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.

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

II. Trabalhos Relacionados

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 [4].

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.

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**

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.

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 IV.

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

IV. 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.

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

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 estão descritos e resumidos na Tabela III.

Já os atribu tos categóricos, estão descritos na Tabela II

2) Visualização: Para a hipótese da influência da idade, na presença de doenças cardíacas, foi utilizada a análise de boxplot para os dados. A Figura 1 ilustra essa relação.

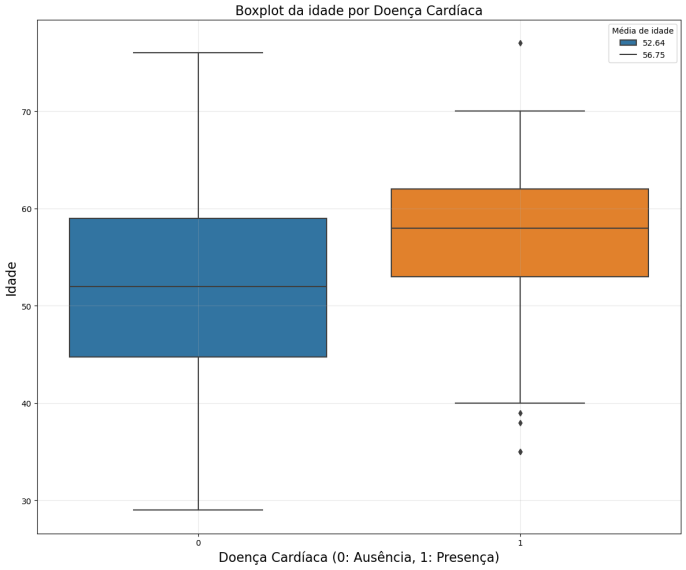

Figura 1. Distribuição da variável *age* (idade)

A relação entre a frequência máxima atingida e a presença ou ausência de doenças cardíacas, pode ser visto no $boxplot$ gerado na Figura 2

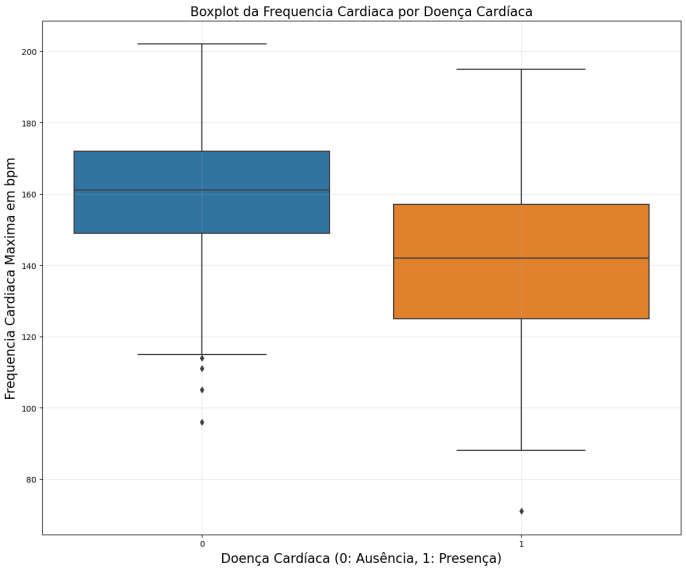

Figura 2. Distribuição da variável *trestbps* (frequência máxima)

Complementar às análises de boxplot, foi realizada a geração da matriz de correlação entre as variáveis. A Figura 3 demonstra a correlação entre as variáveis. Para essa análise, a variável alvo num também está contida na matriz de correlação.

VI. Resultados Obtidos

Comparar os resultados obtidos no modelo, com os de outros trabalhos.

Comparar e interpretar os resultados das questões analiticas.

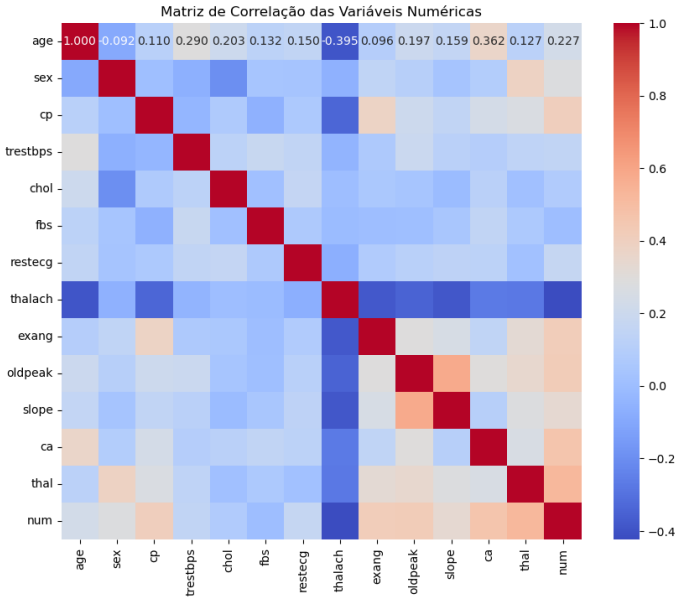

Figura 3. Matriz de correlação entre as variáveis.

Pela Figura **??**, nota-se uma idade média de 56,75 anos para a presença de doenças cardíacas e 52,64 anos para a ausência. Pode-se portanto, confirmar a primeira hipótese, de que a idade avançada contribui para doenças cardíacas.

VII. 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.

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] 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
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 III
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 IV
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

Tabela V
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 VI
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 VII
MEDIDAS DE PERFORMANCE POR MODELO

Métrica	**Discriminant**	**SVM**	**Tree**	**Bayes**	**Forest**
Accuracy	0.9901	0.7888	0.8350	0.9340	0.7657
Specificity	0.0000	0.0000	0.0000	0.0000	0.0000
Precision	0.9901	0.7888	0.8350	0.9340	0.7657
Recall	1.0000	1.0000	1.0000	1.0000	1.0000
F-measure	0.9950	0.8819	0.9101	0.9659	0.8673