



UNIVERSIDADE FEDERAL DE PERNAMBUCO  
CENTRO DE INFORMÁTICA  
GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

JOÃO RAFAEL SANTOS CAMELO

**Análise de Dados Públicos da COVID-19 em Recife  
Utilizando Aprendizagem de Máquina**

Recife

2021

JOÃO RAFAEL SANTOS CAMELO

**Análise de Dados Públicos da COVID-19 em Recife  
Utilizando Aprendizagem de Máquina**

Trabalho apresentado ao Programa de Graduação em Sistemas de Informação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação

**Orientador:** Fernando Maciano de Paula Neto

Recife

2021

## AGRADECIMENTOS

À minha família, que com seu incentivo e apoio incondicional possibilitaram todas minhas conquistas presentes e futuras.

Ao meu orientador, Fernando Maciano de Paula Neto, pelo suporte neste pouco tempo disponível, pelo seu direcionamento, correções e prontidão.

Aos amigos e companheiros de curso que fizeram parte da minha formação e que vão continuar presentes em minha vida.

À Prefeitura do Recife, pela disponibilização das estatísticas que foram de grande utilidade para a elaboração deste trabalho científico.

Aos profissionais de saúde, que serviram em linha de frente em momentos de crise.

*Without data you're just another person with an opinion.*

— William Edwards Deming

## RESUMO

O combate à COVID-19 se tornou um grande desafio da saúde mundial, sendo documentados mais de 600 mil casos em Recife, Pernambuco, até setembro de 2021, onde mais de 7 mil destes resultaram no óbito do paciente. Por meio da ampla coleta de dados demográficos e sintomáticos realizados pela Prefeitura do Recife, este trabalho comparou a efetividade de diferentes métodos de classificação por aprendizagem de máquina na análise de fatores de risco para casos graves da doença e óbito do paciente. Dados relacionados à vacinação foram utilizados de modo a identificar o progresso da vacinação na ocorrência de cada caso. O modelo *XGBoost* alcançou uma acurácia média de 92% na previsão de casos graves, e 95% na previsão de óbitos. Foram investigados também cenários sem dados sintomáticos, representando pacientes pré-clínicos, e filtragens por progresso da vacinação, para identificar mudanças nos fatores de risco. Interpretações dos modelos gerados foram discutidos, percebendo-se idade elevada, doenças cardiovasculares, hipertensão, diabetes e obesidade como maiores riscos para casos graves e óbitos, sendo a presença da vacinação um fator decisivo na diminuição da severidade dos casos.

**Palavras-chaves:** Aprendizagem de máquina. COVID-19. Fatores de risco. Vacinação.

## ABSTRACT

Combating COVID-19 has become a major global health challenge, with more than 600,000 cases having been documented in Recife, Pernambuco, until September 2021, where more than 7,000 of these resulted in the patient's death. Through the extensive collection of demographic and symptomatic data carried out by the City of Recife, this work compared the effectiveness of different classification methods by machine learning in the analysis of risk factors for severe cases of the disease and patient's death. Data related to vaccination were used to identify the progress of vaccination in the occurrence of each case. The *XGBoost* model achieved an average accuracy of 92% in predicting severe cases, and 95% in predicting death. Scenarios without symptomatic data, representing pre-clinical patients, and screening for vaccination progress were also investigated to identify changes in risk factors. Interpretations of the generated models were discussed, noting old age, cardiovascular diseases, hypertension, diabetes and obesity as greater risks for severe cases and deaths, with the presence of vaccination being a decisive factor in reducing the severity of cases.

**Keywords:** COVID-19. Machine learning. Risk factors. Vaccination.

## LISTA DE FIGURAS

Figura 1 – Exemplo de kNN com k = 3 e k = 6 . . . . .	15
Figura 2 – Exemplo de Árvore de Decisão . . . . .	16
Figura 3 – Abstração de uma Random Forest . . . . .	17
Figura 4 – Abstração de Gradient Boosting . . . . .	18
Figura 5 – Curva ROC de uma execução de XGBoost . . . . .	23
Figura 6 – Gráfico SHAP de interpretação de relevância de atributos. . . . .	38
Figura 7 – Gráfico da porcentagem de casos graves para idades e vacinação . . . . .	41
Figura 8 – Gráfico da porcentagem de óbitos para idades e vacinação . . . . .	41
Figura 9 – Gráfico SHAP de interpretação de relevância de atributos na classificação para o modelo XGBoost no conjunto de dados balanceado com AUC-ROC de 92,50% . . . . .	48
Figura 10 – Gráfico SHAP de interpretação de relevância de atributos na classificação para o modelo Gradient Boosting no subconjunto sem colunas de sintomas com AUC-ROC de 75,11% . . . . .	49
Figura 11 – Gráfico SHAP de interpretação de relevância de atributos na classificação para o modelo Light Gradient Boosting no subconjunto sem vacinação com AUC-ROC de 88,29% e vacinação acima de 30% com AUC-ROC de 98,33% . . . . .	50
Figura 12 – Gráfico SHAP de interpretação de relevância de atributos na classificação para o modelo XGBoost no subconjunto de óbitos com AUC-ROC de 96,24% . . . . .	53
Figura 13 – Gráfico SHAP de interpretação de relevância de atributos na classificação para o modelo XGBoost no subconjunto de óbitos sem colunas de sintomas com AUC-ROC de 86,57% . . . . .	53

## LISTA DE TABELAS

Tabela 1 – Exemplo de matriz de confusão binária . . . . .	19
Tabela 2 – Exemplo de métricas de uma matriz de confusão binária . . . . .	20
Tabela 3 – Colunas do conjunto de dados de casos leves . . . . .	28
Tabela 5 – Colunas do conjunto de dados de casos graves . . . . .	29
Tabela 6 – Colunas do agrupamento de vacinação por dia . . . . .	30
Tabela 7 – Categorias de sintomas . . . . .	33
Tabela 8 – Categorias de doenças . . . . .	34
Tabela 9 – Contagem das categorias de idade . . . . .	35
Tabela 10 – Contagem das categorias de vacinação . . . . .	35
Tabela 11 – Quantidade de casos leves, graves e óbitos por progresso de vacinação . . .	40
Tabela 12 – Quantidade de casos leves, graves e óbitos por idade . . . . .	41
Tabela 13 – Médias de métricas de algoritmos de classificação na etapa de teste . . . .	42
Tabela 14 – Matriz de confusão de uma execução do <i>XGBoost</i> . . . . .	43
Tabela 15 – Médias de métricas de algoritmos de classificação na etapa de testes com classes balanceadas . . . . .	43
Tabela 16 – Matriz de confusão de uma execução do <i>XGBoost</i> com classes balanceadas	43
Tabela 17 – Comparação de algoritmos sem dados sintomáticos . . . . .	44
Tabela 18 – Comparação de algoritmos prevendo óbito . . . . .	45
Tabela 19 – Comparação de algoritmos prevendo óbito sem dados sintomáticos . . . .	45
Tabela 20 – Comparação de algoritmos prevendo severidade do caso antes da vacinação	46
Tabela 21 – Comparação de algoritmos prevendo óbitos antes da vacinação . . . . .	46
Tabela 22 – Comparação de algoritmos prevendo severidade dos casos após 30% da população vacinada . . . . .	47
Tabela 23 – Comparação de algoritmos prevendo óbitos após 30% da população vacinada	47
Tabela 24 – Valor SHAP médio de cada fator na classificação de casos graves de acordo com o progresso da vacinação em execuções de Light Gradient Boosting . .	51
Tabela 25 – 10 valores SHAP mais relevantes para a classificação de casos graves de acordo com a vacinação em execuções de Light Gradient Boost . . . . .	52

# SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>11</b>
1.1	MOTIVAÇÃO	11
1.2	OBJETIVOS	12
1.2.1	Gerais	12
1.2.2	Específicos	12
<b>2</b>	<b>CONTEXTO</b>	<b>13</b>
2.1	APRENDIZADO DE MÁQUINA NA SAÚDE	13
2.2	ALGORITMOS DE CLASSIFICAÇÃO	14
2.2.1	k-Nearest Neighbors	14
2.2.2	Decision Tree	15
2.2.3	Random Forest	16
2.2.4	Gradient Boosting	17
2.2.4.1	<i>XGBoost</i>	18
2.2.4.2	<i>Light Gradient Boosting</i>	18
2.3	MÉTRICAS DE DESEMPENHO	19
2.3.1	Matriz de Confusão	19
2.3.2	Acurácia	20
2.3.3	Precisão	20
2.3.4	Valor Preditivo Negativo	20
2.3.5	Precisão Macro	21
2.3.6	Sensibilidade	21
2.3.7	Especificidade	21
2.3.8	Sensibilidade Macro	22
2.3.9	F1-Score	22
2.3.10	F1-Score Macro	22
2.3.11	AUC ROC	23
2.4	TRABALHOS RELACIONADOS	23
2.4.1	Trabalhos Internacionais	24

2.4.2	Trabalhos no Brasil . . . . .	25
<b>3</b>	<b>METODOLOGIA . . . . .</b>	<b>26</b>
3.1	FERRAMENTAS UTILIZADAS . . . . .	26
3.2	CONJUNTOS DE DADOS . . . . .	27
3.2.1	Casos Leves . . . . .	27
3.2.2	Casos Graves . . . . .	28
3.2.3	Vacinômetro . . . . .	29
3.3	PRÉ-PROCESSAMENTO DE DADOS . . . . .	31
3.3.1	Formatação . . . . .	31
3.3.2	Filtragem . . . . .	31
3.3.2.1	<i>Colunas dos Casos Leves e Graves</i> . . . . .	31
3.3.2.2	<i>Colunas da Vacinação</i> . . . . .	32
3.3.3	Interpretação . . . . .	32
3.3.3.1	<i>Sintomas</i> . . . . .	33
3.3.3.2	<i>Doenças Preexistentes</i> . . . . .	34
3.3.3.3	<i>Interpretando a Vacinação</i> . . . . .	34
3.3.4	Categorização . . . . .	35
3.3.5	Normalização . . . . .	36
3.3.6	Conjunto de Dados Final . . . . .	36
3.4	EXPERIMENTO . . . . .	36
3.4.1	Etapa de Treinamento . . . . .	36
3.4.2	Etapa de Teste . . . . .	37
3.4.3	Cálculo de Métricas de Desempenho . . . . .	37
3.4.4	Otimização de Parâmetros . . . . .	37
3.4.5	Geração de Gráficos . . . . .	38
3.4.6	Subconjuntos . . . . .	39
3.4.6.1	<i>Omitir dados sintomáticos</i> . . . . .	39
3.4.6.2	<i>Possibilidade de óbito</i> . . . . .	39
3.4.6.3	<i>Progresso da vacinação</i> . . . . .	39
<b>4</b>	<b>RESULTADOS . . . . .</b>	<b>40</b>
4.1	ANÁLISE DO CONJUNTO DE DADOS . . . . .	40
4.2	COMPARAÇÃO ENTRE ALGORITMOS . . . . .	42
4.3	COMPARAÇÃO ENTRE ALGORITMOS EM SUBCONJUNTOS . . . . .	44

4.3.1	Omitindo Dados Sintomáticos . . . . .	44
4.3.2	Prevendo Óbitos . . . . .	45
4.3.3	Filtrando por Progresso de Vacinação . . . . .	46
4.3.3.1	<i>Casos antes da vacina</i> . . . . .	46
4.3.3.2	<i>Casos com 30% da população vacinada</i> . . . . .	47
4.4	GRÁFICOS SHAP . . . . .	47
4.4.1	Conjunto de dados balanceado - XGBoost . . . . .	48
4.4.2	Subconjunto de dados sem colunas de sintomas - Gradient Boosting	49
4.4.3	Subconjunto de dados de vacinação - Light Gradient Boosting . . .	50
4.4.4	Subconjunto de dados de óbitos - XGBoost . . . . .	52
<b>5</b>	<b>CONCLUSÃO</b> . . . . .	<b>54</b>
5.1	TRABALHOS FUTUROS . . . . .	55
	<b>REFERÊNCIAS</b> . . . . .	<b>56</b>
	<b>APÊNDICE A – MAPA DE CALOR DE CORRELAÇÃO</b> . . . . .	<b>60</b>
	<b>APÊNDICE B – MAPA DE CALOR DE CORRELAÇÃO ANTES</b> <b>DA VACINAÇÃO</b> . . . . .	<b>61</b>
	<b>APÊNDICE C – MAPA DE CALOR DE CORRELAÇÃO COM 30%</b> <b>DE VACINAÇÃO</b> . . . . .	<b>62</b>

*Capítulo 1*

## INTRODUÇÃO

### 1.1 MOTIVAÇÃO

O novo coronavírus, chamado síndrome respiratória aguda grave coronavírus 2 (SARS-CoV-2), que causa a doença COVID-19 [1], foi declarado uma pandemia pela Organização Mundial de Saúde (WHO) em março de 2020 [2]. O vírus causou sintomas em mais de 260 milhões de pessoas ao redor do mundo, trazendo mais de 5 milhões de mortes ao total, até novembro de 2021 [3].

No Brasil, o terceiro país mais afetado mundialmente, houveram mais de 22 milhões de casos registrados, com cerca de 610 mil mortes registradas [3]. Enquanto em Recife, Pernambuco, até setembro de 2021, 610 mil casos foram documentados pela Prefeitura do Recife [4], com 30 mil sendo classificados como casos graves, e cerca de 7 mil destes resultando no óbito do paciente [5]. A vacinação é uma das medidas que a Organização Mundial de Saúde (OMS) propõe para a prevenção da doença, e está em progresso no Recife desde janeiro de 2021 [6]. É estimado que até setembro de 2021, cerca de 65% da população [7] tivesse a dosagem recomendada da vacina, com duas doses ou doses únicas.

Embora os sintomas da doença sejam tosse seca, falta de ar, febre, dor de cabeça, dor de garganta e fadiga, uma parte indefinida dos casos é assintomática, ou seja, não apresentam sintomas, mas ainda carregam e espalham o vírus. Certos casos da doença podem evoluir em gravidade, similar a um caso de pneumonia grave, podendo causar morte. Percebe-se que casos mais graves da doença geralmente ocorrem em pessoas idosas ou com doenças crônicas, como diabetes, hipertensão arterial, doenças cardíacas e respiratórias [8].

Neste contexto de ampla coleta de dados pela Prefeitura do Recife, é possível então aplicar modelos de aprendizagem de máquina com o intuito de compreender e prever a doença e sua possível evolução em certos pacientes. Esta abordagem se provou eficaz em diversos estudos nacionais e internacionais, discutidos na seção 2.4. Porém, não foram identificados estudos focados na população da cidade do Recife ou que apliquem aprendizagem de máquina

considerando o progresso da vacinação em uma região específica como atributo.

## 1.2 OBJETIVOS

### 1.2.1 Gerais

Este trabalho tem como objetivo geral avaliar a eficácia de diferentes algoritmos de classificação por aprendizagem de máquina na identificação de variáveis que refletem em uma maior chance de casos de COVID-19 alcançarem uma severidade grave ou resultarem em óbito, utilizando dados demográficos, sintomáticos, e o progresso da vacinação na cidade do Recife.

### 1.2.2 Específicos

- Aplicar algoritmos de classificação por aprendizagem de máquina nas bases de dados disponibilizadas pela Prefeitura do Recife;
- Analisar os resultados e comparar a eficácia dos métodos implementados;
- Identificar fatores de risco e sintomas apresentados, de forma a auxiliar na tomada de decisão do tratamento dos pacientes;
- Verificar o impacto da vacinação em progresso nos fatores de risco identificados.

## Capítulo 2

### CONTEXTO

Este capítulo introduz temas que servem como base para a análise e o desenvolvimento presentes nesse estudo.

A seção 2.1 reflete sobre os avanços, tanto atuais quanto possíveis, providos pela aplicação de inteligência artificial na área de saúde. Em seguida, a seção 2.2 explica o funcionamento teórico dos modelos de aprendizado de máquina utilizados no desenvolvimento deste trabalho. Por fim, a seção 2.4 contém alguns trabalhos relacionados ao presente estudo, servindo como base acadêmica.

#### 2.1 APRENDIZADO DE MÁQUINA NA SAÚDE

Aprendizado de máquina é um ramo da inteligência artificial e ciência da computação que foca no uso de dados e algoritmos para imitar como os humanos aprendem, melhorando gradativamente sua precisão [9]. Algoritmos de aprendizado de máquina são então treinados para reconhecer padrões em dados ou mídias. A partir disso, é possível obter resultados como classificação de dados que não conheciam anteriormente, ou produção de conteúdo se baseando nos padrões percebidos.

Duas áreas da medicina que se beneficiam do aprendizado de máquina são as de diagnóstico e prognóstico de doenças. Diagnóstico consiste em avaliar o estado atual do paciente, tendo visto avanços, por exemplo, utilizar fotos para identificar câncer de pele. Prognóstico consistem em prever a evolução da doença no paciente, como utilizar a coleta de dados de tecnologias vestíveis em pacientes com diabetes para ajudar no tratamento [10].

Uma conquista recente para a medicina veio da rede de inteligência artificial da *Google*, *DeepMind*. Seu algoritmo, *AlphaFold2*, conseguiu prever precisamente a estrutura tridimensional de proteínas a partir da sequência de aminoácidos, um problema enfrentado na área há décadas [11]. Este avanço possibilita um melhor entendimento de como proteínas se comportam e pode trazer uma mudança de paradigma na fabricação de remédios e na compreensão

---

do funcionamento de células [12].

O uso de aprendizado de máquina depende da análise de grandes quantidades de dados, mas a emergência de tecnologias vestíveis e sistemas mais distribuídos trazem bons prospectos para seu uso na área de saúde [13].

## 2.2 ALGORITMOS DE CLASSIFICAÇÃO

Um algoritmo de classificação pode ser definido como uma função que recebe como entrada um conjunto de registros com atributos discretos, incluindo um atributo escolhido como classe, e pesa esses atributos de forma a devolver como saída registros separados em classes. O classificador é treinado para identificar os pesos que resultam na classificação mais precisa, então testado por suas previsões de classe.

Estudos anteriores mostram que é difícil encontrar algoritmos de classificação que tenham bom desempenho em qualquer cenário, então é desejável que suas características sejam levadas em consideração ao tentar identificar um algoritmo que traga boas previsões ao problema em mãos [14].

Então, se baseando nos trabalhos relacionados ao tema, é considerado relevante para o atual estudo analisar o desempenho dos seguintes algoritmos de classificação:

### 2.2.1 k-Nearest Neighbors

kNN, ou *k-Nearest Neighbors*, é um algoritmo de classificação baseado na ideia de que registros de classes similares ficariam aglomerados quando colocados em um vetor de n-dimensões.

Em um vetor cujas dimensões são definidas pelo número de atributos, todos os registros do conjunto de dados de treinamento são colocados em posições conforme os valores de seus atributos, idealmente próximos a outros registros da mesma classe. Então, na etapa de teste, os registros sem classe têm sua posição definida, e por meio da distância dos seus vizinhos mais próximos, é predita a classe de cada registro. A distância pode ser euclidiana, *Hamming*, *Manhattan* ou qualquer outra distância entre dois vetores, porém a euclidiana é a mais comum, e pode ser calculada com a equação abaixo [15].

$$\text{Distância Euclidiana} = \sqrt{\sum_{i=1}^n (x_i - x_j)^2} \quad (2.1)$$

O número de vizinhos mais próximos é definido pelo parâmetro  $k$ , que é um valor inteiro positivo, idealmente ímpar de modo a evitar a predição de classes como ámbas quando há um número igual de classes vizinhas. Este valor muitas vezes é escolhido por tentativa e erro, seguindo as métricas de desempenho do algoritmo [16].

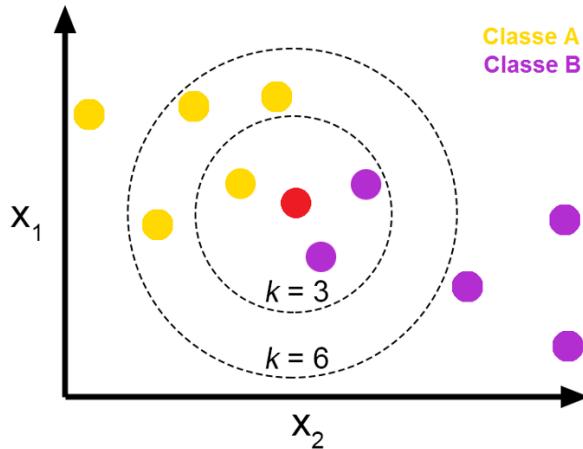


Figura 1 – Exemplo de kNN com  $k = 3$  e  $k = 6$

Fonte: KNN (K-Nearest Neighbors) #1 - Italo José (Medium)

Como os valores são utilizados como posição, kNN se beneficia de uma distribuição uniforme de valores para os atributos, sendo importante realizar a normalização e escala dos dados. O algoritmo também é considerado preguiçoso, não tendo uma etapa de treinamento propriamente dita, também tendo baixa eficiência, pois todos os registros são testados pela sua distância, mas costuma ter um desempenho satisfatório em alguns casos [17].

### 2.2.2 Decision Tree

*Decision Tree*, ou Árvore de Decisão, é um algoritmo interpretável de classificação que divide o conjunto de dados em subconjuntos, através de critérios de decisão, e a partir desses subconjuntos são criadas novas subárvores, até que todos os registros sejam classificados[18].

Os critérios de decisão são funções que dividem os dados entre Verdadeiro e Falso, a partir de um dos atributos dos registros, por exemplo, criando uma subárvore, ou galho, onde todos os registros têm idade maior ou igual a 60, e um outro galho onde os registros têm idade menor que 60. Essa divisão acontece até que o algoritmo decida que não é mais necessário dividir os subconjuntos, criando uma folha, onde todos os registros são classificados como a mesma classe. Os galhos e folhas também são chamados de nós; sendo a árvore construída a

partir de um nó raiz [19].

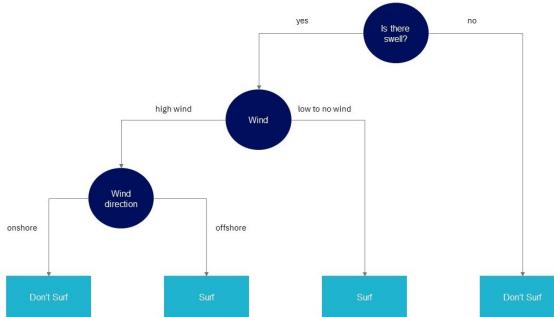


Figura 2 – Exemplo de Árvore de Decisão

Fonte: What is Random Forest - IBM Cloud Education

Contrário ao kNN, a *Decision Tree* possui uma etapa de treinamento, onde o modelo utiliza um conjunto de treinamento para construir a árvore, usando métricas como o índice *Gini* ou *Entropy* para definir o critério de divisão. Então, na etapa de teste, onde os registros novos serão classificados, a árvore é percorrida, através de suas ramificações, até que se encontre um nó folha, onde a classe do registro é predita.

*Overfitting* é um dos maiores problemas enfrentados pelo *Decision Tree*, podendo ser causado por um número excessivo de ramificações da árvore. Isto faz com que o modelo fique dependente demais do conjunto de treinamento, perdendo a flexibilidade necessária para identificar as classes do conjunto de treinamento. *Overfitting* pode ser amenizado ou evitado por técnicas como o *pruning*, que elimina ramificações que não são necessárias, ou limitando o tamanho da árvore [18].

Ainda assim, além de ser simples e satisfatoriamente precisa em muitos casos, um grande ponto positivo da *Decision Tree* é que é um algoritmo interpretável, ou seja, é possível gerar o conjunto de critérios de decisão em forma de árvore, que pode então ser visualizada como um diagrama de decisão.

### 2.2.3 Random Forest

*Random Forest* é um algoritmo de classificação que utiliza uma combinação de *Decision Trees*, denominada floresta, para classificar os dados.

A floresta é composta de um número fixo de árvores de decisão, sendo todas elas construídas a partir do conjunto de treinamento, porém com diferentes subconjuntos de atributos

aleatórios, para que exista uma baixa correlação entre cada árvore. As árvores podem ser configuradas a partir de parâmetros similarmente à *Decision Tree*, além de parâmetros para a floresta, como o número de árvores e o número de atributos aleatórios [20].

Um *ensemble* é um conjunto de classificadores fracos combinados para formar um classificador mais robusto. *Ensembles* são utilizados em outros algoritmos de classificação explicados a seguir, como o *Gradient Boosting*.

O *Random Forest* cria um *ensemble* de Árvores de Decisão na sua etapa de treinamento. Então, na etapa de teste, as predições de várias árvores são combinadas, e a classe predita é a escolhida pela maioria das árvores da floresta. Esta abordagem permite que as árvores cubram seus erros individuais, mas requer mais recursos e sacrifica a interpretabilidade das árvores individuais [21].

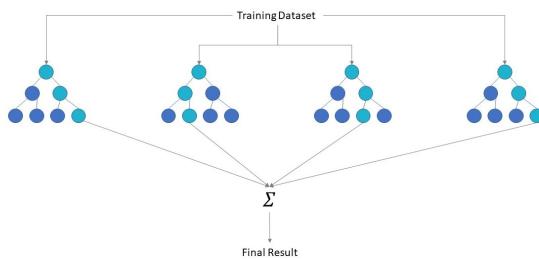


Figura 3 – Abstração de uma Random Forest

Fonte: What is Random Forest - IBM Cloud Education

#### 2.2.4 Gradient Boosting

*Machine Learning Boosting* é um método de criação de *ensembles* de classificadores, onde um modelo inicial é treinado com um conjunto de dados de treinamento, então um segundo modelo é construído com pesos nas predições erradas, de forma que a combinação deles resulte em um classificador melhor que ambos sozinhos, e esse processo é repetido por um número específico de instâncias ou até que o desempenho alcance um nível aceitável ou deixe de melhorar. As predições do *ensemble* são combinadas, e a classe predita é a escolhida pela soma pesada das predições [22].

*Gradient Boosting* é um tipo de *Machine Learning Boosting*, onde *Decision Trees* são construídas de maneira aditiva e sequencial, uma de cada vez, sem alterar as árvores anteriores, e são julgadas por gradientes em uma função de perda, que serve como medida indicativa do

seu desempenho. Um ponto positivo do *Gradient Boosting* é que a sua função de perda é genérica e pode ser escolhida, permitindo que o algoritmo seja adaptado a diferentes tipos de problemas, como usar erro quadrático para regressão e perda logarítmica para classificação [23].

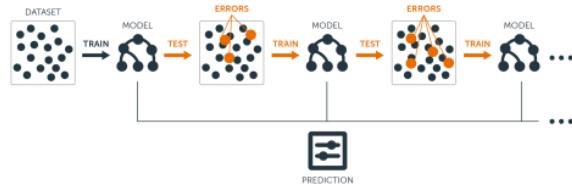


Figura 4 – Abstração de Gradient Boosting

Fonte: Hands-On Machine Learning with R (AI Wiki)

É possível configurar as árvores geradas da mesma forma que as *Decision Trees*, além de parâmetros como a taxa de aprendizado, servindo como um peso para cada adição, geralmente sendo um valor baixo entre 0,1 e 0,3, o que aumenta o número de árvores adicionadas ao modelo, diminuindo o impacto de cada árvore, mas deixando o modelo mais lento. Há também a possibilidade de utilizar *Gradient Boosting* estocástico, utilizando subconjuntos de elementos ou atributos, diminuindo a correlação das árvores similar ao *Random Forest*, mas com um custo computacional maior [23].

#### 2.2.4.1 XGBoost

*Extreme Gradient Boosting*, ou *XGBoost*, é uma variação do *Gradient Boosting* que, contrário à maneira sequencial do *Gradient Boosting*, constrói as árvores paralelamente com uma estratégia por níveis, varrendo os valores de gradiente e usando as somas parciais como avaliação da qualidade de cada divisão possível. Há também várias outras otimizações de *cache*, computação distribuída e adaptações para conjuntos de dados maiores que a capacidade de memória da máquina [24].

#### 2.2.4.2 Light Gradient Boosting

Outra variação do *Gradient Boosting* é a sua versão mais leve, *Light Gradient Boosting*, uma biblioteca de código aberto que busca ser mais eficiente e efetiva, muitas vezes alcan-

çando essa visão até em conjuntos de dados de larga escala [25]. As diferenças vêm de um foco em gradientes maiores (*Gradient-based One Side Sampling*) e a adição de uma seleção automática de características (*Exclusive Feature Bundling*) [26]. Em certos casos, o *Light Gradient Boosting* consegue uma pequena melhora na acurácia em relação ao *XGBoost*, mas consegue ser até 7 vezes mais rápido [27].

## 2.3 MÉTRICAS DE DESEMPENHO

Para comparar os resultados obtidos com os algoritmos de classificação, são utilizadas métricas de desempenho. As predições dos algoritmos são comparadas com os valores reais do conjunto de dados, e a métrica de desempenho é calculada com base nesses valores. As subseções a seguir explicam cada métrica com um exemplo.

### 2.3.1 Matriz de Confusão

A matriz de confusão é uma tabela que mostra a quantidade de acertos e erros de cada algoritmo para cada classe [28].

	<b>Predição: LEVE</b>	<b>Predição: GRAVE</b>
<b>Real: LEVE</b>	114631	85
<b>Real: GRAVE</b>	1949	4198

Tabela 1 – Exemplo de matriz de confusão binária

O valor na primeira célula, **Predição: LEVE** e **Real: LEVE**, indica a quantidade de acertos para a classe LEVE, ou Verdadeiro Positivo (TP), com 114631 acertos. Assim, o valor da segunda célula, **Predição: GRAVE** e **Real: LEVE**, indica quantos casos leves foram preditos como graves, chamados de Falsos Positivos (FP), totalizando 85 erros nesta classe.

Da mesma forma, o valor de **Predição: GRAVE** e **Real: GRAVE**, indica que houveram 4198 acertos na classe de casos graves, chamados de Verdadeiros Negativos (TN). Igualmente, **Predição: LEVE** e **Real: GRAVE** indica quantos casos graves foram preditos como leves, chamados de Falsos Negativos (FN), com 1949 casos graves preditos como leves, uma quantidade considerável e de grande relevância para o problema em questão.

Destes valores, é possível, por meio de cálculos, obter as seguintes métricas.

	<b>Positivo Preditivo</b>	<b>Negativo Preditivo</b>	
<b>Positivo Preditivo</b>	TP 114631	FN 85	<i>Sensibilidade</i>
<b>Negativo Preditivo</b>	FP 1949	TN 4198	<i>Especificidade</i>
	<i>Precisão</i>	<i>Valor Preditivo Negativo</i>	<i>Acurácia</i>

Tabela 2 – Exemplo de métricas de uma matriz de confusão binária

### 2.3.2 Acurácia

A acurácia indica a quantidade de casos que foram preditos corretamente.

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.2)$$

Portanto, a execução do *Random Forest* de exemplo, com a matriz de confusão acima, obteve uma acurácia de 0,98317, ou 98,317%.

### 2.3.3 Precisão

A precisão indica a quantidade de casos leves que foram preditos corretamente, dividido pelo total de previsões positivas. Essa métrica serve para o julgamento da veracidade dos casos leves preditos, porém ignora os casos graves.

$$\text{Precisão} = \frac{TP}{TP + FP} \quad (2.3)$$

Assim, a execução de exemplo obteve uma precisão de 98,328%.

### 2.3.4 Valor Preditivo Negativo

A precisão também pode ser chamada de valor preditivo positivo, então o valor preditivo negativo é a precisão dos casos negativos, ou casos graves, indicando a quantidade de casos graves que foram preditos corretamente, dividido pelo total de previsões negativas. Da mesma forma, a métrica ignora os casos leves, mas traz confiança que os casos graves são realmente graves.

$$\text{Valor Preditivo Negativo} = \frac{TN}{TN + FN} \quad (2.4)$$

A execução de exemplo obteve um valor preditivo negativo de 98,015%

### 2.3.5 Precisão Macro

É possível utilizar a precisão de cada classe para calcular a precisão macro. Em conjuntos com apenas duas classes, a precisão macro é a média da precisão e do valor preditivo negativo, sendo mais dinâmica.

$$\text{Precisão Macro} = \frac{1}{N} \sum_{i=1}^N \frac{TP_i}{TP_i + FP_i} \quad (2.5)$$

A média das métricas então resulta em 98,171%.

### 2.3.6 Sensibilidade

*Recall*, ou sensibilidade, indica a quantidade de casos leves que foram preditos corretamente, dividido pelo total de casos leves. A sensibilidade então serve como uma medida de que as previsões não vão resultar em alarmes falsos, prevendo casos leves como graves.

$$\text{Sensibilidade} = \frac{TP}{TP + FN} \quad (2.6)$$

Obtendo um valor de 99,925%, com uma baixa quantidade de falsos negativos.

### 2.3.7 Especificidade

A especificidade indica a quantidade de casos graves que foram preditos corretamente, dividido pelo total de casos graves. É equivalente à sensibilidade dos casos graves, medindo os casos graves que foram julgados como LEVE, com grande relevância para o conjunto de dados em questão.

$$\text{Especificidade} = \frac{TN}{TN + FP} \quad (2.7)$$

Obtendo um valor de 68,293%, relativamente baixa confiança em casos graves, significando que mais de 30% dos casos graves não foram classificados corretamente.

### 2.3.8 Sensibilidade Macro

Similar à precisão macro, a sensibilidade macro é a média da sensibilidade e da especificidade em conjuntos de dados binários, medindo os falsos positivos e falsos negativos.

$$\text{Sensibilidade Macro} = \frac{1}{N} \sum_{i=1}^N \frac{TP_i}{TP_i + FN_i} \quad (2.8)$$

A média então resulta em 84,109%, que, em comparação com a sensibilidade, informa que a especificidade está baixa.

### 2.3.9 F1-Score

O *F1-Score*, ou F1, é uma métrica composta que utiliza a média harmônica da precisão e da sensibilidade, de forma a medir tanto predições corretas quanto predições falsas dos casos leves.

$$\text{F1-Score} = 2 \times \frac{\text{Precis.} \times \text{Sensib.}}{\text{Precis.} + \text{Sensib}} \quad (2.9)$$

Alternativamente:

$$\text{F1-Score} = \frac{2TP}{2TP + FP + FN} \quad (2.10)$$

Assim obtendo um valor de 99,120%, um valor alto, demonstrando que as predições de casos leves são confiáveis. A mesma métrica pode ser aplicada à classe de casos graves, resultando em 80,5%, um valor bem mais baixo em comparação.

### 2.3.10 F1-Score Macro

É possível então utilizar o *F1-score* de cada classe para calcular o *F1-score* macro, levando em conta não só a precisão e a sensibilidade, mas também o valor preditivo negativo e a especificidade, sendo uma média entre os *F1-scores* das classes.

$$\text{F1-Score Macro} = \frac{1}{N} \sum_{i=1}^N \frac{2TP_i}{2TP_i + FP_i + FN_i} \quad (2.11)$$

Obtendo-se assim um valor que considera todas as métricas anteriores, 89,810%, trazendo uma visão mais ampla do desempenho do algoritmo em questão.

### 2.3.11 AUC ROC

Além do *F1-score*, é possível utilizar as métricas calculadas para plotar a curva Característica de Operação do Receptor (ROC), tendo a sensibilidade no eixo Y contra a especificidade invertida no eixo X. A curva permite não só visualizar o desempenho do classificador e encontrar o ponto ótimo da sensibilidade em função da especificidade, sendo o ponto mais próximo da esquerda superior [29].

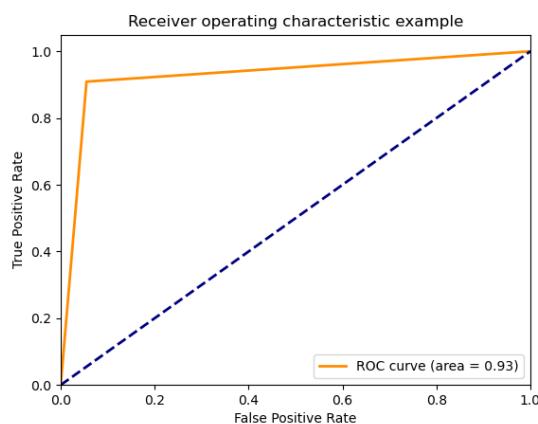


Figura 5 – Curva ROC de uma execução de XGBoost

Com a curva, é possível então calcular a Área Abaixo da Curva (AUC), onde um modelo com 0% de acurácia teria um AUC de 0, e um modelo com 100% de acurácia teria um AUC de 1. Na figura 5, o modelo de exemplo alcançou um AUC de 93,768%.

## 2.4 TRABALHOS RELACIONADOS

Considerando a relevância de aplicar inteligência artificial na área de saúde e a urgência em escala global da pandemia de COVID-19, uma miríade de estudos sobre o tema foram publicados, tanto internacionalmente quanto no Brasil. Alguns trabalhos relacionados ao presente estudo são descritos a seguir. O valor *AUC-ROC*, discutido na subseção 3.4.3, será utilizado como principal métrica de desempenho entre trabalhos.

#### 2.4.1 Trabalhos Internacionais

O trabalho [30] analisou os dados de 214 pacientes de Wuhan, China, com informações sobre suas comorbidades, sintomas e resultados de testes de laboratório, a fim de prever a severidade de casos, de modo similar ao presente trabalho. Teve-se como objetivo manter a interpretabilidade da previsão, portanto, mesmo utilizando outros algoritmos como kNN, redes neurais e *naive Bayes*, decidiu-se focar em *Random Forest*, obtendo um *AUC-ROC* de 99% com os dados laboratoriais, e de 90% com as comorbidades e sintomas.

Utilizando *Gradient Boosting*, o trabalho [31] analisou dados de 99.232 registros de indivíduos testados por COVID-19 em Israel, onde 8393 deles foram confirmados como portadores da doença. O objetivo em questão foi identificar portadores da doença por meio de registros de grupos de idade, sexo, 5 sintomas iniciais e se houve contato com alguém infectado, para que fosse possível priorizar recursos de teste limitados. Foi alcançado um *AUC-ROC* de 90% utilizando *Gradient Boosting*. Como esperado pelos autores, o contato com alguém infectado foi o fator mais importante. Foi observado também que o Ministério da Saúde de Israel não registrou dados sintomáticos suficientes.

Conduzindo uma meta-análise acadêmica, o trabalho [32] buscou determinar o grau em que comorbidades associaram-se com casos graves e óbitos a fim de auxiliar em medidas de tratamento, planejamento e provisionamento. Com um total de 26 estudos analisados e 13.400 amostras, identificaram que doenças pulmonares obstrutivas crônicas, cerebrovasculares, cardiovasculares, diabetes, câncer e hipertensão arterial foram as comorbidades mais significantes para casos graves de COVID-19. Também perceberam que nos estudos analisados, mesmo idade e sexo sendo fortes preditores de mortalidade, em termos de relação entre sintoma e comorbidades, hipertensão e diabetes se relacionaram a pneumonia, enquanto hipertensão se relacionou à síndrome da insuficiência respiratória aguda.

Sendo um dos trabalhos mais similares ao trabalho atual, a pesquisa [33] propôs analisar características dos pacientes, sintomas, diagnósticos e evolução da doença para descobrir os melhores preditores de diagnóstico precoce, possibilitando decisões rápidas nas necessidades de tratamento e isolamento. Como conjunto de dados, utilizou-se dados abertos de 6.512 pacientes de províncias da China. Foi alcançado uma *AUC-ROC* de 89% no algoritmo *XGBoost*. Além deste, também foram utilizados *Gradient Boosting*, *Support Vector Machine*, *Decision Tree* e *Random Forest*, porém com desempenho inferior. Houve um esforço na filtragem dos dados em relação à idade, analisando o desempenho dos algoritmos em grupos de idade

---

diferentes, com acurácia variável, mas conseguiu-se identificar as características mais relevantes para cada grupo.

Em comparação com estes trabalhos internacionais, o presente trabalho analisa casos confirmados, tanto prevendo a severidade quanto o óbito. Não serão utilizados dados laboratoriais, que precisam de mais recursos e tempo para obter. Ainda assim, o conjunto de dados disponibilizado pela Prefeitura do Recife, além de muito mais vasto, possui uma ampla variedade de sintomas e comorbidades, o que pode se mostrar um ponto positivo.

#### 2.4.2 Trabalhos no Brasil

O trabalho [34] analisou 217.580 pacientes de Alagoas (AL), Espírito Santo (ES) e Santa Catarina (SC), de modo a prever casos onde o paciente precisaria de hospitalização. Os dados incluíram idade, sexo, raça, sintomas, comorbidades e a evolução do caso. Foram utilizados os algoritmos de *Decision Tree*, redes neurais e *Support Vector Machine*, alcançando *AUC-ROC* médios de 87%, 90% e 91% para AL, ES e SC. Um viés identificado pelos autores foi o de que estados com melhor infraestrutura geram dados mais confiáveis, refletido no desempenho maior do algoritmo conforme o Índice de Desenvolvimento Humano (IDH) dos estados. Também foi teorizado que a quantidade de leitos disponíveis tenha sido um fator de barulho relevante.

O trabalho [35] envolveu todas as regiões do Brasil, com 113.214 pacientes, onde 50.387 resultaram em óbito, teve como objetivo prever a mortalidade dos casos de COVID-19 no Brasil. Os dados foram similares aos do trabalho anterior [34], tendo também informações de tratamento. Por meio de *Support Vector Machine*, Regressão Logística, *Gradient Boosted Decision Trees* e *Random Forest*, foi possível obter *AUC-ROCs* na predição da mortalidade e necessidade de hospitalização de 79% e 69% respectivamente. A região do hospital foi observada como um fator relevante. A região Nordeste teve a maior razão de probabilidade de mortalidade (2,185) entre todos os fatores, sendo mais que o dobro da região Sudeste (1,030).

Neste contexto, o presente trabalho analisará dados somente de Recife, na região Nordeste, deixando de lado diferenças de desenvolvimento. Serão analisados ambos os cenários de severidade e de óbito, como nestes estudos discutidos. Um fator não estudado anteriormente, tanto no Brasil quanto em outros países, é a vacinação em progresso, que receberá foco neste trabalho.

## Capítulo 3

# METODOLOGIA

Neste capítulo serão discutidas a proposta do trabalho e a metodologia utilizada para a aplicação dos algoritmos de classificação discutidos na seção 2.2.

Na seção 3.1 serão apresentadas as ferramentas e bibliotecas utilizadas no desenvolvimento do trabalho. A seção 3.2 apresenta os conjuntos de dados obtidos e a seção 3.3 apresenta o processo de limpeza e interpretação dos mesmos. A seção 3.4 propõe maneiras de utilizar os algoritmos de classificação discutidos no conjunto de dados, e como obter informações relevantes quanto a ambos.

### 3.1 FERRAMENTAS UTILIZADAS

O desenvolvimento e execução do projeto foi realizado por meio da linguagem *Python*, na versão 3.8.5 [36]. A motivação da escolha foi devido à sua relevante presença acadêmica na área de inteligência artificial. Além disso, há uma ampla disponibilidade de bibliotecas no tema, das quais as seguintes foram usadas:

- *NumPy*: [37] - Ferramentas de manipulação de matrizes e vetores;
- *Pandas*: [38] - Ferramentas de manipulação de dados;
- *ScikitLearn*: [39] - Algoritmos de aprendizado de máquina;
- *SHAP*: [40] - Ferramentas de análise de impacto e geração de gráficos;
- *Seaborn*: [41] - Geração de gráficos de correlação;
- *MatPlotLib*: [42] - Geração de gráficos diversos;

### 3.2 CONJUNTOS DE DADOS

Todos os conjuntos de dados utilizados neste trabalho foram coletados e disponibilizados pela Prefeitura do Recife e Secretaria de Saúde do Recife e são de público acesso.

No total foram utilizados três conjuntos de dados:

#### 3.2.1 Casos Leves

No Portal de Dados Abertos da Prefeitura do Recife foi disponibilizado um conjunto de dados contendo as notificações de casos leves de COVID-19 dos residentes do Recife, realizados pela rede de saúde no sistema e-SUS Notifica do DATASUS desde abril de 2020 [4]. O conjunto de dados foi disponibilizado em 10 de julho de 2021, atualizado periodicamente, sendo que a última atualização até o desenvolvimento deste trabalho foi em 27 de setembro de 2021.

Os dados são em sua maioria texto, com 16 colunas e um total de 574.292 linhas. O conjunto de dados está disponível em formato CSV na seguinte URL: <<http://dados.recife.pe.gov.br/dataset/casos-leves-covid-19>>

Coluna	Tipo	Descrição
sexo	texto	Sexo informado pelo usuário.
idade	texto	Idade em anos.
data_notificacao	data	Data da notificação do caso no e SUS Notifica.
data_inicio_sintomas	data	Data de início dos sintomas do caso no e SUS Notifica.
sintomas	texto	Sintomas apresentados pelo usuário.
outros_sintomas	texto	Demais sintomas apresentados pelo usuário.
evolucao_caso	texto	Cura, Em tratamento domiciliar, Ignorado, Internado/Internado, UTI, Óbito
em_tratamento_domiciliar	texto	Confirmação Laboratorial/Confirmado Clínico Epidemiológico/Confirmado Clínico-Imagem/Confirmado por Critério Clínico/Descartado/Síndrome Gripal Não Especificada
doencas_preexistentes	texto	Comorbidades prévias informadas pelo usuário
raca_cor	texto	Raça/Cor informada pelo usuário.
etnia	texto	Tipo da etnia se a categoria Raça/Cor for preenchida com "Indígena"
profissional_saude	texto	Área de atuação médica do usuário.
cbo	texto	Área de atuação militar do usuário.
municipio_notificacao	texto	Município da notificação.
bairro	texto	Bairro da notificação.
ds	texto	Distrito Sanitário.

Tabela 3 – Colunas do conjunto de dados de casos leves

### 3.2.2 Casos Graves

Similar ao conjunto de dados anterior, no Portal de Dados Abertos da Prefeitura do Recife foi disponibilizado um conjunto de dados contendo as notificações de Síndrome Respiratória Aguda Grave de residentes do Recife, realizados pela rede de saúde no sistema Notifica PE desde março de 2020 [5]. O conjunto de dados de casos graves foi disponibilizado e atualizado nos mesmos dias que o conjunto de dados de casos leves.

Não é informado como os casos foram categorizados como leves ou graves, tendo a Secretaria de Saúde do Recife e profissionais de saúde a responsabilidade pelo julgamento. Por meio de análise, é possível observar que os sintomas desempenham um papel crucial nesta categorização, em especial a saturação de oxigênio estando acima ou abaixo de 95%.

Os dados são em sua maioria texto, com 16 colunas e um total de 30.740 linhas. O conjunto de dados está disponível no formato CSV na URL: <<http://dados.recife.pe.gov.br/dataset/casos-graves-covid-19>>

Coluna	Tipo	Descrição
data_notificacao	data	Data da notificação do caso no Notifica PE.
sexo	texto	Sexo declarado pelo paciente.
idade	texto	Idade em anos.
data_inicio_sintomas	data	Data de início dos sintomas do caso no Notifica PE.
raca	texto	Raça/cor declarada pelo paciente.
etnia	texto	Tipo da etnia se a categoria Raça/Cor for preenchida com "Indígena"
sintomas_apresentados	texto	Sintomas apresentados pelo paciente.
outros_sintomas	texto	Demais sintomas apresentados pelo usuário.
doencas_preeexistentes	texto	Comorbidades prévias informadas pelo paciente.
outras_doencas_preeexistentes	texto	Demais comorbidades prévias informadas pelo paciente.
evolucao	texto	Evolução clínica do paciente (Internado leito de isolamento, Internado em UTI, Isolamento domiciliar, Óbito e Recuperado)
classificacao_final	texto	Classificação do caso de acordo com critérios laboratoriais (Confirmado, Descartado e Em Análise)
data_obito	data	Data do óbito do paciente.
profissional_saude	texto	Área de atuação médica do paciente.
categoria_profissional	texto	Área de atuação profissional do paciente.
municipio_notificacao	texto	Município da notificação.
bairro	texto	Bairro da notificação.
ds	texto	Distrito Sanitário.

Tabela 5 – Colunas do conjunto de dados de casos graves

### 3.2.3 Vacinômetro

Para fins de identificar possíveis impactos do progresso da vacina na cidade, decidiu-se utilizar um conjunto de dados que contém informações sobre a vacinação da população do Recife, realizados pela Secretaria de Saúde do Recife.

Inicialmente foi utilizado o conjunto de dados da Relação de pessoas vacinadas - COVID-19, também disponível no Portal de Dados Abertos da Prefeitura do Recife [43], contendo informações de cada dosagem de vacinação aplicada, assim como dados demográficos de cada pessoa vacinada. Porém, se observou que os dados não estavam de acordo com as informações disponibilizadas no medidor oficial de vacinação do Recife, o Vacinômetro [6].

O Vacinômetro é mantido pela Prefeitura do Recife, contendo dados do *App Recife Vacina* e *Google Forms*, atualizado diariamente. Sendo assim, foi decidido utilizar o conjunto de dados disponibilizado pelo Vacinômetro, que contém informações de maneira mais organizada e completa.

A estrutura do conjunto de dados é a seguinte:

- Número de doses recebidas por tipo de vacina;
- Consolidado do esquema vacinal;
- Número de doses aplicadas da vacina contra a COVID-19 segundo sexo;
- Número de doses aplicadas da vacina contra a COVID-19 segundo raça cor;
- Número de doses aplicadas da vacina contra a COVID-19 por distrito sanitário;
- Número de doses aplicadas da vacina contra a COVID-19 por grupo prioritário;
- Número de doses aplicadas da vacina contra a COVID-19 por dia;
- Controle das doses distribuídas e aplicadas segundo locais de vacinação e tipo de vacina.

Foi decidido utilizar o subconjunto agrupado por datas "Número de doses aplicadas da vacina contra a COVID-19 por dia", com a seguinte estrutura:

<b>Coluna</b>	<b>Tipo</b>	<b>Descrição</b>
Data de Vacinação	data	Dia considerado.
Dose 1	número	Contagem de primeiras doses aplicadas.
Dose 2	número	Contagem de segundas doses aplicadas.
Dose de Reforço	número	Contagem de doses de reforço aplicadas.
Dose única	número	Contagem de doses únicas aplicadas.
Total	número	Soma de todas as doses aplicadas no dia.

Tabela 6 – Colunas do agrupamento de vacinação por dia

Os dados são em sua maioria numéricos e datas, com formatos variados. Todo o conjunto de dados está disponível na página e em formato ODS e PDF na seguinte URL: <<https://conectarecife.recife.pe.gov.br/vacinometro/>>

### 3.3 PRÉ-PROCESSAMENTO DE DADOS

Observando os conjuntos de dados, foi possível perceber inconstância na apresentação dos dados, assim como uma grande quantidade de dados vazios. Devido a esses fatores, considerou-se necessário aplicar alguns procedimentos de pré-processamento para que os dados fossem adequadamente tratados. Também foi identificada a necessidade de mesclar os diferentes conjuntos de dados para possibilitar seu uso nos algoritmos de classificação.

#### 3.3.1 Formatação

Foi observado que muitos dados textuais similares tinham diferenças de caixa, acentuação e pontuação, dificultando a interpretação dos dados pelos algoritmos de classificação. Portanto, todos os dados textuais foram convertidos para caixa alta, assim como normalizados para o padrão unicode NFKD (*Normalization Form Kompatible Decomposition*), e então decodificado para UTF-8 (*UCS Transformation Format 8*), removendo assim caracteres especiais e acentuação. Também foram removidos pontuação e espaçamento extra, no começo, meio ou fim dos textos. Deste modo, os dados se tornaram mais uniformes e interpretáveis.

#### 3.3.2 Filtragem

##### 3.3.2.1 *Colunas dos Casos Leves e Graves*

Os conjuntos de dados de Casos Leves e Casos Graves contavam com uma quantidade considerável de dados vazios, sendo lidados de acordo com a relevância da coluna. Uma grande quantidade de células possuía textos como "IGNORADO", "IGN", "NENHUMA" ou "0", que foram então convertidos para dados nulos.

As colunas **Raça**, **Etnia**, **Bairro**, **Município**, **Distrito Sanitário**, **Área de Atuação Profissional** e **Militar** tiveram entre 10% e 90% de dados nulos cada uma, sendo decidido remover essas colunas do conjunto de dados final. As colunas **Evolução do Caso**, **Tratamento**

**Domiciliar** e **Classificação Final** não foram consideradas relevantes por sua inconsistência e quantidade de dados nulos, 6,1%, 16,6% e 95% respectivamente, portanto também foram removidas do conjunto de dados final. A coluna **Data de Início de Sintomas** possuía cerca de 2,5% de dados nulos, enquanto a **Data de Notificação** estava sempre presente, portanto, foi preferida como data dos casos. Porém, colunas como **Idade** e **Sexo**, dados demográficos básicos, tiveram somente cerca de 0,08% de dados nulos, decidindo-se remover as linhas com dados nulos. O conjunto de dados de **Casos Graves** possuía duas colunas a mais que o de **Casos Leves**, sendo elas **Outros Sintomas** e **Outras Doenças Preexistentes**, que foram mesclados com suas respectivas colunas, **Sintomas** e **Doenças Preexistentes**.

Restando assim, nos conjuntos de dados de **Casos Leves** e **Graves**, as colunas de **Data de Notificação**, **Sexo**, **Idade**, **Sintomas**, **Doenças Preexistentes**, **Data de Óbito** e **Categoria de Profissional de Saúde**.

### 3.3.2.2 *Colunas da Vacinação*

Devido à natureza da vacina, foi decidido ignorar a coluna de **Primeira Dose**, pois a mesma não é considerada eficaz tendo somente uma dosagem parcial. Considerando isto, a contagem de **Segundas Doses** e **Doses Únicas** então serviram como indicador de população devidamente vacinada. Devido à redundância, as colunas **Dose de Reforço** e **Total de Doses no Dia** foram removidas.

### 3.3.3 **Interpretação**

Tendo os conjuntos de dados formatados e filtrados, foi possível interpretar os dados de forma a trazer valor para a análise.

A coluna de **Severidade** então foi criada de acordo com a classificação da Secretaria de Saúde, para indicar a gravidade do caso. O conjunto de dados de **Casos Leves** resultou em severidade LEVE, enquanto os casos no conjunto de dados de **Casos Graves** resultou em severidade GRAVE. Além disso, de acordo com sua presença, se observou possível transformar a coluna de **Data de Óbito** em uma classificação de **Severidade: ÓBITO**.

A **Categoria de Profissional de Saúde** foi simplificada, sendo Falsa caso nula, e Verdadeira caso contrário, indicando se o paciente era um profissional de saúde. **Sintomas** e **Doenças Preexistentes** estavam em extenso, com separadores não uniforme, muitos erros

de digitação e nomenclaturas diferentes, portanto necessitaram esforço específico, descrito a seguir.

### 3.3.3.1 Sintomas

A coluna **Sintomas** possuía separadores diversos, como ",", "E", "+" e "/", que foram usados para separar os sintomas em listas de texto. O texto foi limpo por espaços e pontuação, tendo então os valores únicos contados e ordenados por frequência de ocorrência. Com base nessa contagem, foram percebidas as seguintes categorias de sintomas relevantes na tabela abaixo.

Categoria	Qtd. / 604.315	Descrição
Dor de Cabeça	189425	Cefaléia.
Febre	177898	Febre.
Dor de Garganta	166967	Odinofagia.
Coriza	131478	Rinorreia e secreção.
Anosmia ou Hiposmia	84599	Perda de olfato ou paladar.
Dispneia	81545	Falta de ar e insuficiência respiratória.
Dor no Corpo	49507	Mialgia e dores musculares.
Tosse	33500	Tosse e hemoptise.
Diarreia	30183	Diarreia.
Astenia	25489	Fraqueza e cansaço.
Baixa Saturação de $O_2$	14556	Saturação de Oxigênio <95%.
Náusea	11129	Enjoo e vômitos.
Desconforto Respiratório	10235	Desconfortos ao respirar.
Aperto Torácico	9452	Aperto no peito e dor torácica.
Congestão Nasal	6468	Obstrução nasal.
Espirros	6435	Espirros e sintomas gripais.
Dor Abdominal	1613	Dor epigástrica.
Inapetência	917	Falta de apetite.
Rebaixamento de Consciência	880	Prostraçao e confusão.

Tabela 7 – Categorias de sintomas

O maior número possível de variações de nomenclatura, forma de escrita e erros de digitação foram considerados para cada categoria, abrangendo-se ao máximo os sintomas mais comuns. Os outros sintomas com menor frequência e entradas que não eram sintomas totalizaram 174.143 ocorrências nos 604.315 casos totais. Cada categoria se tornou então uma coluna no

conjunto de dados final com valor Verdadeiro caso estivesse presente nos sintomas do paciente, e Falso caso contrário.

### 3.3.3.2 *Doenças Preexistentes*

A coluna **Doenças Preexistentes** passou pelo mesmo processo dos sintomas, com ",", "E", "+", "/" e ";" como separadores, que foram usados para formar listas de textos e valores únicos. As seguintes categorias de doenças foram percebidas na tabela a seguir.

Categoria	Qtd. / 604.315	Descrição
Doenças Cardiovasculares	34388	Doenças cardíacas ou vasculares.
Diabetes	20335	Diabetes mellitus.
Doenças Respiratórias Crônicas	12748	Asma, tuberculose, etc.
Obesidade	5131	Obesidade ou sobrepeso.
Imunossupressão	3824	Imunodepressão ou deficiência.
Doenças renais	2365	Doenças renais crônicas.
Hipertensão.	2080	Hipertensão arterial.
Tabagismo	1097	Fumante ou ex-fumante.
Doenças Neurológicas	713	Alzheimer, esquizofrenia, etc.
Etilismo	344	Alcoolatra ou ex-alcoolatra.
Doenças Hepáticas	202	Doenças hepáticas crônicas.

Tabela 8 – Categorias de doenças

O maior número possível de variações foi considerado para cada categoria, deixando um total de 15.709 doenças não categorizadas. Similar aos sintomas, cada categoria se tornou então uma coluna lógica no conjunto de dados final.

### 3.3.3.3 *Interpretando a Vacinação*

De modo a identificar possíveis impactos da vacinação em andamento nos fatores de risco e efeitos da doença, foi criada uma nova coluna no conjunto de dados final: **População Vacinada**.

Este valor se deu por um cálculo utilizando os dados da **Data de Notificação** e o conjunto de dados de Vacinação, considerando a população estimada do Recife em 2021 como 1.661.017 pessoas [7]. Sendo assim, foi encontrado um valor aproximado da porcentagem da população vacinada para cada dia, relacionando-o a cada caso como um valor de **População Vacinada**,

que representa o progresso da vacinação no dia do caso.

### 3.3.4 Categorização

De modo a facilitar o processo dos algoritmos de classificação, certos dados extensos foram categorizados em menor número.

A idade foi agrupada em 9 grupos, com espaçamento de 10 anos entre elas, sendo 0 todas as idades abaixo de 10 anos, e 8 todas aquelas idades acima de 80 anos. Esse valor categorizado tomou o lugar do valor extenso na coluna de Idade.

<b>Idades</b>	<b>Qtd. / 604.315</b>
30-39	133961
40-49	115602
20-29	109439
50-59	89035
60-69	56755
10-19	36626
70-79	26312
0-9	23666
80+	12919

Tabela 9 – Contagem das categorias de idade

De maneira similar, o valor extenso da coluna de **População Vacinada** foi categorizado em 6 grupos, com espaçamento de 15% entre cada um. Foi utilizada a técnica de *One-Hot Encoding* para transformar os valores em colunas lógicas em formato de termômetro [44].

<b>Vacinação</b>	<b>Qtd. / 604.315</b>
>0%	272144
>15%	260258
>30%	188579
>45%	112003
>60%	36072
>75%	0

Tabela 10 – Contagem das categorias de vacinação

### 3.3.5 Normalização

Por fim, de forma que possam ser melhores utilizados como valores de entrada nos algoritmos de classificação, todos os valores foram normalizados em uma escala de 0 a 1. Os dados lógicos tendo Verdadeiro como 1 e Falso como 0 e os dados numéricos tendo 0 como seu valor mínimo e 1 como seu valor máximo. **Severidade**, a única coluna textual restante, foi transformada em números: 0 para LEVE, 1 para GRAVE e 2 para ÓBITO, quando presente.

### 3.3.6 Conjunto de Dados Final

O conjunto de dados resultante do pré-processamento é composto de 40 colunas, sendo **Severidade**, **Sexo**, **Idade** e **Profissional de Saúde**, seguidas de 20 colunas de sintomas, 11 colunas de doenças preexistentes e 5 colunas de vacinação. Esse conjunto de dados foi salvo em um arquivo CSV, que pode ser utilizado como base para os algoritmos de classificação.

## 3.4 EXPERIMENTO

Após a obtenção e pré-processamento dos dados, foi realizado o experimento, dividido em seis etapas interligadas, discutidas detalhadamente a seguir.

O conjunto de dados é dividido em treinamento e teste de forma semi-aleatória, onde a frequência de cada classe é respeitada, separando-as proporcionalmente de acordo com a **Severidade**. Para isso são usados algoritmos de divisão de conjunto de dados do *ScikitLearn* com estratégias de estratificação [39]. O resultado deste processo são dois conjuntos de dados separados, com 80% dos dados no conjunto de treinamento e os 20% restantes no conjunto de teste.

### 3.4.1 Etapa de Treinamento

Na etapa de treinamento, os algoritmos de classificação são treinados utilizando o conjunto de treinamento e parâmetros de configuração. Estes parâmetros devem ser ajustados apropriadamente na etapa de otimização, explicada na subseção 3.4.4, sendo repetidas estas etapas e as seguintes conforme necessário. O resultado desta etapa são modelos especializados no problema de classificação apresentado pelo conjunto de dados.

### 3.4.2 Etapa de Teste

A etapa de teste é onde os modelos gerados pelos algoritmos utilizam o conjunto de testes para tentar prever corretamente a classificação de cada caso. O conjunto de estes são os casos restantes após a separação do conjunto de treinamento. A entrada desta etapa é o conjunto de testes e o modelo gerado na etapa de treinamento, enquanto sua saída é uma lista de previsões, corretas ou não.

### 3.4.3 Cálculo de Métricas de Desempenho

A etapa de cálculo de métricas de desempenho recebe a lista de previsões dos algoritmos e os valores reais do conjunto de dados, que são usados para gerar diversas métricas. Cada uma das métricas é discutida na seção 2.3

De modo a acomodar as diferenças entre cada execução, as etapas de treinamento e teste são executadas 20 vezes para cada coleta de métricas, com conjuntos de treinamento e teste aleatórios em cada execução. O máximo, a média e o desvio padrão das métricas são calculados para cada algoritmo, e servem como referência na otimização e resultados do projeto. Este método é conhecido como *k-fold cross-validation*, e permite avaliar os algoritmos de classificação de maneira mais consistente [45].

### 3.4.4 Otimização de Parâmetros

A etapa de otimização de parâmetros consiste em ajustar os parâmetros de configuração dos algoritmos de classificação para que sejam o mais adequados o possível aos dados de treinamento. Para avaliar isso, são usadas as métricas obtidas na etapa anterior. De acordo com a necessidade, as etapas anteriores são executadas múltiplas vezes de modo a obter o melhor resultado.

Para facilitar esse processo, foi utilizada a técnica de otimização de parâmetros (*grid search*) do *ScikitLearn* [39]. O *grid search* tenta encontrar os melhores parâmetros de forma automatizada, executando uma busca exaustiva com diversos parâmetros de cada algoritmo [46].

Notou-se que o conjunto de dados estava desbalanceado, ou seja, a quantidade de casos leves era 20 vezes maior que a quantidade de casos graves. Portanto, se viu necessário aplicar a

técnica de balanceamento de conjunto de dados *RandomUnderSampler*. Os registros da classe mais comum são removidos aleatoriamente para balancear as classes. Então a quantidade casos leves e torna igual à de casos graves no conjunto de dados balanceado.

### 3.4.5 Geração de Gráficos

Tendo um modelo do algoritmo de classificação satisfatório, foram gerados então diversos gráficos interpretáveis para a análise dos dados. Em especial os gráficos SHAP para entender a interpretação do modelo sobre o conjunto de dados na importância de cada variável [40].

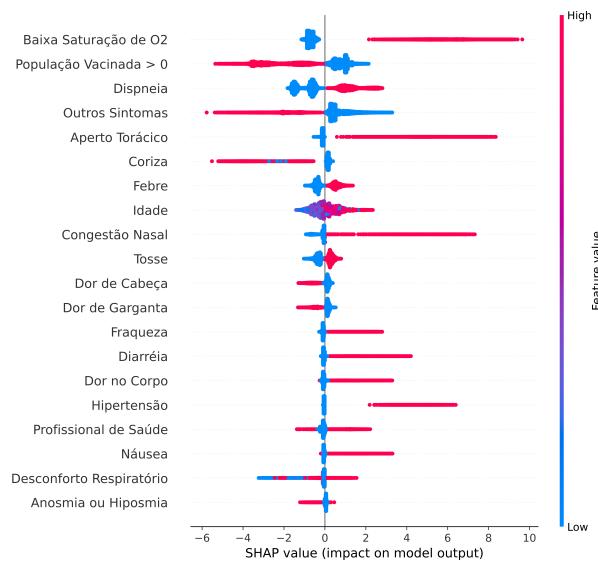


Figura 6 – Gráfico SHAP de interpretação de relevância de atributos.

Fonte: *SHAP documentation*

Gráficos SHAP de pontos como na figura 6 de exemplo combinam gráficos de dispersão com estimativas de densidade. A cor dos pontos significa o valor do atributo, onde pontos azuis significam que o valor do atributo é baixo ou falso, enquanto pontos vermelhos significam que o valor é alto ou verdadeiro. A posição horizontal dos pontos representa o impacto do atributo no resultado da classificação, onde pontos na esquerda significam que aquele valor é importante na classificação negativa, e pontos na direita na classificação positiva.

Também é possível gerar gráficos de análise do conjunto de dados, como gráficos de Mapas de Calor de Correlação [47] e Plotagens de Densidade das Colunas [48], que também podem servir como fonte de informação para a análise dos dados.

### 3.4.6 Subconjuntos

Foi observado que o conjunto de dados permite diferentes cenários por meio de subconjuntos, podendo oferecer *insights* sobre os resultados, fatores de risco e impactos das variáveis. Portanto, as seguintes análises são propostas.

#### 3.4.6.1 *Omitir dados sintomáticos*

O conjunto de dados utilizado no experimento possui 20 colunas de dados sintomáticos apresentados pelos pacientes após confirmado o diagnóstico de COVID-19. Ao omitir essas colunas, é possível observar como os algoritmos predizem os casos leves e graves se baseando somente nos dados demográficos, doenças preexistentes e progresso da vacinação. Idealmente, será possível perceber fatores de risco demográficos e clínicos antes mesmo do paciente contrair COVID-19, permitindo uma melhor priorização de tratamento.

#### 3.4.6.2 *Possibilidade de óbito*

Cerca de 1/4 dos casos graves resultaram em óbito do paciente, como demonstrado na coluna **Data de Óbito**. Por meio desta, é possível comparar casos onde não houve óbito com casos que resultaram em óbito, em um conjunto de dados binário. Sendo assim, é possível identificar especificamente fatores de risco que podem ocasionar no óbito do paciente.

#### 3.4.6.3 *Progresso da vacinação*

Filtrando o conjunto de dados pelas colunas de vacinação, é possível identificar mudanças nos fatores de risco, de modo a perceber os impactos da vacinação no Recife de acordo com seu progresso. Se propõe então observar a situação antes do início da vacinação, assim como depois de um certo progresso da vacinação, para identificar possíveis impactos.

## Capítulo 4

# RESULTADOS

Neste capítulo são apresentados e discutidos os resultados obtidos na realização do experimento descrito em detalhes no Capítulo 3. Será feita uma análise preliminar do conjunto de dados na seção 4.1. Então, na seção 4.2, as métricas obtidas na execução dos algoritmos descritos anteriormente serão utilizados para avaliar a qualidade dos resultados. Seguido então pela seção 4.3, que analisará métricas obtidas na aplicação dos algoritmos nos subconjuntos de dados discutidos na subseção 3.4.6. Gráficos *SHAP* gerados na execução dos algoritmos serão analisados na seção 4.4, de forma a obter percepções sobre os resultados.

## 4.1 ANÁLISE DO CONJUNTO DE DADOS

Antes de aplicar os algoritmos descritos anteriormente, é possível obter algumas informações do conjunto de dados. Existem 604.315 casos registrados, onde 94,91% são casos leves e os restantes são considerados graves, destes, 25,15% resultaram em óbito. Destes casos, embora o total seja de 43,5% para o sexo masculino, eles consistem em 49,3% dos casos graves. O sintoma mais comum foi a tosse, presente em 38% de todos os casos e 72% dos casos graves. Mapas de calor de correlação estão disponíveis nos apêndices A, B e C. As tabelas abaixo descrevem a distribuição da severidade dos casos e óbitos, de acordo com o Progresso de Vacinação e Idade.

Vacinação	Total	%	Leves	%	Graves	%	Óbitos	%
<b>=0%</b>	272144	45.64%	251882	91.03%	15689	5.67%	4573	1.65%
<b>&gt;0%</b>	332171	54.97%	321697	95.93%	7315	2.18%	3159	0.94%
<b>&gt;15%</b>	260258	43.07%	252035	95.93%	5749	2.19%	2474	0.94%
<b>&gt;30%</b>	188579	31.21%	182667	95.95%	4118	2.16%	1794	0.94%
<b>&gt;45%</b>	112003	18.53%	108533	95.97%	2381	2.11%	1089	0.96%
<b>&gt;60%</b>	36072	5.97%	34891	95.76%	816	2.24%	365	1.00%

Tabela 11 – Quantidade de casos leves, graves e óbitos por progresso de vacinação

<b>Idade</b>	<b>Leves</b>	<b>%</b>	<b>Graves</b>	<b>%</b>	<b>Óbitos</b>	<b>%</b>
<b>0-9</b>	21260	89.68%	2365	9.98%	41	0.17%
<b>10-19</b>	36142	98.63%	465	1.27%	19	0.05%
<b>20-29</b>	107641	98.27%	1702	1.55%	96	0.09%
<b>30-39</b>	130232	97.02%	3457	2.58%	272	0.20%
<b>40-49</b>	111150	95.68%	3882	3.34%	570	0.49%
<b>50-59</b>	83865	92.95%	3979	4.41%	1191	1.32%
<b>60-69</b>	51814	88.49%	3145	5.37%	1796	3.07%
<b>70-79</b>	22328	79.54%	2224	7.92%	1760	6.27%
<b>80+</b>	9147	61.25%	1758	11.77%	2014	13.49%

Tabela 12 – Quantidade de casos leves, graves e óbitos por idade

As figuras 7 e 8 comparam a distribuição da porcentagem de casos graves e óbitos, respectivamente, pelas categorias de idade e progresso da vacinação.

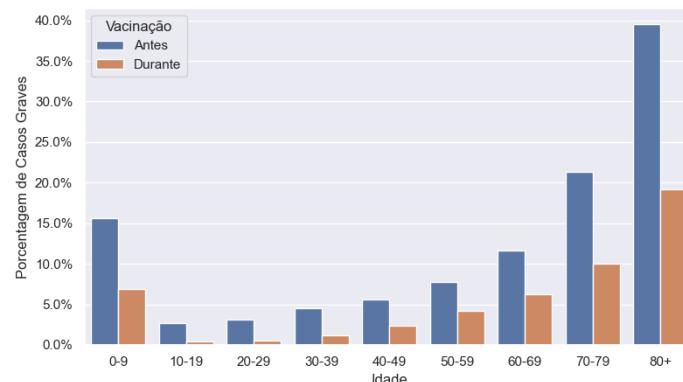


Figura 7 – Gráfico da porcentagem de casos graves para idades e vacinação

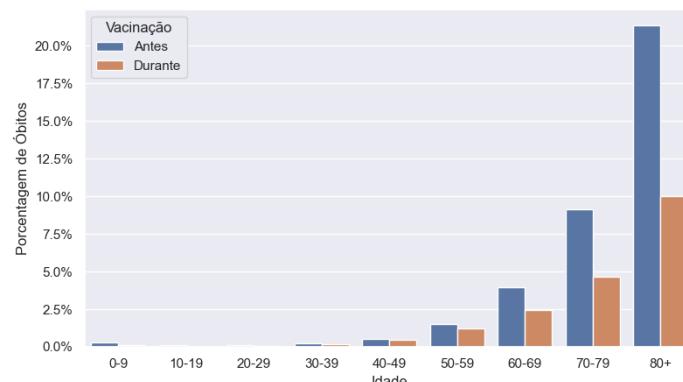


Figura 8 – Gráfico da porcentagem de óbitos para idades e vacinação

## 4.2 COMPARAÇÃO ENTRE ALGORITMOS

Após uma certa quantidade de iterações do processo de otimização de parâmetros como especificado na subseção 3.4.4, os modelos gerados pelos algoritmos de classificação *k-Nearest Neighbors* (kNN), *Decision Tree* (DT), *Random Forest* (RF), *Gradient Boosting* (GB), *Light Gradient Boosting* (LGB) e *XGBoost* (XGB) na etapa de treinamento foram aplicados no conjunto de dados de teste.

Cada algoritmo foi executado 20 vezes, aleatorizando o conjunto de dados em cada uma delas, coletando então suas métricas de acurácia, precisão, sensibilidade, *F1-Score* e *AUC-ROC* para cada execução. Como discutido na subseção 3.4.3, as métricas macro representam a média dos resultados de cada classe, que neste conjunto de dados são casos leves e casos graves. A média e desvio padrão de cada métrica foi calculada, e as métricas consideradas relevantes para comparação se encontram na tabela 13, onde o maior valor para cada uma está representado em negrito.

	<b>Acurácia</b>	<b>Precisão Macro</b>	<b>Sensib. Macro</b>	<b>F1-Score Macro</b>	<b>AUC-ROC</b>
<b>kNN</b>	0,9790±0,0001	0,9475±0,0007	0,8228±0,0020	0,8739±0,0011	0,8228±0,0020
<b>DT</b>	0,9767±0,0015	0,9558±0,0133	0,7913±0,0094	0,8534±0,0101	0,7913±0,0094
<b>RF</b>	0,9828±0,0001	<b>0,9795±0,0006</b>	0,8395±0,0012	0,8962±0,0009	0,8395±0,0012
<b>GB</b>	0,9835±0,0002	0,9727±0,0022	0,8513±0,0015	0,9020±0,0012	0,8513±0,0015
<b>LGB</b>	0,9835±0,0002	0,9748±0,0016	0,8498±0,0015	0,9018±0,0001	0,8498±0,0015
<b>XGB</b>	<b>0,9838±0,0003</b>	0,9707±0,0010	<b>0,8571±0,0031</b>	<b>0,9052±0,0022</b>	<b>0,8571±0,0031</b>

Tabela 13 – Médias de métricas de algoritmos de classificação na etapa de teste

Levando em consideração a métrica *AUC-ROC* como métrica de avaliação, é possível observar que, enquanto todos os algoritmos alcançaram um valor próximo ou acima de 80%, o algoritmo *XGBoost* obteve o melhor desempenho, bem próximo dos outros algoritmos de *Gradient Boosting*. O *F1-Score Macro* também é um indicador de desempenho satisfatório, e todos os algoritmos alcançaram valores acima de 85%, com os algoritmos baseados em *Gradient Boosting* alcançando valores acima de 90%.

Embora a Precisão Macro tenha alcançado valores altos, até acima de 97%, a Sensibilidade Macro ficou sempre abaixo dos 90%, devido a uma quantidade considerável de falsos negativos, casos graves classificados como leves. É possível observar este resultado na tabela 14, que mostra a matriz de confusão do *XGBoost* com melhor *F1-Score* das 20 execuções, 90,97%.

	Predição: LEVE	Predição: GRAVE
Real: LEVE	114511	205 (0,18%)
Real: GRAVE	1664 (27,07%)	4483

Tabela 14 – Matriz de confusão de uma execução do *XGBoost*

Tomando a tabela 14 como exemplo, somente 0,18% de casos leves foram preditos como graves, um número ínfimo de falsos negativos. Isso significa que existe uma confiabilidade satisfatória nos casos preditos como graves.

Contudo, 27% dos casos graves foram erroneamente preditos como leves, significando que cerca de 1/4 dos casos graves não foram percebidos pelo algoritmo. Este fenômeno se mostrou presente em todos os algoritmos analisados, sendo a causa da diminuição da Sensibilidade Macro observada na tabela 13.

Este problema ocorre devido ao desbalanceamento do conjunto de dados, onde existem 20x mais casos leves que graves. Portanto, durante a etapa de otimização de parâmetros descrita na subseção 3.4.4, técnicas de balanceamento de dados foram utilizadas. O processo de balanceamento consiste em igualar a quantidade de classes no conjunto de dados, excluindo registros aleatórios da classe mais comum. Os resultados se encontram na tabela 15.

	Acurácia	Precisão Macro	Sensib. Macro	F1-Score Macro	AUC-ROC
<b>kNN</b>	0,9047±0,0027	0,9050±0,0026	0,9047±0,0027	0,9047±0,0027	0,9047±0,0027
<b>DT</b>	0,8896±0,0102	0,8905±0,0096	0,8896±0,0102	0,8895±0,0103	0,8896±0,0102
<b>RF</b>	0,9150±0,0019	0,9158±0,0020	0,9150±0,0019	0,9150±0,0019	0,9150±0,0019
<b>GB</b>	0,9234±0,0022	0,9238±0,0022	0,9234±0,0022	0,9234±0,0022	0,9234±0,0022
<b>LGB</b>	0,9233±0,0009	0,9239±0,0009	0,9233±0,0009	0,9233±0,0009	0,9233±0,0009
<b>XGB</b>	<b>0,9243±0,0016</b>	<b>0,9249±0,0013</b>	<b>0,9243±0,0016</b>	<b>0,9242±0,0016</b>	<b>0,9243±0,0016</b>

Tabela 15 – Médias de métricas de algoritmos de classificação na etapa de testes com classes balanceadas

A Sensibilidade Macro teve um aumento considerável, chegando acima dos 90% na maioria dos algoritmos, enquanto a Precisão Macro sofreu uma queda. O *F1-Score* consequentemente teve um aumento, por ser a média harmônica dessas medidas. A acurácia caiu significativamente, não mais inflada pelos casos leves, enquanto o valor *AUC-ROC* cresceu proporcionalmente. Demonstram-se essas diferenças mais detalhadamente na tabela 16, uma matriz de confusão do XGBoost com *F1-Score* de 92,67%.

	Predição: LEVE	Predição: GRAVE
Real: LEVE	5770	377 (6,13%)
Real: GRAVE	524 (8,52%)	5623

Tabela 16 – Matriz de confusão de uma execução do *XGBoost* com classes balanceadas

Neste cenário, a diferença entre a proporção de falsos positivos e falsos negativos se equilibra, ainda que se mantenha maior nos casos graves. Existe uma chance de que casos leves sejam preditos como graves, mas a chance de que casos graves sejam preditos como leves é muito menor que anteriormente. Considerando a natureza médica do problema, uma menor taxa de casos graves perdidos pode ser considerada uma melhoria significativa, mesmo que exista um aumento na quantidade de alarmes falsos para casos leves [49]. Portanto, assimilando este julgamento ao aumento do valor *AUC-ROC*, escolhido como métrica de desempenho, balancear o conjunto de dados por *RandomUnderSampler* se evidencia como uma melhoria.

### 4.3 COMPARAÇÃO ENTRE ALGORITMOS EM SUBCONJUNTOS

Como discutido na subseção 3.4.6, a aplicação de certos filtros no conjunto de dados permite observar diferentes cenários, que podem trazer novas percepções ao problema. Sendo assim, os algoritmos foram utilizados em subconjuntos de dados, para obter métricas que podem ser comparadas entre eles. As métricas foram obtidas usando os mesmos métodos que as métricas da seção 4.2, tendo o conjunto de dados balanceado por *RandomUnderSampler*. Gráficos *SHAP* gerados na execução dos algoritmos serão analisados e discutidos na seção 4.4.

#### 4.3.1 Omitindo Dados Sintomáticos

Remover as colunas de dados sintomáticos permite observar como os algoritmos predizem os casos leves e graves se baseando somente nos dados demográficos, doenças preexistentes e progresso da vacinação.

	<b>Acurácia</b>	<b>Precisão Macro</b>	<b>Sensib. Macro</b>	<b>F1-Score Macro</b>	<b>AUC-ROC</b>
<b>kNN</b>	$0,7125 \pm 0,0162$	$0,7144 \pm 0,0172$	$0,7125 \pm 0,0162$	$0,7119 \pm 0,0159$	$0,7125 \pm 0,0162$
<b>DT</b>	$0,7383 \pm 0,0051$	$0,7415 \pm 0,0036$	$0,7383 \pm 0,0051$	$0,7375 \pm 0,0056$	$0,7383 \pm 0,0051$
<b>RF</b>	$0,7383 \pm 0,0025$	$0,7402 \pm 0,0029$	$0,7383 \pm 0,0025$	$0,7378 \pm 0,0024$	$0,7383 \pm 0,0025$
<b>GB</b>	<b><math>0,7463 \pm 0,0044</math></b>	<b><math>0,7481 \pm 0,0042</math></b>	<b><math>0,7463 \pm 0,0044</math></b>	<b><math>0,7459 \pm 0,0045</math></b>	<b><math>0,7463 \pm 0,0044</math></b>
<b>LGB</b>	$0,7453 \pm 0,0019$	$0,7470 \pm 0,0019$	$0,7453 \pm 0,0019$	$0,7448 \pm 0,0019$	$0,7453 \pm 0,0019$
<b>XGB</b>	$0,7451 \pm 0,0047$	$0,7470 \pm 0,0045$	$0,7451 \pm 0,0047$	$0,7446 \pm 0,0047$	$0,7451 \pm 0,0047$

Tabela 17 – Comparação de algoritmos sem dados sintomáticos

A tabela 17 apresenta as métricas da aplicação dos algoritmos nos subconjuntos de dados sem dados sintomáticos. É possível perceber uma queda em todas as métricas, consistente

entre os algoritmos, devido à importância dos dados sintomáticos no julgamento. Todos os valores do *Gradient Boosting* se mantêm em torno dos 75%, incluindo o *AUC-ROC*, utilizado como métrica de desempenho, com uma queda média de 18% comparado à sua execução no conjunto de dados completo. Com isso, embora menos precisos, é possível considerar válida a aplicação dos algoritmos nos subconjuntos de dados sem dados sintomáticos.

#### 4.3.2 Prevendo Óbitos

Separando o conjunto de dados entre casos leves ou graves e óbitos, é possível observar como os algoritmos predizem óbito do paciente de acordo com seus dados demográficos, sintomas, doenças preexistentes e progresso da vacinação.

	<b>Acurácia</b>	<b>Precisão Macro</b>	<b>Sensib. Macro</b>	<b>F1-Score Macro</b>	<b>AUC-ROC</b>
<b>kNN</b>	$0,9354 \pm 0,0060$	$0,9356 \pm 0,0059$	$0,9354 \pm 0,0060$	$0,9354 \pm 0,0060$	$0,9354 \pm 0,0060$
<b>DT</b>	$0,9228 \pm 0,0069$	$0,9231 \pm 0,0067$	$0,9228 \pm 0,0069$	$0,9227 \pm 0,0069$	$0,9228 \pm 0,0069$
<b>RF</b>	$0,9435 \pm 0,0044$				
<b>GB</b>	$0,9499 \pm 0,0038$				
<b>LGB</b>	$0,9465 \pm 0,0027$				
<b>XGB</b>	<b><math>0,9547 \pm 0,0057</math></b>				

Tabela 18 – Comparação de algoritmos prevendo óbito

A tabela 18 apresenta as métricas da aplicação dos algoritmos nos subconjuntos de dados de óbito. Neste caso, todos os algoritmos alcançaram métricas acima de 92%, sendo o *XGBoost* o único que alcançou uma média acima de 95%, mais preciso que no conjunto de dados completo. Acredita-se que isso ocorre devido a melhor separação das classes.

Similar ao discutido na subseção 4.3.1, é possível observar como os algoritmos predizem óbito do paciente sem os sintomas, usando seus dados demográficos, doenças preexistentes e progresso da vacinação.

	<b>Acurácia</b>	<b>Precisão Macro</b>	<b>Sensib. Macro</b>	<b>F1-Score Macro</b>	<b>AUC-ROC</b>
<b>kNN</b>	$0,8450 \pm 0,0069$	$0,8460 \pm 0,0066$	$0,8450 \pm 0,0069$	$0,8449 \pm 0,0070$	$0,8450 \pm 0,0069$
<b>DT</b>	$0,8468 \pm 0,0094$	$0,8475 \pm 0,0091$	$0,8468 \pm 0,0094$	$0,8467 \pm 0,0095$	$0,8468 \pm 0,0094$
<b>RF</b>	$0,8473 \pm 0,0036$	$0,8476 \pm 0,0036$	$0,8473 \pm 0,0036$	$0,8473 \pm 0,0036$	$0,8473 \pm 0,0036$
<b>GB</b>	$0,8598 \pm 0,0021$	$0,8601 \pm 0,0022$	$0,8598 \pm 0,0021$	$0,8598 \pm 0,0021$	$0,8598 \pm 0,0021$
<b>LGB</b>	$0,8554 \pm 0,0030$	$0,8557 \pm 0,0028$	$0,8554 \pm 0,0030$	$0,8553 \pm 0,0030$	$0,8554 \pm 0,0030$
<b>XGB</b>	<b><math>0,8597 \pm 0,0050</math></b>	<b><math>0,8602 \pm 0,0051</math></b>	<b><math>0,8597 \pm 0,0050</math></b>	<b><math>0,8596 \pm 0,0050</math></b>	<b><math>0,8597 \pm 0,0050</math></b>

Tabela 19 – Comparação de algoritmos prevendo óbito sem dados sintomáticos

A tabela 19 apresenta as métricas da aplicação dos algoritmos nos subconjuntos de dados de óbito com dados sintomáticos omitidos. As métricas se mantêm em torno de 85%, tendo um aumento se comparado à predição de casos leves e graves sem os dados sintomáticos. É possível considerar satisfatório aplicar os algoritmos de classificação na predição de óbito do paciente somente usando seus dados demográficos, doenças preexistentes e progresso da vacinação.

#### 4.3.3 Filtrando por Progresso de Vacinação

Filtrando o conjunto de dados por progresso de vacinação, é possível observar possíveis mudanças no perfil de risco dos pacientes em casos leves e graves. Esta abordagem foi então aplicada no conjunto de dados completo e no subconjunto de dados de óbito.

##### 4.3.3.1 Casos antes da vacina

	<b>Acurácia</b>	<b>Precisão Macro</b>	<b>Sensib. Macro</b>	<b>F1-Score Macro</b>	<b>AUC-ROC</b>
<b>kNN</b>	0.8557±0.0053	0.8571±0.0055	0.8557±0.0053	0.8555±0.0053	0.8557±0.0053
<b>DT</b>	0.8433±0.0099	0.8456±0.0102	0.8433±0.0099	0.8430±0.0099	0.8433±0.0099
<b>RF</b>	0.8716±0.0056	0.8737±0.0056	0.8716±0.0056	0.8714±0.0056	0.8716±0.0056
<b>GB</b>	0.8792±0.0040	0.8815±0.0039	0.8792±0.0040	0.8791±0.0041	0.8792±0.0040
<b>LGB</b>	<b>0.8806±0.0019</b>	<b>0.8832±0.0016</b>	<b>0.8806±0.0019</b>	<b>0.8804±0.0020</b>	<b>0.8806±0.0019</b>
<b>XGB</b>	0.8798±0.0026	0.8817±0.0028	0.8798±0.0026	0.8796±0.0026	0.8798±0.0026

Tabela 20 – Comparação de algoritmos prevendo severidade do caso antes da vacinação

	<b>Acurácia</b>	<b>Precisão Macro</b>	<b>Sensib. Macro</b>	<b>F1-Score Macro</b>	<b>AUC-ROC</b>
<b>kNN</b>	0,9188±0,0059	0,9191±0,0058	0,9188±0,0059	0,9187±0,0059	0,9188±0,0059
<b>DT</b>	0,9015±0,0090	0,9022±0,0091	0,9015±0,0090	0,9015±0,0090	0,9015±0,0090
<b>RF</b>	0,9270±0,0049	0,9273±0,0049	0,9270±0,0049	0,9270±0,0049	0,9270±0,0049
<b>GB</b>	0,9362±0,0058	0,9364±0,0059	0,9362±0,0058	0,9362±0,0058	0,9362±0,0058
<b>LGB</b>	<b>0,9376±0,0039</b>	<b>0,9377±0,0038</b>	<b>0,9376±0,0039</b>	<b>0,9376±0,0039</b>	<b>0,9376±0,0039</b>
<b>XGB</b>	0,9348±0,0050	0,9349±0,0049	0,9348±0,0050	0,9348±0,0050	0,9348±0,0050

Tabela 21 – Comparação de algoritmos prevendo óbitos antes da vacinação

As tabelas 20 e 21 apresentam as métricas da aplicação dos algoritmos nos subconjuntos de dados filtrados por casos que ocorreram antes do início da vacinação, tanto predizendo severidade quanto óbito. Há uma queda significativa em todas as métricas, para ambas as

abordagens. Isto ocorre por razão da importância dos dados de vacinação, ou possível incongruência na documentação dos casos no início da pandemia.

#### 4.3.3.2 Casos com 30% da população vacinada

	<b>Acurácia</b>	<b>Precisão Macro</b>	<b>Sensib. Macro</b>	<b>F1-Score Macro</b>	<b>AUC-ROC</b>
<b>kNN</b>	0.9702±0.0030	0.9704±0.0030	0.9702±0.0030	0.9702±0.0030	0.9702±0.0030
<b>DT</b>	0.9636±0.0042	0.9636±0.0042	0.9636±0.0042	0.9635±0.0042	0.9636±0.0042
<b>RF</b>	0.9781±0.0030	0.9782±0.0030	0.9781±0.0030	0.9781±0.0030	0.9781±0.0030
<b>GB</b>	0.9811±0.0012	0.9811±0.0012	0.9811±0.0012	0.9811±0.0012	0.9811±0.0012
<b>LGB</b>	<b>0.9816±0.0016</b>	<b>0.9816±0.0016</b>	<b>0.9816±0.0016</b>	<b>0.9816±0.0016</b>	<b>0.9816±0.0016</b>
<b>XGB</b>	0.9800±0.0021	0.9801±0.0021	0.9800±0.0021	0.9800±0.0021	0.9800±0.0021

Tabela 22 – Comparação de algoritmos prevendo severidade dos casos após 30% da população vacinada

	<b>Acurácia</b>	<b>Precisão Macro</b>	<b>Sensib. Macro</b>	<b>F1-Score Macro</b>	<b>AUC-ROC</b>
<b>kNN</b>	0,9538±0,0054	0,9543±0,0054	0,9538±0,0054	0,9538±0,0054	0,9538±0,0054
<b>DT</b>	0,9378±0,0083	0,9379±0,0083	0,9378±0,0083	0,9378±0,0083	0,9378±0,0083
<b>RF</b>	0,9666±0,0075	0,9667±0,0076	0,9666±0,0075	0,9666±0,0075	0,9666±0,0075
<b>GB</b>	<b>0,9719±0,0040</b>	<b>0,9721±0,0039</b>	<b>0,9719±0,0040</b>	<b>0,9719±0,0040</b>	<b>0,9719±0,0040</b>
<b>LGB</b>	0,9696±0,0055	0,9697±0,0055	0,9696±0,0055	0,9696±0,0055	0,9696±0,0055
<b>XGB</b>	0,9619±0,0043	0,9620±0,0043	0,9619±0,0043	0,9619±0,0043	0,9619±0,0043

Tabela 23 – Comparação de algoritmos prevendo óbitos após 30% da população vacinada

Da mesma maneira, as tabelas 22 e 23 apresentam as métricas da aplicação dos algoritmos nos subconjuntos de dados filtrados por casos que ocorreram após a vacinação alcançar 30% da população, predizendo tanto a severidade quanto óbito. Desta vez há um aumento em todas as métricas, novamente em ambas abordagens, alcançando até uma média de 98% na predição de severidade. É possível que o aumento do desempenho se dê por efeitos da vacinação ou menor quantidade de casos registrados.

## 4.4 GRÁFICOS SHAP

Como discutido na seção 3.4.5, os gráficos *SHAP* gerados na execução dos algoritmos nas seções 4.2 e 4.3 serão analisados para obter novas percepções sobre os resultados obtidos.

#### 4.4.1 Conjunto de dados balanceado - XGBoost

A figura 9 mostra o gráfico *SHAP* de pontos da melhor execução dos algoritmos descritos na seção 4.2.

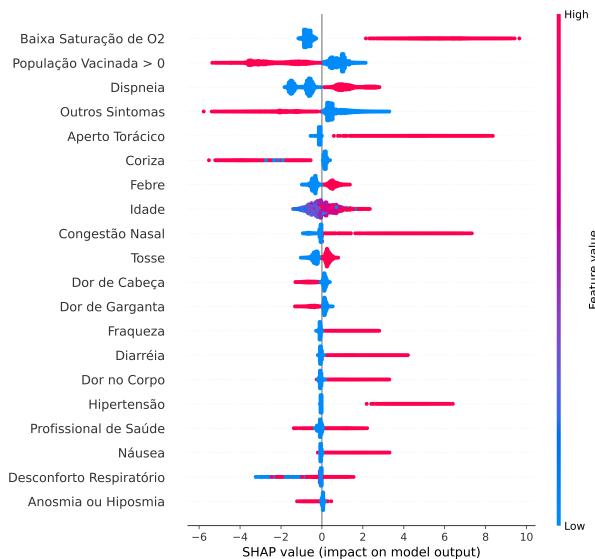


Figura 9 – Gráfico SHAP de interpretação de relevância de atributos na classificação para o modelo XGBoost no conjunto de dados balanceado com AUC-ROC de 92,50%

Baseando-se na interpretação do conjunto de dados pelo modelo, é possível elaborar certas observações. O fator mais decisivo para definir um caso grave foi a Baixa Saturação de  $O_2$ , acredita-se que o sintoma seja um dos critérios escolhidos pela Secretaria de Saúde de Recife para identificar casos graves de COVID-19. A variável de População Vacinada > 0 mostra que quando seu valor é verdadeiro, casos leves são mais comuns, e o contrário para casos graves, demonstrando que foi identificada uma relação entre a vacina e uma diminuição de severidade dos casos. Dispneia, ou falta de ar, se apresenta como terceiro fator relevante para os casos graves. Outros Sintomas, os sintomas não classificados, indicam uma menor chance de caso grave, e é possível interpretar que seja devido a sintomas menores que não foram agrupados. A falta de Aperto Torácico ou Congestão Nasal não é indicativos suficientes de caso leve, mas sua presença influencia significativamente na probabilidade de caso grave. A presença de Coriza, por outro lado, indica maior chance de caso leve, assim como Dor de Cabeça e Dor de Garganta, possivelmente ofuscados por sintomas mais graves durante a documentação de casos graves. Evidenciou-se que uma Idade mais elevada indica maior chance de caso grave, enquanto uma Idade mais baixa indica maior chance de caso leve.

#### 4.4.2 Subconjunto de dados sem colunas de sintomas - Gradient Boosting

Nota-se uma grande influência dos sintomas no julgamento da severidade dos casos, sinalizando menor importância dos fatores demográficos. Portanto, o gráfico *SHAP* de pontos da melhor execução dos algoritmos descritos na subseção 4.3.1 na figura 10 se prova útil para a discussão.

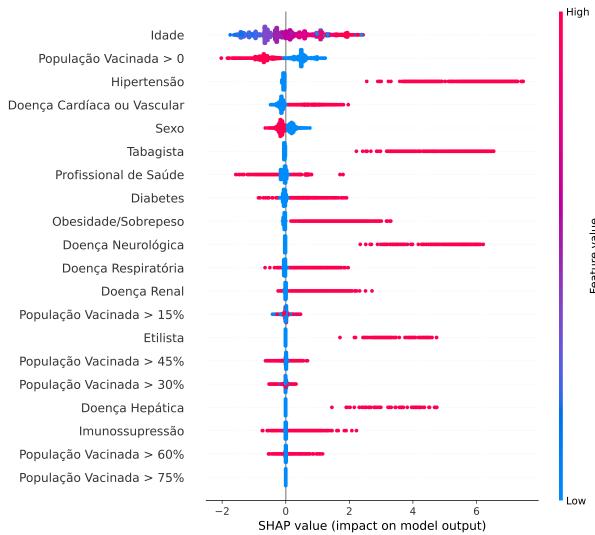


Figura 10 – Gráfico SHAP de interpretação de relevância de atributos na classificação para o modelo Gradient Boosting no subconjunto sem colunas de sintomas com AUC-ROC de 75,11%

É possível então elaborar algumas observações sobre a interpretação do modelo em relação aos dados demográficos e comorbidades, mantendo em mente que seu desempenho não foi excepcional. A Idade dessa vez é a variável mais relevante, mantendo o padrão de que idades mais altas estão em maior risco de caso grave, e idades mais baixas em menor risco. A vacinação estar em progresso também indica menos casos graves. Porém, agora é possível observar comorbidades como a Hipertensão, Doenças Cardiovasculares e Diabetes sendo as mais influentes em casos graves. Sexo se mostra relevante, com homens sendo mais propensos a desenvolverem casos graves, e mulheres a casos leves, embora não tenham sido observadas correlações de sexo com idade ou comorbidades na seção 4.1. Dados relacionados a estilo de vida como Tabagismo, Etilismo e Obesidade também se evidenciaram como influentes na classificação dos casos graves. Outras estatísticas de vacinação se mostraram inconclusivas neste modelo.

#### 4.4.3 Subconjunto de dados de vacinação - Light Gradient Boosting

Filtrando os dados de vacinação, os gráficos *SHAP* de pontos das melhores execuções dos algoritmos descritos na subseção 4.3.3 na figura 11 podem ser analisados.

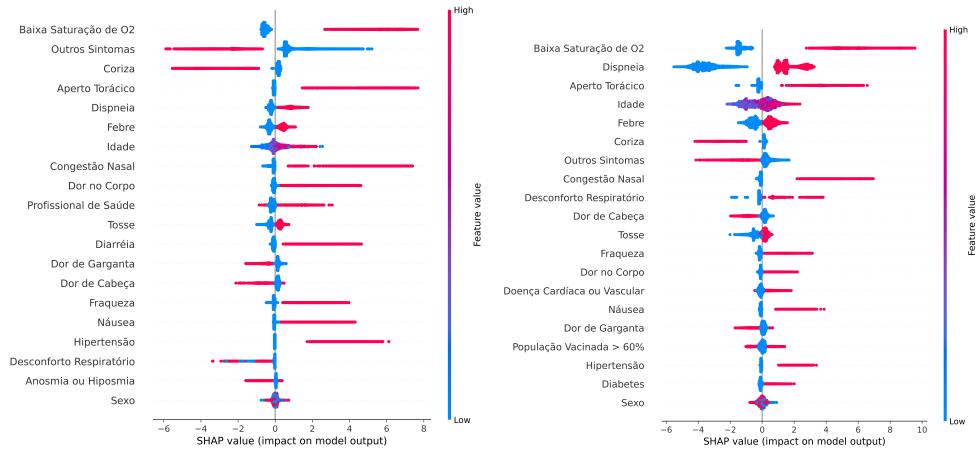


Figura 11 – Gráfico SHAP de interpretação de relevância de atributos na classificação para o modelo Light Gradient Boosting no subconjunto sem vacinação com AUC-ROC de 88,29% e vacinação acima de 30% com AUC-ROC de 98,33%

Ambos os gráficos *SHAP* de pontos são similares à figura 9, com pequenas diferenças em dimensões opostas. Além das diferenças na ordenação das relevâncias, é possível observar mais casos graves em geral no cenário antes da vacinação, e uma maior separação na influência de casos graves e casos leves no cenário da vacinação acima de 30%. A Idade se torna um fator mais consistente ao decorrer da vacinação, sendo observada também uma diminuição na idade de risco para casos graves.

Vacinação	0%	>0%	>30% e <30%
<b>Idade</b>	2.9%	2.6%	9.2%
<b>Sexo</b>	0.0%	0.3%	0.5%
<b>Profissional de Saúde</b>	-2.6%	0.0%	0.0%
<b>Baixa Saturação de <math>O_2</math></b>	-30.7%	-38.2%	-43.4%
<b>Aperto Torácico</b>	14.7%	16.4%	28.8%
<b>Dispneia</b>	10.3%	-2.1%	-14.9%
<b>Desconforto Respiratório</b>	-12.0%	6.1%	11.5%
<b>Tosse</b>	1.2%	0.0%	-5.5%
<b>Congestão Nasal</b>	8.5%	4.5%	2.9%
<b>Febre</b>	0.4%	0.0%	-1.7%
<b>Coriza</b>	-4.3%	-0.2%	-1.7%
<b>Dor no Corpo</b>	4.7%	0.1%	1.3%
<b>Náusea</b>	0.0%	0.0%	1.2%
<b>Diarréia</b>	2.6%	0.0%	1.2%
<b>Dor de Garganta</b>	0.5%	-0.3%	0.8%
<b>Dor de Cabeça</b>	4.1%	-0.6%	-0.6%
<b>Anosmia ou Hiposmia</b>	-1.5%	1.8%	0.5%
<b>Perda de Apetite</b>	0.0%	0.0%	0.3%
<b>Dor Abdominal</b>	0.0%	0.5%	0.2%
<b>Fraqueza</b>	-2.1%	0.3%	-0.1%
<b>Rebaixamento de Consciência</b>	0.0%	-0.5%	0.0%
<b>Espirros</b>	0.0%	0.0%	0.0%
<b>Outros Sintomas</b>	1.8%	2.2%	-1.5%
<b>Doença Cardíaca ou Vascular</b>	-0.2%	-0.1%	2.9%
<b>Obesidade/Sobrepeso</b>	-1.7%	0.0%	-2.0%
<b>Diabetes</b>	0.1%	0.1%	1.8%
<b>Doença Renal</b>	0.0%	-0.3%	0.9%
<b>Doença Neurológica</b>	0.0%	0.0%	0.7%
<b>Tabagista</b>	0.1%	0.0%	0.4%
<b>Hipertensão</b>	0.5%	0.2%	0.3%
<b>Etilista</b>	0.1%	0.1%	-0.2%
<b>Doença Respiratória</b>	0.0%	0.0%	0.0%
<b>Doença Hepática</b>	0.0%	0.0%	0.0%
<b>Imunossupressão</b>	0.0%	0.0%	0.0%
<b>Outras Doenças</b>	2.6%	-0.4%	6.3%

Tabela 24 – Valor SHAP médio de cada fator na classificação de casos graves de acordo com o progresso da vacinação em execuções de Light Gradient Boosting

Valores *SHAP* são uma quantificação do que é demonstrado nos gráficos *SHAP*, e mostram a relevância de uma variável para a classificação de classes. O valor *SHAP* máximo é 50%, significando que, independente do valor da variável, um valor *SHAP* alto teve grande influencia na classificação da classe positiva, enquanto valores negativos, até -50%, influenciaram na classe negativa. Portanto, nesta análise, um valor *SHAP* de 50% significa que a variável possibilitava a certeza de classificação como caso grave.

A tabela 24 mostra o valor SHAP médio de cada fator na classificação de casos graves de acordo com o progresso da vacinação. Valores negativos mostram uma importância na classificação de casos leves. Em especial, se observa que a Idade se tornou um fator mais relevante para os casos graves de acordo com a vacinação. Infere-se também que a Baixa

Saturação de  $O_2$  se tornou menos presente em casos leves ao decorrer da vacinação, diminuindo seu valor *SHAP*. Doenças Cardiovasculares, Diabetes e Doenças Renais também tiveram um aumento na sua relevância para os casos graves, mostrando que ao decorrer da vacinação, casos graves dependeram mais destas comorbidades.

Vacinação 0%	0%	Vacinação >0% e <30%	>0% e <30%	Vacinação >30%	>30%
Baixa Saturação de $O_2$	-30.7%	Baixa Saturação de $O_2$	-38.2%	Baixa Saturação de $O_2$	-43.4%
Aperto Torácico	14.7%	Aperto Torácico	16.4%	Aperto Torácico	28.8%
Desconforto Respiratório	-12.0%	Desconforto Respiratório	6.1%	Dispneia	-14.9%
Dispneia	10.3%	Congestão Nasal	4.5%	Desconforto Respiratório	11.5%
Congestão Nasal	8.5%	Idade	2.6%	Idade	9.2%
Dor no Corpo	4.7%	Outros Sintomas	2.2%	Outras Doenças	6.3%
Coriza	-4.3%	Dispneia	-2.1%	Tosse	-5.5%
Dor de Cabeça	4.1%	Anosmia ou Hiposmia	1.8%	Congestão Nasal	2.9%
Idade	2.9%	Dor de Cabeça	-0.6%	Doença Cardíaca ou Vascular	2.9%
Profissional de Saúde	-2.6%	Dor Abdominal	0.5%	Obesidade/Sobrepeso	-2.0%

Tabela 25 – 10 valores *SHAP* mais relevantes para a classificação de casos graves de acordo com a vacinação em execuções de Light Gradient Boost

A tabela 25 mostra os 10 fatores mais relevantes para a classificação de casos graves de acordo com a vacinação, organizados por ordem decrescente de valor *SHAP*. Baixa Saturação de  $O_2$  e Aperto Torácico, assim como a Idade, cresceram em relevância ao decorrer da vacinação. Por outro lado, a relevância de sintomas menores como Congestão Nasal e Dor de Cabeça diminuiu. Desconforto Respiratório deixou de ser relevante em casos leves e passou a ser relevante em casos graves.

#### 4.4.4 Subconjunto de dados de óbitos - XGBoost

Dividindo o conjunto de dados entre casos leves ou graves e óbitos na subseção 4.3.2, é possível analisar os fatores que influenciam na chance de morte conforme o *XGBoost* com melhor resultado, na figura 12.

Assim como na figura 9, a Baixa Saturação de  $O_2$  é o fator mais relevante, embora mais proeminente neste cenário de óbitos. A Idade sobe em relevância e em detalhamento, havendo uma maior separação entre idades em relação à chance de óbito. A influência da vacinação se mantém, porém, um pouco menos relevante. Os efeitos dos sintomas como Dispneia, Coriza e Febre na chance de óbito se tornam mais claros, enquanto comorbidades como Doenças Cardiovasculares, Hipertensão e Diabetes têm um aumento considerável na sua relevância.

Para observar melhor a relevância dos fatores demográficos e comorbidades, similar à subseção 4.4.2, os gráficos *SHAP* do subconjunto de dados sem colunas de sintomas descrito

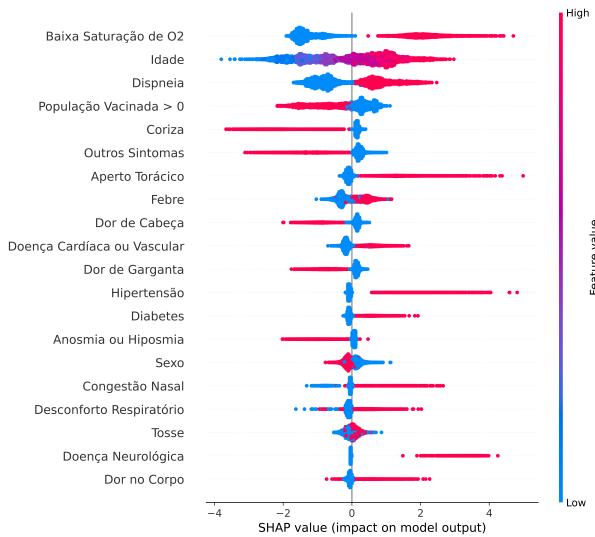


Figura 12 – Gráfico SHAP de interpretação de relevância de atributos na classificação para o modelo XGBoost no subconjunto de óbitos com AUC-ROC de 96,24%

na subseção 4.3.2 pode ser analisado na figura 13.

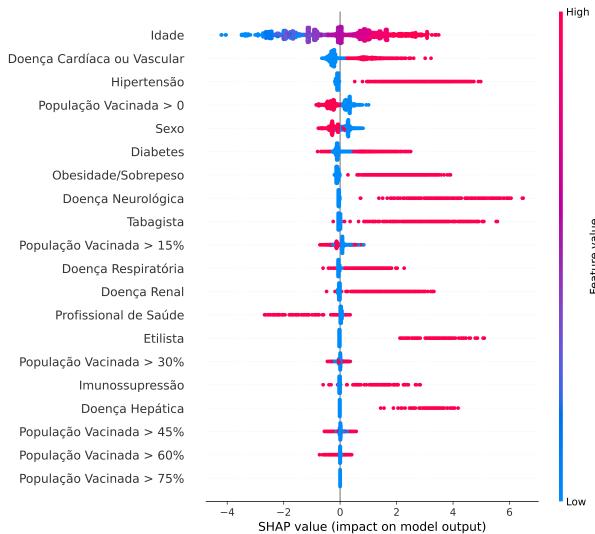


Figura 13 – Gráfico SHAP de interpretação de relevância de atributos na classificação para o modelo XGBoost no subconjunto de óbitos sem colunas de sintomas com AUC-ROC de 86,57%

A Idade novamente fica acima dos outros fatores, sendo o principal fator demográfico a ser considerado. Tendo as comorbidades de Doenças Cardiovasculares, Hipertensão, Diabetes e Doenças Neurológicas como principais influências. Estilo de vida também é um fator relevante, com Obesidade, Tabagismo e Etilismo aumentando as chances de óbito. População Vacinada e Sexo possuem grande relevância, porém menos influência que nos cenários de severidade.

*Capítulo 5*

## CONCLUSÃO

O objetivo principal deste trabalho foi avaliar a eficácia de algoritmos de classificação por aprendizagem de máquina na previsão de casos graves de COVID-19 e possível óbito, utilizando conjuntos de dados disponibilizados pela Prefeitura do Recife. Além disso, a evolução da doença na cidade e os efeitos da vacinação em progresso foram investigados por meio dos modelos gerados. O estudo foi motivado pelo cenário de pandemia e oportunidade de aproveitamento da abundância de dados coletados para estudar possíveis peculiaridades da doença em Recife e os impactos da vacinação.

Os algoritmos foram avaliados no conjunto de dados completo, balanceado por classes e filtrado com certos critérios como progresso de vacinação ou omissão de dados sintomáticos. As médias e desvios das métricas coletadas foram apresentadas em forma de tabela, utilizando *AUC-ROC* como métrica de desempenho, auxiliada por *F1-Score*. Os resultados mostram que todos os algoritmos de classificação experimentados são capazes de classificar os casos da doença e possível óbito. Foram alcançados médias de 92% e 95% de acurácia, *F1-Score* e *AUC-ROC* na predição de casos graves e óbitos respectivamente, um desempenho superior ao esperado quando inicialmente comparado com os trabalhos relacionados na seção 2.4, que alcançaram entre 79% e 95% acurácia com outros conjuntos de dados. Filtrando por casos onde a vacinação estava mais avançada, a predição de óbitos obteve métricas ainda melhores, alcançando 98% e 97% nas médias de acurácia, *F1-Score* e *AUC-ROC* na predição de severidade e óbitos respectivamente. Métodos baseados em *Gradient Boosting*, em especial o *XGBoost*, mostraram um melhor desempenho entre os algoritmos de classificação avaliados.

Os algoritmos escolhidos foram interpretáveis, de modo que a investigação do funcionamento dos modelos gerados fosse possível. Gráficos *SHAP* foram utilizados para identificar fatores de risco como idade avançada, doenças cardiovasculares e estilos de vida não saudáveis como tabagismo e alcoolismo. O progresso da vacinação se mostrou relevante na diminuição da severidade dos casos, como também na redução da probabilidade de óbito, interagindo com fatores de risco de maneira generalizada, em especial foi percebido uma diminuição dos casos

graves em idades mais baixas.

Portanto, se conclui que é possível predizer com satisfatório grau de precisão a evolução de casos de COVID-19 em Recife utilizando dados públicos, por meio de algoritmos de classificação por aprendizagem de máquina. Fatores de risco e efeitos da vacinação podem ser identificados com auxílio dos modelos treinados.

## 5.1 TRABALHOS FUTUROS

Com o propósito de posteriormente estender o escopo do estudo, podem ser utilizados outros algoritmos de classificação como Redes Neurais Artificiais e *Support Vector Machine*, vistos em estudos relacionados. A fim de expandir a análise dos conjuntos de dados disponíveis e melhorar o discernimento de fatores de risco e detecção de impacto de certos fatores como a vacinação, podem ser realizados experimentos de regressão ou aplicação de *Deep Learning*. Pode também ser realizada melhor aplicação de técnicas estatísticas na avaliação dos resultados, por exemplo, testes de hipótese e análises de variância. Ajustes nos parâmetros de configuração dos algoritmos e pré-processamento dos conjuntos de dados podem também ser realizados para melhorar a precisão dos resultados obtidos, sendo possível também investigar a continuação da vacinação na cidade do Recife nos dados futuros, com uma maior porcentagem da população vacinada e informações de doses de reforço da vacina.

## REFERÊNCIAS

- 1 World Health Organization. *Naming the coronavirus disease (COVID-19) and the virus that causes it*. 2020. Disponível em: <[https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-\(covid-2019\)-and-the-virus-that-causes-it](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it)>. Acesso em: 12 dez. 2021.
- 2 World Health Organization. *WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020*. 2020. Disponível em: <<https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>>. Acesso em: 12 dez. 2021.
- 3 E., D. H. D.; L., G. An interactive web-based dashboard to track covid-19 in real time. In: ELSEVIER. *The Lancet Infectious Diseases*. [S.I.], 2020. v. 20, p. 533–534.
- 4 Secretaria de Saúde do Recife. *Casos Leves - Covid-19*. 2020. Disponível em: <<https://dados.recife.pe.gov.br/dataset/casos-leves-covid-19>>. Acesso em: 07 dez. 2021.
- 5 Secretaria de Saúde do Recife. *Casos Graves - Covid-19*. 2020. Disponível em: <[http://dados.recife.pe.gov.br/dataset/casos-graves-covid-19](https://dados.recife.pe.gov.br/dataset/casos-graves-covid-19)>. Acesso em: 07 dez. 2021.
- 6 Prefeitura do Recife. *Vacinômetro - Covid-19*. 2020. Disponível em: <<https://conectarecife.recife.pe.gov.br/vacinometro/>>. Acesso em: 07 dez. 2021.
- 7 IBGE. *IBGE - Cidades - Pernambuco - Recife - Panorama*. 2021. Disponível em: <<https://cidades.ibge.gov.br/brasil/pe/recife/panorama>>. Acesso em: 07 dez. 2021.
- 8 SAWIKR, Y. Transmission and pathogenesis of coronavirus disease (covid-19) outbreak. *Journal of Drug Delivery and Therapeutics*, v. 10, n. 6, p. 239–241, Nov. 2020. Disponível em: <<http://jddtonline.info/index.php/jddt/article/view/4577>>.
- 9 IBM. *Machine Learning*. 2020. Disponível em: <<https://www.ibm.com/in-en/cloud/learn/machine-learning>>. Acesso em: 12 dez. 2021.
- 10 SIDEY-GIBBONS, J. A. M.; SIDEY-GIBBONS, C. J. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, v. 19, n. 1, p. 64, Mar 2019. ISSN 1471-2288. Disponível em: <<https://doi.org/10.1186/s12874-019-0681-4>>.
- 11 JUMPER, J. e. a. Highly accurate protein structure prediction with alphafold. *Nature*, 2021.
- 12 VARADI, M. e. a. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 2021.
- 13 CHOWDHURY, M. E. H.; KHANDAKAR, A.; QIBLAWEY, Y.; REAZ, M. B. I.; ISLAM, M. T.; TOUATI, F. *Machine Learning in Wearable Biomedical Systems*. IntechOpen, 2020. ISBN 9781838803926. Disponível em: <<https://www.intechopen.com/chapters/72859>>.

- 14 GAMA, J.; BRAZDIL, P. Characterization of classification algorithms. In: PINTO-FERREIRA, C.; MAMEDE, N. J. (Ed.). *Progress in Artificial Intelligence*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995. p. 189–200. ISBN 978-3-540-45595-0.
- 15 ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, Taylor & Francis Group, v. 46, n. 3, p. 175–185, 1992.
- 16 JOSE, I. Knn (k-nearest neighbors) 1 - aibrasil - medium. *aibrasil*, Sep 2018. Disponível em: <<https://medium.com/brasil-ai/knn-k-nearest-neighbors-1-e140c82e9c4e>>. Acesso em: 23 dez. 2021.
- 17 PATWARDHAN, S. *Simple understanding and implementation of KNN algorithm!* 2021. Disponível em: <<https://www.analyticsvidhya.com/blog/2021/04/simple-understanding-and-implementation-of-knn-algorithm/>>. Acesso em: 13 dez. 2021.
- 18 SUTHAHARAN, S. Decision tree learning. In: \_\_\_\_\_. *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*. Boston, MA: Springer US, 2016. p. 237–269. ISBN 978-1-4899-7641-3. Disponível em: <<https://doi.org/10.1007/978-1-4899-7641-3\%5F10>>.
- 19 z\_ai. Decision Trees Explained. *Medium*, set. 2021. Disponível em: <<https://towardsdatascience.com/decision-trees-explained-3ec41632ceb6>>.
- 20 SUTHAHARAN, S. Random forest learning. In: \_\_\_\_\_. *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*. Boston, MA: Springer US, 2016. p. 273–288. ISBN 978-1-4899-7641-3. Disponível em: <<https://doi.org/10.1007/978-1-4899-7641-3\%5F11>>.
- 21 YIU, T. *Understanding Random Forest*. 2021. Disponível em: <<https://towardsdatascience.com/understanding-random-forest-58381e0602d2>>. Acesso em: 13 dez. 2021.
- 22 SINGH, H. Understanding gradient boosting machines. *Towards Data Science*, Nov 2018. Disponível em: <<https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab>>. Acesso em: 14 dez. 2021.
- 23 BROWNLEE, J. *A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning*. 2016. Disponível em: <<https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/>>. Acesso em: 14 dez. 2021.
- 24 CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2016. (KDD '16), p. 785–794. ISBN 978-1-4503-4232-2. Disponível em: <<http://doi.acm.org/10.1145/2939672.2939785>>.
- 25 BROWNLEE, J. 2020. Disponível em: <<https://machinelearningmastery.com/light-gradient-boosted-machine-lightgbm-ensemble/>>. Acesso em: 14 dez. 2021.
- 26 SHREYANSI. *LightGBM (Light Gradient Boosting Machine)*. 2020. Disponível em: <<https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/>>. Acesso em: 14 dez. 2021.

- 27 KHANDELWAL, P. 2017. Disponível em: <<https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/>>. Acesso em: 14 dez. 2021.
- 28 TING, K. M. Confusion matrix. In: \_\_\_\_\_. *Encyclopedia of Machine Learning and Data Mining*. Boston, MA: Springer US, 2017. p. 260–260. ISBN 978-1-4899-7687-1. Disponível em: <<https://doi.org/10.1007/978-1-4899-7687-1\%5F50>>.
- 29 METZ, C. E. Basic principles of roc analysis. In: ELSEVIER. *Seminars in nuclear medicine*. [S.I.], 1978. v. 8, p. 283–298.
- 30 CHEN, Y.; OUYANG, L.; BAO, F. S.; LI, Q.; HAN, L.; ZHU, B.; GE, Y.; ROBINSON, P.; XU, M.; LIU, J.; CHEN, S. An Interpretable Machine Learning Framework for Accurate Severe vs Non-Severe COVID-19 Clinical Type Classification. Rochester, NY, n. ID 3638427, jun. 2020. Disponível em: <<https://papers.ssrn.com/abstract=3638427>>.
- 31 ZOABI, Y.; DERI-ROZOV, S.; SHOMRON, N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *npj Digital Medicine*, v. 4, n. 1, p. 3, dez. 2021. ISSN 2398-6352. Disponível em: <<http://www.nature.com/articles/s41746-020-00372-6>>.
- 32 AKTAR, S.; TALUKDER, A.; AHAMAD, M. M.; KAMAL, A. H. M.; KHAN, J. R.; PROTIKUZZAMAN, M.; HOSSAIN, N.; AZAD, A. K. M.; QUINN, J. M. W.; SUMMERS, M. A.; LIAW, T.; EAPEN, V.; MONI, M. A. Machine learning approaches to identify patient comorbidities and symptoms that increased risk of mortality in covid-19. *Diagnostics*, v. 11, n. 8, 2021. ISSN 2075-4418. Disponível em: <<https://www.mdpi.com/2075-4418/11/8/1383>>.
- 33 AHAMAD, M. M.; AKTAR, S.; RASHED-AL-MAHFUZ, M.; UDDIN, S.; LIÒ, P.; XU, H.; SUMMERS, M. A.; QUINN, J. M.; MONI, M. A. A machine learning model to identify early stage symptoms of sars-cov-2 infected patients. *Expert Systems with Applications*, v. 160, p. 113661, 2020. ISSN 0957-4174. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0957417420304851>>.
- 34 MIRANDA, I.; CARDOSO, G.; PAHAR, M.; OLIVEIRA, G.; NIESLER, T. Machine Learning Prediction of Hospitalization due to COVID-19 based on Self-Reported Symptoms: A Study for Brazil\*. fev. 2021. Disponível em: <<https://www.techrxiv.org/articles/preprint/Machine\%5FLearning\%5FPrediction\%5Fof\%5FHospitalization\%5Fdue\%5Fto\%5FCOVID-19\%5Fbased\%5Fon\%5FSelf-Reported\%5FSymptoms\%5FA\%5FStudy\%5Ffor\%5FBrazil\%5F/13736698/1>>.
- 35 WOLLENSTEIN-BETECH, S.; SILVA, A. A. B.; FLECK, J. L.; CASSANDRAS, C. G.; PASCHALIDIS, I. C. Physiological and socioeconomic characteristics predict COVID-19 mortality and resource utilization in Brazil. *PLOS ONE*, v. 15, n. 10, p. e0240346, out. 2020. ISSN 1932-6203. Disponível em: <<https://dx.plos.org/10.1371/journal.pone.0240346>>.
- 36 ROSSUM, G. V.; DRAKE, F. L. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN 1441412697.
- 37 HARRIS, C. R.; MILLMAN, K. J.; WALT, S. J. van der; GOMMERS, R.; VIRTANEN, P. et al. Array programming with NumPy. *Nature*, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, set. 2020. Disponível em: <<https://doi.org/10.1038/s41586-020-2649-2>>.

- 38 TEAM, T. pandas development. *pandas-dev/pandas: Pandas*. Zenodo, 2020. Disponível em: <<https://doi.org/10.5281/zenodo.3509134>>.
- 39 PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B. et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, v. 12, n. Oct, p. 2825–2830, 2011.
- 40 LUNDBERG, S. *SHAP documentation*. 2018. Disponível em: <<https://shap.readthedocs.io/en/latest/index.html>>. Acesso em: 07 dez. 2021.
- 41 WASKOM, M.; BOTVINNIK, O.; O'KANE, D.; HOBSON, P.; LUKAUSKAS, S. et al. *mwaskom/seaborn: v0.8.1 (September 2017)*. Zenodo, 2017. Disponível em: <<https://doi.org/10.5281/zenodo.883859>>.
- 42 HUNTER, J. D. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, IEEE, v. 9, n. 3, p. 90–95, 2007.
- 43 Secretaria de Saúde do Recife. *Relação de pessoas vacinadas - Covid 19*. 2020. Disponível em: <<http://dados.recife.pe.gov.br/dataset/relacao-de-pessoas-vacinadas-covid-19>>. Acesso em: 07 dez. 2021.
- 44 BROWNLEE, J. *Why One-Hot Encode Data in Machine Learning?* 2017. Disponível em: <<https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/>>. Acesso em: 07 dez. 2021.
- 45 REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. In: \_\_\_\_\_. *Encyclopedia of Database Systems*. New York, NY: Springer New York, 2016. p. 1–7. ISBN 978-1-4899-7993-3. Disponível em: <<https://doi.org/10.1007/978-1-4899-7993-3\%5F565-2>>.
- 46 LAVALLE, S. M.; BRANICKY, M. S.; LINDEMANN, S. R. On the relationship between classical grid search and probabilistic roadmaps. *The International Journal of Robotics Research*, SAGE Publications, v. 23, n. 7-8, p. 673–692, 2004.
- 47 BENESTY, J.; CHEN, J.; HUANG, Y.; COHEN, I. Pearson correlation coefficient. In: *Noise reduction in speech processing*. [S.I.]: Springer, 2009. p. 37–40.
- 48 HOLTZ, Y. *Density - From Data to Viz*. 2018. Disponível em: <<https://www.data-to-viz.com/graph/density.html>>. Acesso em: 07 dez. 2021.
- 49 WOOD, T. *Measuring The Accuracy Of AI For Healthcare?* 2020. Disponível em: <<https://fastdatascience.com/measuring-the-accuracy-of-ai-for-healthcare/>>. Acesso em: 18 dez. 2021.

## *APÊNDICE A -*

## MAPA DE CALOR DE CORRELAÇÃO



## *APÊNDICE B –*

## MAPA DE CALOR DE CORRELAÇÃO ANTES DA VACINAÇÃO



APÊNDICE C -

MAPA DE CALOR DE CORRELAÇÃO COM 30% DE VACINAÇÃO

