Individual coursework

- Please submit your coursework on Moodle **by Midday on 26/03**.
- Please upload your text answers and plots to Questions 1 and 2 in **one pdf file**.
- Please also upload one R script for Question 1 and one R script for Question 2.
- Make sure that you have included **sufficient** comments in the codes to make them **readable** by other people. There should be **no error messages** shown when I run your R scripts. You can assume that I have installed all required packages.

Question 1 [**7 marks**]

Use the German credit data. Split the data to a training set (70%) and a test set (30%).

(1) Fit a decision tree to the training data with an optimal tree size determined by 10-fold cross-validation. Create a plot of the pruned tree and interpret the results. If cross-validation does not lead to selection of a pruned tree, then create a pruned tree with five terminal nodes. Compute the test error rate of the pruned tree.

[**2 marks**]

(2) Fit a random forest to the training data with parameter tuned by 10-fold cross-validation. Compute the test error rate and comment on the results. Create a plot showing variable importance for the model with the tuned parameter and comment on the plot. [**3 marks**]

(3) Draw one plot with two ROC curves for the test predictions in (1) and (2). Comment on the plot. [**2 marks**]

Question 2 [**8 marks**]

Simulate a three-class dataset with 50 observations in each class and two features. Make sure that this dataset is not linearly separable. Make a scatter plot to show your dataset. Split the dataset to a training set (50%) and a test set (50%). Train the support vector machine with a polynomial kernel and an RBF kernel, as well as a support vector classifier on the training data. The parameters associated with each model should be tuned by 5-fold cross-validation. Test the models on the test data and comment on the results.