Individual coursework

- You can use either R or Python to complete this task, but please choose one and do not use both.
- Please upload your report in **one pdf file** plus **one code script**.

Please download the heart-disease.csv dataset from Moodle. Your aim is to predict the coronary heart disease (chd: 1/0) for males in a heart-disease high-risk region of the Western Cape, South Africa, described by the following nine features:

- sbp: systolic blood pressure
- tobacco: cumulative tobacco (kg)
- ldl: low density lipoprotein cholesterol
- adiposity
- famhist: family history of heart disease
- typea: type-A behavior
- obesity
- alcohol: current alcohol consumption
- age

Complete the following tasks and write a report to discuss and summarise your findings.

1. Conduct an exploratory data analysis on this dataset.
2. Fit a logistic regression with ridge penalty to classify the patients and explain the results.
3. Explore other classifiers discussed in this module and find one that gives you the highest accuracy. Explain your procedure to achieve the final chosen classifier and discuss the results.

Further requirements are on the next page!

Note:

1. The report should be no more than 1000 words and 5 pages, with font size of at least 12, margin size of at least 1 inch and 1.5 line spacing.
2. Please do not submit .ipynb or .rmd files as the report.
3. The coursework will be marked based on the quality of the report, e.g. logic, structure, writing, correctness, presentation etc., and the quality of the codes, e.g. enough comments for readability. You will not be given a good mark if you achieve the highest accuracy, but do not write a good report and high-quality codes.
4. You may not be able to include all classifiers that you have tried in the report, given the word and page limits. Thus, please give clear motivation and explanations of why the classifiers in the report are chosen.