Homework 5 – Anthony Olivieri

# Data Preparation

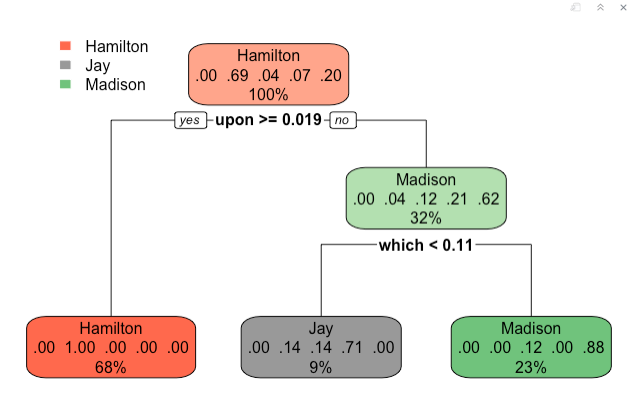
Normal data preparation techniques were used, but unnecessary. For example, evaluating the dataset for missing or outlier values returned no results. The dataset is complete and the values were all within a reasonable range.

Splitting the dataset into training data and testing data for this assignment was different than the technique typically used for other datasets. For other datasets, the data is usually randomly split into a training set consisting of ~2/3 of the data for training and ~1/3 of the data for testing. If the dataset for federalist papers statistics were split in this fashion, both training and testing sets would include the “disputed” author. Modeling a decision tree with this sort of training data, and then applying it to the testing data would yield a prediction of “disputed” in some cases.

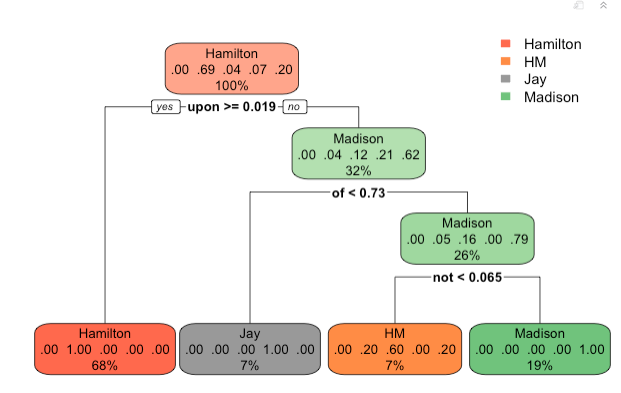
Instead, the federalist papers word dataset was split into the known authors for training and the disputed authorship for testing.

# Build and Tune Decision Tree Models

The original model with no control parameters created a short tree which combined author “HM” (Hamilton and Madison) into the other authors.



Adjusting the complexity parameter “cp” to .05 and the “minbucket” to 5 shows the final classification of authors to be “Hamilton,” “Madison,” “HM,” and “Jay.”



# Prediction

Using the final decision tree model to predict the disputed author results in a decision of “Madison.”