Middle Tennessee State University
Department of Mathematical Sciences

DATA 3550 Fall 2023– Homework 5
Applied Predictive Modeling

Instructor: Ramchandra Rimal

Due: Friday November 24, 2023 at 5 pm

Name: _____
Student Number: _____

The homework 5 contains 3 pages (including this cover page) and 2 questions. Total of points is 30. Please **show all work with proper justification and answer each question**. Your work as well as the presentation of your work will be graded. You can discuss the problems with your classmates.

Please answer each parts of the problems clearly. Please put the comment on your code next to each line to explain what it does. Once your work is completed, do not forget to rename your file as your LastName.FirstInitial-DATA3550-HW5. Then upload your .ipynb file in the dropbox folder for the Homework 5.

Distribution of Marks

Question	Points	Score
1	20	
2	10	
Total:	30	

1. (20 points) For this problem you need to use the Heart dataset and seek to classify the patient have AHD or not using classification trees and related approaches.

    1. (a) Preprocess the data, you may need to convert the categorical variables to dummy variables, impute or drop the missing values. (3 points)
    (b) Split the data set into a training set and a test set in which training set consists of 70% of the data and the remaining data on the test set. (1 points)
    2. (a) Fit a tree to the training data, with AHD as the response and the other variables as predictors. (1 points)
    (b) Calculate the accuracy, sensitivity and specificity on the training data. Describe the results obtained. (1 points)
    (c) Calculate the accuracy, sensitivity and specificity on the test data. Describe the results obtained. (2 points)
    (d) Create a variable importance plot. (1 points)
    3. (a) Using the result from part 1, Fit a tree to the training data, with AHD as the response and the most important 7 variables as predictors. (2 points)
    (b) Calculate the accuracy, sensitivity and specificity on the training data. Describe the results obtained. (2 points)
    (c) What is the training accuracy? Is there any difference in score compared to previous model that uses all predictors? (1 points)
    (d) Create a plot of the tree, and interpret the results. (2 points)
    (e) Pick one of the terminal nodes, and interpret the information displayed. (2 points)
    4. (a) Predict the response on the test data, and produce and plot a confusion matrix comparing the test labels to the predicted test labels.
    (b) What is the test accuracy? What would be the accuracy score without the model (i.e if we just classify the observation to the class with higher count)? Based on the results, is the model useful or not? (1 points)
    (c) Plot a ROC curve and explain your observation. (2 points)

2. (10 points) For this problem you need to use the preprocessed Heart dataset used in Part 1 of Question 1 and seek to classify the patient have AHD or not using enseble methods.

    1. (a) Fit a Random Forest Classifier to the training data, with AHD as the response and all other variables as predictors(not only the 7 predictors chosen before). (2 points)
    (b) Calculate the accuracy, sensitivity and specificity on the test data. Describe the results obtained. (2 points)
    (c) Is it better than the results obtained from the single decision tree? (1 points)
    2. (a) Apply the 10 fold grid search cross validation to the training data to determine the best criterion, number of trees to use in a forest, maximum depth of the tree and maximum number of features to use; for the best prediction accuracy. Report the best value of the parameters selected by 10 fold grid search cross validation. (2 points)

(b) Run the repeated 5 fold cross-validation on the whole dataset using the best value of the parameters selected by 10 fold grid search cross validation. The number of repetition is 30. (1 points)

(c) Report the average accuracy obtained on the hold out fold. Is it different than the accuracy obtained in previous part. If so, why? (2 points)

3. (BONUS) Do something with the classification modeling of the data provided. Note: You may want to explore new visualization/plot you haven't already tried for the previous homework.

   (a) You can implement any new techniques you learned outside this course and explain why this method is useful? (3 points)

   (b) Or, you can visualize the data to provide some interesting insights that we do not know yet and explain it. (3 points)

   (c) Or, you can create the plot to analyze the classification error obtained from the best model you have developed for this homework and explain what unique insights you get from that visualization. (3 points)