UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

DANIEL BORTOT DE SALLES
Orientadora: Fernanda Sumika Hojo de Souza

PREDIÇÃO DE DESFECHO DESFAVORÁVEL EM PACIENTES COM COVID-19 USANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA

Ouro Preto, MG
2023

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

DANIEL BORTOT DE SALLES

PREDIÇÃO DE DESFECHO DESFAVORÁVEL EM PACIENTES COM COVID-19 USANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientadora: Fernanda Sumika Hojo de Souza

Ouro Preto, MG
2023

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Daniel Bortot de Salles

Predição de desfecho desfavorável em pacientes com COVID-19 usando técnicas de aprendizado de máquina

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 22 de Agosto de 2023.

Membros da banca

Fernanda Sumika Hojo de Souza (Orientadora) - Doutora - Universidade Federal de Ouro Preto
Anderson Almeida Ferreira (Examinador) - Doutor - Universidade Federal de Ouro Preto
Vander Luis de Souza Freitas (Examinador) - Doutor - Universidade Federal de Ouro Preto

Fernanda Sumika Hojo de Souza, Orientadora do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 22/08/2023.

Documento assinado eletronicamente por **Fernanda Sumika Hojo de Souza**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 23/08/2023, às 14:09, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0577015** e o código CRC **FAB0021D**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Resumo

As vacinas contra a COVID-19, desenvolvidas e aprovadas para uso emergencial em tempo recorde, desempenharam um papel fundamental em reduzir o número de casos graves, hospitalizações e mortes. Contudo, o surgimento de variantes e subvariantes do SARS-CoV-2 em função das mutações do vírus, a perda gradual da imunidade induzida por vacinas ou infecção prévia, assim como o escape imune dessas variantes, propiciam a ocorrência de reinfecções e infecções disruptivas sintomáticas ou assintomáticas. O objetivo do presente estudo foi desenvolver modelos para prognóstico ou identificação precoce de pacientes com risco aumentado para desfecho desfavorável no atual cenário de predominância de variantes de preocupação e cobertura vacinal relativamente ampla. Dados de pacientes com COVID-19 hospitalizados durante a predominância da variante Ômicron e suas subvariantes foram extraídos da base de dados SIVEP-Gripe, visando detectar subgrupos mais vulneráveis que requerem maior atenção. Uma caracterização preliminar da base de dados foi realizada, objetivando entender e preparar os dados. Posteriormente, por meio de técnicas de aprendizado de máquina, foram criados modelos para predição de prognóstico desfavorável, bem como para identificar as variáveis mais associadas com tais desfechos. O conjunto de dados analisado continha 36 atributos e 107.138 registros que englobam dados sociodemográficos, sintomas, comorbidades, número de doses da vacina, admissão em UTI, necessidade de suporte ventilatório, além da evolução do caso. Os resultados demonstraram que o modelo com melhor desempenho, construído com o algoritmo *Gradient Boosting Classifier*, foi capaz de predizer a evolução do paciente com *ROC-AUC* de 0,82 e *F1-Score* de 0,72. As variáveis de predição mais importantes para o modelo foram o uso de suporte ventilatório, a idade e se o paciente foi admitido em UTI. O presente trabalho apresentou o desenvolvimento de um modelo de aprendizado de máquina para predição de mortalidade dos pacientes com COVID-19 em 2022 no Brasil. Através de modelagens sucessivas, incluindo abordagens de seleção de atributos, reamostragem e calibragem de parâmetros, foi possível direcionar os modelos na melhoria da precisão da classe de interesse, visando priorizar pacientes com prognóstico desfavorável.

Palavras-chave: COVID-19. Predição. Aprendizado de Máquina. Variantes do SARS-CoV-2. Vacinação.

Abstract

COVID-19 vaccines, developed and approved for emergency use in record time, have played a key role in reducing the number of severe cases, hospitalizations, and deaths. However, the emergence of SARS-CoV-2 variants and subvariants due to virus mutations, the gradual waning immunity induced by vaccines or previous infection, and the immune escape of these variants, favor the occurrence of reinfections and symptomatic or asymptomatic disruptive infections. The present study aimed to develop models for prognosis or early identification of patients at increased risk for an unfavorable outcome in the current scenario of the predominance of variants of concern and relatively comprehensive vaccination coverage. Data from patients with COVID-19 hospitalized during the prevalence of the Omicron variant and its subvariants were extracted from the SIVEP-Gripe database, aiming to detect more vulnerable subgroups that require greater attention. A preliminary characterization of the databse was performed, seeking to understand and prepare the data. Subsequently, using machine learning techniques, models were created to predict an unfavorable prognosis, as well as to identify the variables most associated with such outcomes. The analyzed dataset contained 36 attributes and 107,138 records that include sociodemographic data, symptoms, comorbidities, number of vaccine doses, admission to the ICU, need for ventilatory support, in addition to the evolution of the case. The results showed that the model with the best performance, built with the GradientBoostingClassifier algorithm, was able to predict the patient's evolution with ROC-AUC of 0.82 and F1-Score of 0 .72. The most important predictive variables for the model were the use of ventilatory support, age and whether the patient was admitted to the ICU. The present work presented the development of a machine learning model to predict the mortality of patients with COVID-19 in 2022 in Brazil. Through successive modeling, including attribute selection approaches, resampling and parameter calibration, it was possible to direct the models to improve the accuracy of the class of interest, aiming to prioritize patients with an unfavorable prognosis.

Keywords: COVID-19. Prediction. Machine Learning. SARS-CoV-2 variants. Vaccination.

Lista de Ilustrações

Lista de Tabelas

Sumário

1 Introdução

A doença do coronavírus 19 (COVID-19), causada pelo agente etiológico *severe acute respiratory syndrome coronavirus-2* (SARS-CoV-2), é uma doença respiratória infecciosa aguda, que foi identificada pela primeira vez em dezembro de 2019, em Wuhan (China) (JIN et al., 2020). Devido ao alto potencial de disseminação, o vírus espalhou-se rapidamente e atingiu diversos países em todos os continentes. A Organização Mundial de Saúde (OMS) admitiu formalmente que se tratava de uma pandemia apenas em março de 2020 (WHO, 2020), quando já havia muitos casos e óbitos em diversos países. Até 09 de agosto de 2023, o Brasil acumulou 37.717.062 casos e 704.659 mortes por COVID-19 (WHO, 2023), números sem precedentes devido a qualquer doença infecciosa que atingiu o país anteriormente.

Os sintomas da doença variam amplamente, desde um quadro assintomático, que atinge cerca de 35% dos casos (SAH et al., 2021), até casos sintomáticos moderados, graves e críticos (WU; MCGOOGAN, 2020). De modo geral, os subgrupos da população mais vulneráveis são idosos acima de 60 anos e portadores de condições médicas subjacentes (comorbidades), tais com cardiopatia e diabetes (SOUZA et al., 2021b), além de profissionais da saúde que se encontram mais expostos ao atender pacientes com infecção.

Por se tratar de uma infecção humana com um patógeno até então desconhecido e para o qual não havia medicamentos profiláticos e terapêuticos específicos, diversas intervenções não farmacêuticas (NPIs) foram adotadas, tais como uso de máscaras faciais, distanciamento social, higienização das mãos e de ambientes, assim como confinamento (*lockdown*) em alguns casos mais extremos. A produção de vacinas em tempo recorde, usando diferentes plataformas tais como vírus atenuado (Coronavac/Sinovac-Butantan), adenovírus (Vaxzevria/Astrazeneca-Oxford-Fiocruz e Janssen/Johnson&Johnson) e mRNA (Comirnaty/Pfizer–BioNTec e Moderna), contribuíram para atenuar a pandemia (BRASIL, 2022). Entretanto, o elevado número de casos de COVID-19 propiciou o surgimento de variantes de preocupação (*variants of concern*, VOC) do SARS-CoV-2 ancestral reportado na China, provocando novos surtos da doença.

O Brasil foi atingido por quatro ondas sucessivas até o final de 2022. A primeira onda, iniciada em março de 2020, foi causada pelo SARS-CoV-2 ancestral/Linhagem B.1 que prevaleceu ao longo desse ano. A segunda onda foi provocada pelo surgimento da variante Gama/P.1, que surgiu no estado do Amazonas (SABINO et al., 2021) e se mostrou altamente infectiva e mais letal, atingindo até mesmo pessoas mais jovens (SOUZA et al., 2021c). A terceira e quarta ondas foram devido à emergência da variante Ômicron e suas subvariantes (BA.4/BA.5), que predominaram ao longo de 2022. A variante Ômicron mostrou-se altamente transmissível (RAHIMI; ABADI, 2022) com número elevado de casos, porém causou menos hospitalização e óbitos, possivelmente devido à elevada cobertura vacinal da população e/ou por se tratar de variante com características

intrínsecas menos letais (CHAKRABORTY; BHATTACHARYA; SHARMA, 2021).

A pandemia causada pela COVID-19, sem dúvida, impactou fortemente os sistemas de saúde dos países, independentemente do nível econômico, os quais continuam buscando meios e ferramentas tecnológicas para auxiliar na tomada de decisões e planejamento estratégico que auxiliem na saúde e bem-estar da população, além de reduzir a sobrecarga no atendimento (SCHULLER et al., 2022). A emergência de variantes e subvariantes de SARS-CoV-2, capazes de escapar da imunidade induzida por vacinas ou infecção prévia (GARCIA-BELTRAN et al., 2021), além da redução da imunidade ao longo do tempo (FEIKIN et al., 2022) sugerem possíveis ocorrências de novos surtos de COVID-19. Um estudo realizado com pacientes brasileiros hospitalizados, de fato, mostrou que infecção disruptiva de vacina ocorre e pode desencadear desfecho desfavorável, principalmente para pacientes idosos portadores de condições médicas subjacentes (JESUS et al., 2022).

Nesse contexto complexo, estratégias tecnológicas baseadas em Inteligência Artificial (IA) e Aprendizado de Máquina (AM), dentre elas, prognóstico precoce de desfecho desfavorável baseados em informações contidas em bancos de dados, capazes de predizer evolução de paciente hospitalizado com COVID-19, podem ser promissoras para auxiliar nas tomadas de decisões antecipadas e atendimento adequado.

1.1 Justificativa

Dados preliminares indicam que a pandemia da COVID-19 poderá se tornar endemia, com surtos periódicos da doença. Nesse contexto, estratégias baseadas em inteligência artificial e aprendizado de máquina, capazes de prever desfechos desfavoráveis de pacientes hospitalizados devido à doença, podem ser estratégias tecnológicas importantes e altamente promissoras para auxiliar na tomada de decisões antecipadas por gestores públicos. Atendimento antecipado e adequado, redução na sobrecarga hospitalar, e redução de óbitos intra-hospitalar visando a proteção da saúde e bem-estar dos pacientes pode ser auxiliado por novas tecnologias e inovações que têm agregado contribuições relevantes na área de saúde. A variante Ômicron é a VOC mais divergente observada, tendo originado muitas subvariantes, além de apresentar maior capacidade de evasão a anticorpos derivados de infecções prévias e/ou vacinas. Poucos estudos focaram na variante Ômicron utilizando dados do Brasil. Sendo assim, o estudo proposto visa realizar predição de desfechos de pacientes hospitalizados com COVID-19 num cenário complexo da pandemia devido à emergência de variantes de preocupação, que poderá contribuir para tomada de decisões adequadas no atendimento de pacientes, assim como sua aplicação para eventos futuros.

1.2 Objetivos

1.2.1 Geral

O objetivo geral deste trabalho foi desenvolver um modelo baseado em técnicas de aprendizado de máquina visando predizer o desfecho de pacientes diagnosticados com COVID-19 na prevalência da variante Ômicron e suas subvariantes.

1.2.2 Específicos

- Obter um conjunto de dados consistentes com as informações relevantes ao problema.
- Desenvolver modelos de predição eficazes por meio de algoritmos de aprendizado de máquina.
- Identificar atributos importantes para o modelo de predição.
- Compreender os resultados alcançados em comparação com aqueles descritos na literatura.

1.3 Organização do Trabalho

O restante deste trabalho foi organizado em capítulos, conforme segue. O Capítulo 2 consiste na revisão bibliográfica, envolvendo o arcabouço teórico, além de trabalhos relacionados ao tema presentes na literatura. No Capítulo 3, são detalhados os procedimentos aplicados no desenvolvimento das atividades. No Capítulo 4, são apresentados os resultados obtidos pela metodologia proposta, juntamente com análises comparativas. Finalmente, no Capítulo 5 são descritas as considerações finais do trabalho, propostas para trabalhos futuros e possíveis consequências positivas que esta área pode produzir para a sociedade.

2 Revisão Bibliográfica

A rápida evolução da pandemia da COVID-19 sobrecarregou os sistemas de saúde, motivando o uso de inteligência artificial e aprendizado de máquina como recursos auxiliares para fins de diagnóstico, predição de risco para doença grave, estimativas de taxa de infecção, novos surtos, processamento de imagens, etc. Os modelos de predição são ferramentas capazes de aprender com dados históricos e auxiliar em cenários futuros, contribuindo para melhor alocação de recursos e antecipação de medidas que possam mitigar eventos desfavoráveis. Aplicações da IA na área de biomedicina durante a pandemia geraram uma rica literatura referente a COVID-19, de forma que alguns estudos são mencionados a seguir. Além disso, um breve arcabouço teórico sobre as técnicas de AM e IA são apresentadas na sequência.

2.1 Trabalhos Relacionados

Uma visão geral das aplicações da IA em vários campos, incluindo diagnóstico da doença por meio de diferentes tipos de testes e sintomas, monitoramento de pacientes, identificação da gravidade de um paciente, processamento de exames de imagem relacionados à COVID-19, epidemiologia, estudos farmacêuticos, etc, é apresentada por Tayarani-N. (2021). Visando realizar uma pesquisa abrangente sobre as aplicações da IA no combate às dificuldades causadas pelo surto, foram levantadas diversas formas pelas quais as abordagens de IA foram empregadas nos trabalhos presentes na literatura. No artigo, foram revisadas aplicações que atingiram resultados satisfatórios, porém, ainda existe espaço para melhorias.

O artigo de Lalmuanawma, Hussain e Chhakchhuak (2020) apresenta uma revisão abrangente sobre o papel da inteligência artificial e aprendizado de máquina como uma área promissora para triagem, predição, rastreamento de contatos, desenvolvimento de medicamentos e vacinas para COVID-19, embora ainda constituam modelos não implantados suficientemente no mundo real, mas com potencial para combater a epidemia. O artigo aborda estudos recentes que aplicam algoritmos como *Neural Network*, *Support Vector Machine, XGBoost* e *Random Forest* para enfrentar a pandemia da COVID-19. Os autores também abordam alguns erros e desafios ao usar tais algoritmos em problemas reais.

O trabalho realizado por El-Rashidy et al. (2021) possui o objetivo principal de estudar o papel da inteligência artificial como uma tecnologia no combate à pandemia da COVID-19. Conforme o artigo, cinco aplicações significativas foram encontradas: diagnósticos usando diversas bases de dados, estimativa de propagação da enfermidade, associação entre a infecção e as características do paciente, desenvolvimento de vacina ou medicamentos e desenvolvimento de aplicações de suporte. Este estudo também introduz uma comparação entre as bases de dados.

O artigo de Souza et al. (2021a) apresenta um estudo para predizer prognósticos desfavoráveis decorrentes da COVID-19 utilizando de técnicas de aprendizado de máquina como *Logistic Regresion, Linear Discriminant Analysis, Naive Bayes, K-Nearest Neighbors, Decision Trees, XGBoost* e *Support Vector Machine*. O conjunto de dados apresenta 8.443 registros, com informações de cada paciente e a evolução dos casos. Os experimentos mostram que o desfecho da doença pode ser predito com uma ROC AUC de 0,92, sensibilidade de 0,88, e especificidade de 0,82, havendo possibilidade de aperfeiçoar os resultados com a inclusão de dados posteriores.

Utilizando dados obtidos na universidade *The University of Texas Medical Branch* de pacientes com teste positivo para COVID-19, o trabalho de Booth et al. (2021) constrói um modelo de prognóstico de mortalidade dos infectados. Para este propósito, foi desenvolvido um modelo do algoritmo *Support Vector Machine* usando parâmetros bioquímicos de 398 pacientes, para predizer o óbito dos pacientes em até quarenta e oito horas da coleta de material para análise. A utilidade de tal modelo provavelmente está mais em sua capacidade de integrar valores laboratoriais citados como preditores interativos de mortalidade e para orientar a discussão sobre porque tais características acabam sendo relevantes.

Uma análise dos fatores de risco de mortalidade em pacientes com COVID-19 em Wuhan, na China foi desenvolvida por Zhou et al. (2020). Informações demográficas, clínicas, de tratamento e laboratorial, incluindo diversas amostras de RNA, foram extraídas e comparadas entre sobreviventes e não sobreviventes. Resultados apontam alto risco de óbito para idosos, pacientes com risco de falência orgânica e D-Dímero superior a 1 $\mu g/mL$.

Algoritmos de aprendizado de máquina (*Logistic Regression* e *XGBoost*) para identificar fatores que contribuem para o aumento do risco de mortalidade pela infecção do SARS-CoV-2 no Brasil foram utilizados por Baqui et al. (2021). O estudo permitiu concluir que fatores socioeconômicos, geográficos e estruturais, assim como etnia, são mais importantes do que comorbidades individuais para predição de desfechos desfavoráveis para pacientes com COVID-19. Este estudo foi realizado antes do plano de vacinação da população brasileira pelo Ministério da Saúde e estudos futuros podem apresentar mudanças nos resultados apresentados.

Lodato et al. (2022) desenvolveram um modelo de aprendizado de máquina para triagem de pacientes com COVID-19 baseados em dados de prontuário e resultados de exames visando categorizar o nível de gravidade da doença e prever a mortalidade de pacientes. Foi utilizado um conjunto de dados com 429 amostras e diversos atributos adquiridos de testes sanguíneos. Os classificadores *Random Forest* e *Gradient Boosting* foram altamente precisos em prever a mortalidade dos pacientes (acurácia média de 99%), bem como categorizar os pacientes conforme o nível de gravidade da doença (acurácia média de 91%).

Jiang et al. (2020) apresentaram fatores preditivos de casos graves de COVID-19 utilizando dos algoritmos *Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest e Support Vector Machine*. O conjunto de dados aborda dados demográficos, clínicos, laboratoriais e radiográficos de pacientes com doença confirmada. Os resultados obtidos mostraram precisão

de $70 - 80\%$ nas predições.

2.2 Aprendizado de Máquina

Segundo Mitchell (1997), o aprendizado de máquina se preocupa com a questão de como construir *softwares* que melhoram automaticamente com a experiência. O desenvolvimento de modelos de aprendizado de máquina voltado para diversos domínios de aplicação demonstra seu valor prático e potencial, inclusive no domínio médico. Tais aplicações se mostram especialmente úteis em problemas onde os bancos de dados podem conter regularidades implícitas que podem ser descobertas automaticamente. Por exemplo, considere diagnosticar pacientes a partir de suas características. Após aprender um modelo a partir de dados históricos, essa aplicação deve conseguir diagnosticar corretamente novos pacientes, por sua capacidade de generalização (propriedade do AM).

Um problema de classificação consiste em identificar a qual de um conjunto de categorias pertence uma nova instância, dado um histórico usado para treinamento, que contém instâncias cuja associação de categoria é conhecida. Esse tipo de problema é resolvido por meio do aprendizado supervisionado. A tarefa de classificação pode ser binária ou multiclasse. A classificação binária é aquela em que existem dois valores de saída, como 0 ou 1, verdadeiro ou falso. Por outro lado, três ou mais valores são admitidos no contexto multiclasse.

Neste trabalho, alguns algoritmos de aprendizado de máquina supervisionado foram aplicados a um conjunto de dados contendo informações de pacientes com teste positivo para COVID-19 visando criar modelos computacionais capazes de prever o desfecho da doença (óbito ou cura). Assim, trata-se de uma tarefa de classificação binária. Para avaliar o desempenho dos modelos, diversas métricas podem ser utilizadas para mensurar aspectos diferentes dos algoritmos. A seguir, as principais métricas utilizadas na tarefa de classificação são apresentadas.

2.2.1 Avaliação

- Matriz de Confusão: é apresentada por meio de uma matriz (Figura 2.1), que apresenta as frequências de classificação para cada classe do modelo.

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

Figura 2.1 – Matriz de Confusão

  - **Verdadeiro positivo** (*true positive — TP*) ocorre quando, no conjunto real, a classe que estamos buscando foi prevista corretamente.
  - **Falso positivo** (*false positive — FP*) ocorre quando, no conjunto real, a classe que não estamos buscando prever foi prevista incorretamente.
  - **Verdadeiro negativo** (*true negative — TN*) ocorre quando, no conjunto real, a classe que não estamos buscando prever foi prevista corretamente.
  - **Falso negativo** (*false negative — FN*) ocorre quando, no conjunto real, a classe que estamos buscando prever foi prevista incorretamente.

- Acurácia (*Accuracy*): indica o desempenho geral do modelo, ou seja, dentre todas as classificações, quantas o modelo classificou corretamente:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FN + FP)}. \tag{2.1}$$

- Precisão (*Precision*): dentre todas as classificações de classe Positivo que o modelo fez, quantas estão corretas:

$$Precision = \frac{(TP)}{(TP + FP)}. \tag{2.2}$$

- Revocação (*Recall*): dentre todas as situações de classe Positivo como valor esperado, quantas estão corretas:

$$Recall = \frac{(TP)}{(TP + FN)}. \tag{2.3}$$

- *F1-Score*: média harmônica entre precisão e revocação:

$$\text{F1-Score} = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)}. \tag{2.4}$$

- *AUC-ROC (Area Under the Curve of the Receiver Operating Characteristic)*: corresponde à área sob a curva ROC, a qual é calculada através da razão de verdadeiros positivos, em relação à razão de verdadeiros negativos. É uma forma de mensurar a eficiência do modelo, quanto maior for a área sob a curva, melhor será a capacidade do modelo de distinguir a classe positiva da negativa. Uma pontuação de 1 na métrica significa a perfeição do modelo.

  A Figura 2.2 ilustra a representação da curva ROC em relação a um classificador aleatório. Pode-se observar que uma reta no gráfico significa que o modelo apresenta características de classificação aleatória. Quanto maior a curva e mais próxima de 1, melhor a eficiência na predição do modelo e quanto mais próximo de 0 estiver a curva, denota que o modelo está classificando a classe positiva como negativa ou reciprocamente.

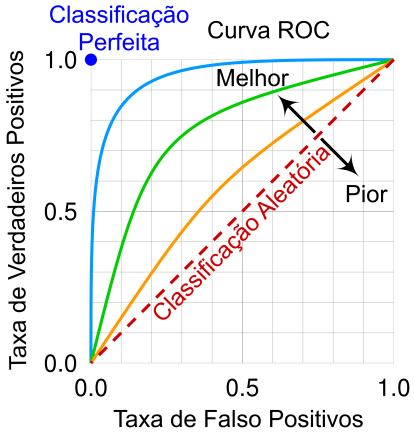


Figura 2.2 – Imagem representativa da Curva ROC

2.2.2 Padronização (*Standard*)

Os modelos de aprendizado de máquina e inteligencia artificial tendem a se comportar melhor e gerar resultados preferíveis quando a base de dados está normalizada ou padronizada sobre uma determinada métrica.

- *Z-Score Standard*: *Z-Score*, ou pontuação padrão, é uma maneira de descrever um ponto de dados em termos de sua relação com a média e o desvio padrão de um grupo de pontos. O objetivo de obter *Z-scores* é remover os efeitos da localização e da escala dos dados, permitindo que diferentes conjuntos de dados sejam comparados diretamente. A intuição por trás do método Z-Score de detecção de *outliers* é que, uma vez que tenhamos centralizado e redimensionado os dados, qualquer valor que esteja muito longe de zero deve ser considerado um *outlier*. O valor de *Z-Score* é calculado conforme a equação abaixo.

$$\text{Z-Score} = \frac{x - \text{Média}}{\text{Desvio Padrão}}. \tag{2.5}$$

2.2.3 Algoritmos

- Regressão Logística (*Logistic Regression*): A Regressão Logística é um algoritmo de classificação usado para atribuir obervações em um grupo de classes discretas. O algoritmo é utilizado para estimar a probabilidade que uma instância da base pertença a uma classe em particular. O objetivo no treinamento do modelo é encontrar o conjunto de parâmetros que melhor estima as instâncias positivas com uma alta probabilidade e as negativas com uma baixa.

  Diferente da regressão linear que tem como saída um número contínuo, a regressão logística transforma a saída usado a função *Sigmoid* para retornar o valor da probabilidade que pode mapear as classes discretas.

$$S(z) = \frac{1}{1 + e^{-z}}$$

A saída da função *Sigmoid* retorna o valor da probabilidade entre 0 e 1. Para mapear as classes discretas (Verdadeiro/Falso), precisamos selecionar um valor limite que irá classificar os valores entre duas classes (0 e 1), conforme o exemplo a seguir:

$$\begin{aligned} p \geq 0.5, \text{classe} = 1 \\ p < 0.5, \text{classe} = 0. \end{aligned} \tag{2.6}$$

Para a Regressão Logística com múltiplas classes, podemos selecionar a classe com maior probabilidade predita (STOLTZFUS, 2011).

- Árvore de Decisão (*Decision Tree*): Uma árvore de decisão é um algoritmo de aprendizado supervisionado utilizado para tarefas de classificação e regressão. Possui uma estrutura hierárquica em árvore, que consiste em um nó raiz, ramificações, nós internos e nós folhas. Uma árvore de decisão começa com um nó raiz, que não possui nenhuma ramificação de entrada. As ramificações de saída do nó raiz alimentam os nós internos, também conhecidos como nós de decisão.

  Uma árvore de decisão funciona dividindo sucessivamente a base de dados em segmentos menores até o subconjunto ter somente uma variável de decisão, ou até a base de dados não poder mais ser dividida. É um algoritmo guloso, ou seja, ele encontra a melhor decisão no dado momento sem se preocupar com a otimização global.

  A construção de uma árvore de decisão segue essas regras:

  1. Atribuir todas as instâncias de treinamento para a raiz da árvore.
  2. Encontrar o melhor atributo ou valor de divisão baseado em um critério previamente determinado.
  3. Particionar todas as instâncias em seus nós conforme a instância ou valor de divisão.
  4. Denotar cada partição como um nó folha do nó atual.
  5. Para cada nó filho:
     - Se for um nó "puro" (Somente um atributo presente no nó), definir o nó como nó folha e retornar.
     - Caso contrário, definir o nó filho como nó atual e retornar a etapa 2.

  As árvores de decisão podem variar segundo o critério de decisão (Ganho, Taxa de ganho ou Coeficiente de Gini MK-Gurucharan (2022)), o tipo das instâncias da base de dado (Categóricos e/ou Numéricos), o tipo do problema (Classificação e/ou Regressão) e o tipo da árvore (Binária ou N-ária) (QUINLAN, 1986).

- Floresta Aleatória (*Random Forest*): O *Random Forest* é um algoritmo de aprendizado de máquina frequentemente utilizado que combina a saída de múltiplas árvores de decisão em um único resultado. Estas árvores de decisão são geradas randomicamente ou seguindo um critério previamente definido. O algoritmo prediz a saída calculando a média dos valores das árvores.

  Sua facilidade de uso e flexibilidade acentua sua adoção, por lidar com problemas de classificação e regressão. Comparada a uma única árvore de decisão, a *Random Forest* diminui suas limitações, reduz o *Overfitting* (não generalização do modelo) e aumenta a precisão (LOUPPE, 2015).

- *Naive Bayes*: O classificador *Naive Bayes* é um algoritmo de aprendizado de máquina supervisionado, usado para tarefas de classificação. Também faz parte de uma família de algoritmos de aprendizado generativo, ou seja, visa modelar a distribuição de entradas de uma determinada classe ou categoria.

  Ao contrário dos classificadores discriminativos, como a regressão logística, ele não aprende quais recursos são mais importantes para diferenciar as classes. Para os algoritmos *Naive Bayes* de classificação, todos os atributos são independentes do valor de outros recursos, portanto, o algoritmo considera que cada um dos atributos contribui para a probabilidade na classificação de forma independente e igual.

- *Gradient Boosting Classifier*: O *Gradient Boosting Classifier* é um algoritmo de aprendizado de máquina, baseado em árvore de decisão que utiliza uma estrutura de *Gradient Boosting*. Cada árvore de decisão é adicionada ao modelo por vez e ajustadas para corrigir os erros de predição realizados pelos modelos anteriores. Este é um tipo de modelo em aprendizado de máquina conhecido como *boosting* (PEDREGOSA et al., 2011).

2.2.4 Treinamento e Testes

Para estimar a capacidade de generalização dos modelos, são geralmente adotadas duas estratégias de divisão da base.

- *Hold-out*: O método de *Hold-out*, como apresentado na Figura 2.3, é uma técnica utilizada no treinamento de modelos de aprendizado de máquina que se resume a dividir a base de dados em subconjuntos independentes, um para a etapa de treinamento, um para validação e outro para a etapa de teste. O método de *Hold-out* é usado para verificar a qualidade do modelo de aprendizado de máquina no desempenho em novos dados, garantindo o poder de generalização dos modelos. Geralmente a divisão é de 70%-30%, onde os 70% da base é usada para treinamento e 30% da base usado para testar os modelos Bennett et al. (2022).

- Validação cruzada (*Cross Validation*): é uma técnica para avaliar os modelos por meio de treinamento em subconjuntos de dados de entrada disponíveis e avaliação dos mesmos no

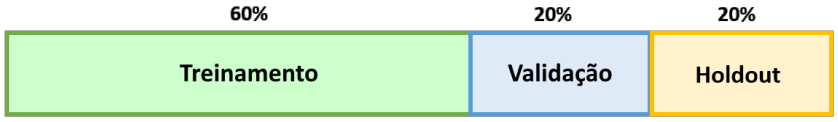


Figura 2.3 – Hold-Out

subconjunto complementar dos dados. A validação cruzada é usada para evitar sobreajuste (*overfitting*), ou seja, a não generalização de um padrão no modelo. Como apresentado na Figura 2.4, um número definido de iterações é realizada na base de dados a dividindo em blocos aleatórios que serão separados entre treino e teste. Os blocos de teste não são repetidos durante as iterações e a métrica usada na avaliação é definida previamente.

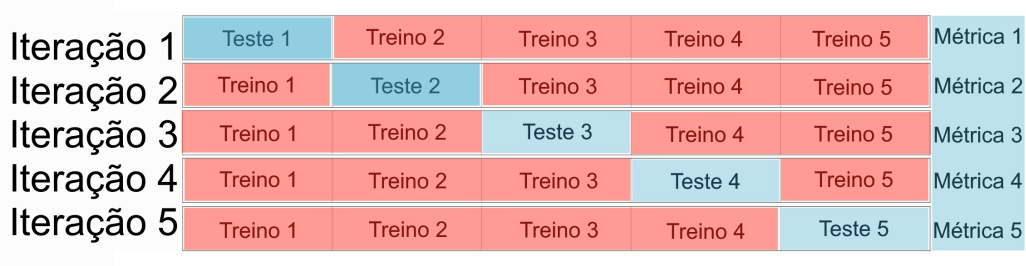


Figura 2.4 – Validação Cruzada

2.2.5 Seleção de Atributos

A seleção de atributos (*Feature Selection*) é um processo de redução do número de variáveis no desenvolvimento de um modelo de predição. É desejável a redução das variáveis de entrada para reduzir o custo computacional do modelo e, em alguns casos, aprimorar os resultados do modelo. Uma estratégia é apresentada a seguir:

- *RFE - Recursive Feature Elimination*: Com o auxílio de um estimador externo que atribui pesos para os atributos, o objetivo da RFE é selecionar atributos da base recursivamente, considerando subconjuntos decrescentes de atributos. Primeiramente o estimador é treinado com um conjunto inicial de atributos, e assim, a importância de cada atributo é obtida. Em seguida, o atributo de menor impacto na predição é eliminado do subconjunto atual e este processo é repetido até que o conjunto (sub)ótimo de atributos é encontrado (PEDREGOSA et al., 2011).

2.2.6 Balanceamento de Classes

Em casos onde a base de dados apresenta desbalanceamento da classe alvo, pode ser interessante a aplicação de alguma técnica para aproximar o tamanho dos dois conjuntos. Algumas possíveis técnicas consistem na remoção e/ou adição de novas instâncias/amostras.

- *Random Under Sampling*: O *Random Under Sampling* é uma técnica para balanceamento de base de dados que remove randomicamente amostras da classe majoritária até obter um equilíbrio de amostras (LEMAÎTRE; NOGUEIRA; ARIDAS, 2017).

- *Random Over Sampling*: O *Random Over Sampling* é uma técnica para balanceamento de base de dados que adiciona randomicamente cópias das amostras da classe minoritária até obter um equilíbrio de amostras (LEMAÎTRE; NOGUEIRA; ARIDAS, 2017).

- *SMOTE (Synthetic Minority Over-sampling Technique)*: *SMOTE* é uma abordagem de *over-sampling* na qual a classe minoritária é sobreamostrada criando exemplos "sintéticos" ao longo da linha de segmentos que se juntam a qualquer ou todos os k vizinhos mais próximos da classe minoritária. Ela cria dados de treinamento extras realizando essas operações em dados reais. Dependendo da quantidade de sobreamostragem necessária, vizinhos dos k-vizinhos mais próximos são aleatoriamente escolhidos. Essa técnica não garante um equilíbrio na quantidade de amostras das classes minoritárias com as classes majoritárias (LIKEBUPT; PETERCLU; CHMCCL, 2022; CHAWLA et al., 2002).

2.2.7 Otimização dos Hiper Parâmetros

Os modelos em aprendizado de máquina apresentam diversos parâmetros que podem ser alterados conforme a base de dados ou com as necessidades dos modelos. Existem alguns algoritmos que realizam modificações nesses parâmetros a fim de melhorar o desempenho do modelo a cerca de métricas pré-definidas.

- *Grid Search*: O *grid search* é um algoritmo que recebe por parâmetro um dicionário de parâmetros para o estimador, e realiza um busca por força bruta em todos as possíveis combinações de parâmetros. Para aprovação dos resultados ele embute da validação cruzada e a métrica de decisão precisa ser pré-definida (LIASHCHYNSKYI; LIASHCHYNSKYI, 2019; PEDREGOSA et al., 2011).

2.2.8 Comparação dos Modelos

Para realizar a comparação entre os modelos, testes estatísticos podem ser utilizados para avaliar o nível de significância das diferenças de desempenho entre eles.

- *ANOVA*: Análise da Variância (*ANOVA*) é um método para testar a igualdade de três ou mais médias populacionais, baseado na análise das variâncias amostrais. O algoritmo compara as médias de diferentes populações para verificar se essas populações possuem médias iguais ou não.

O método se baseia nas seguintes hipóteses:

$$H_0 : Grupo_1 = Grupo_2 = ... = Grupo_k$$
$$H_A : \text{Ao menos uma média dos grupos difere}$$

O resultado da análise (Teste F) pode ser calculado usando a seguinte fórmula:

$$TesteF = \frac{MS_b}{MS_w} \tag{2.7}$$

onde, MS_{b} seria a variância média entre amostras e MS_{w} a variância média nas amostras.

A hipótese de igualdade H_0 é aceita caso o resultado do Teste F seja maior que o teste F crítico (ST»HLE; WOLD, 1989).

- *P-Value*: Em estatística, o p-valor é a probabilidade de obter os resultados observados de um teste, assumindo que a hipótese nula está correta. É o nível de significância marginal em um teste de hipótese estatístico, que representa a probabilidade da ocorrência de um determinado evento. Um valor de p menor significa haver uma evidência mais forte a favor da hipótese alternativa.

  Esta é uma das métricas usadas na Análise da Variância (*ANOVA*), para calcular se os valores de dois grupos de resultados se diferem (NAHM, 2017).

3 Metodologia

A linguagem utilizada nas atividades foi Python (versão 3.10.12) devido à facilidade de tratamento e manipulação de dados, além de existirem diversas bibliotecas relacionadas a ela. As principais bibliotecas utilizadas foram Numpy [1] (versão 1.23.5) para manipulação de listas e cálculos, Pandas [2] (versão 1.5.3) para análise de dados, Matplotlib [3] (versão 3.7.1) no auxílio para geração dos gráficos e Seaborn [4] (versão 0.12.2) como o principal gerador de gráficos. Para modelagem dos algoritmos e comparações, as bibliotecas utilizadas foram Scikit-learn [5] (versão 1.2.2) na geração dos modelos de aprendizado de máquina, Imbalanced-learn [6] (versão 0.10.1) para a modelagem em bases de dados desbalanceadas, Yellowbrick [7] (versão 1.5) na análise de importância dos atributos, Scikit-plot [8] (versão 0.3.7) para auxiliar na construção dos gráficos e Scipy [9] (versão 1.10.1) para realizar o teste ANOVA.

O desenvolvimento deste trabalho foi estruturado e organizado tendo como base a metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*) (AZEVEDO; SANTOS, 2008), idealizada em meados dos anos 90, e ainda válida nas tarefas atuais de ciência de dados. Essa metodologia é subdividida em seis etapas principais: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação, sendo que a última não fez parte do escopo desse trabalho.

3.1 Entendimento do negócio

O primeiro passo provém da ideia de utilizar o legado dos bancos de dados epidemiológicos de Síndrome Respiratória Aguda Grave (SRAG), disponibilizados pelo Ministério da Saúde por meio do Sistema de Informação de Vigilância Epidemiológica da Gripe (SIVEP-Gripe). Além de englobar a vigilância da Influenza e outros vírus respiratórios, o sistema teve os dados da vigilância da COVID-19 incorporados a partir de 2020 (BRASIL, 2019). Na revisão da literatura, foi constatado que o uso de técnicas de aprendizado de máquina no prognóstico de pacientes vêm sendo adotado como forma de auxílio à tomada de decisão e também para melhor compreender as variáveis mais associadas a desfechos desfavoráveis. Assim, sob posse desses dados, o objetivo é propor um modelo de aprendizado de máquina e inteligência artificial capaz de predizer a

[1] <https://numpy.org/>
[2] <https://pandas.pydata.org/>
[3] <https://matplotlib.org/>
[4] <https://seaborn.pydata.org/>
[5] <https://scikit-learn.org/stable/>
[6] <https://imbalanced-learn.org/stable/>
[7] <https://www.scikit-yb.org/en/latest/>
[8] <https://pypi.org/project/scikit-plot/>
[9] <https://scipy.org/>

evolução do caso dos pacientes que testaram positivo para o SARS-CoV-2 e suas variantes.

3.2 Entendimento dos dados

A base de dados do SIVEP-Gripe possui 166 atributos e na data de seu *download*[10], contava com 557.476 registros de hospitalização por SRAG (Síndrome Respiratória Aguda Grave) em todo o Brasil. Esses dados não estão necessariamente atrelados à COVID-19, ou seja, parte dos registros e atributos consistem em informações relacionadas a outras doenças respiratórias causadas por agentes tais como vírus da influenza, vírus sincicial respiratório, adenovírus, entre outros. Após uma análise prévia do dicionário de dados, parte dos atributos foram descartados por não contribuírem diretamente na proposta deste trabalho.

A Tabela 3.1 sumariza os atributos considerados relevantes para o estudo proposto e detalha cada um deles.

Tabela 3.1 – Tabela de variáveis retiradas da base de dados do SIVEP-Gripe

Nome	Tipo da variável	Descrição	Valores
TP_IDADE	Data	Data de nascimento do paciente.	DD/MM/AAAA
NU_IDADE_N	Quantitativa contínua	Idade informada pelo paciente quando não se sabe a data de nascimento.	0 a 100
SG_UF_NOT	Qualitativa Nominal	Unidade Federativa onde está localizada a Unidade que realizou a notificação.	AC, AL, AM, AP, BA, etc.
CS_RACA	Qualitativa Nominal	Identificador de raça	Branca, Preta, Amarela, Parda, Indígena, Não Informado
DOSE_1_COV	Data	Data em que o paciente recebeu a 1ª dose da vacina COVID-19.	DD/MM/AAAA
DOSE_2_COV	Data	Data em que o paciente recebeu a 2ª dose da vacina COVID-19.	DD/MM/AAAA
DOSE_REF	Data	A data em que o paciente recebeu a dose reforço.	DD/MM/AAAA
DT_INTERNA	Data	Data em que o paciente foi hospitalizado.	DD/MM/AAAA
DT_SIN_PRI	Data	Data de 1º sintomas do caso.	DD/MM/AAAA
DT_EVOLUCA	Data	Data da alta ou óbito.	DD/MM/AAAA
VACINA_COV	Binário Assimétrico	Se o paciente recebeu vacina COVID-19	1-Sim, 2-Não e 9-Ignorado
UTI	Qualitativa Ordinal	O paciente foi internado em UTI?	1-Sim, 2-Não e 9-Ignorado
SUPORT_VEN	Qualitativa Ordinal	O paciente fez uso de suporte ventilatório	1 - Sim, invasivo 2 - Sim, não invasivo, 3 - Não e 9 - Ignorado
EVOLUCAO	Binário Assimétrico	Evolução do caso	Cura ou Óbito
FEBRE	Binário Assimétrico	Se o paciente apresentou febre	Sim ou Não
TOSSE	Binário Assimétrico	Se o paciente apresentou tosse	Sim ou Não
GARGANTA	Binário Assimétrico	Se o paciente apresentou dor de garganta	Sim ou Não
DISPNEIA	Binário Assimétrico	Se o paciente apresentou dispneia	Sim ou Não
DESC_RESP	Binário Assimétrico	Se o paciente apresentou dificuldade de respiração	Sim ou Não
SATURACAO	Binário Assimétrico	Se o paciente apresentou saturação O2 < 95%	Sim ou Não
DIARREIA	Binário Assimétrico	Se o paciente apresentou diarreia	Sim ou Não
VOMITO	Binário Assimétrico	Se o paciente apresentou vômitos	Sim ou Não
DOR_ABD	Binário Assimétrico	Se o paciente apresentou dores abdominais	Sim ou Não
FADIGA	Binário Assimétrico	Se o paciente apresentou fadiga	Sim ou Não
PERD_OLFT	Binário Assimétrico	Se o paciente apresentou perda de olfato	Sim ou Não
PERD_PALA	Binário Assimétrico	Se o paciente apresentou perda de paladar	Sim ou Não
OUTRO_SIN	Binário Assimétrico	Se o paciente apresentou outros sintomas	Sim ou Não
PUERPERA	Binário Assimétrico	Paciente é puérpera ou parturiente	Sim ou Não
CARDIOPATI	Binário Assimétrico	Paciente possui Doença Cardiovascular Crônica	Sim ou Não
HEMATOLOGI	Binário Assimétrico	Paciente possui Doença Hematológica Crônica	Sim ou Não
SIND_DOWN	Binário Assimétrico	Paciente possui Síndrome de Down	Sim ou Não
HEPATICA	Binário Assimétrico	Paciente possui Doença Hepática Crônica	Sim ou Não
ASMA	Binário Assimétrico	Paciente possui Asma	Sim ou Não
DIABETES	Binário Assimétrico	Paciente possui Diabete mellitus	Sim ou Não
NEUROLOGIC	Binário Assimétrico	Paciente possui Doença Neurológica	Sim ou Não
PNEUMOPATI	Binário Assimétrico	Paciente possui outra pneumopatia crônica	Sim ou Não
IMUNODEPRE	Binário Assimétrico	Paciente possui Imunodeficiência ou Imunodepressão	Sim ou Não
RENAL	Binário Assimétrico	Paciente possui Doença Renal Crônica	Sim ou Não
OBESIDADE	Binário Assimétrico	Paciente possui obesidade	Sim ou Não
OUT_MORBI	Binário Assimétrico	Paciente possui outro(s) fator(es) de risco	Sim ou Não

3.3 Preparação dos dados

Ao trabalhar com uma base de dados ao nível nacional, é provável se deparar com dados inconsistentes, ausentes e incorretos. Portanto, para obter resultados válidos e precisos, é imprescindível que os dados passem por um processo de filtragem.

[10] A base de dados é referente ao dia 03/04/2023 e compreende dados do ano de 2022.

A primeira etapa da filtragem foi manter somente registros contendo os valores de cura e óbito por COVID-19 em evolução do caso (essa coluna também apresentava óbito por outras causas e valores ausentes). Em seguida foi filtrada a idade dos pacientes, mantendo apenas aqueles maiores de 18 anos. Essa decisão é baseada no esquema vacinal adotado no Brasil, no qual jovens e crianças iniciaram a vacinação mais tardiamente. Como a base é referente a casos de hospitalizados por síndrome respiratória aguda grave, existem registros e atributos de pacientes internados não relativos à COVID-19 que foram, portanto, descartados. Foram considerados somente pacientes que realizaram testes RT-PCR ou Antígeno para o SARS-CoV-2, visando confirmações mais precisas da doença.

Nem todos os atributos estavam diretamente disponíveis na base de dados. Foram criadas novas colunas por meio de organização e agrupamentos. A unidades federativas onde foram realizadas as notificações foram agrupadas nas cinco regiões do país: Norte, Nordeste, Centro-Oeste, Sudeste e Sul. Sobre o suporte ventilatório, foi atribuído a VMI (Ventilação Mecânica Invasiva), a VNI (Ventilação Não-Invasiva) e a não necessidade de suporte ventilatório em um único atributo.

O número de sintomas é um atributo adicionado derivado da soma dos sintomas apresentados pelo paciente. Este atributo, teoricamente, pode variar de 0 a 13 e aparenta ser extremamente correlacionado com a taxa de óbitos. O número de comorbidades segue a mesma linha de raciocínio. Este atributo, teoricamente, pode variar de 0 a 13 e também pode auxiliar na predição dos pacientes.

Outro atributo adicionado é a semana do ano que o paciente detectou seu primeiro sintoma. Como a COVID-19 possui ondas durante o ano, podemos detectar em qual momento o paciente se infectou da doença. Essa métrica pode ser útil na predição dos modelos.

Os dias decorridos pós vacinação é um atributo adicionado a fim de tentar incrementar na base de dados o fator de diminuição da imunidade da vacina. Se o paciente apresenta muitos dias da última dose, seu organismo tende a perder a imunização. Este fator será considerado na predição.

A vacinação dos pacientes foi uma forma de filtrar inconsistências na base. Nesses casos, pacientes que apresentavam incoerências nas informações de vacinação foram removidos. Outra filtragem conduziu-se através das datas de vacinação, de início da vacinação e de doses de reforços, também visando remover registros inconsistentes. O atributo *Doses* foi criado com base nos atributos de datas de vacinação.

Como resultado, obteve-se um conjunto de dados com 36 atributos e 107.138 registros que englobam dados sociodemográficos, sintomas, comorbidades, número de doses da vacina, admissão em UTI, necessidade de suporte ventilatório, além da evolução do caso. A Tabela 3.2 sumariza os atributos do conjunto de dados final.

Algumas técnicas de transformação foram aplicadas na preparação dos dados visando

Tabela 3.2 – Tabela de variáveis filtradas e com os novos atributos

Nome	Tipo da variável	Descrição	Valores
Sexo	Binário Simétrico	Identificador de sexo	M ou F
Idade	Quantitativa contínua	Idade do paciente	18 a 114
Região	Qualitativa Nominal	Identificador de região do paciente	Southeast, South, Midwest, North, Northeast
Doses	Qualitativa Ordinal	Quantidade de doses da vacina	0, 1, 2, 3
UTI	Binário Assimétrico	Se o paciente foi para a UTI	Sim ou Não
Ventilação	Qualitativa Nominal	Ventilação Mecânica Não-Invasiva ou Ventilação Mecânica Invasiva	Não, NIV ou IMV
Semana do primeiro sintoma	Quantitativa contínua	Qual semana do ano o paciente sentiu o primeiro sintoma	0 a 52
Número de sintomas	Quantitativa contínua	Quantidade de sintomas sentido pelo paciente	0 a 13
Número de comorbidades	Quantitativa contínua	Quantidade de comorbidades do paciente	0 a 13
Dias decorridos pós vacinação	Quantitativa contínua	Quantidade em dias que o paciente tomou a ultima dose da vacina	0 a 685
Evolução	Binário Assimétrico	Evolução do caso	Cura ou Óbito
Febre	Binário Assimétrico	Se o paciente apresentou febre	Sim ou Não
Tosse	Binário Assimétrico	Se o paciente apresentou tosse	Sim ou Não
Dor de Garganta	Binário Assimétrico	Se o paciente apresentou dor de garganta	Sim ou Não
Dispneia	Binário Assimétrico	Se o paciente apresentou dispneia	Sim ou Não
Dificuldade de Respiração	Binário Assimétrico	Se o paciente apresentou dificuldade de respiração	Sim ou Não
Saturação Sanguínea	Binário Assimétrico	Se o paciente apresentou saturação O2 < 95%	Sim ou Não
Diarreia	Binário Assimétrico	Se o paciente apresentou diarreia	Sim ou Não
Vômito	Binário Assimétrico	Se o paciente apresentou vômitos	Sim ou Não
Dores Abdominais	Binário Assimétrico	Se o paciente apresentou dores abdominais	Sim ou Não
Fadiga	Binário Assimétrico	Se o paciente apresentou fadiga	Sim ou Não
Perda de Olfato	Binário Assimétrico	Se o paciente apresentou perda de olfato	Sim ou Não
Perda do Paladar	Binário Assimétrico	Se o paciente apresentou perda de paladar	Sim ou Não
Outros Sintomas	Binário Assimétrico	Se o paciente apresentou outros sintomas	Sim ou Não
Cardiopatia	Binário Assimétrico	Paciente possui Doença Cardiovascular Crônica	Sim ou Não
Hematológica	Binário Assimétrico	Paciente possui Doença Hematológica Crônica	Sim ou Não
Síndrome de Down	Binário Assimétrico	Paciente possui Síndrome de Down	Sim ou Não
Hepática	Binário Assimétrico	Paciente possui Doença Hepática Crônica	Sim ou Não
Asma	Binário Assimétrico	Paciente possui Asma	Sim ou Não
Diabetes	Binário Assimétrico	Paciente possui Diabete mellitus	Sim ou Não
Neurológica	Binário Assimétrico	Paciente possui Doença Neurológica	Sim ou Não
Pneumopatia	Binário Assimétrico	Paciente possui outra pneumopatia crônica	Sim ou Não
Imunodepressão	Binário Assimétrico	Paciente possui Imunodeficiência ou Imunodepressão	Sim ou Não
Renal	Binário Assimétrico	Paciente possui Doença Renal Crônica	Sim ou Não
Obesidade	Binário Assimétrico	Paciente possui obesidade	Sim ou Não
Outras Comorbidades	Binário Assimétrico	Paciente possui outro(s) fator(es) de risco	Sim ou Não

obter melhores resultados nos modelos. A técnica de padronização dos dados utilizada foi a *Z-Score Standard* (subseção 2.2.2). Dessa forma, os valores são estrategicamente padronizados para descrever um ponto de dados em termos de sua relação com a média e o desvio padrão de um grupo de pontos.

3.4 Modelagem

Nesta etapa, foram construídos modelos baseados nos seguintes algoritmos: *Logistic Regression*, *Decision Trees*, *Random Forest*, *Naive Bayes*, e *Gradient Boosting Classifier* (subseção 2.2.3). A escolha desses algoritmos é baseada na literatura e por apresentarem princípios de funcionamento diversificados. Apesar dos dados terem sidos preparados anteriormente, ainda existem algumas técnicas de modelagem a fim de tratar desbalanceamento de dados, seleção de atributos e calibragem de parâmetros.

Primeiramente, os dados foram separados em dois conjuntos, um conjunto de teste e um conjunto de treinamento. Dessa forma, podemos garantir uma maneira de testar os modelos. A taxa de divisão escolhida foi 70% para a base de treinamento e 30% para a base de teste. O *overfitting* acontece quando um modelo estatístico se ajusta muito bem ao conjunto de dados anteriormente observado, mas se mostra ineficaz para prever novos resultados, não permitindo uma generalização dos dados. Alternativamente, pode-se utilizar *cross-validation* (subseção 2.2.4).

Seguidamente, foi utilizado o processo de *Feature Selection*. Essa técnica pode usar

heurísticas de seleção de atributos da base, ordenando-as entre mais e menos importante para o modelo. Assim, pode-se remover atributos desnecessários para o modelo, diminuindo a complexidade sem perder a precisão. Outras técnicas baseadas em análise de variância e correlação podem ser também empregadas. Para fins de comparação, foram utilizadas abordagens com e sem *Feature Selection* (subseção 2.2.5).

É esperado que bases de dados reais apresentem desbalanceamento nos dados. Nesse caso, existem técnicas de *undersampling* e *oversampling* para redução ou aumento no número de instâncias, respectivamente. Para contornar essa questão, foram utilizadas as técnicas *Random Over Sampler*, *Random Under Sampler* e *SMOTE* (subseção 2.2.6).

Por fim, é preciso otimizar os hiper parâmetros do modelo, ou seja, encontrar os melhores parâmetros que se adaptam a base de dados e o objetivo especificado. A técnica utilizada foi o *GridSearch*. O algoritmo realiza uma busca completa por todas as possibilidades em um subconjunto de hiper parâmetros do modelo de treinamento e retorna a que apresentou melhores resultados (subseção 2.2.7).

3.5 Avaliação do modelo

Na etapa de avaliação dos modelos, foram utilizadas métricas a fim de compará-los (subseção 2.2.1). Assim, mediante a testes estatísticos é possível comprovar quais modelos apresentaram os resultados mais satisfatórios (subseção 2.2.8).

4 Resultados

Neste capítulo são apresentados resultados do estudo proposto. A seção 4.1 consiste na caracterização do conjunto de dados estudado, analisando cada uma das variáveis da base de dados em gráficos, juntamente com sua frequência em relação à variável de evolução.

A seção 4.2 apresenta os resultados dos modelos desenvolvidos. Neste caso, são apresentados gráficos que explicitam a eficiência dos modelos e os comparam quanto ao desempenho de predição, juntamente com a importância dos atributos.

4.1 Caracterização do conjunto de dados

Observando cada uma das variáveis da base após sua preparação, é possível extrair informações que ajudam a entender melhor os dados. Por meio de visualização de dados, a evolução dos pacientes (variável resposta) é explorada para as diversas variáveis. Os gráficos apresentam barras azuis que se referem à taxa de cura para aquela variável (em porcentagem) e as barras laranjas à taxa de óbito.

Na Figura 4.1, é possível observar que temos dados com proporções bem distribuídas quanto ao atributo sexo do paciente. Outra observação a ser investigada é porcentagem de óbito de pessoas do sexo masculino ser superior ao feminino, mesmo apresentando uma quantidade menor de casos.

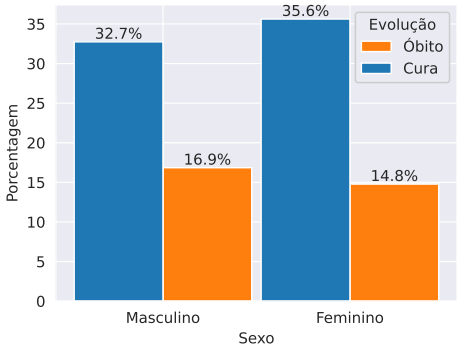


Figura 4.1 – Distribuição do sexo atrelada a taxa de óbitos

A Figura 4.2 apresenta a taxa, em porcentagem, de pacientes hospitalizados de acordo com sua idade, assim podemos comparar a taxa de óbito e cura sobre essa variável. Neste caso, podemos concluir que a COVID-19 atinge, em maioria, pessoas acima de 50 anos e o grupo de risco são pessoas acima de 60 anos, devido à elevação na taxa de óbito para pacientes nessa faixa de idade.

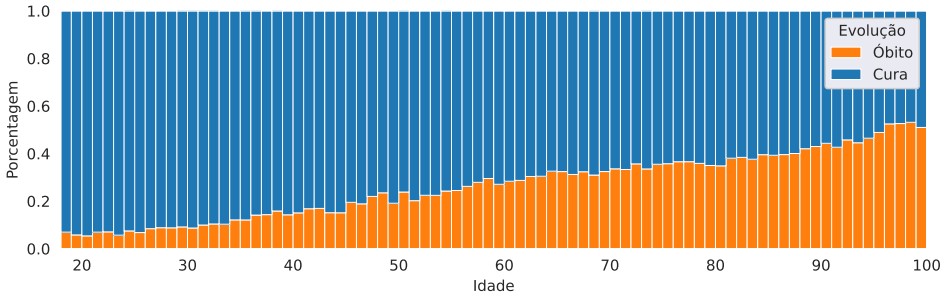


Figura 4.2 – Distribuição da idade atrelada a taxa de óbitos

Na Figura 4.3, temos a quantidade de pacientes internados conforme a região do país. A região com mais casos é a sudeste, seguida da região Sul. Esse comportamento é esperado, devido aos tamanhos populacionais das regiões. Outra observação a ser investigada seria a taxa de letalidade superior na região Nordeste, onde 37.7% dos casos progrediram ao óbito.

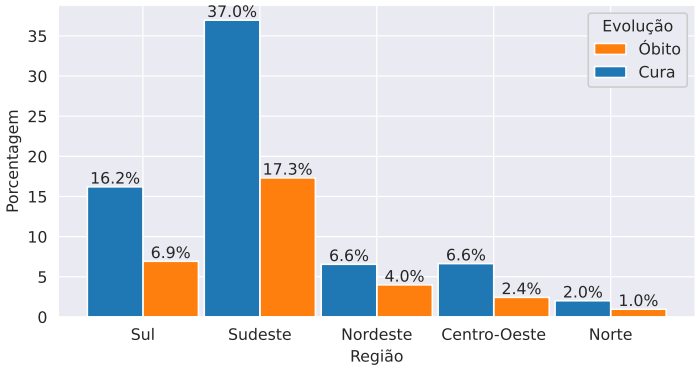


Figura 4.3 – Distribuição das regiões atrelada a taxa de óbitos

Na Figura 4.4, podemos observar a quantidade de pacientes que vieram a óbito segundo a quantidade de doses da vacina contra a COVID-19. Observam-se muitas pessoas não vacinadas na base de dados de 2022. Outra observação a ser investigada é a taxa semelhante de óbito para pacientes que não se imunizaram e aqueles que se protegeram com 2 doses, 33,6% e 32,8% respectivamente, porém, há uma leve redução na taxa de letalidade em pacientes com 3 doses, 28,9%.

Na Figura 4.5, podemos observar a quantidade de pacientes internados que precisaram ser admitidos em uma unidade de terapia intensiva (UTI). Observa-se uma elevada taxa de letalidade nesse subgrupo, logo que, mais pacientes que adentraram a UTI vieram a óbito do que a cura, portanto, sendo considerado os casos mais graves apresentados. Outra observação a ser investigada é a qualidade hospitalar das UTI em garantir a saúde de um paciente em estado grave.

Na Figura 4.6, temos pacientes que precisaram de Ventilação Mecânica Invasiva (VMI), Ventilação Não-Invasiva (VNI) ou nenhum tipo de ventilação mecânica. Nesses casos podemos considerar a utilização da Ventilação Mecânica Invasiva nos casos extremamente graves, acarretando uma elevada taxa de letalidade e a utilização da Ventilação Não-Invasiva em casos menos

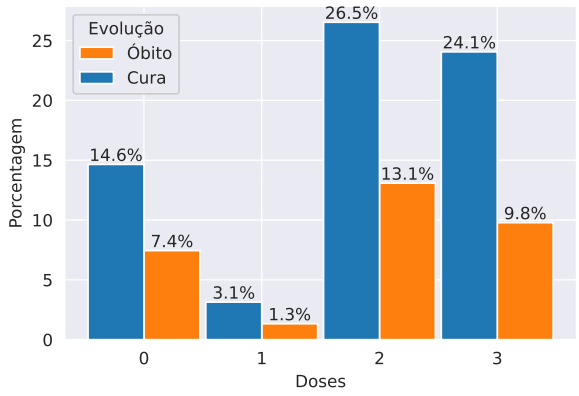


Figura 4.4 – Distribuição da quantidade de doses da vacina atrelada a taxa de óbitos

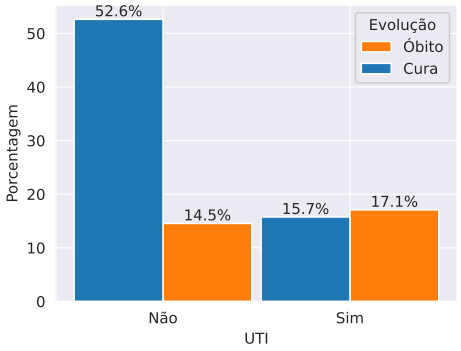


Figura 4.5 – Distribuição de pacientes que necessitaram de UTI atrelada a taxa de óbitos

graves, apesar da elevação na taxa em comparação aos pacientes que não precisaram de nenhum tipo de ventilação mecânica. Outra observação a ser investigada é se esse tipo de ventilação está realmente ajudando os enfermos ou se esse mecanismo é utilizado tardiamente.

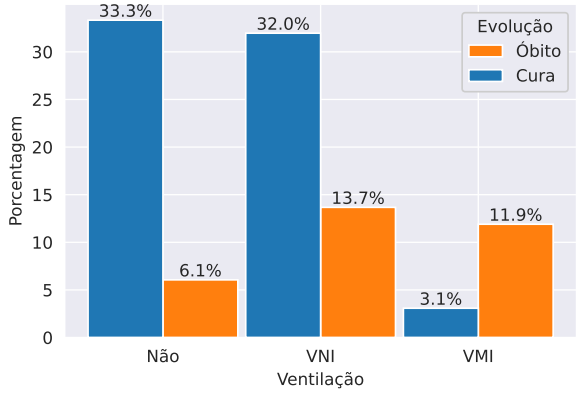


Figura 4.6 – Distribuição de pacientes que necessitaram de VMI, VNI ou nenhuma ventilação mecânica atrelada a taxa de óbitos

Na Figura 4.7, podemos conferir qual semana do ano o paciente teve seu primeiro sintoma detectado. É possível observar que na base de dados, temos 3 épocas do ano com elevadas taxas de infecção. Outra observação é a não existência de uma diferença brusca na taxa de óbito, mesmo

em pacientes que buscaram um centro hospitalar nesses picos, porém essas são as épocas com maiores quantidades de óbito.

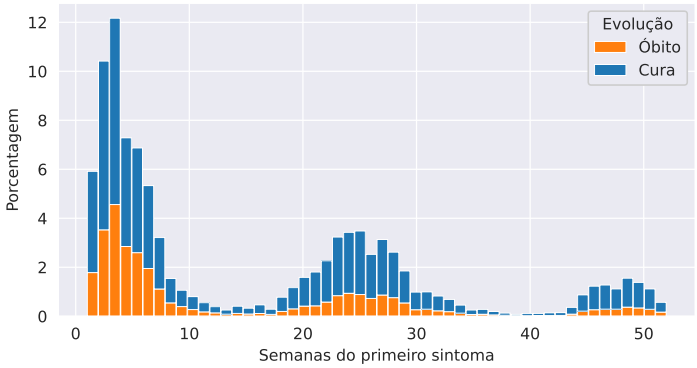


Figura 4.7 – Distribuição da semana do ano na qual o paciente teve o primeiro sintoma, atrelada a taxa de óbitos

Na Figura 4.8, podemos conferir a diferença em dias que o paciente se imunizou com uma dose da vacina contra a COVID-19, com a data de entrada em uma unidade de saúde, ocasionando em uma diminuição da imunidade no paciente. Outra observação a ser investigada é a inexistência de uma variação na taxa de óbitos para pessoas que decorreram mais dias da sua última vacinação.

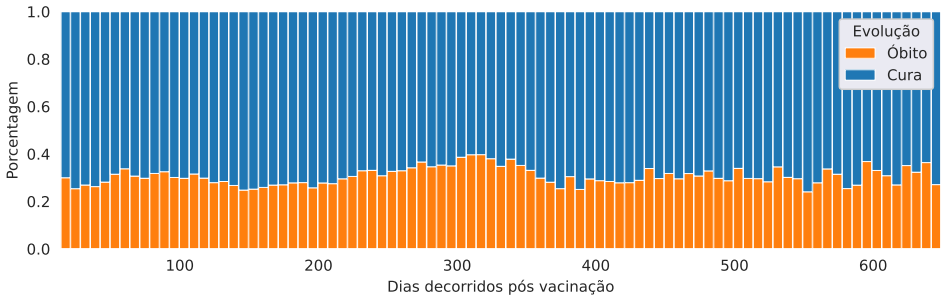


Figura 4.8 – Distribuição de quantos dias o paciente está sem tomar uma dose da vacina atrelada a taxa de óbitos

Na Figura 4.9, podemos observar a quantidade de sintomas que um único enfermo apresentou durante o período de infecção da COVID-19. No gráfico, a maioria dos pacientes registrou de 3 a 4 sintomas diferentes e é possível analisar a correlação do aumento da taxa de óbito com o elevado número de sintomas.

Na Figura 4.10, podemos observar a ocorrência de sintomas nos pacientes, entre os quais, tiveram mais ocorrências na taxa de letalidade. É possível observar que os sintomas mais recorrentes são, respectivamente: tosse, dispneia, baixa saturação de oxigênio no sangue, desconforto respiratório e febre. Os sintomas com as maiores taxas de letalidade são, respectivamente: desconforto respiratório, baixa saturação de oxigênio no sangue, dispneia e fadiga.

Na Figura 4.11, podemos observar a quantidade de comorbidades diferentes que um único enfermo possui. No gráfico, a maioria dos pacientes apresenta de nenhuma a 2 comorbidades e é

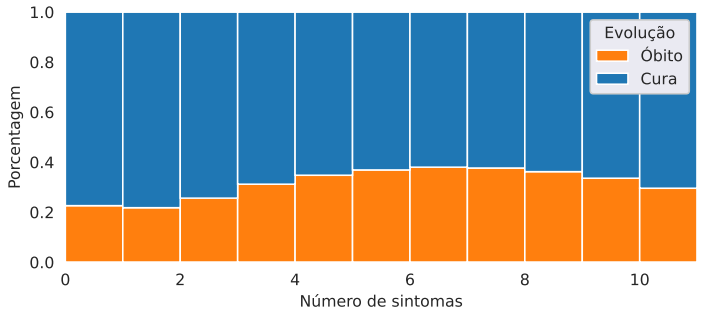


Figura 4.9 – Distribuição da quantidade de sintomas apresentado pelo paciente atrelada a taxa de óbitos

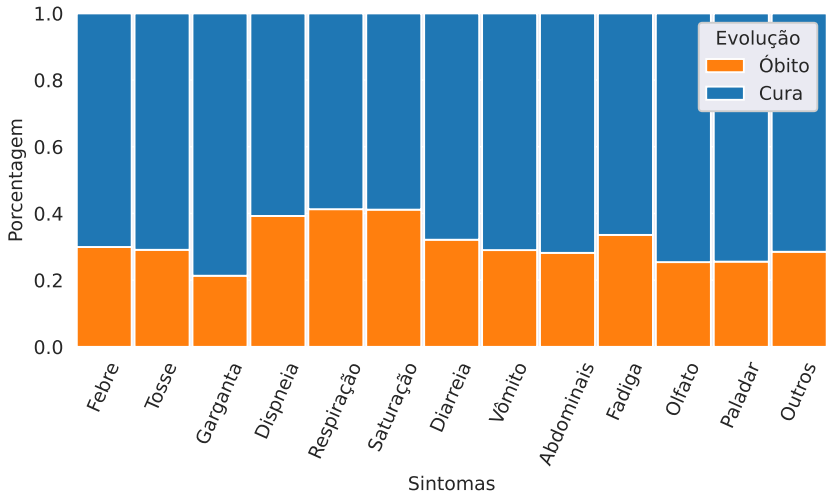


Figura 4.10 – Distribuição de sintomas apresentado pelo paciente atrelada a taxa de óbito

possível analisar a alta correlação do aumento na taxa de óbito com a quantidade de comorbidades.

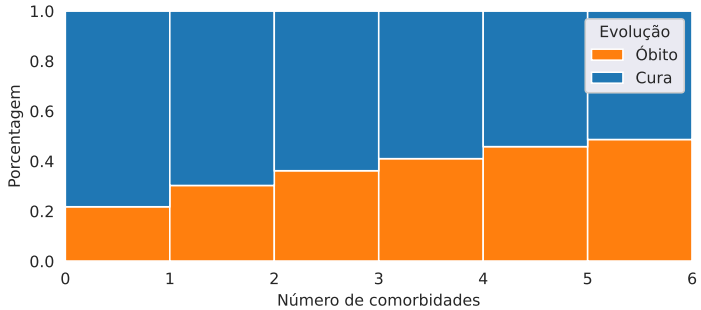


Figura 4.11 – Distribuição da quantidade de comorbidades do paciente atrelada a taxa de óbito

Na Figura 4.12, temos a proporção de pacientes com comorbidades hospitalizados atrelada à taxa de óbito. As comorbidades mais frequentes são, respectivamente: cardiopatia, diabetes e as que foram marcadas como outras comorbidades no registro. É possível observar que as comorbidades associadas a taxas mais altas de letalidade são, respectivamente: doença Hepática crônica, doença renal crônica, imunodeficiência e pneumopatia crônica.

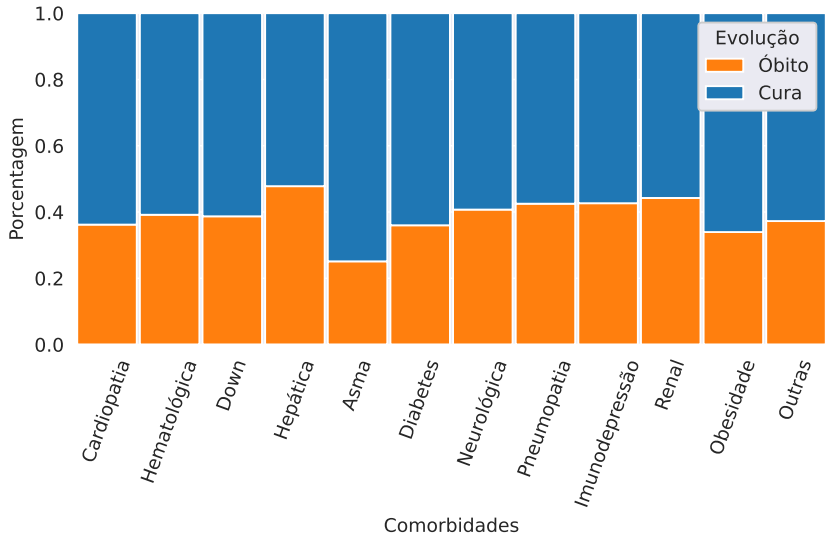


Figura 4.12 – Distribuição das comorbidades dos pacientes atrelada a taxa de óbito

4.2 Modelos de predição

Esta seção apresenta informações dos resultados dos modelos baseados em cada um dos algoritmos citados anteriormente. As métricas envolvidas na avaliação dos modelos foram a precisão, acurácia, *recall* e *f1-score*. Para auxiliar seu entendimento, foram gerados gráficos das saídas obtidas na validação cruzada, a matriz de confusão da predição na base de teste, assim como a curva ROC e o *p-value* do teste de variância *ANOVA*.

4.2.1 Primeira Modelagem

Para a primeira modelagem, foram utilizados algoritmos para predizer a evolução do paciente na base de dados pré-processada sem nenhum incremento de seleção de atributos, balanceamento de classes ou otimização dos hiper parâmetros. Esta é uma etapa inicial para encontrar os melhores modelos para o problema e iniciar os processos comparativos.

Na Figura 4.13, pose-se observar a média dos resultados apresentados pelos modelos na validação cruzada.Os modelos que apresentaram melhores métricas foram, respectivamente: *Gradient Boosting Classifier*, *Logistic Regression* e *Random Forest Classifier*. Outra obervação a ser investigada é o baixa valor na métrica de *recall*. Por termos menos amostras para valores de óbito do que cura, estamos causando um desbalanceamento na predição.

Por fim temos a Figura 4.14, que realiza comparações entre as métricas de saída da validação cruzada dos modelos utilizando o teste *ANOVA*, garantido a não igualdade dos modelos. Os resultados apontam igualdade entre os modelos *Random Forest Classifier* e *Logistic Regression* na acurácia, *Random Forest Classifier* e *Gradient Boosting Classifier* na revocação, *Random Forest Classifier* e *Logistic Regression* na *F1-Score*, e *Gradient Boosting Classifier* e *Logistic Regression* na *F1-Score*. Podemos concluir que estes modelos além de apresentarem métricas

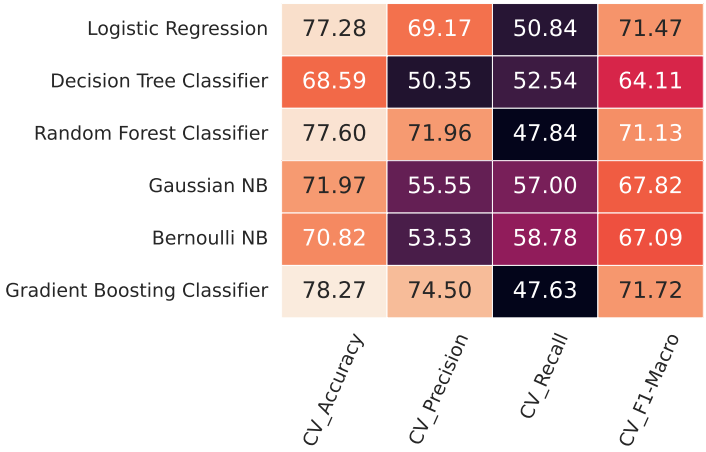


Figura 4.13 – Média das métricas na validação cruzada de cada modelo na primeira modelagem

melhores, indicam uma similaridade nos resultados, portanto serão os escolhidos para futuros testes.

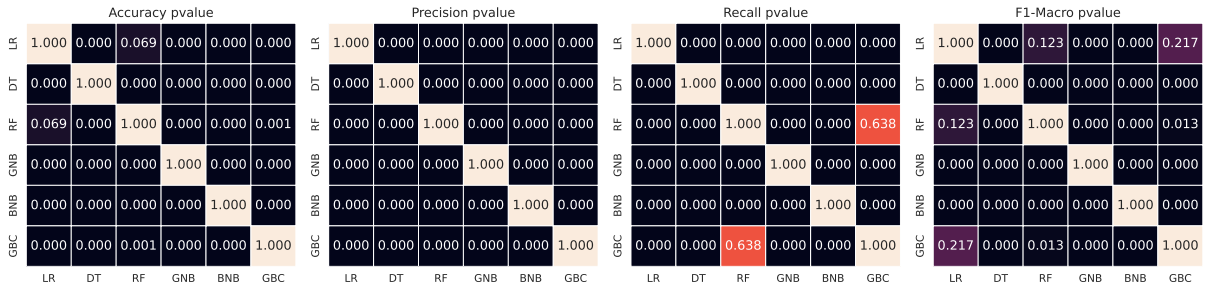


Figura 4.14 – *ANOVA* comparativa de cada modelo na primeira modelagem

4.2.2 Segunda Modelagem usando Seleção de atributos

Para esta segunda modelagem, foram escolhidos como os melhores modelos da primeira modelagem, o *Logistic Regression*, o *Random Forest Classifier* e o *Gradient Boosting Classifier*. Em seguida foi realizada a seleção de atributos utilizando o algoritmo *RFE - Recursive Feature Elimination*. O estimador escolhido foi a *Random Forest Classifier*, por ser um algoritmo sólido para classificação de classes, sem nenhuma otimização dos hiper parâmetros. Os experimentos da seleção de atributos resultaram na seleção de 17 atributos explicitados a seguir:

- Semanas primeiro sintoma
- Número de sintomas
- Dispneia
- UTI
- Diminuição da imunidade
- Outros Sintomas
- Idade
- Diabetes

- Dificuldade de respiração
- Cardiopatia
- Sexo
- Febre
- Número de comorbidades
- Região
- Ventilação
- Doses
- Saturação sanguínea

Em seguida foi aplicada a mesma base de dados da primeira modelagem, porém limitada aos 17 atributos selecionados. As métricas estão focadas em comparar os modelos com a seleção de atributos e sem a seleção de atributos, a fim de decidir seus benefícios aos resultados.

Na Figura 4.15, podemos observar a média dos resultados apresentados pelos modelos com seleção de atributos na validação cruzada. A diferença entre os valores será comparada juntamente com o teste de variância *ANOVA*.

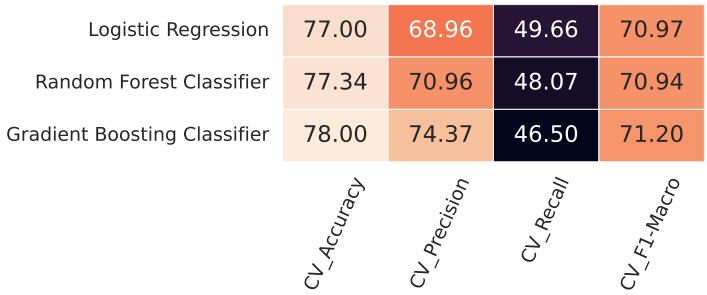


Figura 4.15 – Média das métricas na validação cruzada de cada modelo com seleção de atributos na segunda modelagem

Como podemos observar na Figura 4.16, no teste de variância *ANOVA*, várias métricas apresentaram *p-value* superior a 0,05, portanto podemos deduzir uma similaridade entre os resultados dos modelos com seleção de atributos e sem seleção de atributos. Dessa maneira, a fim de garantir uma otimização dos modelos, será utilizada a base de dados com a seleção de atributos.

Para selecionar o melhor modelo entre os três apresentados na continuação das modelagens, utilizam-se o *F1-Score* da validação cruzada apresentados na Figura 4.15 e o teste de variância *ANOVA* da Figura 4.17. Observa-se que os resultados *ANOVA* gerados da métrica *F1-Score* são superiores a 0,05, portanto podemos deduzir uma similaridade entre os modelos e o escolhido será o *Gradient Boosting Classifier* por apresentar o maior *F1-Score*.

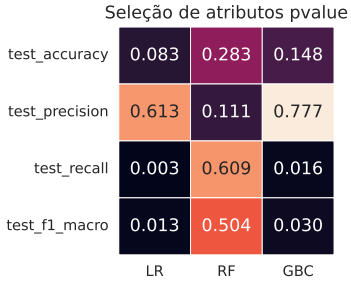


Figura 4.16 – *ANOVA* entre os modelos com seleção de atributos e sem seleção de atributos na segunda modelagem

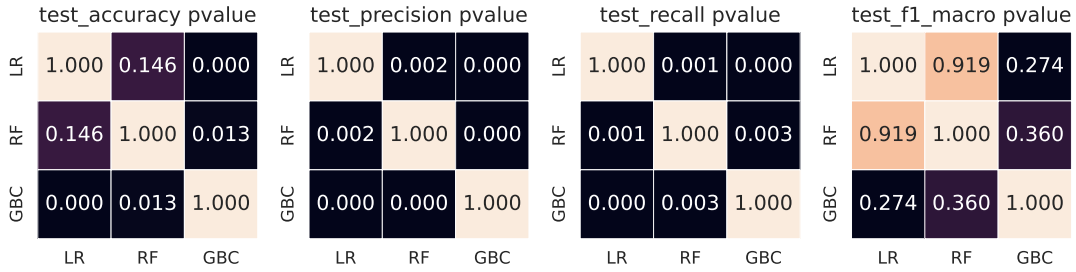


Figura 4.17 – *ANOVA* entre os modelos com seleção de atributos na segunda modelagem

4.2.3 Terceira Modelagem usando Balanceamento de Classes

Como mencionado anteriormente na primeira modelagem, a métrica de *recall* apresentou valores baixos devido ao desbalanceamento da base de dados. Nessa modelagem foram utilizados três algoritmos diferentes para realizar o balanceamento: *Random Over Sampler*, *Random Under Sampler* e *SMOTE*.

Estes algoritmos foram aplicados em um *pipeline* na validação cruzada, a fim de evitar *overfitting* do modelo. Ao se aplicar estas técnicas sem o pipeline, é possível que as cópias das amostras estejam na base de treinamento e na base de teste (MARTIN, 2019).

Como pode ser observado na Figura 4.18, os modelos apresentaram o valor de *F1-Score* semelhantes, isto fica mais evidente na Figura 4.19.

Uma leve queda na acurácia do modelo ocorreu com o balanceamento de classes. Isso se deve ao fato dele estar focado em predizer pacientes que evoluíram ao óbito, portanto, um bom indicativo. Pode-se verificar essa diferenciação de foco da predição com as métricas *precision* e *recall*. A métrica *precision* apresenta queda acentuada nos modelos com balanceamento de classes por predizer erroneamente mais a classe positiva, porém, a métrica *recall* apresenta ascensão acentuada nos modelos com balanceamento de classes por predizer corretamente mais a classe negativa.

Para fins de comparação entre os algoritmos, também foi aplicado o teste de variância *ANOVA* entre as métricas, como pose-se conferir na Figura 4.19. Nas métricas de acurácia, precisão e *recall*, somente os algoritmos *Random Over Sampler* e *Radom Under Sampler* se

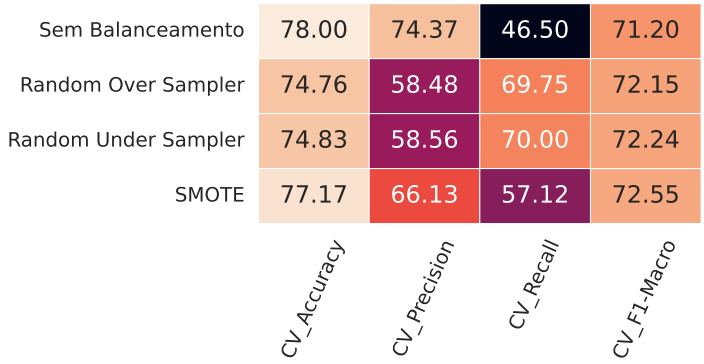


Figura 4.18 – Média das métricas na validação cruzada do modelo *Gradient Boosting Classifier* com balanceamento de classes e sem balanceamento de classes na terceira modelagem

assemelham. Na métrica *F1-Score*, os algoritmos se assemelham, mas diferem da modelagem sem balanceamento de classe.

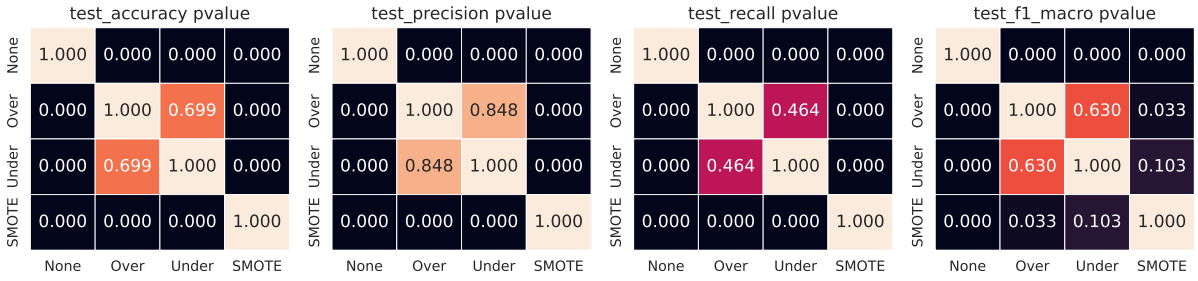


Figura 4.19 – *ANOVA* entre o modelo *Gradient Boosting Classifier* com balanceamento de classes e sem balanceamento de classes na terceira modelagem

Para seguir com a modelagem, foi decidido manter o algoritmo *Random Under Sampler*, por ser superior aos demais na métrica *recall* e apresentar melhores resultados que o sem balanceamento de classes na classe óbito. O algoritmo *SMOTE*, apesar de também apresentar métricas semelhantes e melhores resultados quanto a predição da classe positiva, tem como funcionamento a criação de novas amostras sintéticas, o que para um modelo de aprendizado de máquina na saúde não é ideal. O algoritmo *Random Over Sampler* apresentou resultados semelhantes ao *Random Under Sampler*, porém demanda maior otimização dos modelos por possuir mais amostras.

4.2.4 Quarta Modelagem com Otimização dos Hiper Parâmetros

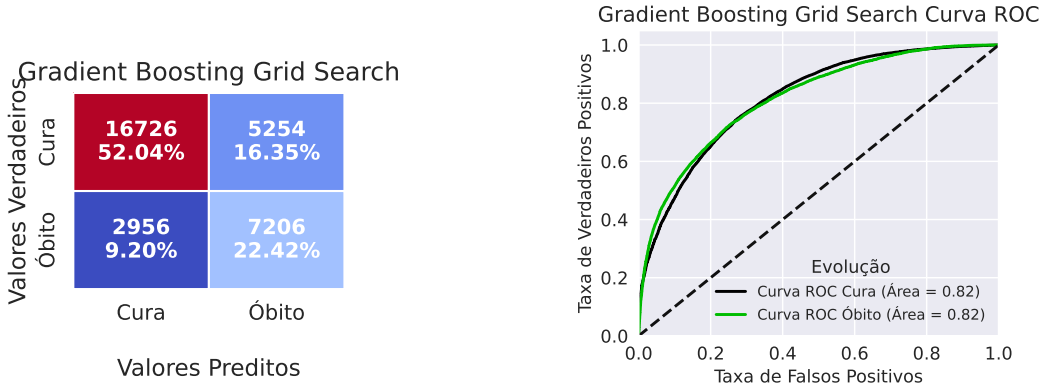
Na quarta e última modelagem, foi utilizado o algoritmo *Gradient Boosting Classifier*, juntamente com a seleção de atributos e balanceamento de classes na otimização dos seus hiper parâmetros com o algoritmo *Grid Search*. A métrica utilizada para comparar os parâmetros foi a *Z-Score*, por apresentar um balanceamento quanto ao poder de predição do modelo quanto as

classes. Os parâmetros utilizados nessa otimização foram baseados nos parâmetros padrão do algoritmo:

- "*learning_rate*": (0.01, 0.1, 0.5),
- "*max_depth*": (3, 10, 20),
- "*subsample*": (0.25, 0.5, 1.0),
- "*n_estimators*": (50, 100, 200)

Como resultados obtivemos um modelo com *F1-Score* na validação cruzada de 72,22%. Os parâmetros encontrados foram *learning_rate* igual a 0,1, *max_depth* igual a 3, *n_estimators* igual a 200 e *subsample* igual a 0,5.

A saída gerada a partir da base de teste está descrita através da matriz de confusão na Figura 4.20a. Pode-se verificar que o modelo prediz similarmente as classes cura e óbito com uma taxa de acerto de 76.1% e 70.9%, respectivamente. Na Figura 4.20b, pode-se verificar que não obteve-se aumento na área sob a curva ROC.



(a) Matriz de confusão

(b) Curva ROC

Figura 4.20 – Matriz de confusão e curva ROC do modelo *Gradient Boosting Classifier* com seleção de atributos, balanceamento de classes e otimização dos hiper parâmetros

Na Figura 4.21, pode-se conferir quais foram os atributos mais importantes na predição da evolução do modelo *Gradient Boosting Classifier* com seleção de atributos, balanceamento de classes e otimização dos hiper parâmetros.

O atributo mais importante para o modelo é se o paciente necessitou de algum tipo de ventilação. Como mostrado na Figura 4.6, estes pacientes apresentaram uma elevação acentuada na taxa de óbitos, principalmente aqueles que necessitaram de ventilação mecânica invasiva. A segunda métrica mais importante é a idade. A Figura 4.2 explicita que o aumento da idade eleva a taxa de óbito. O terceiro atributo mais importante é se o paciente foi admitido em UTI; como explicitado na Figura 4.5, há uma elevada taxa de óbitos para esses pacientes.

Durante todo o processo de modelagem, pode-se verificar que em nenhum momento houve um aumento no poder predição dos modelos. As mudanças estavam restritas as métricas *precision* e *recall*. Portanto, à medida que o modelo se aperfeiçoava na predição da classe óbito,

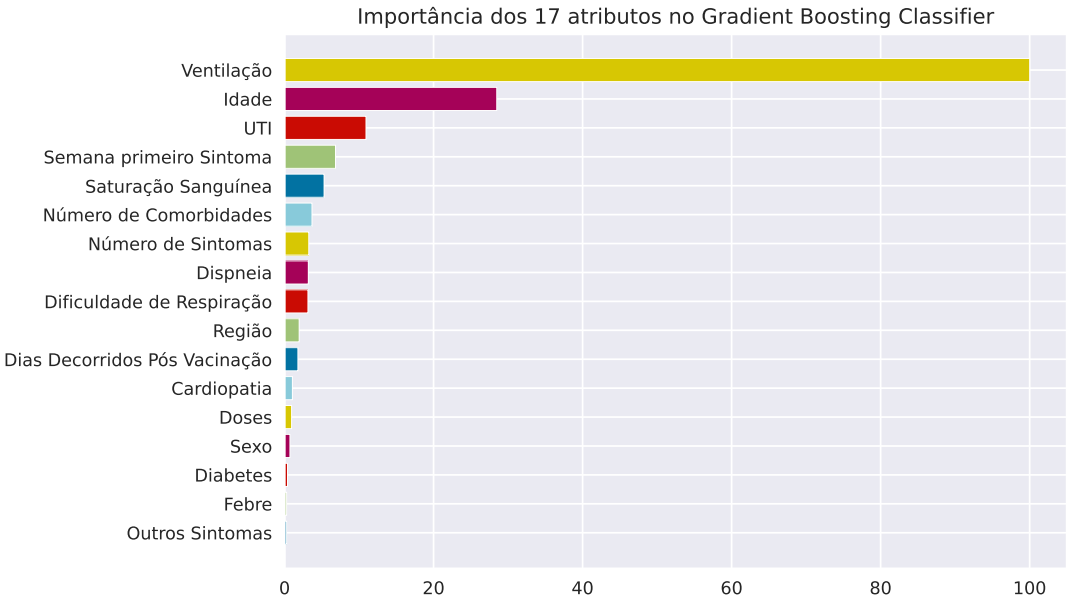


Figura 4.21 – Importância dos atributos do modelo *Gradient Boosting Classifier* com seleção de atributos, balanceamento de classes e optimização dos hiper parametros

prejudicava na predição da classe cura. O objetivo final era encontrar um modelo que prediz igualmente as classes com prioridade na classe óbito.

4.3 Comparação com a literatura

A fim de comparar os modelos da literatura com o apresentado neste trabalho, foi construída uma tabela (Tabela 4.1) comparativa mostrando informações de interesse como: o país de estudo, a base de dados, os modelos de aprendizado de máquina utilizados, os melhores algoritmos e resultados obtidos.

Podemos extrair da tabela que o modelo apresentado neste trabalho, manteve resultados similares aos estudos apresentados no Brasil em anos anteriores, porém, difere dos estudos apresentados em outros países. Um destes fatores é a ausência de dados adicionais, que englobem parâmetros sanguíneos dos pacientes, o que foi demonstrado ser relevante na predição da evolução dos pacientes nestas pesquisas.

Tabela 4.1 – Comparação dos resultados com trabalhos relacionados

Autores	País	Hospitalizados	N. Instâncias	Algoritmos Avaliados	Melhor Algoritmo	Seleção de Atributos	Balanceamento	Hiperparâmetros	ROC-AUC	F1-Macro	Variáveis Importantes
De Souza, et al. (2021)	Brasil	Geral	8.443	Logistic Regression, LDA, Naive Bayes, K-Nearest Neighbors, Decision Trees, XGBoost, SVM	Logistic Regression, LDA XGBoost, SVM	Não	Sim	Sim	0.92	-	Idade Doenças Respiratórias Comorbidades
Baqui, et al. (2021)	Brasil	Hospitalizados	231.112	XGBoost Logistic Regression	XGBoost	Sim	Não	Não	0.813	-	Idade Estado Núm. Comorbidades
Couto, et al. (2022)	Brasil	Hospitalizados	49.197	Random Forest Logistic Regression XGBoost	XGBoost	Sim	Não	Sim	0.803	-	Idade Obesidade Doenças renais
Figuerêdo, et al. (2021)	Brasil	Hospitalizados	232.164	Decision Tree Naive Bayes	Decision Tree	Sim	Sim	Sim	0.8382	-	ICU Idade Suporte Ventilatório
Moulaei, et al. (2021)	Iran	Hospitalizados	1.500	Decision Tree Random Forest XGBoost K-Nearest Neighbors Multilayer Perceptron Logistic Regression Naive Bayes	Random Forest	Sim	Sim	Sim	0.99	-	Idade Sexo Índice de massa corporal
Patel A.B. et al. (2021)	EUA	Hospitalizados	3.597	Logistic Regression Ensemble Gaussian Process Naive Bayes K-Nearest Neighbors SVM Decision Tree Neural Network	Ensemble based	Não	Sim	Sim	-	0.88	Filtração Glomerular Linfócito Granulócito Neutrófilo Suporte Ventilatório
Shanbehzadeh, et al. (2021)	Iran	Hospitalizados	1353	Multilayer Perceptron SVM K-Nearest Neighbors Naive Bayes Random Forest Decision Tree Bayesian Network	Bayesian Network	Sim	Não	Sim	0.97	-	Granulócito Neutrófilo Linfócito Perda de Paladar
Pourhomayoun, et al. (2022)	Diversos	Hospitalizados	2.670.000	SVM Neural Networks Random Forest Decision Tree Logistic Regression K-Nearest Neighbor	Random Forest	Sim	Sim	Sim	0.94	-	-
Esta proposta	Brasil	Hospitalizados	107.138	Logistic Regression Decision Tree Random Forest Gaussian Naive Bayes Bernoulli Naive Bayes Gradient Boosting Classifier	Gradient Boosting Classifier	Sim	Sim	Sim	0.82	0.7222	Suporte Ventilatório Idade UTI

5 Considerações Finais

Este trabalho visou o desenvolvimento um modelo de predição de prognóstico desfavorável em pacientes hospitalizados diagnosticados com COVID-19. Para isso, utilizou-se do banco de dados epidemiológicos de Síndrome Respiratória Aguda Grave (SRAG) do ano de 2022, disponibilizados pelo Ministério da Saúde por meio do Sistema de Informação de Vigilância Epidemiológica da Gripe (SIVEP-Gripe). Essa proposta teve como objetivo contribuir no auxílio à tomada de decisão em momentos de recursos limitados e priorização de grupos mais vulneráveis, buscando identificar pacientes vulneráveis à infecção e, portanto, prioritários para receber atendimento especial para contornar sua possível evolução desfavorável.

Os resultados demonstraram que modelos de aprendizado de máquina e inteligência artificial conseguiram predizer, inicialmente, de 76.1% a 92.2% a classe cura e a de 47.3% a 58.9% a classe óbito. Modelagens posteriores, utilizando abordagens de seleção de atributos e balanceamento de classes, apontaram o *Gradient Boosting Classifier* como melhor algoritmo para o problema. Por fim, foi gerado um modelo capaz de predizer, corretamente, 73.5% dos pacientes. Quanto aos pacientes com evolução desfavorável (óbito), ou seja, a circunstância mais importante, o modelo apresentou uma precisão de 70.9%. Para o modelo, os atributos mais importantes na predição da evolução de paciente foram: ventilação mecânica, idade e UTI.

A abundância e qualidade dos dados coletados utilizados é uma das características mais fortes deste estudo, porém, ele ainda possui limitações. Variáveis associadas à ocupação, nível social e renda familiar poderiam ser fatores importantes para a predição dos modelos (BAQUI et al., 2021). Outros fatores importantes não presentes na base de dados compreendem parâmetros sanguíneos, que mostraram ser atributos relevantes na predição dos modelos ((SUBUDHI; VERMA; AL, 2021; SHANBEHZADEH; OROOJI; KAZEMI-ARPANAHI, 2021)). Por fim, embora o modelo não tenha apresentado um desempenho de excelência, foi possível construir um modelo de aprendizado de máquina para predição de mortalidade dos pacientes com COVID-19 em 2022, com dados sobre a vacinação, e compará-los com outros trabalhos anteriores.

Apesar da vacinação não ter demostrado uma importância excepcional na predição do desfecho dos pacientes, estudos envolvendo mais dados sobre a vacinação seriam interessantes para melhor explicitar seu papel na contenção das formas mais graves da COVID-19. É importante ressaltar que a base de dados em estudo refere-se a pacientes hospitalizados, portanto, aqueles casos já considerados mais graves. Assim, como trabalhos futuros, poderiam ser analisados dados de 2023, uma vez que a vacinação em curso considerou até cinco doses para grupos prioritários. Outros estudos relevantes a serem feitos incluem analisar pacientes não hospitalizados e a utilização de atributos mais específicos, que não estavam disponíveis para este estudo.

Referências

AZEVEDO, A.; SANTOS, M. Kdd, semma and crisp-dm: A parallel overview. p. 182–185, 01 2008.

BAQUI, P.; MARRA, V.; M., A. A.; IOANA, B.; ARI, E.; SCHAAR, M. van der. Comparing covid-19 risk factors in brazil using machine learning: the importance of socioeconomic, demographic and structural factors. *Scientific Reports*, 2021. ISSN 2045-2322. Disponível em: <https://doi.org/10.1038/s41598-021-95004-8>.

BENNETT, M.; NEKOUEI, M.; MEHTA, A. P. R.; KLECZYK, E.; HAYES, K. *Methodology to Create Analysis-Naive Holdout Records as well as Train and Test Records for Machine Learning Analyses in Healthcare*. 2022.

BOOTH; L., A.; ABELS; ELIZABETH; MCCAFFREY; PETER. Development of a prognostic model for mortality in covid-19 infection using machine learning. *Modern Pathology*, v. 34, n. 3, 2021. ISSN 1530-0285. Disponível em: <https://doi.org/10.1038/s41379-020-00700-x>.

BRASIL. *Ministério da Saúde. OpenDataSUS*. 2019. Disponível em: <https://opendatasus.saude.gov.br/>.

BRASIL. *Ministério da Saúde. Plano Nacional de Operacionalização da vacinação contra COVID-19. 12ª Ed.* https://www.gov.br/saude/pt-br/coronavirus/publicacoes-tecnicas/guias-e-planos/plano-nacional-de-operacionalizacao-da-vacinacao-contra-covid-19.pdf: [s.n.], 2022. [Acessado em 13 de março de 2023].

CHAKRABORTY, C.; BHATTACHARYA, M.; SHARMA, A. R. Present variants of concern and variants of interest of severe acute respiratory syndrome coronavirus 2: Their significant mutations in s-glycoprotein, infectivity, re-infectivity, immune escape and vaccines activity. *Reviews in Medical Virology*, Wiley, v. 32, n. 2, jun. 2021. Disponível em: <https://doi.org/10.1002/rmv.2270>.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, AI Access Foundation, v. 16, p. 321–357, jun 2002. Disponível em: <https://doi.org/10.1613\%2Fjair.953>.

EL-RASHIDY, N.; ABDELRAZIK, S.; ABUHMED, T.; AMER, E.; ALI, F.; HU, J.-W.; EL-SAPPAGH, S. Comprehensive survey of using machine learning in the covid-19 pandemic. *Diagnostics*, v. 11, n. 7, 2021. ISSN 2075-4418. Disponível em: <https://www.mdpi.com/2075-4418/11/7/1155>.

FEIKIN, D. R.; HIGDON, M. M.; ABU-RADDAD, L. J.; ANDREWS, N.; ARAOS, R.; GOLDBERG, Y.; GROOME, M. J.; HUPPERT, A.; O'BRIEN, K. L.; SMITH, P. G.; WILDER-SMITH, A.; ZEGER, S.; KNOLL, M. D.; PATEL, M. K. Duration of effectiveness of vaccines against SARS-CoV-2 infection and COVID-19 disease: results of a systematic review and meta-regression. *The Lancet*, Elsevier BV, v. 399, n. 10328, p. 924–944, mar. 2022. Disponível em: <https://doi.org/10.1016/s0140-6736(22)00152-0>.

GARCIA-BELTRAN, W. F.; LAM, E. C.; DENIS, K. S.; NITIDO, A. D.; GARCIA, Z. H.; HAUSER, B. M.; FELDMAN, J.; PAVLOVIC, M. N.; GREGORY, D. J.; POZNANSKY,

M. C.; SIGAL, A.; SCHMIDT, A. G.; IAFRATE, A. J.; NARANBHAI, V.; BALAZS, A. B. Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell*, Elsevier BV, v. 184, n. 9, p. 2372–2383.e9, abr. 2021. Disponível em: <https://doi.org/10.1016/j.cell.2021.03.013>.

JESUS, M. A. S. de; HOJO-SOUZA, N. S.; MORAES, T. R. d.; GUIDONI, D. L.; SOUZA, F. S. H. de. Profile of brazilian inpatients with covid-19 vaccine breakthrough infection and risk factors for unfavorable outcome. *Rev Panam Salud Publica;46, ago. 2022*, 2022. ISSN 1680 5348.

JIANG, X. gao; COFFEE, M.; BARI, A.; WANG, J.; JIANG, X.; HUANG, J.; SHI, J.; DAI, J.; CAI, J.; ZHANG, T.; WU, Z. xing; HE, G.; HUANG, Y. Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *Cmc-computers Materials & Continua*, v. 63, p. 537–551, 2020.

JIN, Y.; YANG, H.; JI, W.; WU, W.; CHEN, S.; ZHANG, W.; DUAN, G. Virology, epidemiology, pathogenesis, and control of COVID-19. *Viruses*, MDPI AG, v. 12, n. 4, p. 372, mar. 2020. Disponível em: <https://doi.org/10.3390/v12040372>.

LALMUANAWMA, S.; HUSSAIN, J.; CHHAKCHHUAK, L. Applications of machine learning and artificial intelligence for covid-19 (sars-cov-2) pandemic: A review. *Chaos, Solitons & Fractals*, v. 139, p. 110059, 2020. ISSN 0960-0779. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0960077920304562>.

LEMAÎTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, v. 18, n. 17, p. 1–5, 2017. Disponível em: <http://jmlr.org/papers/v18/16-365.html>.

LIASHCHYNSKYI, P.; LIASHCHYNSKYI, P. Grid search, random search, genetic algorithm: A big comparison for nas. arXiv, 2019. Disponível em: <https://arxiv.org/abs/1912.06059>.

LIKEBUPT; PETERCLU; CHMCCL v. Smote. 2022. Disponível em: <https://learn.microsoft.com/azure/machine-learning/component-reference/smote>.

LODATO, I.; IYER, A. V.; TO, I. Z.; LAI, Z.-Y.; CHAN, H. S.-Y.; LEUNG, W. S.-W.; TANG, T. H.-C.; CHEUNG, V. K.-L.; WU, T.-C.; NG, G. W.-Y. Prognostic model of COVID-19 severity and survival among hospitalized patients using machine learning techniques. *Diagnostics*, MDPI AG, v. 12, n. 11, p. 2728, nov. 2022. Disponível em: <https://doi.org/10.3390/diagnostics12112728>.

LOUPPE, G. Understanding random forests: From theory to practice. *arXiv*, Cornell University, 2015. Disponível em: <https://doi.org/10.48550/arXiv.1407.7502>.

MARTIN, D. *Stacked turtles*. 2019. Disponível em: <https://kiwidamien.github.io/how-to-do-cross-validation-when-upsampling-data.html>.

MITCHELL, T. M. *Machine learning*. [S.l.]: McGraw-hill New York, 1997. v. 1.

MK-GURUCHARAN. *Gini Index for Decision Trees: Mechanism, Perfect and Imperfect Split With Examples*. 2022. Disponível em: <https://www.upgrad.com/blog/gini-index-for-decision-trees/>.

NAHM, F. S. What the P values really tell us. *Korean J. Pain*, Korean Pain Society, v. 30, n. 4, p. 241–242, oct 2017.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Disponível em: <https://scikit-learn.org/stable/index.html>.

QUINLAN, J. R. Induction of decision trees. *Machine Learning*, v. 1, 1986. ISSN 1573-0565. Disponível em: <https://doi.org/10.1007/BF00116251>.

RAHIMI, F.; ABADI, A. T. B. Omicron: A highly transmissible SARS-CoV-2 variant. *Gene Reports*, Elsevier BV, v. 27, p. 101549, jun. 2022. Disponível em: <https://doi.org/10.1016/j.genrep.2022.101549>.

SABINO, E. C.; BUSS, L. F.; CARVALHO, M. P. S.; PRETE JR, C. A.; CRISPIM, M. A. E.; FRAIJI, N. A.; PEREIRA, R. H. M.; PARAG, K. V.; PEIXOTO, P. da S.; KRAEMER, M. U. G.; OIKAWA, M. K.; SALOMON, T.; CUCUNUBA, Z. M.; CASTRO, M. C.; SANTOS, A. A. de S.; NASCIMENTO, V. H.; PEREIRA, H. S.; FERGUSON, N. M.; PYBUS, O. G.; KUCHARSKI, A.; BUSCH, M. P.; DYE, C.; FARIA, N. R. Resurgence of COVID-19 in manaus, brazil, despite high seroprevalence. *Lancet*, England, v. 397, n. 10273, p. 452–455, jan. 2021.

SAH, P.; FITZPATRICK, M. C.; ZIMMER, C. F.; ABDOLLAHI, E.; JUDEN-KELLY, L.; MOGHADAS, S. M.; SINGER, B. H.; GALVANI, A. P. Asymptomatic SARS-CoV-2 infection: A systematic review and meta-analysis. *Proceedings of the National Academy of Sciences*, Proceedings of the National Academy of Sciences, v. 118, n. 34, ago. 2021. Disponível em: <https://doi.org/10.1073/pnas.2109229118>.

SCHULLER, B. W.; PHAM, P.; DO, S.; PATTICHIS, C. S.; NAIR, P. (Ed.). *Health Technologies and Innovations to Effectively Respond to the COVID-19 Pandemic*. Lausanne: Frontiers Media SA, 2022.

SHANBEHZADEH, M.; OROOJI, A.; KAZEMI-ARPANAHI, H. Comparing of data mining techniques for predicting in-hospital mortality among patients with covid-19. *Journal of Biostatistics and Epidemiology*, v. 7, n. 2, p. 154–173, Jul. 2021. Disponível em: <https://jbe.tums.ac.ir/index.php/jbe/article/view/504>.

SOUZA, F. S. H. D.; HOJO-SOUZA, N. S.; SANTOS, E. B. D.; SILVA, C. M. D.; GUIDONI, D. L. Predicting the disease outcome in covid-19 positive patients through machine learning: A retrospective cohort study with brazilian data. *Frontiers in Artificial Intelligence*, v. 4, 2021. ISSN 2624-8212. Disponível em: <https://www.frontiersin.org/articles/10.3389/frai.2021.579931>.

SOUZA, F. S. H. de; HOJO-SOUZA, N. S.; BATISTA, B. D. d. O.; SILVA, C. M. da; GUIDONI, D. L. On the analysis of mortality risk factors for hospitalized covid-19 patients: A data-driven study using the major brazilian database. *PLOS ONE*, Public Library of Science, v. 16, n. 3, p. 1–21, 03 2021. Disponível em: <https://doi.org/10.1371/journal.pone.0248580>.

SOUZA, F. S. H. de; HOJO-SOUZA, N. S.; SILVA, C. M. da; GUIDONI, D. L. Second wave of COVID-19 in brazil: younger at higher risk. *European Journal of Epidemiology*, Springer Science and Business Media LLC, v. 36, n. 4, p. 441–443, abr. 2021. Disponível em: <https://doi.org/10.1007/s10654-021-00750-8>.

STOLTZFUS, J. C. Logistic regression: A brief primer. *Academic Emergency Medicine*, v. 18, n. 10, p. 1099–1104, 2011. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1553-2712.2011.01185.x>.

ST»HLE, L.; WOLD, S. Analysis of variance (anova). *Chemometrics and Intelligent Laboratory Systems*, v. 6, n. 4, p. 259–272, 1989. ISSN 0169-7439. Disponível em: <https://www.sciencedirect.com/science/article/pii/0169743989800954>.

SUBUDHI, S.; VERMA, A.; AL, P. A. B. et. Comparing machine learning algorithms for predicting icu admission and mortality in covid-19. *npj Digital Medicine*, v. 4, May 2021. ISSN 2398-6352. Disponível em: <https://doi.org/10.1038/s41746-021-00456-x>.

TAYARANI-N., M.-H. Applications of artificial intelligence in battling against covid-19: A literature review. *Chaos, Solitons & Fractals*, v. 142, p. 110338, 2021. ISSN 0960-0779. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0960077920307335>.

WHO. *World Health Organization. WHO announces COVID-19 outbreak a pandemic*. www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic: [s.n.], 2020. [Acessado em 22 de maio de 2020].

WHO. *World Health Organization. WHO Coronavirus (COVID-19) Dashboard.* https://covid19.who.int/region/amro/country/br: [s.n.], 2023. [Acessado em 13 de março de 2023].

WU, Z.; MCGOOGAN, J. M. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in china. *JAMA*, American Medical Association (AMA), v. 323, n. 13, p. 1239, abr. 2020. Disponível em: <https://doi.org/10.1001/jama.2020.2648>.

ZHOU, F.; YU, T.; DU, R.; FAN, G.; LIU, Y.; LIU, Z.; XIANG, J.; WANG, Y.; SONG, B.; GU, X.; GUAN, L.; WEI, Y.; LI, H.; WU, X.; XU, J.; TU, S.; ZHANG, Y.; CHEN, H.; CAO, B. Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study. *The Lancet*, v. 395, n. 10229, p. 1054–1062, 2020. ISSN 0140-6736. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0140673620305663>.