Universidade do Minho
Departamento de Informática

FE04

Curso: Mestrado Integrado em Engenharia Biomédica – Informática Médica
U.C.: Sistemas de Aprendizagem e Extração do Conhecimento

	Folha de Exercícios FE04
Docente	Diana Ferreira
Tema	Explorar o Weka
Turma	PL
Ano Letivo	2019-20 – 2º Semestre
Duração da aula	2 horas

1. Descrição do Problema

O *dataset* usado neste exercício é o *dataset* de doenças cardíacas disponível no ficheiro heart-c.arff, 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 significa nenhuma doença; class > 50_1 indica aumento do nível de doença cardíaca. O principal objetivo deste exercício é prever doenças cardíacas a partir de outros atributos no *dataset*. Obviamente, trata-se de um problema de classificação. O *software* a ser usado é o Weka. No entanto, sinta-se à vontade para tentar qualquer ideia que possa ter para solucionar o problema com qualquer outro *software*. A descrição deste exercício é gradual. Portanto, espera-se que possa entender melhor os vários aspetos e questões envolvidos no processo de KDD.

1.1. Data Understanding

O primeiro passo para abordar o problema é familiarizar-se com os dados. Responder às seguintes perguntas ajudará a entender melhor os dados. O *dataset* heart-c.arff contém algumas informações sobre os dados que armazena. Pode abri-lo num editor de texto. Carregue o *dataset* no Weka.

[1] Para cada atributo, encontre as seguintes informações (pode colocar as respostas numa tabela):

[a] O tipo de atributo, p.e. nominal, ordinal, numérico.

[b] Percentagem de valores ausentes nos dados.

[c] Máx, min, média e desvio padrão.

[d] Existem registos que tenham um valor para um atributo que nenhum outro registo tem?

[e] Estude o histograma no canto inferior direito e descreva informalmente como o atributo parece influenciar o risco de doença cardíaca. O que significam as mensagens *pop-up* que aparecem ao arrastar o rato sobre o gráfico?

[2] Mude para o separador *Visualize,* na parte superior da janela, para visualizar gráficos de dispersão 2D para cada par de atributos. Que atributos parecem estar mais/menos associados a doenças cardíacas? Resuma numa tabela as suas descobertas sobre o valor preditivo de cada atributo.

1.2. Data Preprocessing

A segunda etapa diz respeito ao pré-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.

Investigue a possibilidade de usar o filtro Weka *AttributeSelection* para selecionar um subconjunto de atributos com boa capacidade de previsão. Em seguida, descreva brevemente o(s) filtro(s) usado(s) e compare os resultados obtidos com as conclusões obtidas na seção anterior. Guarde o conjunto de dados com os atributos selecionados no ficheiro *heart-c1.arff*.

[2] Lidar com valores ausentes.

Considere os seguintes métodos para lidar com valores ausentes e investigue cada possibilidade no Weka. Os registos não deverão ser eliminados e é aconselhável atribuir valores onde faltam dados, usando um método adequado.

> [a] Substitua os valores ausentes pela média do atributo, se o atributo for numérico. Caso contrário, substitua os valores ausentes pela moda do atributo (se o atributo for nominal). Use o filtro ReplaceMissingValues para fazer esta transformação. Guarde o conjunto de dados que obteve sem valores ausentes no ficheiro *heart-c2.arff*.
>
> [b] Investigue a possibilidade de usar regressão linear com *crossvalidation* de *10 folds* para estimar os valores ausentes para cada atributo (se possível). Note que a regressão linear apenas pode ser aplicada a atributos numéricos. Apresente a equação obtida e estime os valores em falta através dessa equação. Guarde o conjunto de dados que obteve sem valores ausentes no ficheiro heart-c3.arff.

[3] Eliminar *outliers*.

Elimine os registos discrepantes e guarde o conjunto de dados obtido sem *outliers* no ficheiro *heart-c34.arff*. Investigue a possibilidade de usar o filtro Weka – Unsupervised – Attribute – InterquartileRange para detetar *outliers* e o filtro Weka – Unsupervised – Instance – RemoveWithValues para eliminar *outliers* (não se esqueça de configurar os parâmetros *attributeIndex*, que diz respeito ao índice do outlier, e *nominalIndices*, que corresponde à localização (first ou last) do valor nominal do atributo que se pretende remover).

1.3. Data Mining

O terceiro passo é usar algoritmos de classificação disponíveis no Weka para descobrir padrões ocultos nos dados. Deve repetir as etapas descritas abaixo para cada um dos conjuntos de dados criados durante o pré-processamento, além de usar também o *dataset* original.

[1] Comece com o classificador OneR.

(a) O que pode concluir? Compare as suas conclusões com as conclusões que obteve na seção 1.1.

(b) Compare a precisão do classificador obtida no conjunto de treino (*training set*) com a estimativa de precisão obtida através do método *10 fold-cross validation*. Como explica esta diferença (se existir)?

[2] Use o classificador JRip, ou seja, a versão Weka do classificador de regras RIPPER.

(a) Crie um classificador com e sem *rule pruning*. Qual é o melhor? Justifique a sua resposta.

[3] Use o classificador J48, ou seja, a versão Weka do classificador C4.5 da árvore de decisão.

(a) Explore o uso de diferentes parâmetros de J48, como *pruning* ("unpruned") e número mínimo de registos nas folhas ("minNumObj").

(b) Descreva os padrões que obteve e compare com as conclusões obtidas nas questões anteriores.

1.4. Clustering Tendency

Investigue se existe uma tendência de *clustering* no *dataset*. Pode começar por agrupar os dados com o algoritmo SimpleKMeans, para $2 \leq k \leq 10$.

[1] Não use o atributo *class*, num, para o *clustering*.

[2] Encontre um valor adequado para k, ou seja, o número de *clusters* que vai construir. Justifique a sua escolha de k.

[3] Use o atributo *class* para o *clustering* e certifique-se de que os desvios padrão também são computados para atributos numéricos (displayStdDevs).

[4] Estude as medidas numéricas apresentadas pelo Weka para cada *cluster*. O que pode concluir?

[5] Selecione a opção "Visualize cluster assignments" e tente descobrir uma descrição para cada *cluster*.

[6] Investigue a possibilidade de utilizar as informações do *cluster* para construir um classificador para *num*. Compare os resultados com o que obteve na seção 1.3. Obteve um classificador melhor?
Pista: na janela "Visualize cluster assignments" seleccione no eixo Y "Cluster" e guarde como um novo *dataset*.

1.5. Predicting Performance

Na etapa anterior construiu vários modelos. Por fim, é necessário comparar os diferentes modelos e apresentar as suas conclusões.

[1] O Weka oferece várias medidas de avaliação de desempenho. Escolha algumas medidas de desempenho e justifique a sua escolha.

[2] Resuma numa tabela as medidas de desempenho para cada classificador e cada *dataset*.

[3] O que pode concluir?