Data Mining

COP259

Coursework assignment

Credit value: 100% of the module

This coursework calls for a report on the tasks that you are to carry out giving results and their interpretation on a data set that you have been provided to analyse using the Data Mining Software WEKA.
Please note, this is not a collaborative (group) work, and any kind of plagiarism (including Plagiarism by structure) is forbidden.

Your Task

In this assignment you are not expected to write any software code but use existing tools provided by WEKA.
Perform the following tasks on the dataset provided and produce results and their interpretation in the form of a report.

- You must only use the data set provided.
- In-app screenshots are compulsory to show the results (Not step by step).
- Word count for each part should not exceed 500 words.

Part 1

Pre-process the data, discuss and justify in detail what steps have you taken.

[20 marks]

Part 2

(i) Use linear regression and two other methods to rank each feature (attribute) on its ability to predict the class variable. Tabulate the results and describe how you obtained them. Based on the results, discuss the relative importance of the features and how you would rank them. Note that this part should not be performed using t-test.

[20 marks]

Part 3

(ii) Apply two chosen classifiers to the original data set using 10-fold cross-validation and also apply the same two classifiers to the modified data set containing features of your choice based on the results of **(i)**. Perform *t*-tests to determine whether the differences in performance of the classifiers and data sets are statistically significant from one another. Discuss and justify your chosen classifiers and your findings.

[20 marks]

Continued in the next page

Part 4

(iii) Use the equal-width binning strategy to produce a modified data set. Apply the better performing classifier used in **(ii)** to the modified data set. Discuss the differences in performance between the results in **(ii)** and **(iii)** for the chosen classifier and the merits of using the binning strategy.

[20 marks]

Part 5

(iv) Apply k-means clustering to the original data set. Set the number of desired clusters to the number of classes already known. Assign each instance a class in the clustering output. Discuss the difference between such classification and the available ground-truth.

[20 marks]

Guidelines

Your report on WEKA Exercises should not exceed 3000 words.

The report should contain a section for each of the task. Each section should describe:

(a) Methodology and Data/Tools used and justification;

(b) Experimental Results;

(c) Interpretation of these results.

Marking Criteria

The assignment will be assessed on the following basis:

1. Presentation

- Is the coursework well structured?
- Are the results presented clearly?

2. Technical Content

- Have you used an appropriate process for generating results?
- Are the purposes of the tasks and the results understood and well justified/discussed?