STA5076Z - Supervised Learning 2020
Assignment 3

Instructions

- You may use any typesetting software you wish, but I would encourage you to use R-Markdown or LaTeX.
- Provide complete code for each of the relevant sections under separate headings as an **appendix** to your write-up.
- You may **NOT** provide R output interspersed between your answers! Please typeset relevant elements in the output either in-line, or tabulate results formally. Plots are very useful but use them sparingly and make sure that a given plot is relevant to the question and pertains to text in your answer. Figures are meant to enrich your analysis, not leave it to the reader to analyse. Provide captions for all figures and tables and explanations.
- When you typeset R code use courier or an equivalent `typewriter'-like font.
- All assignments must be accompanied by a signed plagiarism declaration - a template is provided on Vula.
- **Hand-in date is 10th of July, Friday 5pm**. At UCT, marks need to be processed by the 24th of July 2020, which means I need to have the assignments for marking two weeks before this due date. Otherwise I will not be able to finish the marking on time.

Problem Set

Using the training and test data provided on VULA, develop a Support Vector Machine (SVM), and a Neural Network (NN) model to predict the Framingham Cardiovascular Risk Score (T enYearCHD) using the variables available. Details can be found here: https://en.wikipedia.org/wiki/Framingham_Risk_Score and here https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset

The training and test datasets contain the same number of variables, except the Y variable val ues for the test dataset are hidden from you. Make your predictions and write them for the test dataset to a .csv file using the following naming convention (replace `STUDENTNO' with yo ur student number). Do not forget to use the same means and st.devs for the test set if you are considering data standardization.

Your assignment should also include an inspection of your dataset, explaining the problems and solutions that might be necessary before implementing the methods mentioned. Write up a short report comparing the various strategies that you have used. Make sure to include:

- A short description of the dataset and the variables (any data scientist needs to first fully understand the data and there is only one way for this, descriptive statistics and visualizations). Carefully consider the type of the variables and the number of categories available for some of the categorical variables.
- A short description of each technique.
- A comparison of their performance.
- Motivate your choice of regularisation mechanism and hyper-parameters.
- Do not forget to generate predictions for the test data!

You will be required to hand in all materials (R-scripts, R-Markdown and/or write-up) and a .csv file giving your test data classifications.
Good luck!