Aprendizagem Automática I

Projeto de Avaliação

-

Mestrado em Engenharia Informática

Universidade do Minho

Relatório

Grupo	
PG41081	José Alberto Martins Boticas
PG41091	Nelson José Dias Teixeira

10 de Janeiro de 2020

Resumo

Este projeto de avaliação relativo à unidade curricular de Aprendizagem Automática I consiste, globalmente, na aplicação de uma das técnicas abordadas durante as aulas sobre um conjunto de dados. O conjunto de dados mencionado previamente é escolhido sem qualquer tipo de restrição por parte dos elementos do grupo por forma a despoletar o interesse dos mesmos durante a análise estatística dos dados presentes. Como tal, durante a execução deste trabalho prático (cuja unidade curricular integra o perfil de Ciência de Dados), surge uma motivação extra na interpretação dos resultados obtidos.

Conteúdo

Capítulo 1

Introdução

1.1 Apresentação da base de dados escolhida

A base de dados escolhida pelos dois elementos que constituem este grupo diz respeito a registos clínicos ou hospitalares de *Cleveland* com informação relativa à presença ou ausência de doenças cardíacas. Nesta pode-se encontrar exatamente 14 atributos, como por exemplo a idade e o sexo de uma dada pessoa bem como registos da sua pressão arterial.

1.2 Contextualização

Apesar dos progressos consideráveis na luta contra as doenças cardiovasculares, estas continuam a ser a principal causa de morte na Europa. A alimentação inadequada é responsável por cerca de metade das mortes e da incapacidade causada pelas doenças cardiovasculares.

Em Portugal, cerca de 35 mil portugueses morrem anualmente por doenças cardiovasculares, que continuam a ser a principal causa de morte e representam um terço de toda a mortalidade da população, embora muitas dessas mortes e desse sofrimento prolongado pudessem ser evitados por uma mudança simples nos hábitos alimentares.

Com estes dados, os elementos que compõem este grupo ficaram sensibilizados acerca deste tema e mostraram-se interessados em investigar os fatores que de facto influenciam o aparecimento de doenças cardíacas.

1.3 Definição das variáveis

No que diz respeito às variáveis associadas a esta base de dados foi relativamente acessível executar a sua análise. Contudo, o significado dos nomes atribuídos a cada uma das mesmas não é de todo trivial. Desta forma, exibe-se de seguida de forma mais detalhada o que cada uma das variáveis representa:

- **age**: idade de uma determinada pessoa em anos;
- **sex**: sexo de uma determinada pessoa:
  - 1 : masculino;
  - 0 : feminino.
- **cp**: tipo de dor no peito (4 tipos):
  - 1 : angina típica;
  - 2 : angina atípica;

  - 3 : dor não anginal;
  - 4 : assintomática.
- **trestbps**: pressão arterial em repouso (em $mm\ Hg$ de uma determinada pessoa na admissão no hospital);
- **chol**: medição do colesterol de uma determinada pessoa em mg/dl;
- **fbs**: açúcar no sangue em jejum $> 120\ mg/dl$ da pessoa em causa:
  - 1 : verdadeiro;
  - 0 : falso.
- **restecg**: resultados eletrocardiográficos em repouso de uma pessoa:
  - 0 : normal;
  - 1 : com anormalidade da onda ST-T;
  - 2 : com provável ou definida hipertrofia ventricular esquerda pelo critério de Estes.
- **thalach**: frequência cardíaca máxima alcançada;
- **exang**: angina induzida pelo exercício:
  - 1 : sim;
  - 0 : não.
- **oldpeak**: depressão ST induzida por exercício em relação ao repouso;
- **slope**: a inclinação do segmento ST do pico do exercício:
  - 1 : ascendente;
  - 2 : plano (constante);
  - 3 : descendente.
- **ca**: número de vasos principais (0-3);
- **thal**: distúrbio sanguíneo denominado por talassemia:
  - 3 : normal;
  - 6 : defeito fixo;
  - 7 : defeito reversível.
- **target**: doença cardíaca:
  - 1 : sim;
  - 0 : não.

Ainda relativamente às incógnitas presentes na base de dados foi possível identificar tanto as variáveis quantitativas como as variáveis qualitativas ou categóricas. Apresenta-se de seguida o tipo de cada uma das variáveis presentes:

Nome da variável	Tipo de variável
age	Quantitativa discreta
sex	Qualitativa ordinal
cp	Qualitativa ordinal
trestbps	Quantitativa discreta
chol	Quantitativa discreta
fbs	Qualitativa ordinal
restecg	Qualitativa ordinal
thalach	Quantitativa discreta
exang	Qualitativa ordinal
oldpeak	Quantitativa contínua
slope	Qualitativa ordinal
ca	Quantitativa discreta
thal	Qualitativa ordinal
target	Qualitativa ordinal

1.4 Objetivo de análise

Perante esta base de dados, os elementos que integram este grupo procuram, com a informação disponível, prever se uma determinada pessoa possui ou não uma doença cardíaca. Consequentemente, a variável de interesse para realizar este estudo corresponde à incógnita *target*, pelo que o problema em causa é de classificação. Para além deste objetivo, vamos também perceber se existem outros fatores relativos ao coração que permitem prever certos eventos cardiovasculares.

Após traçarmos o objetivo principal na análise sobre a base de dados em causa, surgiram algumas questões da nossa parte. Apresentam-se de seguida as mesmas:

1. Quantas pessoas possuem uma doença cardíaca?
2. Que fatores influenciam o aparecimento de uma doença cardíaca?
3. ...

Consequentemente, por forma a responder a estas questões, é necessário especificar um modelo estatístico que se adequa a este contexto. Como tal, na próxima secção deste documento, é apresentado o modelo requerido.

Capítulo 2

Metodologia

Tal como o nome deste capítulo indica, nesta secção do documento será feita a exposição do modelo adotado pelo grupo para a análise correta dos dados presentes. À semelhança do que foi dito anteriormente, a variável de interesse presente neste conjunto de dados corresponde à incógnita ***target***. Uma vez que a mesma é uma variável qualitativa ordinal, temos que o problema em causa é um problema de **classificação**. Como tal, a técnica de regressão adotada para modelar esta base de dados é nada mais nada menos do que a **regressão logística**.

2.1 Especificação do modelo

Capítulo 3

Resultados

Neste capítulo do relatório são referidos os resultados principais da análise da base de dados escolhida. Como tal, foram escolhidas as tabelas e os gráficos que melhor traduzem os resultados obtidos.

Capítulo 4

Conclusão

Após a apresentação da metodologia e dos resultados obtidos durante análise estatística desta base de dados, dá-se por concluído a realização deste trabalho prático. Com o modelo implementado foi possível dar resposta às perguntas inicialmente traçadas, extraindo informação relevante para a compreensão e interpretação do contexto em causa. Consequentemente, foi também possível alcançar os objectivos delineados na introdução deste documento. De notar também que apesar da dimensão reduzida desta base de dados foi possível especificar um modelo adequado para o efeito.

Capítulo 5

Webgrafia

- *Website* indicado pela docente:
  https://www.kaggle.com/datasets
- *Website* com o resumo da base de dados escolhida:
  https://www.kaggle.com/ronitf/heart-disease-uci
- *Website* com a informação oficial da base de dados escolhida:
  https://archive.ics.uci.edu/ml/datasets/Heart+Disease