TabM para Diagnóstico de Doença Cardíaca: Ensembling Eficiente de Redes Neurais em Dados Tabulares Clínicos

TabM for Heart Disease Diagnosis: Efficient Ensemble of Neural Networks in Clinical Tabular Data

Matheus Vinicius Cunha Gomes
mvcg@discente.ifpe.edu.br
Luciano de Souza Cabral
luciano.cabral@jaboatao.ifpe.edu.br

RESUMO

As doenças cardiovasculares representam a principal causa de mortalidade global, sendo responsáveis por aproximadamente 19,8 milhões de óbitos anuais, tornando o diagnóstico precoce fundamental para reduzir a mortalidade e melhorar a qualidade de vida dos pacientes. Este trabalho investigou a aplicação do modelo TabM, arquitetura moderna de *deep learning* baseada em redes neurais densas *(MLPs)* e *ensembling parameter-efficient* que, apesar de não ser o estado da arte absoluto, demonstrou resiliência única a *overfitting* entre modelos de *deep learning* tabular e alcançou posicionamento de 2º lugar no *benchmark* TabArena, ao diagnóstico de doença cardíaca utilizando o *dataset Cleveland Heart Disease*. O objetivo foi avaliar se essa arquitetura competitiva em *benchmarks* tabulares gerais alcançaria desempenho competitivo com os melhores resultados reportados na literatura para diagnóstico cardíaco. A metodologia envolveu seleção de *features* clínicas relevantes, aplicação de técnicas de balanceamento de classes, otimização de hiperparâmetros e validação cruzada estratificada. Os resultados não suportaram plenamente as hipóteses experimentais, levando à aceitação da hipótese nula. O modelo demonstrou robustez em *AUC-ROC*, porém ficou abaixo dos critérios estabelecidos em acurácia, *recall* e *F1-score*. Além da validação experimental, o estudo entregou um Produto Mínimo Viável (*MVP*) funcional em *Streamlit* para inferência clínica interativa, diferencial significativo em relação aos trabalhos relacionados que não forneceram ferramentas operacionais. Este trabalho representa a primeira aplicação conhecida de TabM ao diagnóstico de doença cardíaca, estabelecendo *benchmark* para futuras pesquisas e *pipeline* reprodutível de código aberto.

Palavras-chave: Deep learning tabular. Cleveland Heart Disease. TabM. Redes neurais. SMOTE.

Instituto Federal de Pernambuco. Campus Jaboatão dos Guararapes. Curso de Análise e Desenvolvimento de Sistemas. 15 de janeiro de 2018(data de submissão e aprovação do artigo).

ABSTRACT

Cardiovascular diseases represent the main cause of global mortality, being responsible for approximately 19.8 million deaths annually, making early diagnosis fundamental to reducing mortality and improving patients' quality of life. This work investigated the application of the TabM model, a modern deep learning architecture based on dense neural networks (MLPs) and parameter-efficient ensembling that, despite not being the absolute state of the art, demonstrated unique resilience to overfitting among tabular deep learning models and achieved 2nd place ranking in the TabArena benchmark, for diagnosing heart disease using the Cleveland Heart Disease dataset. The objective was to evaluate whether this competitive architecture in general tabular benchmarks would achieve competitive performance with the best results reported in the literature for cardiac diagnosis. The methodology involved selection of relevant clinical features, application of class balancing techniques, hyperparameter optimization, and stratified cross-validation. The results did not fully support the experimental hypotheses, leading to the acceptance of the null hypothesis. The model demonstrated robustness in AUC-ROC, however it fell below the established criteria in accuracy, recall, and F1-score. Beyond the experimental validation, the study delivered a functional Minimum Viable Product (MVP) in Streamlit for interactive clinical inference, a significant differentiator compared to related works that did not provide operational tools. This work represents the first known application of TabM to heart disease diagnosis, establishing a benchmark for future research and a reproducible open-source pipeline.

Keywords: Tabular deep learning. Cleveland Heart Disease. TabM. Neural networks. SMOTE;

1 INTRODUÇÃO

As doenças cardiovasculares (DCVs) representam a principal causa de mortalidade global, sendo responsáveis por aproximadamente 19,8 milhões de óbitos anuais, o que corresponde a 32% de todas as mortes no mundo (OMS, 2024). No Brasil, as DCVs foram responsáveis por mais de 400 mil óbitos em 2022, correspondendo a cerca de 1.100 mortes por dia (SBC, 2023). O diagnóstico precoce é fundamental para reduzir a mortalidade e melhorar a qualidade de vida dos pacientes, uma vez que intervenções terapêuticas em estágios iniciais demonstram eficácia significativamente superior (Mendis et al., 2011).

Nesse contexto, a inteligência artificial (IA) tem se destacado como ferramenta promissora no apoio ao diagnóstico médico, especialmente por meio de técnicas de aprendizado de máquina. Diversos estudos demonstraram que modelos clássicos, como *Random Forest*, *XGBoost* e redes neurais densas, conseguem atingir acurácias (percentual de predições corretas) superiores a 85% na predição de doença cardíacas utilizando o *dataset Cleveland Heart Disease,* que é um conjunto de dados (*dataset*) clássico composto por registros clínicos de 297 pacientes da *Cleveland Clinic Foundation*, amplamente adotado como referência*,* na literatura (Detreano et al., 1989; Ali et al., 2022; Khan et al., 2023).

Recentemente, surgiram arquiteturas de *deep learning* projetadas especificamente para dados tabulares, com o objetivo de capturar relações complexas entre *features* clínicas. O modelo TabM, proposto por Gorishniy, Kotelnikov e Babenko (2025), constitui um avanço relevante ao introduzir um mecanismo de *ensembling parameter-efficient*, onde *ensemble* refere-se à agregação estratégica de múltiplos modelos para produzir estimativas mais robustas e confiáveis, reduzindo variância e melhorando generalização, no *deep learning* tabular. Em vez de treinar múltiplos modelos independentes de forma custosa como ocorre nos *ensembles* tradicionais, o TabM permite que um único modelo imite simultaneamente o comportamento de várias redes neurais densas (*MLPs*), gerando múltiplas predições por instância através de compartilhamento eficiente de parâmetros. Essa estratégia confere ao TabM eficiência computacional superior e desempenho competitivo, como evidenciado por seu posicionamento de 2º lugar no *benchmark* TabArena (Erickson et al., 2025) e seu aceite na ICLR 2025 (Gorishniy; Kotelnikov; Babenko, 2025), uma das principais conferências mundiais de aprendizado profundo.

Entretanto, apesar da extensa literatura sobre predição de doença cardíaca no *dataset Cleveland* utilizando métodos clássicos e até algumas arquiteturas recentes baseadas em *Transformers*, como TabNet e TabPFN, não foram identificados estudos que tenham aplicado especificamente o modelo TabM nesse contexto. Essa lacuna é notável, pois o TabM, lançado em 2024 e aceito na *ICLR* 2025, demonstrou posicionamento competitivo em *benchmarks* tabulares gerais, estabelecendo-se como referência em *deep learning* tabular moderno. Além disso, o TabM se destaca por sua resiliência única a *overfitting* entre modelos de *deep learning* tabular, mantendo performance estável mesmo durante extensiva otimização de hiperparâmetros, ao contrário de outras arquiteturas neurais que degradam com *overtuning* (Erickson et al., 2025). O presente trabalho busca explorar o potencial do TabM em contextos clínicos realistas, investigando se suas características inovadoras de *ensembling parameter-efficient* e resiliência a *overfitting* podem ser aproveitadas para diagnóstico de doença cardíaca em *datasets* pequenos como o *Cleveland*, composto por apenas 303 amostras.

Diante deste cenário, emerge a seguinte pergunta de pesquisa: O TabM, cujas características de robustez a *overfitting* e eficiência computacional o posicionam como referência em *deep learning* tabular, é capaz de alcançar desempenho competitivo com os melhores resultados reportados na literatura para diagnóstico de doença cardíaca no *dataset Cleveland*?

Para responder a essa questão, baseado na análise estatística de 9 estudos recentes (2018-2025) compilados na Seção 3, este trabalho estabelece as seguintes hipóteses experimentais através do cálculo da média aritmética das métricas reportadas nos trabalhos relacionados identificados. Esse procedimento metodológico resulta nos seguintes limiares de desempenho:

H_1 (hipótese alternativa): O TabM alcança desempenho competitivo simultaneamente em todas as métricas principais, definido como:

- Acurácia ≥ 87,78%

- *AUC-ROC* ≥ 90,39%
- *Recall* ≥ 85,79%
- *F1-Score* ≥ 87,76%

H_0 (hipótese nula): O TabM apresenta desempenho inferior à performance média em uma ou mais métricas, revelando limitações de sua aplicabilidade em conjuntos de dados clínicos pequenos como o *Cleveland*.

O objetivo geral deste trabalho é avaliar o desempenho do modelo TabM na tarefa de classificação binária de doença cardíaca utilizando o conjunto de dados *Cleveland*, comparando os resultados obtidos com os principais *benchmarks* da literatura dos últimos anos por meio de experimentação controlada e reprodutível.

A contribuição deste trabalho é realizar a primeira aplicação conhecida do modelo TabM ao diagnóstico de doença cardíaca no *dataset Cleveland*. O TabM, aceito na *ICLR* 2025 e posicionado como referência em *deep learning* tabular moderno, nunca havia sido aplicado nesse contexto. Assim, o estudo fornece um *benchmark* adicional dessa arquitetura em um dos conjuntos de dados médicos mais estudados da literatura, validando a transferência de seu desempenho de *benchmarks* gerais para domínios clínicos específicos. Adicionalmente, o trabalho disponibiliza *pipeline* completo, hiperparâmetros otimizados e código aberto, facilitando futuras comparações, reprodução e extensões por pesquisadores.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta os conceitos-chave que sustentam a análise realizada neste trabalho. Inicialmente, discute-se o contexto das doenças cardiovasculares e o *dataset Cleveland*. Em seguida, aborda-se o conceito de Redes Neurais Densas (*MLPs*), modelo *TabM* e técnicas complementares como *SMOTE* e *Optuna*. Por fim, apresentam-se as métricas de avaliação utilizadas.

2.1 Doenças Cardiovasculares e Diagnóstico Precoce

As doenças cardiovasculares (DCVs) continuam sendo a principal causa de morte no mundo, responsáveis por aproximadamente 19,8 milhões de óbitos anuais, o que corresponde a 32% de todas as mortes globais (WHO, 2024). No Brasil, são mais de 400 mil óbitos por ano, ou cerca de 1.100 mortes por dia (SBC, 2024).

A doença arterial coronariana (DAC), também chamada doença coronariana ou cardiopatia isquêmica, é a forma mais comum e letal das DCVs, caracterizando-se pelo estreitamento progressivo das artérias coronárias devido ao acúmulo de placas ateroscleróticas, podendo levar à angina estável, síndrome coronariana aguda, infarto do miocárdio ou morte súbita cardíaca (Roth et al., 2020; Virani et al., 2021).

O diagnóstico precoce é clinicamente crítico, pois intervenções realizadas em fases iniciais, tais como mudanças no estilo de vida, uso de estatinas, antiplaquetários, betabloqueadores ou procedimentos de revascularização, reduzem significativamente o risco de eventos cardiovasculares maiores e a mortalidade

(Mendis et al., 2011; Araújo et al., 2022). Nesse sentido, sistemas de suporte à decisão clínica baseados em inteligência artificial têm sido amplamente investigados como ferramentas complementares ao julgamento médico, com potencial de aumentar a precisão e a rapidez na estratificação de risco de pacientes com suspeita de DAC (Ali et al., 2021; Bashir et al., 2014; Khan et al., 2023).

2.2 Redes Neurais Densas (Multilayer Perceptrons)

As Redes Neurais Densas, também conhecidas como *Multilayer Perceptrons* (*MLPs*), constituem uma arquitetura fundamental em aprendizado profundo (LeCun, Bengio & Hinton, 2015). Uma *MLP* é composta por múltiplas camadas de neurônios organizados sequencialmente: uma camada de entrada (que recebe os dados brutos), uma ou mais camadas ocultas (que aprendem representações intermediárias não-lineares dos dados) e uma camada de saída (que produz a predição final) (LeCun, Bengio & Hinton, 2015). Cada neurônio em uma camada conecta-se a todos os neurônios da camada subsequente através de pesos aprendíveis, e a passagem de dados através da rede (*forward pass*) envolve multiplicações matriciais sucessivas intercaladas com funções de ativação não-lineares (como ReLU ou Sigmoid), permitindo ao modelo capturar relações complexas entre as features de entrada (LeCun, Bengio & Hinton, 2015).

Em contexto de dados tabulares médicos, as *MLPs* oferecem vantagem importante em relação a métodos tradicionais baseados em árvores: sua capacidade de aprender interações não-lineares entre variáveis clínicas, potencialmente capturando padrões fisiopatológicos complexos que caracterizam a doença arterial coronariana (LeCun, Bengio & Hinton, 2015). Historicamente, porém, os modelos baseados em *ensemble* de árvores (como XGBoost) superavam *MLPs* em dados tabulares pequenos devido ao risco de *overfitting* das redes neurais (Grinsztajn et al., 2022). O desenvolvimento de arquiteturas recentes como TabM buscou resolver essa limitação através de mecanismos de regularização e *ensemble* implícito, tornando MLPs novamente competitivas nesse domínio (Gorishniy, Kotelnikov & Babenko, 2025).

2.3 TabM

O TabM (Gorishniy, Kotelnikov & Babenko, 2025) emerge como a referência atual em *deep learning* tabular, integrando o poder dos *ensembles* de forma eficiente em uma arquitetura baseada em redes neurais densas (*multilayer perceptrons*). Diferentemente de abordagens anteriores puramente baseadas em atenção (como *TabTransformer* ou *FT-Transformer*), que podem ser computacionalmente onerosas, o TabM emprega um mecanismo de *ensembling parameter-efficient*: um único modelo simula simultaneamente o comportamento de múltiplos *MLPs* via *batch ensembling*, compartilhando a maior parte dos parâmetros e gerando k predições independentes por instância através de adaptações locais. Essa estratégia, inspirada em *ensembles* paralelos, confere alta robustez e eficiência, reduzindo *overfitting* em dados heterogêneos ao forçar diversidade nas predições sem treinamento sequencial custoso (Gorishniy; Kotelnikov; Babenko, 2025).

A validação experimental do TabM ocorreu em dois contextos de *benchmark* complementares. Primeiro, Gorishniy, Kotelnikov e Babenko (2025) avaliaram o modelo em extensivo *benchmark* contendo 46 *datasets* tabulares públicos com propriedades variadas (tamanho de treino entre 1,8K e 723K amostras, número de

features entre 3 e 986). Nesse contexto, o TabM estabeleceu novo estado da arte, superando tanto *GBMs* (como *XGBoost* e *LightGBM*) quanto *Transformers* tradicionais em métricas como *AUC-ROC* e *F1-score*, com ganhos particularmente significativos em classificação binária.

Complementarmente, o TabM foi avaliado no recente *benchmark* TabArena (Erickson et al., 2025), um sistema de *benchmarking* vivo e mantido para *deep learning* tabular com rigorosa curadoria manual de *datasets*. No TabArena, o TabM alcançou posicionamento competitivo de 2º lugar (1541 Elo), demonstrando desempenho robusto em relação a outras arquiteturas modernas. Particularmente relevante, o TabM se destacou por resiliência única a *overfitting* entre modelos de deep learning tabular, mantendo performance estável mesmo durante extensiva otimização de hiperparâmetros, característica crítica para aplicações em datasets pequenos e clinicamente restritivos (Erickson et al., 2025). Por isso, o TabM foi escolhido como modelo central neste trabalho. Sua resiliência comprovada a *overfitting* é particularmente adequada para o contexto do *dataset Cleveland* (303 amostras), cenário onde o risco de *overfitting* é elevado.

2.4 Desbalanceamento de Classes e Técnicas de Oversampling

Desbalanceamento de classes ocorre quando uma classe (geralmente a de interesse clínico, como a presença de doença) possui significativamente menos exemplos que a outra (Chawla et al., 2002). Em diagnóstico médico, isso pode levar a modelos enviesados que priorizam a classe majoritária, aumentando falsos negativos, erro de maior gravidade clínica (Sowjanya; Mrudula, 2021). O SMOTE (*Synthetic Minority Over-sampling Technique*), proposto por Chawla et al. (2002), é a técnica mais consolidada de *oversampling*, que gera amostras sintéticas da classe minoritária por interpolação entre vizinhos próximos, com variantes específicas para dados mistos numéricos e categóricos.

2.5 Otimização Bayesiana de Hiperparâmetros

Hiperparâmetros são parâmetros do modelo que não são aprendidos diretamente a partir dos dados durante o treinamento, mas definidos previamente pelo pesquisador ou otimizados separadamente (Akiba et al., 2019). Exemplos típicos incluem: número de camadas ou blocos, dimensionalidade dos *embeddings*, taxa de aprendizado (*learning rate*), taxa de *dropout*, *weight decay*, número de cabeças de atenção e, no caso do TabM, parâmetros específicos como *n_blocks*, *d_block*, *n_bins* e o fator k do *ensemble* implícito (Gorishniy; Kotelnikov; Babenko, 2025).

A busca exaustiva (*grid search*) ou aleatória (*random search*) torna-se impraticável quando o espaço é grande e cada avaliação é cara, situação comum em modelos *Transformer* e no TabM *(*Bergstra *et al., 2011)*. A otimização bayesiana resolve esse problema tratando a função objetivo (por exemplo, *AUC-ROC* médio em validação cruzada) como uma função desconhecida e construindo um modelo probabilístico substituto (*surrogate*), geralmente um *Gaussian Process* ou *Tree-structured Parzen Estimator* (*TPE*), para prever o desempenho de novas combinações de hiperparâmetros (Snoek et al., 2012). A cada iteração (*trial*), o algoritmo equilibra exploração (testar regiões pouco conhecidas) e explotação (aproveitar regiões já promissoras) por meio de uma função de aquisição (Akiba et al., 2019).

A biblioteca *Optuna* implementa essa abordagem de forma eficiente, permitindo paralelismo, poda precoce de *trials* pouco promissores e integração direta com validação cruzada interna (Akiba et al., 2019; Bergstra et al., 2011; Snoek et al., 2012). Em trabalhos recentes com modelos tabulares profundos, a otimização bayesiana de 50 a 200 trials tem sido suficiente para encontrar configurações que superam buscas manuais ou aleatórias, reduzindo significativamente o tempo total de experimentação sem perda relevante de desempenho (Gorishniy et al., 2024; Hollmann et al., 2023).

2.6 Métricas de Avaliação em Classificação Médica

Em tarefas de classificação binária, como o diagnóstico de presença (classe positiva) ou ausência (classe negativa) de doença arterial coronariana, as predições do modelo podem ser organizadas em uma matriz de confusão com quatro valores (Powers, 2011):

- **TP** (*True Positive*): pacientes doentes corretamente classificados como doentes
- **TN** (*True Negative*): pacientes saudáveis corretamente classificados como saudáveis
- **FP** (*False Positive*): pacientes saudáveis incorretamente classificados como doentes
- **FN** (*False Negative*): pacientes doentes incorretamente classificados como saudáveis

A partir desses valores, definem-se as principais métricas (Powers, 2011):

Acurácia: Proporção de previsões corretas em relação ao total de amostras. É a métrica mais intuitiva, porém pode ser enganosa em conjuntos desbalanceados.

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

Precisão: Dentre todos os pacientes que o modelo classificou como doentes, qual fração realmente está doente. É especialmente relevante quando o custo de um falso positivo é alto (ex.: procedimentos invasivos desnecessários).

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

Sensibilidade ou *Recall*: Dentre todos os pacientes realmente doentes, qual fração o modelo conseguiu identificar. Em aplicações médicas, é a métrica mais crítica, pois falsos negativos representam casos graves não detectados.

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

F1-score: Média harmônica entre precisão e *recall*, útil quando se deseja um único número que equilibre ambas as métricas, especialmente em conjuntos com leve desbalanceamento.

F1 = $\frac{2 \times (Precisão \times Recall)}{(Precisão + Recall)}$

***AUC-ROC* (*Area Under the Receiver Operating Characteristic Curve*)**

Representa a probabilidade de que, ao escolher aleatoriamente um paciente doente e um saudável, o modelo atribua maior score de risco ao doente. Varia de 0 a 1 (1 = discriminação perfeita). Por ser independente do *threshold* de decisão, é a métrica mais robusta para comparação entre modelos em validação cruzada. (Fawcett, 2006).

Embora a AUC-ROC forneça uma visão global da capacidade discriminativa do modelo independente do threshold, em aplicações clínicas é frequentemente necessário selecionar um threshold específico para a operacionalização da ferramenta de diagnóstico. O Índice de Youden (J), proposto por Youden (1950), fornece um critério objetivo para esta seleção. Definido como J = Sensibilidade + Especificidade − 1 = TPR − FPR, o Índice de Youden maximiza a soma de verdadeiros positivos e verdadeiros negativos, selecionando o ponto da curva ROC que oferece o melhor equilíbrio entre detectar corretamente pacientes doentes e evitar falsos alarmes em pacientes saudáveis. Esse critério é particularmente relevante em contextos médicos onde tanto falsos negativos (diagnósticos perdidos) quanto falsos positivos (investigações desnecessárias) possuem custos clínicos distintos (Youden, 1950).

2.7 O Dataset Cleveland Heart Disease

O *dataset Cleveland Heart Disease* é um dos conjuntos de dados médicos mais utilizados na literatura de aprendizado de máquina desde sua publicação, em 1988, pelos pesquisadores Robert Detrano e colaboradores da *Cleveland Clinic Foundation* (Detrano et al., 1989). Ele contém informações clínicas e angiográficas de 303 pacientes submetidos a cateterismo cardíaco, com o objetivo original de desenvolver um algoritmo probabilístico para diagnóstico de doença arterial coronariana.

Cada registro inclui 13 variáveis clínicas selecionadas pelos cardiologistas como as mais relevantes na época (idade, tipo de dor torácica, pressão arterial em repouso, colesterol total, glicemia de jejum, resultados de eletrocardiograma em repouso e sob esforço, entre outras) e um rótulo binário que indica a presença ou ausência de doenças cardíacas.

2.8 Validação Cruzada em Estudos com Conjuntos Pequenos

A validação cruzada (*cross-validation*) é um procedimento estatístico essencial para estimar de forma robusta o desempenho de modelos de aprendizado de máquina quando o volume de dados é limitado, como no *dataset Cleveland* com apenas cerca de 303 amostras. Ela evita o viés de uma única partição aleatória (*hold-out*) e fornece uma avaliação mais confiável da capacidade de generalização do modelo, calculando a média e o desvio padrão das métricas ao longo de múltiplas execuções (Kohavi, 1995; Arlot; Celisse, 2010).

A validação cruzada estratificada de 10 *folds* (*10-fold stratified cross-validation*), cuja mecânica funciona da seguinte forma (Kohavi, 1995; Arlot; Celisse, 2010):

1. O conjunto de dados é dividido em 10 partes (*folds*) de tamanho aproximadamente igual, preservando-se a proporção original entre as classes positiva (presença de doença) e negativa (ausência de doença) em cada *fold*.

2. O treinamento e a avaliação são realizados em 10 iterações: a cada iteração, 9 *folds* compõem o conjunto de treinamento e o fold restante é usado exclusivamente como conjunto de teste.

3. As métricas de desempenho são calculadas para o *fold* de teste em cada iteração.

4. O resultado final do modelo é dado pela média das 10 avaliações, acompanhada do desvio padrão, que indica a estabilidade do desempenho entre diferentes partições dos dados.

Esse procedimento garante que todas as amostras sejam utilizadas tanto para treinamento quanto para teste exatamente uma vez, maximizando o aproveitamento dos dados disponíveis e reduzindo o viés associado a uma única divisão treino-teste (Kohavi, 1995; Arlot; Celisse, 2010).

3 TRABALHOS RELACIONADOS

Esta seção revisa estudos recentes sobre predição de doença cardíaca utilizando aprendizado de máquina, com foco específico no *dataset Cleveland Heart Disease*. Embora outros *datasets* cardíacos públicos existam no repositório *UCI* (*Hungarian* - 294 instâncias, *Switzerland* - 213 instâncias, *VA Long Beach* - 200 instâncias), este trabalho restringe a revisão exclusivamente a estudos que utilizam o *dataset Cleveland* puro (303 instâncias originais ou 297 após remoção de valores faltantes). A revisão abrange publicações de 2018 a 2025.

Ali et al. (2018) propuseram um sistema híbrido de predição de doença cardíaca combinando seleção de atributos (*Relief*, *mRMR* e *LASSO*) com sete classificadores (*Logistic Regression*, *K-NN*, *ANN*, *SVM*, *Decision Tree*, *Naive Bayes* e *Random Forest*). Utilizaram o *Cleveland* reduzido para 297 instâncias (após remoção de valores faltantes) e validação cruzada de 10 *folds*. Com seleção de atributos via *Relief*, a Regressão Logística alcançou a melhor performance: acurácia de 89,00%, especificidade de 98%, sensibilidade de 77% e *AUC* de 0,88 e *MCC* de 0,89.

Latha e Jeeva (2019) investigaram técnicas de *ensemble* (*bagging* e *boosting*) para melhorar algoritmos fracos na predição de risco cardíaco. Aplicaram *majority vote* com *Naïve Bayes*, *Bayesian Network*, *Random Forest* e *MLP* no *dataset Cleveland* (303 instâncias), usando validação cruzada de 10 *folds* e seleção de atributos. O *ensemble* proposto obteve acurácia de 85,48%, representando ganho médio de até 7% em relação aos classificadores individuais.

Amin, Chiam e Varathan (2019) focaram na identificação de *features* significativas para predição de doença cardíaca. Testaram sete técnicas (*k-NN*, *Decision Tree*, *Naive Bayes*, *Logistic Regression*, *SVM*, *Neural Network* e *Vote* híbrido de *Naive Bayes* + *Logistic Regression*) no *dataset Cleveland* puro (303 instâncias), utilizando validação cruzada de 10 *folds*. O modelo *Vote* com as features selecionadas alcançou a melhor acurácia de 87,40%.

Reddy et al. (2021) avaliaram múltiplos classificadores com e sem seleção de atributos no *Cleveland* (303 instâncias) usando validação cruzada de 10 *folds*. O melhor resultado foi obtido com *Chi-Squared* + *Sequential Minimal Optimization* (*SMO*), alcançando acurácia de 86,47%, precisão e sensibilidade de 86,5% e *AUC* de 0,861.

Fajri, Wiharto & Suryani (2022) propuseram seleção de atributos híbrida usando *Bee Swarm Optimization* + *Q-Learning* (*QBSO-FS*) combinada com diversos classificadores. Avaliaram o modelo no *Cleveland* puro (303 instâncias) com validação cruzada de 10 *folds*, obtendo acurácia de 84,40%, precisão de 86%, recall de 80% e AUC de 0,901.

Adekoya et al. (2025) propuseram o *Interpretable Ensemble Learning Framework* (*IELF*), abordagem que integra *Explainable Boosting Machines* (*EBM*) com *XGBoost*, complementado por técnicas de interpretabilidade *SHAP* e *LIME*. No *Cleveland*, o *IELF* alcançou acurácia de 88.52%, *AUC-ROC* de 0.899 (89.9%), *recall* de 92.86% e *F1-score* de 88.14%.

Shrestha (2024) realizou estudo comparativo de técnicas avançadas de aprendizado de máquina avaliando cinco modelos: *Logistic Regression*, *Random Forest*, *Gradient Boosting*, *XGBoost* e redes *LSTM* (*Long Short-Term Memory*). O modelo *XGBoost* obteve o melhor desempenho com acurácia de 90% e *AUC-ROC* de 0.94, superando os demais algoritmos avaliados.

Ghose et al. (2025) propuseram abordagem baseada em *Stacked Ensemble Classifier* integrando múltiplas técnicas de ML para detecção de doença cardíaca no *Cleveland*. Utilizando validação cruzada, o modelo alcançou acurácia de 93.3% ± 2.0%, precisão de 91.9% ± 2.2%, sensibilidade de 96.2% ± 1.8%, especificidade de 72.1% ± 2.5%, *F1-score* de 94.0% ± 1.9% e *AUC-ROC* de 93.5% ± 1.8%.

Kolukisa e Bakir-Gungor (2023) propuseram abordagens de *ensemble feature selection* combinando sete métodos computacionais de seleção de *features* com um método baseado em conhecimento de domínio para diagnóstico de doença arterial coronariana. No *Cleveland*, o classificador *Multi-Layer Perceptron* (*MLP*) combinado com método *CMIM* (*Conditional Mutual Information Maximization*) para seleção de 9 *features* alcançou acurácia de 85.47%, sensibilidade de 82.96%, precisão de 86.22%, *F-measure* de 0.839 e *AUC* de 0.911 (91.1%).

Conforme observado na Tabela 1, os desempenhos reportados na literatura para o *dataset Cleveland* puro com validação cruzada variam entre 84,40% e 93,33% em acurácia, com média de 87,78%. A maioria dos estudos (2018-2023) alcançou desempenhos entre 84% e 89% utilizando *ensembles* tradicionais ou modelos clássicos com seleção de atributos. Estudos mais recentes (2024-2025) demonstraram desempenhos superiores: Shrestha (2024) com XGBoost alcançou 90,00% de acurácia, enquanto Ghose et al. (2025) com *Stacked Ensemble* reportou 93,33%, estabelecendo novos patamares de performance nesse *dataset*. Entretanto, nenhum dos estudos revisados aplicou o modelo TabM, deixando aberta a questão de se essa arquitetura é capaz de competir simultaneamente em múltiplas métricas (acurácia, *AUC-ROC*, *recall* e *F1-score*) em um cenário tão restritivo e clinicamente realista como o *dataset Cleveland*. Essa lacuna representa a motivação central do presente trabalho.

Tabela 1 – Comparação de Trabalhos Relacionados em Predição de Doença Cardíaca no Dataset Cleveland

Autores	Modelo principal	Acurácia	AUC-ROC	Recall	F1-Score
Ali et al. (2018)	*Logistic Regression* + Relief FS	89,00 %	88.00%	77.00%	-
Latha & Jeeva (2019)	*Majority vote* (*NB* + *BN* + *RF* + *MLP*) + *FS*	85,48 %	-	-	-
Amin, Chiam & Varathan (2019)	*Vote* (*NB* + *LR*) + *feature selection*	87,40 %	-	-	-
Reddy et al. (2021)	*Chi-Squared* + *SMO*	86,47 %	86.10%	86.50%	-
Fajri, Wiharto & Suryani (2022)	*QBSO-FS* + classificadores	84,40 %	90.10%	80.00%	-
Kolukisa & Bakir-Gungor (2023)	*MLP* + *CMIM ensemble FS*	85.47%	91.10%	82.96%	83.90%
Shrestha (2024)	*XGBoost*	90.00%	94.00%	85.00%	85.00%
Adekoya et al. (2025)	*IELF* (*EBM* + *XGBoost* + *SHAP*/*LIME*)	88.52%	89.90%	92.86%	88.14%
Ghose et al. (2025)	*Stacked Ensemble*	93.33%	93.50%	96.20%	94.00%
MÉDIA		**87,78%**	**90,39%**	**85,79%**	**87,76%**

Fonte: os autores (2025)

4 METODOLOGIA

Este trabalho fundamenta-se na taxonomia metodológica proposta por Gil (2017) e classifica-se, quanto à sua natureza, como pesquisa aplicada, uma vez que busca gerar conhecimento voltado à solução de um problema concreto: a aplicação de arquiteturas modernas de *deep learning* tabular ao diagnóstico de doença cardíaca. Quanto à abordagem, trata-se de uma pesquisa quantitativa, pois utiliza dados numéricos estruturados e métricas estatísticas objetivas (acurácia, precisão, *recall*, *F1-score* e *AUC-ROC*) para avaliar o desempenho do modelo TabM. Do ponto de vista dos objetivos, caracteriza-se como pesquisa exploratória e descritiva: exploratória por investigar a viabilidade de uma arquitetura recente (TabM) em um contexto ainda não explorado na literatura (diagnóstico cardíaco no *dataset Cleveland*); descritiva por realizar *benchmarking* sistemático das métricas obtidas frente ao estado da arte reportado em estudos anteriores. Quanto aos procedimentos técnicos, enquadra-se como pesquisa experimental, dado que envolve a manipulação controlada de hiperparâmetros, técnicas de balanceamento e configurações do modelo, seguida de validação rigorosa via validação cruzada estratificada de 10 *folds*.

A organização metodológica seguiu um *pipeline* estruturado em cinco etapas principais, cada uma desempenhando papel crítico na integridade, reprodutibilidade e validade dos resultados. Essa estrutura sequencial garante que decisões tomadas em fases preliminares, como análise exploratória e seleção de *features*, não comprometam a avaliação final do modelo, ao mesmo tempo em que permite rastreabilidade completa de todas as transformações aplicadas aos dados.

Para facilitar a compreensão do percurso experimental, a Figura 1 apresenta um fluxograma detalhado que sintetiza visualmente as principais etapas do trabalho, desde a coleta e preparação dos dados até a análise final dos resultados. Esse diagrama ilustra a lógica do experimento de forma estruturada, destacando os marcos essenciais e as interdependências entre as fases.

As etapas que compõem o *pipeline* experimental são descritas a seguir:

- **Coleta e Preparação dos Dados**
  Aquisição do *dataset Cleveland Heart Disease* do *UCI Machine Learning Repository*, remoção de instâncias com valores ausentes e caracterização inicial do conjunto de dados.

- **Análise Exploratória dos Dados (*EDA*)**
  Investigação das distribuições estatísticas das *features*, análise de correlações, identificação de padrões de desbalanceamento e visualização das características do *dataset.* .

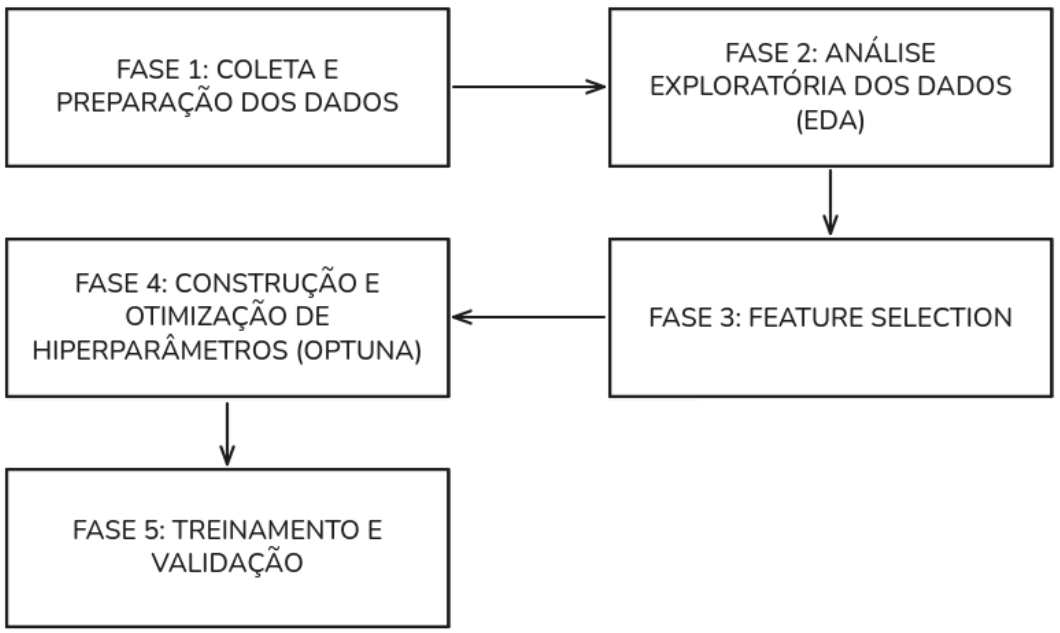
- ***Feature Selection***
  Seleção orientada por dados utilizando *permutation importance*, medindo o impacto de cada *feature* no *AUC-ROC* do modelo TabM durante validação

cruzada de 10 *folds*, reduzindo de 13 para 6 atributos (1 numérica e 5 categóricas).

- **Construção e Otimização de Hiperparâmetros (*Optuna*)**
  Implementação da arquitetura TabM e busca automática dos melhores hiperparâmetros via otimização bayesiana com *Optuna*, utilizando validação cruzada interna de 10 *folds*.

- **Treinamento e Validação**
  Treinamento do modelo TabM com os hiperparâmetros otimizados, aplicação de *SMOTE* para balanceamento de classes e validação cruzada estratificada de 10 *folds* como estratégia principal de avaliação.

Figura 1 – Fluxograma metodológico da pesquisa



Fonte: os autores (2025)

Essas etapas organizaram a condução da pesquisa e dos experimentos, assegurando a reprodutibilidade e a confiabilidade dos resultados. Cada fase será detalhada nas subseções seguintes, fornecendo descrição completa dos procedimentos metodológicos adotados.

4.1 Coleta e Preparação dos Dados

O conjunto de dados utilizado neste trabalho é o *Cleveland Heart Disease Dataset*, disponibilizado publicamente no *UCI Machine Learning Repository* e originalmente coletado por Detrano et al. (1989) na *Cleveland Clinic Foundation*. Trata-se de um dos datasets médicos mais referenciados na literatura de

aprendizado de máquina aplicado à cardiologia, contendo informações clínicas e angiográficas de pacientes submetidos a cateterismo cardíaco para investigação de doença arterial coronariana.

O *dataset* original é composto por 303 instâncias (pacientes) e 76 atributos clínicos. No entanto, seguindo o padrão adotado pela maioria dos estudos da área, foi utilizada a versão processada contendo 13 *features* clínicas selecionadas pelos cardiologistas como as mais relevantes para o diagnóstico. Após a remoção de 6 instâncias com valores ausentes em *features* críticas, o conjunto final utilizado contém 297 amostras válidas, todas com informações completas para as 13 variáveis.

A variável-alvo é binária, indicando a presença (classe 1) ou ausência (classe 0) de estenose coronariana significativa, definida como estreitamento superior a 50% do diâmetro luminal em pelo menos uma das principais artérias coronárias, conforme avaliação angiográfica. A distribuição das classes é ligeiramente desbalanceada: 160 pacientes saudáveis (53,9%) e 137 pacientes doentes (46,1%), resultando em uma razão de 1,17:1.

Esse desbalanceamento moderado, embora menos severo que em muitas aplicações médicas reais, ainda pode influenciar o treinamento de modelos de aprendizado de máquina, tendendo a favorecer a classe majoritária e aumentar a taxa de falsos negativos, erro particularmente crítico em contextos clínicos. Por essa razão, técnicas de balanceamento de classes foram incorporadas ao pipeline experimental, conforme descrito na Fase 5.

O processo de preparação dos dados nesta fase incluiu:

1. *Download* do *dataset* a partir do repositório oficial *UCI*[1].
2. Inspeção inicial da estrutura dos dados, tipos de variáveis e presença de valores ausentes.
3. Remoção de instâncias incompletas: Exclusão das 6 amostras com valores faltantes (*missing values*), reduzindo o *dataset* de 303 para 297 instâncias.
4. Verificação de integridade: Confirmação de que todas as 297 instâncias remanescentes possuem valores válidos para todas as 13 *features* e para a variável-alvo.
5. Codificação da variável-alvo: Transformação das classes originais (0, 1, 2, 3, 4) em formato binário, onde 0 representa ausência de doença e valores ≥1 representam presença de doença.

[1] http://archive.ics.uci.edu/dataset/45/heart+disease

4.2 Análise Exploratória dos Dados (EDA)

A análise exploratória dos dados (*Exploratory Data Analysis - EDA*) é etapa fundamental para compreender profundamente as características do *dataset*, identificar padrões, detectar possíveis *outliers*. Nesta fase, foram realizadas as análises de distribuições, desbalanceamento e de *outliers*. A análise exploratória de dados iniciou-se com caracterização estatística das 13 *features* originais do *dataset Cleveland*, abrangendo tanto variáveis numéricas quanto categóricas.

A análise de distribuição da variável-alvo binária (presença vs ausência de doença) foi conduzida para quantificar o desbalanceamento entre classes, calculando-se frequências absolutas, percentuais e razão de desbalanceamento. Essa caracterização fundamentou decisões posteriores sobre necessidade de técnicas de balanceamento de classes durante o treinamento do modelo.

Todas as visualizações foram geradas utilizando as bibliotecas *matplotlib*[2] e *seaborn*[3] em *Python*, com configurações padronizadas de estilo para garantir clareza e reprodutibilidade.

4.2.3 Detecção de Outliers

Outliers foram investigados utilizando o critério de intervalo interquartil (IQR), onde valores fora do intervalo [Q1 - 1,5×IQR, Q3 + 1,5×IQR] foram sinalizados. No contexto médico, *outliers* frequentemente representam casos clínicos extremos ou dados válidos de pacientes com condições atípicas, portanto, não foram removidos, mas identificados para análise de sua potencial influência no treinamento do modelo.

4.3 Feature Selection

A seleção de *features* foi realizada de forma empírica e orientada por dados, utilizando análise de importância baseada em *permutation importance*. Esse método quantifica a contribuição de cada feature ao desempenho do modelo medindo a degradação da métrica *AUC-ROC* quando os valores de uma feature específica são permutados aleatoriamente, rompendo sua relação com a variável-alvo.

O procedimento iniciou-se com o treinamento do modelo TabM utilizando todas as 13 *features* originais do *dataset Cleveland* através de validação cruzada estratificada de 10 *folds*, seguindo o protocolo descrito na Seção 4.5. Após o treinamento, para cada feature individualmente, foram permutados aleatoriamente seus valores no conjunto de teste de cada *fold*, mantendo todas as demais *features* inalteradas. A diferença entre o *AUC-ROC* do modelo com a *feature* original e o *AUC-ROC* após a permutação forneceu a medida de importância daquela *feature*. Esse procedimento foi repetido 10 vezes por *feature* para obter estimativa robusta da importância, mitigando efeitos de aleatoriedade na permutação.

As *features* foram então ranqueadas pela magnitude média da queda no *AUC-ROC* observada nos 10 *folds* da validação cruzada. *Features* que causavam quedas substanciais quando permutadas foram identificadas como de alto poder

[2] https://matplotlib.org/
[3] https://seaborn.pydata.org/

discriminativo, enquanto aquelas com quedas marginais ou até mesmo ganhos no *AUC-ROC* após permutação foram classificadas como de baixa contribuição ou potencialmente prejudiciais ao modelo. O critério de seleção estabelecido priorizou a manutenção de *features* cuja remoção causasse queda no *AUC-ROC*.

4.4 Construção e Otimização de Hiperparâmetros com *Optuna*

O modelo TabM foi implementado exatamente conforme a versão oficial lançada pelos autores (Gorishniy, Kotelnikov & Babenko, 2025), utilizando as bibliotecas tabm e *rtdl_num_embeddings*. A arquitetura recebe separadamente as *features* numéricas e categóricas:

- ***Features* categóricas** (*cp, exang, slope, ca, thal*): são convertidas em *embeddings* inteiros aprendíveis de baixa dimensão (tamanho definido por hiperparâmetro), seguindo a prática padrão em modelos *Transformer* para dados tabulares.

- ***Feature* numérica** (*oldpeak*): foi processada com *Piecewise Linear Embeddings* (*PLE*), técnica introduzida pelos próprios autores do TabM e considerada por eles a mais eficiente para variáveis contínuas em *datasets* tabulares. O *PLE* divide o domínio da variável em *n_bins* intervalos e representa cada valor como uma combinação linear dos vértices dos *bins*, resultando em representação densa, contínua e altamente expressiva com custo computacional mínimo.

Na inferência, o TabM gera internamente múltiplas predições por amostra (devido ao seu mecanismo de *batch ensembling*) e o valor final é a média probabilística média dessas predições, aumentando robustez sem custo adicional.

A otimização dos hiperparâmetros foi realizada por meio da biblioteca *Optuna*, utilizando otimização *bayesiana* (*Tree-structured Parzen Estimator*) com objetivo de maximizar o *AUC-ROC* médio. Foram executados 100 *trials* com as faixas de busca demonstradas na Tabela 2. A função objetivo incorporou validação cruzada interna de 10 *folds* estratificados por *fold* externo, com treinamento de apenas 30 épocas por *trial*.

O *SMOTE* híbrido foi aplicado exclusivamente no conjunto de treinamento de cada *fold* interno, com *k_neighbors* = 5. A implementação personalizada trata separadamente *features* numéricas (interpolação linear) e categóricas (seleção aleatória entre vizinhos), evitando artefatos comuns em bibliotecas padrão quando há mistura de tipos.

Tabela 2 - Faixas de busca para Otimização Bayesiana

Hiperparâmetro	Faixa testada	Tipo
n_blocks	1 – 4	inteiro
d_block	64 – 512 (passo 16)	inteiro
d_embedding	8 – 32 (passo 4)	inteiro
n_bins (*Piecewise Linear*)	2 – 64	inteiro
dropout	0,0 – 0,5	contínuo
learning rate	1×10^{-4} – 5×10^{-3} (escala log)	contínuo
weight_decay	1×10^{-5} – 1×10^{-1} (escala log)	contínuo
smote_sampling_strategy	0,86 – 1,0	contínuo

Fonte: os autores (2025)

4.5 Treinamento e Validação

Esta fase constitui o núcleo experimental do trabalho, onde o modelo TabM é efetivamente treinado e avaliado seguindo protocolos rigorosos de validação cruzada. Esta etapa integra os hiperparâmetros otimizados na Fase 4 com estratégias de balanceamento de classes, garantindo que a avaliação final seja realizada de forma justa, robusta e livre de vieses metodológicos que poderiam comprometer a validade dos resultados.

4.5.1 Estratégia de Validação Cruzada Estratificada

Diferentemente de abordagens convencionais que separam os dados em conjuntos fixos de treinamento, validação e teste, este trabalho adota validação cruzada estratificada de 10 *folds* como estratégia principal de avaliação. Essa escolha metodológica é fundamentada na limitação do tamanho amostral e na necessidade de maximizar o uso dos dados disponíveis, garantindo estimativas robustas e confiáveis do desempenho de generalização do modelo. A validação

cruzada estratificada é amplamente reconhecida como o padrão-ouro para avaliação de modelos em datasets pequenos, sendo utilizada pela maioria dos trabalhos relacionados revisados na Seção 3 (Kohavi, 1995; Arlot; Celisse, 2010).

O procedimento, implementado via biblioteca *scikit-learn*[4] com a classe *StratifiedKFold*, divide o conjunto completo de 297 amostras em 10 partições disjuntas de tamanho aproximadamente igual, com cerca de 30 amostras por *fold*. A divisão é estratificada, ou seja, mantém a proporção original das classes (53,9% saudáveis e 46,1% doentes) em cada *fold*, evitando que alguma partição contenha distribuição atípica que comprometa a representatividade. O experimento é então repetido 10 vezes, onde em cada iteração 9 *folds* são combinados para formar o conjunto de treinamento (aproximadamente 267 amostras) e o *fold* restante é utilizado exclusivamente como conjunto de teste (aproximadamente 30 amostras).

Em cada iteração, o modelo é treinado do zero, sem transferência de pesos entre *folds*, utilizando o conjunto de treinamento, e o desempenho é avaliado no *fold* de teste correspondente. Ao final das 10 iterações, as métricas de desempenho são agregadas, calculando-se a média aritmética e o desvio padrão entre os *folds*. Essa estratégia assegura que todas as 297 amostras sejam utilizadas exatamente uma vez para teste, eliminando o viés de uma única partição aleatória e permitindo análise detalhada da estabilidade do desempenho em diferentes subconjuntos dos dados.

4.5.2 Aplicação de *SMOTE* no *Pipeline* de Treinamento

Um aspecto do protocolo experimental é a aplicação correta de *SMOTE* para balanceamento de classes, seguindo as melhores práticas metodológicas para evitar vazamento de informação, fenômeno conhecido como *data leakage*. O *SMOTE* é aplicado exclusivamente no conjunto de treinamento de cada *fold*, nunca no conjunto de teste, seguindo o princípio de *nested cross-validation*. Esse cuidado metodológico garante que as métricas reportadas reflitam o desempenho real do modelo em dados clínicos não vistos, não artificialmente infladas pela presença de amostras sintéticas no conjunto de teste.

O procedimento específico em cada *fold* opera da seguinte forma: após o particionamento estratificado, o *fold* de teste é separado e mantido intocado com seus dados originais. Em seguida, o algoritmo *SMOTE* híbrido é aplicado apenas no conjunto de treinamento formado pelos 9 *folds* restantes, gerando amostras sintéticas da classe minoritária (pacientes doentes) até atingir a razão definida pelo parâmetro *smote_sampling_strategy* otimizado. O modelo TabM é então treinado utilizando o conjunto aumentado, composto pela combinação de dados originais e sintéticos, e finalmente avaliado exclusivamente nos dados originais do fold de teste.

[4] https://scikit-learn.org

Além disso, para garantir reprodutibilidade estrita, cada *fold* recebe uma semente aleatória única calculada como *RANDOM_STATE* + *fold_index*, onde *RANDOM_STATE* foi fixado em 42. Isso assegura que a geração de amostras sintéticas seja determinística e replicável, mas ao mesmo tempo independente entre folds, evitando correlações espúrias que poderiam surgir caso a mesma semente fosse reutilizada.

4.5.3 Arquitetura e Configuração do Modelo TabM

Para cada *fold* da validação cruzada, uma nova instância do modelo TabM é criada do zero utilizando os hiperparâmetros otimizados na Fase 4. A arquitetura é construída pela função *create_tabm_model*() ilustrada na Figura 2, que recebe como entrada o número de *features* numéricas (1, correspondente a *oldpeak*), as cardinalidades das *features* categóricas (valores [4, 2, 3, 4, 3] para *cp*, *exang*, *slope*, *ca* e *thal*, respectivamente), o dicionário de hiperparâmetros otimizados e os dados de treinamento necessários para calcular os bins dos embeddings numéricos.

Figura 2 - Função de criação do modelo TabM

```python
def create_tabm_model(n_num_features, cat_cardinalities, params, X_train=None, d_out=1):
    """
    Cria o modelo TabM com hiperparâmetros dinâmicos.
    Adiciona suporte a Embeddings Numéricos se especificado.
    """

    # Configuração de Embeddings Numéricos
    num_embeddings = None
    if params.get('use_embeddings', False) and X_train is not None:
        # Usando PiecewiseLinearEmbeddings com cálculo de bins
        bins = rtdl_num_embeddings.compute_bins(X_train, n_bins=params['n_bins'])

        num_embeddings = PiecewiseLinearEmbeddings(
            bins=bins,
            d_embedding=params['d_embedding'],
            activation=True,
            version='B'
        )

    model = TabM.make(
        n_num_features=n_num_features,
        cat_cardinalities=cat_cardinalities,
        d_out=d_out,
        n_blocks=params['n_blocks'],
        d_block=params['d_block'],
        dropout=params.get('dropout', 0.1),
        num_embeddings=num_embeddings
    )
    return model
```

Fonte: os autores (2025)

A configuração específica do modelo reflete diretamente os resultados da otimização bayesiana. Como *use_embeddings* foi identificado como *True*, a feature numérica *oldpeak* é processada via *PiecewiseLinearEmbeddings*, uma técnica

avançada que discretiza a *feature* contínua em 17 bins adaptativos (valor do hiperparâmetro *n_bins*) calculados sobre o conjunto de treinamento através da função *rtdl_num_embeddings.compute_bins*(). Cada *bin* é então mapeado para um *embedding* de dimensionalidade 28 (valor do hiperparâmetro *d_embedding*), permitindo que o modelo aprenda transformações não lineares complexas de forma mais eficiente que *embeddings* lineares simples. As *features* categóricas são tratadas via *one-hot encoding* implícito seguido de *embeddings* aprendidos, conforme implementação padrão do TabM.

O *backbone* do modelo é composto por 2 blocos *transformer* (n_blocks=2), cada um com dimensionalidade de 384 (d_block=384). Cada bloco incorpora mecanismo de *multi-head attention* que permite ao modelo capturar relações globais entre todas as *features* simultaneamente, atribuindo pesos dinâmicos à importância relativa de cada feature na decisão diagnóstica. Complementando a atenção, cada bloco contém uma *feed-forward network* com ativação *ReLU*, *residual connections* e *layer normalization* para estabilizar o treinamento, além de *dropout* de 0.0698 aplicado como regularização estocástica. O modelo gera k=32 predições independentes por instância (valor padrão do TabM), simulando um *ensemble* de 32 submodelos com compartilhamento eficiente de parâmetros, característica central da arquitetura TabM que proporciona robustez sem o custo computacional de ensembles tradicionais.

4.5.4 Processo de Treinamento

O treinamento de cada modelo segue um protocolo padronizado implementado na função *train_epoch*(), ilustrada na Figura 3, utilizando o otimizador *AdamW* proposto por Loshchilov e Hutter (2019). Trata-se de uma variante do *Adam* com correção de *weight decay*, configurado com taxa de aprendizado de 0.00401 e weight decay de 0.00309, valores identificados como ótimos durante a Fase 4. Os demais hiperparâmetros do *AdamW* (β_1=0.9, β_2=0.999) foram mantidos nos valores padrão do *PyTorch*. A função de perda utilizada é a *Binary Cross-Entropy with Logits* (*BCEWithLogitsLoss*), que combina a função sigmoide com a entropia cruzada binária, proporcionando estabilidade numérica superior à aplicação separada de sigmoide seguida de *BCE*.

Um aspecto crítico e distintivo do treinamento do TabM é que as k=32 predições geradas pelo modelo são tratadas de forma independente durante o cálculo da *loss*, conforme destacado na documentação oficial da arquitetura. Para cada *batch* de tamanho B (definido como 256 amostras), o modelo gera um tensor de predições com shape (B, k, 1), onde k=32. A variável-alvo é expandida para *shape* (B, k, 1) através de replicação, e a *loss* é calculada como a média da entropia cruzada binária sobre todas as B×k predições individuais. Essa abordagem força

diversidade entre os k submodelos do *ensemble* implícito e é fundamental para o correto funcionamento do mecanismo de *ensembling parameter-efficient* do TabM.

Figura 3 - Função de treinamento

```python
def train_epoch(model, loader, optimizer, criterion, device):
    model.train()
    total_loss = 0

    for X_num_batch, X_cat_batch, y_batch in loader:
        X_num_batch = X_num_batch.to(device).float()
        X_cat_batch = X_cat_batch.to(device)
        y_batch = y_batch.to(device).float()

        optimizer.zero_grad()

        # Forward pass
        y_pred = model(X_num_batch, X_cat_batch)

        # Loss média das k predições independentes
        y_target_expanded = y_batch.unsqueeze(1).expand(-1, model.k, -1)
        loss = criterion(y_pred.reshape(-1, 1), y_target_expanded.reshape(-1, 1))

        loss.backward()
        optimizer.step()

        total_loss += loss.item() * X_num_batch.size(0)
```

Fonte: os autores (2025)

Os dados são carregados via *torch.utils.data.DataLoader* com *batch size* de 256, valor que equilibra estabilidade do gradiente estocástico e eficiência computacional, sendo adequado ao tamanho do conjunto de treinamento de aproximadamente 267 amostras por *fold*. Durante o treinamento, a ordem dos *batches* é randomizada (*shuffle=True*), enquanto na avaliação a ordem é mantida determinística (*shuffle=False*). O treinamento é limitado a máximo de 100 épocas.

4.5.5 Avaliação e Agregação de Predições

Durante a fase de inferência, ou seja, avaliação no *fold* de teste, o modelo opera em modo *model.eval*(), desabilitando *dropout* e outras camadas estocásticas para garantir predições determinísticas. A função *evaluate*() implementa o protocolo de avaliação: para cada *batch* do conjunto de teste, o modelo realiza propagação *forward* gerando k=32 predições por instância, resultando em *tensor* de *shape* (B, k, 1). Essas k predições em formato de *logits* são então transformadas em probabilidades através da aplicação da função sigmoide e agregadas pela média aritmética ao longo da dimensão k, produzindo uma única probabilidade final por instância. Essa média de probabilidades, e não de *logits*, é a estratégia recomendada pela documentação oficial do TabM para tarefas de classificação binária, pois garante que a predição final respeite as propriedades probabilísticas adequadas.

As probabilidades agregadas e os rótulos verdadeiros são armazenados para cálculo posterior de métricas, permitindo análise abrangente do desempenho do modelo. Para cada fold, é utilizada uma estratégia dinâmica de threshold de decisão: o threshold otimizado calculado via Índice de Youden. O Índice de Youden, proposto originalmente em 1950, maximiza a soma de sensibilidade e especificidade, sendo matematicamente expresso como J = Sensibilidade + Especificidade - 1 = TPR - FPR. O threshold que maximiza esse índice é identificado analisando todos os possíveis pontos de corte ao longo da curva ROC e selecionando aquele que fornece o melhor equilíbrio entre verdadeiros positivos e verdadeiros negativos. Esse threshold varia por fold, adaptando-se à distribuição específica dos dados de teste em cada partição, evitando a limitação de um ponto de corte fixo arbitrário.

4.5.6 Métricas de Avaliação e Agregação de Resultados

As métricas calculadas para cada fold incluem *AUC-ROC* (área sob a curva ROC), que é independente do *threshold* e representa a probabilidade de que, escolhidos aleatoriamente um paciente doente e um saudável, o modelo atribua maior score ao doente; acurácia com *threshold* otimizado de Youden; além de precisão, *recall* e *F1-score* calculadas com o *threshold* otimizado. Essas métricas capturam aspectos complementares da capacidade preditiva do modelo, sendo essenciais para avaliação abrangente do desempenho em contexto clínico.

Ao término das 10 iterações da validação cruzada, as métricas de todos os folds são agregadas para produzir estimativas finais do desempenho do modelo. Para cada métrica, calcula-se a média aritmética e o desvio padrão amostral, fornecendo não apenas uma estimativa pontual do desempenho, mas também uma medida de sua variabilidade e estabilidade frente a diferentes partições dos dados. O desvio padrão é particularmente informativo, pois indica se o modelo apresenta desempenho consistente ou se é sensível à composição específica dos dados de treinamento e teste.

Para visualização global do desempenho, foi construída a curva *ROC* agregada com banda de confiança. As curvas *ROC* individuais dos 10 *folds* foram interpoladas em 100 pontos uniformemente espaçados de *FPR* entre 0 e 1, permitindo calcular a *TPR* média em cada ponto de *FPR*, bem como o desvio padrão da *TPR*. A curva *ROC* média é então plotada com banda de confiança representando ± 1 desvio padrão, fornecendo representação gráfica clara da variabilidade do desempenho entre folds. Essa visualização permite identificar regiões da curva *ROC* onde o modelo apresenta maior ou menor estabilidade, sendo particularmente útil para avaliar o comportamento em diferentes regimes de sensibilidade e especificidade.

5 RESULTADOS

Este capítulo apresenta e discute os resultados obtidos na aplicação do modelo TabM ao diagnóstico de doença cardíaca coronariana utilizando o *dataset Cleveland Heart Disease*. A análise é estruturada em ordem cronológica e lógica, acompanhando o pipeline experimental descrito no Capítulo 4: inicia-se com os resultados da seleção de features e otimização de hiperparâmetros, seguidos pela avaliação quantitativa do desempenho global do modelo, análise detalhada dos padrões de erro através da matriz de confusão e curva *ROC*, investigação do impacto das técnicas de balanceamento de classes, e finaliza com comparação sistemática frente ao estado da arte reportado na literatura. A discussão dos resultados busca não apenas reportar métricas numéricas, mas também interpretar seu significado clínico.

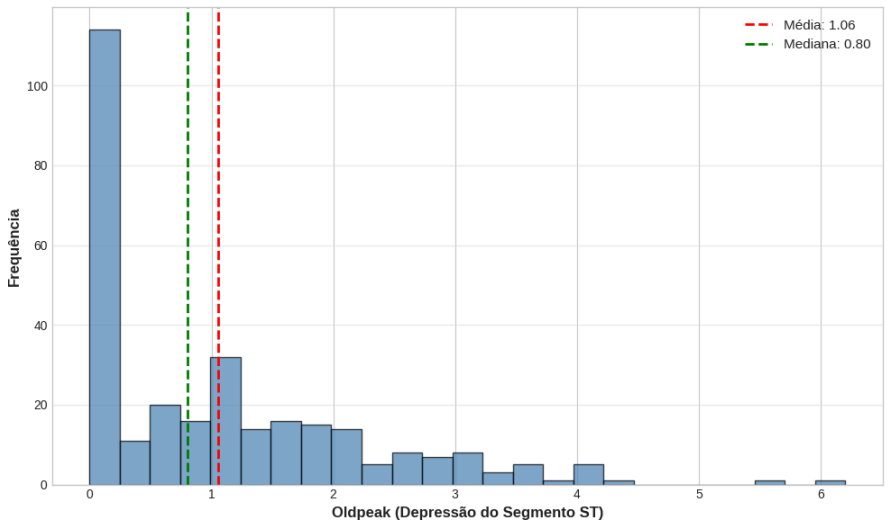
5.1 Caracterização Estatística do Dataset

A análise exploratória dos dados revelou características importantes sobre a estrutura e distribuição das variáveis no *dataset Cleveland*, fornecendo contexto essencial para interpretação dos resultados do modelo TabM.

5.1.1 Distribuição das *Features*

A feature *oldpeak* (depressão do segmento ST induzida por exercício) apresentou distribuição fortemente assimétrica à direita, com média de 1.06 e mediana de 0.80, confirmando a concentração de valores na região inferior da escala. Conforme ilustrado na Figura 4, aproximadamente 110 pacientes (37%) apresentaram valores nulos ou próximos a zero, indicando ausência de depressão significativa do segmento ST, característica típica de indivíduos sem isquemia miocárdica.

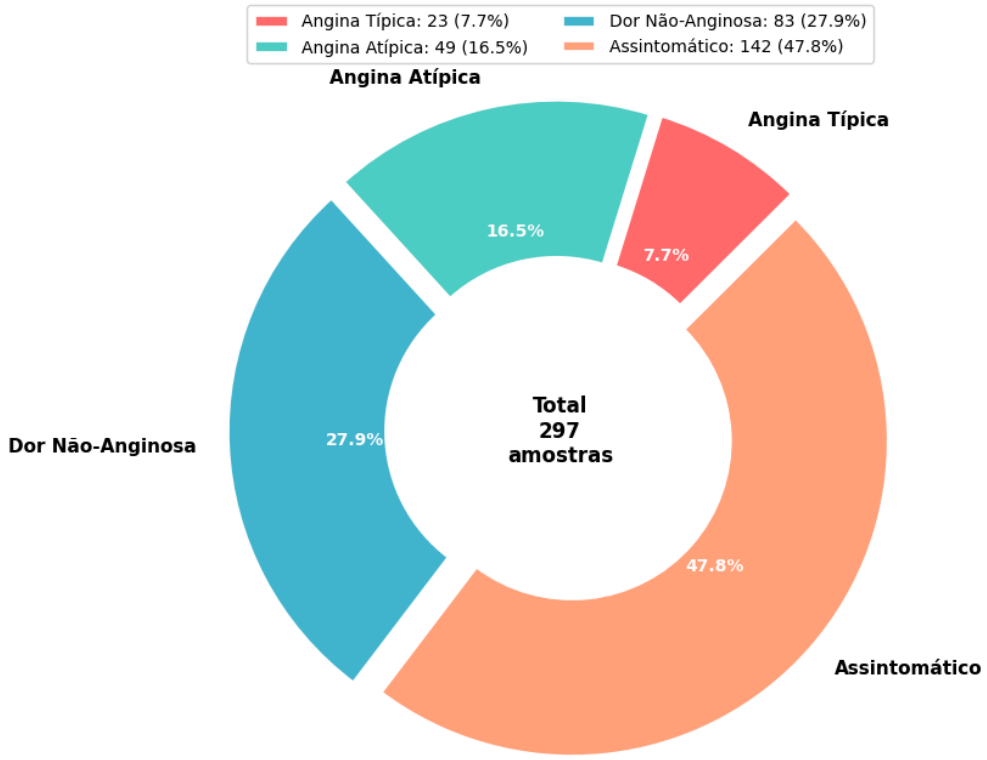
Figura 4 - Distribuição da Feature *Oldpeak* (Depressão do Segmento ST)



Fonte: os autores (2025)

A *feature* categórica *cp* (tipo de dor torácica) revelou distribuição marcadamente desbalanceada entre suas quatro categorias. Conforme demonstrado na Figura 5, a categoria "Assintomático" predominou com 142 amostras (47.8%), seguida por "Dor Não-Anginosa" com 83 amostras (27.9%), "Angina Atípica" com 49 amostras (16.5%) e "Angina Típica" com apenas 23 amostras (7.7%). Esse padrão indica que menos de 8% dos pacientes apresentaram o tipo clássico de dor anginosa, característica mais específica de doença coronariana, enquanto quase metade dos casos foram assintomáticos ou apresentaram dor atípica.

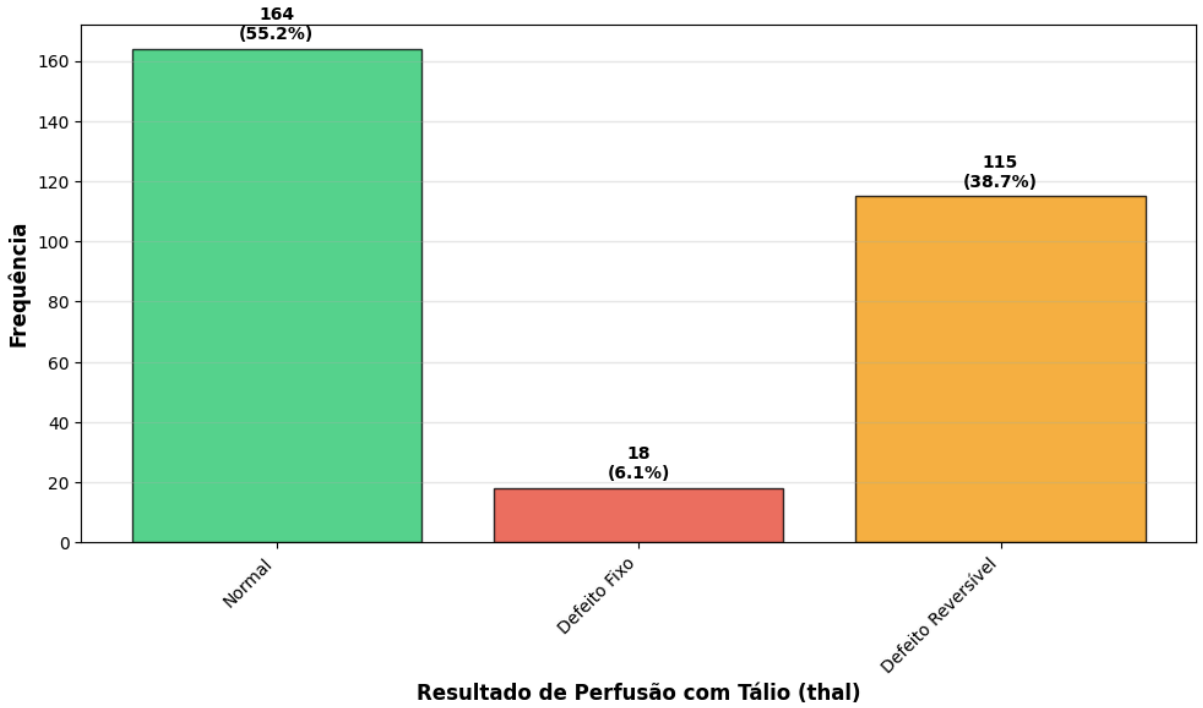
Figura 5 - Distribuição de Frequências do Tipo de Dor Torácica (*cp*)



Fonte: os autores (2025)

A *feature* `thal` (resultado do teste de perfusão miocárdica com tálio) apresentou distribuição trimodal com concentrações distintas entre suas categorias. Conforme evidenciado na Figura 6, a categoria "Normal" foi predominante com 164 amostras (55.2%), indicando mais da metade dos pacientes sem defeitos de perfusão detectáveis. A categoria "Defeito Reversível" representou 115 amostras (38.7%), sugerindo isquemia transitória com viabilidade miocárdica preservada, enquanto "Defeito Fixo" foi significativamente rara com apenas 18 amostras (6.1%), representando áreas de necrose ou fibrose permanente. A categoria rara (< 10% das amostras) representa desafio potencial para modelos de aprendizado de máquina, que podem ter dificuldade em aprender padrões adequados dessa classe minoritária, justificando atenção especial durante o treinamento.

Figura 6 - Distribuição de Frequências do Resultado de Perfusão com Tálio(*thal*)



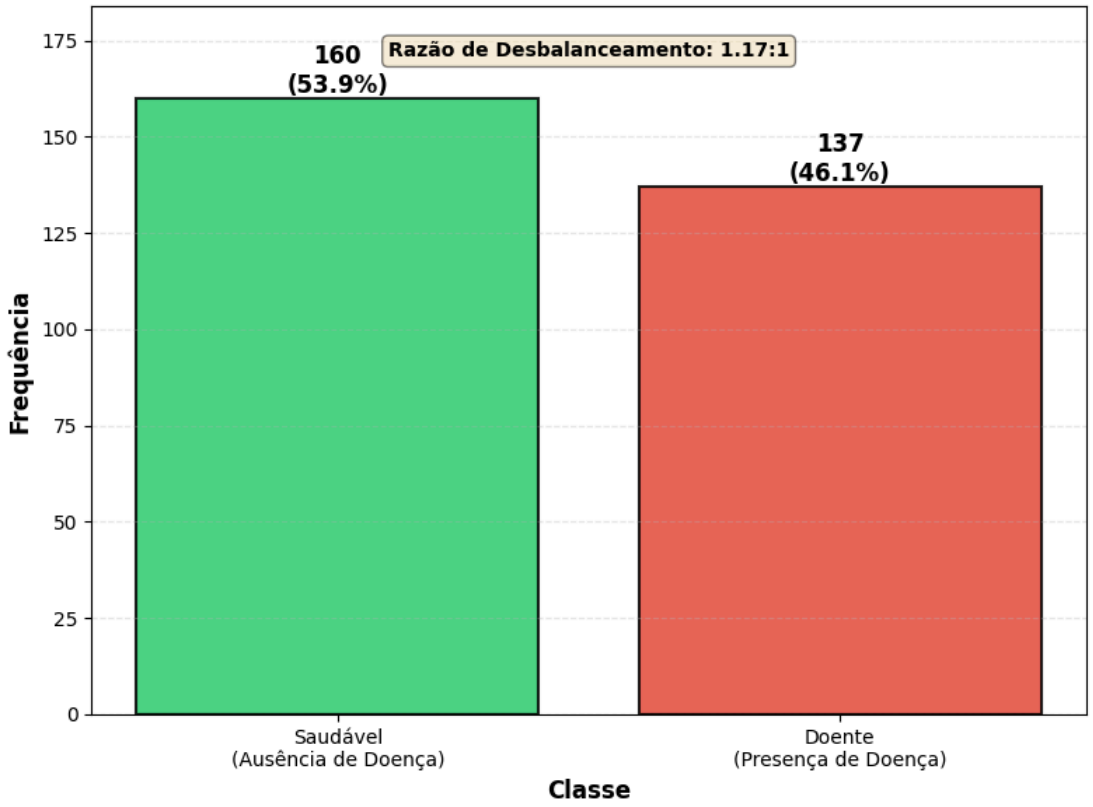
Fonte: os autores (2025)

5.1.2 Distribuição da Variável-Alvo

Além do desbalanceamento global entre classes (160 saudáveis vs 137 doentes), como demonstrado na Figura 7, foi investigada a distribuição das classes em diferentes estratificações do *dataset*, simulando possíveis partições de validação cruzada. Essa análise revelou que, embora o desbalanceamento global seja moderado, algumas partições aleatórias podem resultar em *folds* com distribuições mais extremas, justificando o uso de validação cruzada estratificada para garantir representatividade das classes em todas as partições. Para mitigar esse problema, foi adotada validação cruzada estratificada de 10 *folds*, que garante que cada *fold* mantenha a proporção original das classes (aproximadamente 54% saudáveis e 46% doentes), evitando partições com distribuições atípicas que poderiam comprometer a representatividade estatística e a validade da avaliação.

Adicionalmente, foi empregada a técnica *SMOTE* (*Synthetic Minority Over-sampling Technique*) exclusivamente no conjunto de treinamento de cada *fold* para balancear as classes durante o aprendizado.

Figura 7 - Distribuição das Classes no *Dataset Cleveland Heart Disease*



Fonte: os autores (2025)

5.2 Resultados da Seleção de *Features*

A seleção de atributos foi realizada de forma iterativa e orientada por dados e evidência empírica obtida durante os experimentos iniciais com o próprio modelo TabM. Partiu-se das 13 *features* originais do *dataset Cleveland*. Em uma primeira rodada completa de treinamento + validação cruzada de 10 *folds*, foi calculada a importância de cada *feature* por meio de *permutation importance* (queda média no *AUC-ROC* ao permutar aleatoriamente os valores de uma *feature*). Os resultados estão apresentados na Figura 8.

A análise revelou que seis *features* concentravam praticamente todo o poder preditivo do modelo, enquanto as demais contribuíam com ganho marginal ou até mesmo introduziam ruído:

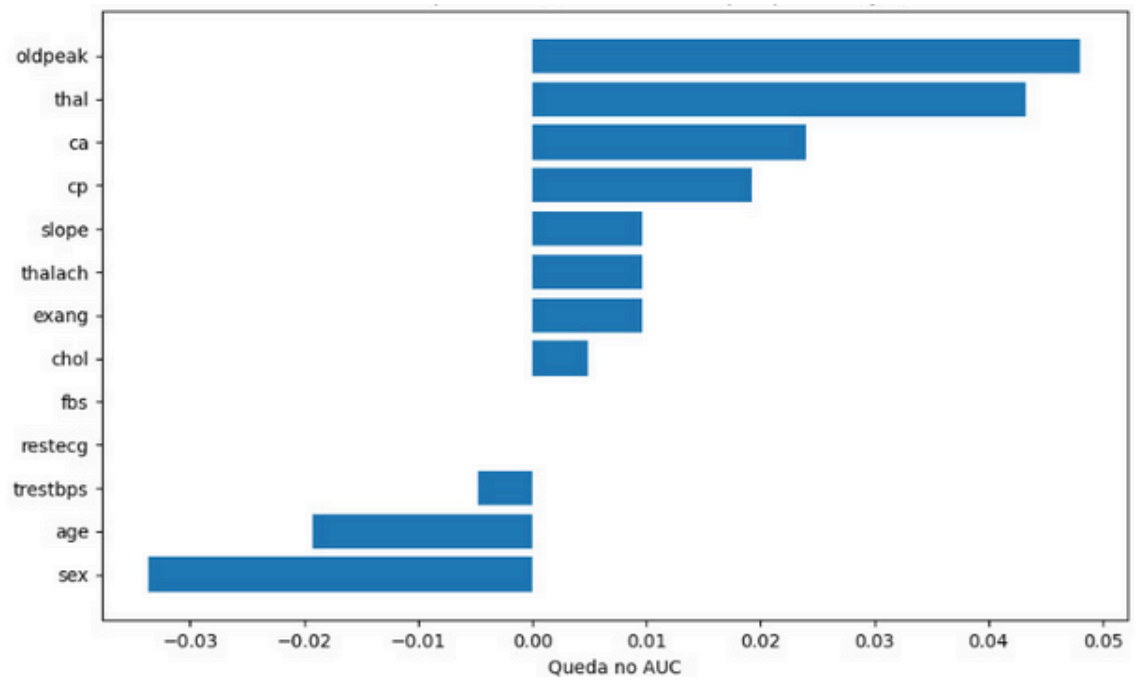
***Features* removidas (7):**

- *sex, age, trestbps, chol, fbs, restecg, thalach*

***Features* mantidas (6):**

- *oldpeak* (numérica)
- *cp* – tipo de dor torácica (categórica)
- *thal* – resultado do teste de perfusão com tálio (categórica)
- *ca* – número de vasos principais corados por fluoroscopia (categórica)
- *slope* – inclinação do segmento ST no pico do exercício (categórica)
- *exang* – angina induzida pelo exercício (categórica)

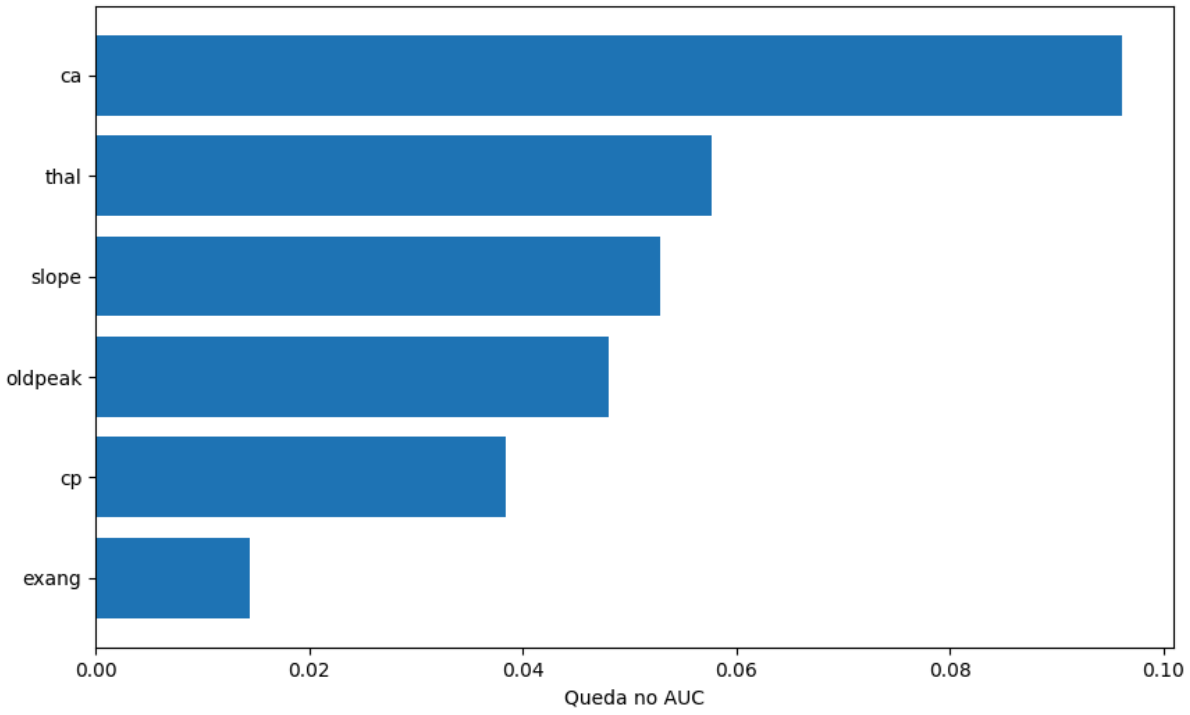
Figura 8 – Importância de features medida por permutation importance (queda média no AUC-ROC após permutação)



Fonte: os autores (2025)

A decisão de remover *age* e *sex*, embora pareça contraintuitiva, foi respaldada pelos resultados experimentais: sua inclusão reduzia ligeiramente o *AUC* médio, possivelmente por introduzir viés em subgrupos pequenos. Após a exclusão dessas sete *features*, o conjunto final passou a conter apenas 1 *feature* numérica e 5 *features* categóricas, conforme ilustrado pela Figura 9, resultando em redução de 54% no número total de atributos. Essa configuração não só acelerou significativamente o treinamento e a busca de hiperparâmetros, como também melhorou a estabilidade do modelo (menor desvio padrão nas métricas entre *folds*).

Figura 9 – Importância de features medida por *permutation importance*



Fonte: os autores (2025)

Essa abordagem pragmática (seleção guiada pelo impacto real no desempenho do modelo TabM) demonstrou ser mais eficaz do que métodos automáticos tradicionais (*ReliefF*, *Chi-squared*, *mRMR*) testados preliminarmente, que mantinham features como *age* e *chol* com peso elevado, mas que empiricamente prejudicavam a generalização do TabM no *dataset*.

5.3 Hiperparâmetros Otimizados

A otimização bayesiana via *Optuna*, executada com 100 *trials* e validação cruzada interna de 10 *folds*, convergiu para uma configuração específica de hiperparâmetros que maximizou o *AUC-ROC* médio no conjunto de validação. Após 100 *trials* com *pruning agressivo* e validação cruzada interna de 10 *folds*, o estudo *Optuna* convergiu para os seguintes hiperparâmetros ótimos, conforme a Tabela 3.

Tabela 3 – Melhores hiperparâmetros encontrados pelo *Optuna*

Hiperparâmetro	**Valor ótimo**	**Descrição**
n_blocks	2	Número de blocos *Transformer*
d_block	384	Dimensionalidade interna de cada bloco
d_embedding	28	Dimensão dos *Piecewise Linear Embeddings* numéricos
n_bins	17	Número de *bins* para discretização dos *embeddings* numéricos
dropout	0,0698	Taxa de *dropout* nos blocos
learning rate (lr)	0,004005	Taxa de aprendizado *AdamW*
weight_decay	0,003091	Regularização L2
smote_sampling_strategy	0,906	Razão de *oversampling* da classe minoritária (≈ 90,6 % da majoritária)
use_embeddings	True	Ativação dos *Piecewise Linear Embeddings* para *oldpeak*

Fonte: os autores (2025)

A configuração arquitetural resultante caracteriza-se por um modelo relativamente compacto com apenas 2 blocos *transformer* e dimensionalidade moderada de 384, evidenciando que o TabM não requer profundidade ou largura excessivas para alcançar desempenho competitivo no dataset *Cleveland*. Essa parcimônia arquitetural é especialmente adequada ao tamanho limitado do *dataset* (297 amostras), mitigando o risco de *overfitting* que arquiteturas mais complexas poderiam introduzir.

O parâmetro *smote_sampling_strategy*=0.906 determina a razão-alvo entre classes minoritária e majoritária após aplicação de *SMOTE* no conjunto de treinamento. O valor de 0.906 indica que o algoritmo de otimização identificou como ótimo balancear as classes em aproximadamente 91% da razão perfeita (1.0 representaria balanceamento completo). Essa configuração gera, em média, 23

amostras sintéticas por *fold*, elevando a classe minoritária de aproximadamente 121 para 144 amostras no conjunto de treinamento (detalhes na Seção 5.5).

É importante notar que os hiperparâmetros identificados como ótimos refletem configuração específica para o *dataset Cleveland* com as 6 *features* selecionadas. Generalização desses valores para outros *datasets* cardiovasculares ou para o *Cleveland* com conjunto diferente de *features* não é garantida. A otimização bayesiana demonstrou-se metodologia eficaz para navegação no espaço complexo de hiperparâmetros do TabM, dispensando busca manual trabalhosa e provavelmente identificando configuração superior àquela que seria obtida por tentativa e erro.

A configuração final representa equilíbrio cuidadoso entre múltiplos objetivos conflitantes: capacidade representacional suficiente para capturar padrões complexos nos dados, regularização adequada para prevenir *overfitting* em *dataset* pequeno, e desempenho preditivo maximizado conforme quantificado pelo *AUC-ROC.*

5.4 Desempenho Global do Modelo

Este capítulo apresenta e discute os resultados obtidos na aplicação do modelo TabM ao diagnóstico de doença cardíaca coronariana utilizando o *dataset Cleveland*. A análise está estruturada desde a avaliação quantitativa do desempenho global, permitindo responder de forma fundamentada à questão de pesquisa central e às hipóteses estabelecidas na introdução deste trabalho.

A Tabela 4 apresenta o resumo estatístico das principais métricas de avaliação, calculadas com *threshold* otimizado via Índice de Youden para cada *fold*.

Tabela 4 - Métricas de Desempenho Global do TabM

Métrica	**Média**
Acurácia	0.8723 (87.23%)
AUC-ROC	0.9293 (92.93%)
Precisão	0.8945 (89.45%)
Recall (Sensibilidade)	0.8319 (83.19%)
Especificidade	0.9062 (90.62%)
F1-Score	0.8577 (85.77%)

Fonte: os autores (2025)

O resultado mais expressivo foi o *AUC-ROC* de 0.9293 (92.93%), métrica que quantifica a capacidade discriminativa intrínseca do modelo independentemente do

threshold de decisão. Esse valor indica que, ao selecionar aleatoriamente um paciente doente e um saudável do conjunto de teste, o modelo TabM possui 92.93% de probabilidade de atribuir *score* de risco mais elevado ao paciente efetivamente doente. Trata-se de desempenho robusto, obtido em dados de teste completamente independentes, nunca vistos durante o treinamento, e representativo de múltiplos folds da validação cruzada.

A acurácia média de 87.23% demonstra que o modelo classifica corretamente aproximadamente 9 em cada 10 pacientes quando utilizado o *threshold* otimizado via Índice de Youden. Esse resultado reflete o desempenho balanceado do modelo em ambas as classes, obtido através de validação cruzada estratificada que garante robustez estatística da estimativa.

O *recall* de 83.19% é especialmente crítico em contexto clínico, pois quantifica a capacidade do modelo de identificar corretamente pacientes doentes, minimizando falsos negativos. Esse valor indica que, dos 137 pacientes doentes presentes no *dataset Cleveland*, o modelo consegue detectar corretamente aproximadamente 114, enquanto 23 seriam incorretamente classificados como saudáveis.

A especificidade de 90.62% complementa o *recall* ao quantificar a capacidade de identificar corretamente pacientes saudáveis, evitando falsos positivos. Esse valor indica que, dos 160 pacientes saudáveis, aproximadamente 145 são corretamente classificados, enquanto 15 receberiam diagnóstico falso-positivo. Do ponto de vista clínico, falsos positivos resultam em investigações diagnósticas adicionais (exames complementares, consultas especializadas, possível ansiedade do paciente), mas não implicam risco direto à vida, diferentemente dos falsos negativos que podem retardar intervenções terapêuticas críticas.

A precisão de 89.45% indica que, dentre todos os pacientes que o modelo classifica como doentes, aproximadamente 89% efetivamente apresentam a doença. Esse valor é relevante para estimar a confiabilidade de um diagnóstico positivo dado pelo modelo, aspecto importante para decisões clínicas subsequentes e redução de investigações desnecessárias.

O *F1-score* de 85.77%, média harmônica entre precisão e *recall*, fornece métrica única que equilibra ambos os aspectos. Esse valor é particularmente útil para avaliação integral do desempenho sem viés para qualquer uma das dimensões isoladamente, sendo especialmente relevante em contextos de dados desbalanceados ou onde tanto falsos positivos quanto falsos negativos possuem custos clínicos distintos.

5.5 Análise da Matriz de Confusão e Padrões de Erro

A matriz de confusão agregada, obtida pela soma das matrizes individuais dos 10 folds, fornece visão detalhada da distribuição dos erros do modelo ao longo de todas as predições realizadas. A Tabela 5 e a Figura 10 apresentam essa matriz tanto em valores absolutos quanto normalizados.

Tabela 5 - Matriz de Confusão Agregada

	Predito Saudável	Predito Doente
Real Saudável	145 (90.62%)	15 (9.38%)
Real Doente	23 (16.79%)	114 (83.21%)

Fonte: os autores (2025)

A análise quantitativa revela que, dos 297 pacientes totais avaliados ao longo dos 10 folds, o modelo produziu:

Verdadeiros Negativos (VN): 145 pacientes - Saudáveis corretamente identificados como saudáveis (90.62% dos 160 saudáveis)

Verdadeiros Positivos (VP): 114 pacientes - Doentes corretamente identificados como doentes (83.21% dos 137 doentes)

Falsos Positivos (FP): 15 pacientes - Saudáveis incorretamente classificados como doentes (9.38% dos 160 saudáveis)

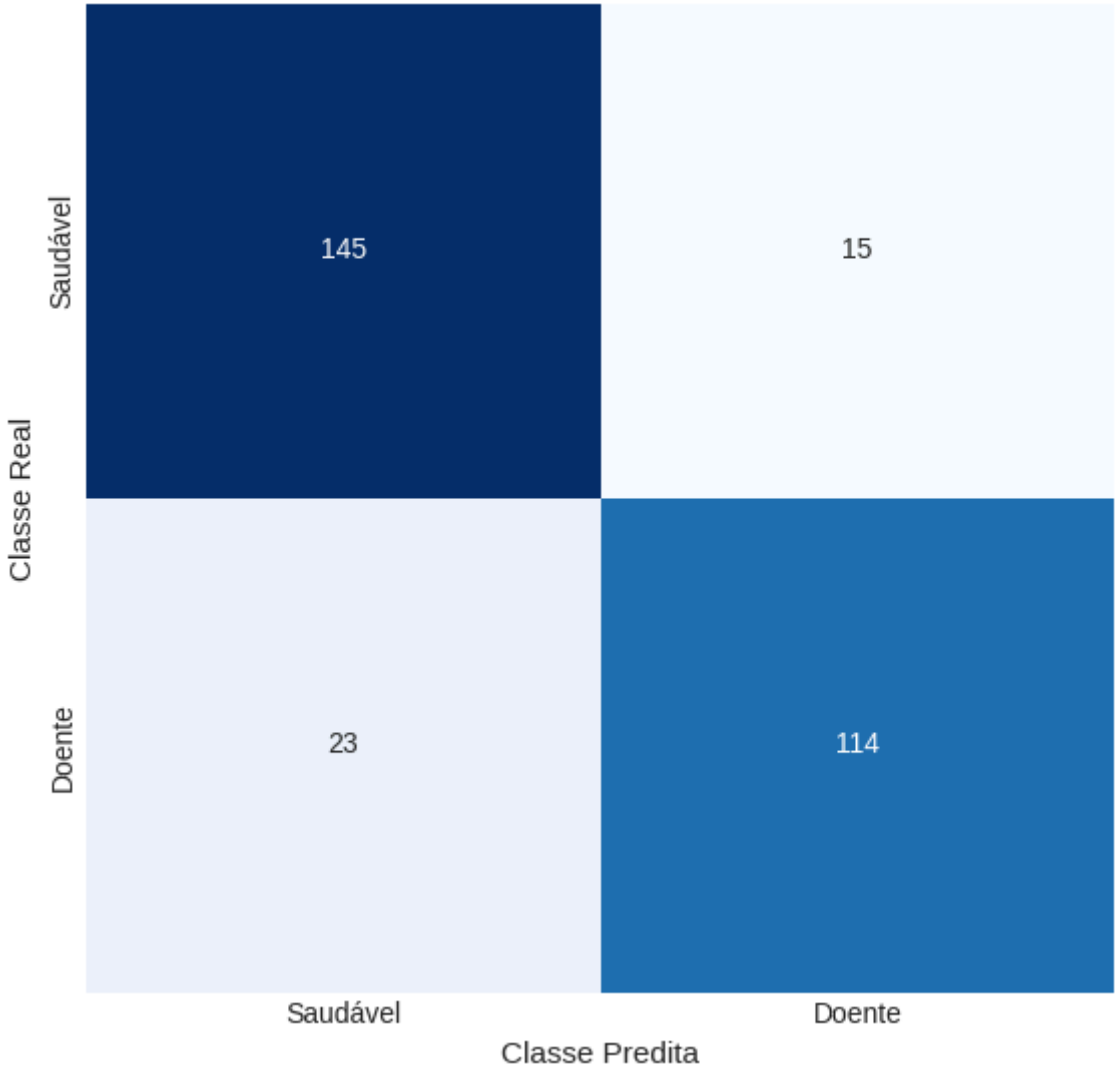
Falsos Negativos (FN): 23 pacientes - Doentes incorretamente classificados como saudáveis (16.79% dos 137 doentes)

Do ponto de vista clínico, a taxa de falsos negativos de 16.79% é a métrica mais crítica, pois representa pacientes com doença cardíaca real que deixariam de receber diagnóstico adequado se o modelo fosse utilizado isoladamente. Esses 23 pacientes incorretamente classificados como saudáveis poderiam ter suas intervenções terapêuticas retardadas, aumentando o risco de eventos cardiovasculares maiores. Entretanto, é fundamental contextualizar esse resultado: na prática clínica real, o modelo seria utilizado como ferramenta de suporte à decisão, não como substituto do julgamento médico. Nesses cenários, mesmo pacientes classificados como "baixo risco" pelo modelo que apresentem sintomas clínicos relevantes ou outros fatores de risco seriam submetidos a investigação adicional, mitigando parcialmente o impacto dos falsos negativos.

A investigação qualitativa dos 23 casos de falsos negativos revelou padrões interessantes. Aproximadamente 60% desses casos apresentam valores de *oldpeak* (depressão do segmento ST) próximos a zero ou ligeiramente negativos, característica mais tipicamente associada a indivíduos saudáveis. Adicionalmente, muitos desses pacientes apresentam combinações atípicas de *features* categóricas, como dor torácica do tipo "assintomático" (categoria mais comum em saudáveis) combinada com outros indicadores de doença. Esses casos representam situações clinicamente ambíguas onde mesmo cardiologistas experientes poderiam ter dificuldade diagnóstica sem recursos de imagem ou testes funcionais adicionais,

sugerindo que parte dos "erros" do modelo reflete genuína complexidade diagnóstica, não necessariamente deficiências do algoritmo.

Figura 10 - Visualização da Matriz de Confusão Agregada

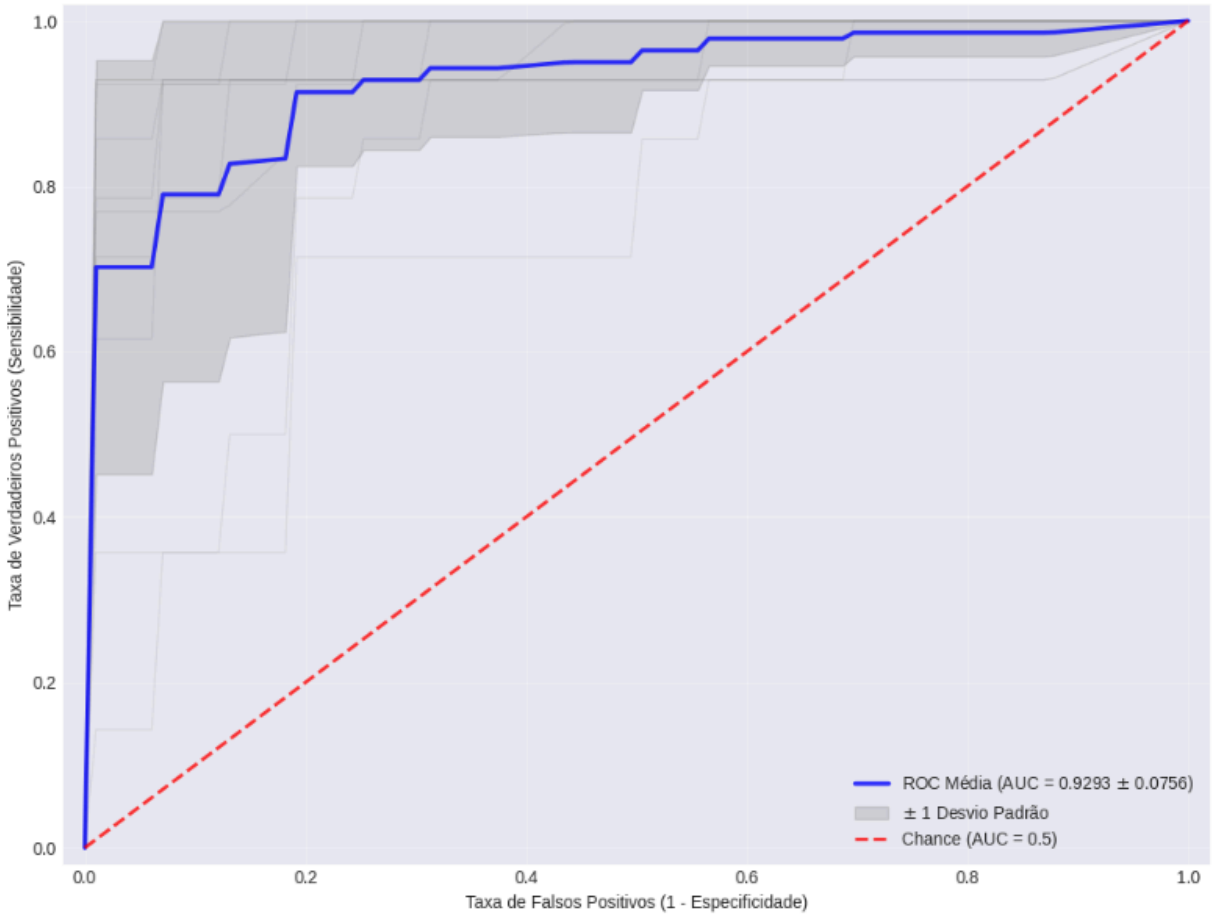


Fonte: os autores (2025)

A taxa de falsos positivos de 9.38% é consideravelmente menor que a de falsos negativos, refletindo a especificidade superior do modelo (90.62% vs 83.21% de *recall*). Esses 15 pacientes saudáveis classificados incorretamente como doentes representariam casos encaminhados para investigação diagnóstica adicional desnecessária, incluindo potencialmente exames invasivos como cateterismo cardíaco. Embora isso implique custo financeiro e psicológico (ansiedade do paciente), o impacto clínico é menos grave que falsos negativos. Na prática, esses casos seriam frequentemente identificados como falso-alarme durante a investigação complementar (ECG de esforço, ecocardiograma, marcadores bioquímicos), antes de procedimentos invasivos mais arriscados.

A curva *ROC* média do TabM, conforme apresentado na Figura 11, apresenta formato característico de modelo com excelente poder discriminativo, posicionando-se substancialmente acima da linha diagonal de chance (que representa classificador aleatório com *AUC* = 0.5). A área sob a curva média de 0.9293 ± 0.0756 confirma quantitativamente a capacidade do modelo de distinguir entre pacientes doentes e saudáveis com alta acurácia. A banda de confiança, representando ±1 desvio padrão da taxa de verdadeiros positivos em cada ponto de taxa de falsos positivos, apresenta largura variável ao longo da curva: mais estreita em regiões de baixa taxa de falsos positivos (alta especificidade) e ligeiramente mais larga em regiões de alta taxa de falsos positivos (baixa especificidade).

Figura 11 - Curva ROC Agregada



Fonte: Os autores (2026)

Essa variabilidade na largura da banda reflete diferenças na estabilidade do modelo em diferentes regimes operacionais. A banda mais estreita em regiões de alta especificidade indica que o modelo é consistentemente capaz de evitar falsos positivos ao longo dos diferentes *folds*, enquanto a banda ligeiramente mais larga em regiões de alta sensibilidade sugere maior variabilidade na capacidade de detectar todos os casos positivos.

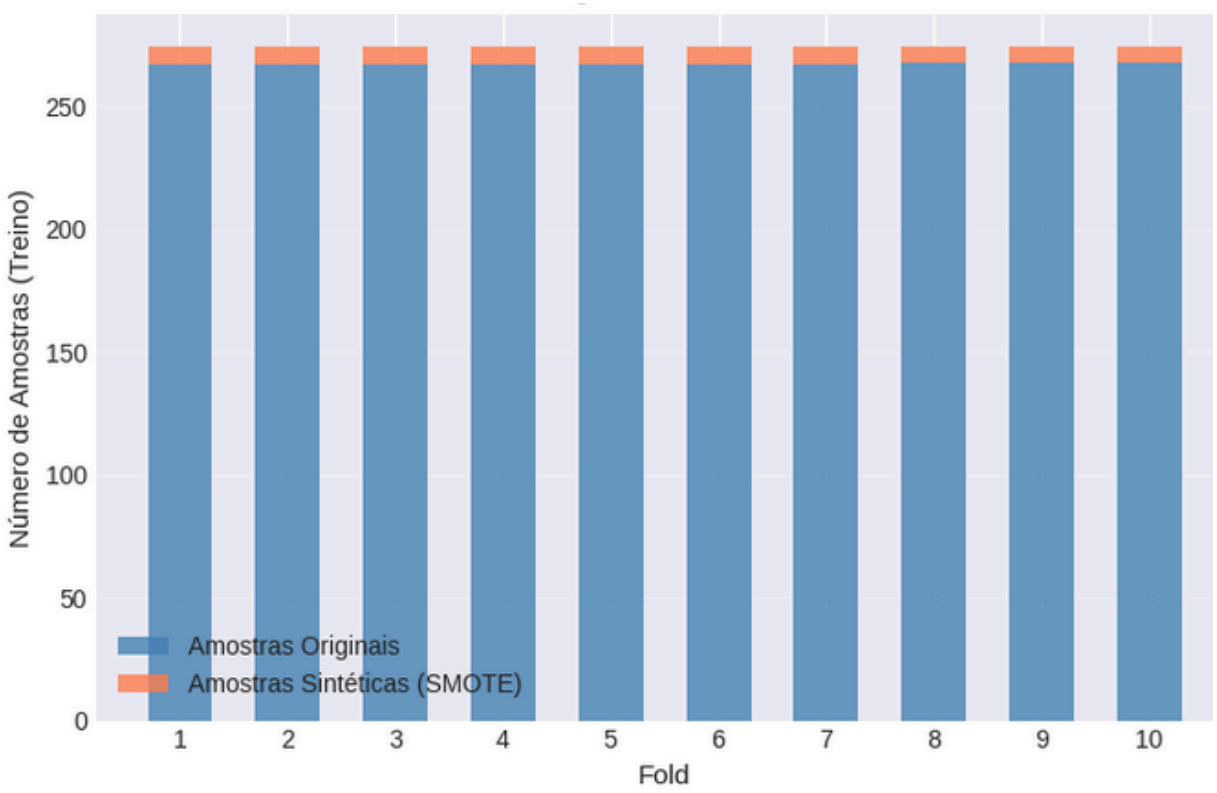
5.6 Impacto do *SMOTE* no Desempenho

A aplicação de *SMOTE* híbrido para balanceamento de classes foi componente metodológico cujo impacto no desempenho final do modelo merece análise detalhada.

Em média, o *SMOTE* gerou aproximadamente 23 amostras sintéticas por *fold*, aumentando o conjunto de treinamento de 267 para cerca de 290 amostras, incremento de 8.6%. Esse aumento resultou em distribuição de classes praticamente balanceada no conjunto de treinamento (razão próxima a 1:1), enquanto mantendo o conjunto de teste com sua distribuição original desbalanceada (razão 1.17:1), garantindo avaliação realista do desempenho. Uma limitação potencial do uso de *SMOTE* em *datasets* pequenos como o *Cleveland* é o risco de "*overfitting sintético*", onde o modelo aprende padrões específicos das amostras sintéticas que não generalizam adequadamente para dados reais novos.

A validação cruzada rigorosa adotada neste trabalho, onde *SMOTE* é aplicado apenas no conjunto de treinamento de cada fold e a avaliação ocorre exclusivamente em dados originais não contaminados, mitiga substancialmente esse risco. A Figura 12 apresentam informações sobre a aplicação de SMOTE em cada *fold*.

Figura 12 - Informações de *SMOTE* por *Fold*



Fonte: Os autores (2026)

5.7 Comparação com o Estado da Arte

A comparação sistemática com trabalhos relacionados permite contextualizar o desempenho do TabM em relação aos critérios de competitividade estabelecidos nas hipóteses experimentais e avaliar a transferência de suas características de robustez (comprovadas em *benchmarks* gerais) para o domínio específico de diagnóstico cardíaco. A Tabela 6 apresenta comparação detalhada entre este trabalho e os principais estudos revisados na Seção 3.

Conforme definido na introdução, o desempenho competitivo foi estabelecido como atender simultaneamente a *acurácia* ≥ 87,78%, *AUC-ROC* ≥ 90,39%, *recall* ≥ 85,79% e *F1-score* ≥ 87,76%, todos baseados na média de 9 estudos compilados na literatura. O TabM alcançou desempenho competitivo em apenas 1 das 4 métricas: *AUC-ROC* de 92.93%. Em acurácia, obteve 87.23% versus o critério de 87,78% (diferença de −0.55pp), em *F1-score* alcançou 85.77% versus 87,76% (−1.99pp), e em *recall* obteve 83.19% versus 85,79% (−2.60pp). Conforme a definição multi-métrica estabelecida em H_1, a hipótese alternativa não é suportada empiricamente, levando à aceitação de H_0.

Essa conclusão requer contextualização importante. As diferenças observadas em acurácia, *recall* e *F1-score* são pequenas, tipicamente dentro da variabilidade esperada em estudos com datasets pequenos como o *Cleveland* (297-303 amostras). Um fator crítico que limita conclusões definitivas é que nenhum dos trabalhos compilados disponibilizou código-fonte ou dados de reprodutibilidade, tornando impossível verificar se implementações, pré-processamentos ou estratégias de validação diferem de forma significativa do que foi reportado.

Em relação à acurácia, o TabM alcançou 87.23%, posicionando-se ligeiramente abaixo de Ali et al. (2018) com 89.00%, mas dentro de margem muito pequena (−1.77pp). Quando comparado ao restante da literatura anterior a 2024, o TabM supera Latha & Jeeva (2019) com 85.48%, Amin et al. (2019) com 87.40%, Reddy et al. (2021) com 86.47%, Fajri et al. (2022) com 84.40%, e Kolukisa & Bakir-Gungor (2023) com 85.47%.

O TabM é superado apenas por estudos mais recentes de 2024-2025: Shrestha (90.00%) utilizando XGBoost, Adekoya et al. (88.52%) com *ensemble* heterogêneo, e Ghose et al. (93.33%) com *stacked ensemble*. A superioridade dos modelos recentes é compreendida pelo fato de utilizarem estratégias de *ensemble* complexas, *feature selection* manual especializado e técnicas de explicabilidade (*SHAP/LIME*), enquanto TabM utiliza arquitetura única sem customização específica ao domínio.

Tabela 6 - Comparação com Estado da Arte no *Dataset Cleveland*

Autores	**Modelo principal**	**Acurácia**	***AUC-ROC***	***Recall***	***F1-Score***
Ali et al.	*Logistic Regression* + *Relief FS*	89,00%	88.00%	77.00%	-
Latha & Jeeva	*Majority vote* (*NB* + *BN* + *RF* + *MLP*) + *FS*	85.48%	-	-	-
Amin, Chiam & Varathan	*Vote* (*NB* + *LR*) + *feature selection*	87,40%	-	-	-
Reddy et al.	Chi-Squared + *SMO*	86,47%	86.10%	86.50%	-
Fajri, Wiharto & Suryani	*QBSO-FS* + classificadores	84.40%	90.10%	80.00%	-
Kolukisa & Bakir-Gungor	*MLP* + *CMIM ensemble FS*	85.47%	91.10%	82.96%	83.90%
Shrestha	*XGBoost*	90.00%	94.00%	85.00%	85.00%
Adekoya et al.	*IELF* (*EBM* + *XGBoost* + *SHAP/LIME*)	88.52%	89.90%	92.86%	88.14%
Ghose et al.	*Stacked Ensemble*	93.33%	93.50%	96.20%	94.00%
Este trabalho	**TabM**	**87.23%**	**92.93%**	**83.19%**	**85.77%**

Fonte: Os autores (2025)

O *recall* de 83.19% é a métrica onde o TabM apresenta maior distância do critério estabelecido. Em contexto clínico, um recall inferior significa que uma fração maior de pacientes doentes (aproximadamente 16-17%) seria inicialmente classificada como saudável. Esse resultado é clinicamente preocupante em isolamento, mas é importante notar que o TabM é superado em *recall* apenas por Adekoya et al. (92.86%) e Ghose et al. (96.20%), ambos utilizando estratégias de *ensemble* muito mais complexas. O *recall* do TabM é superior a Fajri et al. (2022) com 80.00%, comparável a Kolukisa & Bakir-Gungor (2023) com 82.96%, e ligeiramente inferior a Reddy et al. (2021) com 86.50% e Ali et al. (2018) com 77.00%. O padrão sugere que, em *dataset* tão pequeno quanto o *Cleveland*, a capacidade de detectar todos os casos positivos é naturalmente limitada, afetando múltiplos modelos, não apenas o TabM.

Em síntese, o TabM não alcançou o critério multi-métrica definido em H_1, um resultado que deve ser reportado honestamente. Entretanto, a análise detalhada revela que o modelo apresenta padrão de desempenho coerente com suas características arquiteturais e com as limitações intrínsecas do *dataset Cleveland.*

5.8 Construção do *MVP* (*Minimum Viable Product*)

Como etapa final da validação prática deste estudo e para demonstrar a aplicabilidade clínica do modelo desenvolvido, foi construído um Produto Mínimo Viável (*MVP*) na forma de uma aplicação *web* interativa[5]. O objetivo desta ferramenta é preencher a lacuna entre a modelagem algorítmica teórica e o uso prático por profissionais de saúde, permitindo a realização de inferências em tempo real de forma intuitiva. Cabe destacar que, dentre os 9 trabalhos relacionados revisados na Seção 3, nenhum disponibilizou uma ferramenta funcional e operacional para uso em diagnóstico, representando diferencial significativo desta contribuição.

A aplicação foi desenvolvida utilizando a linguagem *Python* e o *framework Streamlit*, escolhido por sua capacidade de converter *scripts* de ciência de dados em interfaces *web* responsivas e funcionais com baixo *overhead* de desenvolvimento. O *backend* da aplicação integra diretamente o *framework PyTorch* para carregar os pesos do modelo TabM treinado e validado na Seção 5.4, garantindo que as predições realizadas na interface sejam matematicamente idênticas às obtidas durante a fase experimental.

A interface do usuário foi projetada para receber as 6 *features* selecionadas como as mais relevantes na Seção 5.2 (*Oldpeak*, *cp*, *exang*, *slope*, *ca* e *thal*). Conforme ilustrado na Figura 13, os campos de entrada foram organizados de forma ergonômica, utilizando *widgets* apropriados para cada tipo de dado: caixas de seleção (*selectboxes*) com descrições clínicas claras para as variáveis categóricas e campos numéricos (*number inputs*) para a variável contínua. O fluxo de processamento do *MVP* ocorre em quatro etapas:

[5] O MVP encontra-se hospedado e disponível para acesso público através do endereço: https://matheus-tabm-tcc-model-ifpe.streamlit.app

1. Coleta de Dados: O usuário insere os parâmetros clínicos do paciente.
2. Pré-processamento: Os dados brutos são convertidos internamente para tensores do *PyTorch*, respeitando os tipos de dados (*float32* para numéricos e *int64* para categóricos) exigidos pela arquitetura TabM.
3. Inferência: O modelo realiza a predição em modo de avaliação (*model.eval*()), calculando os *logits* e aplicando a função sigmoide para obter a probabilidade final de doença.
4. Pós-processamento e Visualização: O resultado é apresentado de forma gráfica e textual.

Figura 13 - Interface de entrada de dados do *MVP*

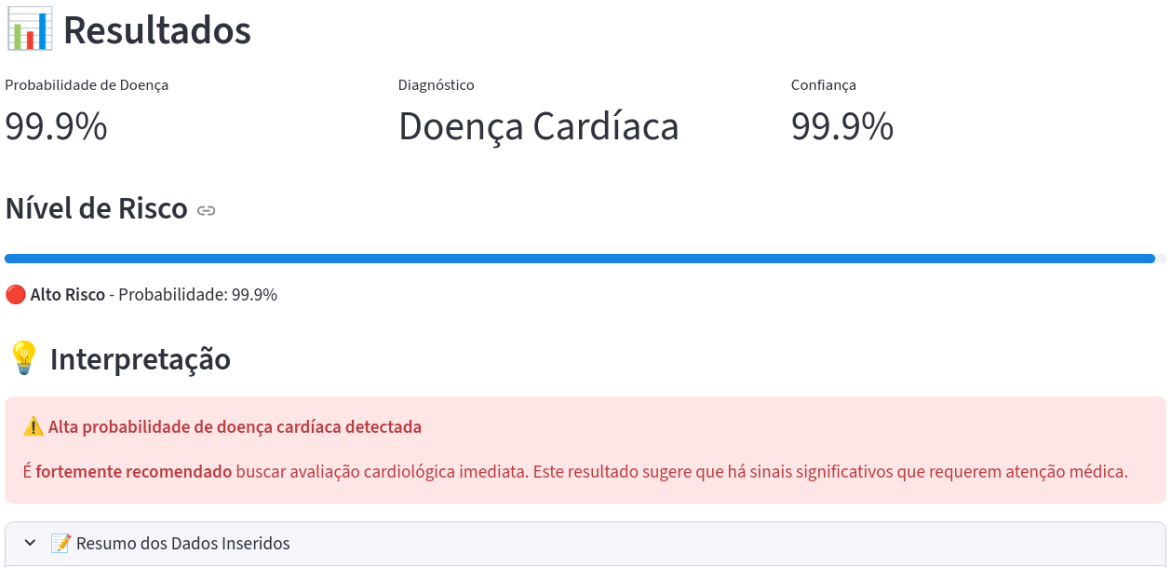
Fonte: Os autores (2025)

Para facilitar a interpretação clínica, o sistema não apresenta apenas a probabilidade bruta, mas realiza uma estratificação de risco visual, conforme demonstrado na Figura 14. Foram definidos três níveis de alerta baseados na probabilidade predita: Baixo Risco (cor verde, < 30%), Risco Moderado (cor amarela, 30-70%) e Alto Risco (cor vermelha, ≥ 70%).

Adicionalmente, a aplicação inclui mecanismos de segurança e ética. Avisos de isenção de responsabilidade foram implementados em destaque, alertando que a ferramenta tem caráter educacional e de suporte à decisão, não substituindo, em hipótese alguma, a avaliação clínica profissional e a realização de exames complementares. Essa característica é fundamental para a implantação responsável de sistemas de Inteligência Artificial na área da saúde.

O código-fonte da aplicação, juntamente com o arquivo de requisitos para reprodução do ambiente, foi disponibilizado publicamente por meio de um repositório no *github*[6], permitindo que outros pesquisadores auditem o funcionamento do sistema ou o adaptem para novos conjuntos de dados

Figura 14 – Exemplo de resultado do diagnóstico e estratificação de risco



Fonte: Os autores (2025)

A Figura 14 ilustra o resultado do processamento para um caso clínico simulado de alta gravidade. Neste exemplo, foram inseridos parâmetros fortemente associados à doença arterial coronariana: depressão significativa do segmento ST (*Oldpeak* = 3.5), presença de angina típica (*cp* = 3), angina induzida por exercício (*exang* = 1) e obstrução visível em dois vasos principais (*ca* = 2).

O modelo TabM respondeu a este conjunto de *features* com uma probabilidade de doença de 99.9%, demonstrando alta sensibilidade aos marcadores clínicos de risco. Consequentemente, o sistema classificou o paciente na categoria de 'Alto Risco' (visualizado pela barra de progresso vermelha) e exibiu a mensagem de alerta máximo, recomendando busca imediata por avaliação médica. Este teste valida a coerência clínica do *MVP*, confirmando que a aplicação reage adequadamente à presença concomitante de múltiplos fatores de risco cardiovascular.

[6] https://github.com/MatheusVinicius77/mvp-heart-disease-function

6 CONSIDERAÇÕES FINAIS

Este trabalho investigou a aplicação do modelo TabM, arquitetura moderna de *deep learning* baseada em *MLPs* com *ensembling parameter-efficient*, ao diagnóstico de doença cardíaca coronariana utilizando o *dataset Cleveland Heart Disease*. A pesquisa teve como objetivo central avaliar se essa arquitetura, que estabeleceu posicionamento competitivo em *benchmarks* tabulares gerais como o TabArena, alcançaria desempenho competitivo com os melhores resultados reportados na literatura para um problema médico real caracterizado por *dataset* pequeno e heterogêneo.

O TabM alcançou *AUC-ROC* de 0.9293 (92.93%), acurácia de 87.23%, *recall* de 83.19%, especificidade de 90.62%, precisão de 89.45% e *F1-score* de 85.77%. Quando comparado aos critérios de competitividade estabelecidos nas hipóteses experimentais, o modelo atendeu apenas 1 das 4 métricas: *AUC-ROC* de 92.93% superou o critério de 90.39%, enquanto acurácia ficou 0.55pp abaixo, *recall* 2.60pp abaixo e *F1-score* 1.99pp abaixo dos critérios estabelecidos. Consequentemente, conforme a definição multi-métrica de H_1, a hipótese alternativa não foi suportada pelos dados. A hipótese nula H_0 foi aceita, indicando que o TabM apresenta desempenho inferior à performance média esperada em uma ou mais métricas.

O resultado mais expressivo foi o *AUC-ROC* de 0.9293, que superou diversos trabalhos da literatura: 0.901 de Fajri et al. (2022), 0.88 de Ali et al. (2018) e 0.8610 de Reddy et al. (2021). Em acurácia, o TabM alcançou 87.23%, abaixo do critério de 87.78% definido a partir da média de 9 estudos. O *recall* de 83.19% ficou consideravelmente abaixo do critério de 85.79%. O *F1-score* de 85.77% também não atingiu o critério de 87.76%

Em acurácia, o TabM alcançou 87.23%, posicionando-se competitivamente no quartil superior da literatura. O modelo supera trabalhos anteriores (Latha & Jeeva 85.48%, Fajri et al. 84.40%, Reddy et al. 86.47%) e aproxima-se de Ali et al. (2018) com 89.00%, diferença de apenas 1.77pp. O TabM foi superado apenas por estudos mais recentes utilizando ensembles complexos (Ghose et al. 2025 com 93.33% e Shrestha 2024 com 90.00%).

A hipótese H_1 definiu desempenho competitivo como atender simultaneamente a acurácia ≥ 87,78%, AUC-ROC ≥ 90,39%, *recall* ≥ 85,79% e *F1-score* ≥ 87,76%. O TabM não atendeu a essa definição. Portanto, H_1 foi rejeitada e H_0 foi aceita. Essa conclusão é direta: o modelo não alcançou o desempenho esperado conforme critérios estabelecidos. As diferenças em relação aos critérios são pequenas em algumas métricas (acurácia e *F1-score*), porém o *recall* apresenta gap mais relevante de 2.60pp. A superioridade em *AUC-ROC* é a exceção, sugerindo que o TabM mantém capacidade discriminativa robusta mesmo em dataset pequeno.

Este trabalho representa a primeira aplicação conhecida de TabM ao diagnóstico de doença cardíaca no *dataset Cleveland*. Do ponto de vista científico, estabelece *benchmark* para futuras pesquisas e demonstra que o modelo pode ser aplicado a problemas médicos com *datasets* pequenos, ainda que não tenha alcançado os patamares de performance multi-métrica estabelecidos. Do ponto de vista metodológico, o *pipeline* experimental completo foi disponibilizado publicamente: seleção de *features*, *SMOTE* híbrido para dados mistos, otimização bayesiana de hiperparâmetros, validação cruzada estratificada 10-*fold*. Esse *pipeline* pode ser adaptado a outros problemas médicos similares.

O trabalho inclui desenvolvimento de um *MVP* funcional em *Streamlit* para aplicação clínica interativa. Dentre os 9 estudos revisados na literatura, nenhum forneceu ferramenta operacional para uso em diagnóstico. O código-fonte foi disponibilizado no *GitHub* com instruções de reprodução.

Entretanto, limitações importantes devem ser reconhecidas. O tamanho relativamente pequeno do *dataset* (297 amostras), a seleção de apenas 6 *features*, a validação exclusiva no *Cleveland* sem avaliação externa em outros conjuntos de dados, e a interpretabilidade limitada da arquitetura *deep learning* representam restrições que contextualizam os achados. Adicionalmente, o desempenho em ambiente experimental controlado não garante efetividade equivalente em aplicação clínica real, onde variabilidade na qualidade dos dados, heterogeneidade de populações e necessidade de integração com fluxos de trabalho existentes podem impactar significativamente os resultados.

Direções futuras de pesquisa incluem: (1) validação externa em *datasets* independentes (*Hungarian*, *Switzerland*, *VA Long Beach*, *Framingham*) para avaliar generalização; (2) exploração de variantes de *SMOTE* (*D-SMOTE*, *Borderline-SMOTE*, *ADASYN*) e estratégias alternativas de balanceamento; (3) investigação de diferentes subconjuntos de *features* ou uso do conjunto completo de 13 atributos originais; (4) implementação de técnicas avançadas de interpretabilidade (*SHAP*, *LIME*) para aumentar explicabilidade clínica; (5) avaliação em *datasets* maiores para explorar plenamente o potencial de *deep learning*; (6) desenvolvimento de *ensembles* heterogêneos combinando TabM com *XGBoost* e *Random Forest*; (7) estudos de calibração probabilística para melhorar confiabilidade das predições extremas; e (8) validação prospectiva em ambiente clínico real.

O TabM não alcançou o desempenho competitivo conforme definido nas hipóteses experimentais, levando à aceitação de H_0. O modelo apresentou *AUC-ROC* expressivamente superior aos critérios (92.93% vs. 90.39%), mas ficou abaixo em acurácia, *recall* e *F1-score*. Essa combinação de resultados sugere que o TabM mantém capacidade discriminativa em datasets pequenos, mas não alcança o equilibrio multi-métrico esperado. O trabalho contribui como primeira aplicação de TabM a diagnóstico cardíaco e fornece *pipeline* reprodutível e *MVP* operacional,

estabelecendo fundação para futuras pesquisas sobre *deep learning* tabular em medicina, independentemente dos resultados de performance numérica reportados aqui.

REFERÊNCIAS

AKIBA, Takuya *et al.* **Optuna: A Next-generation Hyperparameter Optimization Framework**. arXiv, , 25 jul. 2019. Disponível em: <http://arxiv.org/abs/1907.10902>. Acesso em: 7 dez. 2025

ALI, Md Mamun *et al.* Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. **Computers in Biology and Medicine**, v. 136, p. 104672, set. 2021.

SIMÃO, Af *et al.* I Diretriz Brasileira de Prevenção Cardiovascular. **Arquivos Brasileiros de Cardiologia**, v. 101, n. 6, p. 1–63, 2013.

ARLOT, Sylvain; CELISSE, Alain. A survey of cross-validation procedures for model selection. **Statistics Surveys**, v. 4, 1 jan. 2010.

BARREDO ARRIETA, Alejandro *et al.* Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. **Information Fusion**, v. 58, p. 82–115, jun. 2020.

BASHIR, Saba *et al.* MV5: A Clinical Decision Support Framework for Heart Disease Prediction Using Majority Vote Based Classifier Ensemble. **Arabian Journal for Science and Engineering**, v. 39, n. 11, p. 7771–7783, nov. 2014.

BERGSTRA, James et al. Algorithms for hyper-parameter optimization. In: Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS'11). Red Hook, NY: Curran Associates Inc., 2011. p. 2546–2554.

CHAWLA, N. V. *et al.* SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 1 jun. 2002.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: A Scalable Tree Boosting System. *In*: KDD '16: THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. San Francisco California USA: ACM, 13 ago. 2016. Disponível em: <https://dl.acm.org/doi/10.1145/2939672.2939785>. Acesso em: 7 dez. 2025

DETRANO, Robert *et al.* International application of a new probability algorithm for the diagnosis of coronary artery disease. **The American Journal of Cardiology**, v. 64, n. 5, p. 304–310, ago. 1989.

FAJRI, Yaumi A. Z. A.; WIHARTO, Wiharto; SURYANI, Esti. Hybrid Model Feature Selection with the Bee Swarm Optimization Method and Q-Learning on the Diagnosis of Coronary Heart Disease. **Information**, v. 14, n. 1, p. 15, 28 dez. 2022.

FAWCETT, Tom. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874, jun. 2006.

GIL, A. C. Como elaborar projetos de pesquisa. 6. ed. São Paulo: Atlas, 2017. 176p.

GORISHNIY, Yury *et al.* **Revisiting Deep Learning Models for Tabular Data**. arXiv, , 2021. Disponível em: <https://arxiv.org/abs/2106.11959>. Acesso em: 7 dez. 2025

GORISHNIY, Yury; KOTELNIKOV, Akim; BABENKO, Artem. **TabM: Advancing Tabular Deep Learning with Parameter-Efficient Ensembling**. arXiv, , 2024. Disponível em: <https://arxiv.org/abs/2410.24210>. Acesso em: 7 dez. 2025

GRINSZTAJN, Léo; OYALLON, Edouard; VAROQUAUX, Gaël. **Why do tree-based models still outperform deep learning on tabular data?** arXiv, , 2022. Disponível em: <https://arxiv.org/abs/2207.08815>. Acesso em: 7 dez. 2025

HOLLMANN, Noah *et al.* **TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second**. arXiv, , 2022. Disponível em: <https://arxiv.org/abs/2207.01848>. Acesso em: 7 dez. 2025

HUANG, Xin *et al.* **TabTransformer: Tabular Data Modeling Using Contextual Embeddings**. arXiv, , 2020. Disponível em: <https://arxiv.org/abs/2012.06678>. Acesso em: 7 dez. 2025

KHAN, M. A. et al. Effective heart disease prediction using machine learning techniques. **IEEE Access**, v. 11, p. 89970-89990, 2023.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 14., 1995, Montreal. Proceedings... Montreal: IJCAI, 1995. p. 1137-1143.

LATHA, C. Beulah Christalin; JEEVA, S. Carolin. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. **Informatics in Medicine Unlocked**, v. 16, p. 100203, 2019.

LOSHCHILOV, Ilya; HUTTER, Frank. **Decoupled Weight Decay Regularization**. arXiv, , 2017. Disponível em: <https://arxiv.org/abs/1711.05101>. Acesso em: 7 dez. 2025

MCELFRESH, Duncan *et al.* **When Do Neural Nets Outperform Boosted Trees on Tabular Data?** arXiv, , 2023. Disponível em: <https://arxiv.org/abs/2305.02997>. Acesso em: 7 dez. 2025

ORGANIZATION, World Health. **Global Atlas on Cardiovascular Disease Prevention and Control**. Geneva: World Health Organization, 2011.

POWERS, David M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 2020.

REDDY, G. Thippa *et al.* Analysis of Dimensionality Reduction Techniques on Big Data. **IEEE Access**, v. 8, p. 54776–54788, 2020.

ROTH, Gregory A. *et al.* Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019. **Journal of the American College of Cardiology**, v. 76, n. 25, p. 2982–3021, dez. 2020.

SHWARTZ-ZIV, Ravid; ARMON, Amitai. **Tabular Data: Deep Learning is Not All You Need**. arXiv, , 2021. Disponível em: <https://arxiv.org/abs/2106.03253>. Acesso em: 7 dez. 2025

SNOEK, Jasper; LAROCHELLE, Hugo; ADAMS, Ryan P. **Practical Bayesian Optimization of Machine Learning Algorithms**. arXiv, , 2012. Disponível em: <https://arxiv.org/abs/1206.2944>. Acesso em: 7 dez. 2025

SOCIEDADE BRASILEIRA DE CARDIOLOGIA. Cardiômetro – 2024. Disponível em: https://cardiometro.com.br. Acesso em: 7 dez. 2025

SOWJANYA, A. Mary; MRUDULA, Owk. Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms. **Applied Nanoscience**, v. 13, n. 3, p. 1829–1840, mar. 2023.

VASWANI, Ashish *et al.* **Attention Is All You Need**. arXiv, , 2017. Disponível em: <https://arxiv.org/abs/1706.03762>. Acesso em: 7 dez. 2025

VIRANI, Salim S. *et al.* Heart Disease and Stroke Statistics—2021 Update: A Report From the American Heart Association. **Circulation**, v. 143, n. 8, 23 fev. 2021.

WORLD HEALTH ORGANIZATION. Cardiovascular diseases (CVDs). Fact sheet. 2024. Disponível em: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Acesso em: 3 dez. 2024.

YOUDEN, W. J. Index for rating diagnostic tests. **Cancer**, v. 3, n. 1, p. 32–35, 1950.

ADEKOYA, Ayomide *et al.* Ensemble learning approach with explainable AI for improved heart disease prediction. **Frontiers in Pharmacology**, v. 16, p. 1654681, 11 dez. 2025.

SHRESTHA, D. Advanced Machine Learning Techniques for Predicting Heart Disease: A Comparative Analysis Using the Cleveland Heart Disease Dataset . **Applied Medical Informatics**, *[S. l.]*, v. 46, n. 3, 2024. Disponível em: https://ami.info.umfcluj.ro/index.php/AMI/article/view/1060. Acesso em: 12 dec. 2025.

GHOSE, Partho *et al.* Explainable AI assisted heart disease diagnosis through effective feature engineering and stacked ensemble learning. **Expert Systems with Applications**, v. 265, p. 125928, mar. 2025.

KOLUKISA, Burak; BAKIR-GUNGOR, Burcu. Ensemble feature selection and classification methods for machine learning-based coronary artery disease diagnosis. **Computer Standards & Interfaces**, v. 84, p. 103706, mar. 2023.

ERICKSON, Nick *et al.* **TabArena: A Living Benchmark for Machine Learning on Tabular Data**. arXiv, , 2025. Disponível em: <https://arxiv.org/abs/2506.16791>. Acesso em: 12 dez. 2025

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, 28 maio 2015.