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 Sao Paulo
Escola Politécnica
São Paulo, Brasil
email address or ORCID

Rampim Thiago
Universidade de Sao 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 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.

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 contém muitos dados faltantes, o que prejudicaria a confirmação ou descarte das hipoóteses 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 abaixo.

- **age**: Idade do paciente em anos. É uma variável numérica contínua que representa um dos principais fatores de risco para doenças cardíacas.
- **sex**: Sexo biológico do paciente, codificado como 1 = masculino e 0 = feminino. Usado para investigar diferenças de risco cardiovascular entre gêneros.
- **cp**: Tipo de dor torácica apresentada. Classificação:
  - 1: Angina típica
  - 2: Angina atípica
  - 3: Dor não anginosa
  - 4: Assintomático

Esse campo é importante para distinguir a origem cardíaca da dor torácica.

- **trestbps**: Pressão arterial de repouso medida na admissão hospitalar (em mmHg). Ajuda a avaliar hipertensão e sobrecarga cardiovascular.
- **chol**: Nível sérico de colesterol em mg/dL. Valor elevado indica risco aumentado de aterosclerose e doença coronariana.
- **fbs**: Açúcar no sangue em jejum superior a 120 mg/dL (1 = verdadeiro; 0 = falso). Utilizado como indicador indireto de resistência à insulina ou diabetes.
- **restecg**: Resultado do eletrocardiograma de repouso:
  - 0: Normal
  - 1: Anormalidade de onda ST-T (inversão de T ou elevação/depressão do segmento ST)
  - 2: Hipertrofia ventricular esquerda provável ou definitiva (critérios de Estes)

  Indica possíveis distúrbios elétricos ou estruturais cardíacos.
- **thalach**: Frequência cardíaca máxima alcançada durante o teste de esforço. Um bom preditor da capacidade funcional e da resposta cardiovascular ao exercício.
- **exang**: Presença de angina induzida por exercício (1 = sim; 0 = não). Reflete a ocorrência de isquemia durante o esforço físico.
- **oldpeak**: Depressão do segmento ST induzida pelo exercício em relação ao repouso. Mede a severidade da isquemia miocárdica.
- **slope**: Inclinação do segmento ST no pico do exercício:
  - 1: Ascendente
  - 2: Plana
  - 3: Descendente

  Associada ao prognóstico de doença arterial coronariana.
- **ca**: Número de principais vasos coronarianos (0 3) visualizados por fluoroscopia. Valores mais altos indicam maior comprometimento arterial.
- **thal**: Resultados do teste de tálio:
  - 3: Normal
  - 6: Defeito fixo
  - 7: Defeito reversível

  Avalia perfusão miocárdica e presença de áreas isquêmicas.
- **num**: Diagnóstico de doença cardíaca:
  - 0: Menos de 50% de estreitamento do diâmetro dos vasos
  - 1: Mais de 50% de estreitamento (doença significativa)

  É a variável alvo (dependente) usada para identificar presença de doença coronariana.

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 hipoóteses 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 LaTeX. Please observe the conference page limits.

REFERENCIAS

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.

APÊNDICE

Tabela II
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
sex	297	0.68	0.47	0	0	1	1	1
cp	297	3.16	0.96	1	3	3	4	4
trestbps	297	131.69	17.76	94	120	130	140	200
chol	297	247.35	52.00	126	211	243	276	564
fbs	297	0.14	0.35	0	0	0	0	1
restecg	297	1.00	0.99	0	0	1	2	2
thalach	297	149.60	22.94	71	133	153	166	202
exang	297	0.33	0.47	0	0	0	1	1
oldpeak	297	1.06	1.17	0.0	0.0	0.8	1.6	6.2
slope	297	1.60	0.62	1	1	2	2	3

Tabela III
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%