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

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Resumo*—**

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

I. Introdução

As doenças cardíacas representam cerca de 32% das mortes no Brasil []

O objetivo do artigo é analisar a relação entre os fatores de risco associados à presença de doenças cardíacas e analisar 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 [8].

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

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**

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.

Identify applicable funding agency here. If none, delete this.

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

A. Armazenamento

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.

B. Segurança

Os dados da base orginal são públicos. Mas para o propósito acadêmico desse artigo, iremos definir regras de acesso, disponíveis na ferramenta AWS.

C. 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 **??**

- **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.

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

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.

References

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first . . ."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

Referências

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano. "Heart Disease,"UCI Machine Learning Repository, 1989. [Online]. Available: https://doi.org/10.24432/C52P4X.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove the template text from your paper may result in your paper not being published.