Análise de desempenho de modelos de Aprendizado de máquina na base de dados Heart Disease

Cipriani Leonardo
Universidade de Sao Paulo
Escola Politécnica
São Paulo, Brasil
https://orcid.org/0009-0009-5522-3408

Rampim Thiago
Universidade de Sao Paulo
Escola Politécnica
São Paulo, Brasil
https://orcid.org/0009-0002-8710-3880

***Resumo*—Doenças coronarianas são uma das principais causas de morte ao redor do mundo. Por esse motivo, a comunidade científica se debruça sobre as causas, consequências e diagnósticos dessas doenças. Com a evolução de métodos computancionais, o desenvolvimento e pesquisa de ferramentas que pudessem auxiliar no diagnóstico dessas doenças aumentou significativamente com o passar dos anos. Atualmente, técnicas de inteligência artificial estão sendo aplicadas na classificação de doencas coronarianas agudas, como o infarto do miocárdio. Este trabalho tem como objetivo analisar a relação entre fatores de risco e a presença de doenças cardíacas, além de avaliar o desempenho de diversos modelos de aprendizado de máquina na classificação dessas doenças utilizando a base de dados UCI Heart Disease. Foram utilizados modelos como Regressão Logística, Árvore de Decisão, SVM, KNN, Naive Bayes, Random Forest e Gradient Boosting. Os resultados obtidos foram comparados com trabalhos relacionados, destacando as métricas de acurácia e sensibilidade. Os resultados indicam que os modelos de aprendizado de máquina podem ser eficazes na identificação de doenças cardíacas, contribuindo para o avanço das ferramentas de diagnóstico na área da saúde.**

***Index Terms*—Banco de dados, Aprendizado Estatístico, doenças cardíacas**

I. Introdução

De acordo com o Ministério da Saúde, o infarto agudo do miocárdio é a maior causa de morte no Brasil e no mundo [1]. A partir dos anos 60, observa-se no Brasil e no mundo, um aumento das doenças crônicas não transmissíveis. Em 2021, as doenças cronicas cardiovasculares foram responsáveis pela morte de 20,8 milhões de pessoas [2]. Além das vidas perdidas, essas condicoes acarretam em comorbidades que impactam fisicamente e socialmente os pacientes, comprometendo sua qualidade de vida.

O objetivo desse estudo é analisar a relação entre os fatores de risco associados à presença de doenças cardíacas, como a frequência cardíaca máxima, idade e outros. Além disso, realizar uma análise de desempenho de classificadores de machine learning para realizar a identificação de doenças cardíacas.

Será utilizado o conjunto de dados de doenças cardíacas, e realizada a avaliação de desempenho com trabalhos relacionados [3].

II. Trabalhos Relacionados

O trabalho de Detrano et al. [3], apresenta a base de dados utilizada nesse estudo. O conjunto de dados foi coletado em 4 hospitais diferentes, sendo eles: Cleveland Clinic Foundation (CCF), Hungarian Institute of Cardiology, University Hospital Zurich (Switzerland) e Veterans Administration Medical Center (Long Beach, California). O qual deu origem a base de dados UCI Heart Disease, amplamente utilizada em estudos relacionados a doenças cardíacas. O algoritmo proposto foi baseado em uma rede bayesiana, obtendo uma especificidade de 85% para a base de dados de Cleveland.

Já o trabalho de Naser et al. [4], apresenta uma comparação entre diversos modelos de aprendizado de máquina para a identificação de doenças cardíacas. Foram utilizados os modelos de Regressão Logística, SVM, Random Forest e Gradient Boosting. O modelo de Gradient Boosting obteve o melhor desempenho, com uma acurácia média de 81.57%.

O trabalho de Li, et al. [5], propõe um modelo híbrido baseado em Random Forest e XGBoost para a identificação de doenças cardíacas. O modelo proposto obteve uma acurácia de 89.74%, superando os resultados apresentados em outros trabalhos relacionados.

Diversas outras técnicas e trabalhos foram desenvolvidos na área, devido a grande aplicabilidade e importância do tema.

Além da criação e comapração de modelos, outros trabalhos focam na análise dos fatores de risco associados à presença de doenças cardíacas. O trabalho de Regitz-Zagrosek et al. [6], destaca a importância de considerar o sexo biológico como um fator de risco para doenças cardíacas, evidenciando diferenças na prevalência e apresentação clínica entre homens e mulheres.

III. Base de dados

Existem diversas bases de dados disponíveis com informações de doenças cardiovasculares. Para os propósitos apresentados neste trabalho, foi utilizada a base UCI Heart Diseases, disponível em [7].

O conjunto de dados formada por 4 bases de dados, sendo elas: Cleveland, Hungarian, Switzerland e Long Beach. São fornecidos os dados brutos e refinados, contendo ainda os metadados.

Tabela I
TOTAL DE REGISTROS POR BASE DE DADOS

Base de Dados	Total de Registros
Cleveland	303
Hungarian	294
Switzerland	123
Long Beach VA	200
Total Geral	**920**

Cada uma das tabelas contém dados referentes a uma região. A quantidade de dados em cada uma das bases é descrita da Tabela I

Nesse conjunto de dados, para análises de dados e utilização em modelos de aprendizado de máquina, os autores sugerem que se utilize a base de dados de Cleveland, contendo o total de 303 registros.

As outras bases contém muitos dados faltantes, o que prejudicaria a confirmação ou descarte das hipoóteses levantadas.

Para a utilização adequada, seguindo os preceitos do ciclo de vida de dados, foram definidos o armazenamento e acesso conforme as seções a seguir.

Outras fontes de dados podem ser encontrados na PhysioNet, que contempla diversos conjuntos de dados relacionados a saúde [8].

A. Armazenamento e Segurança

O armazenamento dos dados foi feito utilizando a solução AWS S3, que fornece um repositório de objetos. Através dessa ferramenta, define-se políticas de ciclo de vida e de privacidade dos dados.

Não existem informações públicas ou sensíveis nos dados, portanto o acesso pode ser realizado por qualquer pessoa. Camadas de restrição de acesso não precisam ser aplicadas.

B. Descrição dos Dados

Foi utilizada a base de dados da cidade de Cleveland, com um total de 303 registros. A base contém 14 atributos, sendo eles descritos na Tabela VIII.

Além disso, a base de dados foi disponibilizada na plataforma Zenodo, contendo o DOI https://doi.org/10.5281/zenodo.17559614.

IV. QUESTÕES ANALÍTICAS

As primeiras questões a serem respondidas, serão relacionadas a presença ou ausência de doença cardíaca. A variável num será utiliada para partição dos dados.

A idade do paciente pode influenciar na presença de doença cardíaca. Intuitivamente, espera-se uma idade média maior em pacientes com rótulo positivo. Serão utilizados gráficos de boxplot para avaliação dessa hipoótese.

A relação entre sexo biológico do paciente e a presença de doenças cardiovasculares também será avaliada.

A frequência máxima atingida, dada pela variável $thalach$ e a relação com a variável num também será avaliado.

A relação da variável $chol$, que avalia o nível de colesterol presente e a presença ou ausência de doenças cardíacas, também será avaliada.

Um dos sintomas mais comuns associados à doenças do coração, é a dor torácica, dada pela variável cp. A relação dessa variável com a variável num também será observada.

Além dessas questões analíticas, foram desenvolvidos modelos de aprendizado de máquina e comparado os resultados. Através da comparação dos resultados obtidos, com os trabalhos relacionados, será avaliado se houve melhora nos resultados gerais, utilizando métricas de Acurácia e precisão.

V. MÉTODOS E MATERIAIS

A metodologia utilizada, segue o ciclo de vida dos dados, considerando as etapas de planejamento, coleta, processamento, análise e visualização dos dados. Através dessas etapas, busca-se responder as questões analíticas levantadas e avaliar o desempenho de modelos de aprendizado de máquina.

A. Planejamento e Documentação

Conforme direcionado durante as aulas, foi utilizada a ferramenta DMPtool para documentar e acompanhar o desenvolvimento do projeto.

Com a catalogação na ferramenta, é gerado um DOI vinculado ao projeto. O DOI do projeto gerado foi o https://doi.org/10.48321/D12DFBEE7A.

B. Análise dos dados

A análise realizada será dividida na etapa de descrição dos dados, onde descreveremos os dados numéricos e categóricos. Serão apresentadas as métricas clássicas de cada uma das variáveis.

Após a descrição, serão realizadas visualizações específicas sobre alguns conjuntos de dados buscando responder as questões analíticas.

1) Descrição: Os atributos numéricos estão descritos e resumidos na Tabela VII.

Já os atributos categóricos, estão descritos na Tabela VI

2) Visualização: Uma das hipoteses levantadas, é a influência do sexo biológico na presença de doenças cardíacas. A Figura 1 ilustra essa relação.

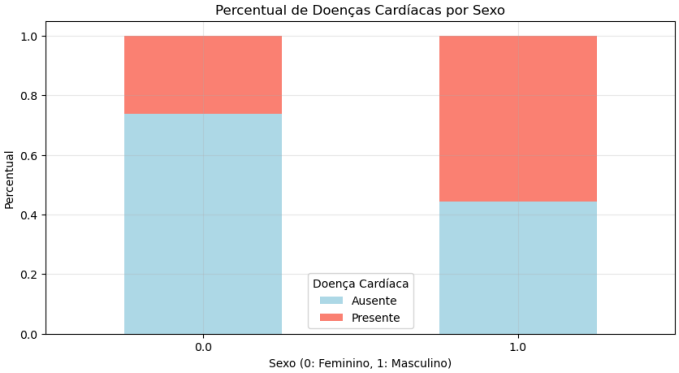

Figura 1. Distribuição da variável *sex* (sexo biológico)

De posse dessa visualização, é possível observar que a presença de doenças cardíacas é maior em pacientes do sexo masculino. O que corrobora com estudos epidemiológicos que indicam maior prevalência de doenças cardíacas em homens [6].

Para a hipótese da influência da idade, na presença de doenças cardíacas, foi utilizada a análise de boxplot para os dados. A Figura 2 ilustra essa relação.

Foi partida a análise entre os sexos, para observar se há diferença na influência da idade entre homens e mulheres. Nota-se uma presença maior de doenças cardíacas em pacientes mais velhos, tanto no sexo masculino quanto no feminino. Ou seja, a idade tem uma influência direta na presença de doenças cardíacas.

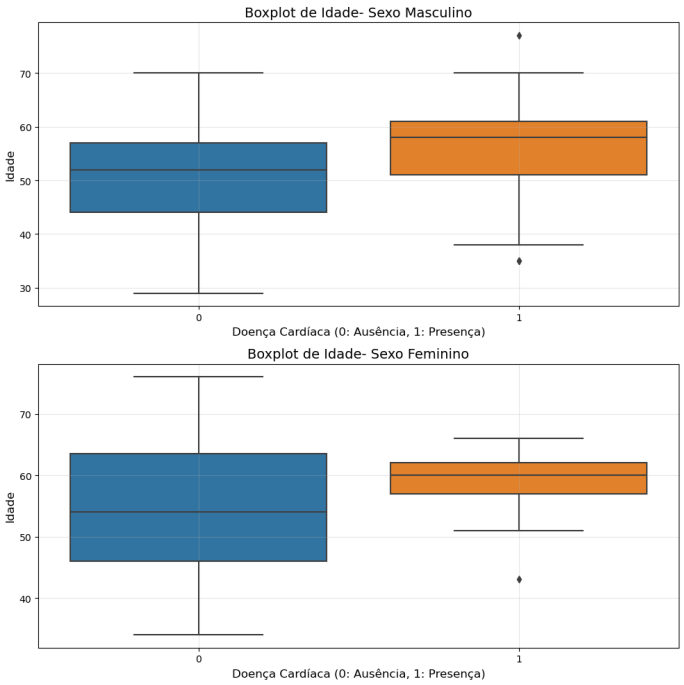

Figura 2. Distribuição da variável *age* (idade) por sexo.

A relação entre a frequência máxima atingida e a presença ou ausência de doenças cardíacas, pode ser visto no $boxplot$ gerado na Figura 3

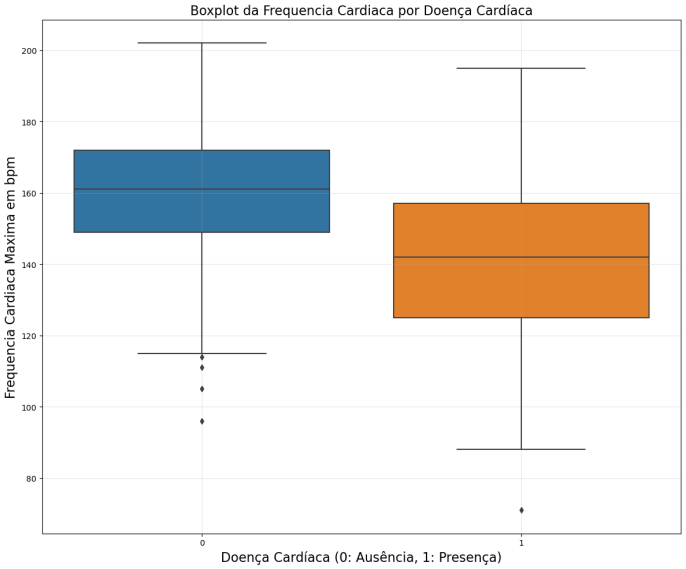

Figura 3. Distribuição da variável *trestbps* (frequência máxima)

A relação entre o tipo de dor torácica e a presença ou ausência de doenças cardíacas, pode ser visto no gráfico de barras gerado na Figura 4

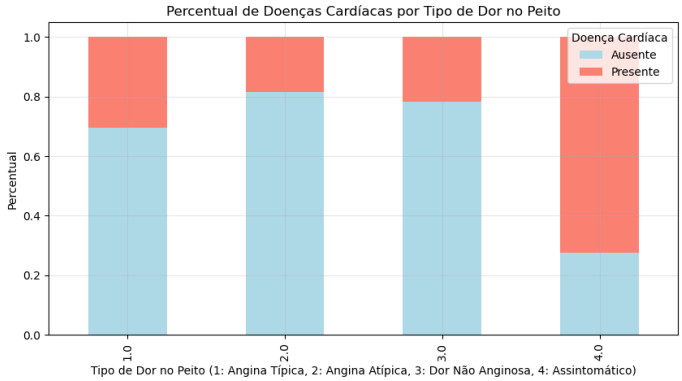

Figura 4. Distribuição da variável *cp* (tipo de dor torácica)

A dominância de dor do tipo 4 (assintomático) em pacientes com doenças cardíacas, pode indicar que a ausência de sintomas não é um indicativo confiável para descartar a presença de doenças cardíacas.

A análise de resultado do ECG de repouso, pode ser vista na Figura 5

Ao observar a Figura 5, nota-se que o resultado 1, que indica alteração do segmento ST-T, está mais presente em pacientes com doenças cardíacas. Isso está alinhado com a literatura médica, que associa alterações no segmento ST-T a condições cardíacas, como isquemia e infarto do miocárdio [9].

Para a relação entre a deformação do segmento ST (oldpeak) e a presença ou ausência de doenças cardíacas, foi realizada uma primeira análise considerando a deformação do segmento ST-T e o percentual de valores com e sem doenças cardíacas. A Figura 6 ilustra essa relação.

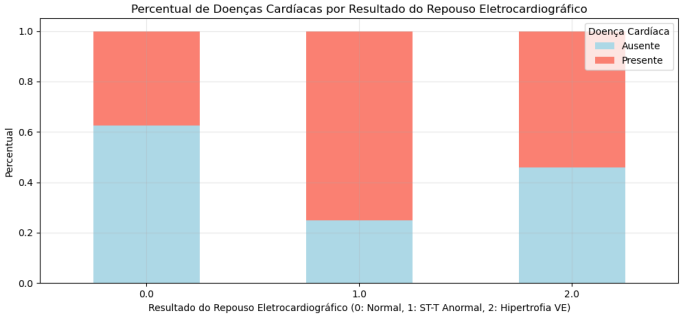

Figura 5. Distribuição da variável *restecg* (resultado do ECG de repouso)

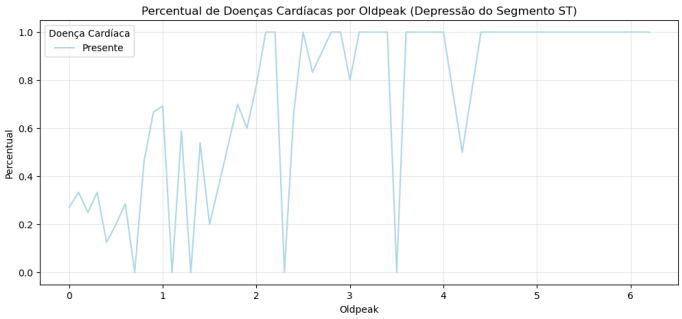

Figura 6. Distribuição da variável *oldpeak* (deformação do segmento ST)

À partir dessas informações, foi desenvolvido um modelo de regressão linear para avaliar a relação entre a deformação do segmento ST e a presença de doenças cardíacas. O modelo indicou uma correlação positiva entre o aumento da deformação do segmento ST e a presença de doenças cardíacas, sugerindo que pacientes com maior deformação têm maior probabilidade de apresentar condições cardíacas. A Figura 7 ilustra o modelo de regressão linear desenvolvido e os dados observados.

Complementar às análises de boxplot, foi realizada a geração da matriz de correlação entre as variáveis. A Figura 8 demonstra a correlação entre as variáveis. Para essa análise, a variável alvo num também está contida na matriz de correlação.

Pode-se observar uma correlação negativa entre idade (age) e thalach (frequência cardíaca máxima), indicando que pacientes mais velhos tendem a ter uma frequência cardíaca máxima menor. ALém disso, a correlação negetaiva entre oldpeak (deformação do segmento ST) e thalach sugere que pacientes com maior deformação do segmento ST tendem a ter uma frequência cardíaca máxima menor.

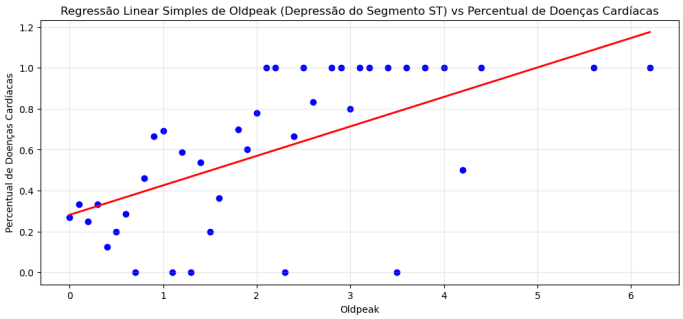

Figura 7. Modelo de regressão linear entre a variável *oldpeak* e a presença de doenças cardíacas.

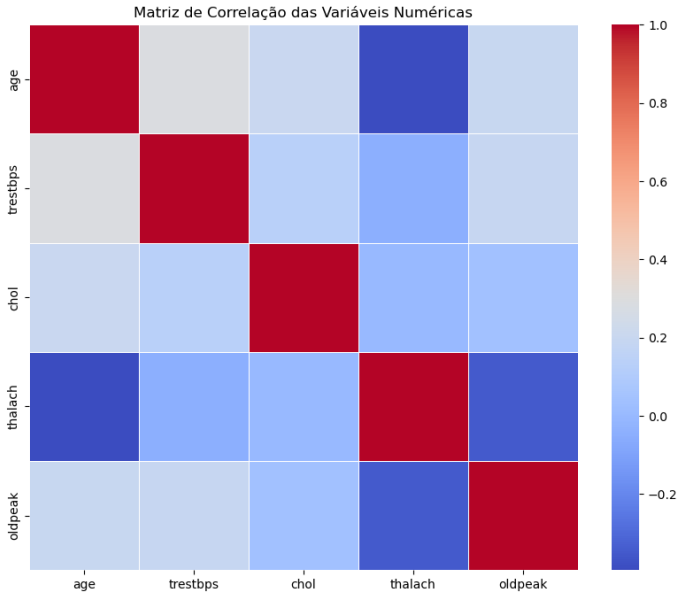

Figura 8. Matriz de correlação entre as variáveis.

C. Modelos de Aprendizado de Máquina

Foram utilizados diversos modelos de aprendizado de máquina para classificação binária, buscando identificar a presença ou ausência de doenças cardíacas. Os modelos utilizados foram: Regressão Logística, Árvore de Decisão, SVM, KNN, Naive Bayes, Random Forest e Gradient Boosting. A avaliação dos modelos foi realizada utilizando a técnica de validação cruzada k-fold, com k=10. As métricas utilizadas para avaliação foram Acurácia e Recall

VI. Resultados Obtidos

A. Modelos desenvolvidos

Os resultados obtidos para cada modelo estão resumidos na Tabela II referente a acurácia e na Tabela III.

Tabela II
ACURÁCIA PARA CLASSIFICADORES PARA KFOLD = 10

Modelo	Min	Max	Média	Q1	Q2	Q3
Logistic Regression	0.667	1.000	0.846	0.800	0.851	0.893
Decision Tree	0.542	0.870	0.728	0.688	0.729	0.776
SVM	0.583	0.708	0.654	0.625	0.660	0.694
KNN	0.417	0.875	0.642	0.583	0.661	0.695
Naive Bayes	0.500	0.957	0.747	0.729	0.769	0.838
Random Forest	0.583	0.870	0.763	0.729	0.787	0.828
Gradient Boosting	0.667	0.913	0.798	0.758	0.809	0.865

Tabela III
SENSIBILIDADE PARA CLASSIFICADORES PARA KFOLD = 10

Modelo	Min	Max	Média	Q1	Q2	Q3
Logistic Regression	0.685	1.000	0.849	0.792	0.871	0.914
Decision Tree	0.458	0.871	0.716	0.657	0.757	0.789
SVM	0.557	0.708	0.656	0.610	0.669	0.696
KNN	0.427	0.864	0.638	0.500	0.671	0.707
Naive Bayes	0.510	0.955	0.777	0.711	0.792	0.879
Random Forest	0.633	0.871	0.788	0.708	0.821	0.846
Gradient Boosting	0.671	0.913	0.796	0.757	0.775	0.856

B. Comparação com Trabalhos Relacionados

Considerando os resultados obtidos, foi realizada uma comparação com os trabalhos relacionados apresentados na seção 2. A Tabela IV apresenta um resumo das métricas de desempenho dos modelos utilizados em outros trabalhos.

Tabela IV
RESUMO DAS MÉTRICAS POR MODELO DE OUTROS TRABALHOS

Modelo	**Mínimo**	**Médio**	**Máximo**
Gradient Boosting	72.38	81.57	89.74
SVM	55.26	65.78	76.31
Random Forest	71.05	80.26	88.15
Logistic Regression	72.36	81.75	89.47

Tabela V
MEDIDAS DE PERFORMANCE DOS CLASSIFICADORES

Métrica	**DT**	**Naïve Bayes**	**RF**	**SVM**
Acurácia	0.9815	0.8037	0.9148	0.7556
Especificidade	0.0000	0.0000	0.0000	0.0000
Precisão	0.9815	0.8037	0.9148	0.7556
Sensibilidade	1.0000	1.0000	1.0000	1.0000
F-medida	0.9907	0.8912	0.9556	0.8608

VII. CONCLUSÃO

Considerando a análise exploratória dos dados e a avaliação dos modelos de aprendizado de máquina, conclui-se que os modelos desenvolvidos apresentam um desempenho competitivo na identificação de doenças cardíacas, com destaque para o modelo de Regressão Logística, que obteve a maior acurácia média entre os modelos testados.

A característica de idade avançada, sugere uma relação positiva com a presença de doenças cardíacas, corroborando com estudos epidemiológicos existentes. Além disso, o sexo biológico masculino mostrou-se mais prevalente entre os pacientes com doenças cardíacas.

Uma frequência cardíaca máxima reduzida também foi associada à presença de doenças cardíacas, indicando que essa variável pode ser um indicador relevante para a avaliação do risco cardiovascular.

Além desses critérios, a análise do tipo de dor torácica como indicador de doenças cardíacas revelou que pacientes assintomáticos ainda podem apresentar condições cardíacas, ressaltando a importância de avaliações clínicas complementares.

A análise do resultado do ECG de repouso indicou que alterações no segmento ST-T estão associadas a doenças cardíacas, alinhando-se com a literatura médica existente. O que dificulta a implementação de sistemas automatizados de diagnóstico, devido a complexidade e demora na análise do ECG, bem como a necessidade de validação clínica rigorosa antes da classificação automatizada.

REFERÊNCIAS

[1] M. da Saude, "Infarto agudo do miocárdio," https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/i/infarto#:~:text=Infarto%20agudo%20do%20miocrdio%20ou,da%20rea%20que%20foi%20obstruda., 2024, acessado em 10/08/2024.

[2] M. Di Cesare, H. Bixby, T. Gaziano, L. Hadeed, C. Kabudula, D. V. McGhie, J. Mwangi, B. Pervan, P. Perel, D. Piñeiro *et al.*, "World heart report 2023: Confronting the world's number one killer," *World Heart Federation: Geneva, Switzerland*, 2023.

[3] R. Detrano, A. Janosi, and Steinbrunn, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American journal of cardiology*, vol. 64, no. 5, pp. 304–310, 1989.

[4] M. A. Naser, A. A. Majeed, M. Alsabah, T. R. Al-Shaikhli, and K. M. Kaky, "A review of machine learning's role in cardiovascular disease prediction: Recent advances and future challenges," *Algorithms*, vol. 17, no. 2, 2024. [Online]. Available: https://www.mdpi.com/1999-4893/17/2/78

[5] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning for healthcare informatics," *arXiv preprint arXiv:2003.08119*, 2020.

[6] V. Regitz-Zagrosek and C. Gebhard, "Gender medicine: effects of sex and gender on cardiovascular disease manifestation and outcomes," *Nature Reviews Cardiology*, vol. 20, no. 4, pp. 236–247, 2023.

[7] S. W. P. Janosi, Andras, "Heart Disease," UCI Machine Learning Repository, 1989, DOI: https://doi.org/10.24432/C52P4X.

[8] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," Circulation 101(23):e215–e220, 2000, https://physionet.org/about/database/.

[9] M. Goldman, *Principles of Clinical Electrocardiography*. Lange Medical Publications, 1960, no. v. 1960. [Online]. Available: https://books.google.com.br/books?id=FgFiU2zq7FYC

APÊNDICE

Tabela VI
DISTRIBUIÇÃO DAS VARIÁVEIS CATEGÓRICAS COM PERCENTUAIS

Variável	**Categoria**	**Frequência**	**Frequência (%)**
sex	1	201	67.7%
	0	96	32.3%
cp	4	142	47.8%
	3	83	27.9%
	2	49	16.5%
	1	23	7.7%
fbs	0	254	85.5%
	1	43	14.5%
restecg	0	147	49.5%
	2	146	49.2%
	1	4	1.3%
exang	0	200	67.3%
	1	97	32.7%
slope	1	139	46.8%
	2	137	46.1%
	3	21	7.1%
ca	0	174	58.6%
	1	65	21.9%
	2	38	12.8%
	3	20	6.7%
thal	3	164	55.2%
	7	115	38.7%
	6	18	6.1%

Tabela VII
RESUMO ESTATÍSTICO DAS VARIÁVEIS NUMÉRICAS

Variável	Contagem	Média	Desvio Padrão	Mínimo	25%	Mediana	75%	Máximo
age	297	54.54	9.05	29	48	56	61	77
trestbps	297	131.69	17.76	94	120	130	140	200
chol	297	247.35	52.00	126	211	243	276	564
thalach	297	149.60	22.94	71	133	153	166	202
oldpeak	297	1.06	1.17	0.0	0.0	0.8	1.6	6.2

Tabela VIII
DESCRIÇÃO DOS ATRIBUTOS DO CONJUNTO DE DADOS COM DOMÍNIOS

Variável	Tipo	Descrição	Domínio
age	Numérica contínua	Idade do paciente em anos.	Não se aplica
sex	Categórica binária	Sexo biológico do paciente.	1 = Masculino; 0 = Feminino
cp	Categórica ordinal	Tipo de dor torácica.	1 = Angina típica; 2 = Angina atípica; 3 = Dor não anginosa; 4 = Assintomático
trestbps	Numérica contínua	Pressão arterial em repouso (mmHg).	Não se aplica
chol	Numérica contínua	Nível sérico de colesterol (mg/dL).	Não se aplica
fbs	Categórica binária	Glicemia de jejum ¿ 120 mg/dL.	1 = Sim; 0 = Não
restecg	Categórica ordinal	Resultado do eletrocardiograma de repouso.	0 = Normal; 1 = Alteração ST-T; 2 = Hipertrofia ventricular esquerda
thalach	Numérica contínua	Frequência cardíaca máxima atingida.	Não se aplica
exang	Categórica binária	Angina induzida por exercício.	1 = Sim; 0 = Não
oldpeak	Numérica contínua	Depressão do segmento ST induzida por exercício.	Não se aplica
slope	Categórica ordinal	Inclinação do segmento ST no pico do exercício.	1 = Ascendente; 2 = Plana; 3 = Descendente
ca	Numérica discreta	Número de vasos principais visualizados por fluoroscopia.	Inteiros: 0–3
thal	Categórica nominal	Resultado do teste de tálio.	3 = Normal; 6 = Defeito fixo; 7 = Defeito reversível
num	Categórica binária (alvo)	Diagnóstico de doença cardíaca.	0 = Sem doença significativa; 1 = Doença Presente