**Tree Methods Project**

### **Project must be done in Jupyter Notebook using R od Python. Save project as .html.**

For this project we will be exploring the use of tree methods to classify schools as Private or Public based off their featuresA data frame with 777 observations on the following 18 variables.

* Private A factor with levels No and Yes indicating private or public university
* Apps Number of applications received
* Accept Number of applications accepted
* Enroll Number of new students enrolled
* Top10perc Pct. new students from top 10% of H.S. class
* Top25perc Pct. new students from top 25% of H.S. class
* F.Undergrad Number of fulltime undergraduates
* P.Undergrad Number of parttime undergraduates
* Outstate Out-of-state tuition
* Room.Board Room and board costs
* Books Estimated book costs
* Personal Estimated personal spending
* PhD Pct. of faculty with Ph.D.’s
* Terminal Pct. of faculty with terminal degree
* S.F.Ratio Student/faculty ratio
* perc.alumni Pct. alumni who donate
* Expend Instructional expenditure per student
* Grad.Rate Graduation rate

## Get the Data

**1. Read in College.csv file**

## EDA

**2. Create a scatterplot of Grad.Rate versus Room.Board, colored by the Private column.**

**3. Create a histogram of full time undergrad students, color by Private.**

**4. Create a histogram of Grad.Rate colored by Private. You should see something odd here.**

**5. What college had a Graduation Rate of above 100% ?**

**6. Change that college's grad rate to 100%**

## Train Test Split

**7. Split your data into training and testing sets 70/30.**

## Decision Tree

**8. Build a decision tree to predict whether or not a school is Private. Remember to only build your tree off the training data.**

**9. Predict the Private label on the test data.**

**10. Create a confusion matrix of your tree model.**

**11. Plot out your tree model.**

## Random Forest

**12. Now use random Forest to build out a model to predict Private class**

**13. What was your model's confusion matrix on its own training set?**

**14. Print the feature importance**

**15. Now use your random forest model to predict on your test set!**

**16. Which model is better and why?**