Data Mining and Discovery

Report Assignment

This task will be assigned to your entire study group and you are encouraged to work on this together. However, your final submission must be entirely your own work. You *do not* need to work on the same datasets as other people in your study group.

Your task is three-fold. First, you must understand the given topic (*Subtopic 1*) in data mining (see below), using either the recommended text or any other resources you can find. Second, you must apply this topic to a data set of your choice **and** compare and contrast your results with those obtained using the second given topic (*Subtopic 2* - see below). Finally, you must demonstrate some level of data preprocessing. You will be able to choose a data set from a selection that will be made available to you. If you wish to use a data set not in this selection, then you need to discuss this with Dr John Evans so that the suitability of this data set can be confirmed. You are also free to use any other areas in data mining you wish (such as an area discussed in class) if you think this will help your extraction of key features from the chosen dataset.

Your report must be a maximum of 2 sides of A4, not including any code that you write, and your font size should be size 11 with font being Times New Roman. Your margins should be no less than 1 inch/2.5cm. All figures, tables, illustrations etc. should be clearly numbered with an appropriate title. Any references used must be clearly listed at the end of the report and are not part of the page limit. You must use the Harvard referencing style. You are encouraged to use LaTeX but this is not required.

Each member of the study group must follow the set topics, but there is no requirement that every member of the study group must use the same data set, or follow the same path. The study group is there to assist with the learning process, and to encourage open discussion and the development of ideas.

Subtopic 1 is as follows: Linear Regression

Subtopic 2 is as follows: Classification

See Appendix D of *Introduction to Data Mining* for a discussion on linear regression. You can use any classification algorithm you like. You are encouraged to find your own resources and applications. While the aforementioned resource may be helpful, you are not required to follow it.