S412/S512/T650 Statistical Learning Project II

- **Important:** This is not a group project. Please work independently.
- Data `proj2data.csv` A retrospective sample of 462 males in a heart-disease high-risk region. The data set contains the following variables
  - Response: `HD`: coronary heart disease (Yes) or not (No)
  - Predictor variables
    - `adiposity`: the accumulation of excessive body fat;
    - `age`: age in years;
    - `alcohol`: current alcohol consumption;
    - `famhist`: family history of heart disease (Yes, No);
    - `ldl`: low density lipoprotein cholesterol;
    - `obesity`: UN:Underweight, HE: Healthy weight, OV: over weight, OB: obesity
    - `sbp`: systolic blood pressure;
    - `tobacco`: cumulative tobacco (kg);
    - `typea`: score of type-A personality;
- Training data: the first 230 cases.
- Testing data: the last 232 cases.
- Apply all of the following models and find a model that has the lowest test data error rate and can be used to classify whether a patient has the coronary heart disease or not. The best final model should be chosen by comparing the test data error rates.
  - 1. LDA
  - 2. QDA
  - 3. Logistic regression
  - 4. Naive Bayes
  - 5. KNN classifier
  - 6. Tree-based Methods
  - 7. Support vector machine

Project Report Format:

- Please typeset your report using R Markdown in RStudio to produce a PDF file. Your report should have the following
  - A title
  - Abstract
  - Introduction
  - Section(s) containing details of your data analysis, include only relevant R codes used for data analysis, graphs and explanation of the methods you used, etc. If random data were generated you must specified a random seed so that your results can be repeated.

  - Summary of Results and Discussion which include a table to summarize the test error rate of the models trained and give the predicting formula of your final model if possible.

- Limit your report within 15 pages. Only report necessary and relevant R code, graphs, and printout. Make your report neat, clean, readable, not like draft. Do not include the data in your report. Remove all unwanted "warnings" and "messages" using `message=FALSE, warning=FALSE` in R chunk or fix the issues that relate to the messages.

- Important: Using AI tools or anything like this is prohibited. Use only the R libraries and functions taught or mentioned in the lectures or the textbook or no credit!