# Assignment 1

Terrence Coy
*CSC 410-01*

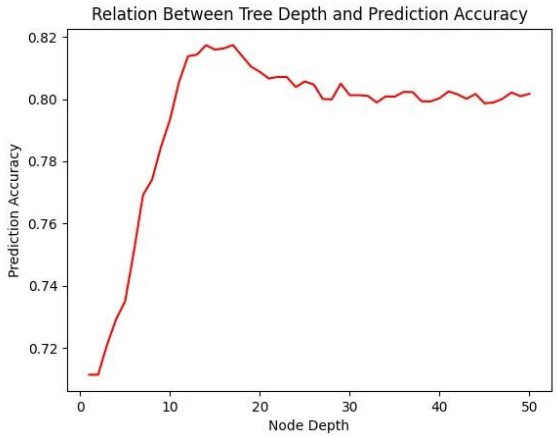
## I. DECISION TREE LEARNER

1A. Using the decision tree classifier with the *entropy* as the node selection criteria the results for each data set was as follows:

| Dataset | Height | Leaves | Prediction Accuracy |
|---|---|---|---|
| Spam | 69 | 6962 | 0.799 |
| Volcanoes | 18 | 126 | 0.776 |
| Voting | 5 | 9 | 0.989 |

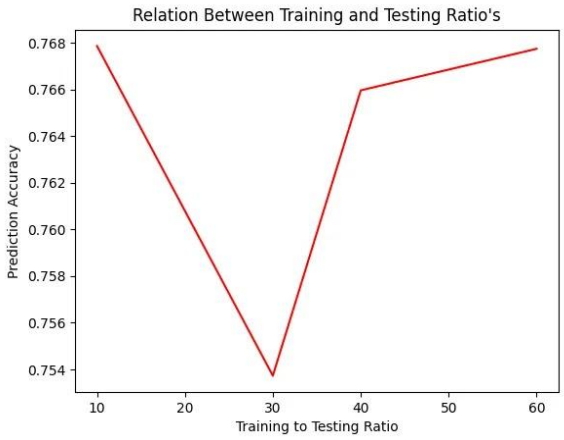
1B. The prediction accuracy for *voting* while using *gini* as the classifier as the node selection yielded a value of 0.989, the same result as before when evaluated with *entropy.*
'Repealing-the-Job-Killing-Health-Care-Law-Act' was the feature with the highest *gini* value from the dataset.

1C. Iterating from a node depth of 1 to 50 the prediction accuracy using *entropy* as the node selection was plotted on a graph:




The increase in prediction accuracy is expected as node depth increases, but it is interesting that diminishing returns appear after a node depths greater than 30. This could possibly be due to overfitting within the model. It's also noted that even with only a few nodes the decision tree's accuracy is still above 70%, which is considered passing by most accounts.

2. When splitting the *volcano* dataset into different training and testing ratios subsets of: 90/10, 70/30, 60/40 and 40/60, the highest reported prediction accuracy is with a 90/10 ratio with an accuracy of 0.7679 which makes sense that the model has more data to train with. It's interesting the 40/60 split group also yielded a higher prediction accuracy compared to the other ratios. Possibly due to the large amount of data available to test on. A smaller dataset could cause this anomaly to not be present.




## II. LINEAR REGRESSION LEARNER

1A. After splitting the *voting* dataset into 80/20 ratio for testing and training subsets the prediction accuracy for the model is 0.977, which is surprisingly lower than the decision tree classifier model.

1B. The probability estimates for *voting* using logistic regression are: 0.0013462, 0.9986538, 0.98982759 and 0.01017241. Using a decision tree classifier the probabilities are 0, 1, 1, 0, which is pretty close to the results given from the logistic regression.

1C. All three of these problems can be modeled with Linear Regression since the goal is to predict a continuous value from the datasets.

## III. REGRESSION ANALYSIS

A. The predicted value for the Day 15 data using Linear Regression model is 43.681minutes and the $R^2$ value is 0.0783.

B. The predicted amount of play time using a Regression Tree model is 44 minutes, with 15 leaves and a height of 5.

C. Loss error is not a good option for the Regression Tree model since the degree of error is not known. When using entropy the amount of information gained from each feature is known and more useful for the prediction model.