Universidade do Minho
Departamento de Informática

FE09

Curso: Mestrado Integrado em Informática – Engenharia do Conhecimento
U.C.: Descoberta do Conhecimento

Folha de Exercícios FE09	
Docente	Cristiana Neto
Tema	RapidMiner – Árvores de Decisão
Turma	PL
Ano Letivo	2019-20 – 2º Semestre
Duração da aula	2 horas

1. Parte I

[1] Quais as características dos atributos de um *dataset* que podem levá-lo a escolher uma metodologia de *Data Mining* de árvore de decisão, em vez de uma abordagem de regressão linear? Porquê?

[2] Para que servem as percentagens de confiança, e por que razão é importante considerá-las, para além de considerar apenas o atributo de previsão?

[3] Como é que é possível manter um atributo, como o nome ou número de identificação de uma pessoa, que não deve ser considerado como preditivo no modelo de um processo, mas que é útil ter nos resultados de *Data Mining*?

[4] Quais as principais vantagens apresentadas na utilização de árvores de decisão comparativamente com outras técnicas de *Data Mining*?

2. Parte II

Com a resolução desta ficha pretende-se criar uma árvore de decisão para prever se você e outras pessoas que conhece seriam sobreviventes ou mortos se estivessem no Titanic. Complete os seguintes passos.

[1] Faça *download* do *dataset* "titanic-training". Importe os dados para o repositório do RapidMiner. Execute a fase de *Data Understanding*.

(a) Qual foi a percentagem de passageiros sobreviventes?

(b) Qual era a principal faixa etária dos passageiros que estavam no Titanic?

(c) Sobreviveram mais crianças ou mais adultos?

[2] Efectue a etapa de *Data Preparation*. Não se esqueça de colocar o operador *Set Role* nos atributos que justifiquem a sua aplicação.

[3] Usando o RapidMiner, crie um primeiro processo utilizando um operador de otimização de parâmetros para descobrir valores otimizados para os parâmetros do operador de *Decision Tree*, tal como se encontra descrito nos slides das aulas.

[4] Numa folha Excel inclua algumas pessoas no *dataset* de teste (titanic-scoring.csv) (pode até usar informações de pessoas que conheça). Guarde esta folha Excel como um ficheiro CSV. Importe-o para o repositório do RapidMiner.

[5] Num novo processo, repita os passos no RapidMiner tal como descritos nos *slides* da aula para aplicar o modelo de árvore de decisão ao *dataset* de teste ("titanic-scoring").

(a) Execute o modelo usando os parâmetros *default*. Após executar o modelo, na secção dos resultados, examine as previsões e as percentagens de confiança no conjunto de teste. Relate os nós da árvore, e discuta se as pessoas que inseriu seriam sobreviventes, falecidos ou desconhecidos.

(b) Volte a executar o modelo, mas agora usando os valores dos parâmetros encontrados no exercício 3. Relate as diferenças na estrutura da sua árvore. Discuta se as suas hipóteses de sobrevivência e das pessoas que conhece aumentam.

(c) Repita os exercícios 3 e 4(b) até que fique satisfeito com os resultados obtidos. Apresente detalhadamente todas as tentativas, bem como os resultados obtidos e respetivas comparações.