INSTITUTO FEDERAL DO ESPÍRITO SANTO
CURSO DE ENGENHARIA ELÉTRICA

IGOR MIRANDA EISENLOHR

USO DE MODELOS DE APRENDIZADO DE MÁQUINA PARA ANÁLISE E PREDIÇÃO DE RECUPERAÇÃO DE PACIENTES DE COVID-19 NO MUNICÍPIO DE VITÓRIA-ES

Vitória
2022

IGOR MIRANDA EISENLOHR

USO DE MODELOS DE APRENDIZADO DE MÁQUINA PARA ANÁLISE E PREDIÇÃO DE RECUPERAÇÃO DE PACIENTES DE COVID-19 NO MUNICÍPIO DE VITÓRIA-ES

Trabalho de Conclusão de Curso apresentado à Coordenadoria do Curso de Engenharia Elétrica do Instituto Federal do Espírito Santo, Campus Vitória, como requisito parcial para obtenção de título de Bacharel em Engenharia Elétrica.

Orientador: Profa. Dra. Mariana Rampinelli Fernandes

Vitória

2022

Dados Internacionais de Catalogação na Publicação (CIP)
(Biblioteca Nilo Peçanha do Instituto Federal do Espírito Santo)

E36u Eisenlohr, Igor Miranda.

Uso de modelos de aprendizado de máquina para análise e predição de recuperação de pacientes de covid-19 no município de Vitória-ES / Igor Miranda Eisenlohr – 2022.

74 f. : il. ; 30 cm

Orientadora: Mariana Rampinelli Fernandes .

Monografia (graduação) – Instituto Federal do Espírito Santo, Coordenadoria de Engenharia Elétrica, Curso Superior de Engenharia Elétrica, 2022.

1. Aprendizado do computador. 2. Mineração de dados (computação). 3. Covid-19 (Doença). 4. Avaliação de riscos de saúde. 5. Algoritmos. 6. Engenharia Elétrica. I. Fernandes, Mariana Rampinelli. II. Instituto Federal do Espírito Santo. III. Título.

CDD 21 – 006.31

Elaborada por Bruno Giordano Rosa – CRB-6/ES – 699

IGOR MIRANDA EISENLOHR

USO DE MODELOS DE APRENDIZADO DE MÁQUINA PARA ANÁLISE E PREDIÇÃO DE RECUPERAÇÃO DE PACIENTES DE COVID-19 NO MUNICÍPIO DE VITÓRIA-ES

Trabalho de Conclusão de Curso apresentado à Coordenadoria do Curso de Engenharia Elétrica do Instituto Federal do Espírito Santo, Campus Vitória, como requisito parcial para obtenção de título de Bacharel em Engenharia Elétrica.

Aprovado em 04 de novembro de 2022.

COMISSÃO EXAMINADORA

(Assinado digitalmente em 14/12/2022 18:51)
MARIANA RAMPINELLI FERNANDES
PROFESSOR DO ENSINO BASICO TECNICO E TECNOLOGICO VIT-CCTE
(11.02.35.01.09.02.19)
Matrícula: 2071531
INSTITUTO FEDERAL DO ESPIRITO SANTO
Orientadora

(Assinado digitalmente em 15/12/2022 13:57)
GABRIEL TOZATTO ZAGO
PROFESSOR DO ENSINO BASICO TECNICO E TECNOLOGICO SER-CGEN (11.02.32.01.08.02)
Matrícula: 2928471
INSTITUTO FEDERAL DO ESPIRITO SANTO
Avaliador

Assinado digitalmente em 15/12/2022 16:03)
SHIRLEY PERONI NEVES CANI
PROFESSOR DO ENSINO BASICO TECNICO E TECNOLOGICO VIT-CCTE
(11.02.35.01.09.02.19)
Matrícula: 1616448
INSTITUTO FEDERAL DO ESPIRITO SANTO
Avaliadora

RESUMO

Com a aparição do novo coronavírus, SARS-CoV-2, o mundo passou a estar em condição de pandemia e 6.341.446 óbitos foram registrados até a data de 27 de junho de 2022. Por meio do avanço da tecnologia, técnicas robustas de análise de dados surgiram e, com o auxílio de profissionais da saúde, cientistas de dados são capazes de analisar as consequências geradas pela Covid-19. A fim de realizar o estudo de como a doença afetou os pacientes confirmados com o vírus no município de Vitória, Espírito Santo, este trabalho realiza o pré-processamento da base de dados com os casos registrados. Após o pré-processamento, os dados são utilizados em modelos de aprendizado de máquina como árvore de decisão, floresta aleatória, XGBoost e AdaBoost, com o intuito de classificar um novo paciente de acordo com sua chance de sobreviver à infecção do novo coronavírus. Essa classificação é realizada a partir das suas condições inicias de saúde, de acordo com as possibilidades de sobreviver ao vírus dadas as características de cada caso.

Palavras-chave: Análise de Dados. COVID-19. Aprendizado de Máquina. Prognóstico.

ABSTRACT

With the appearance of the new coronavirus, SARS-CoV-2, the world became in a pandemic condition and 6,341,446 deaths were recorded up to June 27, 2022. Through the advancement of technology, robust data analysis techniques emerged, and with the help of health professionals, data scientists can analyze the consequences of Covid-19. In order to carry out the study of how the disease affected patients confirmed with the virus in the city of Vitoria, Espirito Santo, this work performs the pre-processing of the database with the registered cases. After pre-processing, the data was used in machine learning models such as decision trees, random forests, XGBoost, and AdaBoost, to classify a new patient according to their chance of surviving the new coronavirus infection. This classification is carried out based on their initial health conditions, according to the possibilities of surviving the virus given the characteristics of each case.

Keywords: Data Analysis. COVID-19. Machine Learning. Prognostic.

LISTA DE FIGURAS

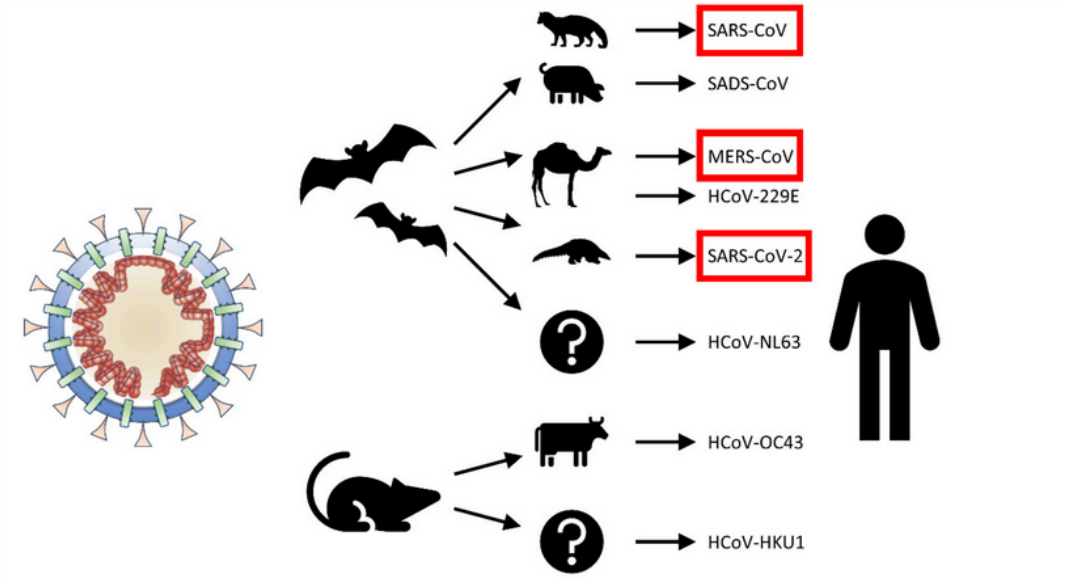
LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Ao longo da história, sete tipos de coronavírus foram identificados em seres humanos (HCoVs), dentre eles, está o SARS-CoV-2, vírus responsável pela COVID-19. Todos os tipos de coronavírus tiveram origem em morcegos ou roedores até infectarem os humanos (RABI et. al., 2020).

Figura 1 – Tipos de coronavírus com destaque os identificados em humanos.



Fonte: Firas A Rabi, Mazhar S Al Zoubi et al/MDPI.com (2020, p. 3)

Em 31 de dezembro de 2019, a Organização Mundial da Saúde (OMS) recebeu informações sobre alguns casos de uma suposta pneumonia na cidade chinesa de Wuhan. Na semana seguinte, as autoridades chinesas haviam identificado um novo tipo de coronavírus que ainda não apresentava registros em seres humanos, tratava-se do SARS-CoV-2 (ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE (OPAS), [entre 2020 e 2022]).

Devido ao elevado grau de contágio do vírus, em 30 de janeiro de 2020 a OMS declarou surto, fato que constituiu uma Emergência de Saúde Pública de Importância Internacional (ESPII). Tal fato, representa o maior nível de alerta da Organização e essa decisão ocorreu na tentativa de controlar e interromper a propagação do vírus (ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE (OPAS), [entre 2020 e 2022]). Os pacientes infectados com a doença passaram a sentir

diversos sintomas, dentre esses os mais comuns foram: febre, tosse, dificuldades respiratórias, fadiga, dores musculares, dor de cabeça, perda do olfato e paladar, dores de garganta, coriza, náusea e diarreia (INSTITUTO BUTANTAN, 2021).

Desde então, a pandemia demarca a história com grande crise de saúde pública e econômica mundial e apresenta 547.507.804 casos confirmados e 6.341.446 mortes registradas no mundo até o período de 27 de junho de 2022. No mesmo período, o vírus afetou o Brasil com 32.434.063 casos e o acumulado de 671.700 mortes (WHO, 2022).

Com números tão expressivos em casos confirmados e casos que se agravaram, por diversas vezes os leitos hospitalares apresentaram níveis elevados de ocupação e muitas cidades enfrentaram superlotação nos hospitais e até o colapso do sistema de saúde. Assim, em muitas ocasiões, houve a necessidade de escolher quais pacientes iriam preferencialmente ocupar os leitos da UTI (Unidade de Terapia Intensiva) (AMB, 2021). Para tanto, mostrou-se necessário avaliar as características de cada paciente internado, para verificar quais, de acordo seus parâmetros vitais, comorbidades e sintomas, apresentavam maiores chances de se recuperar da doença. A partir da análise dessas características, as equipes médicas conseguiam escalonar uma ordem de prioridades para a ocupação dos leitos de UTI com foco nos pacientes que apresentassem maiores chances de se recuperar do estágio grave da doença (AMB, 2021). Dessa forma, com o intuito de auxiliar essa difícil tomada de decisão das equipes médicas e obter decisões mais precisas, este trabalho apresenta modelos de aprendizado de máquina que possam classificar novos pacientes infectados como possíveis sobreviventes ou possíveis casos de óbito. Além dos modelos de aprendizado de máquina, também será apresentado todo o estudo e pré-processamento da base de dados

Neste trabalho, serão utilizados modelos supervisionados voltados para classificação binária do paciente. Pacientes classificados como sobreviventes, são aqueles que, segundo o modelo, possuiriam maiores chances de evoluir para a cura da doença. Os algoritmos escolhidos serão baseados em árvore de

decisão, dentre esses, o próprio algoritmo de árvore de decisão, a floresta aleatória, o XGBoost e o AdaBoost e as métricas para avalição da qualidade dos resultados obtidos serão: matriz de confusão, acurácia, precisão, F1-*score* e ROC AUC (HARRISON, 2020). Por se tratar de algo tão sério como a vida dos pacientes, o modelo necessita apresentar um ótimo nível de avaliação, a fim de auxiliar os médicos no combate do vírus. Cabe salientar, ainda, que os protocolos de ética médica devem ser seguidos e o resultado da classificação do modelo não se apresenta como diagnóstico e nem como a decisão final, que ficará a cargo da equipe médica.

Além disso, destaca-se que o trabalho foi realizado com base nos dados coletados pelo governo do Estado do Espírito Santo, usando apenas os casos provenientes do município de Vitória. A escolha de apenas um município teve por objetivo reduzir o grande volume de dados gerados pelo estado do Espírito Santo. Além disso, Vitória foi escolhida devido à organização dos dados cadastrados dos pacientes na base de dados. Por fim, mostra-se importante ressaltar que cada município é responsável por coletar seus próprios dados, de forma que, em diferentes localizações, os dados coletados podem abordar características distintas para os pacientes.

1.1 OBJETIVOS

O principal objetivo do trabalho é elaborar modelos de aprendizado de máquina capazes de predizer a recuperação dos pacientes infectados pelo novo coronavírus no município de Vitória, Espírito Santo.

Para isso, as seguintes etapas foram realizadas:

- Pré-processamento dos dados:
  - limpar a base de dados, codificar as variáveis categóricas, avaliar as características de maior importância para o modelo e balancear os dados.
- Análise de dados:
  - explorar os dados com estatísticas resumidas e gerar gráficos que auxiliem no entendimento do impacto do vírus nos pacientes do município.

- Modelagem dos algoritmos de aprendizado de máquinas:
  - classificar os pacientes de acordo com sua chance de óbito ou recuperação, a partir das variáveis de maior impacto selecionadas na etapa de pré-processamento.

1.2 ESTRUTURA

O Capítulo 2 apresenta os fundamentos teóricos necessários para a elaboração do trabalho, como técnicas de pré-processamento, algoritmos para a elaboração do modelo de aprendizado de máquina e as métricas para avaliar qual modelo desempenhou melhor. Já no Capítulo 3, os métodos de como aplicar os conhecimentos adquiridos com o referencial teórico serão abordados, com todas as etapas de execução do estudo para obtenção dos resultados. Em seguida, no Capítulo 4, os resultados encontrados serão apresentados e, por fim, no Capítulo 5, o foco será em apresentar discussões relevantes sobre o trabalho, como as limitações encontradas e sugestões para trabalhos futuros.

2 REFERENCIAL TEÓRICO

Neste capítulo serão apresentados conhecimentos fundamentais para a elaboração do trabalho. A princípio, evidencia-se a importância dos dados na atualidade e como os dados podem auxiliar na área da saúde, em seguida, são abordados conceitos sobre técnicas de análise e pré-processamento de dados, de algoritmos de aprendizado de máquina e de métricas para avaliação do desempenho de classificadores.

2.1 IMPORTÂNCIA DOS DADOS

Atualmente, com todo o avanço da tecnologia e a ascensão do mundo digital, cada vez mais os dados ganham relevância. Dessa forma, *websites* rastreiam as atividades dos usuários para armazenar seus interesses de acordo com as páginas e conteúdos acessados, *smartphones* registram localizações e velocidade dos usuários para identificar lugares que frequentam e o tempo gasto em cada local, entre vários outros dispositivos que se baseiam e coletam todo tipo de informação constantemente (GRUS, 2016). Com tais informações, os dados tornaram-se uma indústria de rápido crescimento e extremamente lucrativa, sendo considerado o recurso mais valioso do mundo. Tal fato, deve-se pela possibilidade de as empresas personalizarem diferentes ambientes virtuais de acordo com os interesses de cada usuário e, assim, capitalizar a partir desses dados (THE ECONOMIST, 2017).

2.1.1 ANÁLISE DE DADOS NA ÁREA DA SAÚDE

Não somente para o crescimento de empresas, a análise de dados na área da saúde mostra-se fundamental para auxílio dos profissionais em intervenções clínicas e para desenvolvimento de políticas públicas por parte do Estado (DUARTE et al, 2021). A utilização de dados na área da saúde iniciou-se com John Snow no século XIX, época na qual milhões de pessoas morreram por surtos de cólera. Em 1854, o médico inglês passou a reparar que a maior parte dos casos da doença em Londres estavam concentrados em uma determinada rua. Com essa informação, pode-se encontrar um poço de água que estava no

centro da pandemia (DUARTE et al, 2021). A descoberta teve extrema importância, pois a partir desse fato ocorreu o fechamento do poço e tornou-se possível salvar milhares de pessoas com a redução do contágio da doença.
Além disso, Snow concluiu que "Era entre os pobres, com famílias que viviam, dormiam, cozinhavam, comiam e se asseavam juntas em um único cômodo que a cólera se expandia" (SNOW, 1854 *apud* BOWES, 2020). A partir de tal conclusão, o médico percebeu um padrão específico e, assim, identificou que a doença bacteriana não era transmitida por meio do ar, na verdade ocorria por meio da ingestão de água ou alimentos contaminados, ou de pessoa para pessoa (BOWES, 2020).

Logo, desde o século XIX, a análise de dados destacava-se nas descobertas da área da saúde que auxiliam a salvar milhares de vidas. Até os dias de hoje, epidemiologistas e cientista de dados trabalham na intenção de identificar particularidades e padrões em doenças infecciosas, com a intenção de entender os efeitos e elaborar tecnologias, medicamentos e políticas públicas eficazes no combate à doença (DUARTE et al, 2021).

2.2 ANÁLISE E PRÉ PROCESSAMENTO DOS DADOS

A análise de dados busca explicar as inúmeras informações presentes nas bases de dados e apresentar conclusões precisas que auxiliem a melhorar o impacto nos negócios (HARRISON, 2020). Como, na maioria das vezes, a coleta de dados parte de um grande processo de construção que envolve diversas pessoas, características e respostas individuais, os conjuntos de dados podem apresentar erros e inconsistências que podem prejudicar no seu entendimento (DUARTE et al, 2021).

O primeiro passo necessário para análise de dados e construção de modelos de aprendizado de máquina, é a realização da limpeza dos dados na etapa de pré-processamento, a fim de explorar quais tipos de valores únicos são possíveis para cada variável, realizar a contagem de cada um, buscar por valores estatísticos como média, mediana, quartis, valores mínimos e máximos dos atributos (HARRISON, 2020). Após esse processo, deve-se remover variáveis

com muitos valores ausentes, codificar variáveis categóricas em numéricas e selecionar as variáveis que mais agregam ao modelo no intuito de remover as que não acrescentem no resultado final (HARRISON, 2020). Além disso, na etapa de análise dos dados, a parte de visualização apresenta grande importância, pois facilita o entendimento dos dados por meio da construção de gráficos.

Ademais, como forma de apresentar pesquisas de aprendizado de máquina com múltiplas variáveis relacionadas a prognósticos e diagnósticos de maneira adequada e transparente, há uma lista de verificações denominada TRIPOD. A TRIPOD é uma lista que auxilia o desenvolvimento de etapas necessárias para a elaboração de um modelo de predição que descreva de forma detalhada os processos realizados na etapa de pré-processamento e aborde os resultados e as limitações dos trabalhos (COLLINS *et al*., 2015). O guia TRIPOD também recomenda a separação do conjunto de dados para utilizar parte como treino e validação, a fim de evitar a construção de modelos que apresentem um *overfitting* e seja otimista com os dados futuros (COLLINS *et al*., 2015).

2.2.1 SELEÇÃO DE CARACTERÍSTICAS

Modelos de aprendizado de máquina podem ter dificuldades em lidar com conjuntos de dados que apresentem um grande número de variáveis para cada amostra. Dessa forma, o pré-processamento dos dados, mostra-se imprescindível para a construção de um bom classificador. Com isso, há a preocupação em eliminar variáveis irrelevantes ou redundantes que não apresentem informações tão relevantes para a explicação da variável de saída. Tal fato, pode aumentar a eficiência e a compreensão dos modelos de aprendizado de máquina, além de reduzir a complexidade e o custo computacional dos algoritmos (KUMAR et al., 2014).

Desse modo, algumas técnicas podem ser utilizadas para auxiliar na redução da dimensionalidade do conjunto de dados. Dentre elas, destacam-se a análise individual de cada variável para identificar quais apresentam valores inconsistentes, e a utilização de algoritmos que facilitem na capacidade de

generalização, na velocidade de treinamento e na redução da complexidade dos dados (KUMAR et al., 2014). Uma técnica de destaque para a seleção de características em modelos de aprendizado de máquina que possuem dados categóricos é o teste matemático χ^2. O método tem como objetivo identificar a dependência entre duas variáveis por meio da comparação entre a distribuição da variável observada, como exemplo a Tabela 1, na qual a distribuição observada entre duas variáveis foi representada por “A”, “B”, “C” e “D”. A partir de valores observados com a criação de uma tabela de contingência, torna-se possível o cálculo da distribuição esperada para cada caso, como mostra a Tabela 2.

Tabela 1 – Nomenclatura da tabela de contingência

Nomenclatura tabela de contingência			
	Ocorrência evento X	**Não ocorrência evento X**	TOTAL
Ocorrência evento Y	A	B	A+B
Não ocorrência evento Y	C	D	C+D
TOTAL	A+C	B+D	n=A+B+C+D

Fonte: elaborado pelo autor.

Tabela 2 – Valores esperados para cálculo do χ^2.

Valores esperados			
	Ocorrência evento X	**Não ocorrência evento X**	TOTAL
Ocorrência evento Y	(A+B)*(A+C)/N	(A+B)*(B+D)/N	A+B
Não ocorrência evento Y	(C+D)*(A+C)/N	(C+D)*(B+D)/N	C+D
TOTAL	A+C	B+D	n=A+B+C+D

Fonte: elaborado pelo autor.

Com o cálculo da distribuição esperada para cada evento (Tabela 2) e com os valores observados em cada amostra (Tabela 1), é possível realizar o cálculo matemático $\chi^2 = \sum \frac{(Observado - esperado)^2}{Esperado}$ para avaliar o grau de relação entre as variáveis.

2.2.2 BALANCEAMENTO DOS DADOS

Alguns conjuntos de dados como fraudes em transações bancárias ou prognóstico de doenças, apresentam a variável de saída desbalanceada, ou seja, a quantidade de ocorrências de um evento é consideravelmente maior que o evento oposto. Assim, modelos de aprendizado de máquina apresentam forte tendência a classificar os novos casos de acordo com os casos majoritários encontrados nos conjuntos de dados e, normalmente, a classe minoritária apresenta maior relevância para esses casos de anomalia (BROWNLEE, 2020).

Nesses casos, evidencia-se a necessidade de realizar a distribuição das classes, para isso, existem técnicas de amostragem que podem ser aplicadas no conjunto de dados de treino, como os métodos de *undersampling* e *oversampling.* No caso do *undersampling*, a técnica consiste no processo de separar, de forma aleatória, uma amostra dos dados pertencentes a classe majoritária que tenha o mesmo tamanho de casos que a classe minoritária, resultando em um conjunto de dados de treino com tamanho menor que o original. Já no processo de *oversampling*, ocorre a criação de dados de maneira duplicada da classe minoritária ou a criação de amostras com valores escolhidos de forma aleatória, no intuito de igualar o número de casos de ambas as classes (JOHNSON, 2019).

Além das técnicas descritas, destaca-se um dos algoritmos provenientes do *oversampling* que é denominado de *SMOTE* (*Synthetic Minority Oversampling Technique).* O método consiste na seleção aleatória de uma amostra da classe minoritária e o algoritmo avalia, por padrão, os 5 casos com maiores semelhanças e determina a distância entre eles. Com isso, novos casos são gerados de acordo com a proximidade dos dados originais, garantindo que os dados sintéticos representem de maneira mais fiel os casos reais da base de dados (BROWNLEE, 2020).

2.3 APRENDIZADO DE MÁQUINA

O aprendizado de máquina, ganhou notoriedade juntamente com o avanço da importância dos dados e, consequentemente, das técnicas de ciência de dados.

Essa técnica de programação permite que o próprio algoritmo aprenda a detectar padrões a partir de um conjunto de dados. Dessa forma, com diversas informações e resultados, modelos apropriados de aprendizado de máquina conseguem realizar predições futuras a partir de vários acontecimentos passados (HURWITZ; KIRSCH, 2018).

Esses algoritmos normalmente são complexos e baseiam-se em métodos matemáticos e estatísticos para classificar ou prever resultados a partir de um conjunto de dados. Com as respostas obtidas por meio desses códigos, tornou-se possível antecipar decisões, planos e estratégias, realizar algoritmos de recomendações, de reconhecimento de fala, entre outras diversas aplicações que facilitam a execução de várias tarefas (IBM CLOUD EDUCATION, 2020).

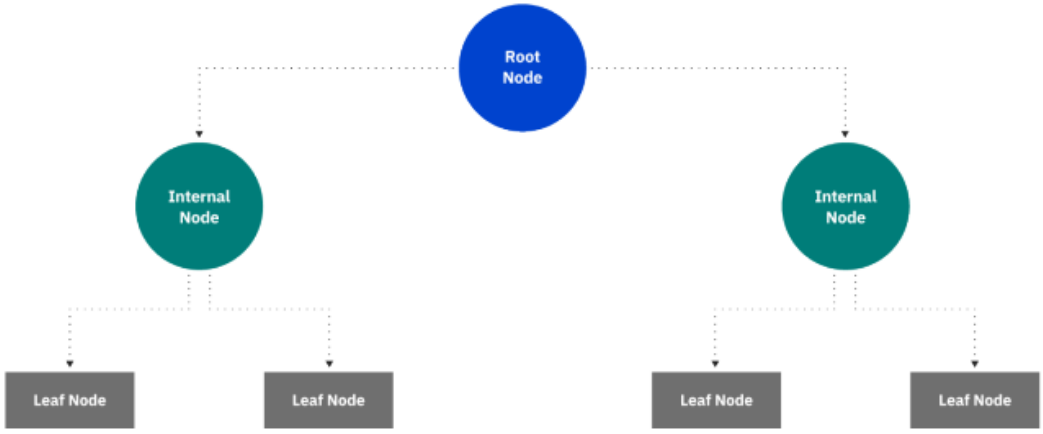
As técnicas de aprendizado de máquina podem ser classificadas em diferentes paradigmas: aprendizado supervisionado, em que se conhece a resposta correta (rótulos) que a técnica deve alcançar; aprendizado não-supervisionado, aplicada a dados não rotulados, ou seja, para os quais não se conhece a resposta *a priori*; aprendizado semi-supervisionado, que mescla os dois primeiros; e o aprendizado por reforço. Dentre os modelos de aprendizado supervisionado, pode-se destacar os algoritmos baseados em decisões, como árvore de decisão, floresta aleatória, AdaBoost e XGBoost.

2.3.1 ÁRVORE DE DECISÃO

O algoritmo de árvore de decisão para classificação é muito utilizado em diversas áreas, como exemplo, diagnósticos médicos, sistemas de reconhecimento de caracteres e de voz. O desenvolvimento do algoritmo é baseado na divisão de uma complexa decisão em diversas outras menores que auxiliem na obtenção do resultado.

A estrutura de uma árvore de decisão é composta de forma hierárquica pela raiz (*root)*, posteriormente, pelos nós (*nodes)* e, por fim, pelas folhas (*leaf)* como pode ser visto na Figura 2. Além disso, destaca-se a profundidade (*depth)* da árvore, ou seja, o número de camadas existentes no algoritmo.

Figura 2 – Representação da árvore de decisão.



Fonte: (IBM CLOUD EDUCATION, 2020).

Os principais objetivos da árvore de decisão são:

- a classificação correta do maior número possível dos dados de treino;
- a generalização dos dados de treino, com a finalidade de classificar de forma correta os dados não vistos ainda;
- ser um algoritmo de estrutura simples, porém com a necessidade de definir a estratégia de decisão que será adotada nos nós internos da árvore.

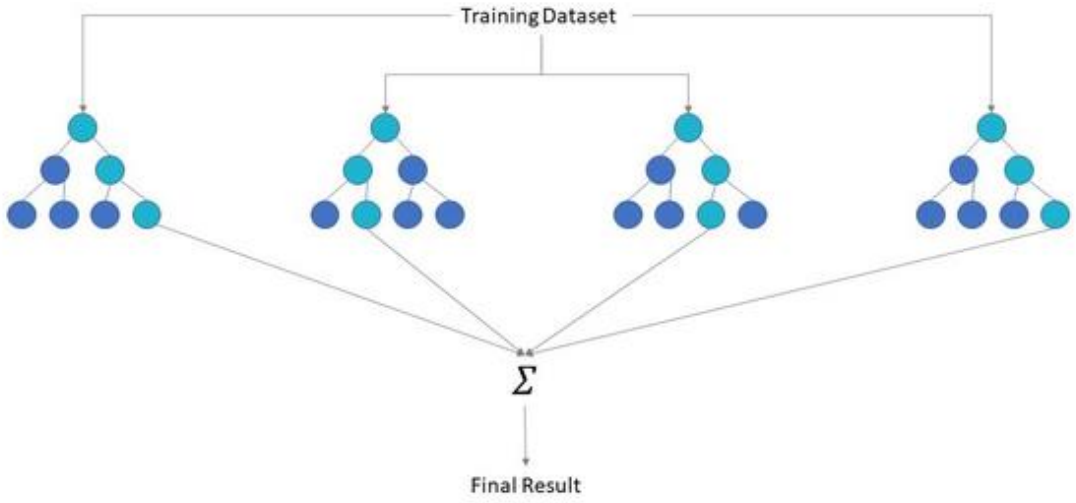
Destaca-se, também, que apesar de eficiente e de fácil compreensão, tipicamente outros algoritmos mais robustos de aprendizado de máquina costumam apresentar maior eficiência. Tal fato, deve-se pela desvantagem do algoritmo ser propenso ao *overfitting* e, assim, não generalizar bem para dados novos apresentando alta variância (IBM CLOUD EDUCATION, 2020).

2.3.2 FLORESTA ALEATÓRIA

O algoritmo de floresta aleatória, muito utilizado nos modelos de aprendizado de máquina, tem como base a utilização de um conjunto de diferentes árvores de decisão como pode ser visto na Figura 3. Essa técnica consiste em identificar o resultado mais comum entre os diversos valores encontrados em cada árvore utilizada. Dessa forma, as predições das florestas aleatórias costumam

apresentar menor variância no resultado e, assim, maior acurácia em relação ao algoritmo de árvore de decisão (IBM Cloud Education, 2020).

Figura 3 – Representação do algoritmo floresta aleatória.



Fonte: (IBM CLOUD EDUCATION, 2020).

Pelo fato de ter como base as árvores de decisão, a estrutura do algoritmo da floresta aleatória é similar ao descrito na Subseção 2.3.1 e apresenta três hiperparâmetros principais, como o número de nós, o número de características que serão utilizadas e o número de árvores de decisão (IBM CLOUD EDUCATION, 2020). Cabe salientar, ainda, alguns dos benefícios na utilização das florestas aleatórias. Primeiramente, destaca-se o fato de o algoritmo reduzir o risco de ocorrência de *overfitting* dos dados. Tal fato, deve-se pelo resultado do algoritmo ser proveniente do resultado individual de diversas árvores. Além disso, o algoritmo pode ser utilizado tanto em problemas de regressão como em problemas de classificação dos dados e, também, a facilidade de avaliar a contribuição das características dos dados para o resultado (IBM CLOUD EDUCATION, 2020).

Em contrapartida, o algoritmo é mais robusto e complexo. Assim, o tempo de execução dessa técnica de programação costumam ser maiores, fator que implica em maior custo computacional e a necessidade de mais recursos para

armazenamento e processamento dos dados, quando comparado com a árvore de decisão (IBM CLOUD EDUCATION, 2020).

2.3.3 ADABOOST

Adaptative Boost, ou AdaBoost, é um algoritmo baseado em um conjunto de árvores de decisão que utiliza o método de treinamento em *boosting*. Ou seja, o método consiste na criação de diversas árvores de decisão com apenas uma dimensão que individualmente não conseguiriam explicar os dados. Porém, o algoritmo atribui pesos maiores às decisões corretas e pesos menores às decisões incorretas. Dessa forma, a cada iteração do algoritmo ocorre a redução do erro entre os valores reais e os previstos, até que o erro apresente um valor limite aceitável (IBM CLOUD EDUCATION, 2020).

2.3.4 XGBOOST

A técnica de aprendizado de máquina denominada como *Xtreme Gradient Boost*, ou XGBoost, também é baseada na árvore de decisão e utiliza o método de treinamento classificado como *gradient boosting.* Isto significa, que o algoritmo tenta gerar resultados precisos inicialmente e o modelo busca otimizar a função de perda de forma iterativa com a finalidade de convergir o valor do erro da função para um valor mínimo e, a cada iteração, o modelo apresentar melhores resultados (IBM Cloud Education, 2020).

2.4 VALIDAÇÃO E HIPERPARÂMETROS

Para a aplicação de modelos de aprendizado de máquina, mostra-se indispensável a separação do conjunto de dados em parte para treino e outra para teste. Na etapa de treino os algoritmos aprendem padrões entre os dados com o objetivo de prever casos novos de maneira eficiente. Já na etapa de teste, os modelos realizam predições com dados ainda não vistos pela parte de treino na tentativa de classificar corretamente cada caso. Dessa forma, é possível realizar a comparação entre as previsões realizadas e as variáveis de saída no

conjunto de dados de teste e avaliar o desempenho de cada algoritmo (GOOGLE MACHINE LEARNING EDUCATION, 2022).

Entretanto, caso o treino dos dados ocorra com a presença dos dados de teste, o modelo, por já ter aprendido os padrões desses dados, apresenta forte tendência em acertar as classificações e apresentar boas métricas de desempenho. Porém, essas métricas apresentariam a falsa sensação de um bom modelo de classificação, pois não seria possível avaliar como o modelo lidaria com novos dados. O fato descrito é denominado de *overfitting* e pode ser reduzido com a separação de parte dos dados para treino, parte para validação e parte para teste (GOOGLE MACHINE LEARNING EDUCATION, 2022).

Na primeira etapa, os algoritmos são treinados apenas com os dados de treino e avaliados com a utilização dos dados destinados para validação do modelo. Posteriormente, os modelos que apresentaram melhor desempenho, devem ser novamente treinados utilizando em conjunto os dados destinados ao treino e à validação para a predição dos dados de teste. Assim, torna-se possível avaliar o desempenho do modelo duas vezes, o que dificulta a chance de *overfitting* dos dados (GOOGLE MACHINE LEARNING EDUCATION, 2022).

Cabe destacar, também, que os algoritmos de aprendizado de máquina possuem parâmetros essenciais para a criação de classificadores. Com a finalidade de construir modelos adequados para a base de dados, há a necessidade de encontrar quais os melhores valores para cada hiperparâmetro apresentado na Tabela 3. A partir do teste realizado, é possível realizar combinações entre os melhores valores de cada parâmetro e obter o melhor modelo.

Tabela 3 – Dicionário explicativo de hiperparâmetros dos algoritmos.

Modelo	**Parâmetros**	**Definição**
Árvore de Decisão	max_depth	Tamanho máximo da árvore
	min_samples_split	Número minímo de amostras para a divisão de um nó
	min_samples_leaf	Número minímo de amostras necessário na folha (*leaf node)*

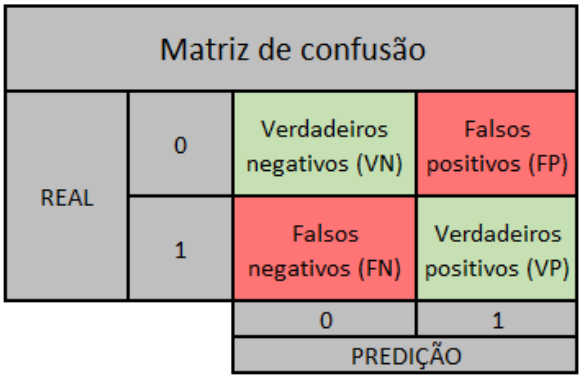
	max_features	Número máximo de características que influenciam na divisão dos nós
Floresta Aleatória	max_depth	Tamanho máximo das árvores
	min_samples_split	Número minímo de amostras para a divisão de um nó
	min_sample_leaf	Número minímo de amostras necessário na folha (*leaf node)*
	n_estimators	Número de árvores que farão parte da floresta
AdaBoost	n_estimators	Número de árvores de decisão fracas (max_depth=1)
	learning_rate	Taxa de aprendizagem do modelo
XGBoost	max_depth	Tamanho máximo da árvore
	n_estimators	Número de árvores que farão parte da floresta
	learning_rate	Taxa de aprendizagem do modelo
	subsampling	Controla o tamanho das amostras de treino

Fonte: elaborado pelo autor

2.5 MÉTRICAS DE AVALIAÇÃO

Após treinados, os modelos passam por testes de classificação que realizam a comparação entre os valores encontrados pelo algoritmo de classificação e a variável de saída. Com isso, há a possibilidade de identificar os valores verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos e, assim, extrair métricas oriundas da matriz de confusão (Figura 4) para avaliar o desempenho do modelo, como acurácia, precisão, *F1-score, recall e* ROC AUC (HARRISON, 2020).

Figura 4 – Apresentando matriz de confusão.



Fonte: elaborado pelo autor.

A matriz de confusão retrata os erros e acertos de classificação do modelo em comparação com o desfecho real dos casos de teste. A partir dessa matriz, as métricas descritas podem ser calculadas para medir a eficiência do modelo da seguinte forma:

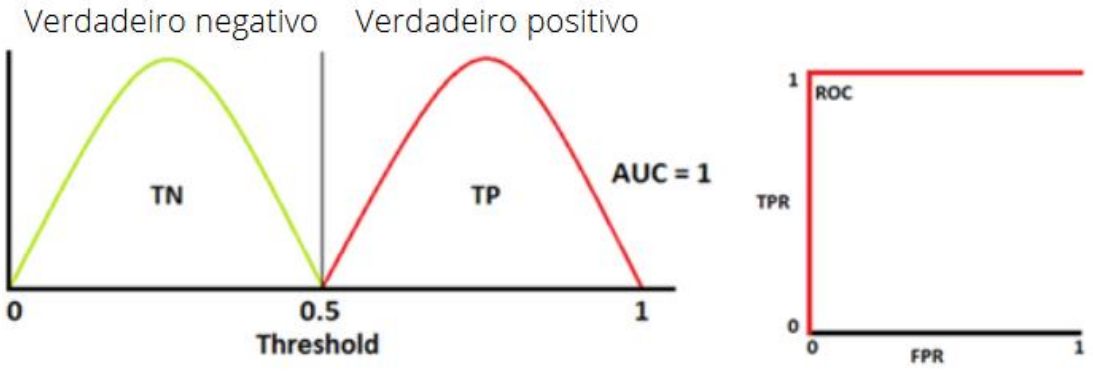
- Acurácia: $\frac{VP+VN}{VP+VN+FP+FN}$
- Precisão: $\frac{VP}{VP+FP}$
- *Recall:* $\frac{VP}{VP+FN}$
- *F1-score*: $\frac{2VP}{2VP+FP+FN}$

Vale destacar, que, em conjuntos de dados desbalanceados, a classe minoritária normalmente apresenta maior importância para o modelo que a classe majoritária. Dessa forma, algumas métricas apresentam maior relevância para a classificação do melhor preditor. Por esse motivo, costuma-se avaliar a área sob a curva ROC que indica a capacidade do classificador de distinguir entre as classes da variável de saída. O gráfico é composto pela relação entre a taxa de verdadeiros positivos (TPR*)* e a taxa de falsos positivos (FPR*)* (BROWNLEE, 2020).

- Taxa verdadeiros positivos: $\frac{VP}{VP+FP}$
- Taxa falsos positivos: $\frac{FP}{FP+VN}$

Com a Figura 5, percebe-se o caso, na qual um algoritmo de aprendizado de máquina foi capaz de distinguir perfeitamente todos os casos da variável de saída e não apresentou valores falsos positivos e nem falsos negativos. Nesse caso, a pontuação ROC AUC, representada pela área da curva ROC, apresentaria valor igual a 1, pelo fato de ter acertado 100% dos casos (NARKHEDE, 2018).

Figura 5 – Caso ideal de ROC AUC com valor 1.



Fonte: (NARKHEDE, 2018).

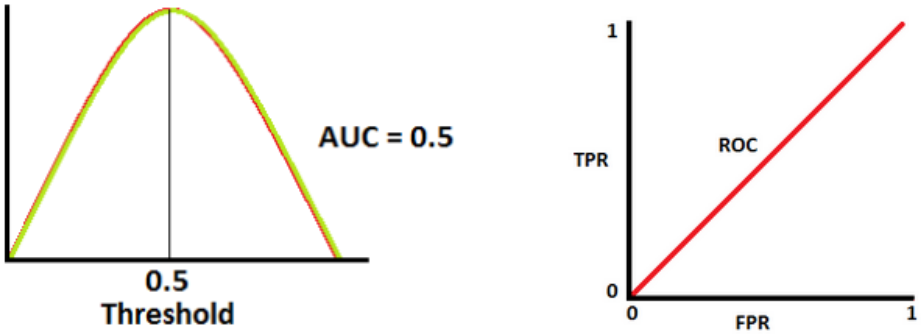
Outro exemplo, mais comum em situações reais, o classificador costuma apresentar casos falsos positivos e casos falsos negativos, com isso, a probabilidade de o modelo distinguir as classes diminui, como pode ser visto na Figura 6, cujo modelo hipotético apresentou 70% de chance de classificar uma nova amostra de maneira correta (NARKHEDE, 2018).

Figura 6 – Caso com a pontuação ROC AUC de 0,7.

Fonte: (NARKHEDE, 2018).

Por fim, a Figura 7, mostra o caso no qual o modelo apresenta 50% de chance de acertar a nova amostra, ou seja, o modelo não é considerado capaz de realizar a classificação. Isso porque em problemas de classificação binária, existe 50% de chance de acertar o resultado da variável de saída, mesmo sem a construção de um modelo de aprendizado de máquina (NARKHEDE, 2018).

Figura 7 - Caso com a pontuação ROC AUC de 0,5.



Fonte: (NARKHEDE, 2018).

Logo, pelo fato da pontuação ROC AUC avaliar a probabilidade de o modelo classificar corretamente os casos da classe majoritária e os casos da classe minoritária, a métrica costuma ser aplicada para comparação de desempenho de diferentes modelos em conjuntos de dados desbalanceados.

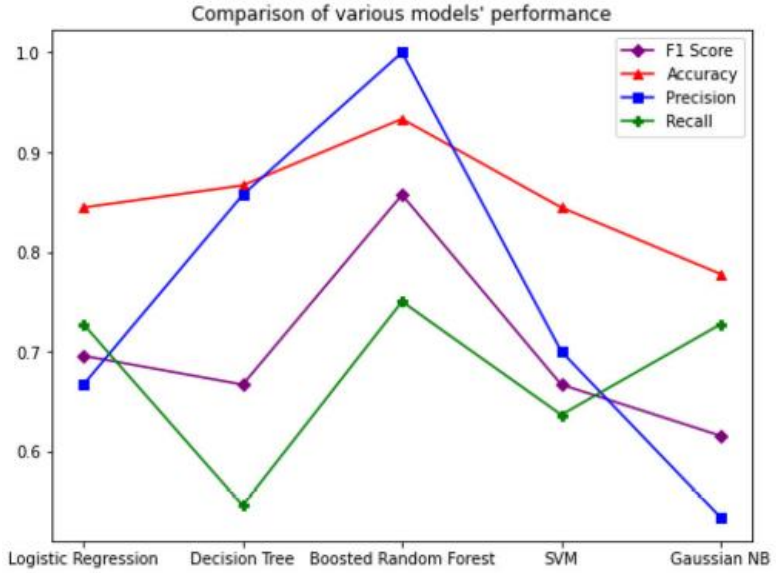
2.6 ESTADO DA ARTE

Desde o início da pandemia, diversos estudos foram realizados com a finalidade de compreender os dados coletados durante esses anos e de prever a severidade da doença em cada paciente. Com isso, alguns trabalhos serviram de referência para contextualizar a pesquisa e auxiliar durante o desenvolvimento do projeto.

Em (IWENDI, et al., 2020), uma base de dados de pacientes testados para Covid-19 na cidade chinesa de Wuhan foi utilizada para auxiliar a identificar o resultado mais provável de recuperação ou óbito do paciente. Os métodos aplicados consistem no pré-processamento dos dados, a fim de codificar variáveis categóricas e substituir valores nulos pela *string* “NA” demarcando os pacientes que apresentam valores faltantes e identificar a correlação entre as variáveis. Posteriormente, com a análise de dados, foi possível observar quais sintomas eram mais comuns nos pacientes e qual a idade e o sexo mais afetados pelo vírus. Por fim, algumas técnicas de aprendizado de máquina como regressão logística, árvore de decisão, floresta aleatória intensificada com o

algoritmo de AdaBoost, SVM, e *naive bayes* foram aplicadas a fim de obter a maior pontuação dentre as seguintes métricas de avaliação: acurácia, precisão, revocação e *F1-score.* Com o resultado melhor no modelo de floresta aleatória intensificada, houve a otimização dos parâmetros apenas para esse caso (IWENDI et al., 2020). Com a Figura 8, torna-se possível avaliar os resultados encontrados para cada métrica aplicada. Assim, pôde-se perceber que os algoritmos baseados em árvores de decisão apresentaram melhores resultados de acurácia e precisão e o método da floresta aleatória instensificada apresentou, também, um bom valor de *F1-score* (0,86), entre as métricas aplicadas, a mais importante para a classificação dos pacientes.

Figura 8 – Pontuação dos algoritmos nas métricas aplicadas.



Fonte: Iwendi et al. (2020, p. 7)

Em (YADAW et al., 2020), o objetivo foi desenvolver um modelo de predição de óbitos por Covid-19 e identificar quais características clínicas mais contribuem com o resultado. Como base do trabalho, o autor utilizou dados demográficos e clínicos dos pacientes de um determinado hospital da cidade de Nova Iorque. Para alcançar o objetivo inicial, apenas pacientes infectados pelo vírus foram selecionados, dados faltantes foram eliminados e, para selecionar características de maior relevância, utilizou-se um algoritmo de seleção de características recursivo a fim de reduzir a dimensionalidade dos dados. Após o pré-processamento, os modelos de aprendizado de máquina como Regressão

Logística, SVM, Floresta aleatória e XGBoost foram aplicados em dois conjuntos de dados de teste. O conjunto de dados, para evitar o *overfitting,* foi dividido em treino, validação e teste do modelo. Durante a etapa de treino, pôde-se perceber que o algoritmo mais eficiente na base de dados foi o XGBoost e que a partir de três principais características o resultado era muito similar ao usar todas as 17 características presentes no trabalho. Por fim, a pontuação na métrica ROC-AUC foi entre 0,91 e 0,94 com a utilização de um intervalo de confiança de 95% (YADAW et al., 2020). Nesse caso, percebe-se que o algoritmo XGBoost apresentou o maior valor de ROC AUC e, além disso, com a seleção de características foi possível reduzir a dimensionalidade dos dados de 17 variáveis para 3 e, ainda sim, apresentar resultado satisfatório, classificando mais que 90% dos casos de maneira correta.

Em (RODRIGUES, 2021), o trabalho foi realizado com dados do estado brasileiro Rio Grande do Sul. O trabalho teve como objetivo criar modelo preditivo baseado em florestas aleatórias para classificação do risco de óbito em pacientes infectados pela Covid-19. O pré-processamento dos dados demográficos, clínicos e temporais teve como base a eliminação de valores nulos e a codificação de variáveis categóricas. Após os ajustes iniciais dos dados, o autor se propôs a realizar uma análise descritiva para encontrar as características que mais influenciavam no risco de óbito dos pacientes a partir de técnicas como a correlação de Pearson e o grau de impureza de Gini. Para a parte de aprendizado de máquina, utilizou-se o algoritmo de floresta aleatória com a utilização da técnica de *undersampling* e para validação do modelo utilizou-se, também, a métrica ROC-AUC e obteve-se a pontuação 0,981, configurando-se como um bom modelo de classificação.

Em (L WANG et al., 2022), utilizou-se dados de pacientes que testaram positivo para o SARS-Cov-2 e permaneceram em acompanhamento no hospital *Houston Methodist at Texas Medical Center* os dados coletados são referentes ao período de 26 de fevereiro de 2020 até a data de junho de 2020. No total, a amostra para o estudo foi de 1027 pacientes com 19 características a respeito de comorbidades já existentes nos pacientes, 8 características demográficas, 15 complicações (condições existentes por complicações provenientes do vírus,

dentre elas a necessidade de internação do paciente) e alguns sinais vitais dos pacientes (como exemplo a sadutação do oxigênio – Sp02). Com isso, estudou-se a significância que as diversas características apresentavam no prognóstico da doença por meio de análises de correlação. Os modelos de aprendizado de máquina foram testados de maneira iterativa, ou seja, inicialmente o modelo utilizou dados entre fevereiro e abril de 2020 para treino e o teste ocorreu com os dados entre 10 de abril e 16 de abril, em seguida, esses dados de teste foram incluídos no conjunto de treino e o teste ocorreu com os dados entre 17 de abril e 23 de abril. Esse processo ocorreu 8 vezes até incluir os últimos dados coletados do dia 3 de junho de 2020. Por fim, utilizou-se o algoritmo de aprendizado de máquina *Lasso Logistic Regression* para 5 conjuntos de dados diferentes e para analisar o desempenho de cada modelo utilizou-se a métrica ROC AUC. Os resultados encontrados foram:

1. Modelo 1: utilização apenas dos dados demográficos (AUC = 77,6%)
2. Modelo 2: Modelo 1 + comorbidades mais significantes (AUC = 79,2%)
3. Modelo 3: Modelo 2 + complicações mais significantes (AUC = 90,7%)
4. Modelo 4: Modelo 3 + sinais vitais do paciente (AUC = 91,6%)
5. Modelo 5: Modelo 4 + parâmetro a quantidade de dias que o paciente apresentou o nível de saturação do oxigênio ruim (AUC = 94,8%)

Percebe-se, então, que a inserção das complicações dos pacientes no modelo apresentou grande impacto para a variável de saída, apresentando um aumento significativo na pontuação ROC AUC (de 79,2% para 90,7%). Dessa forma, os Modelos 3,4 e 5 apresentaram resultados satisfatórios no prognóstico da covid-19, fator que poderia auxiliar a admissão de pacientes identificados como estado crítico na *ICU* (UTI) para tratamento prioritário da doença.

Com isso, destaca-se a relevância do assunto em diversas regiões do mundo e que a aplicação de diversas técnicas de pré-processamento e de análise de dados auxiliam no entendimento do vírus. Além disso, percebe-se que a elaboração de diferentes modelos de aprendizado de máquina facilita a comparação dos resultados para selecionar qual algoritmo apresentou melhor classificação. Por fim, pôde-se observar que nos estudos realizados, o

desempenho da maioria dos modelos foi baseado na pontuação ROC AUC e apresentaram resultados satisfatórios de classificação para o prognóstico da covid-19.

3 MÉTODO

3.1 LINGUAGEM E AMBIENTE DE PROGRAMAÇÃO

Para a realização de todas as etapas de pré-processamento, análise de dados, visualização de dados e elaboração dos modelos de aprendizado de máquina, a linguagem de programação escolhida foi o *Python*. Tal fato, deve-se pelas inúmeras bibliotecas presentes na linguagem que facilitam a execução dos processos descritos. Além disso, o ambiente de programação utilizado foi o *Google Colaboratory,* pela razão das bibliotecas que foram utilizadas durante o trabalho já estarem previamente disponíveis na plataforma.

3.2 BASE DE DADOS

A escolha da base de dados do munícipio de Vitória – ES, justifica-se, não somente por ser a região que se localiza a Instituição de ensino, mas, também, pelo fato do estado do Espírito Santo se apresentar como exemplo quando se trata de *open data*. Ou seja, os dados a respeito da Covid-19 são apresentados com transparência e encontram-se disponíveis e atualizados no portal de transparência do estado (ANDRADE, 2020). Cabe salientar, ainda, o fato de os dados apresentarem um dicionário explicativo, conforme apresentado na Figura 9, sobre cada característica abordada para cada caso registrado, desde os dados do perfil do cidadão, quanto os sintomas, as comorbidades e as informações a respeito dos testes para confirmação do SARS-CoV-2. Além disso, a Secretaria de Saúde, órgão responsável pela coleta dos dados, disponibiliza a metodologia utilizada para definição de alguns parâmetros que possam gerar dúvidas, como a definição de um caso suspeito de infecção pelo vírus (SECRETÁRIA DE SAÚDE DO ESTADO DO ESPÍRITO SANTO, 2021). Logo, a escolha da base de dados foi feita em razão da transparência, qualidade e quantidade de características descritas para cada caso, fator primordial para entender como a Covid-19 afetou a população de Vitória por diversas perspectivas.

Figura 9 – Dicionário dos dados da Covid-19 em Vitória.

Coluna	Descrição
Bairro	Bairro do cidadão
Classificacao	Classificação do caso: Suspeito, Confirmados, Descartados
ComorbidadeCardio	Comorbidade
ComorbidadeDiabetes	Comorbidade
ComorbidadeObesidade	Comorbidade
ComorbidadePulmao	Comorbidade
ComorbidadeRenal	Comorbidade
ComorbidadeTabagismo	Comorbidade
CriterioConfirmacao	Critério utilizado para confirmar o caso
DataCadastro	Data de cadastro no sistema
DataNotificacao	Data da notificação do agravo
DataColeta_RT_PCR	Data da realização da coleto do exame
DataColetaSorologia	Data da realização da coleto do exame
DataColetaSorologiaIGG	Data da realização da coleto do exame
DataColetaTesteRapido	Data da realização da coleto do exame
DataDiagnostico	Data do diagnóstico
DataEncerramento	Data do encerramento
DataObito	Data do óbito
Evolucao	Evolução do caso
IdadeNaDataNotificacao	Idade do cidadão na data da notificação
FicouInternado	Identifica se ficou internado
Municipio	Município do cidadão
Escolaridade	Perfil do cidadão: escolaridade
FaixaEtaria	Perfil do cidadão: Faixa etaria
RacaCor	Perfil do cidadão: Raça/Cor
ProfissionalSaude	Perfil do cidadão: Se é profissional da saúde
MoradorDeRua	Perfil do cidadão: Se é um morador de rua
PossuiDeficiencia	Perfil do cidadão: Se possui deficiência
Sexo	Perfil do cidadão: Sexo
ViagemInternacional	Realizou alguma viagem internacional
ViagemBrasil	Realizou alguma viagem no Brasil
Cefaleia	Sintoma
Coriza	Sintoma
Diarreia	Sintoma
DificuldadeRespiratoria	Sintoma
DorGarganta	Sintoma
Febre	Sintoma
Tosse	Sintoma
StatusNotificacao	Situação da notificação: Em Aberto, Encerrado
ResultadoRT_PCR	Teste: Identifica o resultado do teste RT_PCR
ResultadoSorologia	Teste: Identifica o resultado do teste Sorologia
ResultadoSorologia_IGG	Teste: Identifica o resultado do teste Sorologia_IGG
ResultadoTesteRapido	Teste: Identifica o resultado do teste Teste Rapido
AmostraRT_PCR	Teste: Identifica se realizou o teste RT_PCR
AmostraSorologia	Teste: Identifica se realizou o teste Sorologia
AmostraSorologia_IGG	Teste: Identifica se realizou o teste Sorologia_IGG
AmostraTesteRapido	Teste: Identifica se realizou o teste Teste Rapido

Fonte: SECRETARIA DE SAÚDE DO ESTADO DO ESPÍRITO SANTO, 2020

O banco de dados foi dividido em três partes: uma parte para treino, outra para validação e a última para teste e avaliação do desempenho do modelo. Essa divisão é uma forma de evitar que a validação e teste dos modelos sejam realizados com os mesmos dados vistos pela parte de treino do algoritmo e,

assim, evitar o *overfitting*, ou seja, um ajuste muito adaptado aos dados de treino que apresentaria predições muito otimistas (COLLINS *et al.,* 2015).

Como forma de simular o desempenho dos modelos de aprendizado de máquina na prática, a divisão dos dados foi realizada de forma cronológica. Dessa forma, considerou-se que apenas os casos de 2020 e 2021 aconteceram e, por isso, esses dados foram destinados para treinar os algoritmos. Posteriormente, os dados de janeiro de 2022 ocorreram e os modelos realizaram as classificações dos pacientes como forma de validar o desempenho dos modelos para esses casos. Por fim, pôde-se utilizar os casos de treino (2020 e 2021) e os de validação (janeiro 2022) para treinar novamente os algoritmos e prever como os modelos classificariam os novos pacientes entre fevereiro de 2022 e 26 de junho de 2022 para testar e avaliar o desempenho final do projeto.

3.3 PARTICIPANTES

Com o arquivo disponibilizado pelo Estado, foi possível realizar a leitura do banco de dados armazenado desde o dia 27/02/2020 até a data de 26/06/2022, no qual 467.196 casos de pacientes anônimos foram registrados a partir de 45 tipos diferentes de características iniciais. Como o objetivo geral do trabalho é realizar o prognóstico da Covid-19 em futuros pacientes da doença no município de Vitória, os casos foram restringidos para apenas os 121.257 de casos de pacientes confirmados com a doença. Eliminando, assim, as amostras dos pacientes com suspeita da doença ou que testaram negativo.

3.3.1 ANÁLISE E PRÉ PROCESSAMENTO DOS DADOS

A primeira parte da etapa de pré-processamento dos dados foi a criação de duas novas variáveis para o modelo, com base na data de diagnóstico do paciente para dividir de maneira temporal a ocorrência de cada caso. Assim, as variáveis “Mês” e “Semestre” foram acrescentadas ao conjunto de dados para facilitar na divisão dos dados de treino, validação e teste, e apresentam valores como “2020M1” para o primeiro mês do ano de 2020, e “2020S1” no caso do primeiro semestre do ano de 2020.

Feito isso, iniciou-se o processo de análise dos dados de treino para identificar valores ausentes ou inconsistentes do modelo. Verificou-se, então, que diversas variáveis apresentavam todos os valores preenchidos, porém, parte desses dados apresentavam como valor a palavra “Ignorado” ou “Não Informado”. Dessa forma, houve a substituição dos valores ignorados para “NaN” que, na linguagem *Python*, representa a ausência de um número.

Na busca de casos inconsistentes, foram encontrados 420 casos confirmados de maneira laboratorial, mas que não apresentavam nenhuma coleta de exame com resultado positivo, 197 casos de pacientes com todas as comorbidades e todos os sintomas ignorados, 225 casos em aberto que não apresentavam o prognóstico do paciente e 2 casos confirmados por Covid-19 em janeiro de 2020, mês que não apresentou casos confirmados no Brasil. Por esses motivos, tais pacientes foram removidos no banco de dados, resultando em 73.363 dados para treino.

Após a remoção desses casos, iniciou-se a avaliação individual das variáveis que apresentavam valores ausentes. A princípio, pôde-se perceber que a maioria das datas de realização dos exames e os resultados dos mesmos como, “DataColeta_RT_PCR” e “ResultadoRT_PCR”, “DataColetaTesteRapido” e “ResultadoTesteRapido”, “DataColetaSorologia” e “ResultadoSorologia”, “DataColetaSorologiaIGG” e “ResultadoSorologia_IGG”, apresentavam de forma majoritária dados ausentes e, dessa forma, não agregam muita informação ao classificador. Cabe salientar, ainda, que casos confirmados com os resultados dos exames negativos foram removidos e, assim, todos os casos restantes apresentam confirmação por algum exame. Com isso, pelo fato de possuírem poucos registros, essas características não agregam informações suficientes que influenciem no prognóstico do paciente, por isso, removeu-se do conjunto de dados.

Além disso, mais de 90% dos pacientes apresentaram características como “ViagemBrasil”, “ViagemInternacional”, “ProfissionalSaude”, “MoradorDeRua” e “Gestante” com o valor negativo (“Não”) e percebeu-se, também, que

independente do paciente ter apresentado ou não essas características, a taxa de sobrevivência foi similar e superiores a 97% (conforme a Tabela 4). Dessa forma, concluiu-se que essas características não influenciavam para o prognóstico da covid-19.

Tabela 4 – Proporção de sobreviventes e óbitos para pacientes que apresentavam ou não as características selecionadas

Características	Valor	Total de Casos	Sobreviveu Covid	Casos Registrados	%
Morador de rua	Não	72135	Sobreviveu	70823	98,18%
			Óbito	1312	1,82%
	Sim	85	Sobreviveu	83	97,65%
			Óbito	2	2,35%
Profissional da Saúde	Não	58227	Sobreviveu	56948	97,80%
			Óbito	1279	2,20%
	Sim	5615	Sobreviveu	5591	99,57%
			Óbito	24	0,43%
Viagem brasil	Não	45340	Sobreviveu	44063	97,18%
			Óbito	1277	2,82%
	Sim	1584	Sobreviveu	1576	99,49%
			Óbito	8	0,51%
Viagem Internacional	Não	45340	Sobreviveu	44049	97,17%
			Óbito	1282	2,83%
	Sim	1584	Sobreviveu	57	100%
			Óbito	0	0%

Fonte: Elaborado pelo autor.

Vale destacar, também, a variável “Escolaridade” que além de apresentar a maioria dos dados ausentes (44%), os valores restantes mostravam-se de forma inconsistente (Tabela 5) e a variável “Gestante” que apresentou apenas 206 casos registrados de 73.421 amostras de treino (Tabela 6). Sendo assim, as duas características não pareceram relevantes para o prognóstico da doença.

Tabela 5 – Amostra dos casos por grau de escolaridade.

Escolaridade	Amostras registradas
Educação superior completa	12724
Ensino médio completo (antigo colegial ou 2º grau)	12506
5ª à 8ª série incompleta do EF (antigo ginásio ou 1º grau)	2910
Educação superior incompleta	2789
Ensino médio incompleto (antigo colegial ou 2º grau)	2714
Não se aplica	2542
Ensino fundamental completo (antigo ginásio ou 1º grau)	2231
1ª a 4ª série incompleta do EF (antigo primário ou 1º grau)	1418
4ª série completa do EF (antigo primário ou 1º grau)	928
Analfabeto	494

Fonte: Elaborado pelo autor.

Tabela 6 – Amostras dos casos de gestantes.

Gestante	Amostras registradas
Não	39860
Não se aplica	33355
3º trimeste	97
2º trimeste	50
1º trimeste	43
Idade gestacional ignorada	16

Fonte: Elaborado pelo autor.

Outras variáveis, que também foram retiradas do conjunto de dados foram o “Município”, que apresentava apenas um valor possível para todos os casos (“Vitoria”) e a “IdadeNaDataNotificacao” por apresentar tipo de texto (*string)* com a idade exata do paciente no formato (‘Idade, mês e dia’), o que gerou 24.784

rótulos diferentes para os dados de treino e, dessa forma, optou-se por priorizar a variável “FaixaEtaria” que facilita a padronização dos dados.

Em seguida, houve a necessidade de preencher os dados ausentes de alguns casos registrados, dentre esses, apenas 6 casos não apresentavam a “FaixaEtaria” e foram preenchidos com o valor mais frequente (“30 a 39 anos”). Já para os casos que indicam se o paciente necessitou de internação (“FicouInternado”), verificou-se que os valores (‘Ignorado’ e ‘Não Informado’) apresentavam índices de sobrevivência superiores a 99,86%. Com relação aos casos de pacientes que não ficaram internados, a porcentagem de sobrevivência foi de 99,64%, ou seja, a taxa de sobrevivência foi muito similar se comparada com os casos ‘Ignorado’ e ‘Não Informado’. Por sua vez, os pacientes que necessitaram de internação apresentaram índice de sobrevivência de apenas 37,28% (Tabela 7). Logo, pôde-se concluir que os pacientes com a informação ausente, pelo elevado nível de sobrevivência, dificilmente foram casos de internação e, assim, os valores faltosos foram preenchidos como (‘Não’), conforme a Tabela 8.

Tabela 7 – Taxa de sobrevivência para cada caso da característica “FicouInternado”.

FicouInternado	Total de Casos	SobreviveuCovid	Casos Registrados	%
Ignorado	5706	Sobreviveu	5698	99,86%
		Óbito	8	0,14%
Não	42966	Sobreviveu	42810	99,64%
		Óbito	156	0,36%
Não Informado	23583	Sobreviveu	23567	99,93%
		Óbito	16	0,07%
Sim	1808	Sobreviveu	674	37,28%
		Óbito	1134	62,72%

Tabela 8 – Taxa de sobrevivência após processamento da característica “FicouInternado”.

FicouInternado	Total de Casos	SobreviveuCovid	Casos Registrados	%
Não	72255	Sobreviveu	72075	99,75%
		Óbito	180	0,25%
Sim	1808	Sobreviveu	674	37,28%
		Óbito	1134	62,72%

Por fim, com o conjunto de dados resultante, aplicou-se um codificador *label encoder* que é responsável pela representação de valores categóricos por valores numéricos. Ou seja, características que apresentam valores binários como “Não” e “Sim”, são codificadas para 0 e 1, tal processo, mostra-se essencial para a elaboração dos modelos de aprendizado de máquina que utilizam de processos matemáticos e necessitam de variáveis numéricas para identificarem padrões. Com a transformação dos dados em numéricos, retirou-se, também, a variável “Bairro” pelo fato de apresentar 80 valores diferentes e, dessa forma, dois pacientes com os mesmos agravantes poderiam ter chances muito discrepantes de sobrevivência apenas pelo fato de residir em bairros distintos.

3.4 VARIÁVEL DE SAÍDA

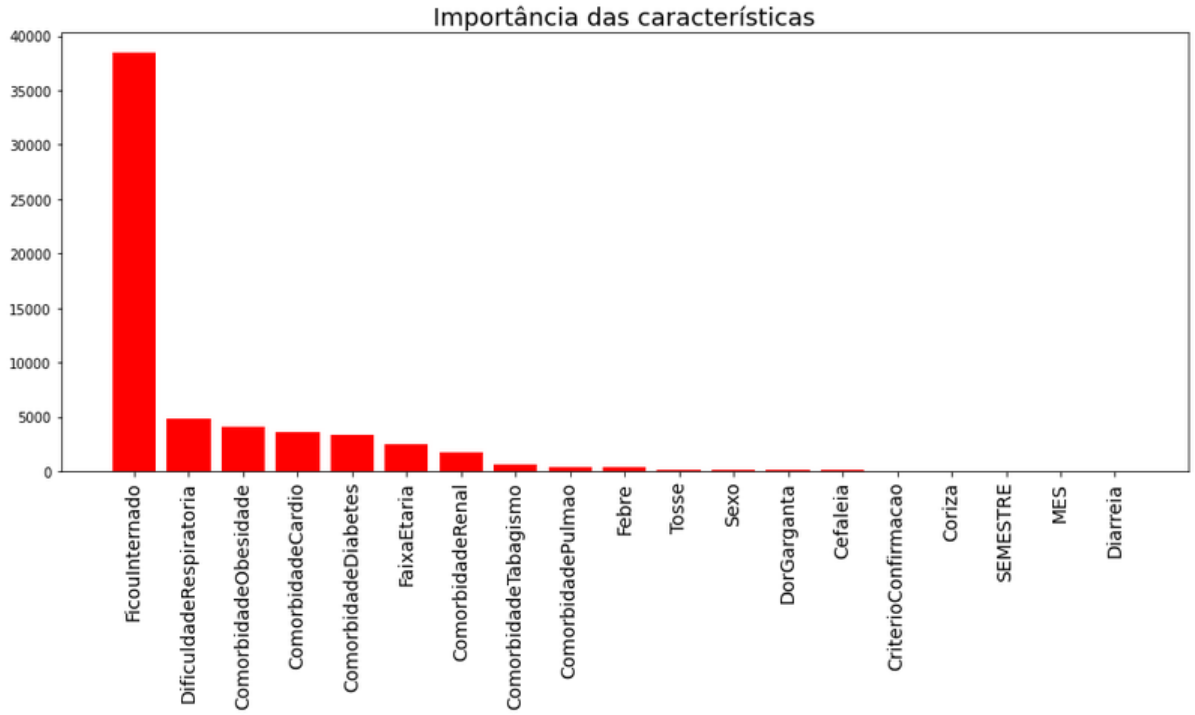
Inicialmente, é imperioso destacar que no conjunto de dados havia a variável “Evolução”, que possuía os seguintes valores: “cura”, “-“, “óbito pelo COVID-19”, “ignorado”, “óbito por outras causas”. Considerando que o presente trabalho busca analisar apenas a sobrevivência ou óbito pela COVID-19, foi necessário criar uma nova variável, a variável de saída “SobreviveuCovid”, na qual foram considerados apenas os valores referentes a “cura” ou “óbito pelo COVID-19”.

Ou seja, a variável “Evolução” foi excluída, com a finalidade de que os valores de “-“, “ignorado” e “óbito por outras causas”, fossem desconsiderados no conjunto de dados final, visto que em nada agregavam ao objetivo final do estudo.

3.5 SELEÇÃO DE CARACTERÍSTICAS

Após a etapa de limpeza dos dados inconsistentes e das características que apresentavam muitos valores faltosos, as variáveis foram submetidas ao teste matemático do χ^2 para analisar quais apresentavam maior relevância para a variável de saída do modelo, ou seja, quais que, de maneira individual, conseguem explicar melhor as chances de o paciente sobreviver. A Figura 14, mostra de forma gráfica a importância de cada característica do conjunto para a previsão de sobrevivência do paciente, dessa forma, pode-se perceber que, a sobrevivência do paciente está muito relacionada a necessidade de internação.

Figura 10 – Importância das características.



Fonte: elaborado pelo autor.

Com isso, percebe-se que características como “Diarreia”, “Mês”, “Semestre”, “Coriza” e “CriterioConfirmacao”, apresentaram pouco peso sobre o resultado de

o paciente sobreviver ou não ao Covid-19. Dessa forma, eliminou-se essas variáveis, com exceção da variável “Mês” que permaneceu no conjunto de dados com o objetivo de manter uma marca temporal dos casos.

Ao fim dessa etapa, 16 características foram selecionadas para o desenvolvimento dos modelos, como pode ser visualizado na Tabela 9. Também é apresentado um dicionário representando os valores possíveis para cada variável, após a transformação realizada pelo *label encoder.*

Tabela 9 – atributos utilizados para construção do modelo.

CARACTERÍSTICA	VALORES TRANSFORMADOS
Ficou internado	{'Não': 0, 'Sim': 1}
Faixa etária	{'0 a 4 anos': 0, '05 a 9 anos': 1, '10 a 19 anos': 2, '20 a 29 anos': 3, '30 a 39 anos': 4, '40 a 49 anos': 5, '50 a 59 anos': 6, '60 a 69 anos': 7, '70 a 79 anos': 8, '80 a 89 anos': 9, '90 anos ou mais': 10}
Dificuldade respiratória	{'Não': 0, 'Sim': 1}
Comorbidade cardíaca	{'Não': 0, 'Sim': 1}
Comorbidade diabetes	{'Não': 0, 'Sim': 1}
Obesidade	{'Não': 0, 'Sim': 1}
Febre	{'Não': 0, 'Sim': 1}
Sexo	{'Feminino': 0, 'Indefinido': 1, 'Masculino': 2}
Tosse	{'Não': 0, 'Sim': 1}
Cefaleia	{'Não': 0, 'Sim': 1}
Dor de garganta	{'Não': 0, 'Sim': 1}
Mês	{'2020M10': 0, '2020M11': 1, '2020M12': 2, '2020M3': 3, '2020M4': 4, '2020M5': 5, '2020M6': 6, '2020M7': 7, '2020M8': 8, '2020M9': 9, '2021M1': 10, '2021M10': 11, '2021M11': 12, '2021M12': 13, '2021M2': 14, '2021M3': 15, '2021M4': 16, '2021M5': 17, '2021M6': 18, '2021M7': 19, '2021M8': 20, '2021M9': 21,

	'2022M1': 22, '2022M2': 23, '2022M3': 24, '2022M4': 25, '2022M5': 26, '2022M6': 27}
Comorbidade Pulmonar	{'Não': 0, 'Sim': 1}
Comorbidade Tabagismo	{'Não': 0, 'Sim': 1}
Comorbidade Renal	{'Não': 0, 'Sim': 1}
Possui Deficiência	{'Não': 0, 'Sim': 1}

3.6 BALANCEAMENTO DOS DADOS E BUSCA DOS MELHORES HIPERPARÂMETROS PARA OS DADOS DE VALIDAÇÃO

Com o fim da etapa de pré-processamento dos dados e com melhor entendimento de como as variáveis afetaram a população do município de Vitória durante o período, iniciou-se a etapa da criação dos modelos de aprendizado de máquina. Para facilitar a execução dos modelos, criou-se uma função que recebe, o algoritmo que será utilizado, os dados destinados para treino do modelo, os dados destinados para predição e a função tem como saída os resultados das métricas do algoritmo e a representação gráfica da matriz de confusão.

Inicialmente, destaca-se, que a variável resposta é desbalanceada, ou seja, a variável “SobreviveuCovid” apresenta muito mais casos de sobrevivência que casos de óbito, como visto na Figura 11. Dessa forma, o caso em questão encaixa-se como anomalia e necessita da utilização de técnicas de balanceamento para a etapa de treino dos algoritmos de aprendizado de máquina.

Figura 11 – Quantidade de casos da variável saída.

Fonte: elaborado pelo autor.

Por isso, nessa etapa, apenas os dados de treino foram balanceados com a utilização da técnica de *undersampling*, fator que reduziu o número de amostras para 2.616, sendo 1.308 casos de óbito e 1.308 casos de sobreviventes. Assim, cada algoritmo de aprendizado de máquina proposto (árvore de decisão, floresta aleatória, AdaBoost e XGBoost) foi apenas treinado com os dados balanceados, de modo que, na etapa de validação, os dados utilizados estavam desbalanceados.

Além disso, nessa etapa de validação, os hiperparâmetros de cada um dos algoritmos foram testados de maneira individual e analisou-se quais apresentavam melhor desempenho na pontuação ROC AUC. Com isso, realizou-se a combinação dos valores encontrados para cada hiperparâmetro, com a finalidade de visualizar qual conjunto de valores proporcionaria o melhor desempenho do modelo.

Posteriormente, o mesmo processo foi realizado, porém os dados de treino foram balanceados com a técnica de *oversampling* denominada SMOTE, gerando um outro conjunto de dados para treino de 144.110 amostras com 72.055 casos de óbito e 72.055 casos de sobreviventes. Após esses processos, foi possível selecionar o melhor modelo de cada algoritmo com a utilização das duas técnicas distintas de balanceamento dos dados.

Com o intuito de encontrar em qual técnica de balanceamento cada algoritmo apresentou o melhor desempenho, estabeleceu-se a mesma semente (100), número que controla divisões aleatórias realizadas pela biblioteca do Scikit-Learn (PEDREGOSA et al., 2011), para garantir a mesma base comparativa na implementação dos modelos. Assim, os algoritmos de árvore de decisão, de floresta aleatória, de AdaBoost e de XGBoost foram implementados no conjunto de dados destinado para treino (casos registrados em 2020 e 2021) e testados com o conjunto de dados destinado para validação do modelo (casos registrados no mês de janeiro de 2022). Logo, os dados previamente estipulados como “Teste” (casos registrados no ano de 2022, com exceção do mês de janeiro), não foram utilizados nesse experimento.

3.6.1 ÁRVORE DE DECISÃO

O primeiro algoritmo a ser utilizado foi o de árvore de decisão. Os hiperparâmetros (*max_depth, min_samples_split, min_samples_leaf e max_features*) foram testados de maneira individual com os intervalos apresentados na Tabela 10, ou seja, primeiramente o modelo utilizou apenas o parâmetro de *max_depth* com valores de 1 a 100 e armazenou os resultados de ROC AUC que o modelo produziu.

Em seguida, dentre os valores armazenados, foram selecionados os 5 que obtiveram melhor desempenho. Após isso, repetiu-se o processo para os outros 3 hiperparâmetros.

Tabela 10 – Intervalos testados para cada hiperparâmetro no algoritmo árvore de decisão.

Modelo	Balanceamento	Hiperparâmetros	Intervalos testados
Árvore de decisao	Undersampling	max_depth	(1 a 100)
		min_samples split:	(2 a 25)
		min_samples_leaf	(1 a 100)
		max_features	(1 a 16)
Árvore de decisao	SMOTE	max_depth	(1 a 100)
		min_samples split:	(2 a 25)
		min_samples_leaf	(1 a 100)
		max_features	(1 a 16)

Fonte: elaborado pelo autor.

Em seguida, realizou-se, em cada técnica de balanceamento *(undersampling e SMOTE)* a combinação entre os 5 melhores valores de cada hiperparâmetro. Ou seja, foram realizadas 625 combinações em cada técnica, a fim de encontrar o modelo que apresentasse melhor desempenho

3.6.2 FLORESTA ALEATÓRIA

O segundo algoritmo a ser testado foi o classificador por floresta aleatória. Os hiperparâmetros (*max_depth, min_samples_split, min_samples_leaf, max_features e n_estimators*) foram testados de maneira individual com os intervalos apresentados na Tabela 11, ou seja, primeiramente o modelo utilizou apenas o parâmetro de *max_depth* com valores de 1 a 100 e armazenou os resultados de ROC AUC que o modelo produziu.

Em seguida, dentre os valores armazenados, foram selecionados os 5 que obtiveram melhor desempenho. Após isso, repetiu-se o processo para os outros 4 hiperparâmetros.

Tabela 11 – Intervalos testados para cada hiperparâmetro no algoritmo floresta aleatória.

Modelo	Balanceamento	Hiperparâmetros	Intervalos testados
Floresta Aleatória	Undersampling	max_depth	(1 a 100)
		min_samples split:	(2 a 100)
		min_samples_leaf	(1 a 100)
		max_features	(1 a 16)
		n_estimators	(50 a 150)
Floresta Aleatória	SMOTE	max_depth	(2 a 25)
		min_samples split:	(2 a 100)
		min_samples_leaf	(1 a 100)
		max_features	(1 a 16)
		n_estimators	(50 a 150)

Fonte: elaborado pelo autor.

Em seguida, realizou-se, em cada técnica de balanceamento *(Undersampling e SMOTE)* a combinação entre os 5 melhores valores de cada hiperparâmetro. Ou seja, foram realizadas 3.125 combinações em cada técnica, a fim de encontrar o modelo que apresentasse melhor desempenho

3.6.3 ADABOOST

O terceiro algoritmo a ser utilizado foi o AdaBoost. Os hiperparâmetros (*learning_rate e n_estimators*) foram testados de maneira individual com os intervalos apresentados na Tabela 12, ou seja, primeiramente o modelo utilizou apenas o parâmetro de *learning_rate* com valores de 0,1 a 1 e armazenou os resultados de ROC AUC que o modelo produziu.

Em seguida, dentre os valores armazenados, foram selecionados os 5 que obtiveram melhor desempenho. Após isso, repetiu-se o processo para o outro hiperparâmetro (*n_estimators)*.

Tabela 12 – Intervalos testados para cada hiperparâmetro no algoritmo AdaBoost.

Modelo	Balanceamento	Hiperparâmetros	Intervalos
AdaBoost	Undersampling	learning_rate	(0.1 a 1 a cada 0.1)

		n_estimators	(1 a 100)
AdaBoost	SMOTE	learning_rate	(0.1 a 1 a cada 0.1)
		n_estimators	(1 a 100)

Em seguida, realizou-se, em cada técnica de balanceamento *(Undersampling e SMOTE)* a combinação entre os 5 melhores valores de cada hiperparâmetro. Ou seja, foram realizadas 25 combinações em cada técnica, a fim de encontrar o modelo que apresentasse melhor desempenho.

3.6.4 XGBOOST

Por fim, utilizou-se o algoritmo de XGBoost. Os hiperparâmetros (*learning_rate, n_estimators, max_depth e subsampling*) foram testados de maneira individual com os intervalos apresentados na Tabela 13, ou seja, primeiramente o modelo utilizou apenas o parâmetro de *learning_rate* com valores de 0,1 a 1 e armazenou os resultados de ROC AUC que o modelo produziu.

Em seguida, dentre os valores armazenados, foram selecionados os 5 que obtiveram melhor desempenho. Após isso, repetiu-se o processo para os outros 3 hiperparâmetros.

Tabela 13 – Intervalos testados para cada hiperparâmetro no algoritmo XGBoost.

Modelo	Balanceamento	Hiperparâmetros	Intervalos
XGBoost	Undersampling	learning_rate	(0.1 a 1 a cada 0.1)
		n_estimators	(1 a 100)
		max_depth	(1 a 10)
		subsampling	(0.5 a 1)
XGBoost	SMOTE	learning_rate	(0.1 a 1 a cada 0.1)
		n_estimators	(1 a 100)
		max_depth	(1 a 10)
		subsampling	(0.5 a 1)

Fonte: elaborado pelo autor.

Em seguida, realizou-se, em cada técnica de balanceamento *(Undersampling e SMOTE)* a combinação entre os 5 melhores valores de cada hiperparâmetro. Ou

seja, foram realizadas 625 combinações em cada técnica, a fim de encontrar o modelo que apresentasse melhor desempenho.

3.7 MÉTRICAS DE DESEMEPENHO DOS MELHORES MODELOS DE VALIDAÇÃO NA PREDIÇÃO DOS DADOS DE TESTE.

Após encontrar os melhores hiperparâmetros e a melhor técnica de balanceamento dos dados para cada algoritmo durante o Experimento 1, os modelos foram novamente treinados, porém os dados de validação que foram utilizados para teste na etapa anterior, foram acrescentados como parte dos dados de treinamento. Com isso, os dados de treino balanceados com a técnica *undersampling* passaram a contar com 2.750 amostras com 1375 sobreviventes e 1375 óbitos e os dados de treino balanceados com a técnica *SMOTE* foram 206.748 amostras com 103.374 casos de sobreviventes e 103.374 casos de óbito.

Em seguida, os modelos que tiveram melhor desempenho com a base de validação, foram testados com a base de dados destinada para o teste. Dessa forma, foi possível comparar as predições geradas pelos classificadores com o rótulo de saída real do paciente e, assim, quantificar o número de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos com a representação da matriz de confusão. Com isso, determinou-se os valores das métricas de acurácia, ROC AUC, *recall,* precisão e F1-score de cada modelo.

Cabe salientar, ainda, que os dados de teste permanecem desbalanceados, assim, algumas métricas como a acurácia podem levar à falsa interpretação de um bom modelo. Tal fato, deve-se pelo motivo de uma classe apresentar muito mais amostras que a outra, sendo assim, se o modelo classificasse todos os casos como sobreviventes, o modelo acertaria a maior parte dos casos, o que acarretaria em uma alta acurácia.

Por esse motivo, a escolha da métrica para selecionar o modelo que obteve melhor desempenho não é simples. No entanto, em casos de anomalia, apesar de ambas as classes serem de extrema importância, a classe minoritária (óbitos)

apresenta maior relevância quando comparada com a majoritária (sobreviventes) na obtenção da métrica.

Dessa maneira, o resultado será determinado de acordo com a pontuação ROC AUC, que apresentará a relação de acertos ao prever os casos de óbito e de sobrevivência dos pacientes. Assim, deve-se calcular a taxa de acerto do número de óbitos previsto pelo modelo e o número total de óbitos e a taxa de acerto entre o número de sobreviventes previsto pelo modelo e o número total de sobreviventes. Com isso, a pontuação ROC AUC consegue determinar o desempenho do modelo em identificar corretamente ambas as classes da variável resposta. Logo, essa será a principal métrica de avaliação para a escolha do melhor modelo de aprendizado de máquina.

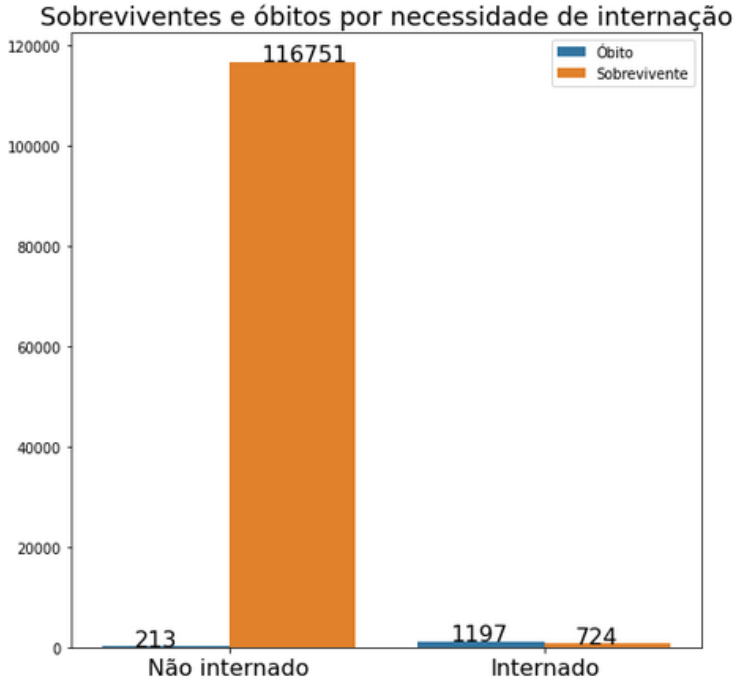
4 RESULTADOS

4.1 ANÁLISE E PRÉ PROCESSAMENTO DOS DADOS

Com a finalidade de compreender como o vírus afetou a população do município de Vitória, elaborou-se diversos gráficos que auxiliam no entendimento de como as características selecionadas para os modelos de aprendizado de máquina afetaram nos casos confirmados de Covid-19. A princípio, analisou-se a necessidade de internação dos pacientes, como mostra a Figura 12, e é possível concluir que na grande maioria dos casos de óbito, os pacientes necessitaram de internação. Isso, deve-se ao fato de que quando o paciente está em estado grave, necessita-se de internação.

Contudo, mesmo nos casos em que o paciente foi internado e, por conta disso, recebeu mais cuidados médicos, a taxa de mortalidade foi de 62%, conforme a Figura 12. Dessa forma, é possível compreender que a maioria dos casos de internação eram pacientes com extremo risco de vida.

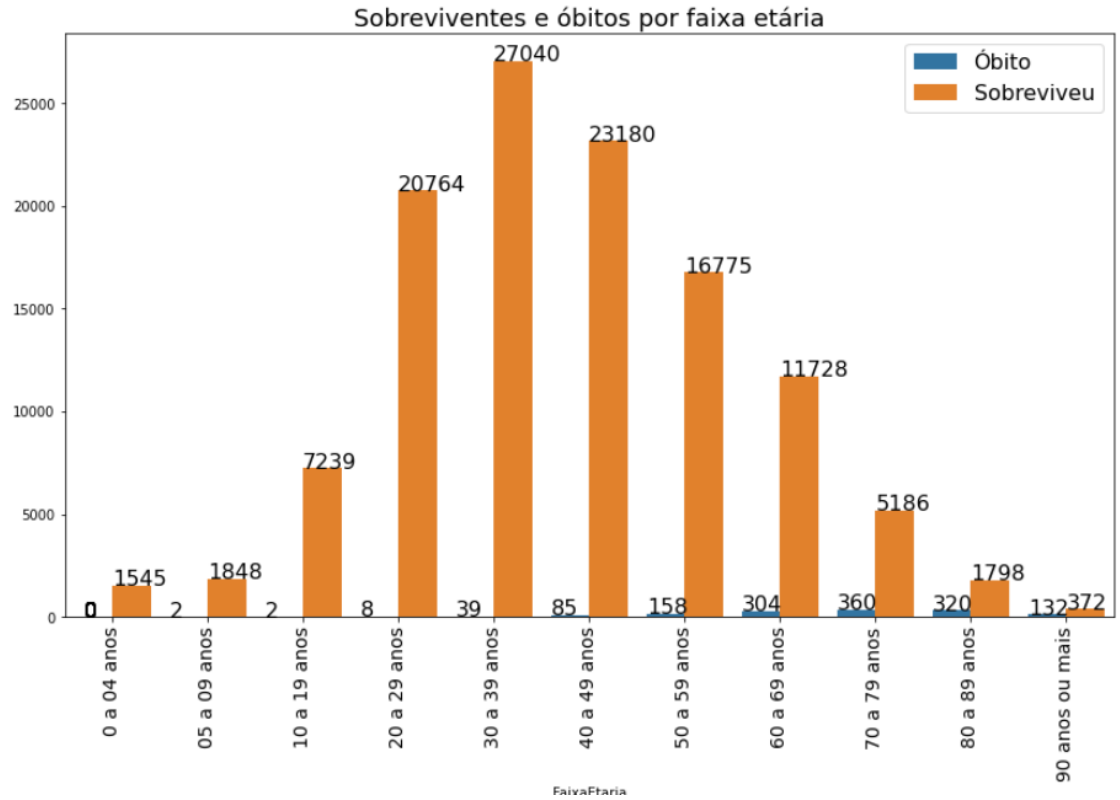
Figura 12 – Casos com necessidade de internação.



Fonte: elaborado pelo autor.

Posteriormente, buscou-se entender quais as características mais afetaram os pacientes e intensificaram as chances de óbito. Para isso, analisou-se características como: faixa etária, comorbidades e sintomas. Dessa forma, foi possível verificar que pacientes com idade mais elevada correm mais risco de óbito. De acordo com a Figura 13, pode-se observar que os pacientes acima de 60 anos apresentaram as maiores taxas de mortalidade, além disso, importante destacar que a taxa de mortalidade aumentou proporcionalmente com a faixa etária e o caso mais preocupante foi o de pacientes com 90 anos ou mais (26,19% de mortalidade).

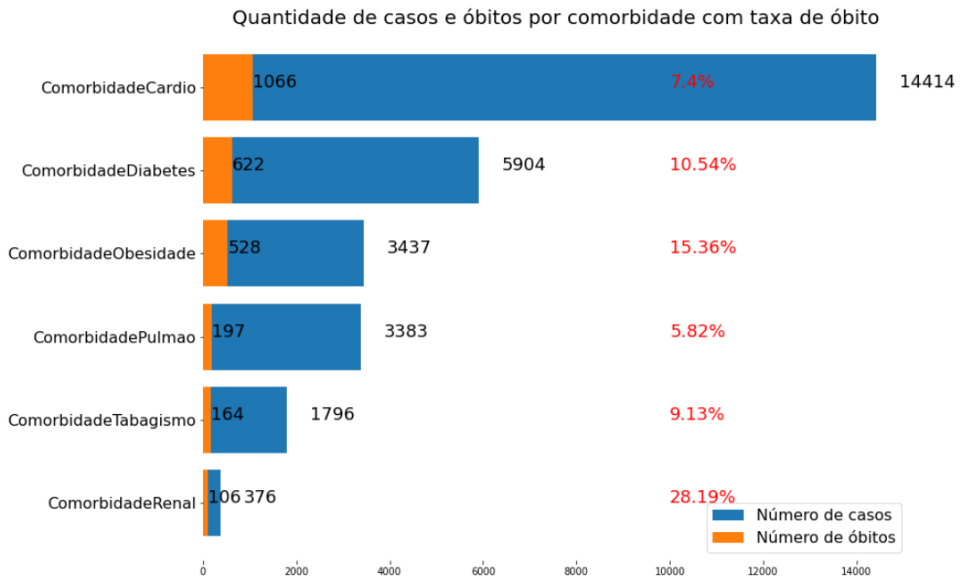
Figura 13 – Número de sobreviventes e óbitos por faixa etária.



Fonte: elaborado pelo autor.

Em relação às comorbidades, percebeu-se que a mais frequente entre os pacientes foi a cardíaca, porém, as comorbidades renais, a obesidade, a diabetes e o tabagismo, geraram maiores riscos à vida dos pacientes. Tal fato, deve-se a maior taxa de mortalidade registrada nesses casos, conforme representado na Figura 14.

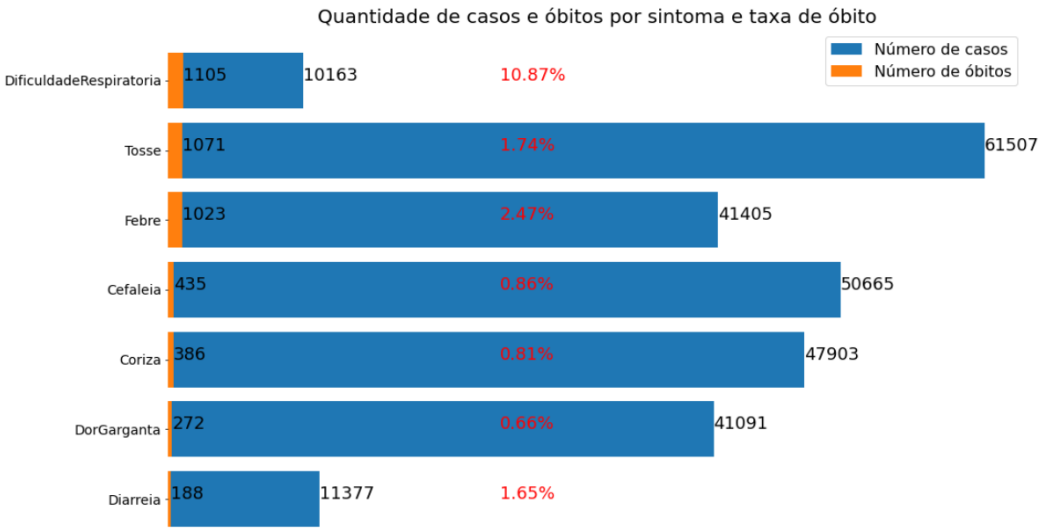
Figura 14 – Número de óbitos que apresentavam cada comorbidade.



Fonte: elaborado pelo autor.

Posteriormente, a mesma análise fora feita em relação aos sintomas de cada paciente, com o intuito de visualizar quais foram agravantes para os casos de óbito, como mostra a Figura 15. Feito isso, percebeu-se que a dificuldade respiratória foi o sintoma mais prejudicial à saúde dos pacientes, com uma taxa de mortalidade (10,87%) muito superior ao restante dos sintomas.

Figura 15 – Número de óbitos que apresentavam cada sintoma.



Fonte: elaborado pelo autor.

Cabe destacar, que diversos pacientes apresentaram várias comorbidades ou vários sintomas e esse fator pode influenciar nas taxas de óbito. Porém, essas características foram analisadas de maneira individual, sem considerar possíveis agravantes. Ou seja, os gráficos da Figura 14 e da Figura 15 consideram apenas se o paciente apresentava determinado sintoma ou determinada comorbidade e quantos sobreviveram ou não.

Por fim, pode-se identificar quais características ocasionaram maior risco à vida dos pacientes. Percebe-se, então, que as maiores taxas de óbito foram em casos que os pacientes apresentaram:

1. Faixa etária elevada (superior a 60 anos)
2. Comorbidade renal
3. Dificuldade respiratória
4. Necessidade de internação

4.2 RESULTADO DA MELHOR TÉCNICA DE BALANCEAMENTO PARA CADA ALGORITMO DURANTE A ETAPA DE VALIDAÇÃO

Nessa seção, serão apresentados os resultados obtidos a partir da realização de cada etapa referente à Seção 3.6. Nesse experimento, os hiperparâmetros de cada algoritmo foram testados de maneira individual, com a finalidade de encontrar os 5 melhores valores de cada. Posteriormente, foram realizadas diversas combinações entre os hiperparâmetros, com o objetivo de encontrar o melhor modelo para cada técnica de balanceamento. A partir dos resultados encontrados, avaliou-se o desempenho de cada modelo, a fim de selecionar em qual técnica de balanceamento o algoritmo apresentou a maior pontuação ROC AUC.

4.2.1 ÁRVORE DE DECISÃO

Com o teste individual de cada hiperparâmetro da árvore de decisão, os melhores valores encontrados para cada técnica de balanceamento podem ser visualizados na Tabela 14.

Tabela 14 – Melhores valores encontrados para cada hiperparâmetro.

Modelo	Balanceamento	Hiperparâmetros	Melhores Valores
Árvore de decisao	Undersampling	max_depth	[23, 48, 95, 39, 60]
		min_samples split:	[3, 4, 5, 7, 6]
		min_samples_leaf	[13, 8, 9, 10, 11]
		max_features	[6, 9, 7, 12, 15]
Árvore de decisao	SMOTE	max_depth	[2, 3, 5, 4, 6]
		min_samples split:	[21, 19, 18, 22, 24]
		min_samples_leaf	[23, 46, 47, 48, 44]
		max_features	[9, 3, 6, 4, 5]

Fonte: elaborado pelo autor.

A partir dos melhores valores individuais, foram realizadas diversas combinações entre os hiperparâmetros e foi possível obter o melhor valor para cada hiperparâmetro em cada técnica de balanceamento, conforme a Tabela 15.

Tabela 15 – Melhor valor encontrado com a combinação de cada hiperparâmetro.

Modelo	Balanceamento	Hiperparâmetros	Melhor Valor
Árvore de decisao	Undersampling	max_depth	48
		min_samples split:	7
		min_samples_leaf	13
		max_features	15
Árvore de decisao	SMOTE	max_depth	5
		min_samples split:	19
		min_samples_leaf	48
		max_features	9

Fonte: elaborado pelo autor.

Dito isso, encontrou-se o melhor modelo de aprendizado de máquina para a predição dos dados de validação com as diferentes técnicas de balanceamento. Assim, foi possível avaliar o desempenho dos modelos com a utilização do ROC AUC (Tabela 16).

Tabela 16 – Desempenho dos modelos de árvore de decisão na etapa de validação.

Modelo	ROC AUC
ArvoreDecisao Undersampling Validacao	92,9308
ArvoreDecisao SMOTE Validacao	93,7489

Fonte: elaborado pelo autor.

Dessa forma, pode-se concluir que no caso do algoritmo de árvore de decisão, o modelo apresentou melhor desempenho com os hiperparâmetros encontrados no balanceamento dos dados com a utilização da técnica *SMOTE*, apresentando o valor de aproximadamente 93,75%.

4.2.2 FLORESTA ALEATÓRIA

Com o teste individual de cada hiperparâmetro de floresta aleatória, os melhores valores encontrados para cada técnica de balanceamento podem ser visualizados na Tabela 17.

Tabela 17 – Melhores valores encontrados para cada hiperparâmetro.

Modelo	Balanceamento	Hiperparâmetros	Melhores valores
Floresta Aleatória	Undersampling	max_depth	[10, 35, 42, 17]
		min_samples split:	[22, 4, 12, 8, 10]
		min_samples_leaf	[21, 19, 1, 24, 3]
		max_features	[1, 3, 2, 4, 6]
		n_estimators	[112, 82, 24, 43]
Floresta Aleatória	SMOTE	max_depth	[6, 5, 3, 4, 2]
		min_samples split:	[62, 45, 66, 30, 65]
		min_samples_leaf	[42, 45, 20, 30, 21]

		max_features	[2, 6, 5, 7, 14]
		n_estimators	[55, 103, 77, 84]

Fonte: elaborado pelo autor.

A partir dos melhores valores individuais, foram realizadas diversas combinações entre os hiperparâmetros e foi possível obter o melhor valor para cada hiperparâmetro em cada técnica de balanceamento, conforme a Tabela 18.

Tabela 18 – Melhor valor encontrado com a combinação de cada hiperparâmetro.

Modelo	Balanceamento	Hiperparâmetros	Melhor Valor
Floresta Aleatória	Undersampling	max_depth	35
		min_samples split:	8
		min_samples_leaf	1
		max_features	2
		n_estimators	82
Floresta Aleatória	SMOTE	max_depth	6
		min_samples split:	62
		min_samples_leaf	45
		max_features	6
		n_estimators	55

Fonte: elaborado pelo autor.

Dito isso, encontrou-se o melhor modelo de aprendizado de máquina para a predição dos dados de validação com as diferentes técnicas de balanceamento. Assim, foi possível avaliar o desempenho dos modelos com a utilização do ROC AUC (Tabela 19).

Tabela 19 – Desempenho dos modelos de floresta aleatória na etapa de validação.

Modelo	ROC AUC
FlorestaAleatoria Undersampling Validacao	96,6030
FlorestaAleatoria SMOTE Validacao	95,9206

Fonte: elaborado pelo autor.

Dessa forma, pode-se concluir que no caso do algoritmo de floresta aleatória, o modelo apresentou melhor desempenho com os hiperparâmetros encontrados no balanceamento dos dados com a utilização da técnica *undersampling*, apresentando o valor de aproximadamente 96,60%.

4.2.3 ADABOOST

Com o teste individual de cada hiperparâmetro do algoritmo AdaBoost, os melhores valores encontrados para cada técnica de balanceamento podem ser visualizados na Tabela 20.

Tabela 20 – Melhores valores encontrados para cada hiperparâmetro.

Modelo	Balanceamento	Hiperparâmetros	Melhores Valores
AdaBoost	Undersampling	learning_rate	[0.1, 0.3, 0.2, 0.4, 0.5]
		n_estimators	[9, 19, 20, 18, 16]
AdaBoost	SMOTE	learning_rate	[0.1, 0.2, 0.3, 0.7, 0.8]
		n_estimators	[3, 7, 4, 5, 9]

Fonte: elaborado pelo autor.

A partir dos melhores valores individuais, foram realizadas diversas combinações entre os hiperparâmetros e foi possível obter o melhor valor para cada hiperparâmetro em cada técnica de balanceamento, conforme a Tabela 21.

Tabela 21 – Melhor valor encontrado com a combinação de cada hiperparâmetro.

Modelo	Balanceamento	Hiperparâmetros	Melhor Valor
AdaBoost	Undersampling	learning_rate	0.5
		n_estimators	19
AdaBoost	SMOTE	learning_rate	0.8
		n_estimators	9

Fonte: elaborado pelo autor.

Dito isso, encontrou-se o melhor modelo de aprendizado de máquina para a predição dos dados de validação com as diferentes técnicas de balanceamento. Assim, foi possível avaliar o desempenho dos modelos com a utilização do ROC AUC (Tabela 22).

Tabela 22 – Desempenho dos modelos de AdaBoost na etapa de validação.

Modelo	ROC AUC
AdaBoost Undersampling Validacao	94,4808
AdaBoost SMOTE Validacao	89,4054

Fonte: elaborado pelo autor.

Dessa forma, pode-se concluir que no caso do algoritmo de AdaBoost, o modelo apresentou melhor desempenho com os hiperparâmetros encontrados no balanceamento dos dados com a utilização da técnica *undersampling*, apresentando o valor de aproximadamente 94,48%.

4.2.4 XGBOOST

Com o teste individual de cada hiperparâmetro do algoritmo XGBoost, os melhores valores encontrados para cada técnica de balanceamento podem ser visualizados na Tabela 23.

Tabela 23 – Melhores valores encontrados para cada hiperparâmetro.

Modelo	Balanceamento	Hiperparâmetros	Melhores Valores
XGBoost	Undersampling	learning_rate	[0.13, 0.09, 0.05, 0.97]
		n_estimators	[5, 18, 19, 20 , 52]
		max_depth	[2, 6, 7, 8]
		subsampling	[0.6, 0.9, 0.7, 0.5]
XGBoost	SMOTE	learning_rate	[0.05, 0.09, 0.01, 0.17]
		n_estimators	[69, 68, 50, 66, 67]
		max_depth	[2, 1, 3, 5, 4]
		subsampling	[0.9, 0.6, 0.7, 0.8]

Fonte: elaborado pelo autor.

A partir dos melhores valores individuais, foram realizadas diversas combinações entre os hiperparâmetros e foi possível obter o melhor valor para cada hiperparâmetro em cada técnica de balanceamento, conforme a Tabela 24.

Tabela 24 – Melhor valor encontrado com a combinação de cada hiperparâmetro.

Modelo	Balanceamento	Hiperparâmetros	Melhor Valor
XGBoost	Undersampling	learning_rate	0.13
		n_estimators	52
		max_depth	6
XGBoost	SMOTE	learning_rate	0.09
		n_estimators	68
		max_depth	3

Fonte: elaborado pelo autor.

Dito isso, encontrou-se o melhor modelo de aprendizado de máquina para a predição dos dados de validação com as diferentes técnicas de balanceamento. Assim, foi possível avaliar o desempenho dos modelos com a utilização do ROC AUC (Tabela 25).

Tabela 25 – Desempenho dos modelos de XGBoost na etapa de validação.

Modelo	ROC AUC
XGBoost Undersampling Validacao	93,7122
XGBoost SMOTE Validacao	94,5718

Fonte: elaborado pelo autor.

Dessa forma, pode-se concluir que no caso do algoritmo de XGBoost, o modelo apresentou melhor desempenho com os hiperparâmetros encontrados no balanceamento dos dados com a utilização da técnica *SMOTE*, apresentando o valor de aproximadamente 94,57%.

4.3 RESULTADO DAS MÉTRICAS DOS MELHORES MODELOS NA PREDIÇÃO DOS DADOS DE TESTE

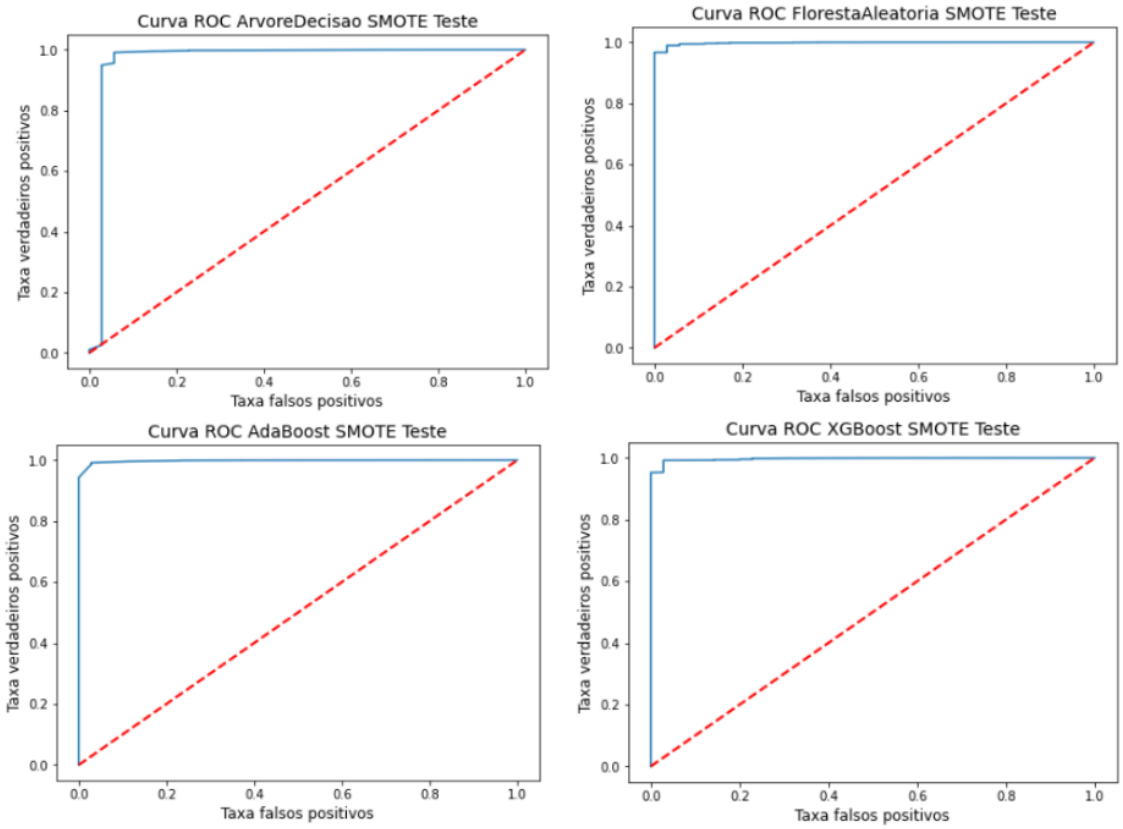
Com a escolha dos melhores resultados provenientes da Seção 4.2, os modelos foram novamente treinados com a utilização dos dados de treino e validação balanceados conforme a melhor técnica encontrada para cada modelo. Com isso, o teste foi realizado com o conjunto de dados destinado para teste e pôde-se observar as métricas de desempenho dos modelos, conforme a Tabela 26. Além disso, é possível visualizar o comportamento da curva ROC na Figura 16 e analisar o comportamento da taxa de verdadeiros positivos e da taxa falsos positivos para diferentes limiares em cada modelo

Tabela 26 – Métricas dos melhores algoritmos nos dados de teste.

Número Modelo	Modelo	Acurácia	ROC AUC	Recall	Precisão	F1-score
1	ArvoreDecisao SMOTE Teste	97,74	97,44	97,74	54,83	58,21
2	FlorestaAleatoria Undersampling Teste	99,14	96,72	99,15	60,78	67,36
3	AdaBoost Undersampling Teste	99,65	96,97	99,66	70,36	78,36
4	XGBoost SMOTE Teste	99,34	89,70	99,39	62,26	68,63

Fonte: elaborado pelo autor.

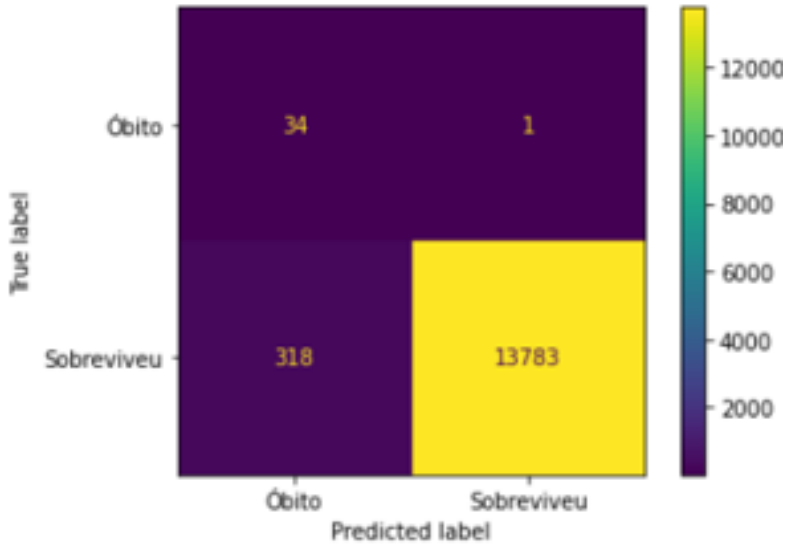
Figura 16 – Curva ROC dos melhores modelos.



Fonte: elaborado pelo autor.

Além disso, gerou-se os gráficos referentes à matriz de confusão dos modelos, com a finalidade de auxiliar o entendimento das métricas. Dessa forma, mostra-se possível a análise de quantos casos de óbito e de sobreviventes os modelos classificaram corretamente, e quantos casos os modelos classificaram de maneira equivocada. Na Figura 17, referente ao modelo de árvore de decisão, percebe-se que de 35 óbitos que ocorreram no período de teste, 34 foram classificados corretamente e apenas 1 caso foi classificado como sobrevivente. Já nos casos de sobreviventes, de 14.101 casos, o modelo classificou corretamente 13.783, porém 318 casos foram classificados como óbito.

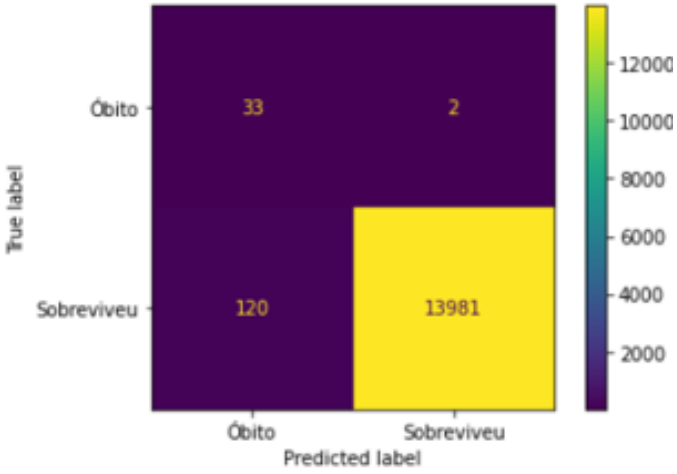
Figura 17 – Matriz de confusão do Modelo 1 (“ArvoreDecisao SMOTE Teste”).



Fonte: elaborado pelo autor.

Já na Figura 18, referente ao modelo de floresta aleatória, percebe-se que de 35 óbitos que ocorreram no período de teste, 33 foram classificadas corretamente e 2 casos foram classificados como sobreviventes. Porém, nos casos dos sobreviventes, de 14.101 casos, o modelo classificou corretamente 13.981, e o erro diminuiu para 120 casos classificados como óbito.

Figura 18 – Matriz de confusão do Modelo 2 (“FlorestaAleatoria Undersampling Teste”).

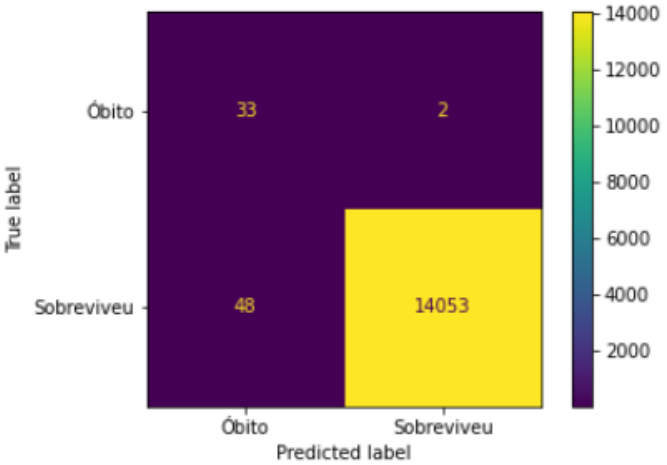


Fonte: elaborado pelo autor.

No modelo do AdaBoost (Figura 19), percebe-se que de 35 óbitos que ocorreram no período de teste, 33 foram classificadas corretamente e 2 casos foram

classificados como sobreviventes. Porém, nos casos dos sobreviventes, de 14.101 casos, o modelo classificou corretamente 14.053, e apenas 48 casos foram classificados como óbito, ocasionando na melhor taxa de precisão entre os modelos.

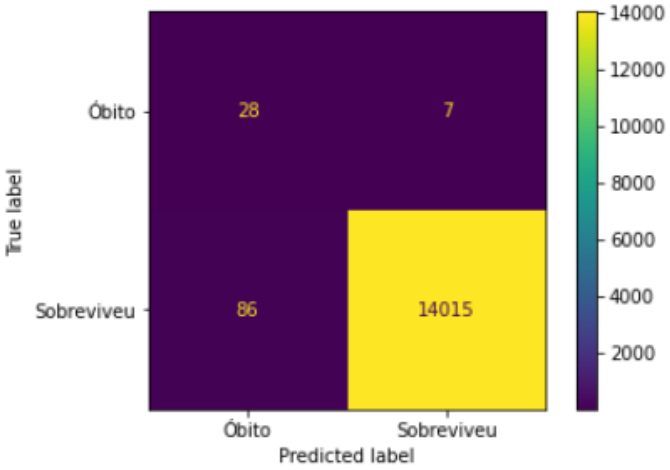
Figura 19 – Matriz de confusão do Modelo 3 (“AdaBoost Undersampling Teste”)



Fonte: elaborado pelo autor.

Por fim, no modelo de XGBoost (Figura 20), percebe-se que de 35 óbitos que ocorreram no período de teste, apenas 28 foram classificados corretamente e 7 casos foram classificados como sobreviventes, apresentando a pior classificação para a classe de maior importância. Já nos casos dos sobreviventes, de 14.101 casos, o modelo classificou corretamente 14.015, e 86 casos foram classificados como óbito.

Figura 20 – Matriz de confusão do Modelo 4 (“XGBoost SMOTE Teste”).



Fonte: elaborado pelo autor.

A partir dos resultados apresentados, percebe-se que o Modelo 1, Modelo 2 e Modelo 3, apresentaram níveis satisfatórios para a pontuação ROC AUC, ou seja, os modelos conseguiram classificar corretamente a maior parte dos casos nos quais os pacientes faleceram. Porém, percebe-se que todos modelos apresentados não apresentaram valores altos para as métricas de precisão e, consequentemente, f1-score, isso significa que, os modelos realizaram diversas predições como óbito para pacientes que, felizmente, conseguiram sobreviver. A hipótese para a maior quantidade de erros para casos de falsos negativos será abordada na Seção 4.5, sobre as limitações do projeto.

Dessa forma, é possível concluir que o Modelo 1 (“ArvoreDecisao SMOTE Teste”) apresentou melhor pontuação ROC AUC (97,44%) e, conforme abordado ao longo do trabalho, esta métrica pode ser considerada a mais importante para avaliação de desempenho do modelo. Porém, as outras métricas do Modelo 1, apresentaram desempenho bem inferior ao comparar com os outros modelos.

Assim, destaca-se que o Modelo 3 (“AdaBoost Undersampling Teste”), apesar de apresentar ROC AUC de 96,97%, ou seja, valor de 0,47% de diferença para o Modelo 1, com relação às outras métricas apresentou desempenho muito superior. Com relação a precisão, o Modelo 3 apresentou 70,36% (15,53% superior ao Modelo 1), já com relação ao *f1-score* o Modelo 3 apresentou 78,36% (20,15% superior ao Modelo 1).

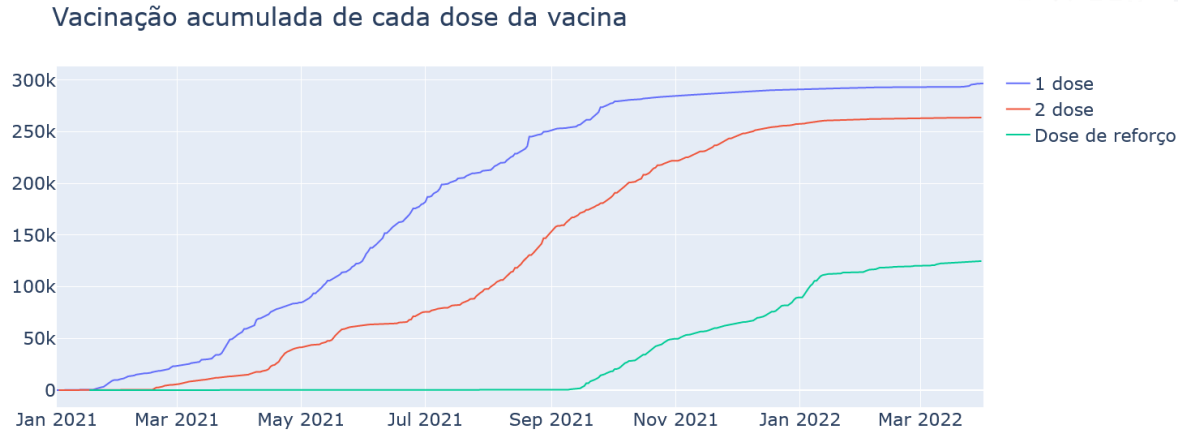
4.4 LIMITAÇÕES

Uma das limitações encontradas durante a execução do trabalho foi a falta do atributo que indicasse se o paciente de cada caso estaria vacinado contra a Covid-19. Como os casos começaram em fevereiro de 2020 e a vacinação, apenas no começo de 2021, percebe-se que incluir os dados sobre a vacinação poderia ser uma dificuldade para as prefeituras, visto que a vacinação começou de forma gradativa, atendendo apenas casos específicos, até apresentar disponibilidade para toda a população.

Porém, o estado oferece os dados referentes às doses de cada etapa da vacinação e a quantidade de pessoas registradas para cada dose. Dessa forma, foi possível a elaboração de gráficos complementares, que indicam números relacionados as doses distribuídas no município de Vitória.

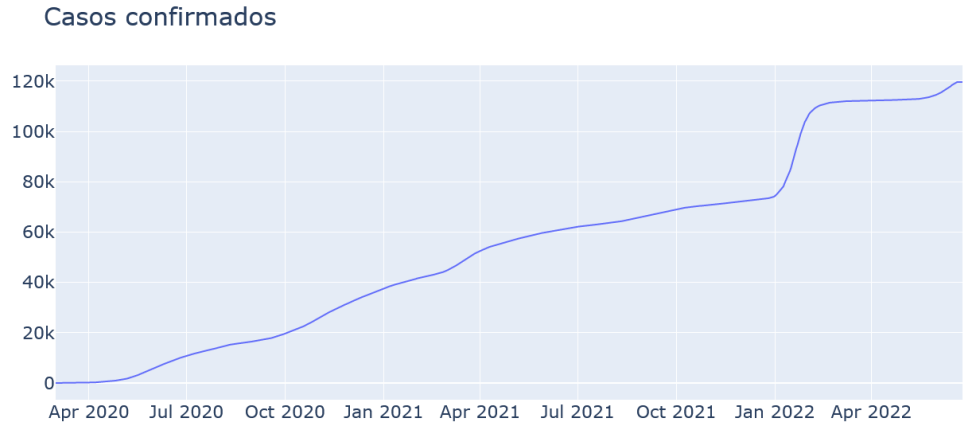
Ao comparar os dados referentes as doses aplicadas da vacina (Figura 21) com os números de casos confirmados acumulados (Figura 22) e óbitos acumulados (Figura 23), percebe-se que a vacinação pode ser um fator que influenciou nos casos de óbitos por covid-19. A partir da análise dos dados, foi possível verificar maior estabilidade nos casos confirmados no ano de 2021 e um elevado número de casos confirmados em janeiro de 2022. Contudo, a curva de óbitos começa a estabilizar com o aumento das doses aplicadas e, mesmo com um pico de casos confirmados no início de 2022, os maiores índices de óbitos ocorreram antes de julho de 2021, ou seja, antes da grande maioria da população receber a imunização.

Figura 21 – Total de doses aplicadas da vacina acumulativa ao longo do tempo.



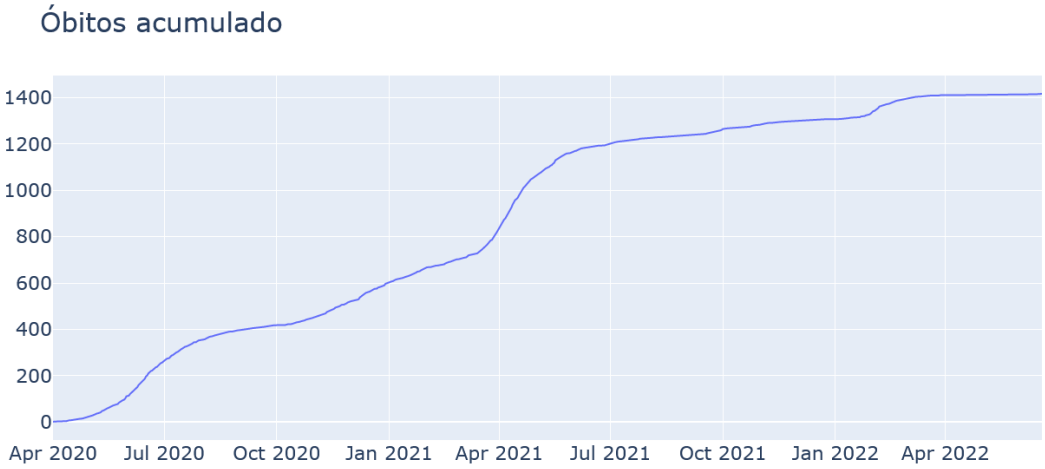
Fonte: elaborado pelo autor.

Figura 22 – Casos confirmados de covid-19 acumulativo.



Fonte: elaborado pelo autor.

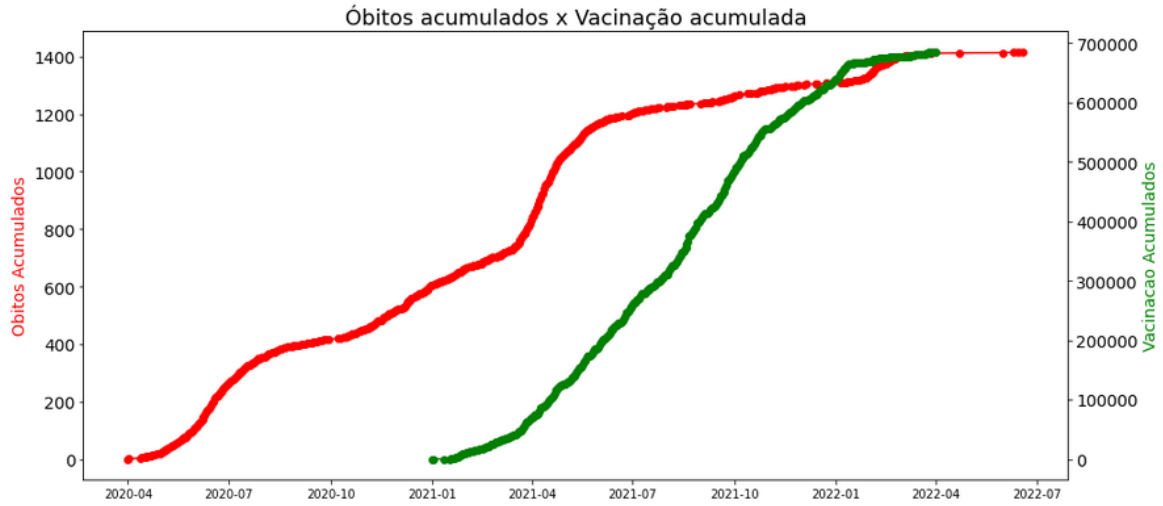
Figura 23 – Óbitos confirmados acumulado ao longo do tempo.



Fonte: elaborado pelo autor.

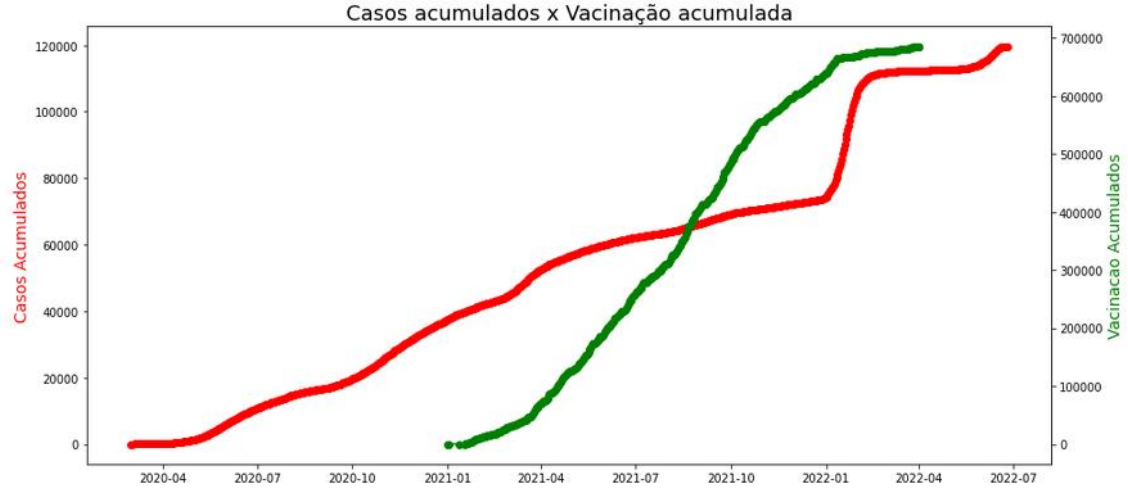
Verifica-se, ainda, o gráfico gerado em duas escalas diferentes com a comparação do comportamento entre o número de óbitos acumulados e de doses aplicadas (Figura 24), no qual há a estabilização do número de óbitos com a crescente da vacinação. E, também, o gráfico da Figura 25, que faz o comparativo entre os casos confirmados e as doses da vacinação e nota-se que o maior pico de casos confirmados foi após grande parte dos pacientes estarem imunizados.

Figura 24 – Comparativo de óbitos e vacinação.



Fonte: elaborado pelo autor.

Figura 25 – Comparativo de casos e vacinação.



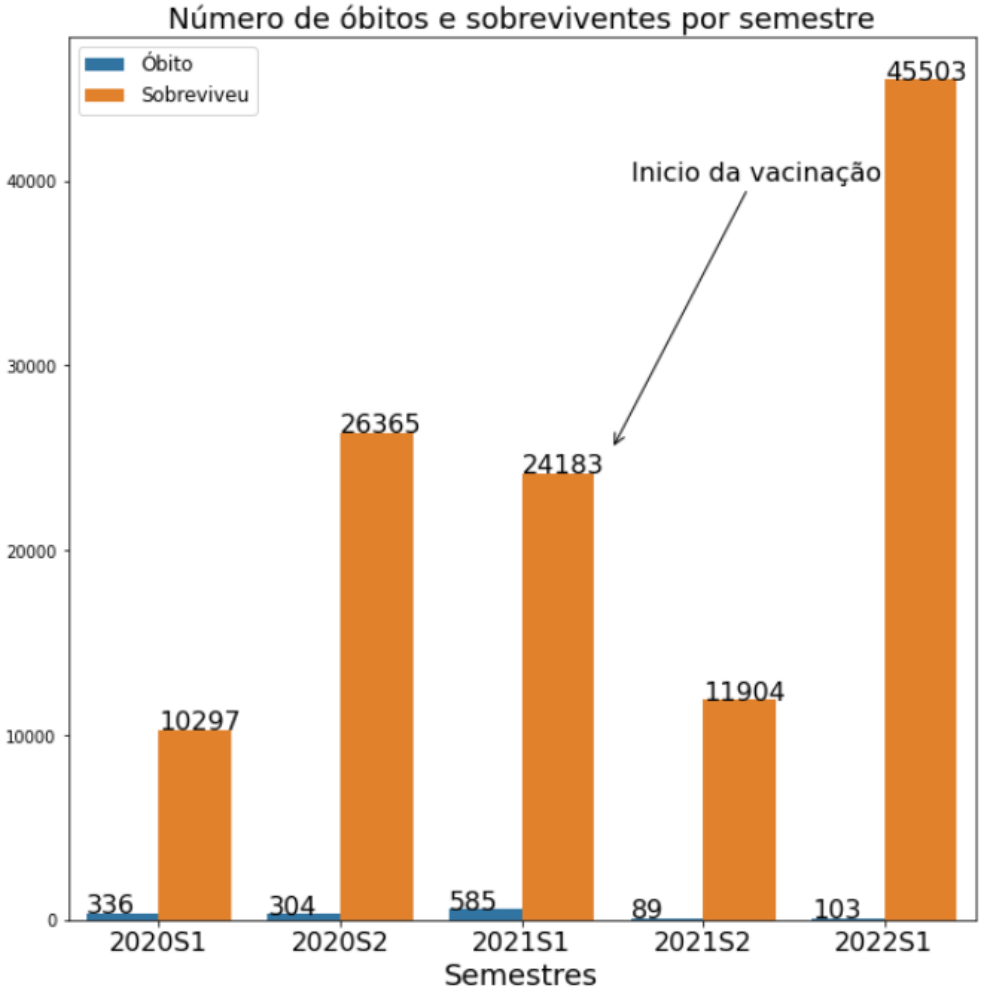
Fonte: elaborado pelo autor.

Outra forma de visualização foi elaborada na Figura 26, que mostra a quantidade de casos de sobrevivência e de óbito por semestre. Nesse gráfico, percebe-se que nos semestres que apresentavam maior parte da população vacinada (2021S2 e 2022S1) o número de óbitos foi reduzido, mesmo com o aumento de casos encontrados no início do ano de 2022.

Porém, além da vacinação, outro fator que pode ter influenciado na taxa de mortalidade foi o fato do vírus mudar com o tempo e sofrer mutações, fator que gerou diversas variantes do coronavírus com taxas de transmissão e de

mortalidade diferentes. Sendo assim, a falta da informação de qual variante cada paciente foi infectado, também se enquadra como uma limitação para melhores resultados. Dessa forma, cabe destacar que inicialmente a pandemia foi impulsionada pelas linhagens B.1.1.28 e B.1.1.33 que foram as mais frequentes até outubro de 2020. A partir disso, originou-se, no Brasil, duas novas variantes P.1 e P.2 e, com apenas quatro meses em circulação, corresponderam a 75% dos casos no Brasil (MICHELON,2021). Por fim, a última cepa identificada do vírus no ano de 2021, foi a Ômicron, variante altamente transmissível e com grande número de mutações que infectou pacientes já imunizados ou que já haviam se recuperado da doença.

Figura 26 – Total de casos e óbitos por semestre.



Fonte: elaborado pelo autor.

Dado os fatos, acredita-se que, apesar de um aumento nos casos com o contágio da variante Ômicron, a crescente da vacinação e o aumento do conhecimento médico acerca da Covid-19, o número de óbitos reduziu em comparação com o primeiro ano da pandemia. Dessa maneira, estima-se que a presença de uma variável que representasse o esquema vacinal do paciente e outro atributo para informar a variante do vírus poderia contribuir positivamente com o modelo.

No modelo apresentado, a fase de treino aconteceu com casos de um período em que não havia vacinação (ou a vacinação estava em estágio inicial) e os pacientes estavam infectados com as primeiras variantes do vírus. Já as fases de validação e teste, ocorreram em um período com elevada taxa de vacinação e outras variantes predominavam no Brasil. Logo, com a separação temporal dos dados, o modelo pode ter classificado como óbito, pacientes que felizmente sobreviveram devido ao novo cenário em que se encontrava a doença.

5 CONCLUSÕES

Após a execução do projeto, pôde-se perceber que, para o conjunto de dados em questão, as técnicas para pré-processamento dos dados auxiliaram na remoção de características inconsistentes sobre os casos de Covid-19 no município de Vitória, ES. Além disso, verificou-se que o balanceamento dos dados por meio da técnica de *undersampling* foi mais eficaz para os algoritmos de floresta aleatória e AdaBoost e que com a técnica *SMOTE* os algoritmos de árvore de decisão e XGBoost desempenharam melhor.

Em seguida, reparou-se que o modelo que apresentou a melhor métrica ROC AUC (97,44%) foi Modelo 1, referente ao (“ArvoreDecisao SMOTE Teste”). Porém, o Modelo 3 (“AdaBoost Undersampling Teste”), apesar de apresentar pontuação ROC AUC um pouco inferior (96,97%), apresentou desempenho significativamente melhor em relação às métricas de precisão *e f1-score.* Dessa forma, destaca-se que ambos os modelos apresentaram valores de ROC AUC dentro do esperado quando comparado com os resultados obtidos nos outros trabalhos abordados no Estado da arte (Seção 2.5).

Apesar da base de dados e as etapas de pré-processamento dos trabalhos serem distintas e não utilizarem muitos casos após o período de vacinação, o valor ROC AUC foi próximo ao encontrado em (YADAW et al., 2020) que apresentou o valor de 0,91-0,94 com o intervalo de confiança de 95%, ao resultado encontrado por (RODRIGUES, 2021) de 0,981 e ao resultado encontrado pelo melhor modelo de (L WANG et al., 2022), com o valor de 94,81%. Porém, quando comparado com os resultados de (IWENDI, et al., 2020), notou-se diferença, visto que o trabalho utilizou a métrica de *f1-score* para avaliar o desempenho do modelo e não a métrica ROC AUC.

Cabe salientar, ainda, que o estudo feito na Seção de Limitações (4.4) mostrou a importância da vacina e o impacto das variantes do vírus na redução dos casos de óbito por covid-19 no município de Vitória. Com isso, a falta dessas informações pode ter interferido nos resultados de precisão dos modelos.

Vale ressaltar, também, que os modelos de classificação apresentados serviriam como uma ferramenta complementar para auxiliar equipes médicas com dados. Visto que, os conhecimentos e a ética médica são indispensáveis na realização de diagnósticos e prognósticos a respeito de qualquer doença, e o tratamento para cada caso fica a cargo da equipe médica.

Por fim, uma vez que cada município é responsável por coletar seus próprios dados, em diferentes localizações, os dados coletados podem abordar características distintas dos pacientes distinta das coletadas em Vitória-ES. Logo, não é possível afirmar que os modelos de aprendizado de máquinas apresentados seriam eficientes em classificar o prognóstico do paciente infectado com o vírus em questão.

5.1 SUGESTÕES TRABALHOS FUTUROS

Como forma de aprimorar o trabalho realizado e verificar o desempenho dos algoritmos de aprendizado de máquina na base de dados escolhida, é possível realizar a separação dos conjuntos de dados de treino, validação e teste com os dados separados de maneira aleatória, ou seja, sem levar em consideração a data de diagnóstico de cada paciente. Dessa forma, as amostras destinadas para treino, por não estarem separadas de maneira cronológica, apresentariam casos de diferentes variantes do vírus e de pacientes em diferentes etapas de vacinação. Assim, os algoritmos poderiam apresentar melhor desempenho na classificação dos pacientes, principalmente para as métricas de precisão e *f1-score.* Além disso, existe a possibilidade de utilizar base de dados de outros estados/municípios que forneçam dados associados aos casos confirmados do novo coronavírus no local, como forma de verificar se o vírus apresentou impactos semelhantes ao da população de Vitória-ES.

REFERÊNCIAS

AMB. **[Coletiva de imprensa – 09/04/2021] Recomendações para triagem de pacientes em UTIs no atual momento da pandemia – CEM-Covid_AMB**, 2021. Disponível em: < https://amb.org.br/cem-covid/protocolo-para-triagem-de-pacientes-em-utis//>. Acesso em: 23/01/2022

ANDRADE, M. **Espírito Santo é exemplo de 'open data' contra a Covid-19**, 2020. Disponível em: <https://whitepaperdocs.com/2020/07/espirito-santo-e-exemplo-de-open-data-contra-a-covid-19/>. Acesso em: 09/01/2022

BOWES, C. **O médico que descobriu como a cólera se espalha (e impediu a doença de causar mais mortes)**, 2020. Disponível em: <https://www.bbc.com/portuguese/geral-53376925>. Acesso em: 14/01/2022

BROWNLEE, J. **Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning**. [s.l.] Machine Learning Mastery, 2020.

COLLINS, G. S. et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod) the tripod statement. **Circulation**, Am Heart Assoc, v. 131, n.2, p. 211-219, 2015.

DU, R.-H. et al. Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: a prospective cohort study. **European Respiratory Journal**, v. 55, n. 5, p. 2000524, 2020.

DUARTE, G. et al. **Guia brasileiro de análise de dados: armadilhas & soluções**. Brasília: ENAP. p. 83–96. 2021.

FILHO, C. **Nota de Esclarecimento – Critérios para escolha de pacientes para atendimento em UTI,** 2020. Disponível em: < https://cremers.org.br/nota-de-esclarecimento-criterios-para-escolha-de-pacientes-para-atendimento-em-uti/>. Acesso em: 04/02/2022

GARCÍA, S.; LUENGO, J.; HERRERA, F. Feature selection. **Intelligent Systems Reference Library**, v. 72, n. 6, p. 163–193, 2015.

GOOGLE MACHINE LEARNING EDUCATION. **Validation set.** Machine Learning Crash Course, 2022

GRUS, J. **Data science from Scratch: First Principles with Python**. Sebastopol: O'REILLY, 2015

HURWITZ, J; KIRSCH, D. **Machine Learning for dummies.** Hoboken: IBM LIMITED EDITION, 2018.

HUNTER, J. D. Matplotlib: A 2D graphics environment. **Computing in Science \& Engineering**, v. 9, n. 3, p. 90–95, 2007.

HARRISON, M. **Machine Learning Guia de Referência Rápida: Trabalhando com dados estruturados em Python.** São Paulo: O'REILLY, 2020.

IBM CLOUD EDUCATION. **O que é machine learning?**, 2020. Disponível em: <https://www.ibm.com/br-pt/cloud/learn/machine-learning>. Acesso em: 16/01/2022

INSTITUTO BUTANTAN. **Como surgiu o novo coronavírus? Conheça as teorias mais aceitas sobre sua origem**, 2021. Disponível em: <https://butantan.gov.br/covid/butantan-tira-duvida/tira-duvida-noticias/como-surgiu-o-novo-coronavirus-conheca-as-teorias-mais-aceitas-sobre-sua-origem>. Acesso em: 07/01/2022

INSTITUTO BUTANTAN. **Qual a diferença entre SARS-CoV-2 e Covid-19? Prevalência e incidência são a mesma coisa? E mortalidade e letalidade?,** 2021. Disponível em: <https://butantan.gov.br/covid/butantan-tira-duvida/tira-duvida-noticias/qual-a-diferenca-entre-sars-cov-2-e-covid-19-prevalencia-e-incidencia-sao-a-mesma-coisa-e-mortalidade-e-letalidade>. Acesso em: 05/01/2022

INSTITUTO BUTANTAN. **Saiba como diferenciar os sintomas da gripe e da Covid-19 em meio ao surto e à pandemia,** 2021. Disponível em: < https://butantan.gov.br/noticias/saiba-como-diferenciar-os-sintomas-da-gripe-e-da-covid-19-em-meio-ao-surto-e-a-pandemia>. Acesso em: 05/02/2022

IWENDI, C. et al. COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. **Frontiers in Public Health**, v. 8, 2020.

JOHNSON, J. M.; KHOSHGOFTAAR, T. M. Survey on deep learning with class imbalance. **Journal of Big Data**, v. 6, n. 1, p. 27, 2019.

KLUYVER, T. et al. **Jupyter Notebooks -- a publishing format for reproducible computational workflows**. (F. Loizides, B. Schmidt, Eds.)Positioning and Power in Academic Publishing: Players, Agents and Agendas, 2016.

KUMAR, V. Feature Selection: A literature Review. **The Smart Computing Review**, v. 4, n. 3, 2014.

L. Wang et al., "A Time-Series Feature-Based Recursive Classification Model to Optimize Treatment Strategies for Improving Outcomes and Resource Allocations of COVID-19 Patients," in IEEE Journal of Biomedical and Health

Informatics, vol. 26, no. 7, pp. 3323-3329, July 2022, doi: 10.1109/JBHI.2021.3139773.

MANDREKAR, J. N. Receiver operating characteristic curve in diagnostic test assessment. **Journal of Thoracic Oncology,** Elsevier, v. 5, n. 9, p. 1315-1316, 2010.

MCKINNEY, W.. **Data structures for statistical computing in python**. Proceedings of the 9th Python in Science Conference. p. 51-56, 2010.

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE (OPAS). **Folha informativa sobre COVID-19,** [entre 2020 e 2022]. Disponível em: <https://www.paho.org/pt/covid19/historico-da-pandemia-covid-19>. Acesso em: 06/01/2022

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in {P}ython. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

RABI, F. A. et al. SARS-CoV-2 and Coronavirus Disease 2019: What We Know So Far. **Pathogens**, v. 9, n. 3, p. 231, 20 mar. 2020.

REIS, E.A., Reis I.A. **Análise Descritiva de Dados**, 2002. Relatório Técnico do Departamento de Estatística da UFMG. Disponível em: <www.est.ufmg.br>. Acesso em: 12/01/2022

RODRIGUES, G. **Modelo Preditivo para prognóstico de pacientes com COVID-19**. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade Federal do Pampa, Alegrete, 2021.

SARANG NARKHEDE. Understanding AUC - ROC Curve. **Towards Data Science**, p. 6–11, 2019.

SECRETARIA DE SAÚDE. **Primeiro caso de Covid-19 no Brasil permanece sendo o de 26 de fevereiro**, 2020. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/noticias/2020/julho/primeiro-caso-de-covid-19-no-brasil-permanece-sendo-o-de-26-de-fevereiro>. Acesso em: 07/01/2022

SECRETARIA DE SAÚDE DO ESTADO DO ESPÍRITO SANTO. **NOTA TÉCNICA COVID-19 N° 06/2021: Definição de Casos Operacionais e Critérios de Coleta.** Espírito Santo, p. 1. 2021.

SECRETARIA DE SAÚDE DO ESTADO DO ESPÍRITO SANTO. **Dicionário de dados,** 2020. Disponível em: < https://coronavirus.es.gov.br/painel-covid-19-es/> Acesso em:

SINGHAL, R.; RANA, R. Chi-square test and its application in hypothesis testing. **Journal of the Practice of Cardiovascular Sciences**, v. 1, n. 1, p. 69, 2015.

TABLEAU. **What is Data Visualization? A definition, examples, and resources,** s.d Disponível em: <https://www.tableau.com/pt-br/learn/articles/data-visualization>. Acesso em: 15/01/2022

TAMIR, M. **What Is Machine Learning (ML)?**, 2020. Disponível em: <https://ischoolonline.berkeley.edu/blog/what-is-machine-learning/>. Acesso em: 13/01/2022

THE ECONOMIST. **The world's most valuable resource is no longer oil, but data,** 2017. Disponível em: <https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data>. Acesso em 17/01/2022

UNA-SUS. **Organização Mundial de Saúde declara pandemia do novo Coronavírus**, 2020. Disponível em: <https://www.unasus.gov.br/noticia/organizacao-mundial-de-saude-declara-pandemia-de-coronavirus>. Acesso em: 06/01/2022

WASKOM, M. L. seaborn: statistical data visualization. **Journal of Open Source Software**, v. 6, n. 60, p. 3021, 2021.

WHO. **Coronavirus disease (COVID-19),** [entre 2020 e 2022]. Disponível em: <https://www.who.int/health-topics/coronavirus#tab=tab_1>. Acesso em: 11/01/2022

WHO. **WHO Coronavirus (COVID-19) Dashboard**, 2022. Disponível em: <https://covid19.who.int/>. Acesso em: 26/07/2022

YADAW, A.S. et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. **Bmj,** British Medical Journal Publishing Group, v. 369, 2020.