STAT 8330: **Homework 5**

Due on October 19, 2020 (10:01pm)

1. In the lab, you applied a classification tree to the `Carseats` data after converting `Sales` to a qualitative response variable. Here, we seek to predict `Sales` as a continuous variable response. Using `set.seed(1)`, split the data into a training set (of size 200) and corresponding test set.

   (a) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?

   (b) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?

   (c) Use the bagging approach to analyze these data. What test MSE do you obtain? Use the `importance()` function to determine which variables are most important.

   (d) Use boosting to analyze these data. What test MSE do you obtain? Report what you use for "distribution", and how you chose the values for "n.trees" and "interaction.depth" (and, what you selected for those values).

   (e) Use random forests to analyze these data. What test MSE do you obtain? Use the `importance()` function to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error obtained.

   (f) Use the `mda` or `earth` package (or some other one!) to perform a MARS analysis of these data. Describe your approach, including your R commands and report your test MSE and any other information you think is relevant.

2. **Fish analysis**: On the course website is a data file `fish_data3` which is a data set consisting of measurements of fish abundance in lakes in Minnesota, along with lake information and water quality metrics. The goal is to use the abundance of different fish types, lake information, and water quality metrics (columns 1-10) to classify the lake as either a large eutrophic lake or a large mesotrophic lake (LSH7class). The data were collected by the MN Department of Natural Resoucres.

   Use `set.seed(1)` to obtain a training set with 400 observations and a test set with 207 observations.

   ```
   > set.seed(1)
   > training.set=sample(1:nrow(fish),400)
   > fish.test=fish[-training.set,]
   > LSH7class.test=fish.test[,11]
   ```

   Consider your best classification tree (as measured by classification error rate on the test sample after training on the training sample) for CART with bagging, CART with boosting, random forests and MARS. Report any information necessary so that your analysis could be duplicated (e.g., what choices for the parameter values, etc.). Which procedure did the best? What variables did you find to be important.