Aneeka Latif
Hw 4
IST 707

Introduction/Purpose

A precursor to the United states Constitution, was a series of essays calling for its deployment, known as the Federalist Papers. The Federalist Papers consisted of 85 essays; these essays were authored by three Founding Fathers of the United States who were to become the first Secretary of the Treasury, Alexander Hamilton, the First chief Justice, John Jay, and the fourth President, James Madison. Alexander Hamilton is known to have written 51 undisputed essays, Jay is known to have written 5, and Madison is known to have written 15. There are an additional 3 essays that were written by both Hamilton and Madison, but there are 11 essays remaining that are attributed to either Madison or Hamilton. As these essays were crucial to the acceptance of the U.S. Constitution, it is important to investigate the authorship behind these disputed essays in order to better interpret and understand its historical significance.

Section 1: Data preparation

To prepare the data for Weka – the set must be divided into a training set and a test set. The training set is comprised of all documents with distinct known authors; the test set is comprised of observations where the author is disputed. In the training set, unnecessary or unhelpful observations were removed – in this case, documents authored by John Jay and documents with joint authorship of Hamilton and Madison were also removed. The joint authorship examples will not be useful as those will confuse our training model. The final training set is comprised of documents with single authorship by Hamilton or Madison. The final test set is comprised of documents with disputed authorship. The final step in cleaning the test set is modifying the author value of “dispt” to “?”. This results in empty values when the test set is loaded to Weka.

Section 2: Build and tune decision tree models

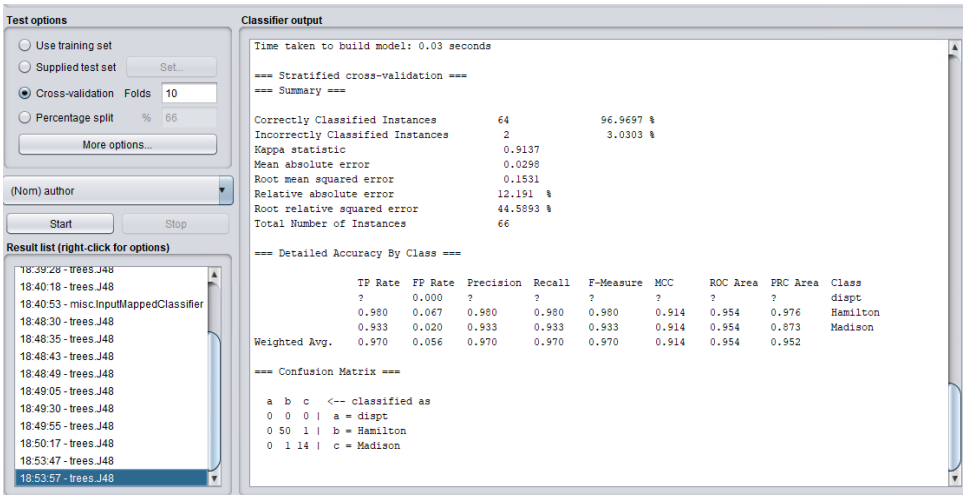
A DT model was first built with default parameters: confidenceFactor: 0.25, numFolds:4. In an attempt to build a better model, models were generated with varying confidenceFactor values as this relates to the aggressiveness of pruning and higher number of folds. The model was generated with unfiltered data, and again with discretized data though it seemed to perform better with unfiltered data.

Using a 66% percentage split and a confidenceFactor of .8, the model seemed to perform the best. There was 100% predictive accuracy and the MSE is very small. The MSE, mean squared error, describes the distance of errors between the data and prediction. An ideal MSE should be close to 0.

Aneeka Latif
Hw 4
IST 707

```
Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          22               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0186
Root mean squared error                  0.0538
Relative absolute error                  7.7115 %
Root relative squared error             16.6767 %
Total Number of Instances               22

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               ?        0.000    ?          ?        ?          ?        ?         ?         dispt
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Hamilton
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Madison
Weighted Avg.  1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b  c   <-- classified as
  0  0  0 |  a = dispt
  0 18  0 |  b = Hamilton
  0  0  4 |  c = Madison
```

Models were created with cross validation set to 10 folds – again, it seemed that higher confidenceFactors resulted in more correct classification and smaller MSE. In this case, there is 3% inaccuracy but it is still overall a fairly good model with 96% correctly classified instances and an MSE of .02.

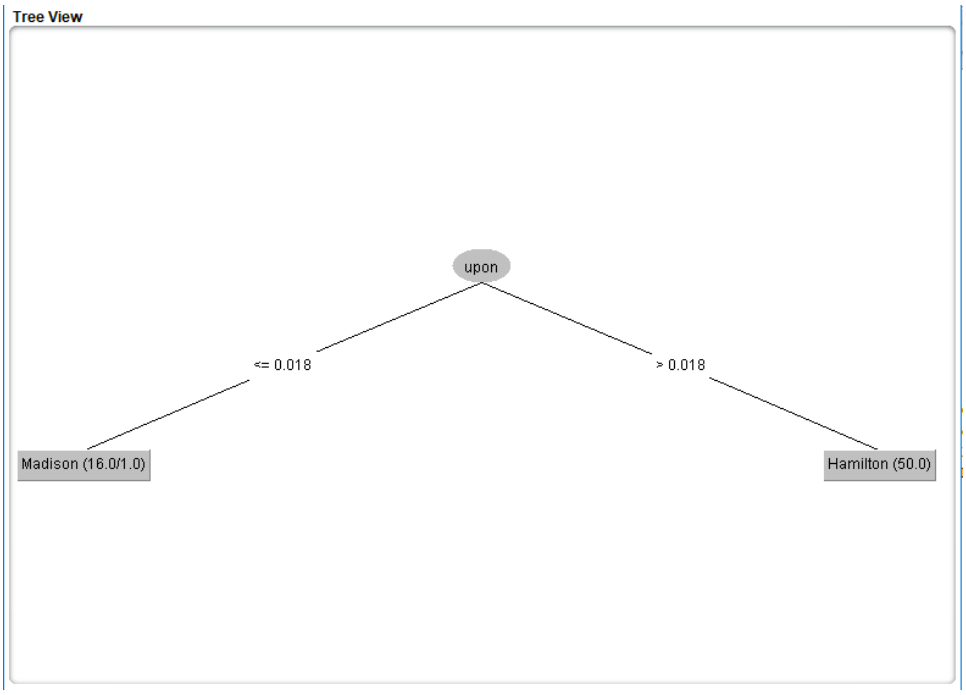


Section 3: Prediction

Aneeka Latif
Hw 4
IST 707

As our test data set does not include definitive answers for document authorship – our prediction model reports 100% incorrectly classified instances.

```
Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances           0               0      %
Incorrectly Classified Instances        11             100      %
Kappa statistic                          0
Mean absolute error                      0.6667
Root mean squared error                  0.7922
Relative absolute error                101.4706 %
Root relative squared error            108.7185 %
Total Number of Instances               11

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.000    ?        ?          0.000    ?          ?        ?         1.000     dispt
               ?        0.000    ?          ?        ?          ?        ?         ?         Hamilton
               ?        1.000    0.000      ?        ?          ?        ?         ?         Madison
Weighted Avg.  0.000    ?        ?          0.000    ?          ?        ?         1.000

=== Confusion Matrix ===

  a  b  c   <-- classified as
  0  0 11 |  a = dispt
  0  0  0 |  b = Hamilton
  0  0  0 |  c = Madison
```



Upon visualizing the tree, the model concludes that the 11 documents with disputed authorship were likely written by Madison. If the frequency value in a document of the word “upon” is less than or equal to .018, the decision tree indicates that the author is Madison, otherwise the author is Hamilton. This is the same conclusion drawn from the k-means clustering techniques in the previous assignment.

Aneeka Latif
Hw 4
IST 707