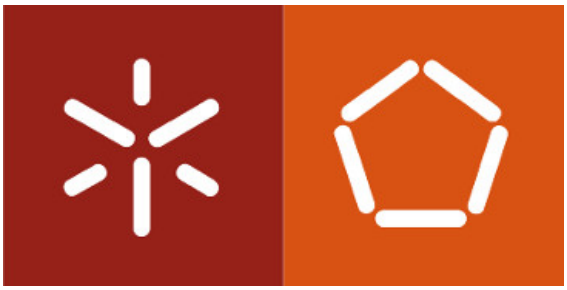

Universidade do Minho

Mestrado Integrado em Engenharia Informática

Engenharia do Conhecimento

Descoberta do Conhecimento

Resolução da Ficha de Exercícios 04

Gonçalo Pinto, A83732

João Diogo Mota, A80791

José Gonçalo Costa, PG42839

José Nuno Costa, A84829

10 de abril de 2021

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

No 2º semestre do 1º ano do Mestrado em Engenharia Informática da Universidade do Minho, existe uma unidade curricular enquadrada no perfil de Engenharia do Conhecimento denominada por Descoberta do Conhecimento, que tem como objetivo a introdução ao conceito de descoberta do conhecimento.

A presente ficha de exercícios enquadra-se na unidade curricular de Descoberta do Conhecimento e tem como objetivo criar competências aos alunos nessa mesma área.

Nesta ficha pretende-se que cada grupo identifique o problema proposto, execute os processos de *Data Understanding, Data Preprocessing, Modeling e Evaluation* enquadrados na metodologia do *CRISP-DM* e por fim que se crie um resumo de todo o trabalho executado.

O CRISP-DM é, no fundo, uma metodologia financiada pela Comunidade Europeia para suportar processos de *Data Mining*. Surgiu com o objetivo de encorajar a utilização de ferramentas interoperáveis ao longo de todo o processo de *Data Mining* de forma a retirar conhecimento valioso durante este processo.

De forma ao seu desenvolvimento foi utilizado uma dessas ferramentas, a escolha recaiu no *software Weka (Waikato Environment for Knowledge Analysis)* que possui uma coleção de algoritmos de *Machine Learning* para execução de tarefas de *Data Mining*. É um software que permite pré-processar grandes volumes de dados, aplicar diferentes algoritmos de *Machine Learning* e comparar vários outputs.

Para este efeito, foi fornecido um conjunto de dados pela equipa docente sobre doenças cardíacas obtido no repositório da UCI. Este *dataset* descreve fatores de risco para doenças cardíacas. O atributo num representa o atributo da classe (binária):

- $class < 50$ - nenhuma doença ;
- $class > 50\_1$ - aumento do nível de doença cardíaca .

O principal objetivo deste problema é prever doenças cardíacas a partir de outros atributos existentes no *dataset*. Obviamente, trata-se de um problema de classificação. A descrição deste exercício é gradual. Portanto, esperamos entender melhor os vários aspetos e questões envolvidos no processo de *KDD* (extração de conhecimento).

Capítulo 2

Data Understanding

A primeira etapa do processo de *Data Mining* passa pelo processo de *Data Understanding*, que tem como objetivo a familiarização com os dados.

Para esta fase, além do uso da ferramenta *WEKA*, recorreu-se à documentação do *dataset* [1] de forma a entender os diversos atributos presentes com o intuito de conseguir obter conclusões mais concretas sobre o problema em questão.

- **[1]** ***Para cada atributo, foi possível encontrar as seguintes informações:***
  - **a)** O tipo (p.e. nominal, ordinal, numérico.) de cada atributo está descrito de seguida:
    - *age* - Numérico
    - *sex* - Nominal
    - *cp* - Nominal
    - *trestbps* - Numérico
    - *chol* - Numérico
    - *fbs* - Nominal
    - *restecg* - Nominal
    - *thalach* - Numérico
    - *exang* - Nominal
    - *oldpeak* - Numérico
    - *slope* - Nominal
    - *ca* - Numérico
    - *thal* - Nominal
    - *num* - Nominal
  - **b)** As percentagens de valores ausentes nos dados são as seguintes:
    - *age* = 0 %
    - *sex* = 0 %
    - *cp* = 0 %
    - *trestbps* = 0 %
    - *chol* = 0 %
    - *fbs* = 0 %
    - *restecg* = 0 %
    - *thalach* = 0 %
    - *exang* = 0 %
    - *oldpeak* = 0 %
    - *slope* = 0 %
    - *ca* = 2 % (5 registos)
    - *thal* = 1 % (2 registos)
    - *num* = 0 %

– **c)** Os valores de máximo, mínimo, média e desvio padrão de cada atributo quando aplicável são os apresentados na tabela abaixo:

	Min.	Máx.	Média	Desvio Padrão
age	29	77	54.366	9.082
sex	Não Aplicável	Não Aplicável	Não Aplicável	Não Aplicável
cp	Não Aplicável	Não Aplicável	Não Aplicável	Não Aplicável
trestbps	94	200	131.624	17.538
chol	126	564	246.264	51.831
fbs	Não Aplicável	Não Aplicável	Não Aplicável	Não Aplicável
restecg	Não Aplicável	Não Aplicável	Não Aplicável	Não Aplicável
thalach	71	202	149.647	22.905
exang	Não Aplicável	Não Aplicável	Não Aplicável	Não Aplicável
oldpeak	0	6.2	1.04	1.161
slope	Não Aplicável	Não Aplicável	Não Aplicável	Não Aplicável
ca	0	3	0.674	0.938
thal	Não Aplicável	Não Aplicável	Não Aplicável	Não Aplicável
num	Não Aplicável	Não Aplicável	Não Aplicável	Não Aplicável

Figura 2.1: Os valores de máx, min, média e desvio padrão dos atributos.

– **d)** No campo *Unique* de cada atributo podemos visualizar o número de registos que tenham um valor para um atributo que nenhum outro registo tem.

1. *age* : 4 registos únicos (1 %)
2. *sex* : 0 registos únicos
3. *cp* : 0 registos únicos
4. *trestbps* : 16 registos únicos (5 %)
5. *chol* : 62 registos únicos (20 %)
6. *fbs* : 0 registos únicos
7. *restecg* : 0 registos únicos
8. *thalach* : 28 registos únicos (9 %)
9. *exang* : 0 registos únicos
10. *oldpeak* : 10 registos únicos (3 %)
11. *slope* : 0 registos únicos
12. *ca* : 0 registos únicos
13. *thal* : 0 registos únicos
14. *num* : 0 registos únicos

– **e)** Um histograma é a representação gráfica em colunas ou em barras de um conjunto de dados previamente tabulado e dividido em classes uniformes ou não uniformes. Na figura 2.2 está apresentado um exemplo de histograma que é apresentado na ferramenta utilizada, este relaciona o atributo *age* e o atributo que permite representar se uma pessoa que apresenta ou não doença (*num*). Os valores apresentados a vermelho correspondem doença e a azul que não existe doença. Desta forma, passando o cursor pelo histograma é possível observar umas mensagens pop-up que significam para um atributo nominal o número de ocorrências desse registo e nos atributos numéricos aparece o intervalo de valores juntamente com o número de ocorrências dentro desse mesmo intervalo. De seguida apresentados algumas conclusões retiradas dos histogramas de cada atributo sobre a influência de cada no risco de doença cardíaca.

  * *age*, é possível verificar que com o aumentar da idade também aumenta o risco de doença. Havendo risco máximo entre os 57 e os 63 anos;
  * *sex*, relativamente ao sexo, existem muitos mais casos de doença em pessoas do sexo masculino do que feminino;
  * *cp*, neste atributo, relativo ao tipo de dor no peito, é difícil a análise pois o número de pessoas analisadas com *assympt* (assintomáticas) é bastante considerável;
  * *trestbps*, possível ver uma distribuição não muito uniforme impossibilitando uma análise concreta, podemos também observar que as pessoas com mais que 147 mm Hg com doença são cada vez menores;
  * *chol*, à medida que o valor de colesterol aumenta, é possível ver que o número de pessoas com doença vai aumentando em relação com as pessoas sem doença;
  * *fbs*, este atributo encontra-se distribuído uniformemente, logo não foi possível retirar nenhuma influência deste atributo no risco de doença cardíaca;
  * *restecg*, de igual forma ao atributo *fbs*, este atributo referente à dor aliviada após descanso, possui uma distribuição idêntica para qualquer das respostas;
  * *thalach*, relativamente ao batimento cardíaco, à medida que o *heart rate* aumenta, o número de pessoas também aumenta, até atingir um pico entre os 154 e os 166, a partir daí, o *heart rate* e o número de pessoas também diminui;
  * *exang*, este atributo refere-se à dor angina causada por exercício, é menos frequente encontrar pessoas com esta dor e terem doença cardíaca;
  * *oldpeak*, quanto à depressão de ST induzida por exercício em relação ao repouso, a percentagem de pessoas com doença cardíaca aumenta à medida que o valor vai aumentando;
  * *slope*, relativamente à inclinação do segmento ST de pico do exercício, verifica-se uma maior percentagem de pessoas com doença cardíaca quando a inclinação é "*flat*"ou "*up*";
  * *ca*, este atributo representa o número de vasos principais coloridos por fluorosopia. Verifica-se uma maior quantidade de pessoas nos valores entre 0.5 e 1, aumentando a percentagem de pessoas com doença cardíaca à medida que o valor vai aumentando;
  * *thal*, com o valor "reversable_defect" apresentam-se mais classes com patologia do que sem;
  * *num*, através da análise do histograma, no *dataset* fornecido existe um total de 165 pessoas sem doença cardíaca e 138 sem.

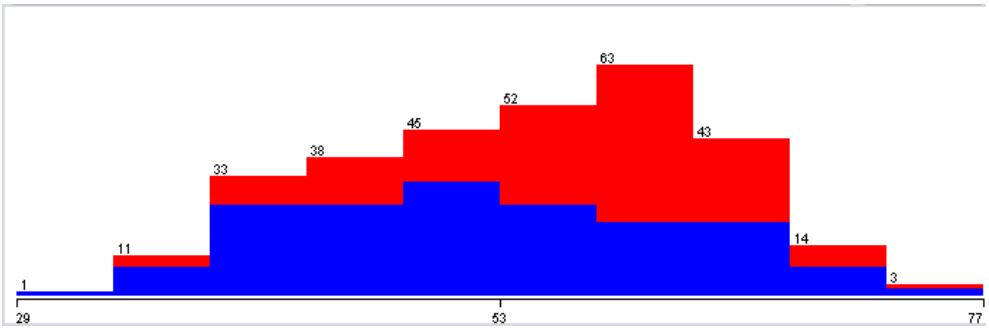

Figura 2.2: Histograma do atributo *age*.

- **[2]** ***Utilizando o separador Visualize, na parte superior da janela, para visualizar gráficos de dispersão 2D para cada par de atributos. Foi possível retirar as seguintes conclusões.***
  - **a)** Na tabela abaixo podemos observar as descobertas do grupo em relação a cada atributo e a sua associação às doenças cardíacas, estas foram baseadas na concentração de valores que representam doença no gráfico de dispersão que relaciona um determinado atributo com ele mesmo.

Atributo	**Maior número de doenças cardíacas**
age	A partir dos 53
sex	Masculino
cp	asympt
trestbps	Não se conseguiu identificar alguma associação
chol	Não se conseguiu identificar alguma associação
fbs	Não se conseguiu identificar alguma associação
restecg	Não se conseguiu identificar alguma associação
thalach	Abaixo dos 136.5
exang	yes
oldpeak	A partir de 3.1
slope	Não se conseguiu identificar alguma associação
ca	Próximo de 1.5
thal	Registos fora do normal

Figura 2.3: Descobertas do grupo em relação a cada atributo e a sua associação às doenças cardíacas.

  - **b)** Tendo em conta os gráficos o par de atributos que parece estar correlacionado é o atributo *age* e o atributo *thalach*.

- **[3]** ***Possíveis associações multivariadas de atributos com o atributo class.***

  De forma a encontrar possíveis associações multivariadas de atributos foi estudado os gráficos de dispersão de dois atributos X e Y e tentou-se identificar possíveis áreas "densas" de doenças cardíacas. Assim foi possível concluir que o par de atributos que parecem estar mais associados é o atributo *age* com o atributo *thalach*, a sua relação pode ser observada na figura 2.4 onde no eixo do X é apresentado o atributo *age* e no eixo do Y o atributo *thalach*. Estes valores parecem estar associados pois à medida que a frequência cardíaca máxima alcançada diminui e existe um aumento da idade, o que leva a uma maior incidência do número de registos com doença cardíaca.

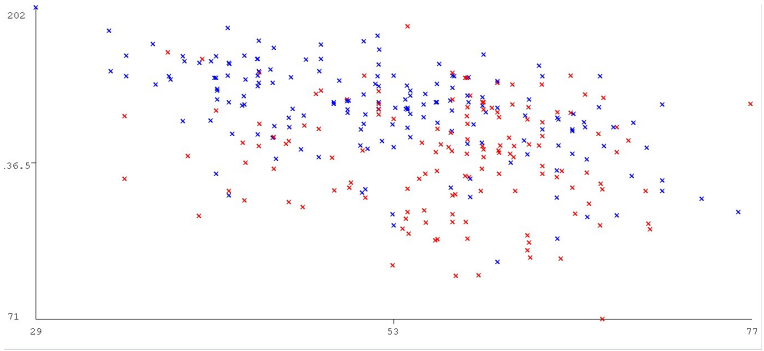

Figura 2.4: Associações multivariadas de atributos.

Capítulo 3

Data Processing

A segunda etapa do processo de *Data Mining* passa pelo processo de *Data Processing*, que diz respeito ao processamento dos dados de modo a que os dados transformados estejam numa forma mais adequada para os algoritmos de *Data Mining*.

- **[1] *Seleção de atributos.***
  De forma a efetuar a seleção de atributos foi utilizado o filtro Weka *AttributeSelection* para selecionar um sub-conjunto de atributos pois este filtro fornece uma boa capacidade de previsão. Assim foi apenas apresentado os atributos *cp, restecg, thalach, exang, oldpeak, ca, thal*. Tendo em conta os resultados obtidos na secção anterior, podemos observar que a maioria dos atributos que foram filtrados foram de encontrando à figura 2.3, isto é, os valores filtrados foram aqueles que o grupo descobriu que existe uma maior associação às doenças cardíacas. Foi guardado o conjunto de dados com os atributos selecionados no ficheiro *heart-c1.arff*.

- **[2] *Lidar com valores ausentes.***
  Considerando os métodos para lidar com valores ausentes no software utilizado, decidiu-se tal como foi recomendado que estes registos não fossem eliminados mas sim que fosse atribuído valores onde faltam dados, usando um método adequado.

  - **a)** Substituição dos valores ausentes pela média do atributo, se o atributo for numérico. Caso contrário, foi substituído os valores ausentes pela moda do atributo (se o atributo for nominal). Foi guardado o conjunto de dados sem valores ausentes no ficheiro *heart-c2.arff*.

  - **b)** De forma a estimar os valores ausentes para cada atributo utilizando regressão (linear) foi aplicado o algoritmo *LinearRegression* disponibilizado pelo *Weka* no separador *Classify*, como existe apenas um atributo numérico com valores ausentes apenas foi aplicado este algoritmo a este atributo. Assim foi obtido a seguinte expressão que permite calcular o valor do atributo:

    $$ca =$$
    $$0.0261 * \text{age} + 0.3133 * \text{fbs=t} + 0.1479 * \text{oldpeak} -$$
    $$0.6343 * \text{slope=down,flat} + 0.4272 * \text{slope=flat} + 0.7098 * \text{num=>50\_1} - 1.138$$

    Após aplicado a fórmula acima apresentada foi guardado o conjunto de dados sem valores ausentes no ficheiro *heart-c3.arff*.

- **[3] *Eliminar outliers.***
  Com o objetivo de eliminar os registos discrepantes foi necessário remover os atributos nominais de forma poder utilizar o filtro *InterquartileRange (Unsupervised - Attribute)* disponibilizado pelo Weka para detetar outliers, posteriormente foi retornado a instância que era outlier que possui no atributo *chol* o valor 564.0. Posteriormente aplicou-se um outro filtro denominado *RemoveWithValues (Unsupervised - Instance)* no sentido de o remover, para isso configurou-se os parâmetros *attributeIndex e nominalIndices* onde o primeiro refere-se ao índice na lista de atributos e o segundo a coluna que queremos remover no outlier (neste caso o outlier possui o valor yes). Foi guardado o conjunto de dados com os atributos selecionados no ficheiro *heart-c4.arff*.

Capítulo 4

Modeling

A terceira etapa do processo de *Data Mining* passa pelo processo de *Modeling*, nesta fase pretende-se usar algoritmos de classificação disponíveis no Weka para descobrir padrões ocultos nos dados. Foi repetido as etapas descritas abaixo para cada um dos conjuntos de dados criados durante o pré-processamento, além de que foi usado também o *dataset* original.

- **[1] *Classificador OneR.***
  - **a)** Aplicado o classificador **OneR** através do método 10 *fold-cross validation* para cada conjunto de dados criados previamente foi obtido os seguintes resultados, as diferenças não foram muito notórias entre eles pois tanto os valores ausentes nos atributos, que foram substituídos pela média/moda ou usando a regressão linear, como os *outliers* são poucos.
    - *dataset original*, foram classificadas corretamente 217 instâncias e 86 incorretamente. Foi possível concluir que o atributo *thal* é o que permite obter um número semelhante ao do atributo da classe ao nível de instâncias corretas e incorretas.
    - *heart-c1*, os resultados obtidos foram semelhantes aos encontrados com *dataset original*.
    - *heart-c2*, foram classificadas corretamente 218 instâncias e 85 incorretamente.
    - *heart-c3*, foram classificadas corretamente 224 instâncias e incorretamente 79 instâncias.
    - *heart-c4*, foram classificadas corretamente 219 instâncias e 83 incorretamente (menos uma instância classificada devido a ter sido removida por ser *outlier*).
  - **b)** Aplicado o classificador **OneR** com conjunto de treino (training set) a cada um dos conjuntos de dados criados previamente foi possível observar que os resultados foram os mesmos ao nível do número de instâncias corretas 232 (equivalendo a uma percentagem de 76.5677 %), comparando com a precisão obtida através do método 10 *fold-cross validation* utilizar o conjunto de treino aumentou a precisão do classificador em cerca de 5 %. Esta diferença pode ser explicada pelo fator de que à medida que se executam os testes, existir uma aprendizagem por parte do algoritmo, fazendo com que se alcancem melhores precisões.

- **[2] *Classificador JRip, ou seja, a versão Weka do classificador de regras RIPPER***
  - **a)** Aplicado o classificador com e sem *rule pruning* para cada conjunto de dados criados previamente foi possível observar o seguinte comportamento.
    * *dataset original*:
      · *Com pruning* = 81.5182 % instâncias corretas e 18.4818 % instâncias incorretas.
      · *Sem pruning* = 77.5578 % instâncias corretas e 22.4422 % instâncias incorretas.
    * *heart-c1*:
      · *Com pruning* = 80.198 % instâncias corretas e 19.802 % instâncias incorretas.
      · *Sem pruning* = 78.2178 % instâncias corretas e 21.7822 % instâncias incorretas.
    * *heart-c2*:
      · *Com pruning* = 81.8482 % instâncias corretas e 18.1518 % instâncias incorretas.
      · *Sem pruning* = 77.8878 % instâncias corretas e 22.1122 % instâncias incorretas.
    * *heart-c3*:
      · *Com pruning* = 77.5578 % instâncias corretas e 22.4422 % instâncias incorretas.
      · *Sem pruning* = 74.5875 % instâncias corretas e 25.4125 % instâncias incorretas.
    * *heart-c4*:
      · *Com pruning* = 77.1523 % instâncias corretas e 22.8477 % instâncias incorretas.
      · *Sem pruning* = 76.8212 % instâncias corretas e 23.1788 % instâncias incorretas.

    Pelos valores apresentados anteriormente podemos observar que para este classificador a melhor opção é utilizar *rule pruning*.

- **[3] *Classificador J48, ou seja, a versão Weka do classificador C4.5 da árvore de decisão.***
  - **a)** No software utilizado foi explorado o uso de diferentes parâmetros no algoritmo J48, como a existência e ausência de *pruning("unpruned")* e variar o número mínimo de registos nas folhas(*"minNumObj"*) utilizando para isso 2, 5 e 10 registos mínimos nas folhas.
  - **b)** De seguida iremos descrever os padrões que obtivemos e a respetiva comparação com os resultados obtidos nas questões anteriores:
    * *dataset original*:
      - *unpruned = true* :
        + *minNumObj = 2* : 77.8878% corretas e 22.1122% incorretas.
        + *minNumObj = 5* : 77.5578% corretas e 22.4422% incorretas.
        + *minNumObj= 10* : 76.5677% corretas e 23.4323% incorretas.
      - *unpruned = false* :
        + *minNumObj = 2* : 77.8878% corretas e 22.1122% incorretas.
        + *minNumObj = 5* : 76.8977% corretas e 23.1023% incorretas.
        + *minNumObj= 10* : 75.9076% corretas e 24.0924% incorretas.

      Para o *dataset original* o uso do atributo *unpruned* não interfere na percentagem de instâncias corretas e incorretas e em relação ao número mínimo de registos nas folhas verificou-se que é 2.

    * *heart-c1*:
      - *unpruned = true* :
        + *minNumObj = 2* : 77.8878% corretas e 22.1122% incorretas.
        + *minNumObj = 5* : 78.2178% corretas e 21.7822% incorretas.
        + *minNumObj= 10* : 76.8977% corretas e 23.1023% incorretas.
      - *unpruned = false* :
        + *minNumObj = 2* : 78.5479% corretas e 21.4521% incorretas.
        + *minNumObj = 5* : 76.2376% corretas e 23.7624% incorretas.
        + *minNumObj= 10* : 77.2277% corretas e 22.7723% incorretas.

      Conclui-se que se conseguem extrair os melhores valores quando se usa o atributo *unpruned=true* com o número médio de folhas (5) e quando se usa *unpruned=false* com o número baixo de folhas(2).

    * *heart-c2*:
      - *unpruned = true* :
        + *minNumObj = 2* : 75.5776% corretas e 24.4224% incorretas.
        + *minNumObj = 5* : 74.9175% corretas e 25.0825% incorretas.
        + *minNumObj= 10* : 78.2178% corretas e 21.7822% incorretas.
      - *unpruned = false* :
        + *minNumObj = 2* : 77.2277% corretas e 22.7723% incorretas.
        + *minNumObj = 5* : 76.2376% corretas e 23.7624% incorretas.
        + *minNumObj= 10* : 77.8878% corretas e 22.1122% incorretas.

      Para o *heart-c2* consegue-se extrair os melhores valores quando se usa o atributo *unpruned=true* com o número elevado de folhas (10).

* *heart-c3*:

  · *unpruned = true* :

  + *minNumObj = 2* : 76.2376% corretas e 23.7624% incorretas.

  + *minNumObj = 5* : 76.5677% corretas e 23.4323% incorretas.

  + *minNumObj= 10* : 76.2376% corretas e 23.7624% incorretas.

  · *unpruned = false* :

  + *minNumObj = 2* : 76.8977% corretas e 23.1023% incorretas.

  + *minNumObj = 5* : 77.5578% corretas e 22.4422% incorretas.

  + *minNumObj= 10* : 75.9076% corretas e 24.0924% incorretas.

  Conclui-se que se conseguem extrair os melhores valores quando se usa o atributo *unpruned=false* com o número médio de folhas (5).

* *heart-c4*:

  · *unpruned = true* :

  + *minNumObj = 2* : 75.1656% corretas e 24.8344% incorretas.

  + *minNumObj = 5* : 75.1656% corretas e 24.8344% incorretas.

  + *minNumObj= 10* : 74.8344% corretas e 25.1656% incorretas.

  · *unpruned = false* :

  + *minNumObj = 2* : 74.8344% corretas e 25.1656% incorretas.

  + *minNumObj = 5* : 76.8212% corretas e 23.1788% incorretas.

  + *minNumObj= 10* : 75.1656% corretas e 24.8344% incorretas.

  Para o *heart-c4* consegue-se extrair os melhores valores quando se usa o atributo *unpruned=false* com o número médio de folhas (5).

- **[4] *Exploração com outros algoritmos de classificação.***

	NaiveBayes - bayes	ZeroR - rules	HoeffdingTree - trees	LWL - lazy
heart-c	83.50%	54.46%	83.83%	73.60%
heart-c1	84.49%	54.46%	83.83%	71.62%
heart-c2	82.84%	54.46%	83.83%	74.59%
heart-c3	83.50%	54.46%	83.83%	73.60%
heart-c4	83.11%	54.30%	83.11%	73.51%

Tabela 4.1: Tabela da exploração com outros algoritmos de classificação.

Capítulo 5

Evaluation

Na quinta e última etapa do processo de *Data Mining* consiste no processo de *Evaluation*, passa por comparar os diferentes modelos obtidos no processo de *Modeling* e obter algumas conclusões.

- **[1]** ***O Weka oferece várias medidas de avaliação de desempenho.***
  O *Weka* fornece várias medidas de avaliação de desempenho, de entre as quais se destacam:
  - Número (Percentagem) de instâncias corretamente classificadas, indica o valor (percentagem) de instâncias que o classificador conseguir identificar corretamente;
  - Número (Percentagem) de instâncias incorretamente classificadas, indica o valor (percentagem) de instâncias que o classificador conseguir identificar incorretamente;
  - *Confusion Matrix*, é uma matriz de valores inteiros, que para cada tipo de classificação, indica o número de vezes que se classificou uma instância com esse mesmo tipo;
  - Número total de instâncias analisadas, indica o caso de estudo;

  O grupo destacou estas medidas de desempenho pois são as principais utilizadas e consideradas durante o processo de *Data Mining*.

- **[2]** ***Tabela com as medidas de desempenho para cada classificador e cada dataset.***

Medida / Classificador	**JRip - Com pruning**	**NaiveBayes - bayes**	**HoeffdingTree - trees**
Instâncias corretamente classificadas(%)	81.5182%	83.50%	83.83%
Instâncias incorretamente classificadas(%)	18.4818%	16.50%	16,17%
Confusion Matrix	Figura 7.1	Figura 7.2	Figura 7.3
Nº total de instâncias	303	303	303

Tabela 5.1: Medidas de desempenho para os melhores 3 classificadores do *dataset* original.

Medida / Classificador	**JRip - Com pruning**	**NaiveBayes - bayes**	**HoeffdingTree - trees**
Instâncias corretamente classificadas(%)	80.198%	84.49%	83.83%
Instâncias incorretamente classificadas(%)	19.802%	15.51%	16,17%
Confusion Matrix	Figura 7.4	Figura 7.5	Figura 7.6
Nº total de instâncias	303	303	303

Tabela 5.2: Medidas de desempenho para os melhores 3 classificadores do *dataset heart-c1*.

Medida / Classificador	JRip - Com pruning	NaiveBayes - bayes	HoeffdingTree - trees
Instâncias corretamente classificadas(%)	81.8482%	82.84%	83.83%
Instâncias incorretamente classificadas(%)	18.1518%	17.16%	16,17%
Confusion Matrix	Figura 7.7	Figura 7.8	Figura 7.9
Nº total de instâncias	303	303	303

Tabela 5.3: Medidas de desempenho para os melhores 3 classificadores do *dataset heart-c2*.

Medida / Classificador	JRip - Com pruning	NaiveBayes - bayes	HoeffdingTree - trees
Instâncias corretamente classificadas(%)	77.5578%	83.50%	83.83%
Instâncias incorretamente classificadas(%)	22.4422%	16.5%	16,17%
Confusion Matrix	Figura 7.10	Figura 7.11	Figura 7.12
Nº total de instâncias	303	303	303

Tabela 5.4: Medidas de desempenho para os melhores 3 classificadores do *dataset heart-c3*.

Medida / Classificador	JRip - Com pruning	NaiveBayes - bayes	HoeffdingTree - trees
Instâncias corretamente classificadas(%)	77.1523%	83.11%	83.11%
Instâncias incorretamente classificadas(%)	22.8477%	16.89%	16,89%
Confusion Matrix	Figura 7.13	Figura 7.14	Figura 7.15
Nº total de instâncias	302	302	302

Tabela 5.5: Medidas de desempenho para os melhores 3 classificadores do *dataset heart-c4*.

- **[3]** ***Conclusões***

  Conclui-se que para o classificador *JRip - Com pruning*, o *dataset heart-c2* apresenta o melhor resultado em termos de instâncias corretamente classificadas. Em relação ao classificador *NaiveBayes*, o *dataset heart-c1* apresenta o melhor resultado em termos de instâncias corretamente classificadas. Por fim, em relação ao classificador *HoeffdingTree* verificou-se que o número de instâncias corretamente classificadas foi aproximadamente igual em todos os *datasets*.

Capítulo 6

Conclusão

O presente relatório descreveu, de forma sucinta, o processo de *Data Mining* na execução de um problema sobre doenças cardíacas cujo objetivo era prever doenças cardíacas a partir de outros atributos existentes no *dataset* disponibilizado.

Após a realização deste problema, foi possível compreender a importância dos processos de *Data Understanding, Data Processing, Modeling* e *Evaluation* sobre os vários aspetos e questões envolvidos no processo de *KDD* (extração de conhecimento). Também foi possível aprofundar o conhecimento de diversos algoritmos de *Data Mining* e a comparação entre estes.

Por fim, esperamos que os conhecimentos obtidos e consolidados sejam de enorme utilidade tendo uma perspetiva futura.

Bibliografia

[1] *uci machine learning repository: heart disease data set*$_2$021. 2021. URL: https://archive.ics.uci.edu/ml/datasets/heart+disease (ver p. 4).

Capítulo 7

Anexos

```
=== Confusion Matrix ===

   a   b   <-- classified as
 136  29 |   a = <50
  27 111 |   b = >50_1
```

Figura 7.1: *Confusion Matrix* para o algoritmo *JRip* sobre o *dataset* original.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 136  29 |   a = <50
  27 111 |   b = >50_1
```

Figura 7.2: *Confusion Matrix* para o algoritmo *NaiveBayes* sobre o *dataset* original.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 144  21 |   a = <50
  28 110 |   b = >50_1
```

Figura 7.3: *Confusion Matrix* para o algoritmo *HoeffdingTree* sobre o *dataset* original.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 142  23 |   a = <50
  37 101 |   b = >50_1
```

Figura 7.4: *Confusion Matrix* para o algoritmo *JRip* sobre o *dataset heart-c1.*

```
=== Confusion Matrix ===

   a   b   <-- classified as
 147  18 |   a = <50
  29 109 |   b = >50_1
```

Figura 7.5: *Confusion Matrix* para o algoritmo *NaiveBayes* sobre o *dataset heart-c1.*

```
=== Confusion Matrix ===

   a   b   <-- classified as
 145  20 |   a = <50
  29 109 |   b = >50_1
```

Figura 7.6: *Confusion Matrix* para o algoritmo *HoeffdingTree* sobre o *dataset heart-c1.*

```
=== Confusion Matrix ===

   a   b   <-- classified as
 138  27 |   a = <50
  28 110 |   b = >50_1
```

Figura 7.7: *Confusion Matrix* para o algoritmo *JRip* sobre o *dataset heart-c2.*

```
=== Confusion Matrix ===

   a   b   <-- classified as
 142  23 |   a = <50
  29 109 |   b = >50_1
```

Figura 7.8: *Confusion Matrix* para o algoritmo *NaiveBayes* sobre o *dataset heart-c2.*

```
=== Confusion Matrix ===

   a   b   <-- classified as
 144  21 |   a = <50
  28 110 |   b = >50_1
```

Figura 7.9: *Confusion Matrix* para o algoritmo *HoeffdingTree* sobre o *dataset heart-c2.*

```
=== Confusion Matrix ===

   a   b   <-- classified as
 134  31 |   a = <50
  37 101 |   b = >50_1
```

Figura 7.10: *Confusion Matrix* para o algoritmo *JRip* sobre o *dataset heart-c3.*

```
=== Confusion Matrix ===

   a   b   <-- classified as
 144  21 |   a = <50
  29 109 |   b = >50_1
```

Figura 7.11: *Confusion Matrix* para o algoritmo *NaiveBayes* sobre o *dataset heart-c3.*

```
=== Confusion Matrix ===

   a   b   <-- classified as
 144  21 |   a = <50
  28 110 |   b = >50_1
```

Figura 7.12: *Confusion Matrix* para o algoritmo *HoeffdingTree* sobre o *dataset heart-c3.*

```
=== Confusion Matrix ===

   a   b   <-- classified as
 139  25 |   a = <50
  44  94 |   b = >50_1
```

Figura 7.13: *Confusion Matrix* para o algoritmo *JRip* sobre o *dataset heart-c4.*

```
=== Confusion Matrix ===

   a   b   <-- classified as
 142  22 |   a = <50
  29 109 |   b = >50_1
```

Figura 7.14: *Confusion Matrix* para o algoritmo *NaiveBayes* sobre o *dataset heart-c4.*

```
=== Confusion Matrix ===

   a   b   <-- classified as
 143  21 |   a = <50
  30 108 |   b = >50_1
```

Figura 7.15: *Confusion Matrix* para o algoritmo *HoeffdingTree* sobre o *dataset heart-c4.*