

## Análise de Dados Clínicos para Predição de Doenças Cardíacas

**Samantha Alves<sup>1</sup>, Matheus Parizoti<sup>2</sup>, João Kleber<sup>3</sup>, Bethina Camargo<sup>4</sup>, Rafael Orsi<sup>5</sup>,  
Aruane Pineda<sup>6</sup>**

<sup>1</sup>Senai Suíço-Brasileira - Santo Amaro, São Paulo, Brasil  
(samantha.caetano@aluno.senai.br)

<sup>2</sup>Senai Suíço-Brasileira - Santo Amaro, Cajamar, Brasil

<sup>3</sup>Senai Suíço-Brasileira - Santo Amaro, Cajamar, Brasil

<sup>4</sup>Senai Suíço-Brasileira - Santo Amaro, São Paulo, Brasil

<sup>5</sup>Senai Suíço-Brasileira - Santo Amaro, São Paulo, Brasil

<sup>6</sup>Senai Suíço-Brasileira - Santo Amaro, São Paulo, Brasil

**Resumo:** Este trabalho utiliza XGBoost para prever doenças cardíacas com base no dataset Heart Disease (UCI). O modelo binário classifica pacientes como saudáveis ou doentes, alcançando acurácia de 86,67% e AUC de 0,943. Variáveis como thal, cp, slope e ca se destacaram, indicando forte relação com fatores clínicos relevantes e potencial aplicação na área da saúde.

**Palavras-chave:** Doença cardíaca; XGBoost; Aprendizado de máquina; Classificação; Análise de dados

### INTRODUÇÃO

As doenças cardiovasculares são uma das principais causas de morte no mundo. Por isso, o diagnóstico precoce dessas enfermidades torna-se essencial é muito importante para diminuir os riscos e melhorar a qualidade de vida dos pacientes. Nesse contexto, técnicas de aprendizado de máquina estão sendo usadas para ajudar os médicos a tomar decisões, identificando padrões importantes em dados médicos e colaborando para diagnósticos mais precisos. Estudos recentes também evidenciam o uso de modelos computacionais e análise de redes para compreender melhor a complexidade dessas doenças e melhorar a capacidade preditiva em sistemas de saúde (Aruane M Pineda, 2023). Além disto, doenças cardiovasculares continuam sendo responsáveis por uma grande parcela da mortalidade global, reforçando a demanda de ferramentas que auxiliem no diagnóstico precoce (World Health Organization, 2021).

Este trabalho usa a base de dados Heart Disease (UCI) (Janosi et al., 1989), sobre doenças cardiovasculares, para desenvolver um algoritmo que prediz se alguém tem doença cardíaca. O modelo de classificação foi fundamentado em XGBoost e

busca responder a uma problema central: a pessoa é saudável ou tem doença cardíaca?

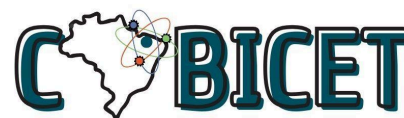
O propósito é criar um modelo que funcione bem e auxilie a prever doenças cardíacas. Para isso, os dados são preparados, o modelo é treinado e seu desempenho é avaliado usando medidas como acurácia, precisão e ROC. Também são verificadas quais variáveis são mais importantes para o modelo tomar suas decisões. Estudos confirmam que algoritmos de machine learning têm apresentado alto desempenho na predição de doenças cardiovasculares, superando métodos tradicionais em diversos cenários (Dritsas & Trigka, 2022).

Isso pode ajudar a melhorar a saúde das pessoas, fornecendo aos médicos ferramentas mais precisas para diagnosticar e tratar doenças cardíacas.

### MATERIAL E MÉTODOS

Os dados foram obtidos do repositório UCI Machine Learning Repository (Heart Disease Dataset). A análise foi realizada no ambiente de programação Jupyter Notebook, utilizando a linguagem Python e as bibliotecas pandas, matplotlib, seaborn e scikit-learn.

A variável alvo original, representada por valores de 0 a 4, foi transformada em uma variável binária, na



qual 0 corresponde a pacientes saudáveis e 1 indica a presença de doença cardíaca.

Para o pré-processamento dos dados, realizou-se o tratamento das variáveis, incluindo renomeação e aplicação de dummies. Em seguida, os dados foram divididos em conjuntos de treino e teste na proporção de 80% e 20%, respectivamente. Essa abordagem é amplamente utilizada em aprendizado de máquina para garantir a avaliação adequada do desempenho do modelo em dados não vistos (Goodfellow et al., 2016). Posteriormente, aplicou-se o modelo XGBoost para a tarefa de classificação. Esse modelo tem sido amplamente utilizado em problemas de classificação médica devido à sua alta performance e capacidade de lidar com dados estruturados (Chen & Guestrin, 2016).

A avaliação do modelo foi realizada por meio das seguintes métricas: acurácia; precisão (precision); Roc; e matriz de confusão.

## RESULTADOS E DISCUSSÃO

O modelo foi desenvolvido com base no conjunto de dados chamado Heart Disease, que é da Universidade da Califórnia, Irvine. Esse conjunto de dados é formado por 303 casos e 13 variáveis clínicas importantes, como a idade do paciente, o tipo de dor no peito que ele relata, a frequência cardíaca mais alta que ele alcança, se ele tem angina quando realiza exercícios e resultados de exames médicos como o thal, a inclinação e a presença de cálcio nas artérias. A informação mais importante que busca-se determinar é se o paciente está clinicamente saudável ou se ele possui alguma doença cardíaca. Essa é uma abordagem simplificada de estudar a doença cardíaca.

Quando avaliado o modelo utilizando dados que não foram usados na etapa de treinamento, observou-se um elevado desempenho, obtendo uma acurácia de 86,67% na detecção de doenças cardíacas. Esse resultado demonstra um desempenho satisfatório. Resultados semelhantes têm sido observados na literatura, onde modelos baseados em aprendizado de máquina apresentam alta acurácia na predição de doenças cardíacas (Alizadehsani et al., 2013). Além disso, ao comparar os resultados obtidos nos grupos de treinamento e teste, observou-se o desempenho do modelo, com acurácia de 96,62% no conjunto de treinamento e 86,67% no conjunto de teste. Esses resultados mostram que o modelo é capaz de aprender padrões a partir dos dados e realizar

previsões precisas, mesmo quando exposto a dados não observados previamente.

## Importância das Variáveis

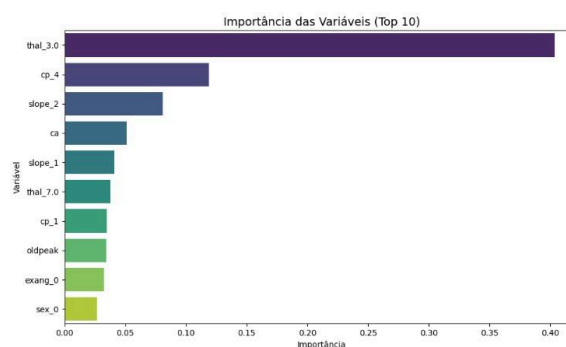


Figura 1. Importância das Variáveis.

A análise da importância das variáveis indica que o modelo atribuiu maior relevância a atributos diretamente relacionados ao funcionamento cardiovascular. A variável *thal* destacou-se como a mais significativa, sendo responsável por avaliar se o fluxo sanguíneo está chegando corretamente ao coração, aspecto fundamental na identificação de anomalias cardíacas. Em seguida, a variável *cp* está associada à presença de angina, caracterizada por dor torácica, um dos principais sintomas clínicos. A variável *slope* corresponde ao exame de eletrocardiograma, que analisa a resposta do coração ao esforço físico, enquanto *ca* representa o número de vasos sanguíneos visíveis, indicando possíveis obstruções nas artérias principais. Esses resultados evidenciam que o modelo foi capaz de identificar padrões consistentes ao conhecimento médico. Estudos clínicos também apontam que fatores como dor no peito, alterações no eletrocardiograma e fluxo sanguíneo são determinantes importantes na identificação de doenças cardíacas (Khatib et al., 2019). Outros estudos também destacam a relevância dessas variáveis como fortes preditores de eventos cardiovasculares (Fihn et al., 2012).

## Matriz de Confusão

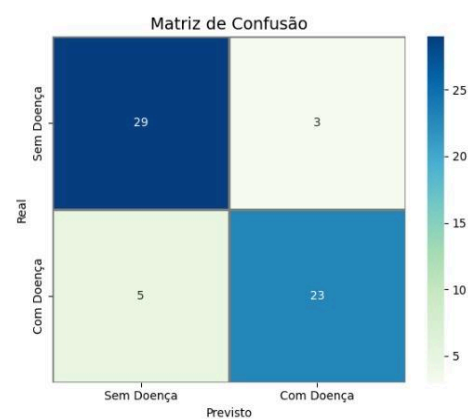


Figura 2. Matriz de Confusão.

A matriz de confusão indica que o modelo acertou 52 classificações em um total de 60 amostras, evidenciando desempenho satisfatório. O modelo identificou corretamente 29 pacientes saudáveis, correspondendo aos verdadeiros negativos, enquanto 23 pacientes com doença cardíaca foram corretamente classificados como verdadeiros positivos. No entanto, houve 3 falsos positivos, representando pacientes saudáveis classificados como incorretamente como doentes, e 5 falsos negativos, que são pacientes doentes não identificados pelo modelo.

Portanto, verifica-se que o modelo apresenta maior capacidade em identificar pacientes sem a doença. Isso pode estar relacionado à distribuição do conjunto de dados, indicando que a maioria dos pacientes no dataset não tem doença. Vale ressaltar a importância dos falsos negativos, pois eles representam pacientes doentes que não foram identificados, o que pode implicar problemas em situações reais, onde a identificação correta da doença é crucial.

A matriz de confusão é uma ferramenta fundamental na avaliação de modelos de classificação, pois permite analisar detalhadamente os acertos e erros do modelo, evidenciando o desempenho em cada classe por meio dos verdadeiros e falsos positivos e negativos (Powers, 2011). Esse tipo de análise é essencial em aplicações médicas, onde diferentes tipos de erro podem ter impactos distintos na tomada de decisão clínica (Sokolova & Lapalme, 2009).

De modo geral, o modelo apresentou baixa taxa de erro, o que mostra que ele é capaz de generalizar bem, indicando que ele pode ser uma ferramenta de apoio à tomada de decisão médica. Entretanto,

destaca-se que o modelo é apenas uma ferramenta complementar à avaliação clínica, não substituindo a análise feita por profissionais da saúde.

### Curva ROC

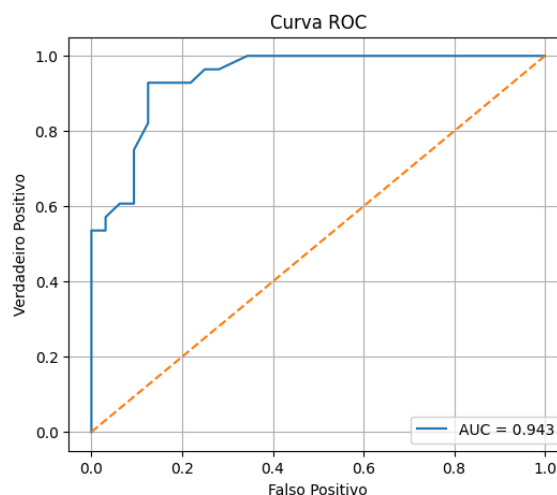
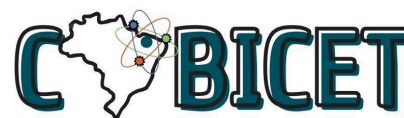


Figura 3. Curva ROC.

A curva ROC mostra uma área sob a curva, também chamada de AUC, de 0,943. Isso indica que o modelo tem uma excelente capacidade de distinguir entre pacientes que têm doença cardíaca e aqueles que não têm. A forma como a curva se comporta demonstra que o modelo consegue manter uma alta taxa de acertos, mesmo quando usamos diferentes níveis para decidir se um paciente tem ou não a doença. Isso mostra que o modelo é muito eficiente em separar os pacientes em dois grupos: os que têm doença cardíaca e os que não têm.

A curva ROC é amplamente utilizada na avaliação de modelos de classificação, pois analisa a relação entre a taxa de verdadeiros positivos e a taxa de falsos positivos, permitindo avaliar a capacidade do modelo em distinguir entre classes de forma independente do limiar de decisão (Hajian-Tilaki, 2013). Além disso, a AUC é considerada uma das métricas mais robustas para avaliar modelos em problemas de classificação médica (Fawcett, 2006).

### Discussão Geral



Em geral, os resultados que obtivemos estão de acordo com o que sabemos sobre clínica e com estudos sobre o assunto. As coisas mais importantes que descobrimos estão relacionadas a aspectos fundamentais, como o fluxo do sangue, a dor no peito, a resposta ao exercício físico e a obstrução das artérias, todos essenciais para diagnosticar doenças do coração. Nosso modelo mostrou um desempenho muito bom e consistente, demonstrando um grande potencial como uma ferramenta para ajudar a tomar decisões na área da saúde. Além disso, estudos recentes indicam que o uso de inteligência artificial e abordagens voltadas à medicina de precisão pode ampliar a eficiência do diagnóstico e do tratamento de doenças cardiovasculares, reforçando a relevância de modelos preditivos nesse contexto (Bansal et al., 2023).

No entanto, ainda há limitações, especialmente quando se trata de erros de classificação em casos mais delicados. Portanto, o modelo deve ser visto como uma ferramenta que complementa, que pode ajudar os profissionais da saúde a escolher os pacientes que precisam de atenção e a analisar o risco, mas não deve substituir exames clínicos ou diagnósticos especializados, sendo essencial manter a responsabilidade profissional, ética e a adequada relação médico-paciente no processo de decisão clínica (Silva et al., 2013).

### CONCLUSÃO

O modelo apresentou desempenho consistente na predição de doenças cardíacas, identificando como principais variáveis relevantes *thal*, *cp*, *slope* e *ca*, associadas à perfusão miocárdica, dor no peito, resposta ao esforço e obstrução arterial.

A abordagem binária mostrou-se eficaz para classificar pacientes entre saudáveis e com risco, indicando potencial aplicação como ferramenta de apoio à triagem clínica.

Apesar do bom desempenho geral, a ocorrência de erros de classificação com impacto clínico relevante evidencia a necessidade de aprimoramentos no modelo.

Nesse contexto, a utilização de modelos preditivos pode contribuir para a identificação precoce de pacientes em risco, auxiliando na tomada de decisão clínica e potencialmente reduzindo a progressão de doenças cardiovasculares, conforme evidenciado em estudos recentes na área (Attia et al., 2019). Dessa

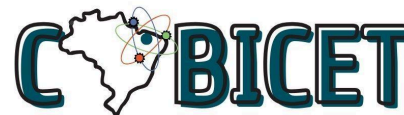
forma, o uso de inteligência artificial na saúde tende a crescer como suporte estratégico na prática médica (Topol, 2019).

### AGRADECIMENTOS

Agradecemos ao Senai suíço-brasileira, pela disponibilidade do ambiente e das tecnologias usadas para a análise.

### REFERÊNCIAS

- Alizadehsani, R.; Habibi, J.; Hosseini, M. J.; Mashayekhi, H.; Boghrati, R.; Ghandeharioun, A.; Bahadorian, B.; Sani, Z. A data mining approach for diagnosis of coronary artery disease. *Journal of Biomedical Informatics*, v. 46, p. 142–151, 2013.
- Attia, Z. I.; Kapa, S.; Lopez-Jimenez, F.; McKie, P. M.; Ladewig, D. J.; Satam, G.; Pellikka, P. A.; Enriquez-Sarano, M.; Noseworthy, P. A.; Lopez-Jimenez, F.; Friedman, P. A. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nature Medicine*, v. 25, p. 70–74, 2019.
- Bansal, A.; et al. Aplicações da inteligência artificial e medicina de precisão em doenças cardiovasculares. *Brazilian Journal of Medical and Health Sciences*, 2023. Disponível em: <https://revista.conaesbrasil.com.br/bjmhs/article/view/6>
- Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 785–794, 2016. Disponível em: <https://arxiv.org/abs/1603.02754>
- Dritsas, E.; Trigka, M. Machine learning methods for heart disease prediction. *Applied Sciences*, v. 12, n. 5, p. 2455, 2022. Disponível em: <https://www.mdpi.com/2076-3417/12/5/2455>
- Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters*, v. 27, n. 8, p. 861–874, 2006.
- Fihn, S. D.; Gardin, J. M.; Abrams, J.; Berra, K.; Blankenship, J. C.; Dallas, A. P.; Douglas, P. S.; Foody, J. M.; Gerber, T. C.; Hinderliter, A. L. 2012 ACCF/AHA guideline for the diagnosis and management of patients with stable ischemic heart disease. *Journal of the American College of Cardiology*, v. 60, n. 24, 2012.
- Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*. MIT Press, Cambridge, 2016. Disponível em: <https://www.deeplearningbook.org/>
- Hajian-Tilaki, K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian Journal of Internal Medicine*, v. 4, n. 2, p. 627–635, 2013. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/24009950/>



Janosi, A.; Steinbrunn, W.; Pfisterer, M.; Detrano, R. Heart Disease Dataset. UCI Machine Learning Repository, University of California, Irvine, 1989. Disponível em: <https://uci-ics-mlr-prod.aws.uci.edu/dataset/45/heart%2Bdisease>

Khatib, R.; McKee, M.; Shannon, H.; Chow, C.; Rangarajan, S.; Teo, K.; Yusuf, S. Availability and affordability of cardiovascular disease medicines and their effect on use in high-income, middle-income, and low-income countries. *The Lancet*, v. 393, p. 61–69, 2019. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6760171/>

Pineda, A. M.; Alves, C. L.; Möckel, M.; Toutain, T. G. L. O.; Porto, J. A. M.; Rodrigues, F. A. Network-based approaches for cardiovascular disease analysis. *Computer Networks*, v. 11, n. 5, 2023. Disponível em: <https://academic.oup.com/comnet/article-abstract/11/5/cnad030/7260365>

Powers, D. M. W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, v. 2, n. 1, p. 37–63, 2011. Disponível em: [https://www.bioinfopublication.org/files/articles/2\\_1\\_1\\_JMLT.pdf](https://www.bioinfopublication.org/files/articles/2_1_1_JMLT.pdf)

Silva, R. M.; et al. Ética, relação médico-paciente e prática clínica. *Revista Médica de Minas Gerais*, 2013. Disponível em: <https://rmmg.org/artigo/detalhes/608>

Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, v. 45, n. 4, p. 427–437, 2009.

Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, v. 25, p. 44–56, 2019. Disponível em: <https://www.nature.com/articles/s41591-018-0300-7>

World Health Organization. Cardiovascular diseases (CVDs). WHO, 2021. Disponível em: [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))