TDS3301 DATA MINING ASSIGNMENT

This assignment is to be attempted in groups of up to 4 people, who are from the same tutorial section. The assignment carries 30% of the marks, and is divided into 3 parts, which are described below. It is advisable to attempt the different parts of the assignment in parallel, as the due date for Parts II and III are very close.

1 PART 1: EXPLORATORY DATA ANALYSIS [DUE: 30/12/2016] – 10%

In Part 1, the group needs to identify and characterize a data set. The datasets can be acquired from resources such as Kaggle, CrowdFlower, and UCI Machine Learning Repository etc.

Prepare a report on the chosen dataset by completing the following tasks:

A. Describe the dataset in your own words.
B. What possible insights can be obtained from mining the chosen dataset?
C. What type of data mining technique (association rule mining, classification or clustering) would be relevant? Give an example, for example, if you think classification is suitable, describe what will be classified and what the possible classes are.
D. Describe data quality issues, and be specific. Identify which attribute (column) has issues, or if the structure of the data has problems.
E. Perform a pre-processing task on the dataset chosen.

Create a team page on a code repository/sharing site such as GitHub and submit the document and R source code used to achieve the tasks above.

2 PART 2: ASSOCIATION RULE MINING [DUE DATE: 13/01/17] - 10%

Your group's task for this part is to identify and perform an association rule mining task. This involves describing the following:

1. Objectives: What is the domain and what are the potential benefits to be derived from association rule mining. This is high level - not find patterns, but what would improve because of the use of the patterns.

2. Data set description: What is in the data, and what preprocessing was done to make it amenable for association rule mining. Where choices were made (e.g., parameter settings for discretization, or decisions to ignore an attribute), describe your reasoning behind the choices.
3. Rule mining process: Parameter settings, choice of algorithm, and the time required.
4. Resulting rules: Summary (number of rules, general description), and a selection of those you would show to a client.
5. Recommendations: What should the client do because of the rules discovered?

Use the Extended Bakery dataset found at https://wiki.csc.calpoly.edu/datasets/wiki/ExtendedBakery

Create a team page on a code repository/sharing site such as GitHub and submit the document and R source code used to achieve the tasks above. You can continue using the same code repository that you created in part me.

3 CLASSIFICATION [DUE DATE: 03/02/2017] 10%

The final part of the assignment requires the group to complete a classification task. As usual, choose a dataset to perform the classification task on. Apply Decision Trees, Naïve Bayes and ANN on the classification task and compare the performance of the classifiers using measures such as accuracy, TPR, FPR etc. and so forth. Choose from the following datasets:

1. Occupation detection dataset https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+#

2. Student Performance dataset https://archive.ics.uci.edu/ml/datasets/Student+Performance

3. Otto product dataset https://www.kaggle.com/c/otto-group-product-classification-challenge

[Challenging choice]

Create a team page on a code repository/sharing site such as GitHub and submit the document and R source code used to achieve the tasks above. You can continue using the same code repository that you created in part me.

You will need to describe the following:

A. Exploratory data analysis
B. Pre-processing tasks
C. Choice of performance measures
D. Performance of the 3 classifiers
E. Suggestion as to why the classifiers behave differently.

4 EXTRA CREDIT: UP TO A MAXIMUM OF 5 MARKS WILL BE ADDED FOR...

1 mark: Completing a Kaggle competition – proven by team score

2 marks: Perform clustering on data sets to discover natural clusters, and to label the data using these clusters.

5 marks: Create a shiny application to show the results of Parts I, II and III.