**Homework 3**

Use caret for all the models except where mentioned to use another package. Review examples provided on Canvas before attempting homework. Please turn in a Quarto (qmd) file or a word/pdf document generated from the qmd file or both.

1. Load the Auto.csv file on Canvas into a dataframe. A description of the dataset can be found at <https://rdrr.io/cran/ISLR/man/Auto.html>
2. Create and add a binary variable, mpg\_high\_low, to the dataset that contains a Yes if mpg is a value above 30, and a No if mpg is a value less than or equal to 30. Make sure the mpg\_high\_low is of type factor.
3. Split the dataset into 75% training and 25% test and use 10 fold cross validation for the models below
4. Fit a logistic regression model to the training set to predict mpg\_high\_low using all the other features/variables except mpg, year, origin, and name. Which variables are statistically significant? Predict the mpg\_high\_low using the test dataset and report the accuracy, specificity, sensitivity, and F measure.
5. Check if the auto data is imbalanced with respect to mpg\_high\_low. Report the percentage of the data that belong to the two classes (Yes and No).
6. Alter the threshold for classifying a No to 0.6 and report the increase in specificity and the corresponding loss in accuracy.
7. Fit a Naïve Bayes model to the training data to predict mpg\_high\_low using all the other features/variables except mpg, year, origin, and name. Predict the mpg\_high\_low using the test dataset. Plot the ROC curve and report the best threshold on the ROC curve plot. Report the AUC on the curve plot as well. Report the accuracy, sensitivity and specificity.
8. Fit a KNN model to the training data to predict mpg\_high\_low using all the other features/variables except mpg, year, origin, and name. Use a grid search between 3 and 10 to find the best value of k. Report the accuracy, sensitivity, specificity, precision and recall
9. Fit a LDA model to the training data to predict mpg\_high\_low using all the other features/variables except mpg, year, origin, and name. Report the accuracy sensitivity and specificity.
10. Summarize the performance of the all the above models by creating a dataframe with 4 columns – Model\_Name, Accuracy, Sensitivity, Specificity. The data frame should contain one row for each model you built above with each of the columns filled in with the appropriate metric. Print out the dataframe. Which model performed the best from an accuracy point of view and which model performed best from a specificity point of view without adjusting for the threshold?