

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

*Note: Sub-titles are not captured in Xplore and should not be used

Cipriani Leonardo
Universidade de São Paulo
Escola Politécnica
São Paulo, Brasil
email address or ORCID

Rampim Thiago
Universidade de São Paulo
Escola Politécnica
São Paulo, Brasil
email address or ORCID

Resumo—

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 crônicas cardiovasculares foram responsáveis pela morte de 20,8 milhões de pessoas [2]. Além das vidas perdidas, essas condições 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.

II. 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 [3].

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.

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 contêm muitos dados faltantes, o que prejudicaria a confirmação ou descarte das hipóteses levantadas.

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

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

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 ??.

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

III. QUESTÕES ANALÍTICAS

Defina questões analíticas e hipóteses sobre os Datasets: técnicas estatísticas aplicadas à seleção e definição de dados a serem aplicados em experimentos computacionais.

Será analisada a relação entre a prática de atividade física e a predominância de doenças cardíacas. Questões levantadas.

- 1. A idade e o sexo estão relacionados com a presença de doença cardíaca?

- 2. Como a frequência cardíaca máxima (thalach) e o nível de colesterol (chol) variam entre pacientes com e sem doença cardíaca?
- 3. Existe diferença no risco de doença cardíaca entre os diferentes tipos de dor no peito (cp)?

IV. MÉTODOS E MATERIAIS

Traduza as questões: em ações e procedimentos a serem adotados em cada uma das etapas do ciclo de vida dos dados: Planejamento, ..., Análise/Visualização/Publicação

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 foram descritos e resumidos na Tabela II.

Já os atributos categóricos, foram descritos na Tabela III

2) *Visualização:*

C.

V. RESULTADOS OBTIDOS

Comparação entre os modelos

VI. CONCLUSÃO

Publicação: Os trabalhos devem ser disponibilizados na comunidade - Big Data Analytics Research Group of Escola Politécnica da Universidade de São Paulo - Zenodo (zenodo.org).

This document is a model and instructions for L^AT_EX. Please observe the conference page limits.

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] S. W. P. M. Janosi, Andras and R. Detrano, “Heart Disease,” UCI Machine Learning Repository, 1989, DOI: <https://doi.org/10.24432/C52P4X>.

Tabela II
RESUMO ESTATÍSTICO DAS VARIÁVEIS NUMÉRICAS

Variável	Contagem	Média	Desvio Padrão	Mínimo	25%	Mediana	75%
age	297	54.54	9.05	29	48	56	
sex	297	0.68	0.47	0	0	1	
cp	297	3.16	0.96	1	3	3	
trestbps	297	131.69	17.76	94	120	130	
chol	297	247.35	52.00	126	211	243	
fbs	297	0.14	0.35	0	0	0	
restecg	297	1.00	0.99	0	0	1	
thalach	297	149.60	22.94	71	133	153	
exang	297	0.33	0.47	0	0	0	
oldpeak	297	1.06	1.17	0.0	0.0	0.8	
slope	297	1.60	0.62	1	1	2	

Tabela III
DISTRIBUIÇÃO DAS VARIÁVEIS CATEGÓRICAS COM PERCENTUAIS

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

APÊNDICE

Apêndice

Tabela IV
DESCRÍÇÃO 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	Tipos de dor torácica.	1 = Angina típica; 2 = Angina atípica; 3 = Dor não anginosa; 4 = Assintomática
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 \geq 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 tálolo.	3 = Normal; 6 = Defeito fixo; 7 = Defeito reversível
num	Categórica binária (alvo)	Diagnóstico de doença cardíaca.	0 = Sem doença significativa; 1 = Doença \geq 50%