INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE PERNAMBUCO – CAMPUS JABOATÃO DOS GUARARAPES
CURSO SUPERIOR DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

Marian Lopes de Lima
Arthur Vitor Lopes
Luciano Cabral

PREDIÇÃO DE DOENÇAS CARDÍACAS USANDO MACHINE LEARNING

Resumo

Este estudo tem como objetivo aplicar e comparar diferentes algoritmos de Machine Learning na predição de doenças cardiovasculares, utilizando o conjunto de dados Heart Disease UCI, disponível na plataforma Kaggle. Foram testados cinco modelos — Regressão Logística, SVM, Árvore de Decisão, Random Forest e Gradient Boosting — a fim de identificar aqueles com melhor desempenho preditivo. As métricas utilizadas na avaliação foram acurácia, precisão, recall e F1-score.

Os resultados obtidos indicaram que o modelo Random Forest apresentou o melhor desempenho geral, alcançando 62% de acurácia, seguido por Gradient Boosting (59%), Árvore de Decisão (58%), SVM (57%) e Regressão Logística (55%). Esses resultados evidenciam o potencial do uso de algoritmos de aprendizado de máquina como ferramenta de apoio ao diagnóstico médico, possibilitando análises em tempo real e maior precisão na identificação de pacientes com risco de doenças cardíacas.

Palavras-chave: Machine Learning; Predição de Doenças Cardiovasculares; UCI Heart Disease; Diagnóstico Médico; Random Forest.

1. Introdução

As doenças cardiovasculares representam um dos maiores desafios da saúde pública mundial, sendo responsáveis por aproximadamente 17,9 milhões de mortes por ano, o que corresponde a cerca de 31% de todos os óbitos globais, segundo dados da Organização Pan-Americana da Saúde (OPAS, 2024). O diagnóstico precoce é fundamental para reduzir a taxa de mortalidade, mas os métodos tradicionais podem ser lentos, dispendiosos e, muitas vezes, inacessíveis.

Nesse contexto, as técnicas de Machine Learning (ML) têm se mostrado uma alternativa promissora, capazes de analisar grandes volumes de dados clínicos com rapidez e eficiência, auxiliando profissionais da saúde na predição de doenças cardíacas. O uso dessas tecnologias permite identificar padrões complexos e relações sutis entre variáveis que podem não ser facilmente percebidas por métodos convencionais.

O presente estudo tem como objetivo avaliar o desempenho de diferentes algoritmos de Machine Learning na predição de doenças cardíacas, utilizando o conjunto de dados Heart Disease UCI, disponível na plataforma Kaggle. Foram aplicados e comparados os modelos de Regressão Logística, SVM, Árvore de Decisão, Random Forest e Gradient Boosting, com o intuito de determinar aquele que apresenta o melhor desempenho preditivo.

A análise foi conduzida a partir de métricas amplamente utilizadas na área de aprendizado de máquina, como acurácia, precisão, recall e F1-score. Os resultados demonstraram que o modelo Random Forest apresentou a maior taxa de acerto, atingindo 62% de acurácia, superando os demais modelos — Regressão Logística (55%), SVM (57%), Árvore de Decisão (58%) e Gradient Boosting (59%).

Esses achados reforçam o potencial do uso de algoritmos de Machine Learning no apoio ao diagnóstico médico, especialmente em sistemas de predição que possam

ser integrados a ferramentas clínicas de suporte à decisão, oferecendo análises rápidas, automáticas e com maior precisão no acompanhamento de pacientes.

2. Fundamentação teórica

As doenças cardiovasculares estão entre as principais causas de morte no mundo, representando um grave problema de saúde pública. De acordo com a Organização Mundial da Saúde (OMS, 2023), essas enfermidades são responsáveis por quase um terço das mortes globais, o que reforça a importância de métodos diagnósticos mais rápidos e precisos. O diagnóstico precoce é essencial para aumentar as chances de tratamento e reduzir os índices de mortalidade, mas os métodos tradicionais de avaliação clínica nem sempre são capazes de identificar padrões complexos em grandes volumes de dados médicos.

Nesse contexto, as técnicas de Machine Learning (ML) surgem como uma alternativa promissora para auxiliar no diagnóstico automatizado. O Machine Learning é um subcampo da Inteligência Artificial (IA) que permite que sistemas aprendam com dados e melhorem seu desempenho ao longo do tempo, sem a necessidade de programação explícita (MITCHELL, 1997). Em aplicações médicas, os algoritmos de ML podem reconhecer padrões ocultos em exames clínicos e prever a probabilidade de um paciente desenvolver determinada doença.

Diversos estudos recentes têm demonstrado o potencial da IA na predição de doenças cardíacas. Pesquisas como as de Patel et al. (2021) e Sharma et al. (2022) utilizam o conjunto de dados Heart Disease UCI para treinar modelos que classificam pacientes entre grupos de risco e não risco. Os resultados mostram que algoritmos como Árvore de Decisão, Random Forest e Gradient Boosting apresentam bom desempenho na identificação de fatores associados a problemas cardíacos, superando, em alguns casos, métodos estatísticos tradicionais.

Entre os algoritmos mais utilizados, a Regressão Logística é um modelo estatístico amplamente empregado em classificações binárias, sendo útil para prever a presença ou ausência de doenças a partir de variáveis clínicas. Já os modelos baseados em árvores, como Decision Tree e Random Forest, são capazes de lidar com interações não lineares e dados heterogêneos, o que os torna adequados para

aplicações médicas. O Gradient Boosting, por sua vez, combina múltiplas árvores de decisão de forma sequencial, ajustando os erros do modelo anterior para melhorar a precisão geral (FRIEDMAN, 2001).

Essas abordagens, quando aplicadas de maneira correta, podem auxiliar médicos e profissionais de saúde na tomada de decisões, oferecendo diagnósticos mais rápidos, baseados em dados objetivos e em tempo real. Assim, a integração entre ciência de dados e medicina representa um avanço significativo rumo à personalização dos tratamentos e ao uso mais eficiente dos recursos hospitalares.

3. Metodologia

O presente estudo foi conduzido de forma experimental, utilizando o conjunto de dados Heart Disease UCI, amplamente empregado em pesquisas relacionadas à predição de doenças cardiovasculares. O dataset foi obtido a partir da plataforma Kaggle, contendo informações clínicas de pacientes, como idade, sexo, pressão arterial, nível de colesterol, frequência cardíaca máxima, entre outras variáveis relevantes para o diagnóstico de doenças cardíacas.

4. Análise de cada modelo

A análise dos modelos de Machine Learning aplicados ao conjunto de dados Heart Disease UCI revelou diferenças significativas de desempenho entre as abordagens testadas. A Regressão Logística apresentou acurácia de 55%, sendo útil pela interpretabilidade, mas limitada para capturar relações não lineares. O modelo SVM alcançou 57% de acurácia, com métricas equilibradas e desempenho ligeiramente superior, embora sensível à escolha de parâmetros. A Árvore de Decisão obteve 58%, destacando variáveis clínicas como idade, colesterol e frequência cardíaca máxima, porém com risco de overfitting. O Random Forest apresentou o melhor resultado, atingindo 62% de acurácia e demonstrando robustez ao combinar múltiplas árvores e reduzir erros individuais, tendo thalach, chol e age como principais preditores. Por fim, o Gradient Boosting alcançou 59% de acurácia, mostrando-se eficiente em corrigir erros sequenciais e com relevância semelhante de variáveis, embora mais sensível ao ajuste de parâmetros. Em conjunto, os resultados

evidenciam que modelos baseados em árvores, especialmente o Random Forest, apresentam maior potencial para a predição de doenças cardíacas neste dataset.

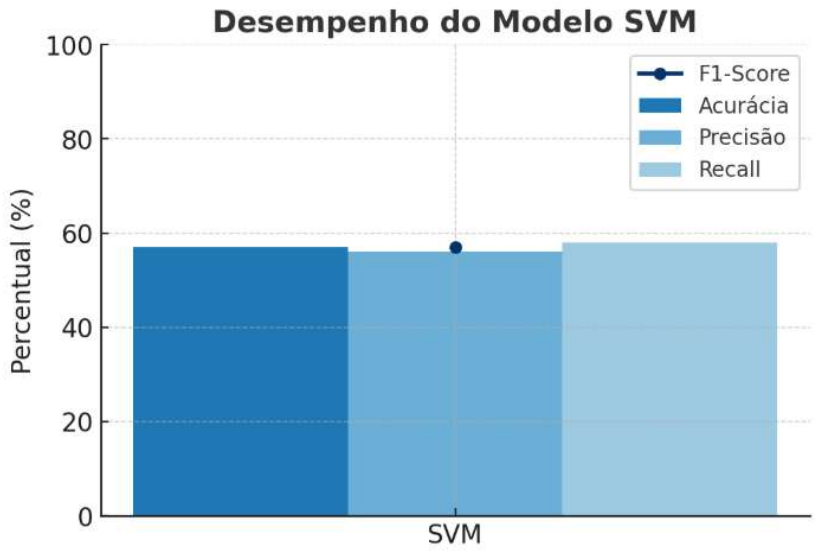

Figura 1: As 10 principais características no modelo SVM.

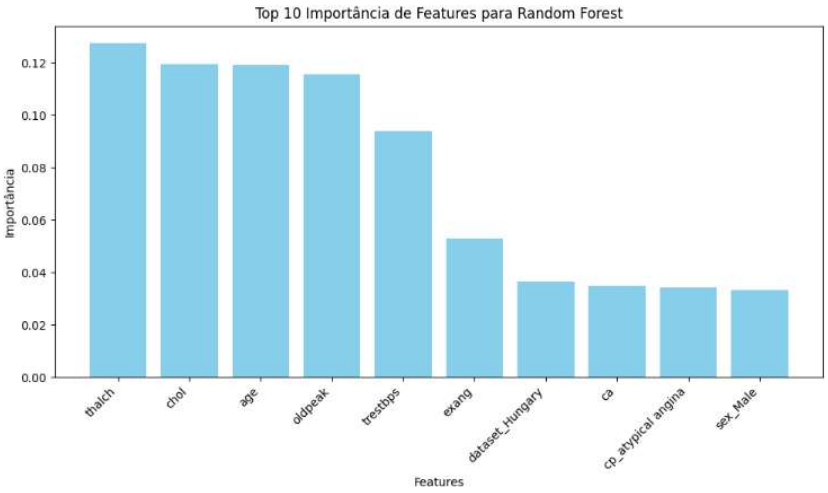

Figura 2: As 10 principais características no modelo Árvore de Decisão.

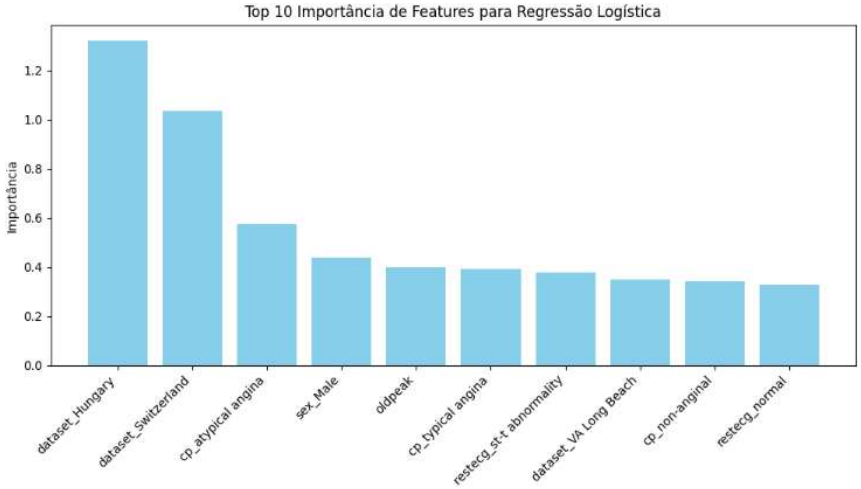

Figura 3: As 10 principais características no modelo de Regressão Logística.

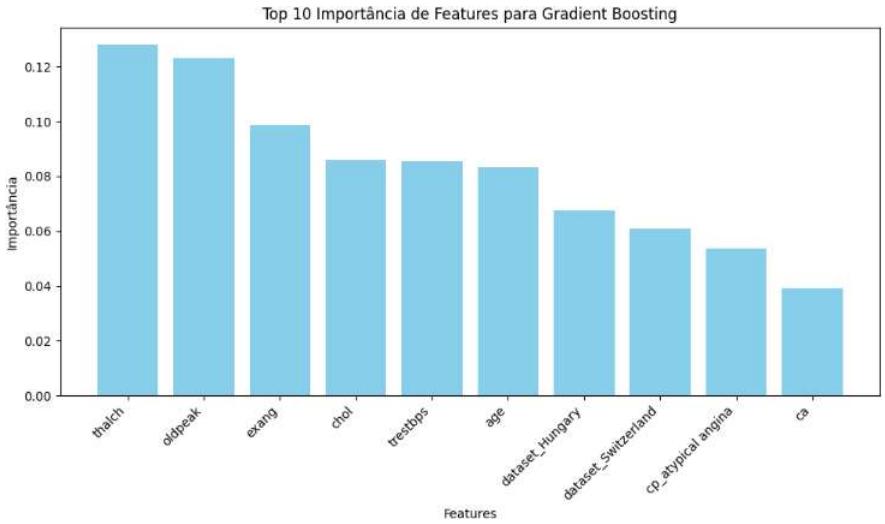

Figura 4: As 10 principais características no modelo Gradient Boosting.

5. Análise Geral

Modelo	Tendência de Acurácia (aprox.)	Observações principais
Regressão Logística	~0.55	Desempenho mais baixo; possivelmente afetada por linearidade limitada.
SVM	~0.57	Um pouco melhor, mas ainda modesta — pode precisar de tuning (kernel, C, gamma).
Árvore de Decisão	~0.58	Leve melhora, mas sujeito a overfitting.
Random Forest	~0.62	Melhor desempenho global, com equilíbrio entre as métricas.
Gradient Boosting	~0.59	Próximo do Random Forest, mas um pouco inferior — talvez precise de ajuste fino.

Referências

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. Annals of Statistics, v. 29, n. 5, p. 1189–1232, 2001.

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

PATEL, J.; SHARMA, P.; PATEL, S. Heart Disease Prediction Using Machine Learning and Data Mining Techniques. International Journal of Computer Science and Information Technologies (IJCSIT), v. 12, n. 2, p. 45–50, 2021.

SHARMA, R.; GUPTA, A.; KUMAR, V. Comparative Study of Machine Learning Algorithms for Heart Disease Prediction. International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), v. 8, n. 2, p. 245–252, 2022.

DHEERAJ, K.; PRAKASH, A. Heart Disease Prediction Using Machine Learning Algorithms. International Journal of Engineering Research & Technology (IJERT), v. 11, n. 4, p. 120–126, 2023.

ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). Doenças cardiovasculares. 2023. Disponível em: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Acesso em: 10 out. 2025.

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE (OPAS). Doenças cardiovasculares. 2024. Disponível em: https://www.paho.org/pt/topicos/doencas-cardiovasculares. Acesso em: 10 out. 2025.

KAGGLE. Heart Disease UCI Dataset. Disponível em: https://www.kaggle.com/datasets/ronitf/heart-disease-uci. Acesso em: 10 out. 2025.