# Coding Assignment 3

Due Date: *Monday, October 13 (11.59pm)*

*Submission: on Gradescope*

***Instructions:***

**The coding assignments may be completed in groups up to 4 students**. If you do so, please make sure that you include everyone's full name, and that you *also select everyone's name when submitting the assignment on Gradescope*. This ensures that each group member will get a grade assigned and have access to the comments from the graders.

For this assignment, the following items need to be submitted:
(1) the **code** in `R` or *Python* (a Markdown file is ok; Jupyter notebook is ok)
(2) a **pdf** file with your code and results (Markdown-style) with all necessary plots and comments.

This coding assignment focuses on tree-based methods using the **wage** data set in the `ISLR2` package available both in R and Python. The goal of the homework is to use regression trees, random forests, and gradient boosting machines (GBM) to model the relationship between worker's **wages** and various demographic variables.

(a) Split the data into training (70%) and testing (30%) data sets. Use seed 598 for reproducibility of the results.

(b) Fit a tree model and visualize it. Interpret at least two splits.

(c) Prune the true using cross-validation. Comment on the performance of the pruning by comparing the MSE of the pruned vs. unpruned tree on the testing data.

(d) Fit a random forest model to the training data.

Choose two different values for `mtry` (in R) or `max_features` (in Python). If you use R, make sure that you include the variable importance (i.e. importance=TRUE in R) when fitting the model. In Python, this is automatically included when you use `RandomForestRegressor`. Comment on the results you obtain.

(e) Based on the random forest model in (d), which seem to be the most important predictors?

(f) Report the testing MSE for the random forest model and compare with the (pruned) regression tree fitted in (c).

(g) Fit a GBM.

Choose at least two learning rates (i.e. shrinkage option). If you use `R`, make sure you choose `distribution="gaussian"` in the gbm package.
Plot the test MSE versus the number of trees, and interpret the two most influential variables.

(h) Compare the performance (i.e. MSE) of GBM on the testing data to the random forest and (pruned) regression tree. Comment on the results.