INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS CORRENTE

CURSO TECNÓLOGO EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

VINÍCIUS MARQUES BEZERRA

UM SISTEMA BASEADO EM APRENDIZADO DE MÁQUINA PARA ASSISTÊNCIA MÉDICA NO DIAGNÓSTICO DE DOENÇAS CARDIOVASCULARES

CORRENTE

2024

VINÍCIUS MARQUES BEZERRA

UM SISTEMA BASEADO EM APRENDIZADO DE MÁQUINA PARA ASSISTÊNCIA MÉDICA NO DIAGNÓSTICO DE DOENÇAS CARDIOVASCULARES

Trabalho de Conclusão de Curso (artigo científico) apresentado como exigência parcial para obtenção do diploma do Curso de Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Corrente.

Orientador: Prof. Jurandir Cavalcante Lacerda Júnior.

Coorientador: Prof. Eutino Junior Vieira Sirqueira.

CORRENTE

2024

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD

Bezerra, Vinicius Marques

B574s Um sistema baseado em aprendizado de máquina para assistência médicano diagnóstico de doenças cardiovasculares / Vinicius Marques Bezerra. - 2024.
34 p.: il. color.

Trabalho de Conclusão de Curso (Análise e Desenvolvimento de Sistemas) - Instituto Federal de Educação, Ciência e Tecnologia do Piauí,Campus Corrente, 2024.
Orientador : Prof Dr. Jurandir Cavalcante Lacerda Júnior..Coorientador
: Prof Esp. Eutino Junior Vieira Sirqueira..

1. Analise e Desenvolvimento de Sistemas - estudo e ensino. 2. Aprendizado de máquina. 3. Doenças cardiovasculares. 4. Modelos preditivos.I.Título.

CDD - 004

Elaborado por Tefischer Huanderson Soares e Sousa CRB 3/1251

UM SISTEMA BASEADO EM APRENDIZADO DE MÁQUINA PARA ASSISTÊNCIA MÉDICA NO DIAGNÓSTICO DE DOENÇAS CARDIOVASCULARES

Vinícius Marques Bezerra[1]
Jurandir Cavalcante Lacerda Júnior[2]
Eutino Junior Vieira Sirqueira[3]

RESUMO

Neste trabalho é investigado o diagnóstico precoce de doenças cardiovasculares (DCV) através da Aprendizagem de Máquina (AM), utilizando a base de dados da *University of California, Irvine* (UCI). Essa base foi validada durante este trabalho com análises que demonstraram consistência com as evidências encontradas na literatura médica. Dentre os modelos de AM utilizados, o *StackingCVClassifier*, integrando *XGBoost*, *K-Nearest Neighbors* (KNN) e *Support Vector Machine* (SVM), destacou-se com a maior acurácia, alcançando 92,64%. Além disso, para priorizar a acessibilidade em locais onde a infraestrutura é limitada, foi desenvolvida uma escala quantitativa para avaliar a dificuldade de coleta de atributos, adaptando o treinamento de modelos utilizando a base de dados com e sem atributos de difícil coleta. Por fim, para facilitar a implementação clínica, uma *Application Programming Interface* (API) e interface de usuário foram desenvolvidas. Dessa forma, este artigo avança na aplicação da AM para o diagnóstico de DCV, provando que a integração de tecnologia, dados validados e acessibilidade pode auxiliar na detecção precoce de condições de saúde complexas.

Palavras-chaves: modelos preditivos; doenças cardiovasculares; aprendizado de máquina.

ABSTRACT

In this paper, the early diagnosis of cardiovascular diseases (CVD) through Machine Learning (ML) is investigated, utilizing the database from the University of California, Irvine (UCI). This database was validated in the course of this work with analyses that demonstrated consistency with evidence found in medical literature. Among the ML models employed, the StackingCVClassifier, integrating XGBoost, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM), was highlighted for achieving the highest accuracy, at 92.64%. Furthermore, to prioritize accessibility in areas where infrastructure is limited, a quantitative scale for assessing the difficulty of attribute collection was developed, adapting model training using the database with and without difficult-to-collect attributes. Lastly, to aid clinical implementation, an Application Programming Interface (API) and user interface were developed. Therefore, this paper advances the application of ML in the diagnosis of CVD, proving that the integration of technology, validated data, and accessibility can aid in the early detection of complex health conditions.

K**eywords:** predictive models; cardiovascular diseases; machine learning.

Data de aprovação: 27/03/2024

[1]Acadêmico do curso de Tecnologia em Análise e Desenvolvimento de Sistemas. Instituto Federal do Piauí - IFPI. E-mail: CACOR.2021121TADS0010@aluno.ifpi.edu.br.

2 Professor Doutor. Instituto Federal do Piauí - IFPI. Orientador do TCC. E-mail: jurandirjr1@gmail.com. [3] Professor Especialista. Instituto Federal do Piauí - IFPI. Co-orientador do TCC. E-mail: eutino.junior@ifpi.edu.br.

1. INTRODUÇÃO

As doenças cardiovasculares (DCV) são um dos principais desafios em saúde pública global, ocasionando um considerável número de mortes e exercendo impacto significativo na morbidade mundial (WORLD HEALTH ORGANIZATION, 2021). Comumente conhecidas como doenças cardíacas, as DCV representam adversidades que afetam o coração e os vasos sanguíneos. A origem dessas patologias está muitas vezes associada a um bloqueio que estreita as artérias, resultando na obstrução do fluxo sanguíneo para o coração ou cérebro (WORLD HEALTH ORGANIZATION, 2021). Em 2019, estima-se que as DCV tenham sido responsáveis por 17,9 milhões de óbitos, correspondendo a 32% do total de mortes naquele ano (WORLD HEALTH ORGANIZATION, 2021). Notavelmente, ataques cardíacos e acidentes vasculares cerebrais representaram predominantemente 85% desses eventos fatais (WORLD HEALTH ORGANIZATION, 2021).

Diante desse cenário, torna-se evidente a necessidade de desenvolver estratégias que possam auxiliar na prevenção e no tratamento das DCV. No âmbito da Ciência da Computação, especificamente em um subcampo da Inteligência Artificial, a Aprendizagem de Máquina (AM) destaca-se na obtenção de informações a partir das bases de dados. De acordo com MÜMTAZ KARATAŞ et al. (2022), a crescente disponibilidade de dados em diversas formas tem impulsionado o papel da AM na informática médica ao longo da última década. A rápida expansão da capacidade de coleta e armazenamento de dados na área médica tem permitido avanços na compreensão de padrões, tendências e correlações, fornecendo uma base sólida para melhorias nos processos de diagnóstico, tratamento e prevenção de doenças.

Nesse contexto, são empregados algoritmos de AM, que extraem conhecimento de conjuntos de dados para realizar previsões a partir do treinamento de modelos estatísticos (BURKOV, 2019). Essa autonomia proporciona aos computadores a capacidade de aprender diretamente dos dados, eliminando a necessidade de os humanos extraírem regras manualmente. Esses modelos se apresentam como uma alternativa eficaz para diversas áreas profissionais, contribuindo para o aprimoramento progressivo da tomada de decisões fundamentadas em dados (BURKOV, 2019).

A utilização de modelos preditivos para auxiliar o diagnóstico precoce de DCV apresenta-se como uma estratégia promissora, possuindo a capacidade de unir dados clínicos estruturados e demais informações pertinentes. Utilizando técnicas de AM, esses modelos identificam padrões e relações complexas entre diferentes tipos de dados, fornecendo previsões objetivas e descobertas úteis sobre os mecanismos subjacentes às DCV, o que auxilia na identificação precoce de indivíduos em risco e no direcionamento de intervenções preventivas e terapêuticas adequadas (CLEOPHAS; ZWINDERMAN, 2015).

Neste artigo, foram explorados 10 modelos de AM para diagnosticar doenças cardíacas, utilizando a base de dados do repositório da UCI (UCI MACHINE LEARNING REPOSITORY, 2013). As técnicas utilizadas no treinamento dos modelos foram: Regressão Logística (*Logistic Regression* - LR), Árvore de Decisão (*Decision Tree* - DT), Floresta Aleatória (*Random Forest* - RF), *K-Nearest Neighbors* (KNN), *Naive Bayes* (NB), Máquina de Vetores de Suporte (*Support Vector Machine* - SVM), *Boosting* Adaptativo (*Adaptive Boosting* - AdaBoost), *eXtreme Gradient Boosting* (XGBoost) e *Extremely Randomized Trees* (*Extra Trees* - ET). Por fim, foi implementado o *Stacking Cross-Validated Classifier* (*StackingCVClassifier*), para integrar previsões de múltiplos modelos como classificadores base. A comparação detalhada deles proporcionou uma compreensão do desempenho, facilitando a escolha do modelo mais eficaz.

Além disso, foi realizada a classificação dos atributos da base de dados segundo a dificuldade de sua coleta, estabelecendo parâmetros de pontuação para determinar quais informações são mais desafiadoras do que outras. Esta categorização considera aspectos práticos fundamentais, como a disponibilidade dos exames no município de Corrente-PI, os custos associados e o tempo necessário para realizar os métodos de extração. Segundo MOURA et al. (2019) frequentemente, a escassa disponibilidade de certos exames pode limitar a coleta de informações, ao passo que custos elevados podem ser um obstáculo para que os pacientes obtenham dados importantes para o diagnóstico

Após identificar os atributos de maior dificuldade de extração, foi realizado um novo treinamento dos modelos com a mesma base de dados omitindo atributos categorizados com pontuação maior que 3 em dificuldade de obtenção. Este procedimento busca simplificar o uso dos modelos para os usuários finais, minimizando a quantidade de dados necessários para gerar previsões confiáveis, priorizando uma aplicação mais prática e menos onerosa em termos de coleta de dados.

Prosseguindo na análise dos atributos, este estudo confirmou a coerência e a relevância dos registros na base de dados, especialmente ao evidenciar padrões que são consistentes com as expectativas clínicas e epidemiológicas conhecidas. Destacando a importância de certos biomarcadores, como o colesterol, cujos altos níveis estão diretamente relacionados ao aumento do risco de DCV (JUNG et al., 2022). A correspondência desses padrões com a literatura científica médica valida a configuração da base de dados, assegurando que as informações são coerentes para a construção de modelos preditivos.

De forma simultânea ao uso do algoritmos de AM, foi desenvolvida uma *Application Programming Interface* (API) para utilizar os modelos com mais facilidade. Esta API oferece uma interface de usuário final, que segue os seis princípios de design de NORMAN (2004): visibilidade, *feedback*, consistência, prevenção de erros, flexibilidade, eficiência, ajuda e documentação. A inserção de informações é realizada por meio de um formulário estruturado em conformidade com os princípios mencionados. O que possibilita a submissão eficiente dos dados, gerando uma predição instantânea em formato percentual, que sinaliza a presença ou ausência de indícios de DCV.

Este artigo está dividido nas seguintes seções: A Seção 2 são revisados os trabalhos relacionados, destacando avanços na literatura. Na Seção 3, detalha-se a metodologia, abordando a seleção e processamento da base de dados do repositório da UCI (UCI MACHINE LEARNING REPOSITORY, 2013). Os resultados e discussões são apresentados na Seção 4, proporcionando uma compreensão do desempenho dos modelos treinados. Por fim, a Seção 5 são discutidas as considerações finais e sugere direções para futuras pesquisas.

2. TRABALHOS RELACIONADOS

Na busca contínua pela melhoria da eficácia diagnóstica e prognóstica das DCV, diversos estudos têm se dedicado ao emprego de técnicas avançadas de AM e modelagem preditiva. MOHAN; THIRUMALAI; SRIVASTAVA (2019) propõem um método híbrido de floresta aleatória com um modelo linear (*hybrid random forest with a linear model* - HRFLM) para melhorar a precisão da previsão de doenças cardíacas. Esse método é aplicado no conjunto de dados *UCI Heart Diseases*, e os resultados mostram um nível de acurácia de 88,7% na previsão de doenças cardíacas. Os autores sugerem que futuros trabalhos possam explorar combinações diversas de técnicas de AM para aprimorar as técnicas de previsão por meio da combinação de modelos.

GUPTA et al. (2020) propõem um modelo, denominado MIFH (*Machine Intelligence Framework for Heart Disease Diagnosis*) que emprega a Análise Fatorial de Dados Mistos para extrair e derivar características do conjunto de dados *UCI Heart Diseases* sobre DCV, treinando modelos preditivos de AM. A validação do MIFH, realizada através do método de validação *holdout*, demonstra um bom desempenho, atingindo uma acurácia de 92,34%. Dessa forma, o MIFH foi capaz de dar suporte a radiologistas e médicos no diagnóstico de pacientes com DCVs.

Outro artigo relevante é o de HAMADA et al. (2021), que tem o objetivo de discriminar indivíduos saudáveis daqueles diagnosticados com DCV, visando elucidar uma compilação de fatores de risco potenciais associados à referida patologia. O estudo adotou uma coleção de medidas biomédicas que abrangem distintas variáveis, incluindo aquelas de natureza comportamental, provenientes do grupo de DCV do *Qatar Biobank* (QBB), um banco de dados com registros de pacientes. Como desfecho, os pesquisadores destacaram o *CatBoost* como o modelo mais eficaz, atingindo uma acurácia de 93%. Adicionalmente, foram identificados novos elementos na lista de fatores de risco, além dos já reconhecidos, tais como distúrbios renais, função hepática, aterosclerose, entre outros.

O estudo realizado por NILS STRODTHOFF et al. (2021), apresentou uma análise comparativa para o conjunto de dados clínicos de Eletrocardiografia (ECG), o PTB-XL (Wagner et al., 2020), introduzindo um recurso valioso para a análise automática de ECG. Eles implementaram modelos de Aprendizado Profundo (*Deep Learning* - DL), destacando o desempenho superior das redes neurais convolucionais, especialmente *ResNet* e *Inception*, em diversas tarefas. Esses resultados, complementados pela consideração da incerteza do modelo e interpretabilidade, reforçam o potencial dos algoritmos baseados em DL na análise de ECG.

Por fim, TIWARI; CHUGH; SHARMA (2022) propuseram uma nova abordagem para prever DCV com a utilização de um *framework ensemble* utilizando o conjunto de dados *UCI Heart Diseases* (UCI MACHINE LEARNING REPOSITORY, 2013), o qual englobou variáveis clínicas de grande importância, tais como a Frequência Cardíaca Máxima Atingida, Colesterol Sérico e tipos de Dor Torácica. O *framework* fez uso de um Classificador de *Ensemble* Empilhado (*Stacked Ensemble Classifier* - CEE), integrando vários algoritmos de AM. Durante a fase de avaliação do modelo desenvolvido, foram usadas diversas medidas de performance, Especificidade, Pontuação F1 (*F1 Score*), Sensibilidade e Acurácia. O modelo demonstrou uma acurácia de 92,44%.

Os estudos enfatizados nesta seção convergem para a utilização de técnicas de AM com o intuito de prever eventos na área da saúde, com foco expressivo em DCV. A tarefa de classificação emerge como um ponto comum entre eles. Todavia, este artigo distingue-se em relação aos trabalhos citados devido às seguintes contribuições:

I. Desenvolve uma escala de atributos baseada na dificuldade de coleta, refletindo a realidade de ambientes clínicos onde determinados exames não estão disponíveis.

II. Treina modelos de AM para classificação binária de DCV, considerando tanto a inclusão quanto a exclusão de variáveis de difícil coleta.

III. Realiza uma análise comparativa e causal, avaliando as relações entre variáveis como sexo e colesterol, com suporte da literatura médica.

IV. Inclui a construção de uma API e interface de usuário, seguindo os princípios de design de NORMAN (2004), para facilitar a adoção clínica.

A Tabela 1 apresenta uma comparação entre diversos trabalhos relacionados, evidenciando suas tarefas, bases de dados utilizadas e as técnicas de predição empregadas.

Tabela 1 - Tabela de comparação dos estudos.

Trabalho	Abordagem	Base de Dados	Métodos de previsão
(MOHAN; THIRUMALAI; SRIVASTAVA, 2019)	Classificação binária	*UCI Heart Diseases*	ANN, SVM, KNN, DT, LR e NB
(GUPTA et al., 2020)	Classificação binária	*UCI Heart Diseases*	LR, KNN, SVM e RF
(HAMADA et al., 2021)	Classificação binária	*Qatar Biobank*	DT, ANN, RF, XGBoost, CatBoost e LR
(STRODTHOFF et al., 2021)	Classificação binária, multiclasse e regressão	PTB-XL	Redes convolucionais e LSTM
(TIWARI; CHUGH; SHARMA, 2022)	Classificação binária	*UCI Heart Diseases*	ET, RF e *XGBoost*
Este trabalho	Classificação binária	*UCI Heart Diseases*	LR, DT, RF, KNN, NB, SVM, AdaBoost, XGBoost, ET, StackingCVClassifier

Fonte: elaborada pelos autores.

3. METODOLOGIA

Nesta seção, é detalhado o processo adotado no estudo, começando pela Subseção 3.1, Seleção de Dados, onde é explicada a escolha do conjunto de dados utilizado. A Subseção 3.2, Pré-processamento de Dados, segue com a descrição das etapas de preparação dos dados para análise. Na Subseção 3.3, Escolha dos Modelos, discutem-se os algoritmos de aprendizado de máquina selecionados para o estudo. A Subseção 3.4, Avaliação dos Modelos, detalha as métricas e métodos utilizados para avaliar e aprimorar a performance dos modelos escolhidos. Cada uma dessas subseções é projetada para proporcionar uma compreensão abrangente das técnicas e processos empregados na pesquisa.

a.OBTENÇÃO DA BASE DE DADOS

Neste estudo, foi utilizada a base de dados *UCI Heart Diseases* que reúne registros de pacientes de quatro instituições médicas de prestígio: *Cleveland Clinic Foundation*, *Hungarian Institute of Cardiology*, *V.A. Medical Center* e *University Hospital, Zurich* (UCI MACHINE LEARNING REPOSITORY, 2013). Essa escolha se deu pela qualidade e integridade dos dados disponíveis, para assegurar a confiabilidade das análises conduzidas. O conjunto de dados selecionado inclui informações de 303 pacientes, categorizados em cinco níveis de risco: 164 casos indicando ausência de doença cardíaca, 55 em risco baixo, 36 em risco médio, 35 em risco alto e 13 em risco grave. Para fins deste estudo, as categorias de risco foram adaptadas para uma análise binária, distinguindo entre a presença e a ausência de doença cardíaca, com base na recomendação de MÜLLER e GUIDO (2016).

Embora a base de dados contenha 76 atributos, apenas um subconjunto de 14 foi utilizado, considerados críticos para o diagnóstico de DCV. Essa primeira seleção foi baseada tanto na disponibilidade desses atributos no conjunto de dados quanto na sua recorrência em pesquisas anteriores, destacando-se os trabalhos de TIWARI; CHUGH; SHARMA (2022) e MOHAN; THIRUMALAI; SRIVASTAVA (2019). Os selecionados atributos são: idade, sexo, tipo de dor no peito, pressão arterial em repouso, colesterol sérico, nível de açúcar no sangue em jejum, resultados eletrocardiográficos em repouso, frequência cardíaca máxima alcançada, angina induzida por exercício, depressão do segmento ST induzida por exercício em relação ao repouso, a inclinação do segmento ST do exercício pico, número de vasos principais coloridos por fluoroscopia, defeito do talio e a presença de DCV como atributo de predição. Posteriormente é feita uma segunda seleção dentro desse subconjunto levando em consideração a dificuldade de coleta.

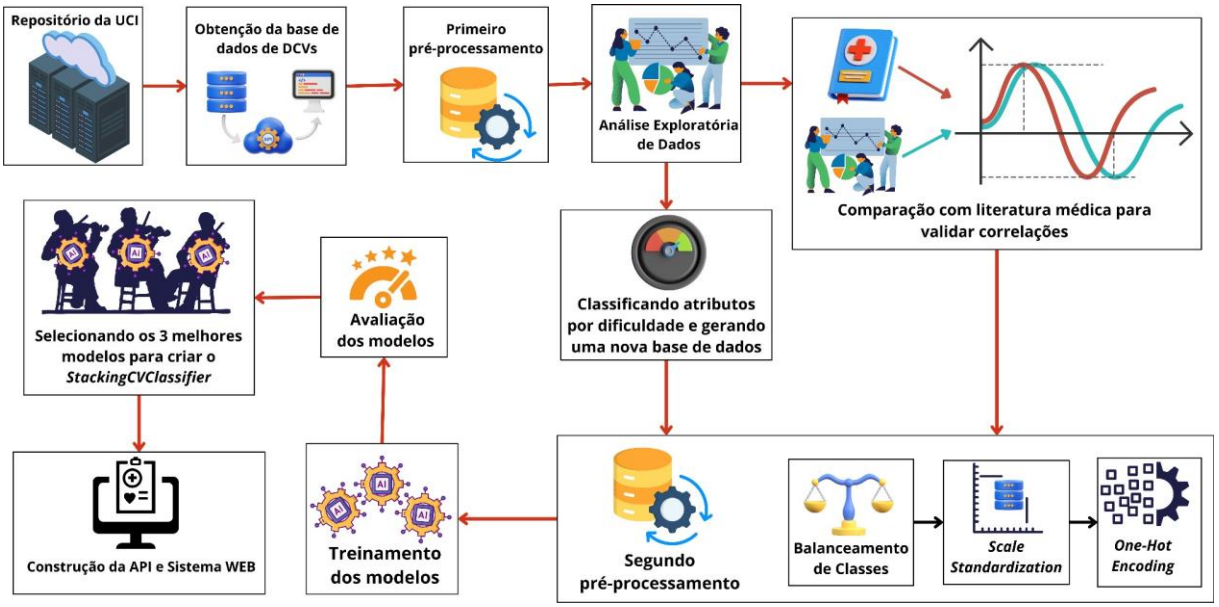
b.PRÉ-PROCESSAMENTO E ANÁLISE DE DADOS

Após a aquisição dos dados ocorreu o pré-processamento e a preparação dos dados. O objetivo foi melhorar a qualidade e a coerência dos dados para as fases posteriores de análise exploratória e modelagem. Primeiramente, utilizando a linguagem *Python (PYTHON SOFTWARE FOUNDATION, 2023)* e a biblioteca *Pandas (PANDAS - PYTHON DATA ANALYSIS LIBRARY, 2023)*, foram eliminadas as linhas que continham valores ausentes, garantindo a integridade das observações. Em seguida foi feita a remoção de duplicatas, assegurando a unicidade das instâncias no conjunto de dados.

Subsequente ao pré-processamento inicial, é realizada uma Análise Exploratória de Dados (*Exploratory Data Analysis* - EDA), utilizando as bibliotecas *Pandas* (*PANDAS - PYTHON DATA ANALYSIS LIBRARY*, 2023), *Seaborn* (*SEABORN: STATISTICAL DATA VISUALIZATION* - 0.13.0, 2023) e *Matplotlib* (*MATPLOTLIB*, 2023) para manipulação, análise e visualização dos dados. Foram elaboradas visualizações específicas para investigar as relações entre as variáveis, abordando aspectos como a distribuição de idade e níveis de colesterol por sexo com sua relação à presença de DCV.

A EDA forneceu uma visão inicial do conjunto de dados, conforme ilustrado na Figura 1, permitindo avaliar sua confiabilidade e natureza ao comparar com a literatura médica, para validar as correlações observadas. O objetivo foi verificar a consistência das correlações com o conhecimento pré-existente, evitando vieses ou inconsistências nos dados que possam comprometer as fases posteriores, evidenciando a validade e confiabilidade dos dados para atingir robustez e interpretabilidade dos resultados, que é fundamental na área de AM, conforme destacado por MÜLLER e GUIDO (2016). Esta etapa é importante para entender os fatores de risco de DCVs e suas implicações para prevenção e tratamento.

Figura 1 - Diagrama da metodologia.



Posteriormente, foi iniciada a segunda etapa de pré-processamento dos dados. Após a seleção inicial, foi utilizada a técnica de balanceamento de classes usando o método SMOTE (*Synthetic Minority Over-sampling Technique*) para evitar possíveis vieses introduzidos pela desproporcionalidade entre classes (CHAWLA et al., 2002). A técnica de *One-Hot Encoding* foi aplicada em variáveis categóricas específicas, segundo GÉRON (2019) sua utilização tem como objetivo representar adequadamente essas variáveis em formatos numéricos, permitindo que os modelos de AM compreendam e processassem essas informações. Em seguida a *Scale Standardization* foi realizada nas variáveis contínuas, garantindo a uniformidade e a equidade na contribuição de cada característica para o modelo (GÉRON, 2019). Essas etapas, que abrangeram desde a limpeza até a transformação das variáveis, contribuíram para a qualidade e a coerência dos dados destinados à análise e à modelagem.

a. MODELOS UTILIZADOS

Considerando a estrutura da base de dados foi feita a seleção de métodos para este estudo, que equilibra técnicas tradicionais, utilizadas pela sua clareza e facilidade de integração clínica, com abordagens avançadas, que ampliam a detecção de padrões complexos nos dados. Esta estratégia garante modelos interpretáveis e predições precisas, otimizando a capacidade de lidar com a complexidade e a variedade dos dados clínicos, visando maximizar a eficácia do sistema de previsão.

i. LR

A LR é reconhecida por sua eficácia em resolver problemas de classificação, atuando através da modelagem da probabilidade de uma variável dependente, particularmente útil para decisões binárias, a principal vantagem da LR reside na sua capacidade de fornecer uma probabilidade clara e interpretação para a ocorrência de um evento específico, conforme descrito por HOSMER; LEMESHOW; STURDIVANT (2013).

A escolha da LR para este estudo baseia-se em sua robustez e simplicidade, tanto em termos de implementação quanto de interpretação. HOSMER; LEMESHOW; STURDIVANT (2013) afirmam que a técnica é eficiente mesmo quando inclui variáveis que não afetam

significativamente o resultado, facilitando o processo de modelagem. O treinamento do modelo foca no ajuste dos coeficientes para refletir com precisão as probabilidades das classes observadas, geralmente utilizando o método do gradiente descendente. Uma vez treinado, o modelo é capaz de prever a probabilidade de novas instâncias pertencerem a uma determinada classe, contribuindo para uma tomada de decisão informada e baseada em dados. Esta característica torna a Regressão Logística uma ferramenta amplamente aplicada em pesquisas focadas em classificação binária, destacando a importância de sua utilização conforme HOSMER; LEMESHOW; STURDIVANT (2013).

ii. DT

As DT são reconhecidas por sua capacidade de abordar tanto problemas de classificação quanto de regressão, destacando-se como ferramentas preditivas devido à sua estrutura intuitiva, que replica o processo de tomada de decisão humana. Segundo CHANDRA e P. PAUL VARGHESE (2009), cada componente da árvore - seja um nó interno que testa um atributo, uma ramificação que representa o resultado de um teste, ou um nó folha que indica uma classe - desempenha um papel importante na determinação da classificação ou do valor de regressão de uma observação.

O processo de desenvolvimento de uma DT inicia com a identificação do melhor atributo, que serve como nó raiz, progredindo através de divisões subsequentes até que cada nó folha represente uma classe distinta ou um valor de regressão. Este método é valorizado por sua abordagem não paramétrica, permitindo uma flexibilidade notável no modelamento das relações entre as variáveis sem a necessidade de premissas sobre sua distribuição. WU e KUMAR (2009) enfatizam a importância da seleção de atributos baseada em métricas como a Entropia e o Ganho de Informação, fundamentais para otimizar a homogeneidade dos grupos resultantes e maximizar a eficácia da árvore na classificação e regressão.

Além disso, a simplicidade conceitual e a transparência das Árvores de Decisão facilitam sua interpretação e explicação, tornando-as atraentes para uma ampla gama de aplicações práticas. Essas características, combinadas com a capacidade de lidar com dados complexos e a flexibilidade para se adaptar a diferentes tipos de problemas analíticos, justificam sua seleção e uso frequente em pesquisas e aplicações de aprendizado de máquina, conforme discutido por CHANDRA (2009) e WU; KUMAR (2009).

iii. RF

O RF emerge como uma metodologia de AM que incorpora múltiplas DTs para formar um modelo preditivo coletivo, com o objetivo de aproveitar a diversidade das DTs individuais, que são treinadas de maneira não correlacionada, para criar um sistema que supera as limitações de um único modelo, oferecendo previsões mais precisas e confiáveis, conforme descrito por BREIMAN (2001).

Um aspecto central da RF, destacado por HASTIE, TIBSHIRANI, e FRIEDMAN (2009), é seu uso eficaz de técnicas de conjunto, como a agregação de *bootstrap*, para melhorar a capacidade de generalização do modelo. Isso é alcançado ao combinar os resultados das diversas árvores, cada uma contribuindo com sua previsão única, resultando em uma decisão final baseada na média dessas previsões. Esse processo não apenas aumenta a precisão na análise de dados, mas também reduz o risco de sobreajuste, tornando o RF apropriado para aplicações em que a robustez e a precisão são críticas.

Sua aplicabilidade em classificação e regressão, como descrito por BREIMAN (2001),não apenas melhora a precisão das previsões e o torna versátil, mas também aumenta a generalização do modelo em diferentes conjuntos de dados. Tal abordagem é fundamental para sua ampla aceitação e uso contínuo em pesquisas e aplicações práticas, consolidando a RF como uma ferramenta indispensável no campo da AM, conforme corroborado pela análise de HASTIE, TIBSHIRANI, e FRIEDMAN (2009).

iv. KNN

O KNN é classificado como um método não paramétrico, uma vez que não faz suposições sobre a distribuição dos dados (HART; STORK; DUDA, 2000). Ele determina a classificação de novos dados com base na proximidade e semelhança a dados já conhecidos, atribuindo-os à classe dos dados mais próximos. É usado tanto para problemas de regressão quanto para problemas de classificação, conhecido como o algoritmo do "aprendiz preguiçoso", pois não aprende imediatamente a partir de um conjunto de dados de treinamento (AHA; KIBLER; ALBERT, 1991). O KNN calcula a distância Euclidiana entre os novos dados A (x1, y1) e os dados previamente acessíveis B(x2, y2), usando a equação

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (1)$$

HART; STORK; DUDA (2000) descreve que a fórmula Euclidiana pode ser usada para avaliar a distância entre dois pontos de dados (x2, x1) e (y2, y1) em um espaço bidimensional, onde KNN coloca os novos dados na classe que tem a menor distância Euclidiana em relação aos novos dados. Apesar da complexidade computacional que torna o KNN inadequado para implementação em ambientes de baixa potência ou em tempo real, a técnica é fundamental nos estudos de aprendizado de máquina devido à sua capacidade de fornecer resultados precisos em classificação, conforme destacado por HART, STORK, e DUDA (2000).

v. NB

O NB é um algoritmo de classificação baseado no teorema de Bayes, com a "ingenuidade" de assumir independência entre os preditores. Apesar de sua simplicidade, o NB demonstrou ser eficaz em várias tarefas de classificação, especialmente em processamento de linguagem natural e análise de sentimentos, de acordo com LOWD; DOMINGOS (2005).

O teorema de Bayes é usado pelo NB para calcular a probabilidade posterior de uma classe, dada uma observação, conforme descrito por LOWD; DOMINGOS (2005). A fórmula básica do teorema de Bayes é:

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)} \quad (2)$$

Onde:

- P(C|X) é a probabilidade posterior da classe (C, alvo) dado o preditor (X, atributos).
- P(C) é a probabilidade de prioridade da classe.
- P(X|C) é a probabilidade de ver os preditores dados a classe.
- P(X) é a probabilidade de prioridade do preditor.

O NB faz a classificação calculando a probabilidade de cada classe para uma dada observação e selecionando a classe com a maior probabilidade. O modelo é chamado de "naive" (ingênuo) porque assume que os atributos são independentes entre si, uma suposição que raramente se verifica na prática, de acordo com LOWD; DOMINGOS (2005). No entanto, essa simplificação contribui para a eficiência do modelo, permitindo-lhe lidar com grandes volumes de dados rapidamente, como mencionado por LOWD; DOMINGOS (2005).

Existem diferentes variantes do NB, dependendo do tipo de distribuição dos dados de entrada, incluindo o Gaussian NB para dados com uma distribuição normal, Multinomial NB para dados de contagem, e Bernoulli NB para variáveis binárias, de acordo com LOWD; DOMINGOS (2005).

vi. SVM

O SVM é um algoritmo de aprendizado supervisionado amplamente utilizado para tarefas de classificação e regressão (MAMMONE; TURCHI; CRISTIANINI, 2009). O objetivo principal do SVM é encontrar o hiperplano que melhor separa as classes de dados em um espaço de características multidimensionais. O hiperplano é escolhido de maneira a maximizar a margem entre as classes, onde a margem é definida como a distância entre o hiperplano e os pontos de dados mais próximos de cada classe, conhecidos como vetores de suporte (MAMMONE; TURCHI; CRISTIANINI, 2009).

Em espaços onde as classes não são linearmente separáveis, o SVM pode ser aplicada usando o truque do *kernel* para mapear os dados para um espaço de características de maior dimensão onde a separação linear é possível, com os *kernel*s mais comuns sendo o linear, polinomial, radial (RBF) e sigmoidal (MAMMONE; TURCHI; CRISTIANINI, 2009).

A SVM é conhecida por sua eficácia em espaços de alta dimensão e quando o número de dimensões é maior que o número de amostras. Além disso, é versátil na escolha de funções de *kernel*, o que a torna adequada para uma variedade de problemas de classificação, que é destacado por MAMMONE; TURCHI; CRISTIANINI (2009).

vii. AdaBoost

O *AdaBoost* é um algoritmo de ensemble que melhora o desempenho de modelos de classificação combinando múltiplos classificadores fracos para criar um classificador forte e robusto (MARGINEANTU; DIETTERICH, 1997). Um classificador fraco é definido como um modelo que tem um desempenho ligeiramente melhor do que a adivinhação aleatória. A ideia central do AdaBoost é ajustar os pesos dos classificadores fracos em cada iteração para se concentrar nos casos mais difíceis, ou seja, nas instâncias de treinamento que foram previamente classificadas incorretamente (MARGINEANTU; DIETTERICH, 1997).

O processo do *AdaBoost* começa com a atribuição de pesos iguais a todas as instâncias de treinamento. Em cada etapa do algoritmo, um classificador fraco é treinado e é calculado o erro de classificação ponderado desse classificador. Baseando-se no erro, os pesos das instâncias são atualizados, de modo que os pesos das instâncias mal classificadas aumentem e os pesos das bem classificadas diminuam. Esse processo é repetido por um número pré-definido de iterações ou até que uma condição de parada específica seja atendida. O resultado final é uma combinação ponderada dos classificadores fracos que forma o modelo final (MARGINEANTU; DIETTERICH, 1997).

A fórmula para atualizar os pesos das instâncias é dada por:

$$peso\_novo = peso\_antigo \times exp(\alpha \times erro) \quad (3)$$

Onde:

α é a taxa de aprendizado, calculada com base no desempenho do classificador fraco.

erro é 1 se a classificação estiver incorreta e 0 se estiver correta.

O *AdaBoost* é particularmente eficaz para problemas de classificação binária e tem sido amplamente utilizado devido à sua simplicidade e eficácia (MARGINEANTU; DIETTERICH, 1997). Ele pode ser usado com qualquer tipo de modelo de classificação como classificador fraco, embora árvores de decisão com profundidade limitada sejam comumente utilizadas.

Um dos principais benefícios do *AdaBoost* é sua capacidade de melhorar a precisão de modelos simples sem exigir ajustes complicados de parâmetros ou extenso pré-processamento de dados. No entanto, é importante notar que, embora o *AdaBoost* possa ser resistente ao *overfitting* em algumas situações, ele pode ser sensível a dados ruidosos e outliers (MARGINEANTU; DIETTERICH, 1997).

viii. XGBoost

O *XGBoost*, uma extensão do *Gradient Boosting* (GB), é uma das técnicas mais eficientes e poderosas de aprendizado de máquina para tarefas de regressão e classificação, destacando-se por sua capacidade de construir modelos preditivos altamente precisos e robustos (NATEKIN; KNOLL, 2013). Desenvolvido como uma implementação otimizada do GB, o *XGBoost* melhora a eficiência computacional e a eficácia preditiva, gerenciando recursos como paralelismo e manuseio de dados esparsos de maneira mais eficiente.

A técnica *XGBoost* tem como base a ideia de combinar uma série de modelos preditivos fracos, geralmente árvores de decisão, em um modelo forte e coeso através de um processo iterativo de aprendizado (NATEKIN; KNOLL, 2013). Este processo inicia com um modelo base simples e, em cada etapa subsequente, ajusta-se adicionando novos modelos que corrigem os erros cometidos pelos modelos anteriores, focando especialmente nos resíduos ou nas previsões incorretas.

O coração do *XGBoost* reside em sua capacidade de minimizar uma função de perda específica, que quantifica a diferença entre as previsões do modelo e os resultados reais, ao mesmo tempo em que regulariza a complexidade do modelo para evitar o sobreajuste. Isso é feito aplicando o gradiente da função de perda para otimizar os parâmetros do modelo, um passo crítico que direciona a melhoria contínua da precisão do modelo (NATEKIN; KNOLL, 2013).

Os principais passos no treinamento do XGBoost incluem:

1. Inicialização com um modelo base que fornece uma estimativa inicial para todas as observações.
2. Iterativamente, cada novo modelo é treinado para corrigir os erros do modelo agregado anteriormente.
3. Aplicação de regularização para controlar a complexidade do modelo, evitando assim o sobreajuste.
4. Otimização da função de perda usando técnicas avançadas como o gradiente estocástico e a otimização baseada em árvore.

O XGBoost é notável por sua eficácia em espaços de alta dimensão e grandes volumes de dados, fornecendo resultados superiores em comparação a outras técnicas de GB, graças a otimizações como a computação paralela, tratamento de dados faltantes e a capacidade de personalizar a função de perda para tarefas específicas (NATEKIN; KNOLL, 2013). Essas características tornam o XGBoost uma escolha preferencial para competições de ciência de dados e aplicações industriais onde a precisão preditiva é de suma importância.

ix. ET

O método ET é um algoritmo de ensemble que pertence à família dos métodos de ensemble baseados em árvores, como as Florestas Aleatórias (RFs). A principal característica que distingue o ET do RF é o maior grau de aleatoriedade introduzido na construção das árvores de decisão, o que pode levar a uma maior diversidade entre as árvores e, potencialmente, a um desempenho melhorado em determinados conjuntos de dados (GEURTS; ERNST; WEHENKEL, 2006).

No ET, as árvores são construídas a partir do conjunto de dados completo. Ao contrário do RF, onde a melhor divisão entre os nós é selecionada a partir de um subconjunto aleatório de características, o ET escolhe pontos de divisão completamente ao acaso. Isso significa que, para cada divisão, tanto a característica a ser dividida quanto o valor de corte são escolhidos de forma aleatória, sem calcular o ganho de informação ou a redução da impureza (GEURTS; ERNST; WEHENKEL, 2006). Após a escolha aleatória do ponto de divisão, a divisão que melhor separa as classes ou minimiza a variância (para regressão) é selecionada para dividir o nó. Esse processo é repetido até que as árvores cresçam até sua profundidade máxima especificada ou até que não se possa mais reduzir a impureza nos nós.

Os principais passos do algoritmo ET são:

1. Para cada iteração de construção de árvore, use todo o conjunto de dados de treinamento, ao invés de uma amostra bootstrap como no RF.
2. Para cada nó da árvore, selecione uma característica e um ponto de divisão de forma aleatória, ao invés de buscar o melhor ponto de divisão.
3. Cresça as árvores até o máximo possível, sem aplicar poda.

Assim como outros métodos de ensemble baseados em árvores, o ET pode lidar eficazmente com variáveis categóricas e numéricas, tornando-o uma escolha versátil para uma ampla gama de problemas de classificação e regressão (GEURTS; ERNST; WEHENKEL, 2006).

x. StackingCVClassifier

O StackingCVClassifier é uma técnica avançada de ensemble que combina as previsões de múltiplos algoritmos de classificação para produzir um modelo final mais poderoso (SMYTH; WOLPERT, 1999). Diferente de métodos de ensemble tradicionais, como bagging ou boosting, o stacking utiliza um meta-classificador que aprende a melhor forma de combinar as previsões dos classificadores base, levando em consideração suas forças e fraquezas em diferentes partes do espaço de dados (SMYTH; WOLPERT, 1999).

No contexto deste estudo, o *StackingCVClassifier* foi construído utilizando três classificadores base distintos: KNN com 5 vizinhos, SVM com kernel linear e XGBoost com 1000 estimadores. Essa combinação diversificada busca explorar a capacidade de generalização de cada modelo e suas características únicas de decisão.

O processo de treinamento do *StackingCVClassifier* envolve o uso de validação cruzada, não apenas para prevenir o sobreajuste, mas também para garantir que o meta-classificador tenha acesso a previsões confiáveis sobre os dados de treinamento (SMYTH; WOLPERT, 1999). Este aspecto é importante, pois o meta-classificador, que nestecaso é um SVC (*Support Vector Classifier*), é treinado com as previsões dos classificadores base como entrada, ao invés dos dados originais. O objetivo é aprender a melhor maneira de integrar essas previsões para fazer a classificação final.

A implementação do *StackingCVClassifier* neste estudo inclui:

1. Classificadores Base: KNN, SVM Linear e XGBoost, selecionados por suas performances individuais e complementaridade.
2. Meta-classificador: SVC, escolhido por sua capacidade de lidar com as previsões combinadas dos classificadores base e realizar a classificação final.
3. Validação Cruzada: Usada tanto na etapa de treinamento dos classificadores base quanto do meta-classificador para maximizar a generalização.

O resultado final do *StackingCVClassifier* é obtido após o meta-classificador ser ajustado com as previsões dos classificadores base, resultando em um modelo que pode capturar padrões complexos e nuances que talvez sejam perdidos ou não totalmente explorados por um único modelo (SMYTH; WOLPERT, 1999).

Para a implementação e treinamento das técnicas descritas anteriormente, foi utilizada a biblioteca *Scikit-learn* (*SCIKIT-LEARN* 1.3.2, 2023) do Python (*PYTHON SOFTWARE FOUNDATION*, 2023), reconhecida pela sua abrangente coleção de algoritmos de Aprendizado de Máquina (AM). A escolha da *Scikit-learn* é justificada pela sua eficiência, flexibilidade e acessibilidade, tornando-a uma ferramenta essencial para a pesquisa e desenvolvimento em AM (PEDREGOSA et al., 2011). Esta biblioteca oferece suporte direto a vários dos modelos discutidos, incluindo LR, DT, RF, KNN, NB, e SVM, facilitando a implementação e experimentação com esses algoritmos.

Para o treinamento específico do modelo *XGBoost*, foi empregada a biblioteca *XGBoost*, também do *Python*, que se destaca por sua eficácia em competições de dados e aplicações práticas devido à sua velocidade e desempenho (CHEN; GUESTRIN, 2016). A *XGBoost* é altamente otimizada para árvores de decisão com boosting, oferecendo recursos avançados de regularização, tratamento de valores ausentes, e otimização de árvores, o que contribui para sua capacidade de gerar modelos preditivos altamente precisos e robustos conforme descrito por CHEN; GUESTRIN (2016).

Além disso, a estratégia de *StackingCVClassifier* foi implementada com o auxílio da biblioteca *mlxtend* do *Python* (*PYTHON SOFTWARE FOUNDATION*, 2023). A *mlxtend* (*Machine Learning Extensions*) é uma biblioteca que complementa as funcionalidades da Scikit-learn, fornecendo uma gama de ferramentas adicionais para a análise de dados, visualização e, especialmente, métodos de *ensemble* como o *StackingCVClassifier* (RASCHKA, 2018). Essa biblioteca facilita a combinação de múltiplos modelos de previsão, permitindo a aplicação de técnicas de validação cruzada para otimizar o desempenho do modelo final (NAIMI; BALZER, 2018).

A seleção dessas bibliotecas e da linguagem *Python* para o desenvolvimento dos modelos foi motivada pela robustez, documentação abrangente, e a ampla comunidade de suporte dessas ferramentas, que são fatores evidenciados por MÜLLER e GUIDO (2016). Além disso, a integração entre as bibliotecas sklearn, *XGBoost*, e mlxtend permite uma

implementação coesa e eficiente das técnicas de AM, desde a preparação dos dados até a avaliação dos resultados, assegurando que os modelos desenvolvidos são baseados nas práticas recomendadas e nos avanços mais recentes da área de AM.

Concluindo, esta subseção expôs a metodologia empregada na seleção, treinamento e avaliação dos modelos de aprendizado de máquina para a predição de DCV, delineando a abordagem sistemática adotada para o tratamento dos dados, a escolha dos algoritmos e a otimização dos modelos.

b. AVALIAÇÃO E OTIMIZAÇÃO DOS MODELOS

No desenvolvimento deste estudo, para avaliar o desempenho dos modelos de classificação, foram escolhidas as métricas: acurácia, precisão e pontuação F1. Selecionadas por sua capacidade de fornecer uma avaliação abrangente e detalhada do desempenho dos modelos sob diferentes perspectivas, conforme explicado por MÜLLER e GUIDO (2016) e corroborado por GÉRON (2019).

A acurácia é destacada por MÜLLER e GUIDO (2016) como a medida que indica a proporção de todas as previsões corretas feitas pelo modelo, enquanto GÉRON (2019) reforça sua importância como um indicador geral da eficácia do modelo, particularmente em contextos onde as classes alvo são distribuídas de maneira equilibrada. É calculada pela fórmula:

$$Acurácia = \frac{(verdadeiros\ positivos + verdadeiros\ negativos)}{(total\ de\ observações)} \quad (4)$$

Precisão, definida por MÜLLER e GUIDO (2016), é a fração de previsões positivas corretas em relação ao total de previsões positivas feitas pelo modelo:

$$Precisão = \frac{verdadeiros\ positivos}{(verdadeiros\ positivos + falsos\ positivos)} \quad (5)$$

GÉRON (2019) aponta a precisão como crucial em aplicações onde o custo dos falsos positivos é significativo, ressaltando a importância de uma seleção meticulosa em cenários que exigem alta especificidade.

Pontuação F1, como explicado por MÜLLER e GUIDO (2016), busca um equilíbrio entre precisão e sensibilidade (*recall*), sendo particularmente valiosa em situações com desequilíbrio entre as classes. A pontuação F1 é o harmônico da precisão e do *recall*, oferecendo um meio de avaliar a capacidade do modelo de identificar corretamente as instâncias positivas enquanto minimiza os falsos positivos:

$$Pontuação\ F1 = \frac{2 * (precisão * recall)}{(precisão + recall)} \quad (6)$$

$$Recall = \frac{verdadeiros\ positivos}{(verdadeiros\ positivos + falsos\ negativos)} \quad (7)$$

GÉRON (2019) enfatiza a F1 como uma métrica composta que ajusta precisão e *recall* em um único indicador, facilitando a compreensão do desempenho do modelo em contextos de desequilíbrio de classe. *Recall*, ou sensibilidade, mede a capacidade do modelo de identificar corretamente todos os positivos reais (GÉRON, 2019).

Essas métricas, ao serem aplicadas conjuntamente, permitem uma visão holística do desempenho dos modelos de classificação. Para assegurar uma avaliação confiável, os

modelos são ajustados e testados nos conjuntos de dados de treino com 80% e teste com 20% recomendados por MÜLLER e GUIDO (2016). Então é realizada uma avaliação desses algoritmos por meio da técnica de validação cruzada de 10 dobras, que tem como objetivo reduzir o viés de treinamento causado pelos dados (MÜLLER; GUIDO, 2016). Nessa técnica, o conjunto de dados é dividido em 10 partes iguais, chamadas de "dobras". O modelo é então treinado em 9 dobras e avaliado na décima. Esse processo é repetido 10 vezes, garantindo que cada dobra seja utilizada como conjunto de teste uma vez (MÜLLER; GUIDO, 2016). Isso assegura que o modelo seja capaz de generalizar bem para dados desconhecidos. Por fim, é calculada a média das avaliações das 10 iterações para fornecer uma estimativa mais precisa do desempenho do modelo (MÜLLER; GUIDO, 2016).

Com base nos resultados da avaliação é possível realizar a otimização, onde são refinados os hiperparâmetros dos modelos para maximizar sua performance (GÉRON, 2019). A importância das características é avaliada para destacar a contribuição relativa de cada variável no modelo. Essas etapas de otimização garantem que os modelos sejam ajustados de forma a obter os melhores resultados possíveis (NAIMI; BALZER, 2018).

4. RESULTADOS E DISCUSSÕES

Nesta seção, são apresentados os resultados alcançados, iniciando com a Subseção 4.1, Pré-processamento e Análise dos Dados, concentra-se na preparação inicial e exploração dos dados, utilizando técnicas estatísticas e de visualização para identificar padrões importantes. Na Subseção 4.2, Treinamento e Avaliação dos Modelos, são discutidos os processos de seleção, otimização e avaliação de desempenho dos modelos, tanto na base de dados completa quanto numa versão simplificada, adaptada para ambientes com recursos limitados. Finalmente, a Subseção 4.3, Desenvolvimento e Implementação da Aplicação, detalha o desenvolvimento de uma aplicação destinada a simplificar a interação dos usuários com o sistema e a apresentação das previsões de forma intuitiva.

a. PRÉ PROCESSAMENTO E ANÁLISE DOS DADOS

Nesta fase, foram aplicadas a síntese estatística e as técnicas de visualização para obter uma compreensão dos dados, identificando tendências. Inicialmente, cada um dos 14 atributos disponíveis foi investigado para identificar suas características e tipos de dados. Ao classificar atributos na análise de dados, foram identificados três tipos principais: contínuos, binários e categóricos.

Atributos contínuos são variáveis quantitativas que podem assumir qualquer valor dentro de um intervalo, representando medições que podem ser infinitamente divididas, como tempo ou distância (MÜLLER; GUIDO, 2016). Atributos binários, também conhecidos como dicotômicos, limitam-se a dois valores possíveis, tais como sim/não ou verdadeiro/falso, usados para representar a presença ou ausência de uma característica (GÉRON, 2019). Atributos categóricos são aqueles que dividem os dados em grupos ou categorias sem uma ordem numérica, como tipos sanguíneos ou categorias de produtos (GÉRON, 2019). Detalhes sobre a natureza de cada atributo e sua relevância para o estudo são sistematizados na Tabela 2.

Tabela 2 - Relação dos atributos usados na Base de Dados.

Atributo	Descrição	Tipo	Valores padrão
age	A idade do paciente em anos.	Contínuo	29–77

sex	O sexo do paciente. (1 se masculino e 0 se feminino)	Binário	1 e 0
cp	Representa o tipo de dor ou desconforto no peito relatado pelo paciente. Os valores podem ser: ● 1: Angina típica (dor no peito relacionada à redução do fluxo sanguíneo para o coração). ● 2: Angina atípica (outros tipos de dor no peito não relacionados à redução do fluxo sanguíneo para o coração). ● 3: Dor não anginal (dor no peito não relacionada à angina). ● 4: Assintomático (sem dor no peito).	Categórico	1, 2, 3 e 4
trestbps	A pressão arterial sistólica em repouso (medida em mm Hg) quando o paciente foi admitido.	Contínuo	94–200
chol	O nível de colesterol sérico medido em mg/dl.	Contínuo	126–564
fbs	Indica se o nível de açúcar no sangue em jejum é maior que 120 mg/dl (1 se verdadeiro, 0 se falso).	Binário	1 e 0
restecg	Mostra os resultados do eletrocardiograma em repouso. Os valores são: ● 0: Normal. ● 1: Anormalidades de ST-T. ● 2: Hipertrofia ventricular esquerda provável.	Categórico	0, 1 e 2
thalach	A frequência cardíaca máxima alcançada durante o teste de esforço.	Contínuo	71–202
exang	Indica se a angina foi induzida pelo exercício. Pode ser 1 para sim ou 0 para não.	Binário	0 e 1
oldpeak	A depressão do segmento ST induzida pelo exercício em relação ao repouso.	Contínuo	0.0–6.2
slope	A inclinação do segmento ST no pico do exercício. Os valores são: ● 1: Ascendente. ● 2: Plano. ● 3: Descida.	Categórico	0, 1 e 2
ca	O número de principais vasos coloridos por fluoroscopia (0-3).	Contínuo	0, 1, 2 e 3
thal	Talassemia. Os valores são: ● 3: Normal. ● 6: Defeito fixo. ● 7: Defeito reversível.	Categórico	3, 6 e 7
num	O diagnóstico de doença cardíaca. Pode ser 0 (sem doença) ou 1 (com doença).	Categórico	0 e 1

Fonte: elaborada pelos autores com adaptações de UCI MACHINE LEARNING REPOSITORY (2013)

Complementando a classificação apresentada, a Tabela 3 aborda a categorização sobre a dificuldade de extração dos atributos de acordo com uma escala de dificuldade de obtenção. Esta escala é quantificada por pontuações que refletem a complexidade envolvida na coleta de cada tipo de dado. Cada pontuação é atribuída com base em cinco critérios: a necessidade de equipamento especializado, a existência de custos financeiros associados, a disponibilidade do procedimento tanto em casa quanto no município de Corrente PI, e a necessidade de preparação especial por parte do paciente.

Tabela 3 - Relação da dificuldade por extração de cada atributo.

Atributo	Modalidade de extração	Não disponível em casa	Não disponível na cidade (Corrente)	Pode ter custo financeiro	Necessita equipamento especializado	Requer preparação especial (Ex: Não fumar 2 horas antes do exame)	Pontuação total
age	Documentação pessoal	0	0	0	0	0	**0**
sex	Documentação pessoal	0	0	0	0	0	**0**
cp	Sensação do paciente	0	0	0	0	0	**0**
trestbps	Esfigmomanômetro	1	0	0	1	0	**2**
fbs	Hemograma	1	0	1	1	0	**3**
chol	Hemograma	1	0	1	1	0	**3**
thal	Hemograma	1	0	1	1	0	**3**
restecg	Eletrocardiograma simples	1	0	1	1	0	**3**
thalach	Teste ergométrico	1	0	1	1	1	**4**
exang	Teste ergométrico	1	0	1	1	1	**4**
oldpeak	Teste ergométrico	1	0	1	1	1	**4**
slope	Teste ergométrico	1	0	1	1	1	**4**
ca	Exame Fluoroscopia ou Radioscopia	1	1	1	1	1	**5**

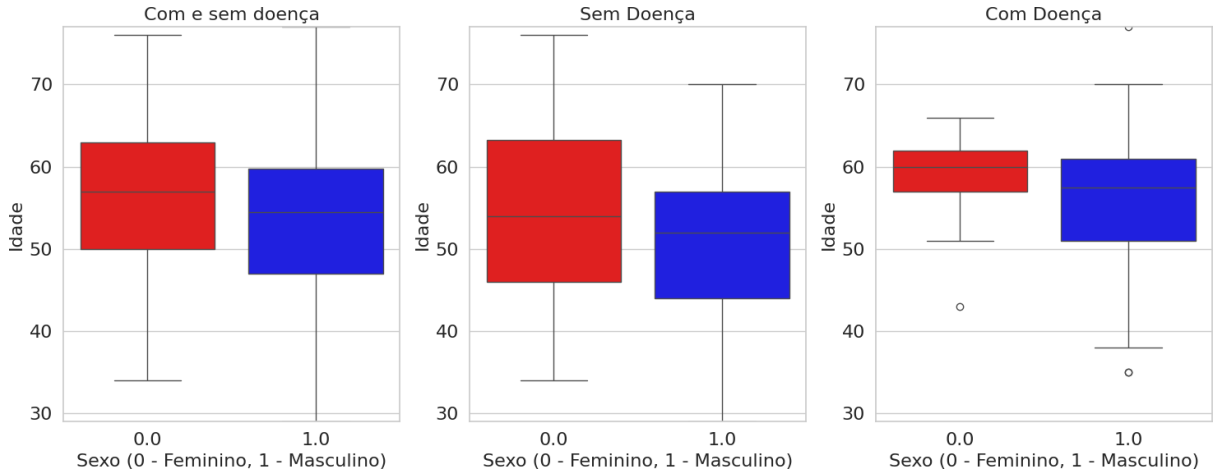
Fonte: elaborada pelos autores.

Analisando os dados fornecidos é possível identificar os atributos com maior dificuldade de extração, o que não só destaca as barreiras práticas enfrentadas por profissionais de saúde em ambientes menos equipados, mas também sinaliza a necessidade de desenvolver abordagens alternativas que possam compensar tais limitações sem comprometer a integridade e a eficácia do diagnóstico. Partindo deste princípio, posteriormente a Tabela 3 servirá de base para simplificar o conjunto de dados utilizado no treinamento dos modelos, com o objetivo de obter resultados adequados com o usuário final fornecendo menos informações.

Com a definição e estruturação do dados, foi iniciado o processo de EDA, que resultou na identificação de padrões na distribuição etária entre os conjuntos de pacientes, categorizados por sexo e presença de doença cardíaca. Três gráficos foram gerados para

ilustrar a prevalência da doença cardíaca em indivíduos do sexo feminino e masculino. Foi evidenciado que, entre as mulheres com a doença, a concentração etária predominante situa-se entre o final dos 50 anos e o início dos 60 anos. Em contraste, a distribuição etária entre os homens com a doença é mais ampla, indicando uma faixa etária mais extensa para a prevalência da doença, conforme ilustrado na Figura 2.

Figura 2 - Distribuição de idades entre os grupos de pacientes com e sem doença cardíaca, separados por sexo.

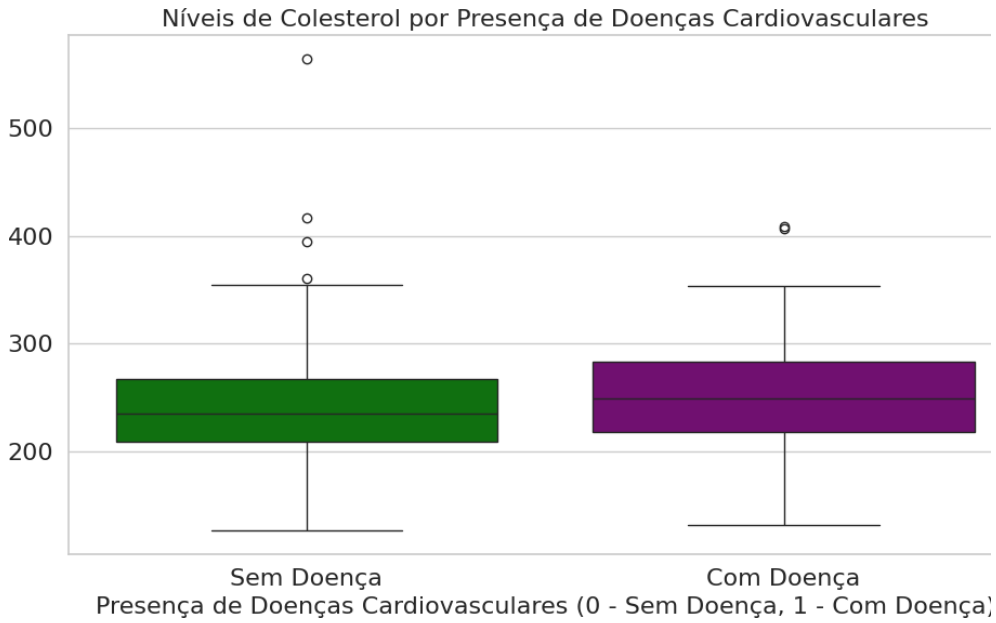


Este fenômeno é corroborado por estudos recentes, como o de REGITZ-ZAGROSEK; GEBHARD (2022), que mostra uma maior prevalência da doença cardíaca em homens. Este fato é atribuído ao efeito protetor dos hormônios sexuais femininos, especialmente o estrogênio, que confere uma certa resiliência cardiovascular às mulheres REGITZ-ZAGROSEK; GEBHARD (2022). No entanto, após a menopausa, as mulheres podem experimentar um aumento no risco de doença cardíaca, uma vez que a proteção hormonal diminui (REGITZ-ZAGROSEK; GEBHARD, 2022).

Além dos aspectos biológicos, REGITZ-ZAGROSEK; GEBHARD (2022) também destacam a influência de fatores comportamentais e sociais na prevalência da doença cardíaca, além dos aspectos biológicos, evidenciando que os homens geralmente apresentam um aumento do risco em idades mais jovens. Isso pode ser atribuído a uma variedade de fatores, incluindo estilos de vida menos saudáveis, como uma maior prevalência de tabagismo e consumo de álcool, bem como maior exposição ao estresse ocupacional e psicossocial (REGITZ-ZAGROSEK; GEBHARD, 2022).

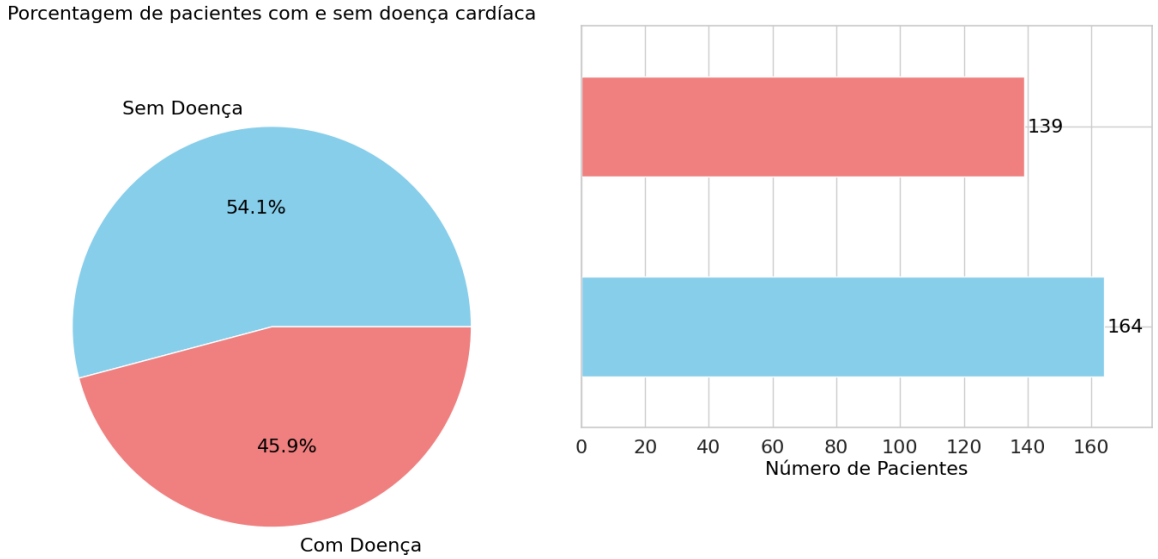
Em sequência, a Figura 3 ilustra a relação entre os níveis de colesterol e a presença de doença cardíaca. De acordo com JUNG et al. (2022), maiores níveis de colesterol estão associados a uma maior prevalência de DCV. O colesterol é um tipo de gordura que é fundamental para o bom funcionamento do organismo, mas ter níveis elevados de colesterol no sangue pode aumentar o risco de problemas cardiovasculares, como infarto ou Acidente Vascular Cerebral (AVC) (JUNG et al., 2022).

Figura 3 - Níveis de Colesterol por Presença de Doença.



Após validar as correlações com a medicina, foi avaliado o equilíbrio dos dados com uma análise quantitativa da distribuição de registros para cada classe de previsão, conforme ilustrado na Figura 4. Os resultados indicaram que a base de dados está adequadamente balanceada para prosseguir com a pesquisa. Especificamente, a classe 0 (sem doença) representa 54% dos registros, totalizando 164 registros, enquanto a classe 1 (com doença) representa 46%, totalizando 139 registros. Segundo GÉRON (2019), essa distribuição é considerada adequada para uma classificação binária.

Figura 4 - Gráficos do balanceamento da base de dados.



Embora as classes da variável de predição estejam relativamente equilibradas, como mostrado na Figura 4, identificou-se a necessidade de implementar a técnica de SMOTE

antes do treinamento dos modelos. Esta técnica, proposta por CHAWLA et al. (2002), é uma abordagem de reamostragem que adiciona novos exemplos sintéticos à classe minoritária, ajudando a equilibrar a distribuição das classes. A implementação do SMOTE é particularmente importante neste contexto para evitar possíveis vieses introduzidos pela desproporcionalidade entre as classes de sexo. Isso ocorre porque, se uma classe estiver sub-representada nos dados de treinamento, o modelo pode ser tendencioso em favor da classe majoritária, resultando em um desempenho de previsão pobre para a classe minoritária.

b. TREINAMENTO E AVALIAÇÃO DOS MODELOS

Dos 10 modelos considerados, cinco terão variações de hiperparâmetros, sendo eles *XGBoost*, KNN, RF, ET e SVM. Esta seleção permite cobrir desde métodos simples até técnicas que lidam com padrões complexos. Segundo RASCHKA e MIRJALILI (2019), usar uma variedade de modelos, com alterações em variáveis internas, é essencial para adaptar as técnicas de AM com os aspectos dos dados analisados.

Para o treinamento dos modelos selecionados, foi executada a metodologia descrita na Seção 3, dividindo o conjunto de dados em 80% para treinamento e 20% para teste. Esta proporção, recomendada por MÜLLER e GUIDO (2016), é utilizada para garantir uma distribuição eficaz que permita aos modelos aprenderem de forma abrangente, ao mesmo tempo em que reservam uma quantidade suficiente de dados não vistos para uma avaliação válida do desempenho. Ademais, foi aplicada a validação cruzada de 10 dobras, conforme sugerido por GÉRON (2019), para obter uma medida mais precisa e confiável do desempenho geral dos modelos.

A otimização destes modelos, seguindo as diretrizes de NAIMI e BALZER (2018), envolveu a análise das métricas de desempenho: acurácia, precisão e pontuação F1, permitindo a identificação das configurações mais adaptadas para cada modelo. Ao escolher 1000 estimadores para o modelo *XGBoost*, a decisão foi baseada na observação de que um maior número de estimadores pode melhorar a capacidade do modelo de aprender padrões complexos nos dados, conforme destacado por GÉRON (2019). No entanto, é crucial equilibrar esse benefício com o risco de sobreajuste, o que pode comprometer a capacidade do modelo de generalizar para novos dados. GÉRON (2019) também aponta que um número excessivo de estimadores pode levar a um aumento desnecessário no tempo de treinamento sem melhorias significativas na performance do modelo.

No caso do KNN, a escolha do número de vizinhos impacta diretamente no desempenho do modelo. MÜLLER e GUIDO (2016) explicam que um número pequeno de vizinhos pode fazer o modelo ser demasiadamente sensível ao ruído dos dados de treinamento, enquanto um número maior pode ajudar a suavizar a decisão do modelo, tornando-o mais robusto ao ruído, mas possivelmente menos preciso em relação às fronteiras de decisão reais dos dados. A decisão de usar 5, 7, 9 e 10 vizinhos no KNN busca um compromisso entre a sensibilidade ao ruído e a capacidade de captura de padrões sutis, uma escolha fundamentada na busca por um modelo equilibrado que possa generalizar bem em dados não vistos, conforme aconselhado por MÜLLER e GUIDO (2016).

Para os modelos baseados em árvores, como RF e ET, o critério de divisão (Gini ou Entropia) e o número de estimadores. GÉRON (2019) destaca que a escolha entre Gini e Entropia influencia diretamente a qualidade das divisões das árvores, uma vez que Gini trabalha para alcançar divisões mais equilibradas, enquanto a Entropia visa produzir subconjuntos mais homogêneos, isto é, grupos com características similares. O número de estimadores, por sua vez, controla a quantidade de árvores na floresta, onde um número

maior pode melhorar a acurácia do modelo até certo ponto, mas com retornos decrescentes em termos de performance versus custo computacional, conforme discutido por GÉRON (2019).

Por fim, quando se trata da seleção de *kernels* no SVM, a decisão entre o *kernel* linear e o RBF reflete a necessidade de adaptar o modelo à natureza dos dados. GÉRON (2019) salienta que o *kernel* linear é eficaz para dados que são linearmente separáveis, oferecendo uma solução simples e computacionalmente eficiente. Por outro lado, o *kernel* RBF é capaz de lidar com casos em que a separação das classes não é linearmente possível no espaço original, transformando os dados em um espaço de maior dimensão onde essa separação se torna viável, conforme GÉRON (2019). A escolha do *kernel* RBF, portanto, é motivada pela sua flexibilidade e capacidade de capturar relações complexas entre as características, uma estratégia recomendada por MÜLLER e GUIDO (2016) quando enfrentamos conjuntos de dados com padrões intrincados e não linearidades. A análise dos resultados, conforme apresentados na Tabela 4, mostra as informações sobre o desempenho dos modelos treinados.

Tabela 4 - Resultados obtidos no treinamento dos modelos.

Modelo	**Acurácia**	**Precisão**	**Pontuação F1**
StackingCVClassifier (XGBoost + KNN + SVM)	92,64%	93,02%	91,18%
XGBoost com 1000 Estimadores	90,53%	86,76%	89,65%
SVM com *kernel* Linear	90,45%	86,95%	89,87%
XGBoost com 100 Estimadores	90,23%	86,64%	89,61%
KNN com 5 Vizinhos	89,01%	88,29%	87,77%
KNN com 7 Vizinhos	88,38%	86,20%	87,71%
ET com 100 Estimadores	88,32%	83,87%	88,13%
RF com Entropia e 100 Estimadores	86,64%	85,71%	85,54%
SVM com kernel RBF (*Radial Basis Function*)	85,34%	85,23%	85,12%

ET com 1000 Estimadores	86,65%	83,33%	86,20%
AdaBoost Padrão	86,69%	83,33%	86,20%
RF com Gini e 100 Estimadores	86,61%	85,71%	85,71%
KNN com 11 Vizinhos	86,67%	85,71%	85,71%
KNN com 9 Vizinhos	86,62%	85,71%	85,71%
NB Padrão	85,97%	84,31%	85,20%
LR Padrão	80,02%	80,18%	79,63%
DT Padrão	71,66%	65,71%	73,01%

O *StackingCVClassifier* emerge como o modelo de maior desempenho, com uma acurácia de 92,64%, precisão de 93,02% e uma pontuação F1 de 91,18%. Esses resultados destacam a eficácia da técnica de *ensemble* em combinar as previsões de múltiplos modelos para melhorar a precisão e a robustez das previsões finais. A alta precisão indica uma forte capacidade do modelo em identificar verdadeiros positivos, enquanto a pontuação F1, uma média harmônica entre precisão e *recall*, sugere um equilíbrio entre a precisão e a capacidade do modelo em detectar todas as instâncias relevantes.

Os modelos *XGBoost*, com diferentes configurações de estimadores, também mostraram resultados promissores, particularmente o *XGBoost* com 1000 estimadores, alcançando uma acurácia de 90,53% e uma pontuação F1 de 89,65%. Isso indica a eficiência do *XGBoost* em lidar com base de dados complexas e sua capacidade de minimizar o *overfitting* através do ajuste de hiperparâmetros.

Entre os modelos de KNN, observa-se uma consistência de desempenho, com KNN com 7 vizinhos e o KNN com 5 vizinhos obtendo acurácias próximas sendo o mais eficaz com 89,01% e pontuações F1 similares. Isso reflete a importância da escolha do número de vizinhos no KNN, onde um equilíbrio adequado pode contribuir para a estabilidade do modelo.

Modelo como o ET também demonstrou competência, embora com nuances em suas métricas específicas. Por exemplo, o ET com 100 Estimadores, apesar de ter uma acurácia comparável a outros modelos (88,32%), possui uma precisão ligeiramente inferior, o que pode influenciar na sua aplicabilidade dependendo do contexto clínico.

Notavelmente, o modelo DT exibe o desempenho mais baixo entre os avaliados, com uma acurácia de apenas 71,66% e uma pontuação F1 de 73,01%. Isso pode ser atribuído à natureza relativamente simples das DT, que pode não capturar a complexidade dos dados relacionados a DCV tão efetivamente quanto técnicas mais sofisticadas.

A Tabela 5 apresenta uma comparação direta das acurácias alcançadas por diferentes trabalhos da literatura sobre diagnóstico de DCV, utilizando a base de dados *UCI Heart Diseases*. O modelo HRFLM, proposto por MOHAN; THIRUMALAI; SRIVASTAVA (2019), obteve uma acurácia de 88,7%. Em seguida, GUPTA et al. (2020) avançaram com o modelo MIFH, alcançando 92,34% de acurácia. TIWARI; CHUGH; SHARMA (2022) introduziram o modelo CEE, que melhorou ainda mais a precisão para 92,44%. O presente trabalho destaca-se ao implementar o *StackingCVClassifier*, superando os estudos anteriores com uma acurácia de 92,64%, demonstrando um incremento contínuo na precisão dos modelos de diagnóstico de doenças cardíacas ao longo do tempo.

Tabela 5 - Tabela de comparação entre a acurácia dos modelos para cada trabalho.

Trabalho	**Modelo**	**Base de Dados**	**Acurácia**
(MOHAN; THIRUMALAI; SRIVASTAVA, 2019)	HRFLM	*UCI Heart Diseases*	88,7%
(GUPTA et al., 2020)	MIFH	*UCI Heart Diseases*	92,34%
(TIWARI; CHUGH; SHARMA, 2022)	CEE	*UCI Heart Diseases*	92,44%
Este trabalho	*StackingCVClassifier*	*UCI Heart Diseases*	92.64%

Na fase de refinamento do estudo, foi tomada uma decisão estratégica para otimizar a análise de predição de doenças cardíacas: a remoção das colunas "thalach", "exang", "oldpeak", "slope" e "ca" da base de dados. Estas colunas foram identificadas com pontuação de dificuldade maior que 3, indicando um maior desafio na sua coleta, seja por necessidade de equipamento especializado, preparação especial do paciente, custos financeiros associados, ou limitações de disponibilidade em determinadas no município de Corrente-PI. A exclusão desses atributos visa simplificar o conjunto de dados, concentrando-se em variáveis mais acessíveis. Esta abordagem visa não apenas facilitar a coleta de dados, mas também potencializar a aplicabilidade prática dos modelos preditivos em ambientes com recursos mais limitados.

A remoção dos atributos com maior dificuldade de coleta reflete uma abordagem pragmática para a construção de modelos de AM, equilibrando a necessidade de precisão preditiva com a viabilidade de implementação. Esse ajuste nos dados levou a uma ligeira redução nas métricas de desempenho dos modelos, o que era esperado devido à diminuição da quantidade de informações disponíveis para o treinamento. No entanto, os resultados continuam demonstrando alta eficácia, reforçando a capacidade dos modelos de manter um

bom desempenho mesmo com um conjunto de dados mais restrito, focando em variáveis de mais fácil obtenção e relevância clínica direta. A Tabela 6 apresenta os resultados após a modificação descrita acima.

Tabela 6 - Modelos usando a base de dados sem os atributos com dificuldade de extração maior que 3.

Modelo	**Acurácia**	**Precisão**	**Pontuação F1**
StackingCVClassifier (XGBoost + KNN + SVM)	88,34%	89,86%	88,44%
XGBoost com 1000 Estimadores	87,81%	84,16%	86,96%
XGBoost com 100 Estimadores	87,52%	84,04%	86,92%
SVM com *kernel* Linear	86,14%	82,98%	84,12%
KNN com 7 Vizinhos	85,73%	83,61%	85,08%
KNN com 5 Vizinhos	85,73%	83,61%	85,08%
ET com 100 Estimadores	85,68%	83,35%	85,49%
RF com Entropia e 100 Estimadores	84,04%	83,14%	83,14%
ET com 1000 Estimadores	84,03%	80,83%	83,61%
RF com Gini e 100 Estimadores	84,01%	83,14%	83,14%
SVM com kernel RBF (*Radial Basis Function*)	83,92%	82,63%	83,19%
AdaBoost Padrão	83,70%	81,83%	83,61%

KNN com 11 Vizinhos	83,35%	83,14%	83,44%
KNN com 9 Vizinhos	83,23%	83,45%	82,19%
NB Padrão	82,89%	81,51%	82,06%
LR Padrão	75,45%	76,62%	75,12%
DT Padrão	69,51%	63,74%	70,82%

c. DESENVOLVIMENTO E IMPLEMENTAÇÃO DA APLICAÇÃO

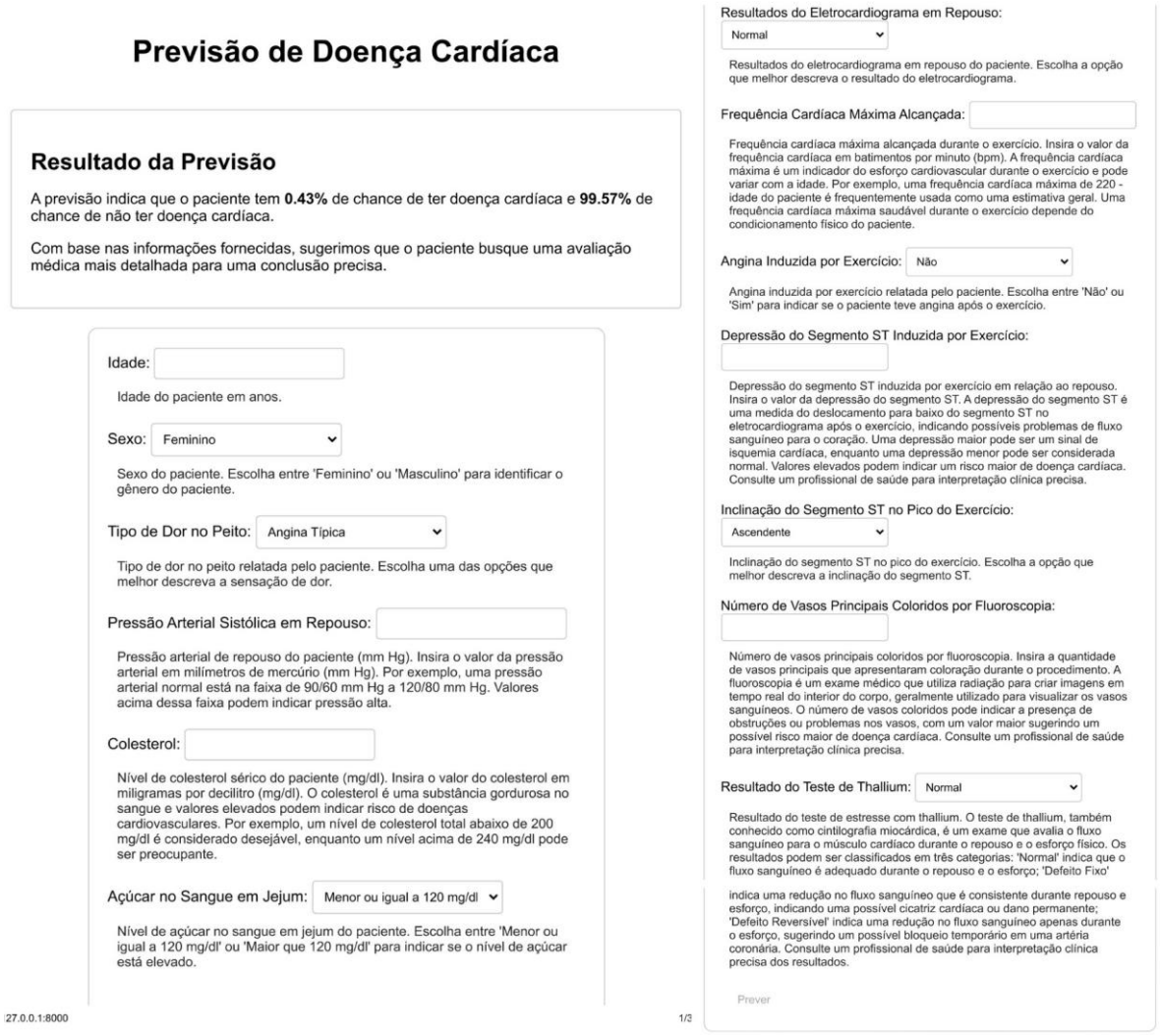
Na fase de desenvolvimento, a construção da aplicação foi planejada e executada seguindo um protocolo técnico. Inicialmente, a API foi desenvolvida utilizando a linguagem de programação *Python*, reconhecida pela sua versatilidade e eficiência. Para este propósito, foram adotados os *frameworks Django* (*DJANGO SOFTWARE FOUNDATION*, 2023) e *Django REST* (CHRISTIE, 2023), escolhidos por sua robustez, segurança, e capacidade de facilitar a criação de *APIs RESTful* escaláveis.

Django, um framework de alto nível para Python, foi utilizado para estruturar a lógica de *back-end* da aplicação, permitindo uma gestão eficiente do banco de dados, autenticação de usuários, e outras funcionalidades essenciais para a operação segura e eficaz do sistema. O *Django REST Framework* complementa o *Django* ao oferecer ferramentas poderosas para a construção de *APIs web*, facilitando a serialização de dados e a comunicação entre a aplicação cliente e o servidor. Este componente atua como uma ponte essencial para a comunicação entre a interface de usuário e os modelos de AM treinados, permitindo a manipulação e o acesso aos dados de forma eficiente e segura.

Para a construção da interface do usuário, foi adotado um desenvolvimento front-end utilizando *HTML* e *CSS*, conforme recomendado pelos *MDN WEB DOCS* (2023), integrando-se harmoniosamente ao ambiente *Django*. Esta escolha permitiu o aproveitamento das funcionalidades do *Django* para gerar formulários dinâmicos, onde a interação do usuário com a aplicação ocorre. O uso de *HTML* e *CSS* foi fundamental para criar uma interface limpa, intuitiva e responsiva, garantindo uma experiência de usuário agradável e acessível.

A integração entre o back-end, construído com *Django* e *Django REST Framework*, e o front-end, desenvolvido com *HTML* e *CSS*, foi realizada de maneira coesa, garantindo que a aplicação funcionasse como um sistema unificado. Essa integração permitiu não apenas a entrada estruturada de dados por meio dos formulários do *Django*, mas também a apresentação dos resultados dos modelos de AM de maneira clara e interativa, promovendo uma interação eficaz e engajadora com o usuário. A Figura 5 mostra a interface desenvolvida.

Figura 5 - Interface da aplicação para o usuário final.



A escolha dessas tecnologias foi baseada em suas capacidades de atender às necessidades específicas do projeto, como escalabilidade, segurança, e facilidade de manutenção (*DJANGO SOFTWARE FOUNDATION*, 2023). A arquitetura da aplicação foi projetada para ser modular, facilitando futuras atualizações e a adição de novas funcionalidades, demonstrando um compromisso com práticas de desenvolvimento sustentável e eficiente recomendadas por MARTIN (2017).

A interação do usuário ocorre por meio de um formulário gerado pelo *Django*, proporcionando uma entrada estruturada de dados. Ao submeter as informações, elas são enviadas para a API, onde passam por uma série de procedimentos. Primeiro, as informações são associadas à base de dados e convertidas para o formato adequado, utilizando técnicas como *one-hot encoding* e normalização. Essas etapas garantem a consistência e uniformidade dos dados. Em seguida, os dados processados são direcionados para o modelo mais eficaz e treinado. O modelo realiza o processamento necessário e retorna um *JavaScript Object Notation* (JSON) para a aplicação que fez a requisição, contendo os resultados da previsão.

O resultado fornecido pela aplicação ao usuário é expresso em percentuais, refletindo a probabilidade estimada de presença ou ausência de DCV. Por exemplo, um percentual de 90% sugere que, de acordo com o modelo, existe uma probabilidade de 90% de presença da

doença, enquanto os restantes 10% apontam para a probabilidade de ausência. É importante enfatizar que estes resultados são indicativos e destinam-se a apoiar a avaliação clínica, não substituindo de forma alguma o diagnóstico médico profissional. A interpretação e o diagnóstico finais devem sempre ser realizados por um médico qualificado, que levará em consideração uma gama mais ampla de fatores clínicos e pessoais.

5. CONSIDERAÇÕES FINAIS

Este artigo apresentou um estudo detalhado sobre a previsão de DCV por meio de técnicas de AM que refletem sobre os avanços significativos alcançados no diagnóstico de DCV através da aplicação de modelos de AM. A análise detalhada da base de dados da UCI, juntamente com a implementação e avaliação de diversos modelos de AM, destacou a potencialidade desta abordagem para melhorar a precisão do diagnóstico precoce de DCV. A metodologia adotada neste estudo demonstrou como a combinação de dados clínicos estruturados e técnicas avançadas de AM pode fornecer resultados para melhorar a tomada de decisões na prática clínica.

A classificação dos atributos de acordo com a dificuldade de sua coleta e a subsequente análise de sua importância para a predição de DCV oferecem uma perspectiva prática que pode orientar a coleta de dados em ambientes clínicos. Isso não apenas facilita a identificação de pacientes em risco, mas também contribui para uma alocação mais eficiente de recursos diagnósticos.

O desenvolvimento e a implementação de uma API e uma interface de usuário, demonstram a viabilidade de integrar essas ferramentas tecnológicas no ambiente clínico. Esta aplicação não só melhora a interação do usuário, mas também serve como uma ferramenta para profissionais de saúde, permitindo a rápida avaliação de riscos de DCV em pacientes.

Embora os resultados deste estudo sejam promissores, é importante reconhecer as limitações e os desafios inerentes ao uso de AM em contextos de saúde. A dependência de dados de alta qualidade junto a necessidade de interpretações claras sobre as informações utilizadas são aspectos que devem ser continuamente abordados. Além disso, a adoção de tais tecnologias no ambiente clínico exige uma abordagem multidisciplinar, envolvendo profissionais de saúde, cientistas de dados e desenvolvedores de software para garantir a eficácia e a ética na aplicação dessas ferramentas.

Para futuras pesquisas, pretende-se realizar a exploração de novas técnicas de AM, a inclusão de conjuntos de dados mais amplos e diversificados, e a realização de estudos longitudinais para avaliar a eficácia dessas abordagens no diagnóstico de DCV ao longo do tempo. Além disso, é crucial desenvolver estratégias para a implementação efetiva dessas tecnologias no sistema de saúde, abordando questões de aceitação por parte dos profissionais de saúde, integração com sistemas de informação em saúde existentes e conformidade com regulamentações de privacidade e proteção de dados.

A aplicação da AM no diagnóstico de DCV representa um avanço no campo da medicina e informática em saúde, com o potencial de melhorar a qualidade dos cuidados ao paciente, otimizar o uso de recursos e contribuir para a prevenção e tratamento de uma das principais causas de morbidade e mortalidade globalmente. A pesquisa e desenvolvimento contínuos nessa área são essenciais para superar os desafios atuais e maximizar o impacto positivo da tecnologia na saúde pública e na prática clínica.

REFERÊNCIAS

AHA, D. W.; KIBLER, D.; ALBERT, M. K. InstanceBased Learning Algorithms. **Machine Learning**, vol. 6, no. 1, p. 37–66, 1991. DOI https://doi.org/10.1023/A:1022689900470.

BREIMAN, L. Random Forests. **Machine Learning**, vol. 45, no. 1, p. 5–32, 1 Jan. 2001. DOI https://doi.org/10.1023/a:1010933404324.

BURKOV, A. **THE HUNDRED-PAGE MACHINE LEARNING BOOK**. Quebec City, QC, Canada: Andriy Burkov, 2019.

CHANDRA, B.; P. PAUL VARGHESE. Moving towards efficient decision tree construction. **Information Sciences**, vol. 179, no. 8, p. 1059–1069, 29 Mar. 2009. DOI https://doi.org/10.1016/j.ins.2008.12.006.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; W. PHILIP KEGELMEYER. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, vol. 16, p. 321–357, 1 Jun. 2002. DOI https://doi.org/10.1613/jair.953.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. 2016. San Francisco, California, USA: Association for Computing Machinery, 2016. p. 785–794. DOI https://doi.org/10.1145/2939672.2939785.

CHRISTIE, T. Home - Django REST framework. 2023. **Django-rest-framework.org**. Disponível em: https://www.django-rest-framework.org/. Acesso em: 10 Nov. 2023.

CLEOPHAS, T. J.; ZWINDERMAN, A. H. Machine Learning in Medicine - a Complete Overview. **SpringerLink**, 2015. DOI https://doi.org/10.1007-978-3-319-15195-3.

DJANGO SOFTWARE FOUNDATION. Django Documentation. 2023. **Django Project**. Disponível em: https://docs.djangoproject.com/en/5.0/. Acesso em: 10 Nov. 2023.

GÉRON, A. **Hands-on machine learning with Scikit-Learn and TensorFlow concepts, tools, and techniques to build intelligent systems**. 2nd ed. [*S. l.*]: O'Reilly Media, Inc., 2019.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine learning**, vol. 63, p. 3–42, 2006.

GUPTA, A.; KUMAR, R.; HARKIRAT SINGH ARORA; RAMAN, B. MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis. **IEEE Access**, vol. 8, p. 14659–14674, 1 Jan. 2020. DOI https://doi.org/10.1109/access.2019.2962755.

HAMADA; MAHMOUD AHMED REFAEE; ATIQ UR REHMAN; MOHAMMAD TARIQUL ISLAM; SAMIR BRAHIM BELHAOUARI; ALAM, T. Risk Factors and Comorbidities Associated to Cardiovascular Disease in Qatar: A Machine Learning Based Case-Control Study. **IEEE Access**, vol. 9, p. 29929–29941, 1 Jan. 2021. DOI https://doi.org/10.1109/access.2021.3059469.

HART, P. E.; STORK, D. G.; DUDA, R. O. **Pattern classification**. [*S. l.*]: Wiley Hoboken, 2000.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning. **SpringerLink**, 2009. DOI https://doi.org/10.1007-978-0-387-84858-7.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. Introduction to the Logistic Regression Model. **Wiley series in probability and statistics**, p. 1–33, 22 Mar. 2013. DOI https://doi.org/10.1002/9781118548387.ch1.

JUNG, E.; SO YEON KONG; YOUNG SUN RO; HYUN HO RYU; SANG DO SHIN. Serum Cholesterol Levels and Risk of Cardiovascular Death: A Systematic Review and a Dose-Response Meta-Analysis of Prospective Cohort Studies. **International Journal of Environmental Research and Public Health**, vol. 19, no. 14, p. 8272–8272, 6 Jul. 2022. DOI https://doi.org/10.3390/ijerph19148272.

LOWD, D.; DOMINGOS, P. Naive Bayes models for probability estimation. 2005. Bonn, Germany: Association for Computing Machinery, 2005. p. 529–536. DOI https://doi.org/10.1145/1102351.1102418.

MAMMONE, A.; TURCHI, M.; CRISTIANINI, N. Support vector machines. **WIREs Computational Statistics**, vol. 1, no. 3, p. 283–289, 1 Nov. 2009. DOI https://doi.org/10.1002/wics.49.

MARGINEANTU, DRAGOS D; DIETTERICH, T. G. Pruning adaptive boosting. 97., 1997. **Citeseer** [...]. [*S. l.*: *s. n.*], 1997. vol. 97, p. 211–218.

MARTIN, R. C. **Clean architecture**. [*S. l.*]: Prentice Hall, 2017.

MATPLOTLIB. 2023. **Matplotlib.org**. Disponível em: https://matplotlib.org/. Acesso em: 3 Dec. 2023.

MDN WEB DOCS. 2023. **MDN Web Docs**. Disponível em: https://developer.mozilla.org/pt-BR/. Acesso em: 10 Dec. 2023.

MOHAN, S.; THIRUMALAI, C.; SRIVASTAVA, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. **IEEE Access**, vol. 7, p. 81542–81554, 2019. https://doi.org/10.1109/access.2019.2923707.

MOURA, C.; LÍGIA GIOVANELLA; ADAUTO EMMERICH OLIVEIRA; NETO, S. Tempo de espera e absenteísmo na atenção especializada: um desafio para os sistemas universais de saúde. **Saúde em Debate**, vol. 43, no. spe5, p. 190–204, 1 Jan. 2019. DOI https://doi.org/10.1590/0103-11042019s516.

MÜLLER, A.; GUIDO, S. **Introduction To Machine Learning With Python**. [*S. l.*]: Oreilly & Associates Inc, 2016.

MÜMTAZ KARATAŞ; LEVENT ERIŞKIN; MUHAMMET DEVECI; DRAGAN PAMUČAR; GARG, H. Big Data for Healthcare Industry 4.0: Applications, challenges and future perspectives. **Expert Systems with Applications**, vol. 200, p. 116912–116912, 1 Aug. 2022. DOI https://doi.org/10.1016/j.eswa.2022.116912.

NAIMI, A. I.; BALZER, L. B. Stacked generalization: an introduction to super learning. **European Journal of Epidemiology**, vol. 33, no. 5, p. 459–464, 10 Apr. 2018. DOI https://doi.org/10.1007/s10654-018-0390-z.

NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial. **Frontiers in neurorobotics**, vol. 7, p. 21, 2013.

NILS STRODTHOFF; WAGNER, P.; SCHAEFFTER, T.; SAMEK, W. Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL. **IEEE Journal of Biomedical and Health Informatics**, vol. 25, no. 5, p. 1519–1528, 1 May 2021. DOI https://doi.org/10.1109/jbhi.2020.3022989.

NORMAN, D. A. **Emotional design : why we love (or hate) everyday things**. New York: Basic Books, 2004.

PANDAS - PYTHON DATA ANALYSIS LIBRARY. 2023. **Pydata.org**. Disponível em: https://pandas.pydata.org/. Acesso em: 3 Dec. 2023.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, vol. 12, p. 2825–2830, 2011.

PYTHON SOFTWARE FOUNDATION. Python's documentation. 2023. **Python.org**. Disponível em: https://www.python.org/doc/. Acesso em: 3 Dec. 2023.

RASCHKA, S.; MIRJALILI, V. **Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2**. [*S. l.*]: Packt Publishing Ltd, 2019.

REGITZ-ZAGROSEK, V.; GEBHARD, C. Gender medicine: effects of sex and gender on cardiovascular disease manifestation and outcomes. **Nature Reviews Cardiology**, vol. 20, no. 4, p. 236–247, 31 Oct. 2022. DOI https://doi.org/10.1038/s41569-022-00797-4.

SCIKIT-LEARN: MACHINE LEARNING IN PYTHON — SCIKIT-LEARN 1.3.2 DOCUMENTATION. 2023. **Scikit-learn.org**. Disponível em: https://scikit-learn.org/stable/. Acesso em: 30 Nov. 2023.

SEABORN: STATISTICAL DATA VISUALIZATION - 0.13.0. 2023. **Pydata.org**. Disponível em: https://seaborn.pydata.org/. Acesso em: 3 Dec. 2023.

SMYTH, P.; WOLPERT, D. Linearly combining density estimators via stacking. **Machine Learning**, vol. 36, p. 59–83, 1999.

TIWARI, A.; CHUGH, A.; SHARMA, A. Ensemble framework for cardiovascular disease prediction. **Computers in Biology and Medicine**, vol. 146, p. 105624–105624, 1 Jul. 2022. DOI https://doi.org/10.1016/j.compbiomed.2022.105624.

UCI MACHINE LEARNING REPOSITORY. 2013. **Uci.edu**. Disponível em: https://archive.ics.uci.edu/. Acesso em: 3 Dec. 2023.

WAGNER, P.; NILS STRODTHOFF; RALF-DIETER BOUSSELJOT; SAMEK, W.; SCHAEFFTER, T. PTB-XL, a large publicly available electrocardiography dataset. 24 Apr. 2020. **Physionet.org**. Disponível em: https://physionet.org/content/ptb-xl/1.0.1/. Acesso em: 8 Nov. 2023.

WORLD HEALTH ORGANIZATION. Cardiovascular diseases. 11 Jun. 2021. **World Health Organization**. Disponível em: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Acesso em: 21 Nov. 2023.

WU, X.; KUMAR, V. **The Top Ten Algorithms in Data Mining**. [*S. l.*]: CRC Press, 2009. Disponível em: https://www.taylorfrancis.com/books/edit/10.1201/9781420089653/top-ten-algorithms-data-mining-xindong-wu-vipin-kumar. Acesso em: 26 Feb. 2024.