INSTITUTO FEDERAL
Pernambuco

UTILIZANDO SHAPLEY ADDITIVE EXPLANATIONS PARA ANALISAR CARACTERÍSTICAS EPIDEMIOLÓGICAS DA COVID-19

USING SHAPLEY ADDITIVE EXPLANATIONS TO ANALYZE EPIDEMIOLOGICAL COVID-19 FEATURES

Paulo André Oliveira da Silva
paos@discente.ifpe.edu.br
Ana Clara Silva Fontes
acsf7@discente.ifpe.edu.br
Flávio Rosendo da Silva Oliveira
flavio.oliveira@paulista.ifpe.edu.br

RESUMO

A pandemia de COVID-19 resultou em milhões de mortes globalmente, tornando a compreensão das características epidemiológicas uma prioridade. Nesse cenário, instituições como a Fiocruz forneceram dados fundamentais para análise, e o uso de técnicas de Aprendizagem de Máquina e Inteligência Artificial Explicável surgiram como uma estratégia para interpretar padrões nesses dados. O presente artigo, por meio de uma abordagem exploratória quantitativa, apresenta uma análise comparativa das características epidemiológicas da COVID-19, no período de 2020 a 2021. Para tal análise, foi aplicada uma técnica de Inteligência Artificial Explicável, SHAP, a modelos de Aprendizagem de Máquina. Por fim, foi possível realizar a análise proposta e identificar tanto as similaridades quanto as dissimilaridades entre os períodos estudados.

Palavras-chave: Epidemiologia; COVID-19; Aprendizagem de máquina; Inteligência Artificial Explicável.

ABSTRACT

The COVID-19 pandemic resulted in millions of deaths globally, making the understanding of epidemiological characteristics a priority. In this context, institutions like Fiocruz provided essential data for analysis, and the use of Machine Learning and Explainable Artificial Intelligence techniques emerged as a strategy to interpret patterns in these data. This article, through a quantitative exploratory approach, presents a comparative analysis of the epidemiological characteristics of COVID-19, from 2020 to 2021. For this analysis, an eXplainable Artificial Intelligence technique, SHAP, was applied to Machine Learning models. Finally, it was possible to conduct the proposed analysis and identify both similarities and dissimilarities between the studied periods.

Keywords: Epidemiology; COVID-19; Machine Learning; eXplainable Artificial Intelligence.

1 INTRODUÇÃO

A pandemia de COVID-19, causada pelo SARS-CoV-2, que se espalhou rapidamente desde seu surgimento em dezembro de 2019, resultou em mais de 6 milhões de mortes em todo o mundo (Cascella *et al.*, 2023). A compreensão das características epidemiológicas da doença, como fatores que influenciam a taxa de transmissão, mortalidade e recuperação, tornou-se uma prioridade para governos, pesquisadores e profissionais de saúde (Freitas; Barcellos; Villela, 2023), podendo auxiliar no monitoramento e evolução da doença, assim como fornecer indicadores em saúde pública (Klokner *et al*., 2021).

Durante esse período, a Fundação Oswaldo Cruz (Fiocruz) desempenhou um papel fundamental ao fornecer dados epidemiológicos que serviram como base primária para estudos e intervenções destinadas a abordar várias questões de saúde pública (Freitas; Barcellos; Villela, 2023). Esses dados, que incluem informações sobre os pacientes e fatores de risco, representam uma fonte para análises. No entanto, para extrair *insights* significativos que permitam a tomada de decisões podem ser necessárias ferramentas computacionais avançadas (Faceli *et al*., 2021).

Nesse contexto, a utilização de técnicas de Aprendizagem de Máquina (AM) e Inteligência Artificial Explicável, do Inglês *Explainable Artificial Intelligence* (XAI), emergiu como uma estratégia promissora para analisar e interpretar padrões complexos em dados de setores como a saúde (Ali *et al*., 2023). Entre as abordagens de XAI, o método *Shapley Additive Explanations* (SHAP) destaca-se por sua capacidade de fornecer explicações detalhadas sobre as contribuições individuais de cada característica para a tomada de decisões de modelos preditivos (Lundberg; Lee, 2017). Aplicar essa tecnologia ao estudo da COVID-19 pode possibilitar identificar e entender as características epidemiológicas mais relevantes que influenciaram a propagação e a gravidade da pandemia.

Diante desse cenário, o presente estudo, por meio de uma abordagem exploratória quantitativa, objetiva realizar uma análise comparativa das características epidemiológicas da COVID-19, no período de 2020 a 2021, utilizando técnicas de XAI, com foco na aplicação do método SHAP. Pretende-se coletar e realizar uma análise exploratória dos dados disponíveis, treinar modelos de classificação e, posteriormente, empregar uma análise SHAP para interpretar os resultados, proporcionando uma compreensão mais profunda dos fatores epidemiológicos que potencialmente contribuíram para o cenário da mortalidade no estado de Pernambuco, no Brasil.

Este artigo está organizado em seis seções. Na Seção 2, Fundamentação Teórica, serão abordados os principais temas que embasam esta pesquisa, como o contexto da pandemia de COVID-19, conceitos fundamentais de Aprendizagem de Máquina, com foco no uso de classificadores e métricas de avaliação, culminando na discussão sobre Inteligência Artificial Explicável. A Seção 3, Trabalhos Relacionados, explorará a aplicação da XAI. Na Seção 4, Metodologia, será descrita a abordagem adotada pelo estudo, detalhando as etapas e técnicas empregadas. A Seção 5, Experimentos e Resultados, apresentará os testes realizados e suas respectivas análises. Por fim, a Seção 6, Conclusão e Trabalhos Futuros, apresentará as considerações finais da pesquisa e delineará possíveis direções para futuras investigações.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 COVID-19

Em Dezembro de 2019, a COVID-19, uma doença causada pelo vírus SARS-CoV-2, teve seus primeiros casos reportados em Wuhan, uma cidade localizada na província chinesa de Hubei (Cascella *et al.*, 2023). Após o surto inicial da doença, em 30 de Janeiro de 2020, a Organização Mundial da Saúde (WHO, 2020) declarou uma emergência global.

Posterior a sua rápida disseminação ao redor do mundo, em 11 de Março de 2020, a OMS declara a COVID-19 uma pandemia (Sharma; Ahmad Farouk; Lal, 2021; WHO, 2020). Enquanto isso, no Brasil, o primeiro caso da COVID-19 foi reportado em 25 de Fevereiro de 2020. O vírus se propagou rapidamente, chegando a mais de 5,4 milhões de casos confirmados em Outubro do mesmo ano (Marinho *et al*., 2021).

Alguns fatores de risco estavam entre os preditores mais prevalentes entre os pacientes hospitalizados com COVID-19, são eles: idade avançada (Guan *et al*., 2020) e comorbidades como hipertensão, doenças cardiovasculares, diabetes, tabagismo, doença pulmonar obstrutiva crônica (DPOC), câncer e insuficiência renal (Emami *et al*., 2020; Klokner *et al*., 2021), entre outros. Alguns dos citados, tidos também como variáveis nas bases de dados selecionadas para o desenvolvimento do atual estudo.

Esses fatores estão ligados diretamente à Epidemiologia da doença, termo que é conceituado como:

> Ciência que estuda o processo saúde-doença em coletividades humanas, analisando a distribuição e os fatores determinantes do risco de doenças, agravos e eventos associados à saúde [...], produzindo informação e conhecimento para apoiar a tomada de decisão no planejamento, administração e avaliação de sistemas, programas, serviços e ações de saúde (Rouquayrol; Gurgel, 2018, p. 70-71).

Tal conceito, reforça o fato da análise das características epidemiológicas ser imprescindível no processo decisório sobre fatores relacionados a doenças (Rouquayrol; Gurgel, 2018) como a COVID-19.

2.2 Aprendizagem de Máquina

A Aprendizagem de Máquina, ou *Machine Learning* em Inglês, pode ser compreendida como uma das áreas da Inteligência Artificial que consiste essencialmente em treinar algoritmos com dados, permitindo que eles façam previsões ou tomem decisões sem a necessidade de realizar uma programação explícita (An *et al*., 2023).

Em adição, é possível afirmar que essa subárea surgiu motivada pela necessidade de lidar com o rápido e crescente volume de dados criados nos mais diversos setores da sociedade (Faceli *et al*., 2021). Ela permite a melhoria do desempenho de um sistema por meio de métodos computacionais que envolvem a experiência, existente na forma de dados, visando gerar modelos que podem atuar bem em novos dados, não apenas naqueles utilizados durante o seu processo de treinamento (Zhou, 2021).

Nesse sentido, os algoritmos de Aprendizagem de Máquina são constantemente utilizados em tarefas distintas (Faceli *et al*., 2021), desempenhando atividades em

variadas áreas e setores. Isso também inclui atividades voltadas ao setor de saúde (Ulrichsen *et al*., 2022), como análises preditivas, diagnóstico de doenças e suporte a decisões clínicas (Kumar *et al*., 2023; An *et al*., 2023).

Além disso, o setor de saúde tem cada vez mais se mostrado como um dos mais interessados em fazer uso dos benefícios de novas tecnologias, como a Inteligência Artificial e Aprendizagem de Máquina (Albahri *et al*., 2023; Kumar *et al*., 2023). Especialmente com a tendência crescente do volume de dados disponíveis no setor, é notável o potencial de modelos de Inteligência Artificial e Aprendizagem de Máquina em uma variedade de aplicações na área (Jia, 2023).

2.3 Classificadores

Os algoritmos de AM podem ser categorizados de diversas formas, de modo geral podem ser categorizados em grandes grupos como aprendizado supervisionado, aprendizado não supervisionado, aprendizado semi-supervisionado e outros (Woodman; Mangoni, 2023), cada um com suas próprias características e especificidades.

No aprendizado supervisionado, o algoritmo é treinado a partir de um conjunto de dados rotulados, onde cada entrada é acompanhada por uma saída esperada (Woodman; Mangoni, 2023). Nesse caso, quando o processo de treinamento de um algoritmo é executado de forma bem-sucedida, o algoritmo deve ser capaz de realizar predições adequadas a novas amostras de dados (Sidey-Gibbons, 2019). Vale ressaltar que as aplicações mais comuns dos algoritmos de aprendizado supervisionado englobam tarefas como classificação e regressão (Sarker, 2021).

Em geral, esses algoritmos consistem em desenvolver uma função capaz de estimar um valor resultante com base em dados de entrada (Sarker, 2021). É possível afirmar que a principal diferença entre as tarefas de classificação e regressão é o resultado obtido a partir das funções desenvolvidas, sendo os resultados obtidos a partir das funções de regressão valores contínuos e nas funções de classificação valores discretos, por vezes referidos como classes, como no caso da base de dados utilizada neste estudo. Os modelos que são capazes de produzir esses valores discretos ou classes, são aqueles chamados de classificadores (Sidey-Gibbons, 2019).

Embora alguns algoritmos de AM possuam construções que permitam sua utilização em atividades de regressão e classificação, a exemplo do *Support Vector Machine* (Cervantes *et al*., 2019) e da Árvore de Decisão (Kushwah, 2021), no presente artigo esses algoritmos devem ser considerados unicamente como classificadores.

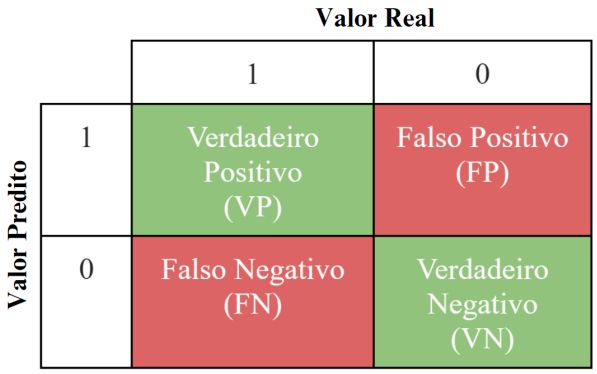
2.4 Métricas

A avaliação do desempenho preditivo dos classificadores é uma questão relevante para avaliar os resultados do processo de classificação (Vanacore; Pellegrino; Ciardelli, 2022). Nesse contexto, faz-se necessário o uso de métricas de avaliação de desempenho.

As métricas mais amplamente empregadas são aquelas baseadas na Matriz de Confusão (Vanacore; Pellegrino; Ciardelli, 2022), uma combinação dos resultados das classificações com quatro atributos principais que apresentam os dados do resultado (Uddin *et al.*, 2022). Numa matriz, levando em consideração uma tarefa de classificação binária, as instâncias de dados são normalmente previstas como positivas ou negativas (Rainio; Teuho; Klén, 2024). Por exemplo, caso o resultado da

classificação seja previsto como 1 e o valor verdadeiro for igual a 1, o resultado é considerado verdadeiro positivo (VP). O mesmo ocorre para o valor 0, sendo considerado verdadeiro negativo (VN). Ademais, se o resultado da previsão for 1 e o valor verdadeiro for igual a 0, o resultado é classificado como falso positivo (FP), e o inverso é chamado de falso negativo (FN) (Uddin *et al.*, 2022), como apresentado na Figura 1.

Figura 1 – Exemplo de Matriz de Confusão



Fonte: Adaptado de Uddin *et al.* (2022), pelos autores

Em relação às métricas, a Acurácia, *Recall*, Precisão e F1 *Score*, com fórmulas listadas na Tabela 1, estão entre as mais comumente utilizadas (Vanacore; Pellegrino; Ciardelli, 2022). A Acurácia avalia o desempenho geral do classificador, considerando toda a matriz de confusão, enquanto Recall, Precisão e F1 *Score* estão concentradas no desempenho de uma classe (Vanacore; Pellegrino; Ciardelli, 2022). *Recall* e Precisão são frequentemente analisadas em conjunto, sendo o F1 *Score* sua média harmônica (Rainio; Teuho; Klén, 2024).

Tabela 1 – Descrição das fórmulas das métricas

Métrica	Fórmula
Acurácia	$\frac{VP + VN}{VP + VN + FP + FN}$
Precisão	$\frac{VP}{VP + FP}$
Recall	$\frac{VP}{VP + FN}$
F1 *Score*	$\frac{2 * Precisão * Recall}{Precisão + Recall}$

Fonte: Elaborada pelos autores (2024)

2.5 Inteligência Artificial Explicável

No contexto da saúde, os algoritmos de IA podem ajudar médicos na leitura e análise de relatórios escritos por patologistas, radiologistas e muito mais (Hulsen, 2023). Porém, existe um risco de uso dos mesmos nessa área (Hulsen *et al*., 2022) devido ao problema da “caixa preta”, que refere-se à dificuldade de um sistema em oferecer uma explicação razoável de como chegou a uma decisão (Ali *et al*., 2023). Tal dificuldade torna a capacidade de explicação dos modelos limitada e acaba reduzindo a confiança de médicos e pacientes neles (Hulsen *et al*., 2022).

Um dos caminhos para a transparência da IA é a XAI, que tem como principal objetivo obter modelos interpretáveis por humanos, especialmente para aplicações em setores sensíveis, como o de saúde, já que os especialistas no domínio precisam de ajuda para resolver problemas com mais eficiência. Sua explicabilidade expressa o que está ocorrendo no modelo, fornecendo uma explicação legível sobre a decisão do mesmo (Ali *et al*., 2023).

Nesse sentido, a *Shapley Additive Explanations*, considerada uma técnica de IA Explicável (Ali *et al*., 2023), foi proposta em função de ajudar usuários na interpretação dos resultados dos modelos (Lundberg; Lee, 2017). Com a SHAP é possível analisar visualmente os atributos de mais influência nas predições (Liu *et al*., 2024), atribuindo a cada característica um valor de importância (Lundberg; Lee, 2017).

3 TRABALHOS RELACIONADOS

A Inteligência Artificial Explicável tem sido aplicada em diversos campos para fornecer interpretações claras e compreensíveis dos resultados gerados por modelos de Inteligência Artificial. Nesta seção, são explorados trabalhos anteriores que abordaram temas semelhantes ao deste estudo, sendo possível observar as contribuições do uso da AM e técnicas de XAI em diversos cenários, incluindo estudos relacionados a COVID-19.

Em relação à previsão de doenças, pode-se destacar o trabalho de Liu *et al*. (2024), que propõe modelos de previsão de risco para Doenças Vasculares Periféricas (DVP) em pacientes com Diabetes Mellitus 2. Além da previsão, o estudo realizou uma análise dos fatores de risco para DVP utilizando o método SHAP, também utilizado no presente artigo.

Similarmente, Romano *et al*. (2024) investigaram a influência da poluição do ar e de fatores socioeconômicos na mortalidade por câncer respiratório na Itália. O estudo analisa a associação entre esses fatores e a taxa de mortalidade padronizada (SMR) por câncer respiratório. Treze algoritmos de AM diferentes são avaliados, e o estudo também aplica SHAP para interpretar os resultados e auxiliar na identificação dos fatores mais importantes na previsão da SMR.

No contexto da COVID-19, no estudo realizado por Rodrigues; Kreutz (2022), o foco está na utilização de um conjunto de dados aberto do estado do Rio Grande do Sul, Brasil, para o desenvolvimento de um modelo preditor para classificar o risco de óbito por COVID-19. Embora também ocorra o desenvolvimento de preditores com essa finalidade, o estudo não utiliza XAI para tornar os resultados mais interpretáveis.

Debjit *et al*. (2022) também abordam a previsão de COVID-19, mas com um conjunto de dados diferente. O estudo utiliza um conjunto de dados disponível publicamente com informações sobre pacientes com COVID-19 e aplica técnicas de otimização de hiperparâmetros para aprimorar a precisão dos modelos de aprendizado de máquina. O artigo compara o desempenho de diversos algoritmos e também utiliza SHAP para analisar a importância das variáveis na previsão. Contudo, o foco do estudo é a previsão da positividade para COVID-19, e não a previsão da evolução do paciente, como no presente estudo.

Em outro estudo relacionado ao período da pandemia, Lee (2022) investiga as mudanças no comportamento de uso do transporte público durante a pandemia de COVID-19. O autor utiliza dados de cartões inteligentes de Seul, Coreia, e aplica o modelo XGBoost para analisar as escolhas de uso do transporte público. A técnica

SHAP também é aplicada para interpretar os resultados e entender os fatores que influenciaram as mudanças de comportamento.

Por fim, Fonseca *et al*. (2023) apresentam um modelo de aprendizado profundo para classificar mosquitos transmissores de doenças com base em sinais de áudio. Os autores utilizam uma rede neural convolucional e espectrogramas de áudio para treinar o modelo. O estudo também utiliza SHAP para interpretar as decisões do modelo, além de entender quais características do áudio são mais relevantes para a classificação.

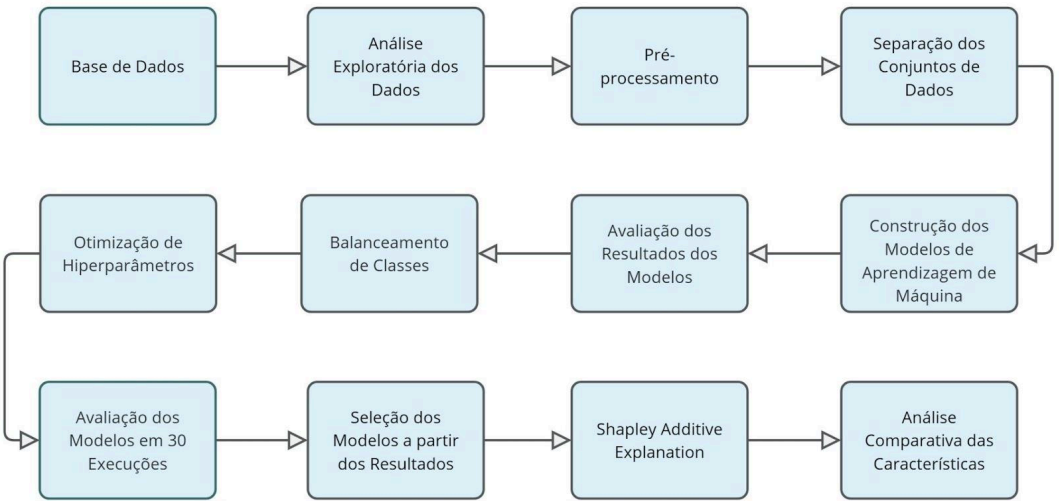
Em suma, os trabalhos explorados apresentam similaridades com este estudo, como a aplicação de AM na área da saúde e o uso de técnicas de XAI. Contudo, cada estudo possui suas particularidades, como o conjunto de dados utilizado e os métodos de AM empregados.

4 METODOLOGIA

O presente estudo adota uma abordagem de pesquisa exploratória quantitativa. O método principal utilizado é a análise dos resultados dos modelos classificadores por meio da técnica SHAP para extrair conhecimento sobre as relações entre características epidemiológicas da COVID-19.

A pesquisa exploratória pode ser considerada uma especialização da exploração de informações, uma classe mais ampla de atividades em que novas informações são buscadas em uma área conceitual definida (White *et al*., 2008). O desenvolvimento e a execução da investigação foram conduzidos conforme o fluxo metodológico ilustrado na Figura 2. As subseções seguintes descrevem em mais detalhes os elementos deste fluxograma.

Figura 2 – Fluxograma metodológico realizado neste estudo



Fonte: Elaborada pelos autores (2024)

4.1 Base de dados

Os dados presentes nestas bases de dados, disponibilizadas pela Fundação Oswaldo Cruz, são resultantes de coletas de dados da evolução de pacientes internados com COVID-19 realizadas pela instituição em três períodos distintos durante a pandemia: março a outubro de 2020, outubro de 2020 a janeiro de 2021 e fevereiro a outubro de 2021.

O conjunto das três bases de dados possui um total de 107.716 amostras. Todas as variáveis são categóricas e representantes de aspectos dos pacientes, conforme a Tabela 2. As características incluem mas não se limitam a raça e sexo, presença de algumas das comorbidades consideradas fatores de risco para a COVID-19 (Cascella *et al*., 2023), segmentação geográfica e faixa etária.

Tabela 2 – Descrição da base de dados

Nº	Nome das variáveis	Descrição	Valores
1	DATA ATUALIZAÇÃO	Data de atualização da evolução do paciente	Variável temporal
2	DATA DA NOTIFICAÇÃO	Data de notificação da doença	Variável temporal
3	SEXO	Sexo da pessoa	0 - MASCULINO; 1- FEMININO; 2- IGNORADO
4	PRO_SAÚDE	Profissional de Saúde	(-1) NaN; 0- Sim; 1- Não
5	GESTANTE	Se a pessoa é gestante	(-1) NaN; 0- Não; 1- Sim
6	VACINADO	Se a pessoa foi vacinada	(-1) NaN; 0- Não; 1- Sim
7	RACA	Raça da pessoa	(-1) NaN; 0- Parda R1; 1- Branca; 2- Ignorado; 3- Preta R2; 4- Amarela R3; 5- Indígena R4
8	EVOLUCAO	Se o paciente conseguiu se recuperar	(-1) NaN; 0- Óbito; 1- Não óbito
9	INTERNADO	Se o paciente foi internado	(-1) NaN; 0- Sim; 1- Não
10	REGIAO	Se o paciente mora no interior ou na RMR	1- Interior; 2- RMR
11	FAIXA_ETARIA	Faixa de idade do paciente	1- (0,9] F1; 2- (9,17] F2; 3- (17,39]; 4- (39,65] F3; 5- (65,75] F4; 6- (75,INF) F5
12	FEBRE	Se o paciente teve febre	1- NÃO; 2- SIM
13	TOSSE	Se o paciente teve tosse	1- NÃO; 2- SIM
14	DISPNEIA	Se o paciente teve dispneia	1- NÃO; 2- SIM
15	DESCONFORTO_RESPIRATÓRIO	Se o paciente teve desconforto respiratório	1- NÃO; 2- SIM
16	APERTO_TORACICO	Se o paciente teve aperto torácico	1- NÃO; 2- SIM
17	SATURACAO_O2	Se o paciente teve Saturação O2	1- NÃO; 2- SIM
18	VASCULARES	Se o paciente tem Doenças cardíacas	1- NÃO; 2- SIM
19	DIABETES	Se o Paciente tem Diabetes	1- NÃO; 2- SIM
20	D_RESPIRACAO_CRONICA	Se o paciente tem Doenças de RESPIRAÇÃO CRÔNICA	1- NÃO; 2- SIM
21	OBESIDADE	Se o paciente é obeso	1- NÃO; 2- SIM
22	HAS	Se o paciente tem HAS	1- NÃO; 2- SIM
23	TABAGISTA	Se o paciente é Tabagista	1- NÃO; 2- SIM

Fonte: Elaborada pelos autores (2024)

Vale pontuar que a divisão da base de dados em três períodos foi realizada para considerar diferentes variantes do SARS-CoV-2 (Cascella *et al.*, 2023) ao longo do tempo. Dessa forma, pode ser possível avaliar se determinados padrões observados nos dados estão associados a características específicas das variantes em circulação.

4.2 Processamento de dados

Na etapa de processamento, primeiramente, foi realizada uma análise exploratória dos dados para identificar valores ausentes, duplicados ou desbalanceamento entre as classes. Para tal análise, foi utilizada a ferramenta Pandas Profiling, que detalha as informações sobre uma base de dados analisada. A análise revelou uma quantidade considerável de valores faltantes e linhas duplicadas. Com base nisso, as linhas duplicadas foram removidas.

Além disso, ocorreu uma análise de componentes principais (PCA, na sigla do inglês *Principal Component Analysis*) com o intuito de identificar padrões nos dados. No entanto, a PCA não se mostrou relevante para a separação dos grupos, uma vez que não foi observada uma diferenciação satisfatória entre os componentes principais gerados.

Ademais, conforme demonstrado anteriormente na Tabela 2, as bases de dados inicialmente continham variáveis temporais, como a data de atualização e a data de notificação, que foram consideradas irrelevantes e, portanto, foram removidas. Além disso, as variáveis VACINADO, GESTANTE, RACA e PRO_SAUDE foram excluídas devido à presença de diversos dados faltantes. No total, as bases de dados continham 23 características, e a remoção dessas variáveis permitiu concentrar a análise nas 17 características restantes, conforme apresentado na Tabela 3.

Tabela 3 – Descrição da base de dados após remoção de variáveis

Nº	Nome das variáveis
1	SEXO
2	EVOLUCAO
3	INTERNADO
4	REGIAO
5	FAIXA_ETARIA
6	FEBRE
7	TOSSE
8	DISPNEIA
9	DESCONFORTO_RESPIRATÓRIO
10	APERTO_TORACICO
11	SATURACAO_O2
12	VASCULARES
13	DIABETES
14	D_RESPIRACAO_CRONICA
15	OBESIDADE
16	HAS
17	TABAGISTA

Fonte: Elaborada pelos autores (2024)

Posterior à seleção das variáveis fundamentais para o estudo, foi preciso fazer a separação dos conjuntos para treinamento e teste. Para esse processo, foi importada a função *train_test_split* da biblioteca Scikit-Learn. Essa biblioteca fornece metodologias diretas e eficazes para tarefas incluindo, mas não se limitando a, classificação, regressão, agrupamento, redução de dimensionalidade, seleção de modelos e pré-processamento (Igual, 2017). Foi definido o tamanho do conjunto de testes como 20% e passado o parâmetro *random_state* com o valor 0 para garantir a reprodutibilidade e consistência dos resultados.

Após os primeiros resultados da classificação, foi necessária uma nova etapa de processamento de dados, onde empregou-se a técnica de *undersampling* aleatório (*Random Undersampling*) com o *random_state* definido como 0 através da biblioteca Scikit-Learn. O *undersampling* consiste em remover aleatoriamente exemplos da classe majoritária para equilibrar a distribuição entre as classes. Desse modo, permite-se o aprimoramento da capacidade de generalização do modelo, priorizando a utilização de dados reais e evitando a geração de dados sintéticos pela natureza do problema, que envolve questões de saúde.

No processo de otimização de hiperparâmetros, para avaliar o desempenho do modelo em dados não vistos durante o treinamento, foi implementado um conjunto de validação. Utilizando o *train_test_split* com o *random_state* = 0, ocorreu a subdivisão do conjunto de treinamento em treinamento e validação, que recebeu 20% do conjunto de treinamento, garantindo uma avaliação do modelo antes da aplicação no conjunto de testes.

4.3 Classificação

Em relação ao processo de classificação binária realizado, foram utilizadas bibliotecas da linguagem de programação Python essenciais ao lidar com dados, com destaque para a biblioteca Scikit-Learn. Nesse sentido, após a execução da análise e do processamento dos dados, foram definidos os algoritmos classificadores a serem utilizados. Ao final deste processo de definição foram apontados 7 classificadores, sendo eles:

a) Support Vector Machine (SVM): Representa uma ferramenta potente na classificação de dados e tem se mostrado eficaz para enfrentar vários desafios do mundo real (Cervantes *et al*., 2019).

b) Regressão Logística: Apesar do seu nome incluir o termo “regressão”, o algoritmo de regressão logística funciona como um classificador (Zhou, 2021).

c) K-nearest neighbors (KNN): O algoritmo KNN é um dos mais frequentemente utilizados, especialmente pela sua boa acurácia e simplicidade (Faceli *et al*., 2021).

d) Rede Neural Artificial (RNA): As Redes Neurais Artificiais são algoritmos que simulam o processo de aprendizagem de seres vivos (Pauli; Kleina; Bonat, 2021).

e) XGBoost: Esse algoritmo notoriamente conquistou amplo reconhecimento em competições, bem como em casos de uso reais (Chen; Guestrin, 2016).

f) Random Forest: O Random Forest é um algoritmo de aprendizado em conjunto (*ensemble*) de árvores de decisão (Schonlau; Zou, 2020).

g) Árvore De Decisão: As árvores são amplamente reconhecidas por permitirem uma melhor visão de seu desempenho, assim como pela sua precisão na classificação (Jijo; Abdulazeez, 2021).

Escolhidos os classificadores, deu-se início à construção dos modelos, definindo o parâmetro *random_state* = 0 para garantir a reprodutibilidade, sem modificar os

demais parâmetros. Para esse processo e para a avaliação dos modelos, foram utilizados métodos da biblioteca Scikit-Learn[1] na versão 1.3.2.

Com base nos resultados iniciais da classificação, foi realizada a otimização de hiperparâmetros dos modelos com o parâmetro *random_state* = 0, utilizando os dados após o balanceamento de classes e separação dos conjuntos de treinamento e validação. Para a otimização, foi utilizada a biblioteca Optuna[2], que permite construir dinamicamente os espaços de busca dos hiperparâmetros (Akiba *et al*., 2019). Em posse dos melhores hiperparâmetros, ocorreu a avaliação dos modelos sem *random_state* fixo em 30 execuções de treino e teste, sendo selecionados os modelos com os melhores resultados para cada período.

4.4 Shapley Additive Explanations

Diante dos classificadores selecionados na etapa anterior, foi aplicada a técnica SHAP, tendo em vista a sua capacidade de auxiliar na interpretação dos resultados dos modelos devido a interpretabilidade desses serem, por vezes, limitadas (Hulsen *et al*., 2022).

A partir da utilização da técnica, através da biblioteca Shap, também da linguagem de programação Python, foram gerados e apresentados os gráficos resultantes das análises aplicadas aos modelos selecionados anteriormente.

Por fim, tornou-se possível analisar comparativamente as diferentes características epidemiológicas da COVID-19, visualizando, de forma clara, quais características apresentaram maior influência positiva ou negativa no resultado final das classificações. Em seguida, foram realizadas análises para explorar e quantificar as diferenças nas contribuições das variáveis entre os períodos estudados, ou seja, entre as três bases de dados.

5 EXPERIMENTOS E RESULTADOS

Após a análise exploratória e o processamento dos dados, mediante os classificadores citados, foram realizados experimentos iniciais com o processo de treinamento e testagem dos modelos. Posteriormente, utilizando métricas comumente aplicadas na avaliação de modelos como Acurácia, F1 *Score*, assim como a Matriz de Confusão (Vanacore; Pellegrino; Ciardelli, 2022), foi realizado o processo de análise do desempenho dos classificadores para os resultados dos conjuntos de dados de treinamento e teste de cada período representado pelas bases de dados, tendo sido ao final considerados os resultados do conjunto de testes, conforme representado nas tabelas: Tabela 4, Tabela 5 e Tabela 6.

[1]Mais informações sobre o Scikit-Learn: https://scikit-learn.org/stable

[2]Mais informações sobre a Optuna: https://optuna.org/

Tabela 4 – Métricas dos classificadores em teste - Março a outubro de 2020

Classificador	Acurácia	F1 *Score* (Classe 0)	F1 *Score* (Classe 1)	Matriz de Confusão
Support Vector Machine (SVM)	0,76	0,38	0,85	[720 1970] [398 6964]
Regressão Logística	0,75	0,43	0,84	[949 1741] [792 6570]
K-nearest neighbors (KNN)	0,74	0,45	0,83	[1079 1611] [1045 6317]
Rede Neural Artificial (RNA)	0,77	0,49	0,85	[1155 1535] [825 6537]
XGBoost	0,77	0,48	0,85	[1081 1609] [752 6610]
Random Forest	0,75	0,47	0,84	[1082 1608] [857 6505]
Árvore De Decisão	0,75	0,48	0,83	[1185 1505] [1055 6307]

Fonte: Elaborada pelos autores (2024)

Tabela 5 – Métricas dos classificadores em teste - Outubro de 2020 a Janeiro de 2021

Classificador	Acurácia	F1 *Score* (Classe 0)	F1 *Score* (Classe 1)	Matriz de Confusão
Support Vector Machine (SVM)	0,73	0,42	0,82	[242 570] [104 1565]
Regressão Logística	0,72	0,48	0,81	[327 485] [214 1455]
K-nearest neighbors (KNN)	0,69	0,45	0,78	[311 501] [273 1396]
Rede Neural Artificial (RNA)	0,73	0,50	0,81	[340 472] [204 1465]
XGBoost	0,71	0,47	0,80	[321 491] [222 1447]
Random Forest	0,70	0,48	0,79	[342 470] [275 1394]
Árvore De Decisão	0,67	0,48	0,76	[377 435] [373 1296]

Fonte: Elaborada pelos autores (2024)

Tabela 6 – Métricas dos classificadores em teste - Fevereiro a outubro de 2021

Classificador	Acurácia	F1 *Score* (Classe 0)	F1 *Score* (Classe 1)	Matriz de Confusão
Support Vector Machine (SVM)	0,74	0,41	0,83	[741 1762] [330 5225]
Regressão Logística	0,74	0,46	0,83	[902 1601] [514 5041]
K-nearest neighbors (KNN)	0,70	0,44	0,80	[946 1557] [843 4712]
Rede Neural Artificial (RNA)	0,74	0,49	0,83	[1020 1483] [609 4946]
XGBoost	0,74	0,47	0,83	[932 1571] [522 5033]
Random Forest	0,73	0,47	0,82	[959 1544] [655 4900]
Árvore De Decisão	0,72	0,48	0,81	[1031 1472] [793 4762]

Fonte: Elaborada pelos autores (2024)

Embora não tenham sido identificados indícios de *overfitting*, ou seja, de um desempenho significativamente superior no conjunto de treinamento em relação ao de teste, os resultados obtidos indicaram um enviesamento em relação à classe 1, o que resultou em prejuízo à representação da classe 0, representante de Óbito. Diante desse cenário, tornou-se necessário realizar experimentos adicionais. Após uma análise, a abordagem definida com base nos resultados obtidos foi *undersampling,* uma estratégia para balanceamento de classes.

Ademais, buscando melhorar os resultados dos classificadores, foi implementada a otimização dos hiperparâmetros utilizando a biblioteca Optuna. Essa técnica permitiu explorar o espaço de hiperparâmetros, conforme o Apêndice A.

Com os hiperparâmetros escolhidos após a otimização, os modelos foram avaliados com inicialização aleatória, por meio de 30 execuções consecutivas de treinamento e teste, coletando as métricas de acurácia, matriz de confusão e F1 *Score* para a classe 0 e classe 1 em teste. Ao final, para cada modelo, foram geradas as médias e os desvios padrões das métricas de acurácia e F1 *Score* para ambas as classes. Tendo em vista que os desvios padrões apresentaram valores inferiores a 0,00, foram consideradas apenas as médias das métricas, conforme Tabela 7, Tabela 8 e Tabela 9.

Tabela 7 – Média das métricas dos classificadores em teste - Março a outubro de 2020

Classificador	Média da Acurácia	Média do F1 *Score* (classe 0)	Média do F1 *Score* (classe 1)
Support Vector Machine (SVM)	0,54	0,27	0,66
Regressão Logística	0,72	0,72	0,71
K-nearest neighbors (KNN)	0,72	0,74	0,70
Rede Neural Artificial (RNA)	0,73	0,75	0,71
XGBoost	0,73	0,74	0,71
Random Forest	0,73	0,75	0,71
Árvore De Decisão	0,72	0,74	0,70

Fonte: Elaborada pelos autores (2024)

Tabela 8 – Média das métricas dos classificadores em teste - Outubro de 2020 a Janeiro de 2021

Classificador	Média da Acurácia	Média do F1 *Score* (classe 0)	Média do F1 *Score* (classe 1)
Support Vector Machine (SVM)	0,70	0,72	0,67
Regressão Logística	0,68	0,69	0,67
K-nearest neighbors (KNN)	0,67	0,69	0,65
Rede Neural Artificial (RNA)	0,69	0,70	0,67
XGBoost	0,69	0,70	0,67
Random Forest	0,69	0,71	0,67
Árvore De Decisão	0,66	0,67	0,64

Fonte: Elaborada pelos autores (2024)

Tabela 9 – Média das métricas dos classificadores em teste - Fevereiro a outubro de 2021

Classificador	Média da Acurácia	Média do F1 *Score* (classe 0)	Média do F1 *Score* (classe 1)
Support Vector Machine (SVM)	0,68	0,67	0,70
Regressão Logística	0,69	0,68	0,69
K-nearest neighbors (KNN)	0,67	0,67	0,67
Rede Neural Artificial (RNA)	0,69	0,69	0,69
XGBoost	0,70	0,70	0,70
Random Forest	0,69	0,69	0,69
Árvore De Decisão	0,70	0,69	0,70

Fonte: Elaborada pelos autores (2024)

Em posse desses resultados, foi possível realizar a seleção dos modelos, sendo escolhidos apenas os dois modelos com os melhores resultados para cada base de dados levando em consideração especialmente as médias das métricas de acurácia e F1 *Score* para as classes 0 e 1, respectivamente, no conjunto de testes. Os modelos selecionados para a continuidade da análise, seguindo para a realização da SHAP, encontram-se indicados na Tabela 10.

Tabela 10 – Classificadores selecionados para a realização da SHAP

Período	Classificadores
Março a outubro de 2020	Random Forest; Rede Neural Artificial (RNA)
Outubro de 2020 a janeiro de 2021	Random Forest; Support Vector Machine (SVM)
Fevereiro a outubro de 2021	Árvore de Decisão; XGBoost

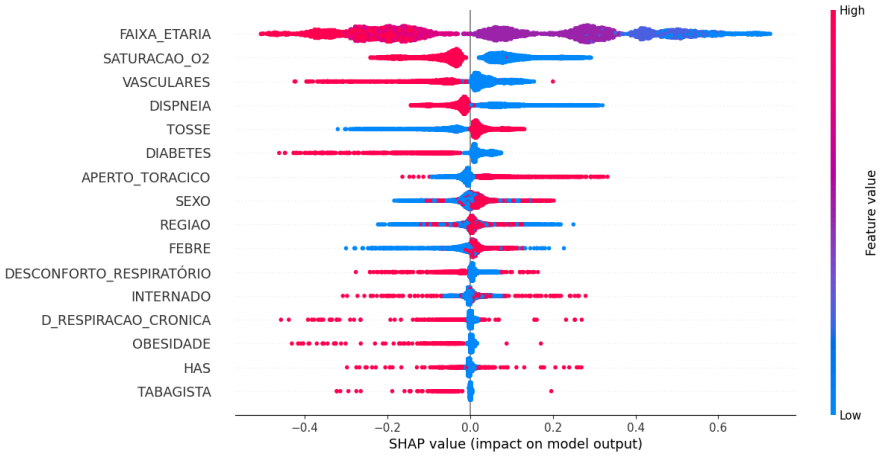
Fonte: Elaborada pelos autores (2024)

Vale ressaltar que no período de fevereiro a outubro de 2021 o XGBoost foi configurado com os hiperparâmetros *learning_rate*=0,2473364407802898, *max_depth*=3 e *subsample*=0,8959573157740724, enquanto a Árvore de Decisão foi configurada com os hiperparâmetros *max_depth*=8, *min_samples_split*=16, *min_samples_leaf*=4. Em outubro de 2020 a janeiro de 2021, o Random Forest foi iniciado com os hiperparâmetros *n_estimators*=114, *max_depth*=10, *min_samples_split*=11, enquanto o SVM foi utilizado com os hiperparâmetros *C*=9,979157445198476, *kernel*='rbf', *max_iter*=9300. Por fim, de março a outubro de 2020, a Rede Neural Artificial foi configurada com *hidden_layer_sizes*=(50, 50), *alpha*=5,850528889632702e-05, *learning_rate_init*=0,00019715551635249314, já a

Random Forest teve seus hiperparâmetros ajustados para *n_estimators*=215, *max_depth*=12, *min_samples_split*=19.

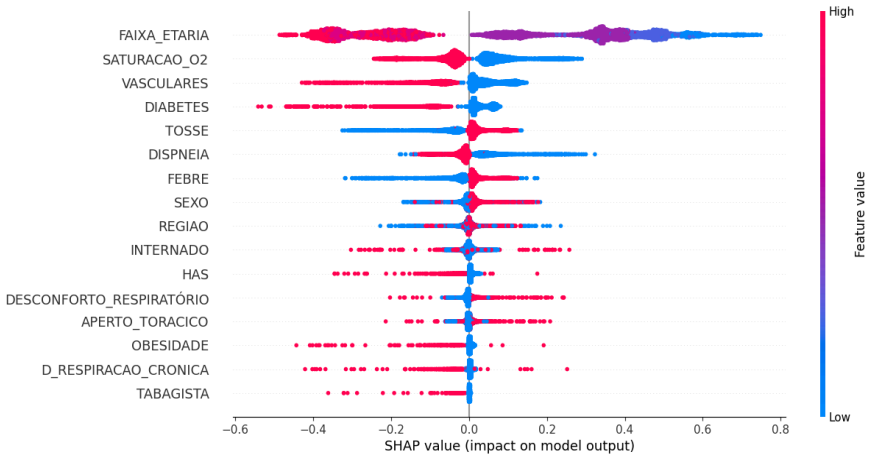
Finalmente, foi realizada uma análise detalhada com o uso do método SHAP para interpretar a importância das variáveis nos modelos selecionados. Nas figuras abaixo são apresentados os gráficos resultantes das análises SHAP aplicadas aos modelos selecionados. As Figuras 3, 4, 5, 6, 7 e 8 ilustram como cada variável contribuiu para a tomada de decisão dos classificadores, permitindo a compreensão do comportamento dos modelos em cada período representado pelas bases de dados.

Figura 3 – SHAP: RNA em março a outubro de 2020



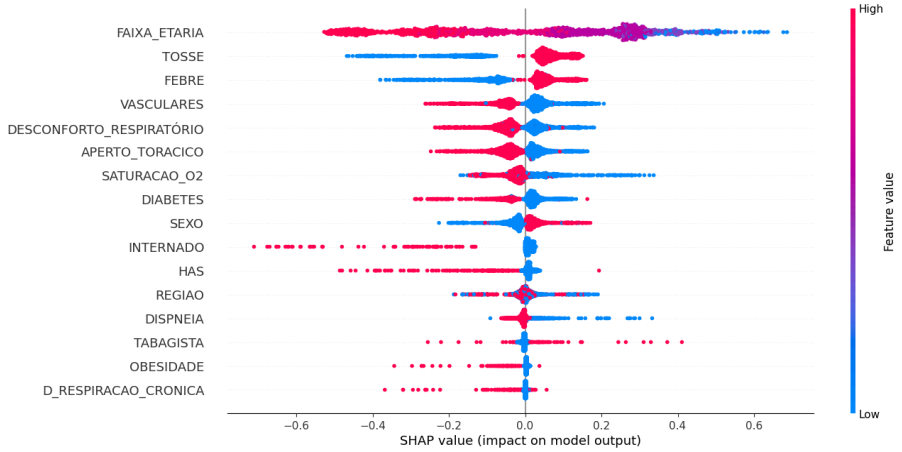
Fonte: Elaborada pelos autores (2024)

Figura 4 – SHAP: Random Forest em março a outubro de 2020



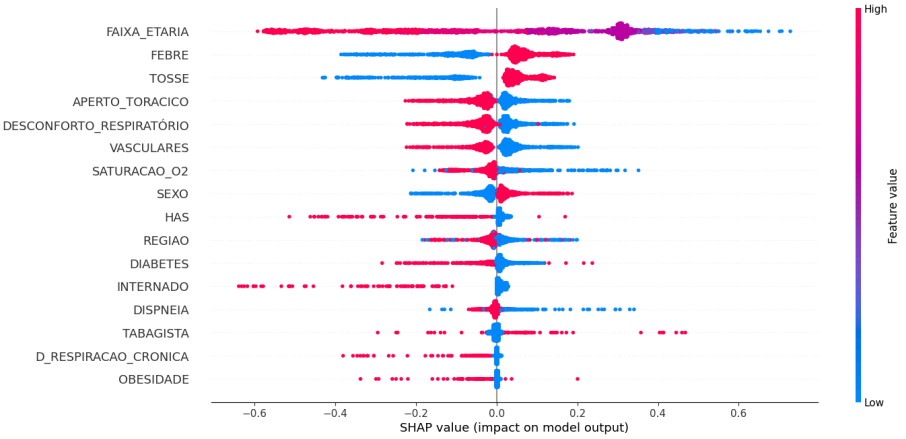
Fonte: Elaborada pelos autores (2024)

Figura 5 – SHAP: Random Forest em outubro de 2020 a janeiro de 2021



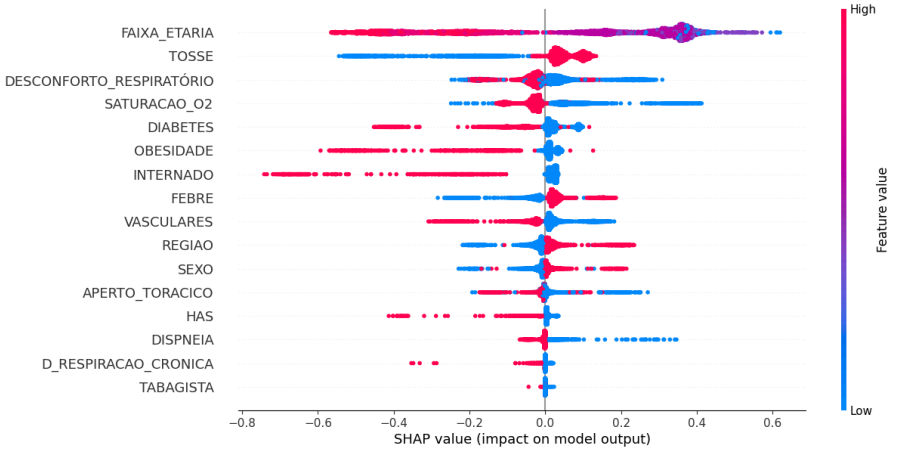
Fonte: Elaborada pelos autores (2024)

Figura 6 – SHAP: SVM em outubro de 2020 a janeiro de 2021



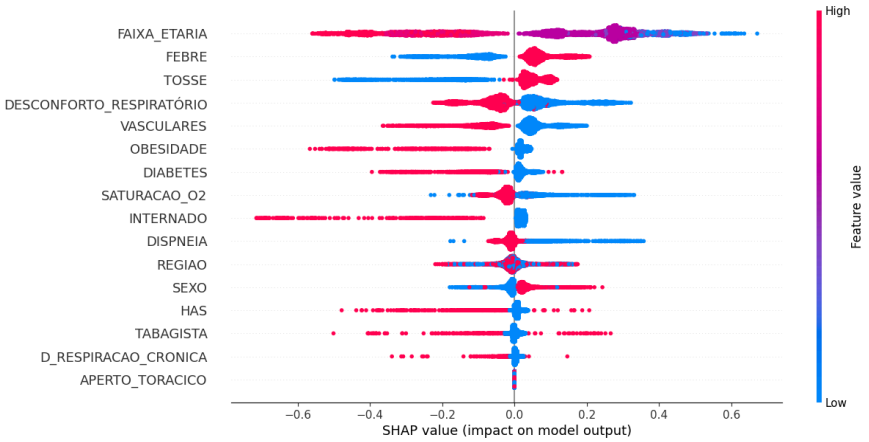
Fonte: Elaborada pelos autores (2024)

Figura 7 – SHAP: Árvore de Decisão em fevereiro de 2021 a outubro de 2021



Fonte: Elaborada pelos autores (2024)

Figura 8 – SHAP: XGBoost em fevereiro de 2021 a outubro de 2021



Fonte: Elaborada pelos autores (2024)

A partir da análise das figuras é possível observar tanto similaridades quanto dissimilaridades nos resultados da técnica SHAP. No período entre março e outubro de 2020, considerando os modelos de Random Forest e RNA, as seis variáveis que mais influenciaram a tomada de decisão para ambos os modelos foram FAIXA_ETARIA, SATURACAO_O2, VASCULARES, DISPNEIA, DIABETES e TOSSE.

Entre outubro de 2020 e janeiro de 2021, considerando os modelos de Random Forest e SVM, as variáveis com maior impacto foram FAIXA_ETARIA, FEBRE, TOSSE, DESCONFORTO_RESPIRATORIO, APERTO_TORACICO e VASCULARES.

No período de fevereiro a outubro de 2021, considerando os modelos de Árvore de Decisão e XGBoost, as variáveis FAIXA_ETARIA, TOSSE, DESCONFORTO_RESPIRATORIO e OBESIDADE estiveram entre as seis mais relevantes para os dois modelos analisados. Contudo, ao comparar os resultados da análise SHAP para os modelos, observou-se uma diferença: na Árvore de Decisão, as variáveis SATURACAO_O2 e DIABETES foram incluídas entre as mais importantes, enquanto no XGBoost, as variáveis FEBRE e VASCULARES ocuparam as posições.

É possível observar que algumas variáveis mantiveram seu impacto. FAIXA_ETARIA e TOSSE foram consistentes em todos os períodos, evidenciando seu impacto na tomada de decisão dos modelos. Nota-se que a FAIXA_ETARIA foi a variável que mais contribuiu para a tomada de decisão em todos os modelos e períodos. Além disso, VASCULARES apresentou relevância nos dois primeiros períodos (março a outubro de 2020 e outubro de 2020 a janeiro de 2021) e exclusivamente no modelo XGBoost em fevereiro a outubro de 2021, enquanto DESCONFORTO_RESPIRATORIO apareceu como variável importante em dois períodos consecutivos (outubro de 2020 a janeiro de 2021 e fevereiro de 2021 a outubro de 2021).

Por outro lado, SATURACAO_O2 foi relevante apenas no período de março a outubro de 2020 e, posteriormente, de forma exclusiva no modelo de Árvore de Decisão em fevereiro de 2021 a outubro de 2021. Embora a FEBRE tenha figurado como relevante no XGBoost em fevereiro de 2021 a outubro de 2021, essa variável

e APERTO_TORACICO emergiram como variáveis importantes somente no período de outubro de 2020 a janeiro de 2021, enquanto OBESIDADE esteve presente entre as 6 variáveis mais relevantes apenas no período de fevereiro a outubro de 2021. Além disso, apenas no período de março a outubro de 2020, a variável DISPNEIA foi identificada entre as mais relevantes. Neste mesmo período e exclusivamente no modelo de Árvore de Decisão em fevereiro a outubro de 2021, a variável DIABETES foi identificada como uma das seis variáveis de maior importância.

Por fim, vale ressaltar que os dois classificadores selecionados em cada período apresentam boa coerência quanto aos seis primeiros atributos apontados pela SHAP como os mais influentes no processo de classificação, sugerindo que esses atributos desempenham um papel relevante na predição, reforçando sua importância nos resultados analisados.

6 CONCLUSÃO E TRABALHOS FUTUROS

Neste estudo, inicialmente foi realizada a coleta e análise exploratória dos dados, seguindo para o treinamento de modelos de classificação. Ao analisar os primeiros resultados obtidos das métricas de Acurácia, F1 *Score* e da Matriz de Confusão, observou-se um viés em relação à classe 1, representando Não Óbito.

Para mitigar esse viés, após uma análise, adotou-se uma técnica de balanceamento de classes, a *undersampling*, que permitiu a melhoria do desempenho preditivo para a classe de Óbito. Adicionalmente, após a realização de uma otimização de hiperparâmetros, foi possível a realização de 30 execuções de treino e teste dos modelos, obtendo as métricas em cada execução. Após isso, foram geradas as médias e os desvios padrões dos resultados. A análise das médias das métricas possibilitou a avaliação e escolha dos modelos que seguiram para a SHAP.

Além disso, a SHAP possibilitou identificar as seis variáveis mais influentes nos processos decisórios dos modelos em cada período analisado, permitindo a realização da análise de similaridades e dissimilaridades nos resultados. Desse modo, o objetivo do presente estudo foi alcançado com a realização de uma análise comparativa das características epidemiológicas da COVID-19 no período de 2020 a 2021 utilizando o método SHAP, tendo sido possível uma compreensão dos fatores epidemiológicos que potencialmente impactaram a evolução da pandemia.

Embora o objetivo principal tenha sido alcançado, ainda existe margem para a melhoria dos resultados obtidos, especialmente em relação às acurácias dos modelos. Como trabalhos futuros, recomenda-se a exploração de estratégias para melhorar a acurácia dos modelos, como a aplicação de métodos de seleção de atributos e a utilização de algoritmos classificadores mais complexos. Ademais, é válido considerar explorar a otimização de hiperparâmetros, incluindo a avaliação de novos hiperparâmetros que não foram considerados neste estudo, bem como a redefinição ou expansão do espaço de busca utilizado.

AGRADECIMENTOS

Agradecemos à Dra. Louisiana Regadas de Macedo Quinino e a FIOCRUZ-PE pela disponibilização das bases de dados utilizadas neste trabalho.

REFERÊNCIAS

AKIBA, Takuya; SANO, Shotaro; YANASE, Toshihiko; OHTA, Takeru; KOYAMA, Masanori. **Optuna: A Next-generation Hyperparameter Optimization Framework**. *In*: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, [*S.l.*], p. 2623-2631, 2019.

ALBAHRI, Ahmed. *et al*. A systematic review of trustworthy and explainable artificial intelligence in healthcare: assessment of quality, bias risk, and data fusion. **Information Fusion**, [*S.l.*], v. 96, p. 156-191, 2023. DOI: 10.1016/j.inffus.2023.03.008.

ALI, Sajid. *et al*. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. **Information Fusion**, [*S.l.*], v. 99, p. 101805, 2023. DOI: 10.1016/j.inffus.2023.101805.

AN, Qi *et al*. A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities and challenges. **Sensors**, [*S.l.*], v. 23, n. 9, p. 4178, 2023. DOI: 10.3390/s23094178.

CASCELLA, Marco *et al*. Features, evaluation, and treatment of coronavirus (COVID-19). **STATPEARLS**, Treasure Island, Flórida, EUA, 2023. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK554776/. Acesso em: 20 de julho de 2024.

CERVANTES, Jair. *et al*. A comprehensive survey on support vector machine classification: applications, challenges and trends. **Neurocomputing**, [*S.l.*], v. 408, p. 189-215, 2020. DOI: 10.1016/j.neucom.2019.10.118.

CHEN, Tianqi; GUESTRIN, Carlos. **XGBoost: a scalable tree boosting system**. *In*: PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. São Francisco, California, EUA, 2016. p. 785-794. Nova Iorque, NI: Association for Computing Machinery. DOI: 10.1145/2939672.2939785.

DEBJIT, Kumar *et al*. An improved machine-learning approach for COVID-19 prediction using Harris Hawks optimization and feature analysis using SHAP. **Diagnostics**, [*S.l.*], v. 12, n. 5, p. 1023, 2022. DOI: 10.3390/diagnostics12051023.

EMAMI, Amir *et al*. Prevalence of underlying diseases in hospitalized patients with COVID-19: a systematic review and meta-analysis. **Archives of Academic Emergency Medicine**, [*S.l.*], v. 8, n. 1, p. e35, 2020.

FACELI, Katti *et al*. **Inteligência Artificial: uma abordagem de aprendizado de máquina**. 2. ed. Rio de Janeiro: LTC, 2021. p. 1, 3, 26, 28, 29, 35, 36, 43, 62

FONSECA, Vinicius L. N. *et al*. **Classification of Tropical Disease-carrying Mosquitoes Using Deep Learning and SHAP**. *In*: SIMPÓSIO BRASILEIRO DE COMPUTAÇÃO APLICADA À SAÚDE (SBCAS), 23. , 2023, São Paulo/SP. Anais Do XXIII Simpósio Brasileiro De Computação Aplicada À Saúde. Porto Alegre:

Sociedade Brasileira de Computação, 2023. p. 25-34. DOI: 10.5753/sbcas.2023.229406.

FREITAS, Carlos M.; BARCELLOS, Christovam; VILLELA, Daniel A. M. Observatório Covid-19 Fiocruz: uma análise da evolução da pandemia de fevereiro de 2020 a abril de 2022. **Ciência & Saúde Coletiva**, [*S.l.*], v. 28, n. 10, p. 2845-2855, 2023. DOI: 10.1590/1413-812320232810.10412023.

GUAN, Wei-jie *et al*. Clinical characteristics of coronavirus disease 2019 in China. **New England Journal of Medicine**, [*S.l.*], v. 382, n. 18, p. 1708-1720, 2020. DOI: 10.1056/NEJMoa2002032.

HULSEN, Tim *et al*. From big data to better patient outcomes. **Clinical Chemistry and Laboratory Medicine (CCLM)**, [*S.l.*], v. 61, n. 4, p. 580-586, 2022. DOI: 10.1515/cclm-2022-1096.

HULSEN, Tim. Explainable artificial intelligence (XAI): concepts and challenges in healthcare. **AI**, [*S.l.*], v. 4, n. 3, p. 652-666, 2023. DOI: 10.3390/ai4030034.

IGUAL, Laura; SEGUÍ, Santi. Regression Analysis. In: Introduction to Data Science. **Undergraduate Topics in Computer Science**. Cham: Springer, 2017. p. 07. DOI: 10.1007/978-3-319-50017-1_6.

JIA, Zhenge *et al*. The importance of resource awareness in artificial intelligence for healthcare. **Nature Machine Intelligence**, [*S.l.*], v. 5, p. 687-698, 2023. DOI: 10.1038/s42256-023-00670-0.

JIJO, Bahzad T.; ABDULAZEEZ, Adnan. M. Classification based on decision tree algorithm for machine learning. **Journal of Applied Science and Technology Trends**, [*S.l.*], v. 2, n. 1, p. 20-28, 2021. DOI: 10.38094/jastt20165.

KLOKNER, Sarah G. M. *et al*. Epidemiological profile and risk factors predictors of COVID-19 in southern Brazil. **Research, Society and Development**, [*S.l.*], v. 10, n. 3, p. e17710313197, 2021. DOI: 10.33448/rsd-v10i3.13197.

KUMAR, Yogesh K. *et al.* Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. **J Ambient Intell Human Comput 14**, [*S.l.*], 8459–8486, 2023. DOI: 10.1007/s12652-021-03612-z

KUSHWAH, Jitendra S. *et al*. Comparative study of regressor and classifier with decision tree using modern tools. **Materials Today: Proceedings**, [*S.l.*], v. 56, p. 3571-3576, 2022. Nota: First International Conference on Design and Materials. DOI: 10.1016/j.matpr.2021.11.635.

LEE, Eun Hak. Exploring transit use during COVID-19 based on XGB and SHAP using smart card data. **Journal of Advanced Transportation**, [*S.l.*], v. 2022, n. 1, p. 6458371, 2022. DOI: 10.1155/2022/6458371.

LIU, Lianhua *et al*. Predictive model and risk analysis for peripheral vascular disease in type 2 diabetes mellitus patients using machine learning and Shapley additive explanation. **Frontiers in Endocrinology**, [*S.l.*], v. 15, 2024. DOI: 10.3389/fendo.2024.1320335.

LUNDBERG, Scott M.; LEE, Su-In. **A unified approach to interpreting model predictions**. *In*: PROCEEDINGS OF THE 31ST INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NIPS 2017). Long Beach, Califórnia, EUA, 2017. p. 4768-4777. Red Hook, NI: Curran Associates Inc.

MARINHO, Pedro R. D. *et al.* Covid-19 in Brazil: a sad scenario. **Cytokine & Growth Factor Reviews**, [*S.l.*], v. 58, p. 51-54, 2021. DOI: 10.1016/j.cytogfr.2020.10.010

PAULI, Suellen T. Z. de; KLEINA, M.; BONAT, W. H. MULTILAYER PERCEPTRON ARTIFICIAL NEURAL NETWORKS: AN APPROACH FOR LEARNING THROUGH THE BAYESIAN FRAMEWORK. **Brazilian Journal of Biometrics**, [*S.l.*], v. 39, n. 1, p. 45–59, 2021. DOI: 10.28951/rbb.v39i1.495.

RAINIO, Oona; TEUHO, Jarmo; KLÉN, Riku. Evaluation metrics and statistical tests for machine learning. **Scientific Reports**, [*S.l.*], v. 14, p. 6086, 2024. DOI: 10.1038/s41598-024-56706-x.

RODRIGUES, Gustavo; KREUTZ, Diego. **Modelo preditivo para classificação de risco de óbito de pacientes com COVID-19 utilizando dados abertos**. *In*: SIMPÓSIO BRASILEIRO DE COMPUTAÇÃO APLICADA À SAÚDE (SBCAS), 22. , 2022, Teresina. Anais Do XXII Simpósio Brasileiro De Computação Aplicada À Saúde. Porto Alegre: Sociedade Brasileira de Computação, 2022. p. 144-155. ISSN 2763-8952. DOI: 10.5753/sbcas.2022.222494.

ROMANO, Donato *et al*. Insights from explainable artificial intelligence of pollution and socioeconomic influences for respiratory cancer mortality in Italy. **Journal of Personalized Medicine**, [*S.l.*], v. 14, n. 4, p. 430, 2024. DOI: 10.3390/jpm14040430.

ROUQUAYROL, Maria Zélia; GURGEL, Marcelo. **Rouquayrol: epidemiologia e saúde**. 8. ed. Rio de Janeiro: MedBook, 2018. 719 p. 70 - 73.

SARKER, Iqbal H. Machine learning: algorithms, real-world applications and research directions. **SN Computer Science**, [*S.l.*], v. 2, p. 160, 2021. DOI: 10.1007/s42979-021-00592-x

SCHONLAU, Matthias; ZOU, Rosie Y. The random forest algorithm for statistical learning. **The Stata Journal**, [*S.l.*], v. 20, n. 1, p. 3-29, 2020. DOI: 10.1177/1536867X20909688.

SHARMA, Anshika *et al*. COVID-19: A review on the novel coronavirus disease evolution, transmission, detection, control and prevention. **Viruses**, Basel, v. 13, n. 2, p. 202, 2021. DOI: 10.3390/v13020202.

SIDEY-GIBBONS, Jenni A. M. *et al.* Machine learning in medicine: a practical introduction. ***BMC Medical Research Methodology***, [*S.l.*], v. 19, p. 64, 2019. DOI: 10.1186/s12874-019-0681-4.

UDDIN, Shahadat *et al*. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. **Scientific Reports**, [*S.l.*], v. 12, p. 6256, 2022. DOI: 10.1038/s41598-022-10358-x.

ULRICHSEN, Felipe C.; SENA, Alexandre C.; PÔRTO, Luís C. M. S.; FIGUEIREDO, Karla. **Avaliação do Uso de IA para Auxiliar no Diagnóstico de Casos de COVID-19 para Diferentes Surtos**. *In*: SIMPÓSIO BRASILEIRO DE COMPUTAÇÃO APLICADA À SAÚDE (SBCAS), 22. , 2022, Teresina. Anais Do XXII Simpósio Brasileiro De Computação Aplicada À Saúde. Porto Alegre: Sociedade Brasileira de Computação, 2022. p. 108-119. DOI: 10.5753/sbcas.2022.222467.

VANACORE, Amalia; PELLEGRINO, Maria S.; CIARDELLI, Armando C. Fair evaluation of classifier predictive performance based on binary confusion matrix. **Computational Statistics**, [*S.l.*], v. 39, p. 363-383, 2024. DOI: 10.1007/s00180-022-01301-9.

WHITE, Ryen W.; MARCHIONINI, Gary; MURESAN, Gheorghe. **Evaluating exploratory search systems**. Information Processing and Management 44, 2 (2008), 433.

WOODMAN, Richard J. *et al*. A comprehensive review of machine learning algorithms and their application in geriatric medicine: present and future. **Aging Clinical and Experimental Research**, [*S.l.*], v. 35, p. 2363-2397, 2023. DOI: 10.1007/s40520-023-02552-2.

WORLD HEALTH ORGANIZATION. **WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020**. [*S.l.*], 2020. Disponível em: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. Acesso em 20 de julho de 2024.

ZHOU, Zhi-Hua. **Machine learning**. Tradução de Shaowu Liu. 1. ed. Singapura: Springer Singapore. p. 2, 4, 63, 2021. DOI: 10.1007/978-981-15-1967-3.

APÊNDICE A - OTIMIZAÇÃO DE HIPERPARÂMETROS

Tabela A.1 – Espaço de busca da otimização de hiperparâmetros

Classificador	Hiperparâmetro	Espaço de busca
Support Vector Machine (SVM)	C	(0.01, 10.0)
	kernel	["linear", "rbf", "poly"]
	max_iter	(500, 10000, step=100)
Regressão Logística	solver	["lbfgs","liblinear","saga"]
	penalty	["l2","l1","none"]
	C	(0.01, 10.0)
K-nearest neighbors (KNN)	n_neighbors	(3, 20)
	weights	["uniform", "distance"]
	metric	["euclidean", "manhattan", "minkowski"]
Rede Neural Artificial (RNA)	hidden_layer_sizes	[(50,), (100,), (50, 50)]
	alpha	(1e-5, 1e-2)
	learning_rate_init	(1e-4, 1e-2)
XGBoost	learning_rate	(0.01, 0.3)
	max_depth	(3, 10)
	subsample	(0.5, 1.0)
Random Forest	n_estimators	(50, 300)
	max_depth	(3, 20)
	min_samples_split	(2, 20)
Árvore De Decisão	max_depth	(3, 20)
	min_samples_split	(2, 20)
	min_samples_leaf	(1, 20)

Fonte: Elaborada pelos autores (2025)