capacidade do modelo em acertar as classificações feitas. Em um cenário ideal um modelo poderia ser validado utilizando o valor da acurácia em acertar as predições, porém, é comum ao trabalhar com modelos de classificação encontrar coleções de dados com a distribuição das amostras desbalanceadas, nesses casos o uso da acurácia pode levar a uma interpretação equivocada do desempenho do modelo. Saber interpretar os diferentes tipos de métricas e qual a ideal em cada situação é essencial para validar a performance do modelo (FERRI; MODROIU, 2006)

A Tabela 2.1 demonstra a estrutura de uma matriz de confusão, que contém informações sobre as classificações atuais e preditas (VISA et al., 2011). Verdadeiro

[1] modelo se ajusta muito bem para os dados de treino mas é ineficaz na previsão de novos resultados

[2] dados que podem prejudicar o modelo de ML

Positivo indica os valores que são verdadeiros de fato, Falso Negativo demonstra os valores que foram classificados como falsos mas são verdadeiros, Falso positivo são os valores classificados como positivos porém são falsos e Verdadeiro Negativo os valores classificados como falsos e são realmente falsos.

	Valor Predito	
	Verdadeiro	**Falso**
Verdadeiro Real	Verdadeiro Positivo (TP)	Falso Negativo (FN)
Falso Real	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Tabela 2.1: *Matriz confusão.*

2.9.1 Acurácia

Acurácia é a mais comum das métricas utilizadas, ela mede o grau de precisão das predições corretas do modelo.

$$acc = \frac{TP+TN}{n}$$

$acc =$ acurácia obtida, $(TP+TN) =$ a soma de Verdadeiros Positivo e Verdadeiros Negativo e $n =$ a soma de todos valores da matriz de confusão.

2.9.2 Média Geométrica

Média Geométrica é ideal para medir o grau de eficiência de modelos de classificação de classes desbalanceadas, a média geométrica faz o calculo da acurácia dos valores positivos e negativos do modelo. A métrica tenta maximizar a acurácia em cada uma das classes enquanto mantém o equilíbrio (BARANDELA J.S. SANCHEZ; RANGEL, 2003).

$$G\_mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}$$

$G\_mean =$ média geométrica, $TP =$ verdadeiros positivo, $TP+FN =$ a soma de verdadeiros positivo e falsos negativo, $TN =$ verdadeiros negativo e $TN+FP =$ soma dos verdadeiros negativo e falsos positivos.

2.9.3 Precisão

A precisão mede a exatidão dos valores positivos, dos exemplos positivos (TP e FP) quantos de fato são positivos (MOHAMMED; RAWASHDEH; ABDULLAH, 2020).

$$p = \frac{TP}{TP+FP}$$

$p =$ precisão, $TP =$ verdadeiros positivo e $TP + FP =$ soma dos verdadeiros e falsos positivo.

2.9.4 Recall

Recall também chamado de *Sensitivity* mede a proporção dos positivos verdadeiro pela soma das previsões positivas.

$$r = \frac{TP}{TP + FN}$$

CAPÍTULO 3

Trabalhos Relacionados

Na literatura existem muitos estudos relacionados as técnicas utilizadas para fazer a análise do desempenho dos modelos de classificação, para garantir que um modelo de aprendizado de máquina tenha um bom desempenho é necessário ficar atento com os dados que irão ser utilizados e também seguir metodologias que irão garantir a qualidade do modelo final. Serão apresentados alguns trabalhos que falam dessas metodologias:

(LONES, 2021) escreve um guia a ser seguido para evitar certas armadilhas que resultam em um modelo de aprendizado de máquina de má qualidade. Para o autor, seu trabalho é um guia para pessoas que possuem o objetivo de desenvolver um modelo com fins academicos, dito isso, seu trabalho são para estudantes que irão realizar pesquisas na área do aprendizado de máquina. O seu guia é dividido em 5 etapas.

1. É importante que você entenda bem os dados que você irá trabalhar, tenha certeza que possui uma boa quantidade de dados para que seja possível criar um bom modelo, separe uma porção aleatória dos seus dados e não faça nenhum tipo de exploração nesses dados separados alem de estudar a área na qual irá desenvolver sua pesquisa.
2. Não permitir que as informações dos seus dados de teste, aquela porção aleatória que foi separada, vaze em seus dados de treino, teste diferentes modelos porém sem usar modelos inapropriados para a sua proposta.
3. Use dados de validação no momento de treinamento do seu modelo, geralmente essa etapa é feita automaticamente pelos modelos selecionados e sempre usar os dados de teste apropriados para validação do seu modelo.
4. Fazer suposições a respeito das métricas obtidas pode ser perigoso, dependendo da métrica observada um valor alto nem sempre pode indicar um bom desempenho. É importante fazer multiplas comparações entre as métricas selecionadas e os modelos utilizados.
5. Ser transparente sobre os resultados do seu modelo é fundamental, mesmo que o modelo desenvolvido não possua um bom desempenho não quer dizer que

isso seja sua culpa. Usar várias métricas é o ideal para demonstrar como é a performance do seu modelo de diferentes formas.

(LEEVY TAGHI M. KHOSHGOFTAAR; SELIYA, 2018) realiza uma pesquisa demonstrando as diferentes abordagens utilizadas para o balanceamento de dados em coleções de dados tradicionais como, *Oversampling*, *Undersampling*, *Feature Selection*, *Cost-Sensitive* e *Hybrid/ensemble* e técnicas de balanceamento usadas em *Big Data* por exemplo, *Random Oversampling*, *Random Undersampling* e *Synthetic Minority Oversampling Technique*, comparando seus resultados pelas métricas obtidas. São discutidas não somente as métricas e os modelos de balanceamento, mas também o tempo de treinamento levado para cada abordagem utilizada.

Em seu trabalho ele aborda dois métodos de se trabalhar com classes desbalanceadas, técnicas de balanceamento a nível de dados e técnicas de balanceamento por algorítmos. Métodos de tratamento a nível de dados são os métodos onde o balanceamento é realizado a partir do redimensionamento das classes, seja pela fabricação de dados sintéticos ou pela remoção de dados da coleção. Os modelos a nível de algorítmo tentam resolver problemas de balanceamento atribuindo pesos as classes, dando um maior peso a classe desbalanceada e uma junção de modelos, também chamados de modelos híbridos, combinando o redimensionamento com uso de váriso modelos preditivos com o redimensionamento a nível de dados e atribuindo pesos para cada classe.

(FERRI; MODROIU, 2006) realiza experimentos comparando a performance de diferentes métricas para modelos de classificação, explicando os benefícios e diferenças de cada métrica utilizada, ressaltando qual a métrica correta em cada situação.

Para o autor, devem ser aplicadas diferentes tipos de métricas nos modelos, levando em conta que cada métrica tem um significado diferente, podendo ser mais ideal para determinado modelo e não fazer sentido para outros. As métricas são classificadas em tres famílias, sendo:

1. Métricas baseadas em um limiar e uma compreensão qualitativa dos erros do modelo, métricas dessa família são a acurácia, *F-Score* e *Kappa statistic*.
2. Métricas baseadas em uma comprrensão probabilística do erro, medindo o desvio previsto em relação aos valores reais, como exemplo dessas métricas temos *mean absolut error*, *mean squared error* e *LogLoss*.
3. Por fim temos as métricas baseadas em quão bem os modelos classificam os exemplos, por exemplo *Area Under Curve*.

(HE YANG BAI; LI, 2008) apresenta o *Adaptive Synthetic Sampling Approach for Imbalanced Learning* (ADASYN), uma abordagem desenvolvida para tratar

problemas de desbalanceamento de classes, onde dados sintéticos da classe minoritária são criados usando uma distribuição por pesos para diferentes classes minoritárias de acordo com a dificuldade do aprendizado. Segundo o autor os benefícios de se usar seu modelo é a redução do viés causado pela classe majoritária e pela forma utilizada para a geração dos dados sintéticos.

(BARANDELA J.S. SANCHEZ; RANGEL, 2003) propõem estratégias a serem adotadas para problemas de desbalanceamento de classes, apresenta técnicas de redução da classe majoritária, modelo de Wilson e k-NCN (*nearest centroid neighborhood*) comparando a média geométrica obtida por cada técnica de balanceamento.

CAPÍTULO 4

Metodologia

Este trabalho tem como objetivo principal desenvolver um modelo de Aprendizado de Máquina para classificar se uma pessoa infectada pelo vírus do COVID19 veio a óbito ou não. É importante também para a proposta do trabalho identificar quais são os fatores mais importantes na classificação de óbito de uma pessoa. Os dados utilizados para realizar esse trabalho foram coletados em postos de atendimento para pessoas suspeitas de infecção pelo vírus do COVID19 no ano de 2020 e estão disponibilizados no site do OpenDataSUS, um site que contém dados abertos dos casos de doenças respiratórias identificadas pelo Brasil no decorrer do ano. Apesar do objetivo principal do trabalho ser desenvolver um modelo de classificação, no decorrer dos processos de tratamento e análise dos dados, notou-se o desbalanceamento dos rótulos das instâncias da coleçào de dados, sendo assim, o trabalho também aborda o uso de diferentes técnicas utilizadas para o tratamento de dados desbanceados, validando essas técnicas comparandos as métricas obtidas após a criação do modelo de Aprendizado de Máquina. Para chegar ao objetivo final do trabalho que é de prever o rótulo de novos dados, foram seguidas algumas etapas, descritas como etapas essenciais para garantir um bom desempenho do modelo de Aprendizado de Máquina (LONES, 2021).

As etapas realizadas são:

- 1-Coleta ou obtenção dos dados
- 2-Tratamento ou pré-processamento dos dados
- 3-Divisão dos dados tratados em dados de treino e dados de teste
- 4-Normalização dos dados
- 5-Uso de técnicas de balanceamento nos dados
- 6-Treinamento do modelo escolhido com os dados de treino após aplicada a técnica de balanceamento
- 7-Aplicação de métricas para avaliar o desempenho do modelo para cada técnica de balanceamento utilizada
- 8- Análise de qual a melhor técnica de balanceamento pelas métricas

A figura 4.1 mostra o fluxograma seguido para a realização do trabalho, quando é feita a validação do modelo, voltamos para o passo de técnicas de balanceamento utilizando uma outra abordagem.

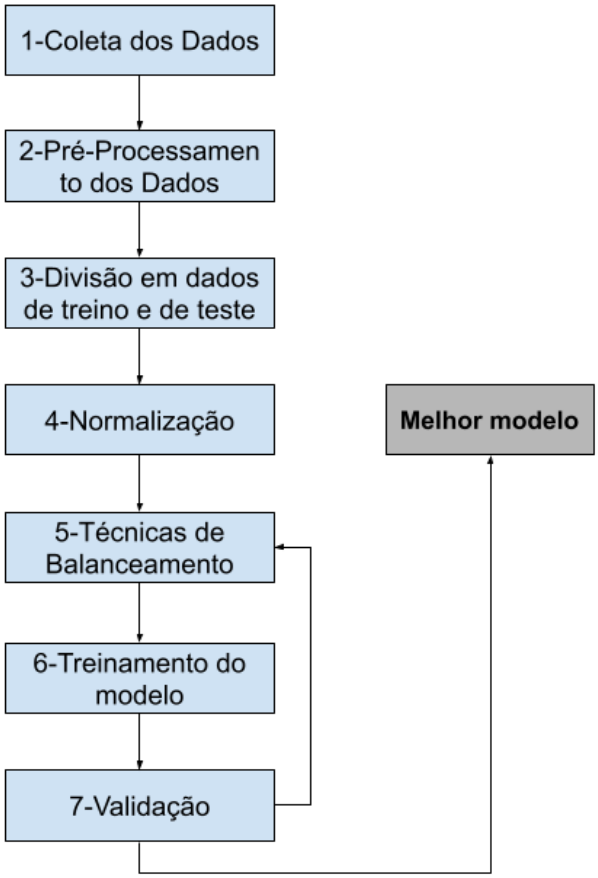


Figura 4.1: *Fluxograma do trabalho.*

Esse trabalho foi feito utilizando a linguagem de programação *Python* na versão 3.7 (ROSSUM; DRAKE et al., 2000). Também foi utilizado a plataforma *Jupyter Notebook* (KLUYVER et al., 2016) para a executar todas as etapas do processo de criação dos modelos. Foram utlizadas as bibliotecas *Pandas* (MCKINNEY, 2012), *Matplotlib* (HUNTER, 2007), *Plotly*, *Scikit-Learn* (PEDREGOSA et al., 2011) e *Imbalanced-Learn* (LEMAÎTRE; NOGUEIRA; ARIDAS, 2017). De toda a coleção de dados, foram usados somente as pessoas diagnosticadas pelo vírus Sars-CoV-2.

CAPÍTULO 5

Experimentos e Resultados Obtidos

5.1 Obtenção dos Dados

Os Dados utilizados para realizar este trabalho foram obtidos diretamente do site OpenDataSUS, o portal contém várias coleções de dados de diversos anos, colhidas pelo Sistema Único de Saúde (SUS) no momento do cadastro de pessoas com possível infecção de alguma doença respiratória. A coleção de dados utilizada neste trabalho é referente aos dados de todo ano de 2020, com a presença dos registros de pessoas infectadas pelo vírus Sars-CoV-2. O conjunto é composto por 154 atributos e 1.196.025 instâncias, em que cada instância se refere a uma pessoa. No portal, os dados podem ser baixados no formato CSV, e junto disponibilizam o dicionário dos dados, contendo as informações de cada atributo, seus possíveis valores e significados.

5.2 Pré-processamento dos Dados

Para fazer o carregamento dos dados, foi utilizado a biblioteca *Pandas*[1] (MCKINNEY, 2008), uma biblioteca gratuita frequentemente usada nas atividades de tratamento dos dados. Após o carregamento dos dados foi iniciado o processo de análise dos dados, a visualização dos valores ausentes e incoerentes, ou seja, valores que não faziam parte daqueles descritos no dicionário de dados.

A primeira etapa foi fazer a eliminação de atributos que continham uma grande quantidade de valores ausentes, a remoção desses atributos foi feita através de um *script* onde selecionava somente os atributos que possuíam 70% ou mais de valores não nulos e descartava os demais. Após essa seleção, restaram ainda 55 atributos, essa coleção de atributos era composta por informações referentes a pessoa, a doença da pessoa, seus sintomas e evolução e os dados geográficos da pessoa.

Como o intuito do trabalho é desenvolver um modelo para classificar óbitos, foi identificado a falta de necessidade de manter atributos relacionados a

[1] biblioteca desenvolvida para a linguagem python

geolocalização da pessoa ou do lugar de atendimento, então foram removidos todos os dados dessa relação, no fim restando 34 atributos, como demonstrado na Figura 5.1. Também foi realizada uma filtragem nas instâncias, o atributo "CLASSI_FIN"contem as informações referentes aos tipos de doenças respiratórias, foram selecionadas as instâncias que possuíam o valor 5, indicando diagnóstico pelo vírus Sars-CoV-2, diminuindo o total de instâncias para 641.705. Todos os atributos restantes após realizado o pré-processamento dos dados são categóricos.

SEM_NOT	SATURACAO
SEM_PRI	DIARREIA
NU_IDADE_N	VOMITO
TP_IDADE	OUTRO_SIN
CS_GESTANT	ANTIVIRAL
CS_RACA	HOSPITAL
CS_ESCOL_N	UTI
CO_PAIS	SUPORT_VEN
CS_ZONA	RAIOX_RES
SURTO_SG	AMOSTRA
NOSOCOMIAL	PCR_RESUL
AVE_SUINO	CRITERIO
FEBRE	HISTO_VGM
TOSSE	EVOLUCAO
GARGANTA	CS_SEXO_F
DISPNEIA	CS_SEXO_M
DESC_RESP	

Figura 5.1: *Atributos restantes após filtragem.*

Para entender melhor os dados finais obtidos, foi feita a plotagem das suas correlações utilizando a biblioteca *Seaborn*[2] (WASKOM, 2012), essa biblioteca possui a função *heatmap* que cria um mapa de calor dos atributos e suas correlações, como demonstrado na Figura 5.2. Essa figura mostra a relação entre todos os dados.

Já a figura 5.3 mostra a correlação dos atributos em relação com o atributo EVOLUCAO, que é o rótulo que indica óbito ou não óbito de uma pessoa com o vírus do COVID19.

Finalizado a seleção dos atributos, foi feito o preenchimento dos valores vazios que ainda existiam, todos os atributos restantes eram compostos por números inteiros que representavam um determinado fenômeno, no próprio dicionário de dados havia um valor, o número 9, que caso o paciente não fornecesse uma informação, deveria ser usado para preencher o formulário. Seguindo as instruções

[2]biblioteca utilizada para a criação de gráficos

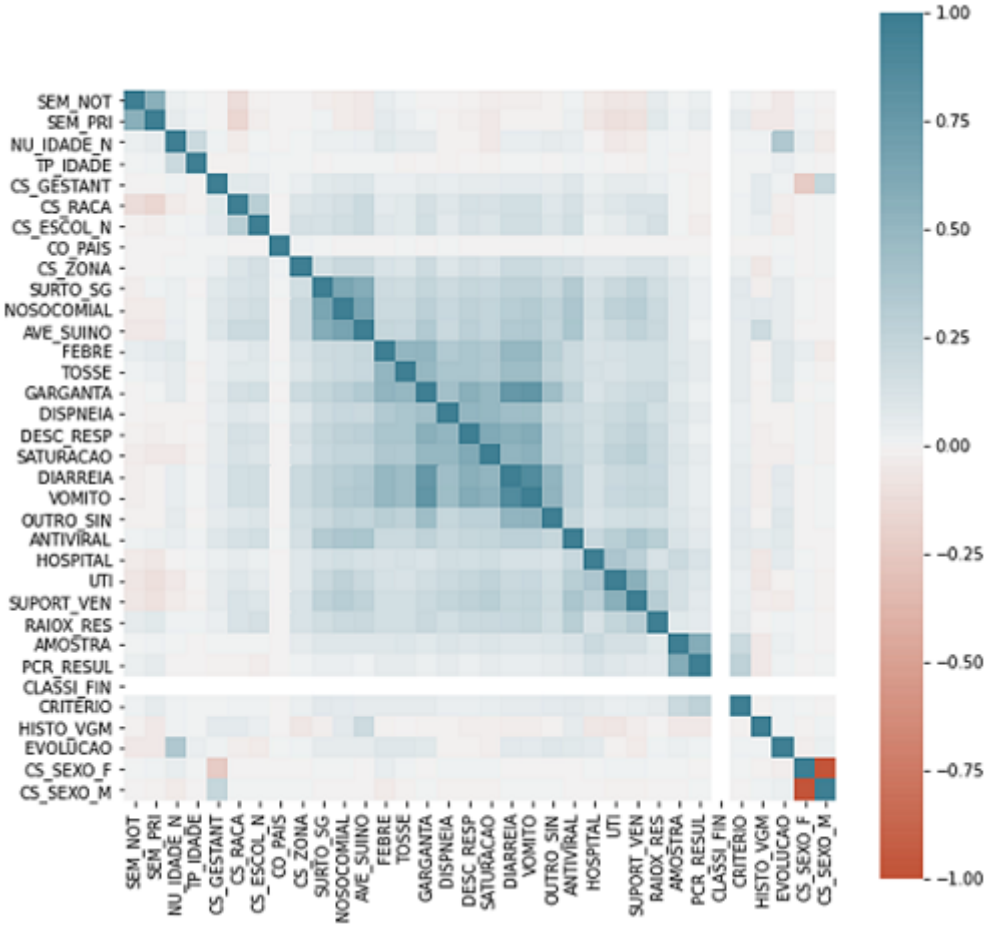


Figura 5.2: *Correlação dos atributos da coleção de dados.*

do próprio dicionário de dados foi feita a inserção desse valor nos atributos que continham valor vazio.

Por fim foi feito a divisão da coleção de dados em dados de treino e dados de teste, os dados de treino são os dados que serão usados no treinamento do modelo de Aprendizado de Máquina, os dados de teste são usados para fazer a validação do modelo, ou seja, medir a capacidade de previsão do modelo criado. Em cada coleção de dados dividido foi feito a separação da variável dependente, ou seja, o atributo de interesse de previsão, neste caso o atributo "EVOLUCAO", os resultados possíveis para esse atributo são 1 para não óbito e 2 para óbito. A coleção dos dados foi dividida em 70% para o treinamento do modelo e 30% para o teste do modelo criado.

5.3 Normalizando os Dados

Para evitar que um atributo tenha maior influência do que outro devido a grande diferença de valores, os dados foram normalizados fazendo o uso de uma

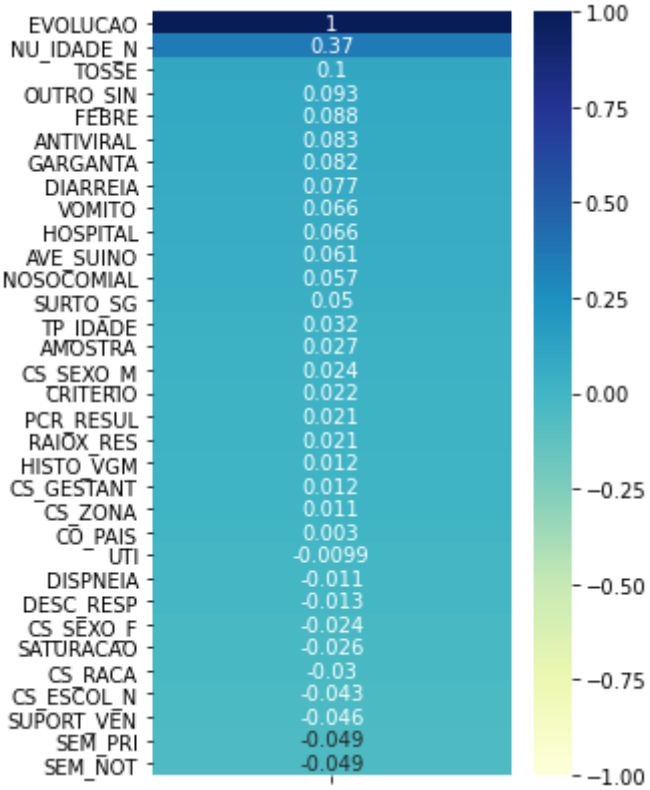


Figura 5.3: *Correlação dos atributos com o rótulo das instâncias.*

biblioteca chamada *Scikit-learn*[3] (COURNAPEAU, 2007), essa biblioteca possui a função *StandardScaler* que normaliza os valores dos dados fazendo a média dos atributos se tornar 0 e o desvio padrão 1, esse método não reescala os valores em uma faixa padrão. É importante ficar atento a quais dados usar nessa etapa, não devem ser usados os dados de teste para que não tenha vazamento de informações no momento da padronização, a biblioteca *scikit-learn* tem a função *fit* que realiza os cálculos necessários para a padronização e deve ser usada somente nos dados de treinamento do modelo. A função *transform* aplica os cálculos realizados e deve ser usada tanto nos dados de treinamento e de teste.

5.4 Balanceando os Dados

Após normalizar os dados, foi feito a contagem dos valores do atributo ”EVOLUCAO”que é o atributo que nosso modelo irá ser treinado para prever, e também a plotagem da sua distribuição. Apesar da plotagem da distribuição dos dados não conter informações relevantes, ficou constatado o desbalanceamento das

[3]biblioteca utilizada para a criação de modelos de ML

informações, com 292.983 instâncias de valor 1, não óbito e 156.210 de valor 2, óbito, demonstrado na Figura 5.4. O ”Counter”na figura indica a quantidade de valores de cada classe dentro da coleção de dados.

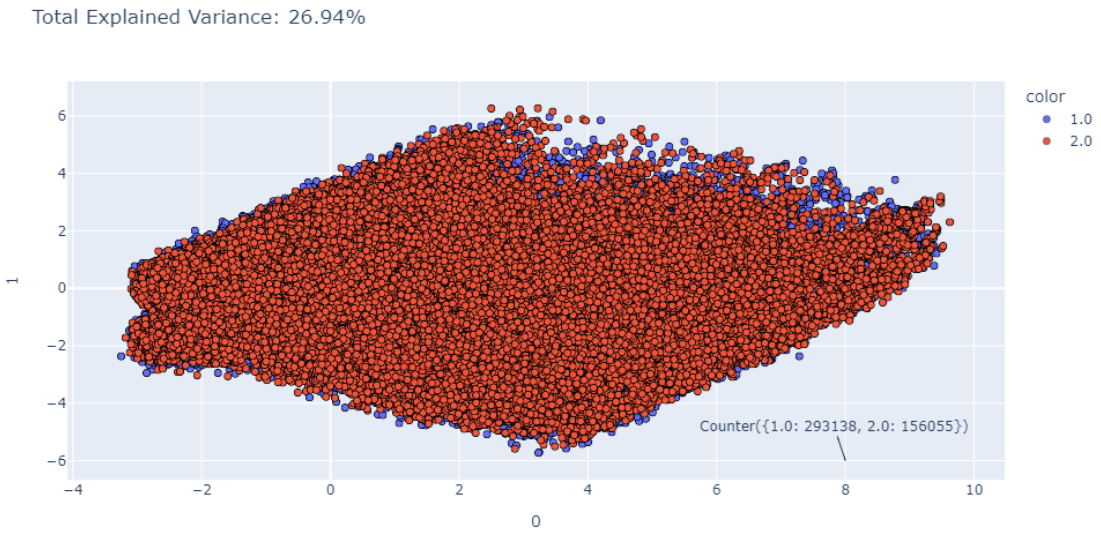


Figura 5.4: *Distribuição dos dados desbalanceados.*

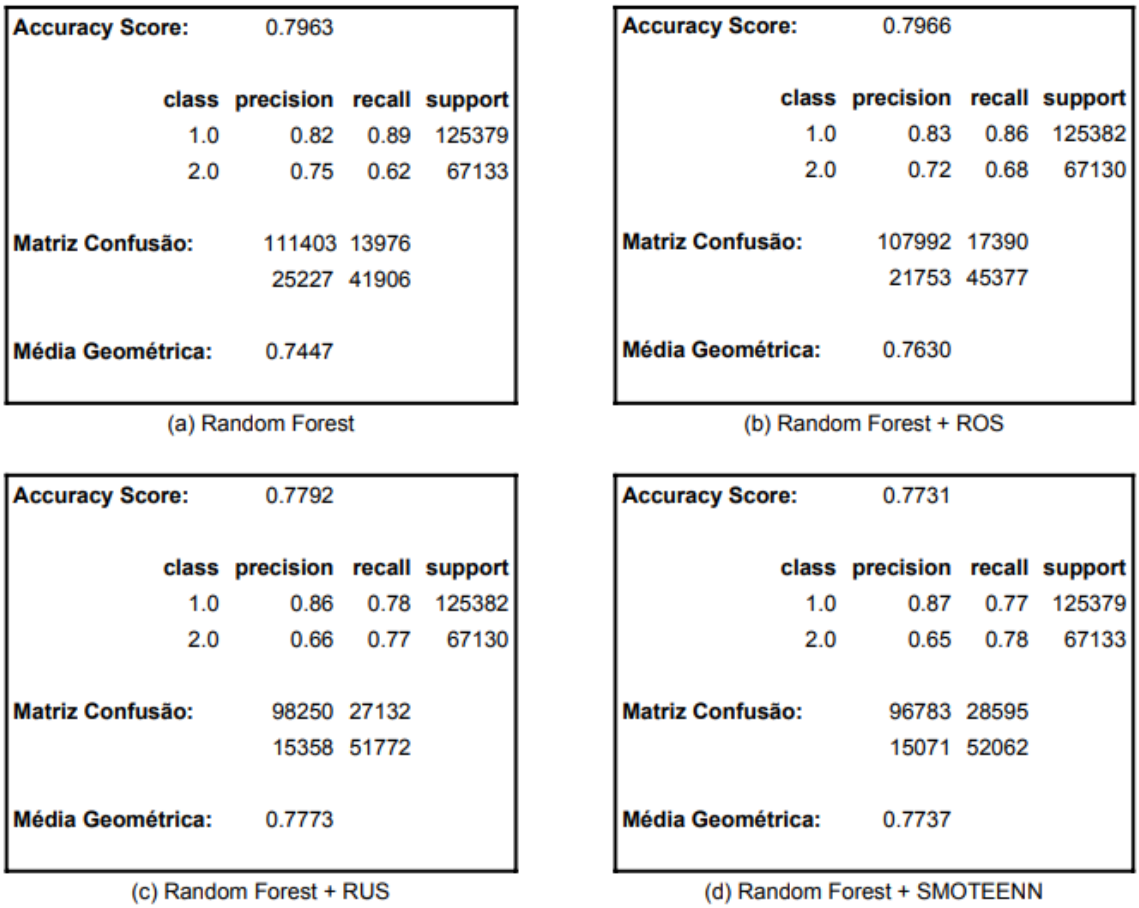
Neste trabalho foram abordadas três estratégias diferentes para resolver o problema de desbalanceamento de dados, *Random Oversampling* (ROS), *Random Undersampling* (RUS) e SMOTEENN, uma técnica híbrida que faz o *Oversampling* e *Undersampling* para eliminar os dados ruidosos.

O modelo de aprendizado de máquina utilizado foi o *Random Forest*, ele foi aplicado em todas as estratégias de balanceamento de dados utilizadas, foram coletadas as informações da matriz de confusão, medido a performance através do uso das métricas e calculado a importância de cada atributo utilizado no processo de criação do modelo de ML. Random Forest foi escolhido por ser um modelo potente e costuma ter bons resultados em aplicações de aprendixado de máquina em que os dados possuam alta complexidade, alem de ser um modelo mais resistente a *Overfitting*.

Para facilitar o uso do *Random Forest*, foram criadas novas coleções de dados para cada técnica de balanceamento usada, e também uma função que recebia esses dados e retornavam todas as informações de interesse, sendo a acurácia, precisão, *recall*, matriz de confusão e média geométrica.

5.5 Resultados

A Figura 5.5 contém as métricas do *Random Forest* de todas as técnicas de balanceamento usadas e está organizada da seguinte forma, a acurácia de cada modelo, sua precisão, *recall*, matriz de confusão e por fim a média geométrica. A medida utilizada para avaliar o desempenho do modelo foi a média geométrica. Das abordagens adotadas a que teve o melhor desempenho foi *Random Forest* utilizando RUS, comparando com *Random Forest*, que é o aprendizado de máquina com os dados sem o uso de técnicas de balanceamento, foi obtido uma melhora de 0.744 para 0.777, apesar de não parecer um ganho expressivo, ao analisar os valores da matriz de confusão, o modelo utilizando ROS teve um aumento de 9.866 acertos para a classe desbalanceada, melhorando a capacidade de previsão do modelo.



(a) Random Forest

(b) Random Forest + ROS

(c) Random Forest + RUS

(d) Random Forest + SMOTEENN

Figura 5.5: *Métricas obtidas por cada estratégia de balanceamento.*

A Tabela 5.1 mostra a importância de cada atributo para o treinamento do modelo de ML utilizando a abordagem RUS, é interessante observar quais os atributos mais relevantes. Para o modelo, a idade é um fator determinante para a classificação de óbito ou não óbito de pessoas infectadas pelo vírus do COVID19. Os atributos referentes ao estado de saúde da pessoa (SUPORT_VEN,

UTI, RAIOX_RES) também apresentam alta importância para o modelo (ver Tabela 5.1).

Importância dos atributos (%)	
NU_IDADE_N	0.20
SUPORT_VEN	0.13
SEM_NOT	0.08
SEM_PRI	0.08
UTI	0.07
RAIOX_RES	0.04
CS_RACA	0.03
CS_ESCOL_N	0.03
CS_GESTANT	0.02
CS_ZONA	0.02
SURTO_SG	0.02
NOSOCOMIAL	0.02
FEBRE	0.02
TOSSE	0.02
GARGANTA	0.02
DESC_RESP	0.02
SATURACAO	0.02
DIARREIA	0.02
OUTRO_SIN	0.02
ANTIVIRAL	0.02
PCR_RESUL	0.02
HISTO_VGM	0.02
AVE_SUINO	0.01
DISPNEIA	0.01
VOMITO	0.01
HOSPITAL	0.01
CRITERIO	0.01
CS_SEXO_F	0.01
CS_SEXO_M	0.01

Tabela 5.1: *Tabela de importância na abordagem RUS*

O tempo de execução obtido varia para cada modelo, como visto na Tabela 5.2, é importante que os modelos possuam um balanceamento entre o resultado e o seu tempo de execução. O modelo que teve o menor tempo de execução foi aquele que fez o uso do *Random Undersampling*, com um tempo de execução total de 53.629 milissegundos(ms), sendo 53.400/ms para o balanceamento dos dados e 229/ms para o treinamento do ML. A abordagem com o maior tempo de execução foi a SMOTEENN com 4.763.700/ms, 4.716.000/ms para o balanceamento dos dados e 47.700/ms para o treinamento do ML.

Tempo de execução(milissegundos)			
	Balanceamento	ML	Total
Random Forest	-	72.000	72.000
Random Oversampling	85.800	215	86.015
Random Undersampling	53.400	229	53.629
SMOTEENN	4.716.000	47.700	4.763.700

Tabela 5.2: *Tempo de execução de cada abordagem.*

CAPÍTULO 6

Conclusões e Trabalhos Futuros

O trabalho apresenta as etapas no desenvolvimento de um modelo de aprendizado de máquina utilizando uma coleção de dados desbalanceados. O intuito do trabalho é avaliar os resultados obtidos pelo uso de diferentes abordagens de balanceamento de dados a partir das métricas obtidas analisando o tempo de execução de cada uma das abordagens usadas e também entender quais são os atributos mais importantes para o modelo de Aprendizado de Máquina.

O trabalho demonstra a importância do balanceamento dos dados para a obtenção de um modelo de ML de alta confiabilidade. É possível perceber como o uso de cada uma das técnicas de balanceamento pode oferecer resultados e desempenho diferentes. Algumas abordagens apresentam soluções complexas, porém o tempo de execução pode acabar impactando na sua escolha quando o ganho obtido é mínimo comparado a abordagens menos complexas e mais rápidas.

Ficou claro que para o modelo a idade da pessoa e o uso de ventiladores respiratórios são os atributos mais importantes no treinamento do modelo preditivo, seguidos de semana da notificação, semana do primeiro sintoma e se a pessoa foi para a UTI. Essas informações podem ser importantes para a criação de protocolos de tratamento para pacientes infectados pelo vírus da COVID19

Através do estudo é possível perceber que, se tratando de modelos desbalanceados o uso de métricas convencionais pode ocasionar uma interpretação equivocada do seu desempenho. No trabalho apresentado, o modelo sem o uso de nenhuma técnica de balanceamento apresentava uma boa acurácia, porém tem um péssimo desempenho na previsão da classe desbalanceada. É importante entender como cada métrica é calculada e fazer a escolha certa para a validação do modelo para evitar a sua má interpretação.

Como trabalhos futuros, pretende-se fazer estudo de outras métricas robustas e adequadas para modelos desbalanceados, como *Cohen's Kappa* e Área Abaixo da Curva ROC (AUC) (JENI; COHN; TORRE, 2013) e utilizar outros modelos de classificação, como *Support Vector Machine* (SVM) e Árvores de Decisão comparando os resultados obtidos.

Bibliografia

ACOCK, A. C. Working with missing values. *Journal of Marriage and Family*, v. 67, n. 4, p. 1012–1028, 2005.

BARANDELA J.S. SANCHEZ, V. G. R.; RANGEL, E. Strategies for learning in class imbalance problems. In: *The Journal of the Pattern Recognition Society*. [S.l.: s.n.], 2003.

BEN-GAL, I. Outlier detection. In: _____. *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer US, 2005. p. 131–146. ISBN 978-0-387-25465-4. Disponível em: <https://doi.org/10.1007/0-387-25465-X_7>.

BREIMAN, L. Random forests. *Mach. Learn.*, Kluwer Academic Publishers, USA, v. 45, n. 1, p. 5–32, out. 2001. ISSN 0885-6125. Disponível em: <https://doi.org/10.1023/A:1010933404324>.

COHEN, G. et al. Learning from imbalanced data in surveillance of nosocomial infection. *Artificial intelligence in medicine*, v. 37 1, p. 7–18, 2006.

COURNAPEAU, D. *scikit-learn*. 01 2007. <https://scikit-learn.org/stable/>. Accessed: 2021-06-08.

CUNNINGHAM, P.; CORD, M.; DELANY, S. J. Supervised learning. In: _____. *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. p. 21–49. ISBN 978-3-540-75171-7. Disponível em: <https://doi.org/10.1007/978-3-540-75171-7_2>.

FERRI, J. H.-O. C.; MODROIU, R. An experimental comparison of performance measures for classification. In: *Pattern Recognition Letters*. [S.l.: s.n.], 2006. v. 30, p. 27–38.

GU, Q.; ZHU, L.; CAI, Z. Evaluation measures of the classification performance of imbalanced data sets. In: CAI, Z. et al. (Ed.). *Computational Intelligence and Intelligent Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 461–471.

HASOKSUZ, M.; KILIç, S.; SARAC, F. Coronaviruses and sars-cov-2. *Turkish journal of medical sciences*, v. 50, 04 2020.

HE YANG BAI, E. A. G. H.; LI, S. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: *Conference: Neural Networks*. [S.l.: s.n.], 2008.

HOX, J.; BOEIJE, H. Data collection, primary versus secondary. *Encyclopedia of Social Measurement*, v. 1, p. 593–599, 12 2005.

HUNTER, J. D. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, IEEE Computer Society, v. 9, n. 03, p. 90–95, 2007.

JENI, L. A.; COHN, J. F.; TORRE, F. D. L. Facing imbalanced data–recommendations for the use of performance metrics. In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. [S.l.: s.n.], 2013. p. 245–251.

KLUYVER, T. et al. *Jupyter Notebooks-a publishing format for reproducible computational workflows.* [S.l.: s.n.], 2016.

LEEVY TAGHI M. KHOSHGOFTAAR, R. A. B. J. L.; SELIYA, N. A survey on addressing high-class imbalance in big data. In: LEEVY, J. L. (Ed.). *Journal of Big Data*. [S.l.: s.n.], 2018.

LEMAÎTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, JMLR. org, v. 18, n. 1, p. 559–563, 2017.

LIU, A.; GHOSH, J.; MARTIN, C. Generative oversampling for mining imbalanced datasets. In: . [S.l.: s.n.], 2007. p. 66–72.

LONES, M. A. *How to avoid machine learning pitfalls: a guide for academic researchers*. 2021.

MAHER, N. A. et al. Passive data collection and use in healthcare: A systematic review of ethical issues. *International Journal of Medical Informatics*, v. 129, p. 242–247, 2019.

MANJU, B.; NAIR, A. Classification of cardiac arrhythmia of 12 lead ecg using combination of smoteenn, xgboost and machine learning algorithms. In: *2019 9th International Symposium on Embedded Computing and System Design (ISED)*. [S.l.: s.n.], 2019. p. 1–7.

MARRELLI, A. Collecting data through case studies. *Performance Improvement*, v. 46, p. 39 – 44, 08 2007.

MCKINNEY, W. *pandas*. 01 2008. <https://pandas.pydata.org/>. Accessed: 2021-06-08.

MCKINNEY, W. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. [S.l.]: "O'Reilly Media, Inc.", 2012.

MOHAMMED, R.; RAWASHDEH, J.; ABDULLAH, M. Machine learning with oversampling and undersampling techniques: Overview study and experimental results. In: . [S.l.: s.n.], 2020. p. 243–248.

NAQA, I. E.; MURPHY, M. J. What is machine learning? In: _____. *Machine Learning in Radiation Oncology: Theory and Applications*. Cham: Springer International Publishing, 2015. p. 3–11. ISBN 978-3-319-18305-3. Disponível em: <https://doi.org/10.1007/978-3-319-18305-3_1>.

ORGANIZATION, W. H. *WHO Coronavirus (COVID-19) Dashboard*. 01 2021. <https://covid19.who.int/region/amro/country/br>. Accessed: 2021-06-08.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, JMLR. org, v. 12, p. 2825–2830, 2011.

QUINLAN, J. R. Induction of decision trees. *Mach. Learn.*, Kluwer Academic Publishers, USA, v. 1, n. 1, p. 81–106, mar. 1986. ISSN 0885-6125. Disponível em: <https://doi.org/10.1023/A:1022643204877>.

RESMINI AURA CONCI, T. B. B. R. Diagnóstico precoce de doenças mamárias usando imagens térmicas e aprendizado de máquina. In: *Revista Eletrônica do Alto Vale do Itajaí*. [S.l.: s.n.], 2012. p. 55–67.

ROSSUM, G. V.; DRAKE, F. L. et al. *Python reference manual*. [S.l.]: iUniverse Indiana, 2000.

SAUDE, S. U. da. *SRAG 2020 - Banco de Dados de Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19*. 01 2021. <https://opendatasus.saude.gov.br/dataset/bd-srag-2020>. Accessed: 2021-06-08.

TALóN-BALLESTERO, P. et al. Using big data from customer relationship management information systems to determine the client profile in the hotel sector. *Tourism Management*, v. 68, p. 187–197, 2018. ISSN 0261-5177. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0261517718300670>.

TSAI, M.-H. et al. A machine learning based strategy for election result prediction. In: *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*. [S.l.: s.n.], 2019. p. 1408–1410.

VARMEDJA, D. et al. Credit card fraud detection - machine learning methods. In: *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*. [S.l.: s.n.], 2019. p. 1–5.

VISA, S. et al. Confusion matrix-based feature selection. In: . [S.l.: s.n.], 2011. v. 710, p. 120–127.

WASKOM, M. *seaborn*. 01 2012. <https://seaborn.pydata.org/>. Accessed: 2021-06-08.