STAT 587 **HOMEWORK # 2** Winter 2026

DATA SCIENCE I

DUE DATE: **February 4, 11:59 pm**

Instructions:

- *You are allowed to use **Python** or a Scientific Calculator where required.*
- *Submit only **one** **PDF** on* CANVAS.

---

1. Consider the `College` data from the ISLP package. Details about the data is described on page 65 of the ISLP textbook for this class (https://islp.readthedocs.io/en/latest/datasets/College.html). We would like to ***predict*** the number of applications received using the other variables. 80% of the data (randomly generated) will be treated as training data. The rest will be the test data.

    a) Fit a linear model using least squares and report the estimate of the test error.
    b) Fit a tree to the data. Summarize the results. Unless the number of terminal nodes is large, display the tree graphically. Report its MSE.
    c) Use Cross validation to determine whether pruning is helpful and determine the optimal size for the pruned tree. Compare the pruned and un-pruned trees. Report MSE for the pruned tree. Which predictors seem to be the most important?
    d) Use a bagging approach to analyze the data with B = 500 and B = 1000. Compute the MSE. Which predictors seem to be the most important?
    e) Repeat (d) with a random forest approach with B = 500 and B = 1000, and m ≈ p = 3.
    f) Compare the results from the various methods. Which method would you recommend?

---

2. Consider the business school admission data available in the `admission.csv`. The admission officer of a business school has used an "*index*" of undergraduate grade point average (GPA,X_1) and graduate management aptitude test (GMAT,X_2) scores to help decide which applicants should be admitted to the school's graduate programs. This index is used to categorize each applicant into one of three groups – admit (group 1), do not admit (group 2), and borderline (group 3). We will take the last ***four*** observations in ***each category*** as test data and the remaining observations as training data.

a) Perform an exploratory analysis of the training data by examining appropriate plots and comment on how helpful these predictors may be in predicting response.
b) Perform an LDA using the training data. Superimpose the decision boundary on an appropriate display of the data. Does the decision boundary seem sensible? In addition, compute the confusion matrix and overall misclassification rate based on both training and test data. What do you observe?
c) Repeat (b) using QDA.
d) Fit a KNN with K chosen optimally using test error rate. Report error rate, sensitivity, specificity, and AUC for the optimal KNN based on the training data. Also, report its estimated test error rate.
e) Compare the results in (b), (c) and (d). Which classifier would you recommend? Justify your conclusions.
