Universidade do Minho
Departamento de Informática

FE09

Course: Mestrado Integrado em Informática – Engenharia do Conhecimento
U.C.: Knowledge Discovery

Exercise Sheet FE09	
Teacher	Cristiana Neto
Theme	RapidMiner – Decision Trees
Class	PL
Year	2019-20 – 2nd Semester
Duration	2 hours

1. Part I

[1] What are the characteristics of a dataset's attributes that can lead you to choose a decision tree data mining methodology instead of a linear regression approach? Why?

[2] What are the confidence percentages for, and why is it important to consider them, in addition to considering only the prediction attribute?

[3] How is it possible to maintain an attribute, such as a person's name or identification number, which should not be considered as predictive in a process model, but which is useful to have in the data mining results?

[4] What are the main advantages presented in the use of decision trees compared to other data mining techniques?

2. Parte II

With the resolution of this exercise, it is intended to create a decision tree to predict whether you and others you know would be survivors or dead if you were on the Titanic. Complete the following steps.

[1] Download the “titanic-training” dataset. Import the data into the RapidMiner repository. Perform the Data Understanding phase.

(a) What was the percentage of passengers surviving?

(b) What was the main age group of passengers on the Titanic?

(c) Did more children or more adults survive?

[2] Perform the Data Preparation step. Don't forget to put the Set Role operator in the attributes that justify its application.

[3] Using RapidMiner, create a first process using a parameter optimization operator to discover optimized values for the Decision Tree operator parameters with the training dataset, as described in the lesson slides.

[4] In an Excel sheet, include some people you know in the scoring dataset (titanic-scoring.csv). Save this Excel sheet as a CSV file. Import it into the RapidMiner repository.

[5] In a new process, repeat the steps in RapidMiner as described in the lesson slides to apply the decision tree model to the test dataset ("titanic-scoring").

(a) Run the model using the default parameters. After running the model, in the results section, examine the predictions and percentages of confidence in the test set. Report the tree nodes and discuss whether the people you entered would be survivors or deceased.

(b) Run the model again, but now using the parameter values found in exercise 3. Report the differences in the structure of your tree. Discuss whether your chances of survival and the people you know increase.

(c) Repeat exercises 3 and 5 (b) until you are satisfied with the results obtained. Present in detail all the attempts, as well as the results obtained and the respective comparisons.