UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de São José do Rio Preto

Heloísa Silveira Bula

Predição de mortalidade por Dengue: uma abordagem de Ciências de Dados e Aprendizado de Máquina

São José do Rio Preto - SP

2025

Heloísa Silveira Bula

Predição de mortalidade por Dengue: uma abordagem de Ciências de Dados e Aprendizado de Máquina

Monografia apresentada como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, junto ao Departamento de Ciências de Computação e Estatística do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

Universidade Estadual Paulista "Júlio de Mesquita Filho" - (UNESP)

Instituto de Biociências, Letras e Ciências Exatas - (IBILCE)

Departamento de Ciências de Computação e Estatística

Bacharelado em Ciência da Computação

Orientador: Lucas C. Ribas

São José do Rio Preto - SP

2025

B933p

Bula, Heloísa

Predição de mortalidade por Dengue: uma abordagem de Ciências de Dados e Aprendizado de Máquina / Heloísa Bula. -- São José do Rio Preto, 2025

79 f. : il., tabs., fotos

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto

Orientador: Lucas Correia Ribas

1. Aprendizado do computador. 2. Dengue. 3. Classificação. 4. Mortalidade. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

Heloísa Silveira Bula

Predição de mortalidade por Dengue: uma abordagem de Ciências de Dados e Aprendizado de Máquina

Monografia apresentada como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, junto ao Departamento de Ciências de Computação e Estatística do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Câmpus de São José do Rio Preto.

Banca Examinadora:

Prof(a). Dr(a). Lucas Correia Ribas
UNESP - Câmpus São José do Rio Preto
Orientador

Prof. Dr. Diego Renan Bruno
UNESP - Câmpus São José do Rio Preto

Prof. Dr. Rodrigo Capobianco Guido
UNESP - Câmpus São José do Rio Preto

São José do Rio Preto - SP
Data da Defesa

À minha família, que embora nem sempre me compreende-se, me apoiaram incondicionalmente.

Agradecimentos

Agradeço em primeiro lugar à Deus, pela capacidade de obter conhecimento.

Dedico esse Trabalho à minha mãe, que me acalmava em relação à Unesp e me amparava quando eu ficava desanimada ou com dúvidas. Ao meu pai, que sempre se orgulhou e falou que tudo ia dar certo no final, me incentivando a acreditar em mim mesma. Ao meu irmão, por ser a pessoa que me fazia rir quando eu tirava nota baixa, mesmo me irritando com suas piadinhas. Obrigada por ser um ponto de equilíbrio.

Aos meu tios, que me acolheram em casa durante toda a faculdade e foram um apoio durante minha estadia.

Aos meus filhos de quatro patas, Flash, Lana e Luna, por serem meu refúgio e minha dose diária de alegria, com ronrons e latidos que tornaram os longos dias de estudo suportáveis.

Ao meu namorado, que foi um apoio fundamental na minha vida acadêmica. Sua paciência, carinho e conhecimento foram essenciais para me auxiliar durante o período acadêmico e nas dúvidas cruciais sobre o desenvolvimento da pesquisa.

Às minhas amigas, que tornaram essa jornada mais leve e divertida, compartilhando as alegrias e tristezas da graduação. Agradeço por ter conhecido cada uma de vocês e não imagino como seria essa jornada sem nossa amizade.

Ao meu orientador, que me ajudou com a definição de tema e com todos os passos até o presente documento. E aos meus professores, que ensinaram-me toda a base em computação.

"Está aí uma coisa que nunca saberei nem compreenderei - do que os humanos são capazes."
('A menina que roubava livros', Markus Zusak)

Resumo

A Dengue representa um dos principais desafios de saúde pública no Brasil, com taxas crescentes de hospitalização e mortalidade nas últimas décadas. Diante desse cenário, este trabalho propõe duas abordagens complementares baseadas em aprendizado de máquina para a predição do risco de mortalidade em pacientes com Dengue, utilizando dados tabulares clínicos e sociodemográficos. A primeira abordagem, de aprendizado supervisionado, envolveu a aplicação de técnicas de pré-processamento, balanceamento de classes e seleção de atributos, seguida da comparação entre diferentes classificadores, como Árvore de Decisão, *Random Forest*, Regressão Logística e *Naive Bayes*. A segunda abordagem concentrou-se na detecção de *outliers*, por meio dos algoritmos *Isolation Forest*, *Local Outlier Factor* e *One-Class SVM*, visando avaliar seu impacto no desempenho dos modelos preditivos. A análise experimental considerou bases referentes aos anos de 2023 e 2024, permitindo avaliar a robustez das abordagens em diferentes cenários. Os resultados evidenciam que tanto as técnicas de balanceamento de classes quanto a detecção de *outliers* contribuem para aprimorar o desempenho obtido, especialmente na identificação de casos fatais. Assim, o estudo demonstra o potencial das abordagens de aprendizado de máquina como ferramentas de apoio à vigilância epidemiológica e à tomada de decisão em saúde pública.

Palavras-chave: Aprendizado de máquina, Dengue, Mortalidade, Classificação.

Abstract

Dengue represents one of the main public health challenges in Brazil, with increasing hospitalization and mortality rates in recent decades. In this context, this study proposes two complementary machine learning approaches to predict the risk of mortality in Dengue patients using tabular clinical and sociodemographic data. The first approach, based on supervised learning, involved preprocessing, class balancing, and feature selection techniques, followed by the comparison of classifiers such as Decision Tree, Random Forest, Logistic Regression, and Naive Bayes. The second approach focused on the detection and removal of outliers using algorithms such as Isolation Forest, Local Outlier Factor, and One-Class SVM, in order to assess their impact on predictive performance. The experimental analysis considered datasets from 2023 and 2024, allowing the evaluation of the robustness of the approaches under different scenarios. The results show that both class balancing techniques and outlier detection contribute to improving model performance, especially in identifying fatal cases. Therefore, this study highlights the potential of machine learning approaches as tools to support epidemiological surveillance and public health decision-making.

Keywords: *Machine Learning, Dengue, Mortality, Classification.*

Sumário

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ACC	*Acurácia*
AM	*Aprendizado de Máquina*
AUROC	*Área sob a Curva ROC*
AUPRC	*Área sob a Curva de Precisão-Recall*
FN	*Falsos negativos*
FP	*Falsos positivos*
IA	*Inteligência Artificial*
iForest	*Isolation Forest*
KNN	*K-nearest Neighbours*
OC-SVM	*One-Class Support Vector Machine*
OMS	*Organização Mundial da Saúde*
PRC	*Curva de Precisão-Recall*
RL	*Regressão Logística*
ROC	*Receiver Operating Characteristic*
SHAP	*SHapley Additive Explanations*
SINAN	*Sistema de Informação de Agravos de Notificação*
SMOTE	*Synthetic Minority Over-sampling Technique*
SVM	*Support Vector Machine*
UMAP	*Uniform Manifold Approximation and Projection*
VN	*Verdadeiros negativos*
VP	*Verdadeiros positivos*
XGBoost	*Extreme Gradient Boosting*

1 Introdução

Uma arbovirose é um termo usado para doenças transmitidas por artrópodes, dentre elas, destaca-se a Dengue, transmitida pelo mosquito *Aedes aegypti*. Essa doença apresenta grande relevância para a saúde pública no Brasil, uma vez que seus indicadores epidemiológicos como morbidade e mortalidade vêm gradualmente aumentando nos últimos 30 anos (CARDOSO, 2024). Diante desse cenário, prever o risco de mortalidade de pacientes diagnosticados com dengue é uma tarefa essencial para a adoção de medidas preventivas e o direcionamento de recursos hospitalares de forma mais eficiente. Nesse contexto, a aplicação de inteligência artificial e, em particular, o uso de modelos de aprendizado de máquina para a predição de desfechos clínicos, emerge como uma ferramenta promissora.

O aumento expressivo de casos de Dengue não apenas eleva os índices de morbidade e mortalidade, como também sobrecarrega o sistema de saúde, resultando em unidades em colapso e agravamento de comorbidades preexistentes. Esses fatores intensificam o impacto socioeconômico dessa doença (BORGES et al., 2023). A infecção pelo vírus da Dengue apresenta um quadro clínico variável, sendo desde casos assintomáticos até formas graves com risco de morte. Nesse contexto, a predição precisa do risco de mortalidade em pacientes diagnosticados torna-se fundamental para otimizar a alocação de recursos médicos, implementar medidas preventivas direcionadas e, em última instância, reduzir o número de fatalidades.

Apesar da crescente aplicação de técnicas de aprendizado de máquina na área da saúde (ISLAM et al., 2023) e do reconhecimento da Dengue como um relevante problema de saúde pública no Brasil como citado anteriormente, a predição específica da mortalidade, ainda apresenta lacunas relevantes. A maioria dos estudos existentes aborda a predição da gravidade da doença de forma geral, sem necessariamente focar na identificação precisa dos pacientes com maior risco de óbito (HUSSAIN et al., 2023).

Diante desse cenário, torna-se relevante investigar e comparar o desempenho de diferentes algoritmos de classificação para pacientes diagnosticados com Dengue, buscando identificar aqueles que melhor predizem a ocorrência de mortalidade. A análise do potencial preditivo de variáveis clínicas e sociodemográficas, juntamente com a avaliação da interpretabilidade dos modelos de Aprendizado de Máquina (AM) supervisionado gerados, podem fornecer informações valiosas para o aprimoramento das estratégias de manejo clínico e a otimização da alocação de recursos da saúde no enfrentamento da Dengue.

Esta monografia aplica e compara diferentes algoritmos de aprendizado de máquina para identificar modelos com bom desempenho preditivo da mortalidade, utilizando métricas específicas para esse fim. Além disso, a análise da importância das variáveis preditoras pode fornecer informações relevantes sobre os fatores clínicos associados à mortalidade, contribuindo para uma compreensão mais aprofundada da progressão da doença e para o desenvolvimento de estratégias de intervenção mais direcionadas.

A literatura recente reforça o potencial do aprendizado de máquina na área da saúde, com estudos como o de Islam et al. (ISLAM et al., 2023), que evidenciam a capacidade desses métodos de identificar padrões complexos em dados clínicos e contribuir com a tomada de decisão e a gestão mais eficiente dos recursos em saúde. A análise das diferentes abordagens de modelagem — incluindo aprendizado supervisionado e detecção de outliers — possibilita avaliar não apenas o desempenho dos algoritmos, mas também o impacto do pré-processamento e da interpretabilidade na predição de mortalidade por Dengue.

1.1 Objetivos

Este trabalho de conclusão de curso tem como objetivo geral investigar e comparar diferentes técnicas de aprendizado de máquina, com foco na avaliação da capacidade preditiva de algoritmos de classificação aplicados à identificação do risco de mortalidade em pacientes com dengue. O estudo foi estruturado em duas abordagens principais, a primeira utilizando modelos supervisionados de classificação, avaliados sob diferentes estratégias de pré-processamento e balanceamento; e a outra avaliando modelos de detecção de *outliers* que foram empregados na identificação de casos potencialmente associados ao óbito.

Complementarmente à abordagem supervisionada, será realizada uma análise exploratória dos dados, com o intuito de compreender padrões relevantes e identificar os modelos com desempenho mais promissor para essa tarefa específica. Para alcançar o objetivo geral proposto, foram realizados os seguintes objetivos específicos:

- Analisar o impacto do desbalanceamento de classes nos dados e aplicar técnicas de tratamento adequadas para garantir uma análise robusta dos modelos supervisionados.
- Implementar, treinar e comparar os algoritmos de classificação adequados para análise de dados tabulares, construindo modelos preditivos de mortalidade.
- Avaliar e comparar o desempenho dos modelos treinados utilizando métricas apropriadas.
- Investigar a influência das variáveis preditoras nos modelos com melhor desempenho, aplicando técnicas de interpretabilidade para identificar os fatores mais relevantes associados à mortalidade.
- Investigar o papel da detecção de *outliers*, analisando sua capacidade de identificar registros associados a desfechos fatais.
- Avaliar e comparar o desempenho das abordagens supervisionadas e baseadas em detecção de anomalias.
- Discutir a aplicabilidade dos resultados, considerando seu potencial para apoiar a identificação precoce de pacientes de alto risco, bem como para subsidiar estratégias de intervenção clínica e vigilância epidemiológica.

1.2 Estrutura do Trabalho

Este trabalho está organizado em 5 capítulos, contando com a introdução. O Capítulo 2 apresenta uma revisão da literatura, abordando pesquisas atuais sobre a Dengue e a respectiva predição de desfechos clínicos. O Capítulo 3 descreve, de modo mais detalhado, a metodologia adotada, incluindo a descrição do conjunto de dados utilizado, assim como o pré-processamento realizado, algoritmos implementados e as métricas de avaliação empregadas. O Capítulo 4 discute os resultados da avaliação comparativa dos modelos preditivos e a análise da importância das variáveis preditoras. Por fim, o Capítulo 5 conclui o trabalho, resumindo as principais descobertas ao discutir as implicações dos resultados para a área da saúde pública e sugerindo possíveis direções para futuras pesquisas.

2 Fundamentação e Revisão da Literatura

Este capítulo apresenta a fundamentação teórica que embasou o presente trabalho de conclusão de curso, abordando os conceitos essenciais relacionados a predição utilizando Aprendizado de Máquina (AM). Adicionalmente, realiza-se uma revisão da literatura científica existente, com foco em estudos que aplicaram técnicas de AM na análise da Dengue.

2.1 Fundamentação Teórica

2.1.1 Aprendizado de Máquina

O Aprendizado de Máquina (AM) é um subcampo da Inteligência Artificial (IA) que se dedica ao desenvolvimento de algoritmos capazes de identificar padrões em dados e realizar previsões ou decisões com base neles, sem que tenham sido explicitamente programados para cada tarefa (ALPAYDIN, 2020). De modo geral, o processo envolve o treinamento de modelos a partir de conjuntos de dados, permitindo que esses modelos generalizem o conhecimento adquirido para novos exemplos.

Os algoritmos de AM possuem diferentes tipos de categorias, a depender do tipo do problema e dos dados disponíveis. Entre os principais, destacam-se o aprendizado supervisionado, que utiliza dados rotulados para ensinar ao modelo as relações entre entradas e saídas; o aprendizado não supervisionado, que busca descobrir estruturas ocultas em dados não rotulados; e o aprendizado por reforço, em que o modelo aprende por meio de interações com um ambiente, recebendo recompensas ou penalidades a cada ação tomada (ALPAYDIN, 2020).

O aprendizado supervisionado é uma das abordagens mais utilizadas em AM e envolve o treinamento de um modelo com um conjunto de dados rotulado, ou seja, com exemplos em que a saída desejada (rótulo) é conhecida. O objetivo do modelo é aprender uma função que projete as entradas para as saídas, de modo que, dado um novo conjunto de dados, ele consiga prever corretamente o rótulo ou a classe associada. Esse modelo de aprendizado é eficaz em problemas de classificação e regressão, onde o objetivo é atribuir uma classe ou prever um valor contínuo a partir das características de entrada (MITCHELL, 1997).

A abordagem supervisionada tem sido aplicada com sucesso em diversas áreas, como diagnóstico médico, reconhecimento de padrões em imagens e previsão de séries temporais (BISHOP, 2006). Em um contexto de saúde, por exemplo, o aprendizado supervisionado pode ser utilizado para prever a probabilidade de um paciente desenvolver uma determinada doença com base em dados clínicos históricos. O sucesso dessa técnica depende da qualidade dos dados rotulados utilizados para o treinamento, sendo fundamental que as amostras de treinamento cubram adequadamente a variabilidade do problema em questão (DOMINGOS, 2012).

Essa abordagem supervisionada é particularmente relevante na presente monografia, uma vez que realizamos a uma predição da mortalidade a partir de atributos clínicos e sociodemográficos de

pacientes, utilizando modelos de classificação treinados com dados rotulados.

2.1.2 Dados Desbalanceados em Problemas de Classificação

Em problemas de classificação, o desbalanceamento de dados ocorre quando a distribuição das classes alvo é significativamente desigual, com uma ou mais classes majoritárias apresentando um número substancialmente maior de instâncias em comparação com as demais (HE; GARCIA, 2009). Essa distribuição assimétrica pode surgir naturalmente em diversos domínios.

O desbalanceamento impõe desafios significativos ao treinamento de modelos de AM destinados à classificação, uma vez que muitos algoritmos priorizam a otimização de métricas de desempenho globais, como a acurácia. Em tais contextos, um modelo pode apresentar desempenho aparentemente satisfatório simplesmente classificando a maioria das instâncias como pertencentes à classe majoritária, ainda que falhe em identificar corretamente exemplos da classe minoritária, a qual contém frequentemente a informação de maior interesse prático (CHAWLA; JAPKOWICZ; KOTCZ, 2004).

O principal impacto do desbalanceamento está na dificuldade do modelo em aprender os padrões característicos da classe minoritária. A escassez de exemplos representativos prejudica a capacidade do algoritmo ajustar seus parâmetros de forma a reconhecer essas instâncias de maneira eficaz(JAPKOWICZ; STEPHEN, 2002). Consequentemente, o modelo tende a apresentar baixa sensibilidade (*recall*) e baixa precisão para a classe minoritária. Em contextos como a predição de mortalidade, onde a identificação correta dos casos fatais é crucial, essa limitação pode comprometer seriamente a utilidade prática do modelo.

Diante disso, torna-se essencial reconhecer e considerar o desbalanceamento durante o desenvolvimento de modelos preditivos, especialmente em aplicações sensíveis como a área da saúde.

2.1.3 Técnicas para Balanceamento de Dados

Para enfrentar os desafios decorrentes do desbalanceamento, diversas técnicas são aplicadas com o objetivo de equilibrar a distribuição das classes no conjunto de dados. Essas abordagens podem ser categorizadas em três tipos principais: *oversampling*, que consiste em aumentar a representatividade da classe minoritária através da adição de novas amostras; *undersampling*, que reduz a dominância da classe majoritária por meio da remoção seletiva de exemplos; e técnicas combinadas, que aplicam simultaneamente estratégias de aumento e redução de dados (CHAWLA; JAPKOWICZ; KOTCZ, 2004; HE; GARCIA, 2009).

O propósito dessas técnicas é permitir que os modelos aprendam de maneira mais equilibrada, aprimorando o desempenho na identificação de todas as classes e reduzindo o viés causado pela disparidade na representação dos dados (JAPKOWICZ; STEPHEN, 2002). Nos tópicos seguintes, serão apresentadas e detalhadas algumas das técnicas utilizadas no tratamento do desbalanceamento.

2.1.3.1 Synthetic Minority Over-sampling Technique (SMOTE)

O SMOTE é uma técnica de *oversampling*, utilizada para balanceamento de dados. Ela visa aumentar a representatividade da classe minoritária em conjuntos de dados desbalanceados através da

criação de amostras sintéticas. Ao invés de simplesmente duplicar os exemplos existentes da classe minoritária, essa técnica gera novas instâncias artificiais interpolando entre exemplos reais próximos, promovendo uma maior diversidade nos dados gerados (CHAWLA et al., 2002).

O algoritmo consiste em, para cada exemplo da classe minoritária, identificar seus *k* vizinhos mais próximos no espaço de características (onde *k* é um parâmetro configurável, sendo 5 um valor comumente utilizado). Novas amostras sintéticas são criadas ao selecionar aleatoriamente um desses vizinhos também da classe minoritária e calcular pontos intermediários na distância entre o exemplo original e o vizinho escolhido. Essa interpolação é feita aplicando-se uma combinação linear ponderada, onde um valor aleatório entre 0 e 1 define a posição do novo ponto ao longo do segmento que une os dois exemplos.

Essa abordagem contribui para expandir a região da classe minoritária, melhorando o aprendizado do modelo ao fornecer exemplos mais variados e evitando o problema de *overfitting* com a simples replicação de dados. Além disso, o SMOTE tem se mostrado eficaz para melhorar o desempenho de classificadores em métricas relacionadas à classe minoritária, como *recall* e *F1-score*, especialmente em contextos críticos como a detecção de doenças ou eventos raros (HE; GARCIA, 2009).

2.1.3.2 Tomek Links

O *Tomek Links* é uma técnica de *undersampling* que atua na redução da sobreposição entre classes - classes diferentes estão muito próximas no espaço de características, dificultando a distinção clara entre eles. A técnica foca na remoção de instâncias da classe majoritária que estão sobrepostas em relação à classe minoritária. Essa limpeza do conjunto de dados ajuda a tornar a fronteira entre as classes mais nítida, facilitando o aprendizado do modelo (TOMEK, 1976).

A técnica se baseia na identificação de pares de amostras, denominados *Tomek Links*, que pertencem a classes distintas e são mutuamente os vizinhos mais próximos uma da outra no espaço de características. Formalmente, um par de exemplos (x_i, x_j) são nomeadas como um *Tomek Links* se:

- x_i e x_j pertencem a classes diferentes;
- x_j é o vizinho mais próximo de x_i;
- x_i é o vizinho mais próximo de x_j.

O uso dessa técnica contribui para aprimorar a separabilidade das classes, diminuindo a complexidade da fronteira de decisão e possibilitando que os algoritmos de aprendizado se concentrem em exemplos mais representativos. Assim, a aplicação dessa técnica pode melhorar o desempenho dos modelos, especialmente em cenários onde as classes apresentam sobreposição significativa (BATISTA; PRATI; MONARD, 2004).

2.1.3.3 Combinação de Técnicas para Balanceamento de Dados

A combinação de estratégias de *oversampling* e *undersampling* emerge como uma abordagem robusta para o balanceamento de dados desbalanceados, buscando otimizar os benefícios de ambas

as metodologias. Enquanto o *oversampling* — como a técnica SMOTE — atua na expansão da representatividade da classe minoritária pela geração de exemplos sintéticos, o *undersampling* — como os *Tomek Links* — foca na remoção de instâncias da classe majoritária que causam sobreposição e ruído.

Uma combinação comumente empregada é a sequência *SMOTE+Tomek Links*. Nesta estratégia, o SMOTE é aplicado inicialmente aumentando o número de registros da classe minoritária, enriquecendo sua representatividade. Em seguida, os *Tomek Links* são identificados e removidos, promovendo uma limpeza da fronteira de decisão. Essa etapa subsequente é crucial, pois elimina instâncias ambíguas da classe majoritária que estão próximas da classe minoritária, tornando a separação entre as classes mais nítida e reduzindo o impacto de ruído no aprendizado do modelo.

A aplicação sequencial dessas técnicas têm como objetivo não apenas corrigir a disparidade numérica entre as classes, mas também refinar a estrutura do conjunto de dados, tornando-o mais informativo e menos propenso a erros de classificação. Essa estratégia combinada pode melhorar significativamente o desempenho preditivo, especialmente em métricas como *F1-score* e *recall* da classe minoritária, que são cruciais em cenários sensíveis, como a área médica (HERNANDEZ; CARRASCO-OCHOA; MARTíNEZ-TRINIDAD, 2013).

2.1.4 Interpretabilidade e visualização de Modelos Preditivos

A interpretabilidade de modelos de Aprendizado de Máquina é um aspecto fundamental, especialmente em domínios sensíveis, onde a tomada de decisão depende da compreensão dos fatores que influenciam a predição. Modelos mais complexos ou com múltiplos atributos podem apresentar comportamento difícil de ser compreendido diretamente, o que dificulta sua aceitação e aplicação prática. Por esse motivo, diversas técnicas têm sido desenvolvidas para explicar como os modelos tomam decisões e de visualizar as relações entre variáveis, promovendo uma análise mais transparente, confiável e útil para os usuários finais.

2.1.4.1 SHapley Additive Explanations (SHAP)

O *SHapley Additive Explanations* (SHAP) é uma técnica de interpretação de modelos baseada na teoria dos valores de *Shapley*, oriunda dos jogos cooperativos. Essa abordagem busca quantificar a contribuição individual de cada variável para a predição realizada por um modelo. A ideia central consiste em avaliar o impacto que uma variável tem no resultado ao ser incluída em diferentes combinações possíveis com as demais variáveis(LUNDBERG; LEE, 2017). Com isso, o SHAP calcula uma média ponderada das diferenças nas predições com e sem a presença da variável, garantindo que cada contribuição seja distribuída de forma justa entre as variáveis.

Uma das principais vantagens desse algoritmo é sua capacidade de fornecer explicações tanto locais — para uma instância específica — quanto globais, refletindo o comportamento geral do modelo. Essa característica o torna especialmente útil em domínios sensíveis, nos quais decisões críticas são tomadas com base em dados. Ao oferecer interpretações transparentes e consistentes das decisões do modelo, o SHAP contribui para maior confiança e aceitação dos resultados por parte dos usuários

finais.

2.1.4.2 Uniform Manifold Approximation and Projection (UMAP)

O *Uniform Manifold Approximation and Projection* (UMAP) é uma técnica para redução de dimensionalidade que tem ganhado destaque nos últimos anos, ela se destaca por suas vantagens em termos de eficiência computacional e qualidade na visualização dos dados.

O UMAP foi desenvolvido com o objetivo de representar dados de alta dimensão em espaços de menor dimensão de forma eficiente, preservando tanto as estruturas locais quanto as globais dos dados (MCINNES; HEALY; MELVILLE, 2018). Isso significa que o método mantém a proximidade entre pontos semelhantes ao mesmo tempo em que busca conservar o formato geral e os agrupamentos presentes no conjunto original, permitindo visualizar padrões mais amplos e relações entre diferentes grupos de amostras.

Por esses motivos, o UMAP tem sido amplamente adotado em áreas como biomedicina, genômica e outras ciências onde a visualização e interpretação de dados complexos são essenciais. Sua capacidade de manter uma boa representação tanto local quanto global facilita a identificação de subgrupos, padrões e relações importantes nos dados, contribuindo para análises mais ricas e intuitivas (BECHT et al., 2019).

Com um conjunto de características definido, a etapa seguinte consiste na escolha e aplicação de um algoritmo de classificação adequado. As subseções a seguir apresentam alguns dos classificadores para modelos de AM.

2.1.5 Classificadores dos Modelos de Aprendizado de Máquina

A classificação consiste na atribuição automática de rótulos a dados com base em suas características, sendo uma etapa fundamental em diversas aplicações de aprendizado de máquina. Para melhorar o desempenho dos classificadores, muitas vezes aplica-se a seleção e avaliação de características, buscando um subconjunto relevante que otimize a capacidade preditiva do modelo. Diversos algoritmos podem ser empregados na tarefa de classificação. Esta seção apresenta uma breve fundamentação sobre alguns dos classificadores mais comuns na literatura, abordando suas principais características e funcionamento.

2.1.5.1 Support Vector Machine (SVM)

O *Support Vector Machine* (SVM) é um algoritmo de AM supervisionado amplamente utilizado em tarefas de classificação e regressão, principalmente por sua eficácia na separação entre classes em diferentes espaços de características (VAPNIK, 2013).

O princípio fundamental do SVM consiste na identificação de um hiperplano ótimo que divida os dados em diferentes classes, de forma que a margem entre as amostras mais próximas de cada classe - chamadas de *vetores de suporte* - seja maximizada. Em um espaço n-dimensional, esse hiperplano é uma superfície de dimensão $n-1$ que separa os dados com a maior margem possível, o que contribui para uma melhor generalização do modelo.

Dado um conjunto de treinamento $(x_i, y_i)_{i=1}^n$, onde $x_i \in \mathbb{R}^d$ representa os vetores de características e $y_i \in \{-1, 1\}$ os rótulos das classes, o objetivo do SVM é determinar o hiperplano ótimo que separa os dados das duas classes, maximizando a margem entre os exemplos mais próximos de cada categoria.

Esse hiperplano é definido por um vetor $\mathbf{w}$ e um valor de viés b, resolvendo um problema de otimização que busca minimizar a norma de $\mathbf{w}$, o que equivale a maximizar a margem, obedecendo à condição de que todos os pontos estejam corretamente classificados, ou o mais próximo disso, em casos de dados não separáveis linearmente.

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{sujeito a} \quad y_i(\mathbf{w} \cdot x_i + b) \geq 1, \forall i, \tag{1}$$

Em casos onde os dados não são linearmente separáveis, o SVM introduz variáveis de folga (ξ_i) e um parâmetro de penalização C, que controla o *trade-off* entre a maximização da margem e a minimização dos erros de classificação. Isso permite certa flexibilidade na classificação ao tolerar erros em alguns exemplos para melhorar a generalização do modelo (VAPNIK, 2013).

Essa flexibilidade é particularmente útil em conjuntos de dados desbalanceados, onde o ajuste do parâmetro C ou o uso de *class_weight* que atribui pesos diferentes às classes durante o treinamento, de modo que o modelo dê mais atenção e penalize mais os erros na classe minoritária, como a mortalidade em problemas de saúde.

Além disso, o SVM permite aplicar uma transformação dos dados para um espaço de maior dimensão, onde a separação linear seja viável. Essa transformação é realizada indiretamente por meio de uma função *kernel*, evitando o custo computacional de projetar os dados explicitamente (SCHÖLKOPF; SMOLA, 2002).

Entre os *kernels* mais utilizados, destacam-se:

- Lineares: $K(x_i, x_j) = x_i \cdot x_j$;
- Polinomiais: Permitem separar os dados com curvas polinomiais de diferentes graus através da equação $(\gamma x_i \cdot x_j + r)^d$;
- *Radial Basis Function* (RBF): Cria fronteiras altamente flexíveis aplicando uma exponencial a $(-\gamma\|x_i - x_j\|^2)$

O *Support Vector Machine* apresenta como vantagens sua robustez contra *overfitting* em espaços de alta dimensionalidade, boa capacidade de generalização e forte embasamento teórico. No entanto, a escolha do *kernel* e de seus parâmetros influencia diretamente a capacidade do modelo de capturar padrões nos dados, exigindo cuidados na fase de pré-processamento e validação.

2.1.5.2 Árvore de Decisão

Entre os métodos mais tradicionais e interpretáveis, destacam-se as Árvores de Decisão. Esse método é um modelo preditivo que utiliza uma estrutura hierárquica em forma de árvore, onde cada

nó interno representa uma decisão baseada em uma variável de entrada, os ramos correspondem aos possíveis resultados dessa decisão e os nós folha correspondem às previsões finais (BREIMAN et al., 2017). O processo de construção de uma árvore envolve a divisão recursiva do conjunto de dados com base em atributos que maximizem a separação entre as classes, segundo critérios como o índice de Gini, entropia ou ganho de informação. Tais critérios visam medir a impureza dos nós e promover divisões que resultem em nós mais homogêneos em relação à variável alvo. Essa estratégia permite que o modelo aprenda regras de decisão simples e facilmente interpretáveis a partir dos dados.

Uma das principais vantagens das árvores de decisão é sua capacidade de lidar com dados heterogêneos (numéricos e categóricos), além de oferecer interpretabilidade ao permitir a visualização das regras de decisão. No entanto, quando usadas isoladamente, tendem a se ajustar demais aos dados de treinamento, sofrendo com o problema de *overfitting*. Por essas razões, árvores de decisão são frequentemente utilizadas como modelos base em métodos de *ensemble*, como o *Random Forest*, que combinam múltiplas árvores para aumentar a robustez e o desempenho preditivo do modelo.

2.1.5.3 Random Forest

Como exposto anteriormente o *Random Forest* é um método de aprendizado supervisionado baseado em um conjunto de árvores de decisão, proposto por Breiman (BREIMAN, 2001). A ideia central do algoritmo é a construção de diversas árvores de decisão independentes e combinar suas saídas, geralmente por votação majoritária, no caso de classificação.

Cada árvore é construída a partir de uma amostra aleatória com reposição (*bootstrap*) do conjunto de dados original. Além disso, durante a divisão de cada nó da árvore, apenas um subconjunto aleatório de atributos é considerado para seleção, o que introduz diversidade entre as árvores e reduz a correlação entre elas. Esse mecanismo de aleatoriedade ajuda a mitigar o problema de *overfitting*, bastante comum em árvores de decisão isoladas, melhorando a capacidade de generalização do modelo.

O processo pode ser descrito da seguinte forma: para cada árvore, escolhem-se aleatoriamente K atributos dentre o total disponível; desses, seleciona-se aquele que produz a melhor separação das classes e define-o como nó de decisão. Esse processo se repete recursivamente para a construção dos ramos e folhas, até que critérios de parada sejam atendidos. A predição final é obtida pela agregação das decisões de todas as árvores.

Entre suas principais vantagens, destacam-se a alta acurácia, a capacidade de lidar com grandes volumes de dados e variáveis ruidosas, e a estimativa interna do erro generalizado por meio do método *Out-of-Bag* (OOB) que utiliza amostras não incluídas no treinamento de cada árvore (FERNÁNDEZ-DELGADO et al., 2014).

Sua robustez e capacidade de capturar interações complexas entre variáveis o tornam particularmente eficaz em dados tabulares, e ele pode ser combinado com as técnicas de amostragem (como *oversampling* ou *undersampling*) para lidar com o desbalanceamento de classes. No entanto, o *Random Forest* pode demandar maior tempo de treinamento e menos interpretativa comparado com uma única árvore (DIAS, 2024).

2.1.5.4 Extreme Gradient Boosting (XGBoost)

O *Extreme Gradient Boosting* (XGBoost) é um algoritmo de AM baseado em árvores de decisão que também pertence à classe dos métodos de *ensemble* por *boosting*, ou seja, treinam os modelos de forma sequencial, de modo que cada novo modelo tente corrigir os erros cometidos pelos modelos anteriores. Essa técnica aprimora a técnica de *Gradient Boosting Decision Trees* (GBDT) com foco em eficiência computacional, escalabilidade e regularização, sendo amplamente reconhecido por seu desempenho superior em tarefas de classificação e regressão, especialmente em dados tabulares (CHEN; GUESTRIN, 2016). Seu funcionamento baseia-se na construção sequencial de árvores de regressão, em que cada nova árvore é treinada para minimizar os erros cometidos pelas anteriores.

Além disso, o XGBoost implementa técnicas como *shrinkage* (similar a uma taxa de aprendizado) e subamostragem de colunas e instâncias, que auxiliam na redução de *overfitting* e melhoram a capacidade de generalização do modelo (SHAW, 2017). O sistema é otimizado para execução paralela, com pré-busca de cache e suporte a computação fora de memória (*out-of-core*), o que o torna eficaz para grandes volumes de dados mesmo em ambientes com recursos computacionais limitados (IBM, 2024).

Para lidar com conjuntos de dados desbalanceados, como é comum em problemas de saúde, o XGBoost oferece a flexibilidade de ajustar o peso das classes minoritárias por meio de parâmetros como *scale_pos_weight*. Esses parâmetros permitem atribuir maior importância à classe positiva (geralmente a minoritária), ajustando o impacto de seus erros durante o treinamento.

O valor de `scale_pos_weight` pode ser calculado como a razão entre o número de exemplos negativos e positivos:

$$scale\_pos\_weight = \frac{\text{n\_negativos}}{\text{n\_positivos}}$$

Esse ajuste força o modelo a focar mais a atenção nos exemplos da classe minoritária durante o cálculo da função de perda, o que é especialmente útil em tarefas como a predição de mortalidade, onde é crítico detectar corretamente os poucos casos positivos, mesmo que isso implique em mais falsos positivos.

Além disso, o uso de `scale_pos_weight` pode ser combinado com outras técnicas como *oversampling* ou *undersampling*, ampliando a capacidade do modelo de capturar padrões relevantes da classe minoritária.

Diversos estudos demonstram a aplicabilidade e eficácia do XGBoost em problemas reais, incluindo previsão de epidemias, análise de risco, detecção de fraudes e diagnóstico médico. A capacidade do XGBoost de lidar com dados heterogêneos e de capturar interações complexas entre variáveis torna-o especialmente adequado para esse tipo de problema (GUO et al., 2017).

Portanto, o XGBoost representa uma evolução significativa nos algoritmos baseados em árvores, combinando alto desempenho preditivo com otimizações computacionais que o tornam aplicável a uma ampla gama de problemas, inclusive no campo da saúde pública.

2.1.5.5 K-nearest Neighbour (KNN)

O *K-Nearest Neighbors* (KNN) é um algoritmo de aprendizado supervisionado baseado em instâncias, que realiza a classificação (ou regressão) de uma nova amostra com base na similaridade entre ela e os exemplos do conjunto de treinamento. O princípio fundamental do KNN é que objetos semelhantes tendem a estar próximos no espaço de atributos, ou seja, instâncias com características similares provavelmente pertencem à mesma classe (COVER; HART, 1967). Para classificar uma nova amostra, o algoritmo identifica os k vizinhos mais próximos — normalmente utilizando uma medida de distância, como a distância Euclidiana — e atribui a classe mais comum nesse espaço.

O KNN não possui fase explícita de treinamento, sendo considerado um método *lazy learning*, pois adia o processo de generalização até o momento da predição, tornando-o simples de implementar. Entre suas vantagens, destacam-se a capacidade de lidar com múltiplas classes e a ausência de suposições fortes sobre a distribuição dos dados (ALPAYDIN, 2020).

Por outro lado, o KNN pode ser computacionalmente custoso em bases de dados grandes, uma vez que requer o cálculo das distâncias para todas as instâncias durante a predição. Além disso, sua performance pode ser prejudicada em cenários com alta dimensionalidade ou com atributos irrelevantes, exigindo técnicas de pré-processamento como normalização e seleção de características para melhores resultados.

2.1.5.6 Naive Bayes

O *Naive Bayes* é um algoritmo baseado no Teorema de *Bayes*, com a suposição de independência condicional entre os atributos. Seu princípio fundamental é calcular a probabilidade de uma instância pertencer a determinada classe com base nas probabilidades individuais dos atributos, assumindo que esses são estatisticamente independentes entre si, dado o valor da classe (MARON, 1961). Apesar dessa suposição simplificadora — frequentemente violada na prática — o algoritmo tem se mostrado eficaz em diversas tarefas de classificação, especialmente em domínios com alta dimensionalidade. A fórmula geral do Teorema de Bayes é:

$$P(C_k \mid x) = \frac{P(x \mid C_k) \cdot P(C_k)}{P(x)}$$

O termo resultante $P(C_k \mid x)$ é a probabilidade da classe C_k ocorrer dado o vetor de atributos x, enquanto $P(x \mid C_k)$ é a probabilidade do vetor de atributos dados a ocorrência da classe C_k, por fim $P(C_k)$ é a probabilidade da classe ocorrer, e $P(x)$ é a evidência.

A simplicidade computacional é uma das maiores vantagens do *Naive Bayes*, uma vez que o treinamento consiste basicamente em estimar as distribuições de probabilidade a partir dos dados. Além disso, o algoritmo é escalável, robusto a dados ruidosos e funciona bem mesmo com conjuntos de dados pequenos.

Por outro lado, a suposição de independência entre atributos pode limitar sua performance em problemas onde existe forte correlação entre variáveis (DOMINGOS; PAZZANI, 1997). Ainda assim, o

Naive Bayes continua sendo uma escolha popular como modelo de linha de base ou quando se busca uma solução rápida e eficiente.

2.1.5.7 Regressão Logística (RL) com penalização

A Regressão Logística (RL) é um algoritmo fundamental de classificação binária que modela a probabilidade de um evento ocorrer. Para isso, ela usa a função sigmoide expressa abaixo:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Ao usar essa função, uma combinação linear de variáveis preditoras é transformada em uma probabilidade com valor entre 0 e 1. Tradicionalmente, os coeficientes do modelo são estimados pela maximização da função de *log-verossimilhança*.

Contudo, a RL padrão pode apresentar desafios em certas situações. Em conjuntos de dados com muitas variáveis (alta dimensionalidade) ou quando se lida com fatores de risco com prevalência muito baixa (ou seja, exposições raras), o método de estimação convencional pode falhar em convergir ou produzir estimativas de risco que sejam confiáveis e plausíveis (FRIEDRICH et al., 2023).

Para superar essas limitações e controlar o *overfitting*, a Regressão Logística Penalizada é empregada. Essa técnica aprimora a estimação dos coeficientes ao adicionar um termo de penalidade à função original (TIBSHIRANI, 1996). O efeito dessa penalidade é encolher os coeficientes dos preditores, aproximando-os de zero, ou até mesmo zerando alguns deles, o que resulta em modelos mais simples e robustos.

A penalização age como uma forma de "informação adicional"que guia o processo de estimação. Ela cria um equilíbrio entre o quão bem o modelo se ajusta aos dados e sua própria complexidade. Ao reduzir a magnitude dos coeficientes, a penalização impede que o modelo atribua uma importância excessiva a pequenas variações nos dados de treinamento, o que é crucial para melhorar a capacidade de generalização para novos dados. O grau de penalidade é controlado por um parâmetro de ajuste, que geralmente é otimizado através de técnicas como a validação cruzada, para garantir o melhor desempenho do modelo.

Em problemas com desbalanceamento de classes, a RL pode ser adaptada usando o parâmetro *class_weight='balanced'*. Essa abordagem ajusta automaticamente os pesos das classes durante o treinamento, dando maior importância à classe minoritária e ajudando o modelo a aprender melhor os padrões associados à ela (scikit-learn developers, 2024).

A Regressão Logística Penalizada é altamente valorizada por sua interpretabilidade, uma vez que os coeficientes ainda podem ser usados para entender a influência de cada variável no desfecho. Sua capacidade de gerar estimativas estáveis em dados esparsos ou com fatores de risco de baixa prevalência, torna-a uma ferramenta robusta e valiosa na pesquisa biomédica e epidemiológica, servindo como um excelente ponto de partida para muitos estudos (FRIEDRICH et al., 2023).

2.2 Métricas para Avaliação de Desempenho

A avaliação rigorosa do desempenho é crucial para garantir a confiabilidade e a aplicabilidade dos modelos de classificação. Para quantificar a qualidade das predições dos classificadores empregados nesta monografia, foram consideradas as seguintes métricas amplamente estabelecidas na literatura: Acurácia (ACC), Precisão e *F1-score*.

2.2.1 Matriz de Confusão

O cálculo das métricas citadas baseia-se na chamada Matriz de Confusão, uma estrutura que resume os resultados de uma classificação binária. A matriz de confusão pode ser esquematizada conforme a Tabela 1.

Tabela 1 – Representação esquemática de uma matriz de confusão

		Classe Predita	
		Positivo	**Negativo**
Classe Real	**Positivo**	VP	FN
	Negativo	FP	VN

Para a correta interpretação da Matriz de Confusão, é importante compreender o significado de cada uma de suas categorias:

- **Verdadeiros Positivos (VP):** instâncias positivas corretamente classificadas como positivas;
- **Verdadeiros Negativos (VN):** instâncias negativas corretamente classificadas como negativas;
- **Falsos Positivos (FP):** instâncias negativas incorretamente classificadas como positivas;
- **Falsos Negativos (FN):** instâncias positivas incorretamente classificadas como negativas.

Conforme mencionado anteriormente, é possível utilizar diversas métricas para mensurar o desempenho dos classificadores. A seguir, são apresentadas as principais métricas adotadas nesta monografia, iniciando pela Acurácia.

2.2.2 Acurácia (ACC)

A métrica Acurácia (ACC) é amplamente utilizada para avaliar o desempenho de modelos de classificação. Ela indica a proporção de classificações corretas do modelo — tanto de instâncias positivas quanto negativas — em relação ao total de instâncias avaliadas. Seu cálculo é baseado nos valores da Matriz de Confusão e é dado por:

$$ACC = \frac{VP + VN}{VP + VN + FP + FN} \tag{2}$$

O resultado obtido pode ser interpretado como a taxa de acertos do modelo, expressa em forma percentual entre 0% e 100%. No entanto, é importante notar que a acurácia pode ser uma métrica enganosa em conjuntos de dados desbalanceados, onde um modelo pode obter alta acurácia simplesmente prevendo a classe majoritária.

2.2.3 Métricas Base

As métricas base são medidas fundamentais utilizadas para avaliar o desempenho de modelos de classificação. Elas quantificam aspectos específicos do resultado do modelo e servem de base para o cálculo de métricas mais compostas e complexas.

Nesta seção, apresentamos as métricas base que avaliam a performance do modelo de classificação: Precisão e *Recall*.

- **Precisão:**

  A precisão mede a proporção de amostras classificadas como positivas pelo modelo e que, de fato, são positivas, ou seja VP. Em outras palavras, entre todas as previsões positivas feitas pelo modelo, abrangendo Verdadeiro Positivo somado a Falso Positivo, quantas realmente estavam corretas. Ela pode ser calculada através da fórmula:

$$P = \frac{VP}{VP + FP} \tag{3}$$

- **Recall:**

  O *recall* (também chamado de sensibilidade) indica a proporção de amostras positivas que foram corretamente identificadas pelo modelo. Ou seja, de todas as amostras que realmente pertencem à classe positiva (VP junto com FN), quantas o modelo conseguiu reconhecer. A fórmula para essa métrica é:

$$R = \frac{VP}{VP + FN} \tag{4}$$

2.2.4 F1-Score

O *F1-Score* combina as últimas duas métricas apresentadas, precisão e *recall*, em uma única métrica por meio da média harmônica. Essa combinação é importante porque considera o equilíbrio entre a capacidade do modelo de ser preciso e de recuperar corretamente as amostras positivas. A fórmula para o cálculo do *F1-Score* é:

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{5}$$

Seus resultados variam entre 0 e 1, sendo que valores próximos a 1 indicam um bom desempenho tanto em precisão quanto em *recall*.

2.3 Métricas para Avaliação de Modelos com dados desbalanceados

Em conjuntos de dados desbalanceados, a acurácia pode ser enganosa, favorecendo a classe majoritária, como explicado anteriormente. Por isso, métricas específicas como a Área sob a Curva ROC (AUROC) e Área sob curva de *Precisão-Recall* (AUPRC) são mais adequadas, pois consideram o desempenho do modelo em ambas as classes, mesmo quando desproporcionais.

2.3.1 Área sob a Curva ROC (AUROC)

A Análise da Curva ROC (*Receiver Operating Characteristic*) é uma ferramenta gráfica e analítica amplamente empregada para avaliar e comparar o desempenho de modelos de classificação binária. O gráfico é constituído pela taxa de Verdadeiros Positivos (VP) no eixo Y contra a taxa de Falsos Positivos (FP) no eixo X, demonstrando o *trade-off* inerente entre essas duas métricas.

Para resumir o desempenho de um classificador em um único valor, utiliza-se a Área Sob a Curva ROC (AUC). O valor da AUC varia de 0 a 1, onde 0.5 indica um classificador sem poder discriminatório (equivalente ao aleatório) e 1.0 representa um classificador perfeito. Uma propriedade estatística fundamental da AUC é que ela pode ser interpretada como a probabilidade de que um classificador atribua uma pontuação de probabilidade mais alta a uma instância positiva escolhida aleatoriamente do que a uma instância negativa também selecionada ao acaso (BRADLEY, 1997). Essa característica torna a AUC uma métrica particularmente robusta para conjuntos de dados desbalanceados, pois seu valor não é influenciado pela proporção ou distribuição das classes (FAWCETT, 2006).

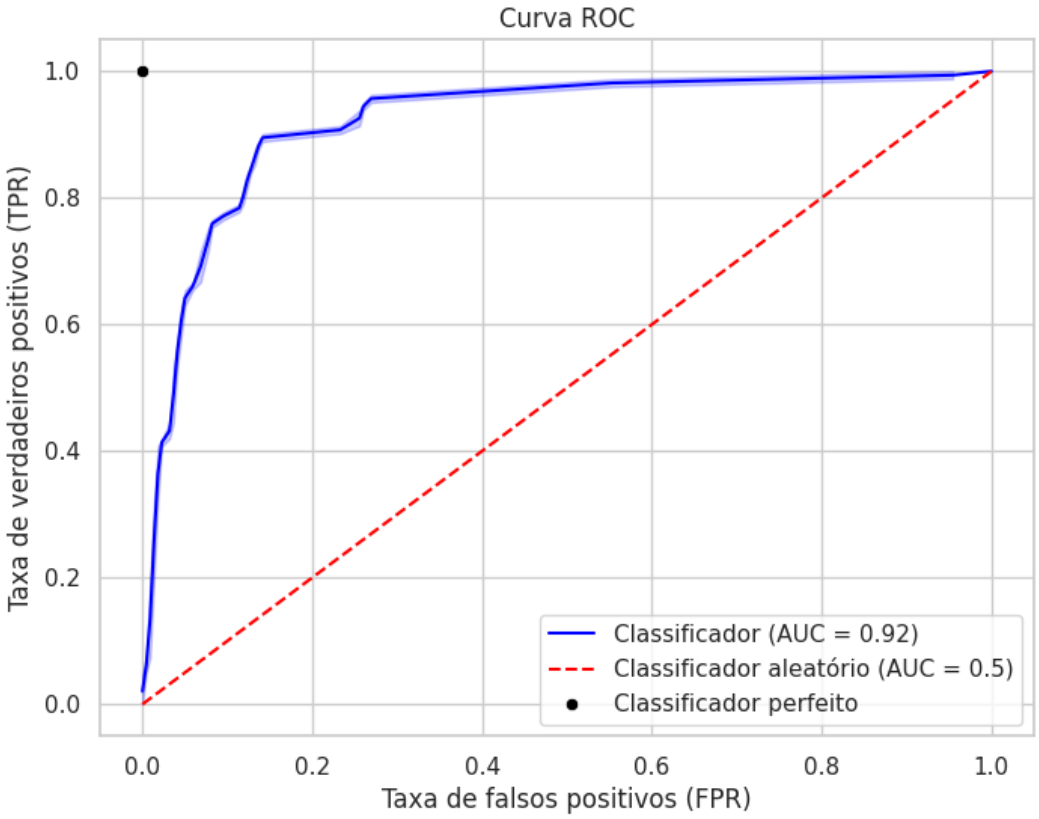
Originada na teoria de detecção de sinais e posteriormente popularizada nas áreas de tomada de decisões médicas e epidemiologia (METZ, 1978). Os gráficos ROC apresentam de forma intuitiva o *trade-off* entre os benefícios e os custos de um classificador em diferentes limiares de decisão. A Figura 1 apresenta um exemplo ilustrativo.

No gráfico ilustrativo, a linha azul sólida representa o desempenho de um modelo classificador, que atingiu uma AUC de 0.92, indicando um bom poder discriminatório. A linha tracejada vermelha, correspondente a $y = x$, ilustra o desempenho de um classificador aleatório, com AUC de 0.5. O ponto preto no canto superior esquerdo (0,1) simboliza o classificador perfeito, que atingiria 100% de Verdadeiros Positivos sem nenhum Falso Positivo.

Classificadores que fornecem apenas um rótulo de classe discreto produzem um único ponto no espaço ROC. No entanto, modelos de AM como a Regressão Logística, que geram uma pontuação de probabilidade para cada instância, permitem que uma curva ROC completa seja traçada. Isso é feito variando-se o limiar de corte, o que representa o desempenho do classificador em todo o espectro de sensibilidade e especificidade (FAWCETT, 2006).

Dada essa robustez a variações na proporção de classes e sua independência dos custos de erro, a AUROC se posiciona como uma métrica superior em diversos contextos clínicos e de pesquisa. Ela fornece uma visão abrangente do poder discriminatório do modelo, permitindo que classificadores sejam selecionados considerando-se o equilíbrio ideal das classificações, de acordo com as necessidades

Figura 1 – Exemplo de Curva ROC e a Área sob a Curva (AUC). O gráfico ilustra o desempenho de um classificador (curva azul) em relação a um classificador aleatório (linha tracejada vermelha) e ao ponto de classificação perfeita (ponto preto).



Fonte: Elaborado pela autora.

específicas da aplicação (FAWCETT, 2006).

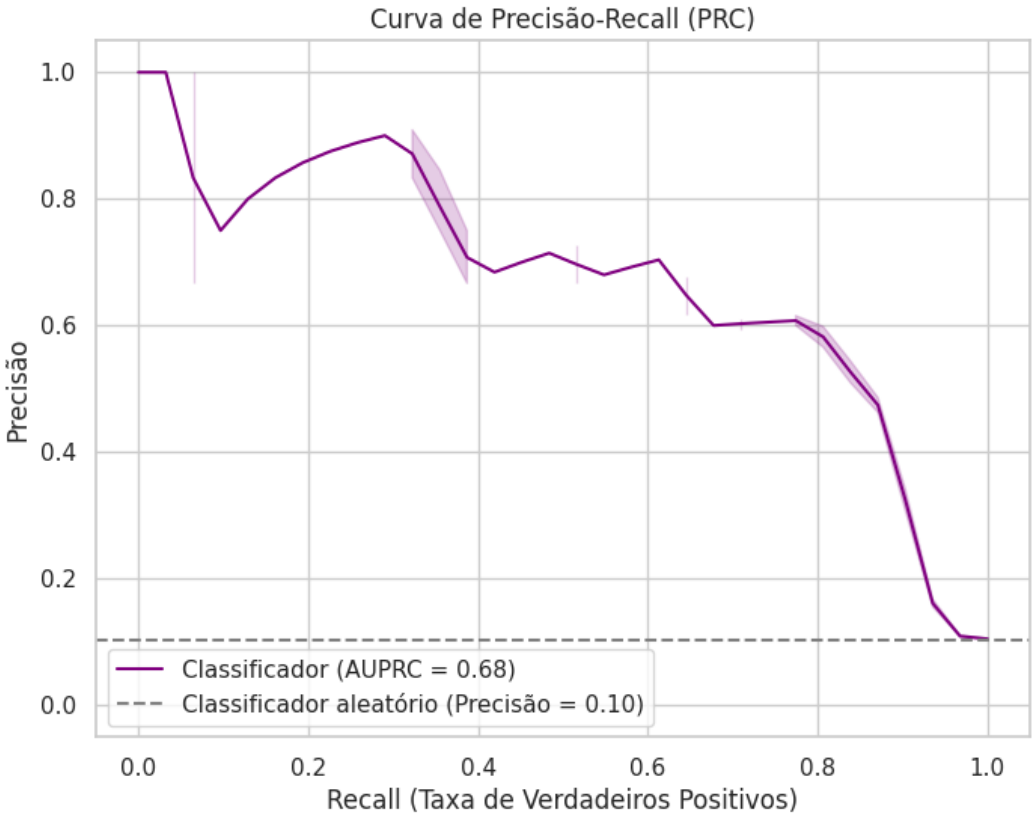
2.3.2 Área sob a Curva de *Precisão-Recall* (AUPRC)

Apesar da ampla utilização da Curva ROC e da AUROC como métricas de avaliação, em cenários de classificação com dados altamente desbalanceados, onde a classe de interesse (positiva) é significativamente minoritária, a Curva de *Precisão-Recall* (PRC) e sua área (AUPRC) oferecem uma perspectiva mais informativa e sensível ao desempenho do classificador (DAVIS; GOADRICH, 2006; SAITO; REHMSMEIER, 2015).

A Curva de *Precisão-Recall* é construída plotando-se a Precisão no eixo Y contra o *Recall* no eixo X. Um exemplo ilustrativo dessa construção, representando o desempenho de um classificador, pode ser visualizado na Figura 2.

Em contraste com a AUROC, que avalia o desempenho geral do classificador em todas as distribuições de custo, a AUPRC foca diretamente na capacidade do modelo identificar corretamente a classe positiva (minoritária). Em cenários com conjuntos de dados desbalanceados, a Curva ROC pode apresentar uma performance aparentemente inflacionada. Isso ocorre porque o grande número de Verdadeiros Negativos, provenientes da classe majoritária, tende a manter o eixo X baixo, mesmo que o modelo gere um número considerável de Falsos Positivos absolutos na classe minoritária. A

Figura 2 – Exemplo de Curva de *Precisão-Recall* (PRC). O gráfico apresenta o desempenho de um classificador (curva roxa, AUPRC = 0.68) em comparação com a linha de base de um classificador aleatório (linha tracejada cinza, Precisão = 0.10), que representa a proporção da classe positiva.



Fonte: Elaborado pela autora.

AUPRC, por outro lado, é extremamente sensível a Falsos Positivos, pois estes impactam diretamente a Precisão (eixo Y da curva PR), tornando-a uma métrica mais confiável para avaliar o desempenho em classes raras (DAVIS; GOADRICH, 2006)(DAVIS; GOADRICH, 2006)

Uma AUPRC mais alta indica que o modelo é capaz de alcançar um alto *recall* (identificando a maioria dos positivos) enquanto mantém uma precisão elevada (evitando muitos falsos positivos). A linha de base para um classificador aleatório nessa métrica é a proporção da classe positiva no conjunto de dados, e não 0.5 como na Curva ROC (SAITO; REHMSMEIER, 2015). Devido à sua sensibilidade a mudanças na classe minoritária e sua capacidade de identificar corretamente os positivos sem introduzir muitos falsos alarmes, a AUPRC é frequentemente recomendada como uma métrica de avaliação superior em problemas de classificação com dados desbalanceados.

2.4 Detecção de Anomalias como Abordagem Complementar

Em cenários onde a classe de interesse é extremamente rara e os padrões das "anomalias"são complexos, as técnicas de detecção de anomalias oferecem uma abordagem valiosa. Diferente da classificação tradicional, elas focam em identificar instâncias que se desviam significativamente do comportamento esperado da maioria dos dados. Isso é particularmente útil em problemas de saúde

pública, como a identificação de casos atípicos ou eventos raros que podem ser precursores de situações graves.

No presente trabalho de conclusão de curso serão abordadas as técnicas *Isolation Forest*, *Local Outlier Factor* e *One-Class SVM*.

2.4.1 Isolation Forest (iForest)

O *Isolation Forest* (*iForest*) é um algoritmo destacado para detecção de anomalias, uma tarefa analítica crítica com vasta aplicação em diferentes áreas, entre as quais estão o diagnóstico médico e a identificação de comportamentos inesperados em dados (CHENG; ZOU; DONG, 2019). Sua concepção difere de métodos tradicionais baseados em distância ou densidade, focando na facilidade com que uma instância pode ser isolada do restante do conjunto de dados (LIU; TING; ZHOU, 2012).

A essência do algoritmo reside na construção de uma coleção de Árvores de Isolamento (*Isolation Trees*). Cada árvore é construída recursivamente particionando-se o conjunto de dados. Em cada etapa da partição, um atributo é selecionado aleatoriamente, e um valor de divisão (*split value*) é escolhido aleatoriamente entre os valores mínimo e máximo daquele atributo no subconjunto de dados atual. Esse processo continua até que uma instância seja isolada (atingindo um nó folha) ou uma profundidade máxima seja alcançada.

O princípio fundamental é que anomalias, sendo pontos atípicos, geralmente exigem menos partições para serem isoladas em uma árvore. Isso se reflete em um comprimento de caminho mais curto da raiz até o nó terminal em comparação com pontos normais, que requerem mais partições para serem separados dos vizinhos densos. Para aumentar a robustez e mitigar a aleatoriedade de uma única árvore, o *iForest* constrói uma *ensemble* de várias Árvores de Isolamento, e a pontuação de anomalia de uma instância é derivada da média dos comprimentos obtidos em todas as árvores.

A pontuação de anomalia $s(x, n)$ para uma instância x em um conjunto de n amostras é calculada utilizando-se a seguinte fórmula, que normaliza o comprimento médio do caminho encontrado para gerar uma pontuação entre 0 e 1:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{6}$$

Onde $E(h(x))$ representa o comprimento médio do caminho da instância x na floresta de árvores. O termo $c(n)$ é um fator de normalização que serve para escalar a pontuação de anomalia. Ele é calculado com base no tamanho da amostra e representa o comprimento médio do caminho para um ponto em uma árvore de isolamento aleatória com *n* instâncias. Seu cálculo é expresso pela equação abaixo:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \tag{7}$$

Com $H(i)$ sendo o número harmônico. Este fator garante que a pontuação de anomalia seja consistente e comparável, independentemente do número de amostras envolvidas na construção das árvores. Baseada na Equação 6, quanto mais próximo de 1 for o resultado obtido para a pontuação

de anomalia, maior a probabilidade de a instância *X* ser considerada uma anomalia, enquanto valores próximos de 0 sugerem que a instância é normal.

O *iForest* se destaca na identificação de *outliers* globais devido à sua baixa complexidade de tempo e alta precisão, tornando-o eficiente para conjuntos de dados em larga escala (LIU; TING; ZHOU, 2012). Além disso, ele é capaz de detectar não apenas anomalias dispersas, mas também aquelas que estão sendo vizinhas de pontos normais. No entanto, sua eficácia pode diminuir ao lidar com *outliers* estritamente locais, o que aponta para a importância de metodologias que complementem essa característica.

2.4.2 Local Outlier Factor (LOF)

Diferentemente de métodos que identificam anomalias com base em características globais do conjunto de dados, o *Local Outlier Factor* (LOF) é um algoritmo de detecção de anomalias baseado em densidade, projetado especificamente para identificar *outliers* locais (BREUNIG et al., 2000). Um *outlier* local é uma instância que é anômala em relação à sua vizinhança imediata, mesmo que possa parecer "normal"no contexto global do dataset.

A ideia central do LOF é comparar a densidade de uma instância com a densidade de seus vizinhos. Instâncias em regiões de baixa densidade em comparação com a densidade de seus vizinhos são consideradas *outliers* locais. Para isso, o LOF define os seguintes conceitos principais:

- *Tamanho da Vizinhança:* Para cada ponto, o LOF considera seus 'k' vizinhos mais próximos, estabelecendo uma vizinhança. A distância até o mais afastado entre esses 'k' vizinhos define o raio dessa vizinhança.
- *Distância de Alcance (Reachability Distance):* Em vez de usar apenas a distância direta entre pontos, o LOF utiliza uma "distância de alcance". Este é o maior de dois valores: a distância direta entre dois pontos ou a distância máxima dentro da vizinhança. Isso suaviza o impacto de pequenas flutuações de distância.
- *Densidade Local (Local Reachability Density(LRD):* O LOF estima a densidade de cada ponto observando o quão integralmente seus vizinhos estão agrupados. Ele calcula a Densidade Local, que é o inverso da média da distância de alcance para seus vizinhos. Uma LRD alta significa que o ponto está em uma área densa; uma LRD baixa significa que está em uma área esparsa.

Com esses conceitos calcula-se o LOF para cada ponto, comparando a Densidade Local de um ponto com as de seus vizinhos. Uma pontuação LOF de aproximadamente 1 indica que a densidade do ponto p é semelhante à densidade de seus vizinhos, sugerindo que p é um ponto normal. Uma pontuação LOF significativamente maior que 1 indica que o ponto p é menos denso do que seus vizinhos, classificando-o como um *outlier* local.

A principal vantagem do LOF é sua capacidade de detectar *outliers* em conjuntos de dados com densidades variadas. No entanto, sua desvantagem reside no seu alto custo computacional, especialmente para grandes conjuntos de dados, uma vez que requer o cálculo de distâncias e a

determinação de vizinhança para cada ponto (CHENG; ZOU; DONG, 2019). Isso o torna menos adequado para análise em larga escala sem abordagens de otimização, como as que utilizam *micro-clusters* (POKRAJAC; LAZAREVIC; LATECKI, 2007).

2.4.3 One-Class SVM (OC-SVM)

O *One-Class Support Vector Machine* (OC-SVM), implementado em bibliotecas de AM como a `sklearn.svm.OneClassSVM`, oferece uma abordagem poderosa para a detecção de anomalias. É particularmente relevante em cenários onde apenas a classe "normal" é bem definida e seus dados superam significativamente os dados anormais, como em aplicações de detecção de intrusões em sistemas de segurança (PIMENTEL et al., 2014). Este algoritmo se diferencia dos SVMs tradicionais, que buscam separar duas categorias distintas, ao focar na modelagem da distribuição da classe majoritária, ou seja, o comportamento considerado normal (SCHÖLKOPF et al., 2001).

A ideia central do OC-SVM é aprender a forma da classe majoritária. Para isso, o método transforma o espaço de características dos dados, para encontrar um hiperplano ótimo que separe o conjunto de dados normais da origem em um espaço de alta dimensionalidade. A origem é tratada como o único membro de uma "segunda classe" para fins de separação(SCHÖLKOPF et al., 2001). A função de *kernel* é fundamental nesse processo, pois permite que o algoritmo capture limites complexos e não lineares, definindo uma fronteira precisa ao redor dos pontos de dados normais. Pontos que caem fora dessa fronteira são, então, classificados como anomalias. A eficácia do OC-SVM reside na determinação de um limiar de distância apropriado a partir do hiperplano, além do qual os pontos são classificados como anômalos.

Essa técnica é notavelmente mais sensível na detecção de anomalias, sendo capaz de identificar desvios sutis que podem não ser óbvios para outros métodos (PIMENTEL et al., 2014). Essa sensibilidade constitui uma vantagem particular, pois permite o reconhecimento de padrões novos ou comportamentos inesperados sem a necessidade de conhecimento prévio de suas características específicas, diferentemente de abordagens que dependem de um catálogo predefinido de anomalias. A capacidade deste método de detectar ocorrências desconhecidas que se afastam do comportamento normal o estabelece como uma ferramenta valiosa para aprimorar a confiabilidade e a integridade de diversos sistemas e dados.

Apesar de suas vantagens, o OC-SVM apresenta desafios em termos de escalabilidade. Para datasets muito grandes, o treinamento pode se tornar computacionalmente intensivo, especialmente quando são utilizados *kernels* não lineares. Embora *kernels* lineares possam mitigar parte desse problema por serem mais eficientes, eles podem não ser adequados para capturar as complexidades não lineares presentes em muitas distribuições de dados. A pesquisa contínua busca aprimorar o desempenho e a aplicabilidade dessa técnica em cenários desafiadores, abordando suas limitações e propondo melhorias para sua eficácia na detecção de anomalias (LI et al., 2003).

2.5 Revisão Bibliográfica

Nesta seção, apresenta-se uma revisão da literatura relacionada ao tema deste trabalho de conclusão, com o objetivo de contextualizar o problema de mortalidade por dengue no Brasil e explorar abordagens baseadas em Aprendizado de Máquina aplicadas à área da saúde, com enfoque posterior na predição de Dengue. Essa fundamentação visa sustentar a relevância e a viabilidade do uso de técnicas de AM na predição da mortalidade por dengue.

2.5.1 Panorama da Dengue no Contexto Brasileiro

A dengue é uma arbovirose causada por vírus do gênero *Flavivirus*, transmitida pelo mosquito *Aedes aegypti*, cuja distribuição tem se expandido globalmente nas últimas décadas, especialmente em regiões tropicais, como o Brasil. Quatro sorotipos distintos do vírus da dengue (*DENV-1*, *DENV-2*, *DENV-3* e *DENV-4*) circulam no país. A infecção por um desses sorotipos confere imunidade apenas contra ele, de modo que uma infecção posterior por um sorotipo diferente pode aumentar o risco de formas graves da doença. As maiores concentrações de casos da doença ocorrem, especialmente, em áreas urbanizadas. Estima-se que milhões de infecções por ano ocorram globalmente, com um número significativo de casos graves e óbitos (ORGANIZATION et al., 2009). No, Brasil, essa expansão está ligada a fatores como mudanças climáticas, mutações virais, crescimento populacional e, sobretudo, à urbanização acelerada e desordenada. Tal cenário resulta em ocupações precárias com ausência de saneamento básico e degradação ambiental — favorecendo a reprodução do vetor (JACOBI, 2004).

No Brasil, o contexto urbano tem papel decisivo na configuração da doença. Enquanto no início do século *XX* o controle do vetor era favorecido por cidades menos densas e com menor geração de resíduos, hoje cerca de 87% da população vive em centros urbanos (IBGE, 2022), onde o acúmulo de lixo, a desigualdade social e a precariedade da infraestrutura contribuem para a manutenção do ciclo de transmissão. A fragilidade das políticas públicas - marcada pela burocracia, escassez de recursos e descontinuidade nas ações - dificulta o controle efetivo da doença, um desafio acentuado pela complexidade socioambiental do país, evidenciando a necessidade de integração entre políticas de saúde, meio ambiente e planejamento urbano.

O espectro clínico da dengue varia de casos assintomáticos ou leves a formas graves que podem levar à morte. A Organização Mundial da Saúde (OMS) classifica a dengue em três categorias: dengue sem sinais de alarme, dengue com sinais de alarme e dengue grave. A dengue grave, também conhecida como dengue hemorrágica, é caracterizada por extravasamento de plasma, sangramentos graves ou comprometimento grave de órgãos, como fígado, coração ou cérebro, e é a forma mais letal da doença (ORGANIZATION et al., 2009).

A identificação de fatores de risco associados à progressão para formas graves e óbito é fundamental para a vigilância e o manejo clínico. A complexidade da fisiopatologia da dengue e a interação de diversos fatores individuais e clínicos do paciente são algumas das características que influenciam a gravidade da doença. Dessa forma, a capacidade de identificar e priorizar esses atributos de risco de forma precoce, a partir dos dados disponíveis, é crucial para intervenções que possam reduzir a mortalidade.

2.5.2 Aprendizado de Máquina aplicado na área da Saúde

Como introduzido anteriormente, o Aprendizado de Máquina (AM) é um subcampo da IA que desenvolve algoritmos capazes de aprender padrões a partir de dados, sem necessidade de programação explícita para cada tarefa. Esses algoritmos têm se destacado como uma ferramenta promissora na área da saúde, transformando a análise e a utilização de dados clínicos, biológicos e epidemiológicos. Sua capacidade de identificar padrões complexos, realizar previsões acuradas e auxiliar na tomada de decisões tem impulsionado avanços significativos. Segundo Cabrera et al (CABRERA et al., 2022), os algoritmos auxiliam desde o diagnóstico precoce até a gestão eficiente de recursos, contribuindo para a melhoria da assistência ao paciente.

O artigo de Alves (ALVES et al., 2021) discute diversas aplicações do AM no contexto da saúde pública, destacando-se a previsão de surtos de doenças infecciosas, a identificação de fatores de risco para condições crônicas e o suporte à alocação eficiente de recursos hospitalares. Os autores evidenciam como o uso de técnicas de AM na análise de dados epidemiológicos tem contribuído para a detecção de padrões de disseminação de doenças, possibilitando medidas preventivas mais direcionadas e eficazes. Além disso, a aplicação desses algoritmos à dados clínicos individuais pode auxiliar na identificação precoce de pacientes com maior risco de desenvolver complicações ou evoluir para óbito.

De forma semelhante, Silveira e Moreira (SILVEIRA; MOREIRA, 2020) investigam o potencial de modelos preditivos baseados em AM para apoiar decisões em saúde pública, especialmente em cenários com recursos limitados. Nesse estudo, os autores demonstram que a utilização de dados históricos e variáveis socioambientais permite antever comportamentos epidêmicos, o que reforça a relevância do AM como ferramenta estratégica no planejamento e na resposta a crises sanitárias.

Apesar do crescente interesse e das aplicações promissoras, alguns estudos como os de Mendes et al (MENDES R. F., 2023) e Groppo (GROPPO M., 2023) abordam questões como a qualidade dos dados disponíveis, a representatividade das amostras e a interpretabilidade dos modelos gerados, que são aspectos cruciais para garantir a confiabilidade e a utilidade das predições em contextos clínicos e epidemiológicos.

2.5.3 Aprendizado de Máquina aplicado à Dengue

A aplicação de técnicas de AM tem se mostrado especialmente promissora no enfrentamento da dengue, tanto na previsão de sua incidência quanto no apoio ao diagnóstico e prognóstico da doença. A seguir, são apresentados estudos relevantes que demonstram essa aplicabilidade em diferentes frentes, destacando também como a presente monografia pode contribuir para ampliar esse campo.

Um estudo de destaque é o de Roster (ROSTER, 2022), que avaliou o potencial de modelos de AM para prever casos mensais de dengue em mais de 200 cidades brasileiras. O estudo comparou diferentes algoritmos — incluindo *ensembles* de árvores de decisão, redes neurais e regressão por vetores de suporte — e concluiu que o modelo *Random Forest* teve o melhor desempenho geral, superando inclusive uma linha de base sazonal ingênua. A análise também investigou o papel de variáveis meteorológicas e epidemiológicas, revelando que, embora variáveis climáticas contribuam em algumas cidades, os registros de casos anteriores de dengue foram consistentemente os preditores mais

relevantes. Este estudo reforça a robustez do uso de AM na vigilância epidemiológica e indica que modelos com entrada limitada de dados, mas bem ajustados, podem ser eficientes e escaláveis.

Apesar da abrangência do trabalho de Roster, identifica-se uma lacuna na literatura quanto à predição de desfechos mais severos. Neste sentido, o presente trabalho de conclusão busca avançar ao aplicar e comparar diferentes classificadores de AM especificamente na predição do risco de mortalidade por dengue, um aspecto que ainda carece de maior exploração. Adicionalmente, esta pesquisa considera a importância da interpretabilidade do modelo e o uso de algoritmos adaptados às características dos dados brasileiros — um ponto crítico evidenciado no estudo de Roster, que ressaltou a variação significativa de desempenho entre diferentes contextos geográficos.

No campo do diagnóstico clínico, Bohm et al. (BOHM et al., 2024) investigaram o uso de AM para triagem de casos suspeitos de dengue a partir de dados do Sistema de Informação de Agravos de Notificação (SINAN) em Minas Gerais e Rio de Janeiro. Após selecionar um conjunto reduzido de variáveis clínicas (como febre, mialgia e idade), os autores aplicaram diversos algoritmos e observaram acurácias superiores a 90%, com destaque para a árvore de decisão e o *Perceptron Multicamadas (MLP)*. A simplicidade e interpretabilidade da árvore de decisão foram apontadas como fatores decisivos para sua aplicação em contextos de baixa infraestrutura. Contudo, limitações como o uso de dados de apenas dois municípios e a dependência de registros secundários indicam oportunidades para estudos mais abrangentes.

A presente monografia dialoga com essa abordagem ao também utilizar dados do SINAN, mas amplia significativamente o escopo ao considerar registros provenientes de todo o território brasileiro. Além disso, enquanto o foco de Bohm et al. está na triagem inicial de casos, a proposta aqui apresentada se volta à predição da mortalidade, o que representa uma perspectiva complementar e estratégica para o apoio à tomada de decisão em saúde pública, especialmente na identificação de populações com maior risco.

Um campo emergente, embora fora do escopo direto deste trabalho, é o uso de dados genômicos para prognóstico da gravidade da doença, como demonstrado por Davi et al. (DAVI et al., 2019). Nesse estudo, os autores aplicaram o algoritmo SVM para selecionar polimorfismos de nucleotídeo único (SNPs, do inglês *Single Nucleotide Polymorphisms*) e Redes Neurais Artificiais (RNAs) para a classificação, alcançando acurácia superior a 86% na previsão de evolução para dengue grave. Apesar de exigir recursos laboratoriais especializados, essa linha de pesquisa reforça o potencial do aprendizado de máquina em contextos clínicos mais amplos e pode futuramente ser integrada a abordagens híbridas que combinem dados clínicos e moleculares.

De forma geral, os estudos analisados mostram que o Aprendizado de Máquina oferece uma gama de possibilidades no combate à dengue, desde modelos preditivos epidemiológicos até ferramentas diagnósticas e prognósticas. O presente trabalho busca contribuir para esse corpo de conhecimento ao focar na predição de mortalidade por dengue utilizando dados tabulares do SINAN, explorando múltiplos classificadores e técnicas de seleção de atributos, de modo a propor modelos otimizados, acessíveis e reprodutíveis. Ao integrar variáveis contextuais, epidemiológicas e demográficas, espera-se identificar padrões capazes de subsidiar a tomada de decisão em saúde pública e contribuir com estratégias de intervenção mais eficazes.

3 Metodologia

A metodologia empregada neste trabalho de conclusão de curso foi delineada para abordar o desafio de predição de óbito em casos de dengue por meio de diferentes abordagens de Aprendizado de Máquina (AM).

Inicialmente, foi conduzido um pré-processamento comum a todas as abordagens, que incluiu a preparação das bases de dados, a verificação de duplicatas e a padronização de variáveis. Após essa etapa geral, cada abordagem foi estruturada de forma independente, com suas próprias estratégias de divisão de dados, tratamento de valores ausentes, balanceamento de classes, construção e ajuste de modelos, além da etapa de avaliação. Dessa forma, os detalhes específicos de cada abordagem são apresentados individualmente nas seções subsequentes.

3.1 Pré-processamento dos Dados

Esta etapa inicial é fundamental para preparar os dados brutos para a modelagem, garantindo sua qualidade, consistência e adequação aos algoritmos de Aprendizado de Máquina. Envolve a descrição dos dados, modificações estruturais, tratamento de duplicatas, padronização de características e gerenciamento de valores faltantes.

Este estudo utiliza dados referentes a casos de dengue notificados no Brasil, disponibilizados publicamente pelo SINAN. As bases de dados originais referem-se aos anos de 2023 e 2024, apresentando um volume substancial de registros.

3.1.1 Características Iniciais da Base de Dados

Ambas as bases de dados anuais, tanto a referente a 2023 quanto a base de 2024, inicialmente possuíam aproximadamente 1,5 milhão de registros com 120 atributos. Devido ao alto volume de dados, a manipulação e o processamento completo dos *datasets* seriam computacionalmente custosos. Para otimizar os recursos computacionais sem comprometer a representatividade dos dados, foram realizadas as seguintes modificações:

- Criação da variável alvo `Obito`: Derivada de `DT_OBITO`, onde datas preenchidas indicam 1 (óbito) e valores nulos indicam 0 (não-óbito).
- Cálculo de `QUANT_DIAS`: Diferença em dias entre `DT_NOTIFIC` e `DT_SIN_PRI`, indicando o tempo da evolução da doença.
- Redução do volume de registros: Cada base foi reduzida para 500.000 registros, de forma estratificada, preservando a proporção da classe alvo e mantendo integralmente a classe minoritária.

- Remoção de atributos: Foram excluídos os atributos completamente vazios, com mais de 50% de valores ausentes, irrelevantes ou redundantes (como `ID_MUNICIP`, `NU_ANO`, atributos específicos de outras arboviroses, e `EVOLUCAO` por redundância com a variável alvo).
- Ajuste de `NU_IDADE_N`: Remoção do prefixo numérico "40" para isolar a idade real do paciente.

Para uma descrição mais detalhada das bases de dados, incluindo a porcentagem de valores faltantes por atributo e outras estatísticas descritivas de cada base, o leitor pode consultar o **Apêndice I**.

As bases de dados resultantes, com tamanho reduzido e variáveis ajustadas, serviram como ponto de partida para as etapas subsequentes. A seleção final dos atributos foi realizada manualmente pelo pesquisador, considerando sua relevância para a predição de óbitos por dengue, a temática do estudo e a qualidade dos dados. A lista completa dos atributos mantidos é apresentada na Tabela 2.

Tabela 2 – Atributos restantes após o pré-processamento das bases de dados.

Atributos	**Descrição resumida**
SEM_NOT	Semana epidemiológica da notificação
SG_UF_NOT	Unidade federativa da notificação
NU_IDADE_N	Idade do paciente
CS_SEXO	Sexo
CS_GESTANT	Gestação
CS_RACA	Raça/cor
FEBRE	Presença de febre
MIALGIA	Presença de mialgia
CEFALEIA	Presença de cefaleia
EXANTEMA	Presença de exantema
VOMITO	Presença de vômito
NAUSEA	Presença de náusea
DOR_COSTAS	Dor nas costas
CONJUNTVIT	Conjuntivite
ARTRITE	Presença de artrite
ARTRALGIA	Presença de artralgia
PETEQUIA_N	Presença de petéquias
LEUCOPENIA	Presença de leucopenia
LACO	Resultado do LACOs
DOR_RETRO	Dor retro-orbital
DIABETES	Comorbidade: diabetes
HEMATOLOG	Comorbidade: doenças hematológicas
HEPATOPAT	Comorbidade: hepatopatias
RENAL	Comorbidade: doenças renais

Atributos	**Descrição resumida**
HIPERTENSA	Comorbidade: hipertensão
ACIDO_PEPT	Comorbidade: acidose péptica
AUTO_IMUNE	Comorbidade: doenças autoimunes
RESUL_SORO	Resultado sorológico
RESUL_NS1	Resultado NS1
RESUL_VI_N	Resultado virológico
RESUL_PCR_	Resultado PCR
SOROTIPO	Sorotipo do vírus
HISTOPA_N	Histórico patológico
IMUNOH_N	Resultado imunohistoquímico
HOSPITALIZ	Hospitalização
UF	Unidade federativa do paciente
CLASSI_FIN	Classificação final da dengue
CRITERIO	Critério de confirmação
DT_ENCERRA	Data de encerramento do caso
ALRM_HIPOT	Alarme: hipotensão
ALRM_PLAQ	Alarme: plaquetopenia
ALRM_VOM	Alarme: vômito persistente
ALRM_SANG	Alarme: sangramento
ALRM_HEMAT	Alarme: hematêmese
ALRM_ABDOM	Alarme: dor abdominal intensa
ALRM_LETAR	Alarme: letargia
ALRM_HEPAT	Alarme: hepatomegalia
ALRM_LIQ	Alarme: extravasamento de líquido
GRAV_PULSO	Gravidade: pulso rápido
GRAV_CONV	Gravidade: convulsões
GRAV_ENCH	Gravidade: edema
GRAV_INSUF	Gravidade: insuficiência orgânica
GRAV_TAQUI	Gravidade: taquicardia
GRAV_EXTRE	Gravidade: extremidades frias
GRAV_HIPOT	Gravidade: hipotensão
GRAV_HEMAT	Gravidade: hematomas
GRAV_MELEN	Gravidade: melena
GRAV_METRO	Gravidade: metrorragia
GRAV_SANG	Gravidade: sangramento
GRAV_AST	Gravidade: AST elevada
GRAV_MIOC	Gravidade: miocardite
GRAV_CONSC	Gravidade: alteração de consciência
GRAV_ORGAO	Gravidade: falência de órgão
QUANT_DIAS	Diferença em dias entre notificação e início dos sintomas

Atributos	**Descrição resumida**
Obito	Variável alvo: óbito (1) ou não (0)

Fonte: Elaborado pela autora.

3.1.2 Verificação e tratamento de registros duplicados

A qualidade dos dados é um fator crítico para o desempenho de modelos de AM. Desta forma, após o pré-processamento inicial, as bases de dados foram submetidas a verificações adicionais para identificar e tratar registros duplicados.

Primeiramente, foi realizada uma verificação da quantidade total de registros duplicados em cada base de dados. Para a base de 2023, foram identificados 834 registros duplicados, enquanto para a base de dados de 2024 o número foi de 498.

Uma análise mais aprofundada revelou que todos os registros duplicados, em ambas as bases, correspondiam a casos de "não-óbito". Dado que a duplicação se concentrava exclusivamente na classe majoritária e não apresentava impactos sobre a classe minoritária (óbito), optou-se pela remoção desses registros. Após a exclusão, a base de dados de 2023 passou a conter aproximadamente 399,9 mil registros únicos e a base de 2024 passou a ter 305,8 mil registros.

3.1.3 Padronização de dados e Divisão dos Conjuntos

Antes de aplicar qualquer transformação, o conjunto de dados foi preparado para garantir a integridade das análises e evitar o vazamento de informações do conjunto de teste na fase de treino, levando a estimativas de desempenho excessivamente otimistas *data leakage*.

Além disso, os dados foram preparados para validação cruzada, utilizando o método *K-fold*, de forma a preservar a representatividade das classes em cada divisão e permitir uma avaliação mais robusta do desempenho dos modelos.

Para algoritmos de Aprendizado de Máquina (AM) sensíveis à escala das características, como o KNN, é crucial padronizar ou normalizar os dados numéricos, garantindo que características com maiores faixas numéricas não influenciem desproporcionalmente o modelo. A padronização foi feita através do `StandardScaler`, devido a natureza desbalanceada dos conjuntos de dados. Esse escalador transforma os dados de modo que tenham média 0 e desvio padrão 1, sendo mais robusto à *outliers* e prevenindo a compressão de dados que poderia achatar as distribuições das características. Para isso, as etapas seguidas foram:

1. Codificação de Características Categóricas: Antes do escalonamento numérico, os atributos categóricos foram convertidos em representações numéricas. Para isso, utilizou-se o `LabelEncoder`. Cada categoria única em um atributo foi associada à um inteiro e quaisquer valores ausentes nas colunas categóricas foram temporariamente transformados em *'missing'* antes da codificação, garantindo que todas as instâncias fossem processadas.

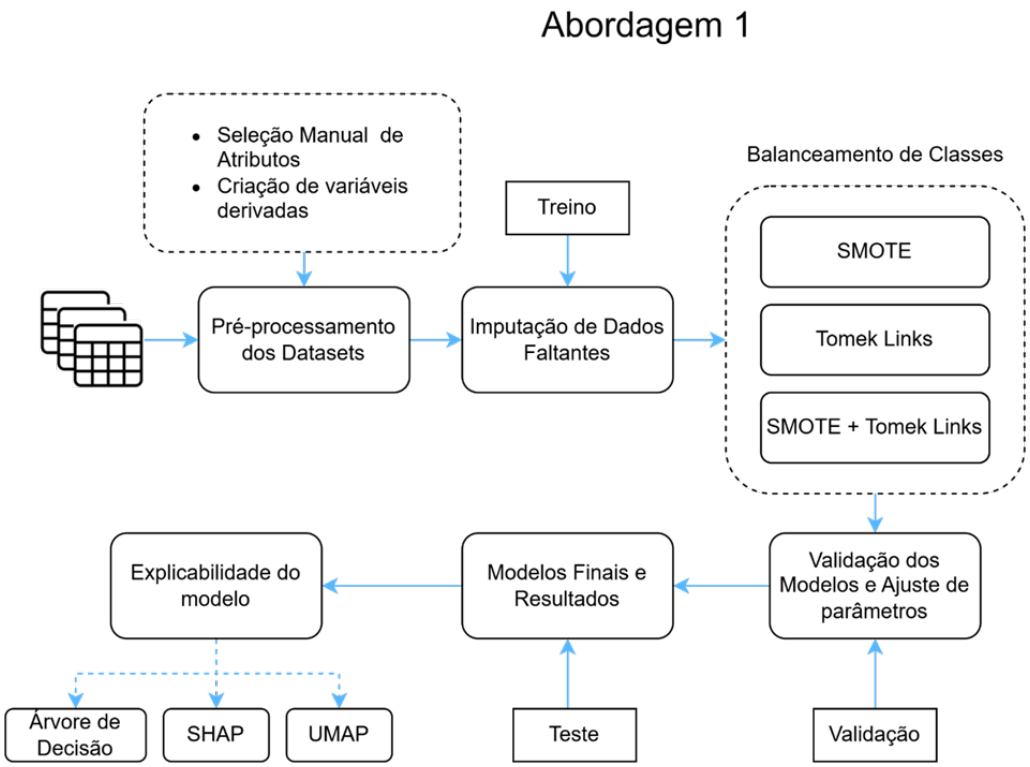
2. Padronização de Características Numéricas: O `StandardScaler` foi aplicado exclusivamente às colunas originalmente numéricas do conjunto de dados de treinamento (`fit()`). Posteriormente, a transformação (`transform()`) foi aplicada às colunas numéricas dos demais conjuntos, assegurando que todos os eles fossem escalonados utilizando os mesmos parâmetros aprendidos, mantendo a consistência dos dados.

A estratégia de divisão dos conjuntos de dados variou significativamente conforme a abordagem. Enquanto os cenários de classificação empregaram a estratificação tradicional (treino, validação e teste) para garantir a representatividade de ambas as classes, a abordagem de detecção de anomalias exigiu uma metodologia distinta. Para esta última, o conjunto de treinamento foi composto exclusivamente pela classe majoritária, isolando a classe minoritária para avaliação. Os detalhes específicos e as proporções de cada divisão são apresentados nas seções metodológicas correspondentes.

3.2 Abordagem 1: Algoritmos de Aprendizado Supervisionado para Classificação de Óbitos

Para a abordagem de AM supervisionado, o fluxo do trabalho seguiu o mostrado pelo diagrama 3.

Figura 3 – Fluxo Metodológico para Abordagem 1.



Fonte: Elaborado pela autora.

Essa estratégia teve os conjuntos de dados separados em treino, validação e teste. Para os conjuntos de treino utilizou-se a validação cruzada (*K-Fold*), com *K=5*, garantindo uma avaliação

robusta do desempenho dos modelos, uma vez que dentro são criados 5 separações de treino com respectivos conjuntos de validação interna, de forma a evitar o viés introduzido por uma única partição aleatória de treino/validação, obtendo uma medida de desempenho mais estável e robusta.

Após a divisão dos dados, o tratamento de atributos faltantes foi aplicado aos *folds* de treinamento. Com esse objetivo foram removidos do conjunto de treino os registros com mais de 50% de atributos ausentes. Para preservar a integridade da classe minoritária (*Óbito*), registros dessa classe foram mantidos independentemente da quantidade de valores ausentes.

O preenchimento dos dados ausentes restantes foi realizada de duas formas distintas, sendo a primeira a partir de medidas estatísticas e a outra baseada no algoritmo KNN.

Imputação baseada em medidas estatísticas

Primeiramente é importante destacar que essas medidas estatísticas foram calculadas separadamente para cada classe da variável alvo (*'Óbito'* $= 0$ e *'Óbito'* $= 1$), garantindo que a imputação refletisse as características específicas de cada classe.

Para atributos numéricos, a escolha entre média ou mediana foi feita com base na assimetria (*skewness*) da distribuição dos dados, se a distribuição era aproximadamente simétrica (assimetria absoluta menor que 1), utilizou-se a média; caso contrário, a mediana foi empregada por ser mais robusta à presença de valores extremos (*outliers*).

Para atributos categóricos ou binários, os valores faltantes foram preenchidos com a moda, ou seja, o valor mais frequente, calculada exclusivamente a partir do conjunto de treino, preservando a integridade do processo e evitando *data leakage*.

Imputação com K-Nearest Neighbors (KNN)

O algoritmo *KNNImputer* foi utilizado para estimar os valores ausentes, baseando-se nos $K{=}3$ vizinhos mais próximos dentro do conjunto de treino de cada *fold*. Este valor de K foi escolhido para priorizar a localidade dos dados, garantindo que a imputação se baseasse em um grupo pequeno e representativo de registros de alta similaridade.

Após o ajuste no conjunto de treino, a transformação foi aplicada aos respectivos conjuntos de validação. É fundamental destacar que a imputação nestes conjuntos foi baseada apenas em informações aprendidas no conjunto treino. Para colunas originalmente categóricas, que foram previamente codificadas numericamente, os valores imputados foram arredondados para o inteiro mais próximo, mantendo a representação numérica utilizada pelos classificadores posteriormente.

Assim, as diferentes combinações de estratégia de remoção e imputação geraram conjuntos de dados distintos para a avaliação dos modelos de aprendizado supervisionado, o que possibilitou analisar o efeito direto de cada pré-processamento no desempenho preditivo.

3.2.1 Balanceamento de Classes

Conforme o diagrama metodológico, a etapa de balanceamento de classes é crucial para lidar com o desbalanceamento significativo da classe minoritária (óbitos). Esse desbalanceamento poderia levar os modelos a apresentarem alta acurácia apenas por favorecerem a classe majoritária, mascarando um desempenho real insatisfatório.

Para mitigar esse problema, foram aplicadas estratégias de amostragem, balanceando a proporção de classes nos conjuntos de dados de treinamento. Foram utilizadas tanto técnicas de *Oversampling* quanto de *Undersampling* aplicadas individualmente e de forma híbrida, como explicado previamente na Seção 2.1.3:

- **Oversampling:** Criação de novas amostras sintéticas para a classe minoritária, óbito, através do SMOTE, gerando instâncias sintéticas da classe minoritária (*Óbito*=1) a partir dos seus vizinhos mais próximos. Isso permitiu ampliar a representatividade da classe sem duplicar registros existentes (HE; GARCIA, 2009).
- **Undersampling:** Consistiu na remoção de instâncias da classe majoritária (*Óbito*=0) que formavam pares de vizinhos mais próximos com registros da classe minoritária. Essa estratégia reduziu sobreposição de fronteiras entre classes e simplificou a distribuição (TOMEK, 1976; BATISTA; PRATI; MONARD, 2004).
- **Métodos Híbridos/Combinados:** Inicialmente aplicou-se SMOTE para aumentar a representatividade da classe minoritária, seguido da remoção de *Tomek Links*, de modo a evitar redundância e melhorar a separabilidade entre as classes (TOMEK, 1976).

Essas técnicas foram aplicadas exclusivamente nos conjuntos de treino de cada *fold*, de forma a evitar vazamento de informação para os conjuntos de validação e teste.

3.2.2 Escolha dos Modelos

Com o objetivo de identificar os algoritmos mais eficazes para a predição de óbitos por dengue, diferentes classificadores supervisionados foram avaliados a partir de diferentes abordagens de pré-processamento e balanceamento de dados. A seleção dos modelos considerou sua relevância na literatura, diversidade de paradigmas de aprendizado e compatibilidade com os objetivos do estudo, especialmente em contextos com desbalanceamento das classes.

Foram considerados os seguintes algoritmos de classificação:

- Árvores de decisão e o *ensemble Random Forest*;
- *Gradient Boosting* (XGBoost);
- *Naive Bayes*;
- *K-Nearest Neighbors* (KNN);

- *Support Vector Machine* (SVM);
- Regressão Logística com Penalização.

Visando verificar o imapcto das técnicas de amostragem, os experimentos foram realizados tanto com os dados originais quanto com as versões balanceadas abordadas.

Para os dados originais, foi explorada a utilização do parâmetro `class_weight='balanced'` nos modelos que oferecem suporte a esse ajuste, como Regressão Logística, SVM, Árvores de Decisão e *Random Forest*. Já para os dados balanceados artificialmente, os modelos foram aplicados sem esse ajuste, de forma a avaliar o impacto direto das técnicas de balanceamento sobre o desempenho dos classificadores.

3.2.3 Explicabilidade do Modelo

A etapa final da metodologia teve como objetivo tornar os modelos de classificação mais transparentes e interpretáveis, aspecto fundamental em contextos clínicos e de saúde pública. A explicabilidade permite entender como as predições são feitas e quais atributos têm maior influência nos resultados, contribuindo tanto na validação científica quanto na tomada de decisão.

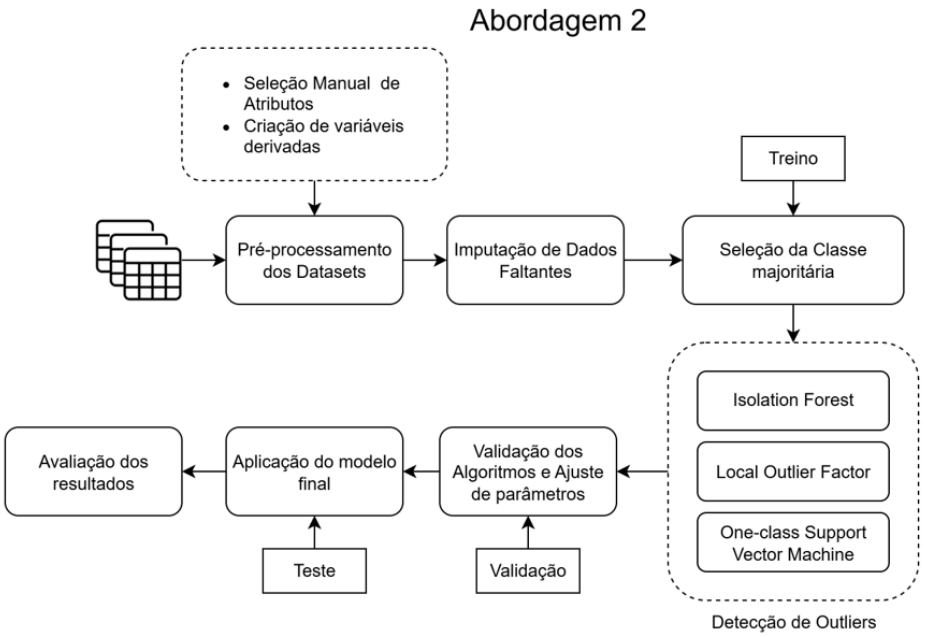
Foram abordadas as seguintes técnicas para análise de interpretabilidade e visualização:

- **SHAP (SHapley Additive exPlanations):** Quantifica a contribuição de cada atributo para predições individuais, destacando a direção e intensidade de seus impactos.
- **UMAP (Uniform Manifold Approximation and Projection):** Ferramenta de redução de dimensionalidade voltada à preservação da estrutura global dos dados, útil na identificação de agrupamentos e padrões latentes.
- **Árvore de Decisão:** Apesar de ter sido avaliada também como classificador, a Árvore de Decisão foi utilizada como ferramenta de explicabilidade intríseca. Sua estrutura hierárquica e naturalmente interpretável permite a derivação e visualização de regras de decisão claras, apresentadas pelos seus nós e caminhos, facilitando a compreensão de como combinações específicas de atributos levam a diferentes desfechos.

Essas abordagens complementares possibilitaram não apenas avaliar o desempenho preditivo dos modelos, mas também compreender os fatores subjacentes às decisões algorítmicas. A utilização conjunta de técnicas locais (como o SHAP), globais (como o UMAP) e intrinsicamente interpretáveis (como a Árvore de Decisão) fortaleceu a confiabilidade dos resultados e forneceu subsídios relevantes para a análise clínica e epidemiológica, promovendo maior transparência e potencial aplicação prática dos modelos desenvolvidos.

3.3 Abordagem 2: Algoritmos de aprendizado não supervisionado para detecção de outliers

Figura 4 – Fluxo Metodológico para Abordagem 2.



Fonte: Elaborado pela autora.

Em função da natureza extremamente desbalanceada do problema, foi desenvolvida uma segunda abordagem baseada em técnicas de detecção de anomalias, conforme ilustrado no Diagrama 4. A ideia central foi tratar os casos positivos (óbitos) como eventos raros, buscando identificá-los entre os registros da classe majoritária (sobreviventes). Essa abordagem foi estruturada nas seguintes etapas:

3.3.1 Seleção dos algoritmos de detecção de anomalias

Foram considerados três métodos para detecção de *outliers*, destacados na metodologia 4 e descritos previamente na Seção 2.4:

- *Isolation Forest (iForest)*: Baseado em árvores de decisão aleatórias, identifica observações que se isolam com poucas divisões — característica que o torna eficiente para detectar registros anômalos em grandes volumes de dados (LIU; TING; ZHOU, 2012).
- *Local Outlier Factor (LOF)*: Calcula a densidade local de cada ponto em relação aos seus vizinhos, classificando como outliers aqueles cuja densidade é significativamente mais baixa. Isso permite identificar casos raros mesmo em regiões densas da distribuição (BREUNIG et al., 2000).
- *One-Class SVM*: Treinado exclusivamente com a classe majoritária, aprende uma fronteira que delimita o comportamento considerado normal. Registros que ultrapassam essa fronteira — como os óbitos — são rotulados como anomalias (SCHÖLKOPF et al., 2001).

3.3.2 Treinamento restrito à classe majoritária

Seguindo o fluxo apresentado no Diagrama 4, todos os modelos foram treinados apenas com registros da classe negativa (sobreviventes). Dessa forma tinha-se como objetivo ensinar aos modelos apenas o padrão típico da população majoritária e focar a detecção nos desvios mais significativos, de modo que registros muito diferentes — como os casos de óbito — fossem naturalmente identificados como anomalias.

Essa análise mostra-se particularmente adequada para problemas altamente desbalanceados, pois complementa as estratégias tradicionais de classificação supervisionada, oferecendo uma perspectiva alternativa sobre os registros raros.

Na sequência, serão detalhadas as etapas de validação e as métricas utilizadas para avaliar o desempenho de ambas abordagens.

3.4 Validação e Avaliação

Para garantir o desempenho ideal de cada modelo, realizou-se a etapa de otimização de hiperparâmetros por meio do método *Grid Search*, aplicada após a seleção do melhor *fold* de treinamento em cada cenário experimental.

3.4.1 Otimização de Hiperparâmetros

A otimização de hiperparâmetros foi conduzida dentro do processo de validação cruzada *(K-Fold)*, utilizando o *F1-Score* como métrica principal de seleção. Essa métrica foi escolhida por equilibrar Precisão e *Recall*, aspecto fundamental diante do desbalanceamento existente entre as classes de interesse (óbito e não óbito).

O método *Grid Search* realizou uma busca sistemática sobre combinações pré-definidas de parâmetros, variando conforme o tipo de classificador avaliado. Por exemplo, nos modelos baseados em árvores de decisão (*Decision Tree* e *Random Forest*), os parâmetros incluíram profundidade máxima, tamanho mínimo de divisão e número de estimadores, enquanto em outros algoritmos foram explorados seus respectivos hiperparâmetros relevantes.

Dessa forma, o procedimento foi padronizado para todos os classificadores, garantindo uma comparação justa e controlada quanto ao impacto do ajuste de parâmetros sobre o desempenho final.

3.4.2 Avaliação Final

Após a otimização, os modelos com as melhores combinações de parâmetros foram retreinados e, em seguida, submetidos à validação. Essa etapa teve como objetivo estimar a capacidade real de generalização dos modelos, evitando o *overfitting* aos dados de treino.

Em todos os conjuntos de avaliação (validação e teste), as seguintes métricas foram utilizadas para mensurar o desempenho dos modelos, conforme detalhado na Seção 2.2:

- **Acurácia**
- **Precisão**
- **Recall (Sensibilidade)**
- **F1-Score**
- **Área sob a Curva ROC (AUC-ROC)**
- **Área sob a Curva Precisão-Recall (AUPRC)**

Os resultados obtidos através dessa metodologia estão apresentados no capítulo seguinte.

4 Resultados

Este capítulo apresenta e discute os resultados obtidos a partir dos modelos de classificação treinados e dos algoritmos de detecção de *outliers*. A avaliação dos modelos foi realizada com base nos dois conjuntos de dados anuais (2023 e 2024), a partir da metodologia descrita previamente. Assim, serão apresentados os resultados obtidos por cada uma das abordagens seguido de uma comparação entre ambas.

4.1 Análise dos resultados para Aprendizado Supervisionado

Conforme detalhado no Capítulo 3, os classificadores foram submetidos a um processo de ajuste de hiperparâmetros utilizando a técnica de *Grid Search* com o objetivo de maximizar o desempenho. A Tabela 3 resume os parâmetros testados para cada um dos classificadores.

Tabela 3 – Parâmetros de Otimização Utilizados no *Grid Search*.

Classificador	**Parâmetro**	**Valores Testados**
Árvore de Decisão		
	`max_depth`	{ `None`, 5, 10}
	`min_samples_split`	{2, 5, 10}
	`min_samples_leaf`	{1, 2, 4}
	`criterion`	{`gini`, `entropy`}
Random Forest		
	`n_estimators`	{100, 200, 300}
	`max_depth`	{`None`, 15, 30}
	`min_samples_split`	{2, 5, 10}
XGBoost		
	`learning_rate`	{0.01, 0.1}
	`n_estimators`	{100, 200, 300}
	`max_depth`	{3, 5, 7}
	`subsample`	{0.8, 1.0}
	`colsample_bytree`	{0.8, 1.0}
Naive Bayes (GaussianNB)		
	`var_smoothing`	$\{10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}\}$
KNN		
	`n_neighbors`	{3, 5, 7}
	`weights`	{`uniform`, `distance`}
	`P`	{1, 2}
SVM		
	`C`	{0.01, 0.1, 1, 10, 100}
	`loss`	{square_hinge}
	`penalty`	{l2}
Regressão Logística		
	`C`	{0.1, 1, 10}
	`penalty`	{`l1` ,`l2`}
	`solver`	`liblinear`

Fonte: Elaborado pela autora.

Essa otimização foi crucial para garantir que cada modelo operasse em sua configuração de melhor performance em combinação com as quatro diferentes abordagens de balanceamento: Normal (sem balanceamento), SMOTE, *Tomek Links* e Híbrido (SMOTE + *Tomek Links*).

4.1.1 Resultados obtidos para a imputação com medidas estatisticas

Nesta seção, são apresentados os resultados alcançados pelos classificadores treinados com os dados imputados por medidas estatísticas (média, mediana e moda). As Tabelas mostram o melhor desempenho alcançado por cada modelo após o processo de otimização dos hiperparâmetros descrito anteriormente.

Os resultados exibidos na Tabela 4 representam cada um dos classificadores com sua melhor configuração de balanceamento de dados (Balanc.). No ano de 2023, o modelo *Random Forest* com a configuração de balanceamento normal (sem aplicação de técnicas de amostragem) apresentou o melhor desempenho geral entre os classificadores avaliados, com o maior *F1-Score* (0.6975) e a maior AUC (0.9751), demonstrando elevada capacidade discriminativa na identificação correta dos casos positivos (óbitos), mesmo diante do desbalanceamento das classes, reforçando sua adequação para tarefas de predição de mortalidade.

Tabela 4 – Melhores Resultados de Classificação por Classificador e Configuração (2023)

Classificador	**Balanc.**	**Otimização**	**F1-Score**	**Precisão**	**Recall**	**AUC**	**AUPRC**
Árvore de Decisão	Normal	Grid Search	0.6305	0.7634	0.5379	0.8203	0.5501
Random Forest	Normal	Grid Search	**0.6975**	**0.8557**	0.5887	**0.9751**	0.6702
XGBoost	Normal	Grid Search	0.6555	0.8041	0.5532	**0.9826**	**0.6791**
Naive Bayes	SMOTE	Default	0.0911	0.0487	**0.9080**	0.7512	0.0450
KNN	Tomek	Grid Search	0.4040	0.7018	0.2837	0.7509	0.2674
SVM	SMOTE	Grid Search	0.5898	0.5050	0.7042	0.8541	0.4568
Regressão Logística	SMOTE	Grid Search	0.1807	0.1014	0.8298	0.9068	0.0845

O *XGBoost*, também em configuração normal, destacou-se pelo maior AUPRC (0.6791), evidenciando sua robustez na classificação em cenários com classes desproporcionais. Por outro lado, o algoritmo *Naive Bayes* obteve o maior *Recall* (0.9080), porém acompanhado de um *F1-Score* extremamente baixo (0.0911) e da pior precisão (0.0487), indicando uma forte tendência do modelo classificar a grande maioria das instâncias como positivas, resultando em um grande número de falsos positivos.

Os demais classificadores, como KNN, SVM e Regressão Logística, apresentaram desempenhos moderados a baixos, sugerindo que a imputação por estatísticas não foi suficiente para extrair as características necessárias para esses algoritmos, os quais são mais sensíveis à distribuição das variáveis.

Os resultados referentes ao ano de 2024 (Tabela 5) mantêm a tendência observada no ano anterior, com os classificadores mais robustos apresentando melhor desempenho geral, mas com uma degradação nos indicadores de performance em comparação a 2023. O *ensemble Random Forest*

permaneceu como o mais eficiente, atingindo o maior *F1-Score* (0.6154), ainda que com uma redução de aproximadamente 8 pontos percentuais em relação ao ano anterior.

Tabela 5 – Melhores Resultados de Classificação por Classificador e Configuração (2024)

Classificador	**Balanc.**	**Otimização**	**F1-Score**	**Precisão**	**Recall**	**AUC**	**AUPRC**
Árvore de Decisão	Tomek	Grid Search	0.5630	0.8444	0.4222	0.9288	0.4620
Random Forest	Normal	Grid Search	**0.6154**	0.8302	0.4889	0.9508	0.6028
XGBoost	SMOTE	Grid Search	0.6040	0.7627	0.5000	**0.9813**	**0.6622**
Naive Bayes	Híbrido	Grid Search	0.4203	0.6042	0.3222	0.7493	0.3525
KNN	SMOTE	Grid Search	0.2055	0.1293	0.5000	0.7856	0.0965
SVM	Tomek	Grid Search	0.4407	**0.9286**	0.2889	0.9703	0.5031
Regressão Logística	SMOTE	Grid Search	0.1137	0.0609	**0.8556**	0.9164	0.0523

A Regressão Logística e o *Naive Bayes* continuaram apresentando os maiores valores de *recall* (respectivamente 0.8556 e 0.3222), reforçando a dificuldade dos algoritmos em balancear a detecção de instâncias positivas (*recall* alto) com a correta identificação das negativas (precisão baixa).

Ao comparar os melhores resultados globais - considerando *F1-Score* e AUPRC - observa-se que os modelos de 2023 superaram os de 2024 em todos os aspectos. Essa diferença de desempenho sugere que as características intrínsecas dos dados coletados em 2023 favoreceram a separabilidade de classes, enquanto que os dados de 2024 apresentaram maior complexidade e desbalanceamento, o que limitou o ganho de desempenho mesmo com a aplicação da imputação estatística e dos métodos de balanceamento.

Em suma, na imputação por medidas estatísticas, os classificadores baseados em árvores de decisão (*Random Forest* e *XGBoost*) mostraram-se as abordagens mais robustas, sobretudo quando empregados sem técnicas de balanceamento e otimizados via *Grid Search*.

4.1.2 Resultados obtidos para os dados imputados por KNN

A imputação de valores ausentes pelo *K-Nearest Neighbors*(KNN) buscou preservar relações locais entre as instâncias, partindo da premissa de que amostras semelhantes possuem valores próximos para as variáveis faltantes.

Contudo, os resultados evidenciaram um desempenho insatisfatório na maioria dos classificadores, com baixos valores para *F1-Score*, precisão e *recall*. Isso indica que a imputação por KNN não conseguiu reconstruir adequadamente as distribuições originais dos dados, possivelmente introduzindo ruídos e dificultando o aprendizado das fronteiras de decisão, especialmente para a classe minoritária (óbitos).

No conjunto de dados de 2023, o classificador *Naive Bayes* obteve o melhor resultado, alcançando um *F1-Score* de 0.1482, com precisão de 0.0820 e *recall* de 0.7730. Esse comportamento

demonstra que o modelo foi capaz de identificar uma parcela considerável dos casos positivos, porém à custas de uma grande quantidade de falsos positivos, refletindo o baixo valor de precisão.

Por outro lado, o *XGBoost* apresentou os maiores valores de AUC (0.8402) e AUPRC (0.4483), indicando que, embora não tenha gerado predições positivas efetivas na classe minoritária, ainda manteve boa capacidade de distinção entre as classes.

Tabela 6 – Melhores Resultados de Classificação por Classificador e Configuração (2023) - Imputado KNN

Classificador	**Balanc.**	**Otimização**	**F1-Score**	**Precisão**	**Recall**	**AUC**	**AUPRC**
Árvore de Decisão	Tomek	Grid Search	0.0120	0.0400	0.0071	0.4969	0.0031
Random Forest	Tomek	Grid Search	0.0000	0.0000	0.0000	0.6805	0.0103
XGBoost	Normal	Grid Search	0.0000	0.0000	0.0000	**0.8402**	**0.4483**
Naive Bayes	Tomek	Grid Search	**0.1482**	**0.0820**	**0.7730**	0.8140	0.0640
KNN	SMOTE	Grid Search	0.0000	0.0000	0.0000	0.5000	0.0022
SVM	Normal	Grid Search	0.0081	0.0041	0.2979	0.5691	0.0028
Regressão Logística	Híbrido	Grid Search	0.0017	0.0009	0.1631	0.1839	0.0189

Fonte: Elaborado pela autora.

Para o ano de 2024, os resultados foram ainda mais limitados, com métricas próximas de zero em praticamente todos os classificadores. O SVM apresentou o melhor *F1-Score* (0.0017) entre os modelos testados, enquanto o *XGBoost* manteve desempenho superior em AUC (0.6805), sugerindo certa estabilidade na distinção entre as classes, mas sem traduzir essa discriminação em predição positiva correta.

Tabela 7 – Melhores Resultados de Classificação por Classificador e Configuração (2024) - Imputado KNN

Classificador	**Balanc.**	**Otimização**	**F1-Score**	**Precisão**	**Recall**	**AUC**	**AUPRC**
Árvore de Decisão	SMOTE	Grid Search	0.0000	0.0000	0.0000	0.5065	0.0018
Random Forest	Normal	Grid Search	0.0000	0.0000	0.0000	0.4152	0.0014
XGBoost	Tomek	Grid Search	0.0000	0.0000	0.0000	**0.6805**	**0.0103**
Naive Bayes	Híbrido	Grid Search	**0.0006**	**0.0003**	**0.1667**	0.0752	0.0010
KNN	Normal	Grid Search	0.0000	0.0000	0.0000	0.5000	0.0017
SVM	Normal	Grid Search	0.0017	0.0009	0.1000	0.4500	0.0016
Regressão Logística	SMOTE	Grid Search	0.0005	0.0002	0.0556	0.0768	0.0011

Fonte: Elaborado pela autora.

De maneira geral, os resultados reforçam que a imputação baseada em KNN não favoreceu o aprendizado dos modelos. O método parece ter preservado variações locais irrelevantes e amplificado

o ruído, dificultando a generalização e a detecção confiável da classe minoritária.

4.1.3 Comparando as técnicas de imputação de dados faltantes

A análise comparativa entre as estratégias de imputação evidenciou diferenças significativas no desempenho dos classificadores avaliados. De forma ampla, a imputação baseada em medidas estatísticas apresentou maior estabilidade, com desempenho superior nas principais métricas de avaliação, como *F1-Score* e *Recall*. Esses resultados sugerem que o preenchimento por estatísticas descritivas teve maior impacto sobre modelos mais sensíveis à variabilidade dos dados, como o *Naive Bayes* e a Regressão Logística, contribuindo para uma melhor identificação da classe minoritária.

Essa sensibilidade pode ser explicada pelo fato de que esses classificadores dependem de pressupostos estatísticos — como independência entre variáveis e a linearidade entre os preditores e a variável resposta. Assim, alterações na distribuição ou na escala dos atributos podem afetar significativamente sua capacidade de generalização. Em contrapartida, modelos baseados em árvores de decisão, como o *Random Forest* e o *XGBoost*, mostraram-se mais robustos a essas variações, devido à sua natureza não paramétrica e a independência de pressupostos sobre a forma funcional dos dados.

Em contraste, a imputação realizada por meio do algoritmo *K-Nearest Neighbors* (KNN), embora metodologicamente mais sofisticada por incorporar relações de similaridade entre as amostras, apresentou desempenho global inferior. Tal resultado pode ser atribuído à alta dimensionalidade dos *datasets* e à presença de ruído nos registros, fatores que dificultam a definição de vizinhanças representativas e comprometem a reconstrução das distribuições originais das variáveis ausentes.

Ainda assim, a imputação pelo algoritmo KNN proporcionou ganhos pontuais em métricas probabilísticas, como a AUC, sugerindo que, mesmo com baixo desempenho em métricas baseadas em classificação direta, alguns modelos mantiveram uma capacidade discriminativa moderada entre as classes. Esses achados reforçam a importância de considerar o contexto e as características intrínsecas dos dados na escolha da técnica de imputação mais adequada a cada cenário analítico.

4.1.4 Explicabilidade dos modelos

A etapa de explicabilidade teve como objetivo compreender o comportamento interno dos modelos de AM e identificar quais variáveis apresentaram maior influência na predição do desfecho de óbito por Dengue. Essa análise permite interpretar de forma mais transparente as decisões dos algoritmos, fornecendo subsídios para a compreensão clínica e epidemiológica dos resultados obtidos.

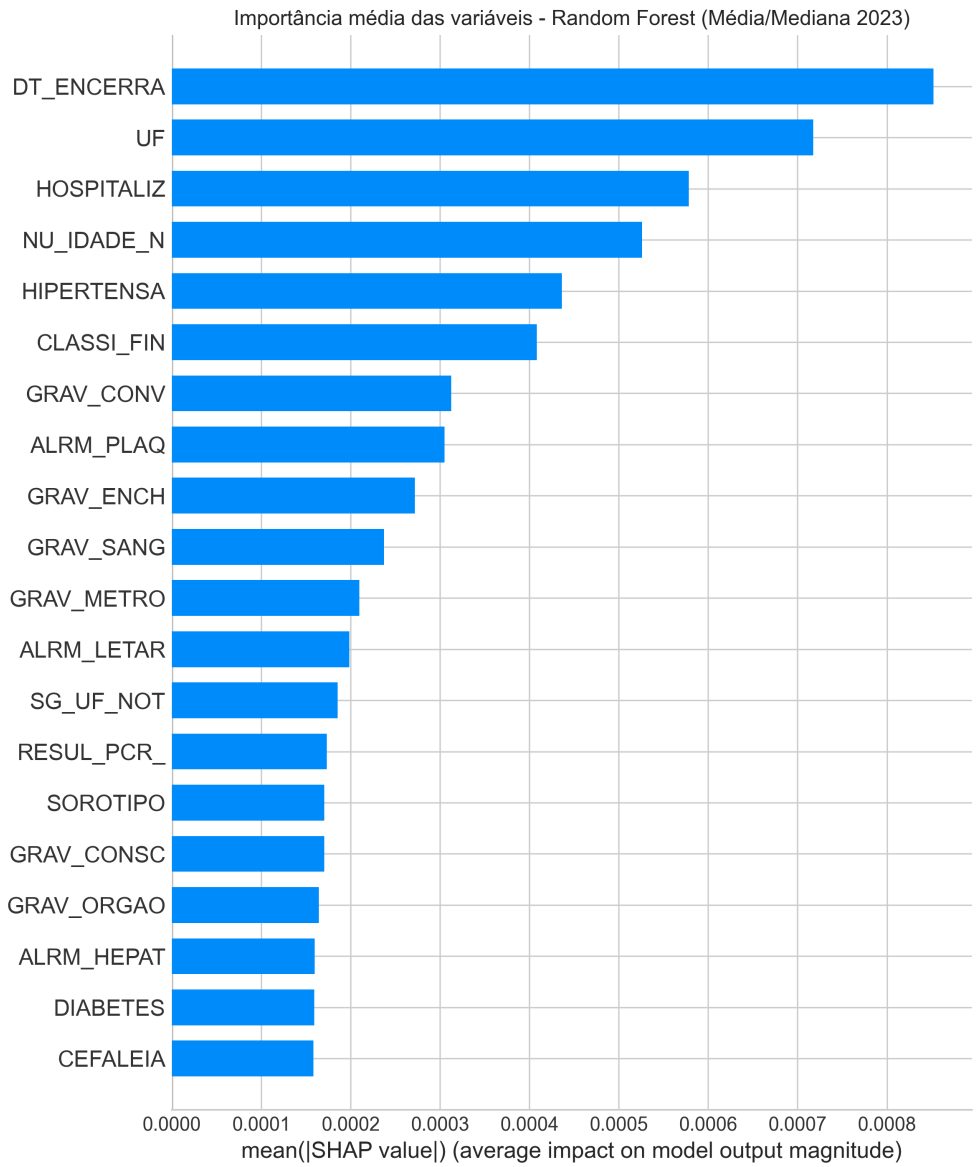
Foram consideradas nesta etapa os classificadores que obtiveram os melhores resultados na primeira imputação, baseada em estatísticas, sendo eles *Random Forest* e *XGBoost*, uma vez que estas se destacaram nas análises de classificação realizadas anteriormente. Optou-se por conduzir a explicabilidade utilizando exclusivamente esse tipo de imputação, dado que essa abordagem apresentou o desempenho mais consistente. A imputação via KNN, por sua vez, apresentou resultados insatisfatórios, o que inviabilizou sua utilização nesta etapa.

Para os classificadores escolhidos, foram empregados métodos de explicabilidade global e local

baseados em *SHAP values* e visualizações via UMAP, com o intuito de avaliar tanto a importância geral das variáveis quanto o comportamento das amostras no espaço de decisão dos modelos.

Primeiro foi realizada a análise do classificador *Random Forest*, com o intuito de identificar as variáveis que mais contribuíram para a predição do desfecho de óbito. A Figura 5 apresenta o gráfico da importância das variáveis para o *dataset* de 2023, no qual é possível observar as características que exerceram maior influência no modelo.

Figura 5 – Importância das variáveis com base nos *SHAP values* para o modelo *Random Forest* (2023).



Fonte: Elaborado pela autora.

O atributo *DT_ENCERRA* apresentou a maior importância na classificação dos casos de óbito, o que é coerente, uma vez que se relaciona diretamente à data do desfecho final do paciente e ao momento de encerramento do caso no sistema de notificação (SINAN). O segundo atributo mais relevante foi a Unidade Federativa (UF) no qual o caso foi registrado, sugerindo que fatores regionais — possivelmente associados às condições de atendimento, infraestrutura hospitalar e políticas públicas

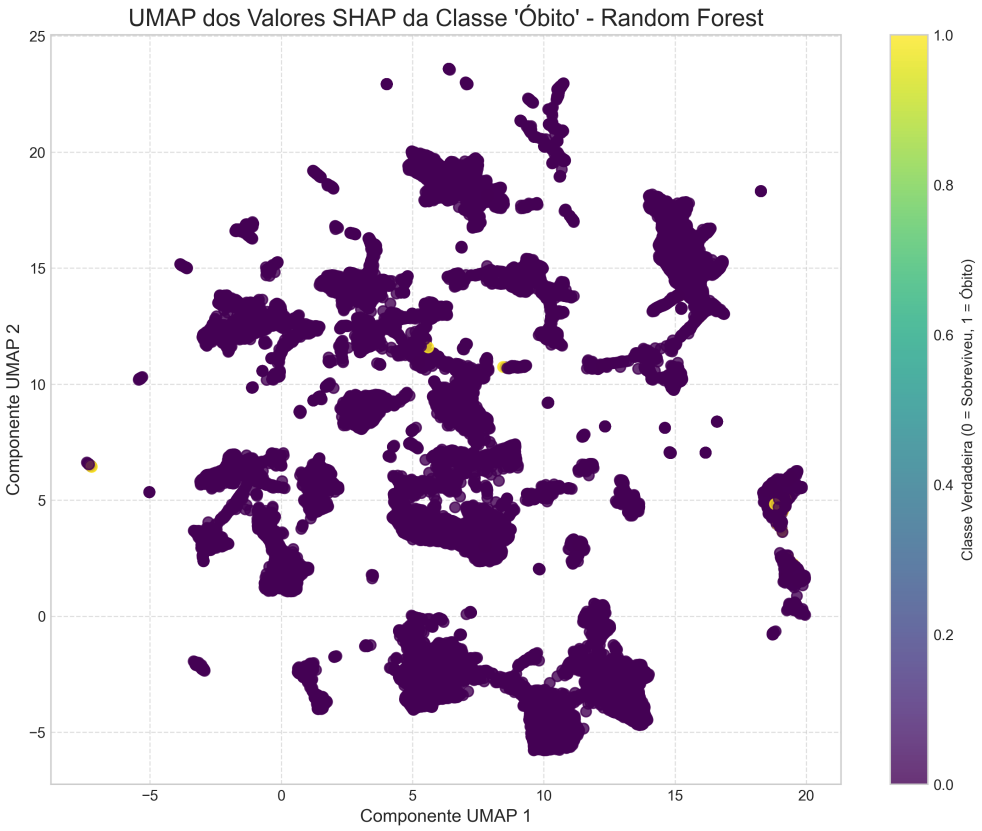
locais — influenciam de forma significativa o risco de mortalidade.

Observa-se, contudo, que algumas variáveis clínicas e laboratoriais normalmente associadas a complicações graves da doença, como presença de sangramento, baixos níveis de plaquetas ou sinais de alarme, não apresentaram destaque expressivo entre os atributos mais importantes. Essa menor relevância pode estar associada à incompletude desses campos nos *datasets* originais ou à sobreposição de informação entre variáveis correlacionadas, o que reduz seu impacto individual no modelo.

Ainda assim, a predominância de atributos administrativos e contextuais reforça a necessidade de melhoria na qualidade e padronização dos registros clínicos, de modo a permitir que características médicas tenham maior peso preditivo nas análises futuras.

Posteriormente, os valores de SHAP obtidos foram utilizados como entrada para a técnica de redução de dimensionalidade UMAP. Essa abordagem permitiu a projeção e visualização da distribuição das amostras no espaço de explicabilidade do modelo, sendo crucial para identificar possíveis padrões de separação entre os casos de óbito e não óbito com base nas suas respectivas contribuições de variáveis.

Figura 6 – Projeção bidimensional via UMAP dos valores de *SHAP* para o modelo *Random Forest* (2023), destacando os casos de óbito.



Fonte: Elaborado pela autora.

O gráfico apresenta a projeção bidimensional obtida a partir do UMAP, aplicada sobre os valores do modelo *Random Forest* para o conjunto de dados de 2023. Cada ponto no gráfico representa uma amostra, projetada em um espaço de duas dimensões construído com base na similaridade entre as explicações locais geradas pelo modelo.

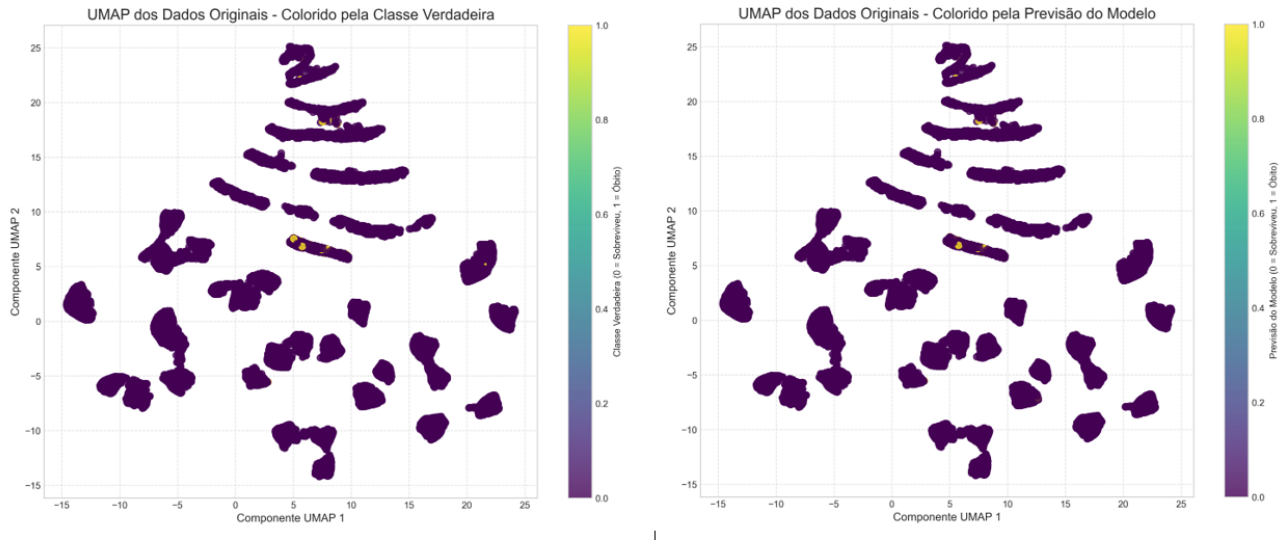
Os eixos do gráfico — Componente 1 e Componente 2 — não representam as variáveis originais diretamente, mas sim as direções de maior variabilidade e contraste nas explicações do modelo. Como o UMAP foi aplicado aos valores SHAP, a interpretação da proximidade é fundamental: amostras que se encontram próximas no plano compartilham padrões de importância e influência das variáveis semelhantes na predição (*i.e.*, um comportamento explicativo comum), enquanto pontos distantes indicam mecanismos de decisão ou fatores distintos.

Na visualização, observa-se a formação de regiões mais densas que agrupam os casos com explicações similares. A concentração de casos de óbito em áreas específicas do mapa sugere que o modelo identificou subconjuntos de indivíduos com perfis explicativos comuns, potencialmente associados a determinados contextos epidemiológicos ou administrativos. Essa separação reforça que o modelo não apenas distinguiu corretamente as classes, mas também capturou nuances nas relações entre atributos.

Essa projeção permite, portanto, compreender de forma mais intuitiva como o modelo organiza internamente suas decisões e como os casos de óbito se distribuem em relação aos demais registros, oferecendo uma perspectiva complementar à análise de importância das variáveis.

Complementarmente, empregou-se a técnica de redução de dimensionalidade para comparar a distribuição original dos dados com as predições realizadas pelo modelo, conforme ilustrado na Figura 7.

Figura 7 – Distribuição original dos dados rotulados e predições realizadas pelo modelo Random Forest (2023).



Fonte: Elaborado pela autora.

A figura apresenta a projeção UMAP dos dados originais coloridos pela classe verdadeira (à esquerda) e a predição realizada pelo modelo *Random Forest* (à direita). Cada ponto representa uma amostra do conjunto de 2023, projetada em duas dimensões a partir das relações de similaridade entre suas características originais.

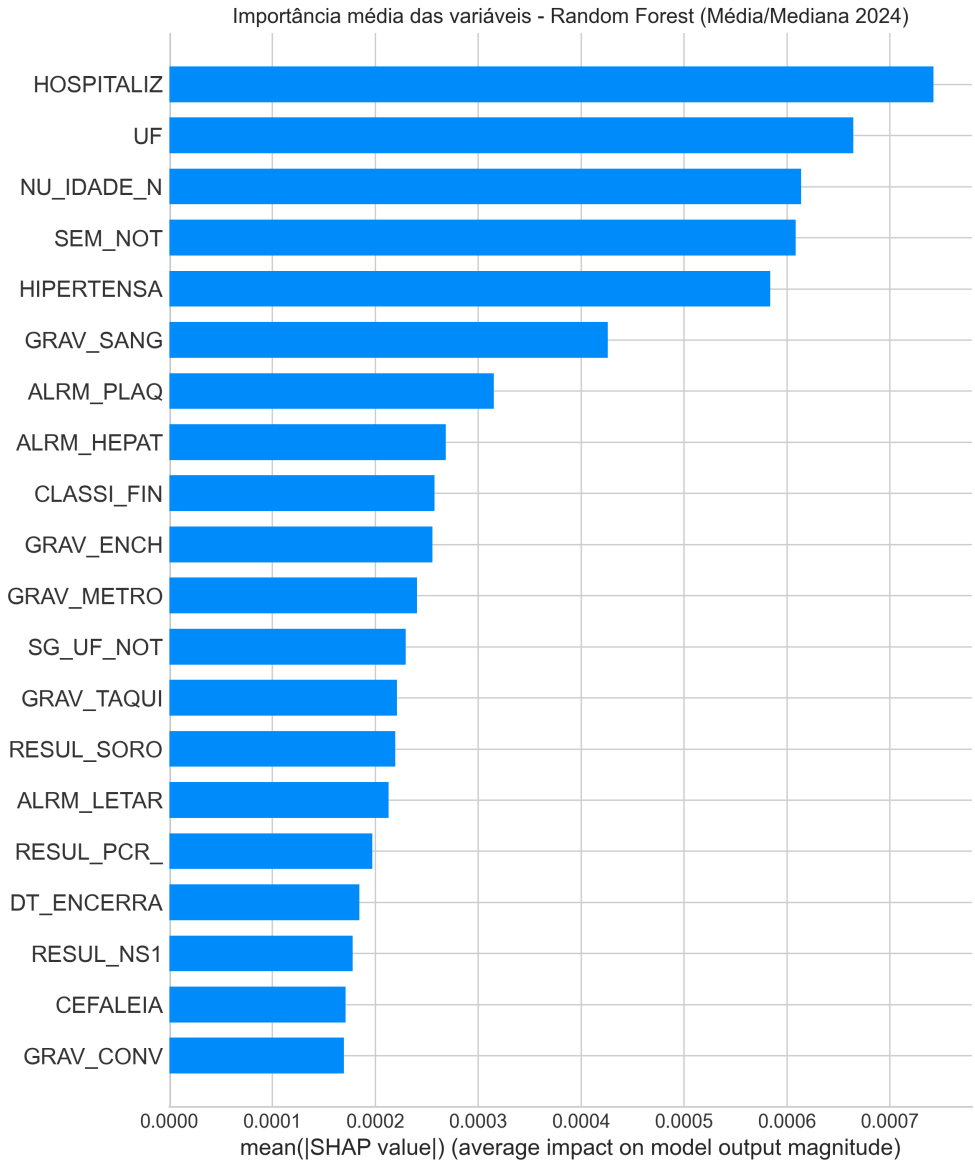
Ambas as projeções exibem distribuições visuais muito semelhantes, indicando que o modelo

foi capaz de reproduzir adequadamente a estrutura dos dados originais. Essa correspondência sugere que as predições do *Random Forest* refletem, de maneira consistente, os padrões presentes nos atributos utilizados para o treinamento. As pequenas variações observadas em regiões pontuais do mapa correspondem a amostras cuja classificação apresentou maior incerteza ou divergência em relação à classe real, o que é esperado em problemas com forte desbalanceamento entre as classes.

A semelhança entre as distribuições reforça a capacidade do modelo em capturar as relações subjacentes entre os atributos e o desfecho (óbito ou não óbito), demonstrando coerência entre o espaço de características aprendido e a estrutura intrínseca dos dados.

O conjunto de dados de 2024 apresentou uma pequena variação entre os atributos mais importantes para a classificação como óbito.

Figura 8 – Importância das variáveis com base nos *SHAP values* para o modelo *Random Forest* (2024).



Fonte: Elaborado pela autora.

Ao contrário de 2023, para o ano de 2024 o tempo de hospitalização passou a ocupar posição de

destaque, representando o atributo com maior influência nas predições do modelo. Essa mudança sugere que, nesse período, a variável associada à internação teve papel mais determinante na diferenciação entre os casos de óbito e não óbito, possivelmente refletindo diferenças no perfil clínico dos pacientes ou na qualidade do registro hospitalar.

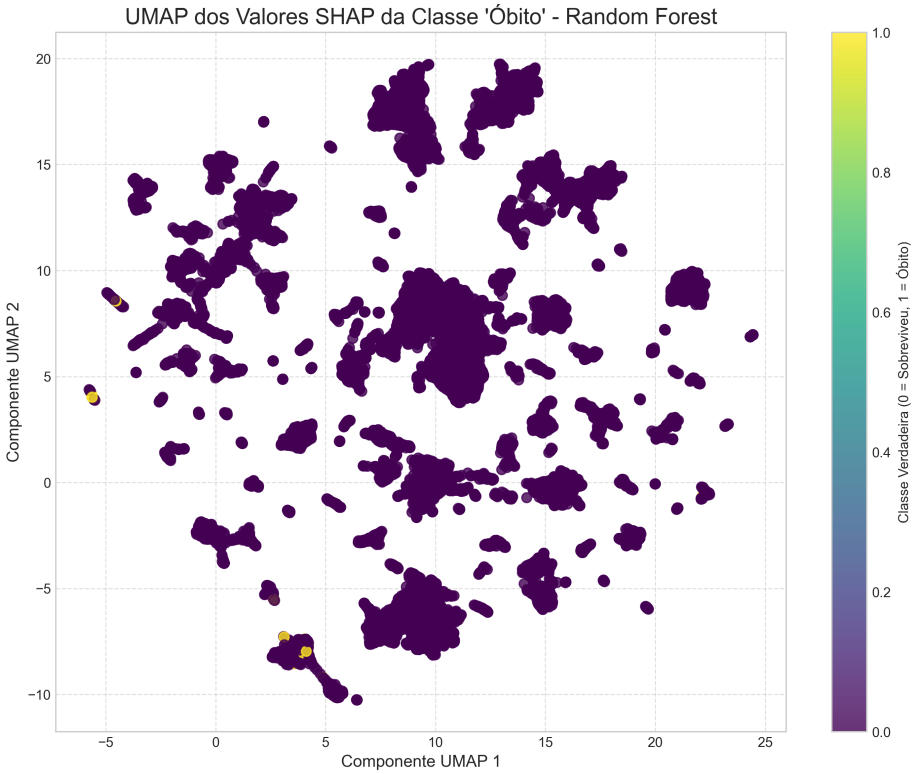
Entre as demais variáveis relevantes, observam-se novamente fatores administrativos e contextuais, como a unidade Federativa (UF) e o número da semana de notificação do caso, acompanhados de características demográficas, como a idade do paciente.

Fatores clínicos como hipertensão e sinais de gravidade relacionados à sangramentos e ao baixo nível de plaquetas também se mantiveram entre os principais preditores, indicando consistência com os achados do ano anterior.

De modo geral, observa-se que as variáveis administrativas e contextuais continuam exercendo maior peso preditivo, embora a influência de atributos clínicos tenha aumentado ligeiramente. Esse comportamento sugere uma possível melhoria na qualidade das informações clínicas registradas ou uma mudança na representatividade dos casos de maior gravidade em 2024.

Para a projeção bidimensional, assim como observado em 2023, nota-se a formação de regiões densas que agrupam amostras com explicações semelhantes. Os casos de óbito permanecem concentrados em áreas específicas do mapa, indicando que o modelo manteve coerência na forma como organiza internamente suas decisões.

Figura 9 – Projeção bidimensional via UMAP dos valores de *SHAP* para o modelo *Random Forest* (2024), destacando os casos de óbito.



Fonte: Elaborado pela autora.

Assim como observado em 2023, as distribuições das classes reais e preditas pelo modelo

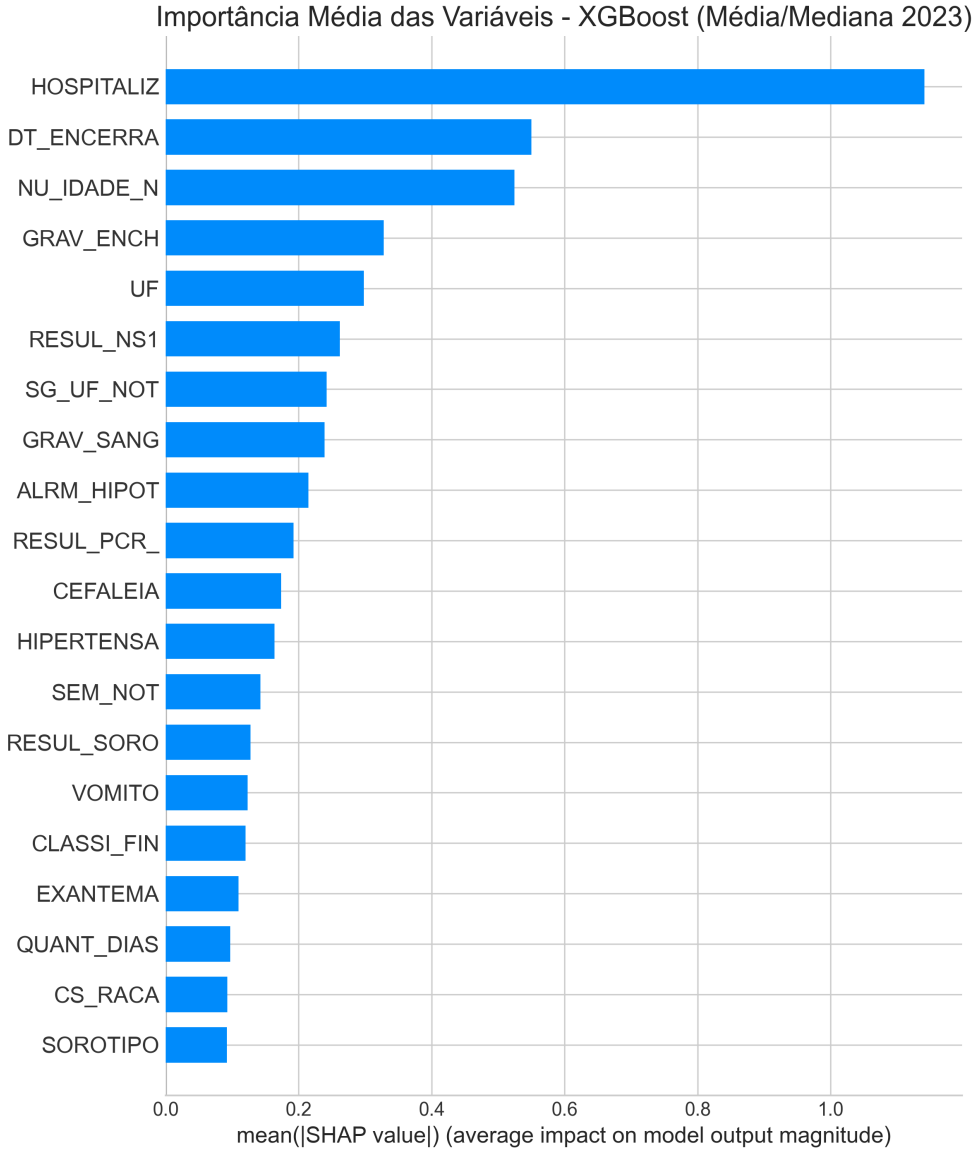
apresentaram grande similaridade no contexto de 2024, indicando consistência no desempenho e na forma como o modelo diferenciou os casos de óbito e não óbito ao longo dos anos.

Essa estabilidade entre os anos reforça a robustez do modelo *Random Forest* e a consistência dos padrões identificados, mesmo diante de pequenas inconstâncias nas variáveis mais influentes.

Após a análise do modelo *Random Forest*, aplicou-se os mesmos algoritmos de explicabilidade para o *XGBoost*. Esse classificador foi selecionado por sua eficiência computacional e por incorporar mecanismos internos de regularização, que tendem a reduzir o sobreajuste e a melhorar a precisão das previsões.

Novamente, foram utilizados os *SHAP values* para identificar as variáveis com maior contribuição para a classificação dos casos de óbito. A Figura 10 apresenta a importância das variáveis obtida para o conjunto de dados de 2023.

Figura 10 – Importância das variáveis com base nos *SHAP values* para o modelo *XGBoost* (2023).



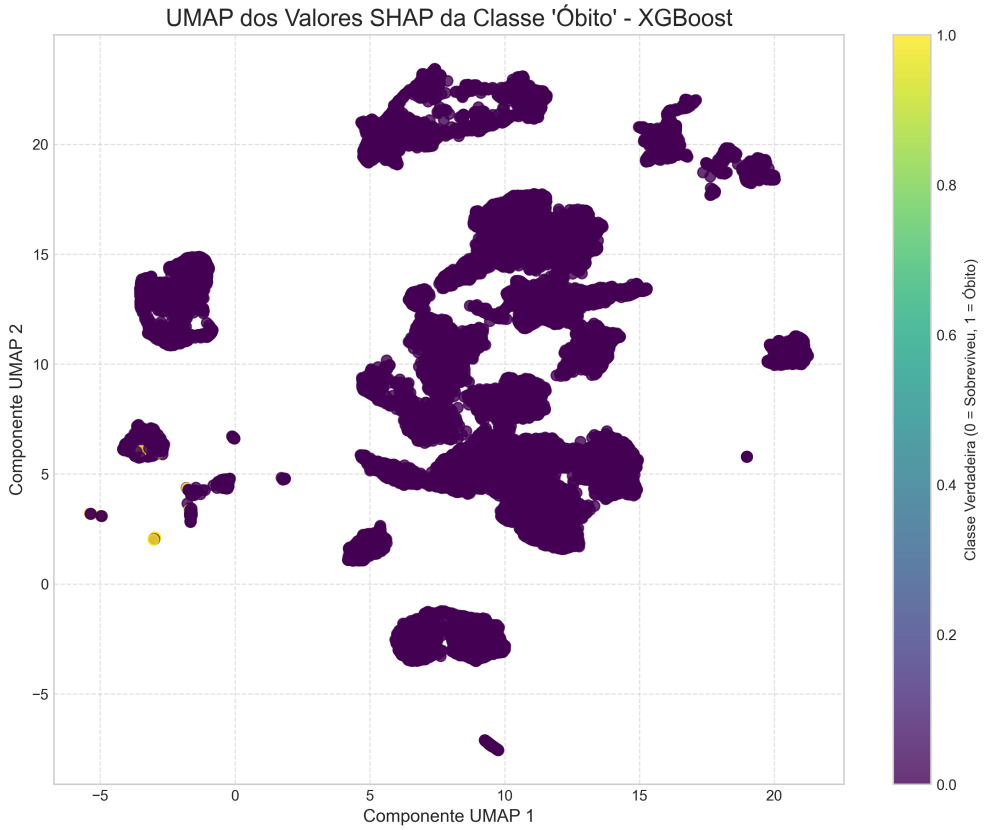
Fonte: Elaborado pela autora.

Observa-se que a variável referente à quantidade de dias em hospitalização destacou-se como a mais influente na predição dos casos de óbito, seguida pela data de encerramento do caso e a idade do paciente. Esse resultado reforça o padrão identificado no modelo *Random Forest*, indicando que esses fatores possuem papel determinante na classificação dos casos.

O UMAP apresentado pelo gráfico 11 mostra a projeção bidimensional dos valores de explicabilidade obtidos a partir dos *SHAP values* do modelo *XGBoost*, considerando apenas a classe "Óbito".

O gráfico exemplifica a formação de agrupamentos bem definidos, as regiões em tons mais claros correspondem aos casos de óbito, que se concentram em áreas específicas do espaço latente, sugerindo que o modelo aprendeu padrões consistentes para diferenciar os indivíduos que evoluíram à óbito dos demais. Essa separação indica que o *XGBoost* foi capaz de capturar relações significativas entre as variáveis clínicas, demográficas e administrativas no processo de predição.

Figura 11 – Projeção bidimensional via UMAP dos valores de *SHAP* para o modelo *XGBoost* (2023), destacando os casos de óbito.



Fonte: Elaborado pela autora.

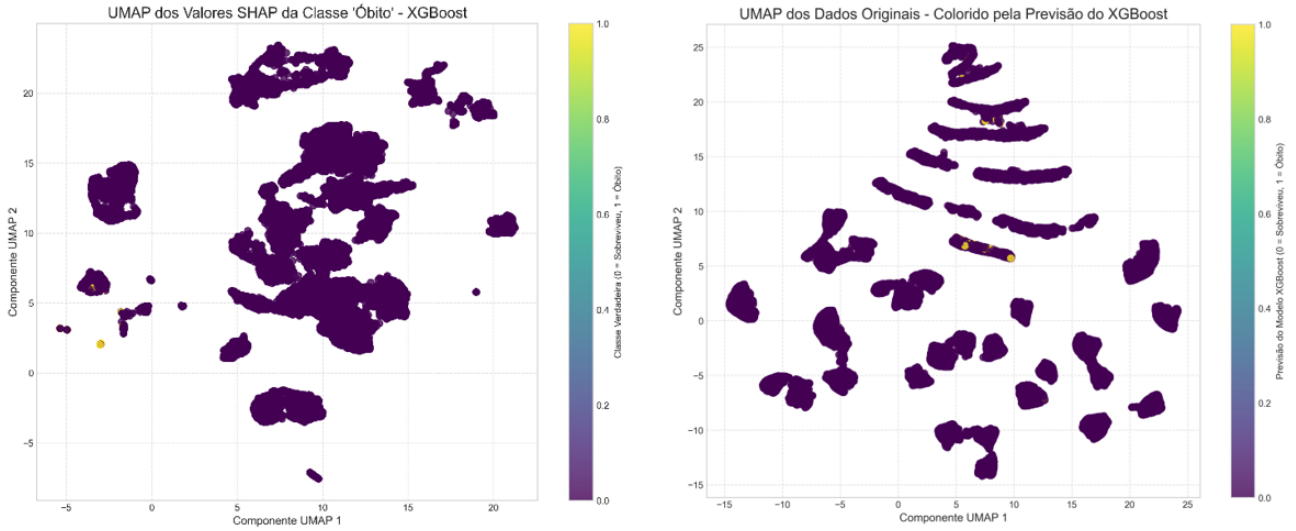
Além disso, o comportamento observado é coerente com o modelo *Random Forest*, indicando que ambos os classificadores reconheceram uma estrutura semelhante nos dados. Essa consistência entre modelos distintos reforça a robustez dos padrões aprendidos e a relevância das variáveis identificadas na análise de importância.

A Figura 12 apresenta uma análise comparativa do modelo *XGBoost*. O painel esquerdo pro-

jeta a contribuição das variáveis (*SHAP values*) para os casos de óbito, onde a visualização demonstra a capacidade do modelo de aprender e distinguir a estrutura da classe minoritária, representada pelos agrupamentos em amarelo.

Contudo, a análise do painel direito, que utiliza a mesma projeção, mas colore os pontos conforme a predição final do modelo, revela um resultado divergente. Nota-se a predominância esmagadora da previsão 0 (Sobreviveu), com a classe 1 (Óbito) sendo erroneamente predita em raros casos.

Figura 12 – Distribuição original dos dados rotulados e predições realizadas pelo modelo *XGBoost* (2023).



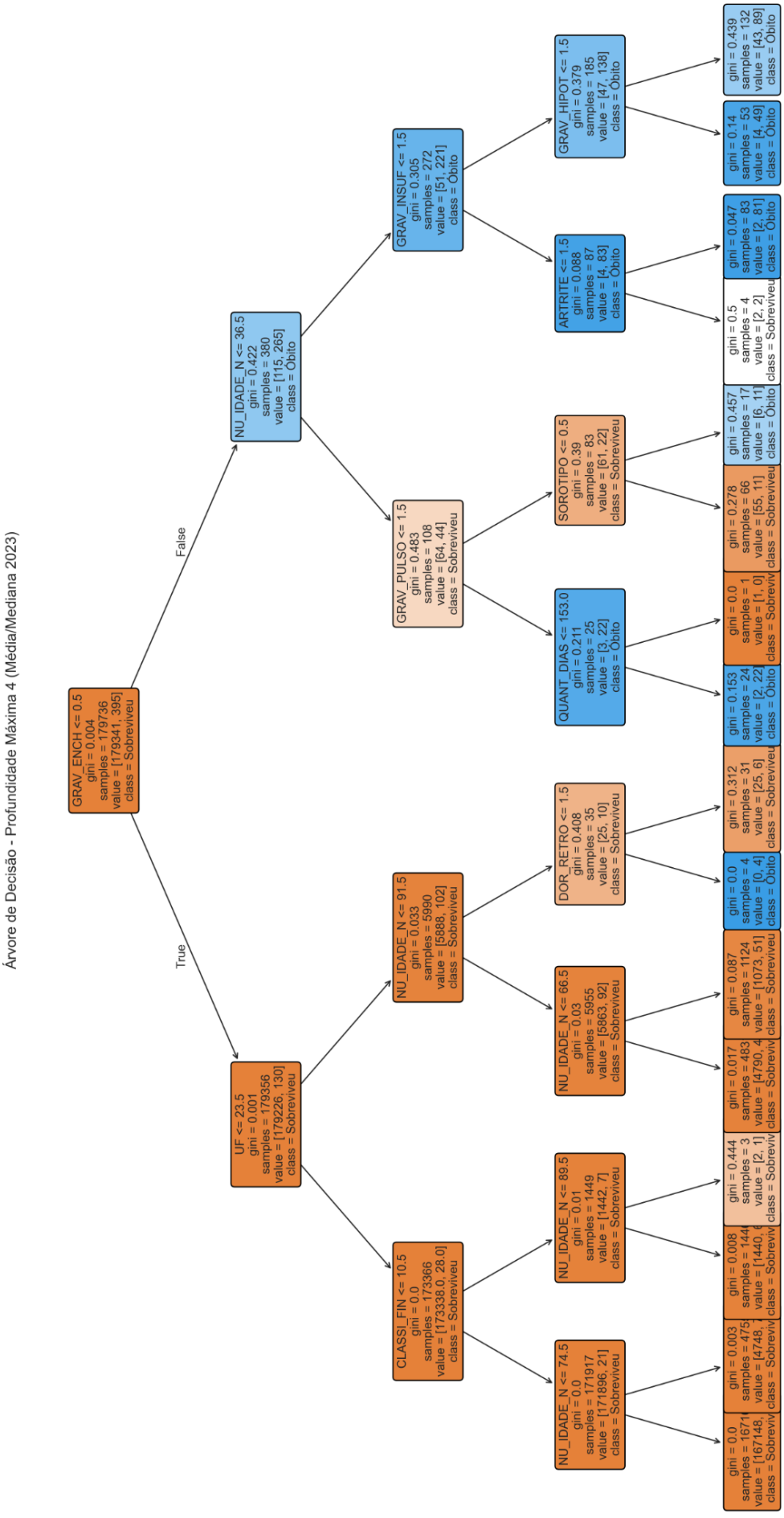
Fonte: Elaborado pela autora.

Esta discrepância indica que, embora o *XGBoost* tenha aprendido padrões significativos para ranquear a probabilidade de óbito, o uso do *threshold* padrão de 0.5 o torna excessivamente conservador.

Conclui-se, portanto, que a performance insatisfatória em métricas como *Recall* e *F1-Score* deve-se à dificuldade de conversão do potencial de ranqueamento em uma classificação binária eficaz, reforçando a necessidade de calibração do limite de corte de probabilidade.

Por fim, para complementar a análise de explicabilidade global e local dos algoritmos anteriores, foi explorada a estrutura interna do modelo de Árvore de Decisão, devido ao seu caráter intrinsecamente interpretável. Essa visualização permite compreender de forma direta as regras de decisão aprendidas pelo modelo, evidenciando como os atributos são combinados para distinguir entre os casos.

Figura 13 – Estrutura da Árvore de Decisão treinada sobre a base Média/Mediana (2023), com profundidade máxima igual a 4.

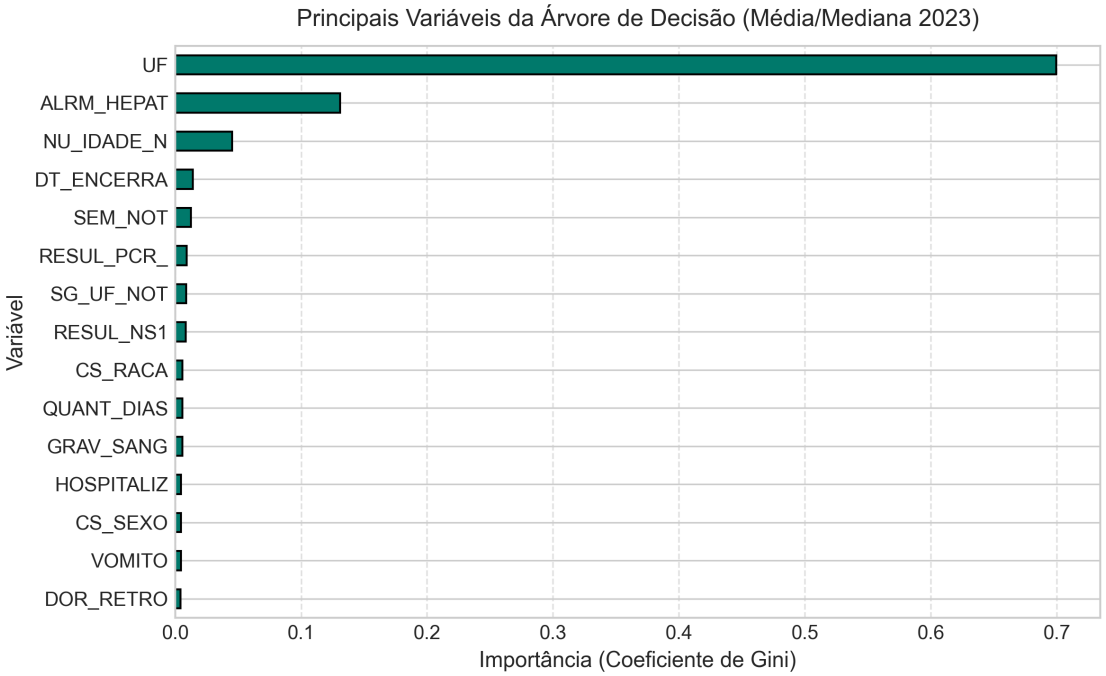


Fonte: Elaborado pela autora.

Na Figura 13, os primeiros nós de decisão envolvem atributos como *GRAV_ENCH*, referente a extravasamento plasmático; a Unidade Federativa (UF) e a idade do paciente *NU_IDADE_N*, que aparecem também entre as variáveis mais relevantes nos modelos de *Random Forest* e *XGBoost*. A predominância de variáveis administrativas e de classificação reforça a hipótese de que fatores contextuais e de notificação influenciam significativamente a predição do desfecho.

Em contrapartida, variáveis clínicas associadas a complicações — como insuficiência, sangramento e hipotensão — aparecem apenas em níveis inferiores da árvore, confirmando que, embora possuam relação com o óbito, sua contribuição para a separação inicial das classes é mais limitada, possivelmente devido à baixa frequência ou à inconsistência no registro desses sintomas.

Figura 14 – Importância das variáveis no modelo de Árvore de Decisão (2023), com base no coeficiente de Gini.



Fonte: Elaborado pela autora.

A Figura 14 reforça esse padrão, destacando a Unidade Federativa (UF) como a mais influente, seguido de comprometimento hepático (*ALRM_HEPAT*) e à idade do paciente. Tal comportamento é coerente com os resultados dos modelos *ensemble*, demonstrando consistência entre as diferentes abordagens de explicabilidade.

Em suma, as análises de explicabilidade evidenciam a coerência entre os diferentes modelos aplicados, demonstrando que tanto os classificadores baseados em conjuntos (*ensemble*) quanto a Árvore de Decisão reconheceram padrões semelhantes. A predominância de variáveis administrativas e epidemiológicas sobre as clínicas ressalta desafios estruturais nos registros de Dengue, indicando a importância de aprimorar a completude e padronização dos dados clínicos para aprimorar futuras abordagens preditivas.

4.2 Detecção de Outliers

Nesta abordagem, foram empregados três algoritmos distintos: *Isolation Forest*, *Local Outlier Factor* (LOF) e *One-Class SVM*. O objetivo foi investigar se tais métodos seriam capazes de identificar os casos de óbito como observações anômalas nos conjuntos de dados. As Tabelas a seguir apresentam os resultados obtidos em termos das principais métricas de desempenho.

Os resultados referentes ao *Isolation Forest*, apresentados na Tabela 4.2, indicam um desempenho limitado na identificação de óbitos, evidenciado pelos baixos valores de *F1-Score* e AUPRC em ambos os anos.

Tabela 8 – Resultados do Isolation Forest nos Testes Finais.

Ano	F1-Score	Recall	Precisão	AUROC	AUPRC
2023	**0.0610**	0.0830	**0.0480**	**0.5440**	**0.0094**
2024	0.0057	**0.7690**	0.0029	0.3900	0.0031

Fonte: Elaborado pela autora.

Embora o modelo de 2024 apresente um *Recall* elevado (0.769), o valor de precisão extremamente baixo (0.0029) revela que muitos casos não pertencentes à classe minoritária foram incorretamente classificados como *outliers*. Esse comportamento sugere que o algoritmo não conseguiu distinguir adequadamente os padrões que caracterizam os óbitos, possivelmente devido à natureza complexa e não linear das variáveis clínicas e sociodemográficas envolvidas.

Em comparação, o algoritmo *Local Outlier Factor* (LOF) apresentou um desempenho ligeiramente superior no ano de 2023, com valores mais altos de *F1-Score*, precisão e AUPRC, como é possível ver na Tabela 9. No entanto, de forma semelhante ao *Isolation Forest*, o modelo de 2024 exibiu um *Recall* elevado (0.8339) acompanhado de uma precisão extremamente baixa (0.0031), evidenciando novamente um grande número de falsos positivos.

Esses resultados sugerem que, embora o LOF tenha identificado corretamente uma proporção maior de casos de óbito, ele também classificou incorretamente muitos registros normais como *outliers*, reforçando a dificuldade desses algoritmos em capturar a complexidade das relações entre as variáveis que diferenciam casos fatais e não fatais.

Tabela 9 – Resultados do Local Outlier Factor (LOF) nos Testes Finais

Ano	F1-Score	Recall	Precisão	AUROC	AUPRC
2023	**0.0757**	0.1109	**0.0575**	**0.5501**	**0.0116**
2024	0.0062	**0.8339**	0.0031	0.4169	0.0032

Fonte: Elaborado pela autora.

Por sua vez, a Tabela 10 revela que o *One-Class SVM* apresentou um comportamento distinto em relação aos demais algoritmos. No ano de 2023, o modelo alcançou um *Recall* expressivamente elevado (0.8849), indicando que a maior parte dos casos de óbito foi identificada como *outlier*. Esse

resultado sugere uma boa sensibilidade do método em reconhecer padrões anômalos associados aos óbitos.

Além disso, o valor de AUROC (0.8444) foi expressivamente superior aos obtidos pelos demais algoritmos, refletindo uma capacidade relativamente maior de separação entre as classes, mesmo em um cenário não supervisionado.

Contudo, a precisão permaneceu baixa (0.0258), o que indica que, embora o modelo tenha conseguido detectar a maioria dos casos positivos, também classificou incorretamente um grande número de registros não relacionados ao óbito como anômalos. Esse comportamento reflete uma tendência do *One-Class SVM* de gerar falsos positivos em contextos de forte desbalanceamento, como o presente neste estudo.

No ano de 2024, o desempenho do algoritmo apresentou uma queda acentuada em todas as métricas, com destaque para a redução do AUROC e do *F1-Score*, que atingiu apenas 0.0042. Essa deterioração aponta para uma perda da capacidade de generalização do modelo diante de novos dados, possivelmente associada à variação temporal das características clínicas e sociodemográficas dos pacientes.

Tabela 10 – Resultados do One-Class SVM nos Testes Finais (2023 vs 2024)

Ano	**F1-Score**	**Recall**	**Precisão**	**AUROC**	**AUPRC**
2023	**0.0502**	**0.8849**	**0.0258**	0.8444	**0.0235**
2024	0.0042	0.5648	0.0021	0.2824	0.0028

Fonte: Elaborado pela autora.

De modo geral, os três algoritmos explorados apresentaram comportamentos distintos em relação à identificação de casos de óbito. Enquanto o *Isolation Forest* se destacou pelo equilíbrio entre precisão e sensibilidade, o *Local Outlier Factor* mostrou-se mais conservador, identificando menos instâncias positivas, porém com maior confiança nas classificações. Por outro lado, o *One-Class SVM* apresentou alta sensibilidade em 2023, mas com queda expressiva nos resultados de 2024, sugerindo menor capacidade de generalização temporal.

Esses resultados revelam que, embora as técnicas de detecção de *outliers* consigam capturar parte dos padrões anômalos, sua eficácia é bastante variável entre os algoritmos e períodos analisados. Assim, a escolha do método mais adequado deve considerar não apenas a performance estatística, mas também a estabilidade do modelo ao longo do tempo e sua interpretabilidade no contexto clínico.

4.2.1 Comparação das abordagens

As abordagens de detecção de *outliers* e de aprendizado supervisionado (AM) apresentaram diferenças substanciais em termos de desempenho e aplicabilidade para o tema do presente estudo.

A detecção de *outliers* teve como objetivo identificar registros atípicos no conjunto de dados, partindo do pressuposto de que os casos de óbito constituem eventos raros e potencialmente detectáveis como anomalias. Apesar de alguns modelos apresentarem *Recall* elevado — indicando boa sensibilidade

para detectar óbitos —, as demais métricas, especialmente precisão e *F1-score*, permaneceram muito baixas. Esse comportamento sugere que os algoritmos de detecção de anomalias não conseguiram discriminar de forma eficaz os casos de óbito em meio aos não óbitos, resultando em grande número de falsos positivos.

Por outro lado, os classificadores supervisionados mostraram-se consideravelmente mais eficazes ao explorar relações complexas entre variáveis clínicas, sociodemográficas e o desfecho de interesse. Esses modelos alcançaram melhor equilíbrio entre sensibilidade e precisão, indicando maior capacidade de generalização. Além disso, sua interpretabilidade foi reforçada pela análise de explicabilidade baseada em *SHAP values* e UMAP, que permitiu compreender o impacto individual das variáveis e a estrutura interna das decisões.

Enquanto a abordagem de detecção de *outliers* mostrou-se limitada em capturar a complexidade multivariada dos dados clínicos e administrativos, os classificadores supervisionados foram capazes de identificar padrões mais robustos e coerentes com o contexto epidemiológico.

Dessa forma, os resultados indicam que a utilização de modelos supervisionados foi mais adequada para a tarefa de predição de mortalidade, especialmente quando associada a estratégias adequadas de imputação e balanceamento de dados.

5 Conclusão

Este trabalho teve como objetivo investigar e comparar diferentes técnicas de aprendizado de máquina aplicadas à predição de mortalidade em pacientes diagnosticados com Dengue, bem como avaliar o desempenho de métodos de detecção de anomalias como estratégia alternativa para identificação de casos de óbito. A partir das análises realizadas, verificou-se que os algoritmos supervisionados apresentaram desempenho significativamente superior aos métodos de detecção de *outliers*, demonstrando maior capacidade de generalização e melhor equilíbrio entre as métricas de precisão e sensibilidade.

A abordagem de detecção de anomalias mostrou-se limitada para o problema em questão. Embora alguns modelos tenham apresentado *Recall* elevado - indicando sensibilidade aceitável na detecção de óbitos - o alto número de falsos positivos comprometeu a precisão e, consequentemente, o desempenho geral. Esses resultados sugerem que os padrões que diferenciam os casos de óbito não são suficientemente capturados por técnicas que assumem uma distribuição majoritária homogênea, o que reforça a complexidade do contexto clínico e epidemiológico da Dengue.

Em contrapartida, os modelos supervisionados apresentaram resultados mais consistentes. Técnicas como *Support Vector Machine* (SVM), *Random Forest* e Regressão Logística conseguiram explorar de forma mais eficiente as interações entre variáveis clínicas e sociodemográficas, alcançando melhor desempenho preditivo.

A aplicação de métodos de balanceamento de classes, como SMOTE e a combinação de SMOTE com *Tomek Links*, contribuiram significativamente para aprimorar a sensibilidade sem comprometer a precisão. Além disso, a análise de interpretabilidade, conduzida por meio de *SHAP values* e UMAP, evidenciou o papel de variáveis como idade, presença de comorbidades e indicadores laboratoriais na definição do risco de mortalidade, oferecendo uma visão mais transparente e útil do processo decisório dos modelos.

Com base nos resultados obtidos, conclui-se que os objetivos propostos foram alcançados. A comparação entre abordagens demonstrou a superioridade dos métodos supervisionados para o problema proposto e ressaltou a importância do pré-processamento adequado e da interpretabilidade dos modelos para aplicações em saúde pública. No entanto, o trabalho apresenta algumas limitações que devem ser consideradas. Primeiramente, os dados utilizados são oriundos de notificações do SINAN, que podem conter inconsistências, subnotificações e variabilidade no preenchimento dos dados, o que pode impactar o desempenho e a generalização dos modelos. Além disso, o forte desbalanceamento entre as classes, característico de estudos de mortalidade, exigiu técnicas artificiais de balanceamento que, embora úteis, podem introduzir vieses ou alterar a distribuição real dos dados. Outro ponto é que o estudo se restringiu apenas aos anos de 2023 e 2024, o que limita a análise temporal e a robustez diante de padrões epidemiológicos futuros.

5.0.1 Trabalhos Futuros

Com base nas limitações observadas ao longo do estudo e nas oportunidades de aprimoramento identificadas durante as análises, alguns caminhos podem ser explorados em trabalhos futuros. Tais direções visam aprofundar a compreensão sobre o comportamento dos dados, refinar o processo de pré-processamento e avaliar metodologias alternativas que possam complementar ou superar as abordagens utilizadas neste trabalho. As principais sugestões para continuidade da pesquisa incluem:

- Investigar o impacto da manutenção dos registros com mais de 50% de valores ausentes, avaliando como a ausência dessa etapa de exclusão prévia pode influenciar o processo de imputação e, consequentemente, o desempenho dos modelos preditivos.
- Aplicar técnicas de aprendizado profundo (*Deep Learning*) e integrar abordagens de explicabilidade temporal, de modo a aprimorar a capacidade preditiva e fortalecer a aplicabilidade dos modelos em sistemas de apoio à decisão e vigilância epidemiológica.
- Explorar métodos mais avançados de detecção de *outliers*, incluindo técnicas de aprendizado não supervisionado e semissupervisionado, bem como abordagens híbridas que combinem diferentes algoritmos.
- Incorporar mecanismos de calibração ou ponderação baseados em características clínicas, visando reduzir o número de falsos positivos observados nesta análise e aumentar a consistência dos resultados.
- Aplicar métodos de detecção de *outliers* em estágios preliminares do fluxo de pré-processamento, possibilitando a identificação de inconsistências ou registros anômalos antes da modelagem supervisionada.

Em síntese, as sugestões propostas para trabalhos futuros têm o potencial de aumentar a robustez dos modelos, aprofundar a compreensão dos padrões clínicos associados à mortalidade por Dengue e fortalecer a aplicabilidade prática das abordagens baseadas em aprendizado de máquina no contexto da vigilância epidemiológica. Avançar nessas frentes pode contribuir para sistemas preditivos mais precisos, interpretáveis e efetivos no apoio à tomada de decisão em saúde pública.

6 Apêndice I - Descrição Detalhada dos *Datasets* reduzidos

Este apêndice é composto pelas tabelas com a distribuição das bases de dados após a redução dos registros. Cada tabela é composta pelos atributos originais dispostos em linhas, as colunas permitem algumas informações sobre cada atributo, sendo elas:

- Tipo de dado: Pode receber os valores 'Numérico', 'Binário' ou 'Categórico'.
- Variável alvo: Essa coluna demonstra se o atributo é ou não nossa classe alvo para predição.
- Valores únicos: Informa quantos valores diferentes existem para o atributo, assim o valor '1' nessa coluna significa que o atributo só possui um único valor para todos os registros, podendo ser irrelevante para a predição.
- Total de valores faltantes: Indica quantos registros vazios ou 'NaN' existem na coluna.
- Porcentagem de valores faltantes: Calcula a porcentagem de valores faltantes no atributo para facilitar a visualização e impacto. É a partir desse valor que definiu-se os atributos irrelevantes para a predição, removendo-os.
- Média dos atributos: Apresenta, para os atributos numéricos, a média dos valores que o compõem.
- Descrição: O que o atributo representa, ou seja, o significado.

Abaixo são exibidas as tabelas de cada *dataset*.

6.0.1 Descrição do *Dataset* 2023

Tabela 11 – Características dos Atributos do Dataset de Dengue (2023).

Atributo	**Tipo**	**É Classe?**	**Valores Únicos**	**Total Faltantes**	**% Faltantes**	**Descrição**
SEM_NOT	Numérico	Não	22	0	0%	Semana epidemiológica da notificação do caso.
SG_UF_NOT	Numérico	Não	26	0	0%	Sigla da Unidade Federativa (Estado) de notificação do caso.
NU_IDADE_N	Numérico	Não	100	4	0%	Idade do paciente no momento da notificação.
CS_SEXO	Categórico	Não	3	2	0%	Sexo biológico do paciente.
CS_GESTANT	Numérico	Não	7	64	0.02%	Situação de gravidez (apenas para o sexo feminino).
CS_RACA	Numérico	Não	6	2	0%	Raça/Cor autodeclarada do paciente.

Tabela 11 – continuação da página anterior

Atributo	Tipo	É Classe?	Valores Únicos	Total Faltantes	% Faltantes	Descrição
FEBRE	Binário	Não	2	1	0%	Presença de febre como sintoma.
MIALGIA	Binário	Não	2	1	0%	Presença de mialgia (dor muscular) como sintoma.
CEFALEIA	Binário	Não	2	1	0%	Presença de cefaleia (dor de cabeça) como sintoma.
EXANTEMA	Binário	Não	2	1	0%	Presença de exantema (erupção cutânea) como sintoma.
VOMITO	Binário	Não	2	1	0%	Presença de vômito como sintoma.
NAUSEA	Binário	Não	2	1	0%	Presença de náusea como sintoma.
DOR_COSTAS	Binário	Não	2	1	0%	Presença de dor nas costas como sintoma.
CONJUNTVIT	Binário	Não	2	1	0%	Presença de conjuntivite como sintoma.
ARTRITE	Binário	Não	2	1	0%	Presença de artrite (inflamação nas articulações) como sintoma.
ARTRALGIA	Binário	Não	2	1	0%	Presença de artralgia (dor nas articulações) como sintoma.
PETEQUIA_N	Binário	Não	2	1	0%	Presença de petéquias (pequenas manchas vermelhas na pele) como sintoma.
LEUCOPENIA	Binário	Não	2	1	0%	Presença de leucopenia (baixa contagem de glóbulos brancos) como sintoma.
LACO	Binário	Não	2	1	0%	Resultado da prova do laço (teste para fragilidade capilar).
DOR_RETRO	Binário	Não	2	1	0%	Presença de dor retro-orbital (atrás dos olhos) como sintoma.
DIABETES	Binário	Não	2	1	0%	Presença de comorbidade: Diabetes.
HEMATOLOG	Binário	Não	2	1	0%	Presença de comorbidade: Doença hematológica.
HEPATOPAT	Binário	Não	2	1	0%	Presença de comorbidade: Hepatopatia (doença do fígado).
RENAL	Binário	Não	2	1	0%	Presença de comorbidade: Doença renal.
HIPERTENSA	Binário	Não	2	1	0%	Presença de comorbidade: Hipertensão.
ACIDO_PEPT	Binário	Não	2	1	0%	Presença de comorbidade: Doença ácido péptica (úlcera, gastrite).
AUTO_IMUNE	Binário	Não	2	1	0%	Presença de comorbidade: Doença autoimune.
RESUL_SORO	Numérico	Não	4	138285	34.51%	Resultado do exame sorológico para Dengue.
RESUL_NS1	Numérico	Não	4	122187	30.49%	Resultado do teste rápido NS1 para Dengue.
RESUL_VI_N	Numérico	Não	4	190238	47.47%	Resultado do isolamento viral.
RESUL_PCR_	Numérico	Não	4	183386	45.76%	Resultado do RT-PCR (Reação em Cadeia da Polimerase) para Dengue.
SOROTIPO	Numérico	Não	4	387704	96.75%	Sorotipo do vírus da Dengue (DENV-1, DENV-2, etc.).

Tabela 11 – continuação da página anterior

Atributo	Tipo	É Classe?	Valores Únicos	Total Faltantes	% Faltantes	Descrição
HISTOPA_N	Numérico	Não	4	212353	52.99%	Resultado do exame histopatológico.
IMUNOH_N	Numérico	Não	4	212108	52.93%	Resultado da imunohistoquímica.
HOSPITALIZ	Numérico	Não	3	66690	16.64%	Indicador de hospitalização do paciente.
UF	Numérico	Não	26	388563	96.96%	Unidade Federativa (Estado) de residência do paciente.
CLASSI_FIN	Numérico	Não	4	17	0%	Classificação final do caso (ex: Dengue Clássica, Dengue com Sinais de Alarme).
CRITERIO	Numérico	Não	3	717	0.18%	Critério de confirmação do caso (ex: laboratorial, clínico-epidemiológico).
DT_ENCERRA	Categórico	Não	472	701	0.17%	Data de encerramento da investigação do caso.
ALRM_HIPOT	Binário	Não	2	393944	98.31%	Presença de sinal de alarme: Hipotensão.
ALRM_PLAQ	Binário	Não	2	393943	98.31%	Presença de sinal de alarme: Plaquetopenia (baixa contagem de plaquetas).
ALRM_VOM	Binário	Não	2	393951	98.31%	Presença de sinal de alarme: Vômitos persistentes.
ALRM_SANG	Binário	Não	2	393946	98.31%	Presença de sinal de alarme: Sangramento de mucosas (nariz, gengiva).
ALRM_HEMAT	Binário	Não	2	393966	98.31%	Presença de sinal de alarme: Hematomas.
ALRM_ABDOM	Binário	Não	2	393942	98.3%	Presença de sinal de alarme: Dor abdominal intensa e contínua.
ALRM_LETAR	Binário	Não	2	393959	98.31%	Presença de sinal de alarme: Letargia ou irritabilidade.
ALRM_HEPAT	Binário	Não	2	393964	98.31%	Presença de sinal de alarme: Hepatomegalia (fígado aumentado).
ALRM_LIQ	Binário	Não	2	393964	98.31%	Presença de sinal de alarme: Acúmulo de líquidos (derrame pleural, ascite).
GRAV_PULSO	Binário	Não	2	400052	99.83%	Presença de sinal de gravidade: Pulso débil ou ausente.
GRAV_CONV	Binário	Não	2	400052	99.83%	Presença de sinal de gravidade: Convulsões.
GRAV_ENCH	Binário	Não	2	400053	99.83%	Presença de sinal de gravidade: Enchimento capilar lento (> 2 seg).
GRAV_INSUF	Binário	Não	2	400053	99.83%	Presença de sinal de gravidade: Insuficiência respiratória.
GRAV_TAQUI	Binário	Não	2	400051	99.83%	Presença de sinal de gravidade: Taquicardia.
GRAV_EXTRE	Binário	Não	2	400051	99.83%	Presença de sinal de gravidade: Extremidades frias e cianóticas.
GRAV_HIPOT	Binário	Não	2	400052	99.83%	Presença de sinal de gravidade: Hipotensão.
GRAV_HEMAT	Binário	Não	2	400052	99.83%	Presença de sinal de gravidade: Sangramento grave.

Tabela 11 – continuação da página anterior

Atributo	Tipo	É Classe?	Valores Únicos	Total Faltantes	% Faltantes	Descrição
GRAV_MELEN	Binário	Não	2	400052	99.83%	Presença de sinal de gravidade: Melena (sangue nas fezes).
GRAV_METRO	Binário	Não	2	400052	99.83%	Presença de sinal de gravidade: Metrorragia (sangramento uterino anormal).
GRAV_SANG	Binário	Não	2	400052	99.83%	Presença de sinal de gravidade: Sangramentos (geral).
GRAV_AST	Binário	Não	2	400053	99.83%	Presença de sinal de gravidade: Ascite (acúmulo de líquido no abdome).
GRAV_MIOC	Binário	Não	2	400052	99.83%	Presença de sinal de gravidade: Miocardite (inflamação do músculo cardíaco).
GRAV_CONSC	Binário	Não	2	400051	99.83%	Presença de sinal de gravidade: Alteração da consciência.
GRAV_ORGAO	Binário	Não	2	400053	99.83%	Presença de sinal de gravidade: Disfunção de múltiplos órgãos.
Obito	Binário	Sim	2	0	0%	Variável alvo: Indica se o paciente foi a óbito (0=Não, 1=Sim).
QUANT_DIAS	Numérico	Não	104	285170	71.16%	Quantidade de dias desde o início dos sintomas até o desfecho.

Fonte: Elaborado pela autora.

6.0.2 Descrição do *Dataset* 2024

Tabela 12 – Características dos Atributos do Dataset de Dengue - 2024

Atributo	Tipo	É Classe?	Valores Únicos	Total Faltantes	% Faltantes	Descrição
NU_IDADE_N	Numérico	Não	100	0	0.0%	Idade do paciente no momento da notificação.
SG_UF_NOT	Numérico	Não	18	0	0.0%	Sigla da Unidade Federativa (Estado) de notificação do caso.
CS_SEXO	Categórico	Não	3	1	0.0%	Sexo biológico do paciente.
CS_GESTANT	Numérico	Não	7	53	0.02%	Situação de gravidez (apenas para o sexo feminino).
CS_RACA	Numérico	Não	6	2	0.0%	Raça/Cor autodeclarada do paciente.
FEBRE	Binário	Não	2	0	0.0%	Presença de febre como sintoma.
MIALGIA	Binário	Não	2	0	0.0%	Presença de mialgia (dor muscular) como sintoma.
CEFALEIA	Binário	Não	2	0	0.0%	Presença de cefaleia (dor de cabeça) como sintoma.
EXANTEMA	Binário	Não	2	0	0.0%	Presença de exantema (erupção cutânea) como sintoma.
VOMITO	Binário	Não	2	0	0.0%	Presença de vômito como sintoma.

Tabela 12

Atributo	Tipo	É Classe?	Valores Únicos	Total Faltantes	% Faltantes	Descrição
NAUSEA	Binário	Não	2	0	0.0%	Presença de náusea como sintoma.
DOR_COSTAS	Binário	Não	2	0	0.0%	Presença de dor nas costas como sintoma.
CONJUNTVIT	Binário	Não	2	0	0.0%	Presença de conjuntivite como sintoma.
ARTRITE	Binário	Não	2	0	0.0%	Presença de artrite (inflamação nas articulações) como sintoma.
ARTRALGIA	Binário	Não	2	0	0.0%	Presença de artralgia (dor nas articulações) como sintoma.
PETEQUIA_N	Binário	Não	2	0	0.0%	Presença de petéquias (pequenas manchas vermelhas na pele) como sintoma.
LEUCOPENIA	Binário	Não	2	0	0.0%	Presença de leucopenia (baixa contagem de glóbulos brancos) como sintoma.
LACO	Binário	Não	2	0	0.0%	Resultado da prova do laço (teste para fragilidade capilar).
DOR_RETRO	Binário	Não	2	0	0.0%	Presença de dor retro-orbital (atrás dos olhos) como sintoma.
DIABETES	Binário	Não	2	0	0.0%	Presença de comorbidade: Diabetes.
HEMATOLOG	Binário	Não	2	0	0.0%	Presença de comorbidade: Doença hematológica.
HEPATOPAT	Binário	Não	2	0	0.0%	Presença de comorbidade: Hepatopatia (doença do fígado).
RENAL	Binário	Não	2	0	0.0%	Presença de comorbidade: Doença renal.
HIPERTENSA	Binário	Não	2	0	0.0%	Presença de comorbidade: Hipertensão.
ACIDO_PEPT	Binário	Não	2	0	0.0%	Presença de comorbidade: Doença ácido péptica (úlcera, gastrite).
AUTO_IMUNE	Binário	Não	2	0	0.0%	Presença de comorbidade: Doença autoimune.
RESUL_SORO	Numérico	Não	4	155358	50.71%	Resultado do exame sorológico para Dengue.
RESUL_NS1	Numérico	Não	4	122592	40.02%	Resultado do teste rápido NS1 para Dengue.
RESUL_VI_N	Numérico	Não	4	181944	59.43%	Resultado do isolamento viral.
RESUL_PCR_	Numérico	Não	4	185223	60.51%	Resultado do RT-PCR (Reação em Cadeia da Polimerase) para Dengue.
SOROTIPO	Numérico	Não	4	199822	65.90%	Sorotipo do vírus da Dengue (DENV-1, DENV-2, etc.).
HISTOPA_N	Numérico	Não	4	198822	64.90%	Resultado do exame histopatológico.
IMUNOH_N	Numérico	Não	4	199654	64.98%	Resultado da imunohistoquímica.
HOSPITALIZ	Numérico	Não	3	29008	9.47%	Indicador de hospitalização do paciente.
UF	Numérico	Não	8	300806	98.19%	Unidade Federativa (Estado) de residência do paciente.
CLASSI_FIN	Numérico	Não	4	96	0.03%	Classificação final do caso (ex: Dengue Clássica, Dengue com Sinais de Alarme).

Tabela 12

Atributo	**Tipo**	**É Classe?**	**Valores Únicos**	**Total Faltantes**	**% Faltantes**	**Descrição**
CRITERIO	Numérico	Não	3	647	0.21%	Critério de confirmação do caso (ex: laboratorial, clínico-epidemiológico).
DT_ENCERRA	Categórico	Não	211	812	0.27%	Data de encerramento da investigação do caso.
ALRM_HIPOT	Binário	Não	2	302014	98.59%	Presença de sinal de alarme: Hipotensão.
ALRM_PLAQ	Binário	Não	2	302010	98.59%	Presença de sinal de alarme: Plaquetopenia (baixa contagem de plaquetas).
ALRM_VOM	Binário	Não	2	302022	98.59%	Presença de sinal de alarme: Vômitos persistentes.
ALRM_SANG	Binário	Não	2	302009	98.59%	Presença de sinal de alarme: Sangramento de mucosas (nariz, gengiva).
ALRM_HEMAT	Binário	Não	2	302032	98.59%	Presença de sinal de alarme: Hematomas.
ALRM_ABDOM	Binário	Não	2	301994	98.58%	Presença de sinal de alarme: Dor abdominal intensa e contínua.
ALRM_LETAR	Binário	Não	2	302026	98.59%	Presença de sinal de alarme: Letargia ou irritabilidade.
ALRM_HEPAT	Binário	Não	2	302038	98.60%	Presença de sinal de alarme: Hepatomegalia (fígado aumentado).
ALRM_LIQ	Binário	Não	2	302034	98.60%	Presença de sinal de alarme: Acúmulo de líquidos (derrame pleural, ascite).
GRAV_PULSO	Binário	Não	2	305838	99.83%	Presença de sinal de gravidade: Pulso débil ou ausente.
GRAV_CONV	Binário	Não	2	305838	99.83%	Presença de sinal de gravidade: Convulsões.
GRAV_ENCH	Binário	Não	2	305831	99.83%	Presença de sinal de gravidade: Enchimento capilar lento (> 2 seg).
GRAV_INSUF	Binário	Não	2	305831	99.83%	Presença de sinal de gravidade: Insuficiência respiratória.
GRAV_TAQUI	Binário	Não	2	305829	99.83%	Presença de sinal de gravidade: Taquicardia.
GRAV_EXTRE	Binário	Não	2	305829	99.83%	Presença de sinal de gravidade: Extremidades frias e cianóticas.
GRAV_HIPOT	Binário	Não	2	305831	99.83%	Presença de sinal de gravidade: Hipotensão.
GRAV_HEMAT	Binário	Não	2	305831	99.83%	Presença de sinal de gravidade: Sangramento grave.
GRAV_MELEN	Binário	Não	2	305831	99.83%	Presença de sinal de gravidade: Melena (sangue nas fezes).
GRAV_METRO	Binário	Não	2	305830	99.83%	Presença de sinal de gravidade: Metrorragia (sangramento uterino anormal).
GRAV_SANG	Binário	Não	2	305831	99.83%	Presença de sinal de gravidade: Sangramentos (geral).

Tabela 12

Atributo	**Tipo**	**É Classe?**	**Valores Únicos**	**Total Faltantes**	**% Faltantes**	**Descrição**
GRAV_AST	Binário	Não	2	305831	99.83%	Presença de sinal de gravidade: Ascite (acúmulo de líquido no abdome).
GRAV_MIOC	Binário	Não	2	305831	99.83%	Presença de sinal de gravidade: Miocardite (inflamação do músculo cardíaco).
GRAV_CONSC	Binário	Não	2	305830	99.83%	Presença de sinal de gravidade: Alteração da consciência.
GRAV_ORGAO	Binário	Não	2	305831	99.83%	Presença de sinal de gravidade: Disfunção de múltiplos órgãos.
Obito	Binário	Sim	2	0	0.0%	Variável alvo: Indica se o paciente foi a óbito (0=Não, 1=Sim).
QUANT_DIAS	Numérico	Não	148	0	0.0%	Quantidade de dias desde o início dos sintomas até o desfecho.

Fonte: Elaborado pela autora.

Referências

ALPAYDIN, E. *Introdução ao aprendizado de máquina*. São Paulo: Editora Blucher, 2020.

ALVES, G. O. et al. Estudo comparativo de técnicas de previsão para casos de dengue. *Revista de Engenharia e Pesquisa Aplicada*, v. 6, n. 3, p. 12–20, 2021.

BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, v. 6, n. 1, p. 20–29, 2004.

BECHT, E. et al. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, Nature Publishing Group, v. 37, n. 1, p. 38–44, 2019.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

BOHM, B. C. et al. Utilization of machine learning for dengue case screening. *BMC Public Health*, Springer, v. 24, n. 1, p. 1573, 2024.

BORGES, M. G. et al. Aspectos clínicos da dengue em crianças e perspectivas quanto às vacinas no brasil. *Brazilian Journal of Health Review*, v. 6, n. 6, p. 3580–3589, 2023.

BRADLEY, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, v. 30, n. 7, p. 1145–1159, 1997. ISSN 0031-3203. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0031320396001422⟩.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

BREIMAN, L. et al. *Classification and regression trees*. [S.l.]: Routledge, 2017.

BREUNIG, M. M. et al. Lof: identifying density-based local outliers. In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. [S.l.: s.n.], 2000. p. 93–104.

CABRERA, M. et al. Dengue prediction in latin america using machine learning and the one health perspective: a literature review. *Journal of Tropical Medicine and Infectious Disease*, v. 7, n. 10, p. 322, 2022. Disponível em: ⟨https://doi.org/10.3390/tropicalmed7100322⟩.

CARDOSO, R. L. Dengue no brasil: uma revisão sistemática. *Revista Foco*, v. 17, n. 3, p. e4640, mar. 2024.

CHAWLA, N. V. et al. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002. Disponível em: ⟨https://www.jair.org/index.php/jair/article/view/10302/24590⟩.

CHAWLA, N. V.; JAPKOWICZ, N.; KOTCZ, A. Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, ACM, v. 6, n. 1, p. 1–6, 2004.

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*. [S.l.: s.n.], 2016. p. 785–794.

CHENG, P.; ZOU, X.; DONG, F. A two-stage outlier detection method for large-scale datasets. *Journal of Big Data*, Springer, v. 6, n. 1, p. 1–17, 2019.

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE transactions on information theory*, IEEE, v. 13, n. 1, p. 21–27, 1967.

DAVI, C. et al. Severe dengue prognosis using human genome data and machine learning. *IEEE Transactions on Biomedical Engineering*, IEEE, v. 66, n. 10, p. 2861–2868, 2019.

DAVIS, J.; GOADRICH, M. The relationship between precision-recall and roc curves. In: *Proceedings of the 23rd International Conference on Machine Learning*. New York, NY, USA: Association for Computing Machinery, 2006. (ICML '06), p. 233–240. ISBN 1595933832. Disponível em: ⟨https://doi.org/10.1145/1143844.1143874⟩.

DIAS, J. L. *Aprendizado de máquina aplicado à predição de doenças crônicas: um estudo de caso de hipertensão arterial*. Dissertação (Dissertação (Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria)) — Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, São Carlos, 2024. Acesso em: 2 nov. 2025. Disponível em: ⟨https://www.teses.usp.br/teses/disponiveis/55/55137/tde-27012025-120434/publico/Jaqueline_Lopes_Dias_ME.pdf⟩.

DOMINGOS, P. A few useful things to know about machine learning. *Communications of the ACM*, ACM, v. 55, n. 10, p. 78–87, 2012.

DOMINGOS, P.; PAZZANI, M. On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, Springer, v. 29, p. 103–130, 1997.

FAWCETT, T. An introduction to roc analysis. *Pattern recognition letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006.

FERNÁNDEZ-DELGADO, M. et al. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, v. 15, n. 1, p. 3133–3181, 2014.

FRIEDRICH, S. et al. Regularization approaches in clinical biostatistics: A review of methods and their applications. *Statistical Methods in Medical Research*, SAGE Publications Sage UK: London, England, v. 32, n. 2, p. 425–440, 2023.

GROPPO M., e. a. Detecção de fraudes no consumo de Água empregando data science e o método shap. *Revista Brasileira de Recursos Hídricos*, v. 28, 2023.

GUO, P. et al. Developing a dengue forecast model using machine learning: A case study in china. *PLoS neglected tropical diseases*, Public Library of Science San Francisco, CA USA, v. 11, n. 10, p. e0005973, 2017.

HE, H.; GARCIA, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 21, n. 9, p. 1263–1284, 2009.

HERNANDEZ, J.; CARRASCO-OCHOA, J.; MARTíNEZ-TRINIDAD, J. F. An empirical study of oversampling and undersampling for instance selection methods on imbalance datasets. In: *Advances in Artificial Intelligence: MICAI 2013 International Conference*. [S.l.: s.n.], 2013. p. 262–269. ISBN 978-3-642-41821-1.

HUSSAIN, Z. et al. Machine learning approaches for dengue prediction: A review of algorithms and applications. *Pak. Geogr. Rev.*, v. 78, p. 15–36, 2023.

IBGE, I. B. de Geografia e E. *Censo Demográfico 2022: Panorama*. 2022. Acesso em: 12 de maio de 2025. Disponível em: ⟨https://censo2022.ibge.gov.br/panorama/⟩.

IBM. *O que é XGBoost?* 2024. ⟨https://www.ibm.com/br-pt/think/topics/xgboost⟩. Acesso em: 22 maio 2025.

ISLAM, M. R. et al. Machine learning with health information technology: Transforming data-driven healthcare systems. *Journal of Medical and Health Studies*, v. 4, n. 1, p. 89–96, 2023.

JACOBI, P. *Gestão participativa e meio ambiente urbano: desafios à governança democrática*. [S.l.]: Annablume, 2004.

JAPKOWICZ, N.; STEPHEN, S. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, IOS Press, v. 6, n. 5, p. 429–449, 2002.

LI, K.-L. et al. Improving one-class svm for anomaly detection. In: IEEE. *Proceedings of the 2003 international conference on machine learning and cybernetics (IEEE Cat. No. 03EX693)*. [S.l.], 2003. v. 5, p. 3077–3081.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, ACM New York, NY, USA, v. 6, n. 1, p. 1–39, 2012.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*. [S.l.: s.n.], 2017. p. 4765–4774.

MARON, M. E. Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM)*, ACM New York, NY, USA, v. 8, n. 3, p. 404–417, 1961.

MCINNES, L.; HEALY, J.; MELVILLE, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

MENDES R. F., e. a. Modelo preditivo de infecção hospitalar utilizando aprendizado de máquina. *Revista Brasileira de Saúde Ocupacional*, v. 48, 2023.

METZ, C. E. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, Elsevier, v. 8, n. 4, p. 283–298, 1978.

MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill, 1997.

ORGANIZATION, W. H. et al. *Dengue: guidelines for diagnosis, treatment, prevention and control*. [S.l.]: World Health Organization, 2009.

PIMENTEL, M. A. et al. A review of novelty detection. *Signal processing*, Elsevier, v. 99, p. 215–249, 2014.

POKRAJAC, D.; LAZAREVIC, A.; LATECKI, L. J. Incremental local outlier detection for data streams. In: IEEE. *2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007)*. [S.l.]: IEEE, 2007. p. 250–256.

ROSTER, K. I. O. *Data science for epidemiology: a case study of dengue in Brazil*. Tese (Doutorado) — Universidade de São Paulo, dez. 2022. Disponível em: ⟨https://www.teses.usp.br/teses/disponiveis/55/55134/tde-27022023-142607/⟩.

SAITO, T.; REHMSMEIER, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, Public Library of Science San Francisco, CA USA, v. 10, n. 3, p. e0118432, 2015.

SCHÖLKOPF, B. et al. Estimating the support of a high-dimensional distribution. *Neural computation*, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 13, n. 7, p. 1443–1471, 2001.

SCHÖLKOPF, B.; SMOLA, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. [S.l.]: MIT press, 2002.

scikit-learn developers. *scikit-learn: Machine Learning in Python*. 2024. ⟨https://scikit-learn.org/stable/⟩. Acesso em 22 de maio de 2025.

SHAW, R. *Xgboost: A concise technical overview*. 2017.

SILVEIRA, F. R. de V.; MOREIRA, L. Y. M. R. Utilização de algoritmos de aprendizagem de máquina na predição de arboviroses transmitidas pelo *aedes aegypti*. *Revista Conexões - Ciência E Tecnologia*, v. 14, n. 1, 2020. Disponível em: ⟨https://periodicos.iftm.edu.br/index.php/conexoes/article/view/1013/1028⟩.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Oxford University Press, v. 58, n. 1, p. 267–288, 1996.

TOMEK, I. Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6, n. 11, p. 769–772, 1976.

VAPNIK, V. *The nature of statistical learning theory*. 2nd. ed. [S.l.]: Springer science & business media, 2013. ISBN 978-1-4757-3264-1.