Assignment 3

Course: STAT2604 Introduction to R Programming and Elementary Data Analysis
Total marks: 100
Due date: 23:59, Dec 03, 2021

Please pack your Rmd source code file together with the output html file as one compressed file, and submit only that compressed file onto Moodle. Name the compressed file in the format `(Name)_(UID)_A3`.

Question 1. Download the UCI machine learning data set file `ionosphere.data` through: "https://archive.ics.uci.edu/ml/datasets/Ionosphere". The goal is to predict high-energy structures in the atmosphere from antenna data. More information can be found in the data description file `ionosphere.names` from the website above. (*Hint: data cleaning might be needed before model construtions*)

a) Load the data into R, split 80% of the data samples into training set, and 20% into testing set. (20 marks)

b) Run the following algorithms using the `caret` package: logistic regressions, KNN, SVM, naïve bayes, decision trees, random forest and glmnet models. Tune the each of the above models (if it does have tunning parameters) using cross-validation, and set `metric="ROC"`, `tuneLength=10`. Compute the area under ROC curve on the testing set for each model. (50 marks)

c) Perform model selection using the function `caret::resamples`, visualize the result, and choose the best model for this data set. (30 marks)