Classification Trees

In this assignment you will complete a variety of tasks related to binary classification with classification trees. The dataset that we will be using is a healthcare dataset which involves the prediction of heart disease based on several characteristics. If you have to spend some time on Google to help you understand what these variables are, please do so.

Libraries: For this assignment you will need the following libraries: tidyverse, tidymodels, caret, rpart, rpart.plot, rattle, and RColorBrewer.

Before beginning the assignment tasks, you should read-in the data for the assignment into a data frame called "heart".
Then **carefully** convert the "sex", "ChestPainType", "RestingECG", "ExerciseAngina", "ST_Slope", and "HeartDisease" variables to factors. Recode the levels of the "HeartDisease" variable from "0" to "No" and "1" to "Yes".

Question 1: Split the data into training and testing sets. Your training set should have 70% of the data. Use a random number (set.seed) of 12345. Stratify your split by the response variable "HeartDisease".

How many rows are in the training set?

Question 2: Create a classification tree to predict "HeartDisease" in the training set (using all of the other variables as predictors). Plot the tree. You do not need to manually tune the complexity parameter (i.e., it's OK to allow R to try different cp values on its own). **Do not use k-folds at this point**.

The first split in the tree is a split on which variable?
A. Sex
B. ST_Slope
C. ChestPainType
D. ExerciseAngina

Question 3: Examine the complexity parameter (cp) values tried by R.

Which cp value is optimal (recall that the optimal cp corresponds to the minimized "xerror" value)? Report your answer to two decimal places.

Question 4: Use a tuning grid (as we did in the Titanic problem) to allow R to try 25 different values for the complexity parameter (cp). R will select reasonable values. Use 5-fold k-fold cross-validation (don't forget to set up your folds). Use a seed of 123 when setting up your folds.

Hint: You can reuse the vast majority of the code that I provided for you. Be careful to change names and you should be "good to go". Note: This model took about two minutes to run on my computer. Your run time will vary by your computational power :) Plot the relationship between the complexity parameter (cp) and model performance (given by accuracy and by ROC AUC). I have provided code in the lectures that use the "collect_metrics" functions to help you do this.

From this plot, what is the accuracy of the model (to two decimal places) if a cp value of 0.1 is selected? You will need to "eyeball" this answer. I have included a bit of a tolerance in the answer on Canvas. As long as you are "close" to the correct accuracy, you will see your answer marked as correct.

Question 5: Which cp value (to four decimal places) yields the "optimal" accuracy value?

Question 6: Plot the tree that corresponds to the cp value from Question 5. Don't forget to finalize your workflow and generate your final fit before trying to plot.

How would you classify a patient that is "Male" with an "ST_Slope" that is "Flat"?

Question 7: What is the accuracy (on the training set) of the "tree" that you generated in Question 6? Take your time and think about how to determine this value. Report your answer to four decimal places.

Question 8 What is the sensitivity of your model from Question 6 (on the training set)? Report your answer to four decimal places.

Question 9 What is the naive accuracy of your model from Question 6 (on the training set)? Report your answer to four decimal places.

Question 10 What is the accuracy of your model from Question 6 on the testing set (to four decimal places)?