ST3189: Assessed Coursework Project

You will undertake a project that will determine your final mark of the course by 30 per cent. The project will require you to analyse one or more real-world datasets of your choice. You can use the OpenML website, the UCI repository, or any other open access domain to select your dataset(s).

The project will consist of completing the following three tasks that can be implemented on one or more of your chosen real-world datasets.

1. Unsupervised Learning: where the problem consists of identifying homogeneous population groups or dimension reduction techniques, which can then be used in the context of the empirical application
2. Regression: where the problem consists of continuous target variable(s).
3. Classification: where problem consists of categorical target variable(s).

You will be expected to present each of the datasets you are analysing, identify research questions that can be addressed by your analysis and, ideally, present relevant existing literature and contrast your results against it. You are expected to use multiple technique for the regression and classification tasks, and compare their results.

In all cases, your analysis should be presented in a paper like format, avoiding highly technical language where possible. It may be helpful to think of your audience as consisting of people with some quantitative background but no prior knowledge of Machine Learning. Your ability to present and interpret the results will be regarded as important as your ability to apply the taught techniques.

The results of the project should be presented in a 10-page article in A4 format. The 10-page limit includes figures and tables but excludes the title page, table of contents and references. Make sure to include your candidate number in the title page and the filename but not your name. If your candidate number has not been generated at the time of submission, this should be your student registration number (SRN). In addition to the 10-page article, which should be submitted via a word or pdf file, your R code should also be submitted with appropriate comments and description via an R script or an RMarkdown file. You may alternatively also use Python code; in which case you should submit a Jupyter notebook or a Spyder script file.

You may choose to conduct all the above three tasks on a single dataset or conduct some of the tasks on separate datasets; this is up to you. Do not submit your data, just provide the open access links in your code files, from which the data can be downloaded.

To sum up the following two files are required where your candidate number (or if not yet available, your student registration number) should be visible:

1. A word or pdf file with your report that should not contain any code (10-page limit applies as mentioned above).
2. Your code in a single file of appropriate format (R script, RMarkdown, Spyder script, Jupyter notebook).