IST 707 – Applied Machine Learning
Professor Wallace
Spring 22

Assignment 5

Dixitha Kasturi
dkasturi@syr.edu

Topic: Decision Trees

Dataset: Federalist Papers

Outcome: Classify disputed documents correctly

Data Division/Preparation:

For the analysis, 2 cases were considered for training

1) Case1 : Data including articles by Hamilton and Madison(HM) , Jay and Disputed articles excluded out
2) Case2 : Data excluding articles by Hamilton and Madison(HM) , Jay and disputed articles

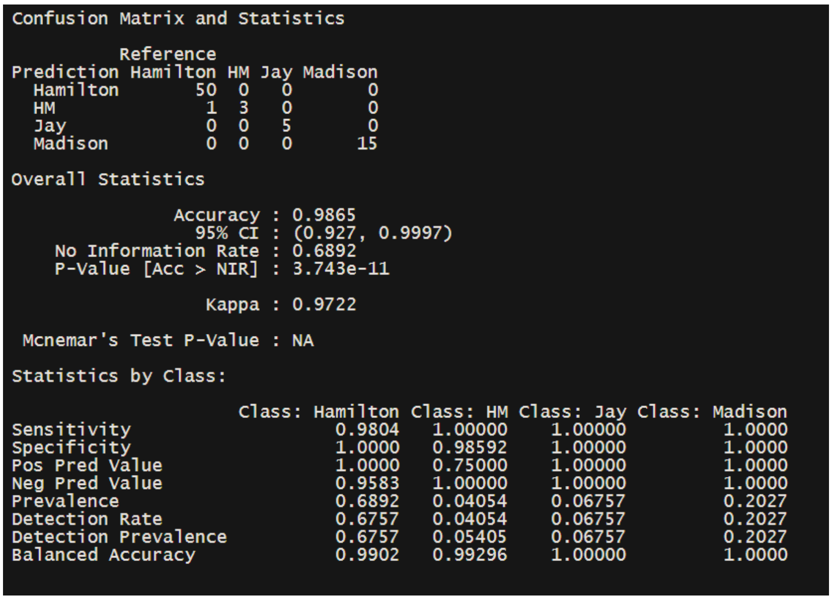
The testing data contains 11 of the disputed articles.

Decision Tree Models:

Four models were generated and tested.

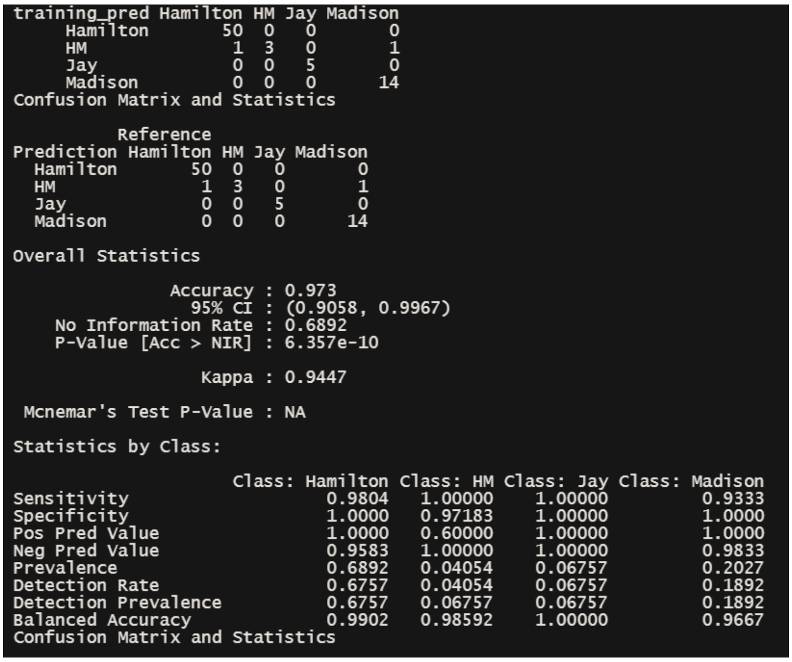
Case1:

Model 1: Simple DT with case 1 data as training and testing. This is to see if we get 100% accuracy. But on generating the model and predicting the same data, there was ambiguity. Usually the decision tree is constructed with a certain depth through splitting. If the depth is not large enough if specified, then a lot of the samples may het accumulated at each leaf without being of the same type. This could lead to not having 100% accuracy even if training and testing data is the same. The leaf nodes may not contain the same classes. For example, one of the articles that was written by Hamilton, was classified as being written by both Hamilton and Madison. This could be because Hamilton was involved in other Collaboratory works with madison and the decision tree got learnt in such a way that one of hamilton's works had more resemblance to the Collaboratory works hence it was misclassified as collaboratory work. We ended up getting 98% accuracy instead of the expected 100%.



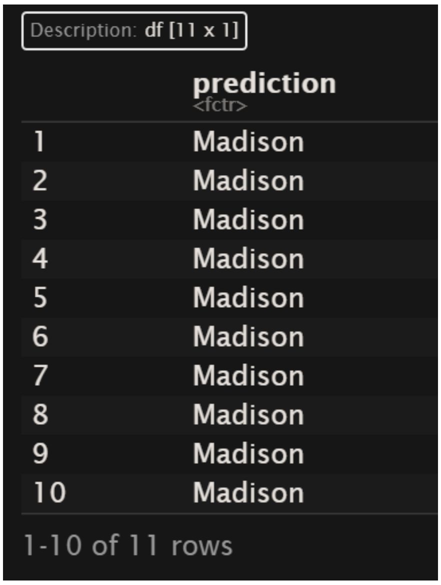
```
Confusion Matrix and Statistics

          Reference
Prediction Hamilton HM Jay Madison
  Hamilton       50  0   0       0
  HM              1  3   0       0
  Jay             0  0   5       0
  Madison         0  0   0      15

Overall Statistics

               Accuracy : 0.9865
                 95% CI : (0.927, 0.9997)
    No Information Rate : 0.6892
    P-Value [Acc > NIR] : 3.743e-11

                  Kappa : 0.9722

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Hamilton Class: HM Class: Jay Class: Madison
Sensitivity                   0.9804   1.00000    1.00000         1.0000
Specificity                   1.0000   0.98592    1.00000         1.0000
Pos Pred Value                1.0000   0.75000    1.00000         1.0000
Neg Pred Value                0.9583   1.00000    1.00000         1.0000
Prevalence                    0.6892   0.04054    0.06757         0.2027
Detection Rate                0.6757   0.04054    0.06757         0.2027
Detection Prevalence          0.6757   0.05405    0.06757         0.2027
Balanced Accuracy             0.9902   0.99296    1.00000         1.0000
```

Model 2: When we go by the optimal model using cross validation, we ended up getting the same results for training data. with an accuracy of 97%



```
training_pred Hamilton HM Jay Madison
     Hamilton       50  0   0       0
     HM              1  3   0       1
     Jay             0  0   5       0
     Madison         0  0   0      14
Confusion Matrix and Statistics

          Reference
Prediction Hamilton HM Jay Madison
  Hamilton       50  0   0       0
  HM              1  3   0       1
  Jay             0  0   5       0
  Madison         0  0   0      14

Overall Statistics

               Accuracy : 0.973
                 95% CI : (0.9058, 0.9967)
    No Information Rate : 0.6892
    P-Value [Acc > NIR] : 6.357e-10

                  Kappa : 0.9447

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Hamilton Class: HM Class: Jay Class: Madison
Sensitivity                   0.9804   1.00000    1.00000         0.9333
Specificity                   1.0000   0.97183    1.00000         1.0000
Pos Pred Value                1.0000   0.60000    1.00000         1.0000
Neg Pred Value                0.9583   1.00000    1.00000         0.9833
Prevalence                    0.6892   0.04054    0.06757         0.2027
Detection Rate                0.6757   0.04054    0.06757         0.1892
Detection Prevalence          0.6757   0.06757    0.06757         0.1892
Balanced Accuracy             0.9902   0.98592    1.00000         0.9667
Confusion Matrix and Statistics
```

When the optimal model was applied on testing data(Disputed files), it gave the following results. All disputed articles were classified to be written by madison.

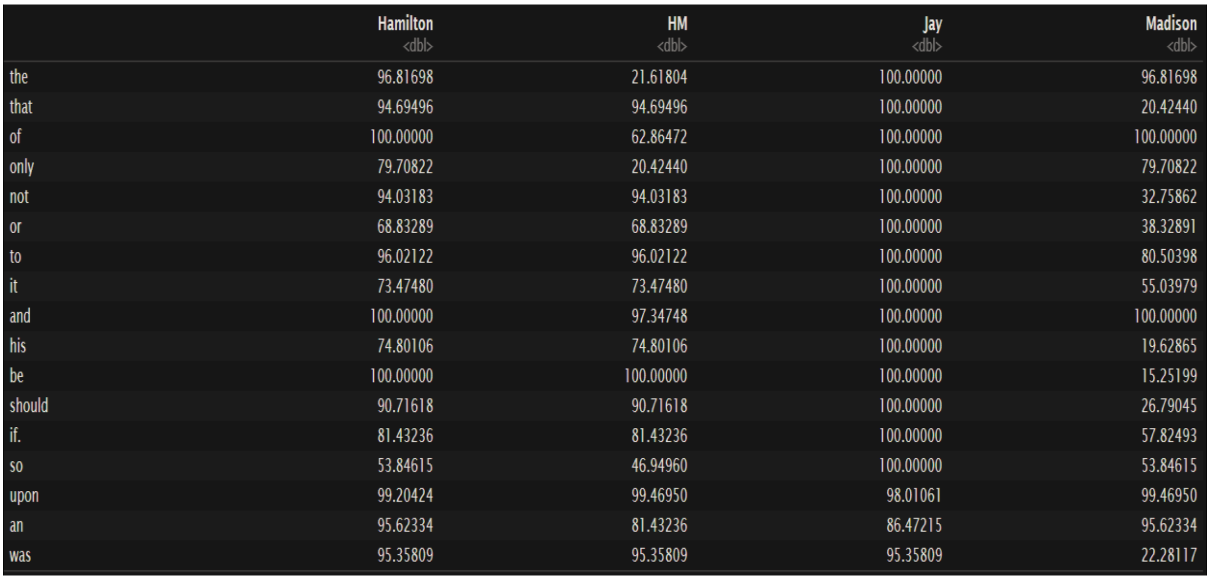


Description: df [11 x 1]

	prediction <fctr>
1	Madison
2	Madison
3	Madison
4	Madison
5	Madison
6	Madison
7	Madison
8	Madison
9	Madison
10	Madison

1-10 of 11 rows

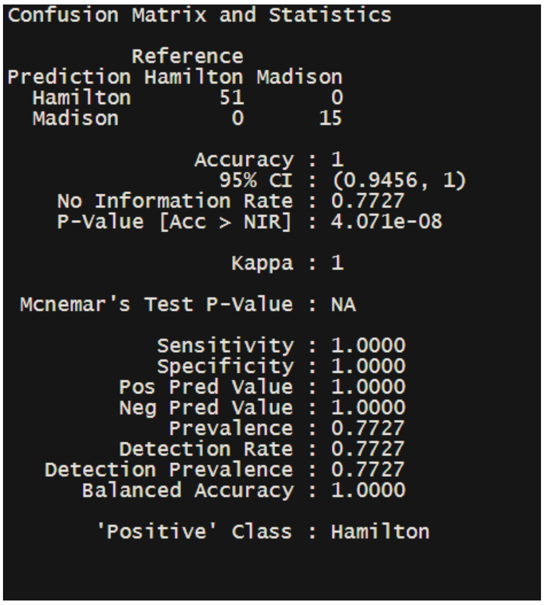
The following attributes were prioritised or worked well with each of the category(author)



	Hamilton <dbl>	HM <dbl>	Jay <dbl>	Madison <dbl>
the	96.81698	21.61804	100.00000	96.81698
that	94.69496	94.69496	100.00000	20.42440
of	100.00000	62.86472	100.00000	100.00000
only	79.70822	20.42440	100.00000	79.70822
not	94.03183	94.03183	100.00000	32.75862
or	68.83289	68.83289	100.00000	38.32891
to	96.02122	96.02122	100.00000	80.50398
it	73.47480	73.47480	100.00000	55.03979
and	100.00000	97.34748	100.00000	100.00000
his	74.80106	74.80106	100.00000	19.62865
be	100.00000	100.00000	100.00000	15.25199
should	90.71618	90.71618	100.00000	26.79045
if.	81.43236	81.43236	100.00000	57.82493
so	53.84615	46.94960	100.00000	53.84615
upon	99.20424	99.46950	98.01061	99.46950
an	95.62334	81.43236	86.47215	95.62334
was	95.35809	95.35809	95.35809	22.28117

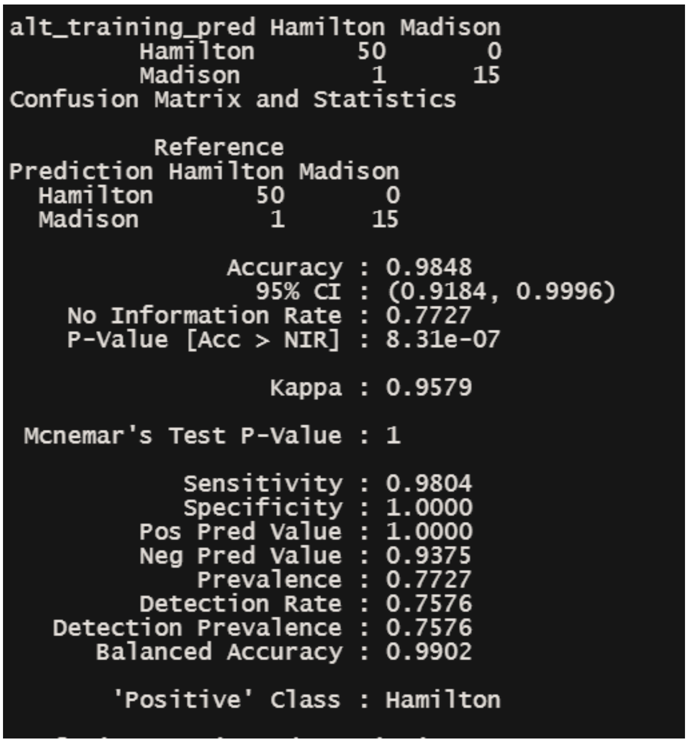
Case2 : excluding HM and Jay

Model 3: Simple DT with case2 data as training and testing. This is to see if we get 100% accuracy. We get 100% accuracy here. So it is understood that in the previous case, the HM category articles was potentially causing distortion.

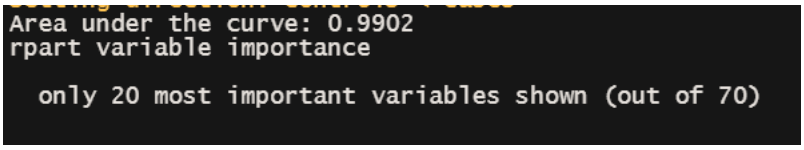


```
Confusion Matrix and Statistics

          Reference
Prediction Hamilton Madison
  Hamilton       51       0
  Madison         0      15

               Accuracy : 1
                 95% CI : (0.9456, 1)
    No Information Rate : 0.7727
    P-Value [Acc > NIR] : 4.071e-08

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.7727
         Detection Rate : 0.7727
   Detection Prevalence : 0.7727
      Balanced Accuracy : 1.0000

       'Positive' Class : Hamilton
```

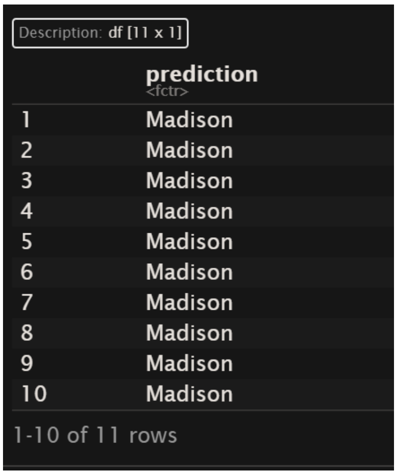
Model4: We tune the model and use rpart model instead of J48 to train with this data to classify the disputed articles. We use the model with 98% accuracy AND 99% AUC and get the following results



```
alt_training_pred Hamilton Madison
         Hamilton       50       0
         Madison         1      15
Confusion Matrix and Statistics

          Reference
Prediction Hamilton Madison
  Hamilton       50       0
  Madison         1      15

               Accuracy : 0.9848
                 95% CI : (0.9184, 0.9996)
    No Information Rate : 0.7727
    P-Value [Acc > NIR] : 8.31e-07

                  Kappa : 0.9579

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9804
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9375
             Prevalence : 0.7727
         Detection Rate : 0.7576
   Detection Prevalence : 0.7576
      Balanced Accuracy : 0.9902

       'Positive' Class : Hamilton
```



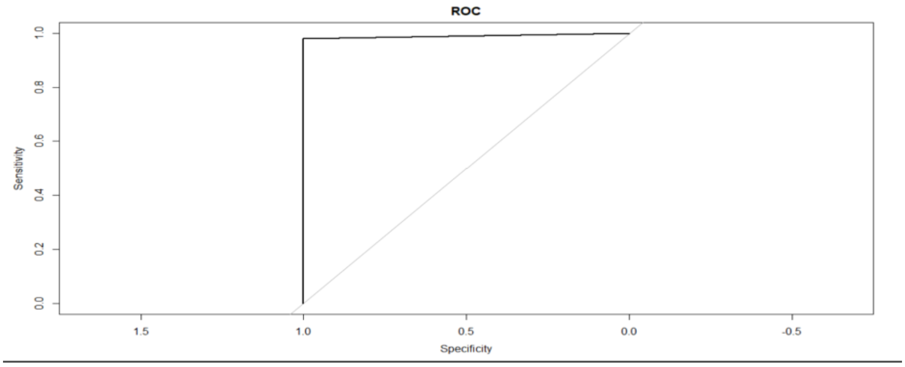
```
Area under the curve: 0.9902
rpart variable importance

  only 20 most important variables shown (out of 70)
```



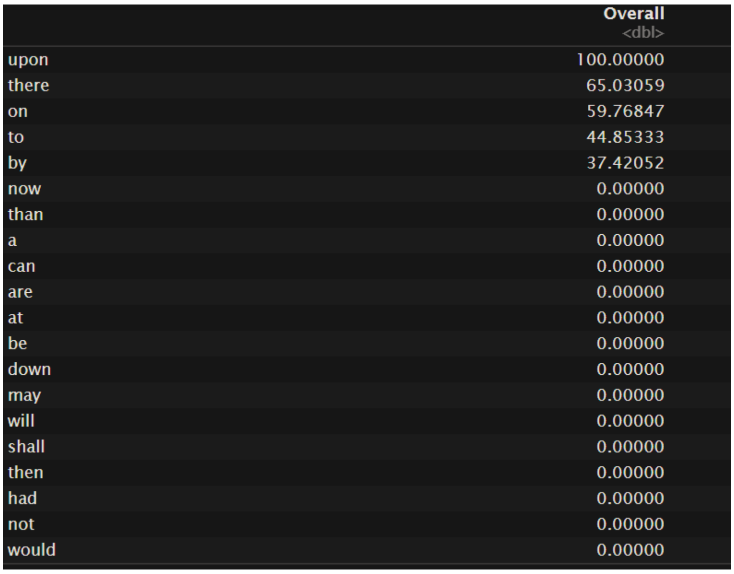
Description: df [11 x 1]

	prediction <fctr>
1	Madison
2	Madison
3	Madison
4	Madison
5	Madison
6	Madison
7	Madison
8	Madison
9	Madison
10	Madison

1-10 of 11 rows



In classifying the articles the following attributes were given more priority:



	Overall <dbl>
upon	100.00000
there	65.03059
on	59.76847
to	44.85333
by	37.42052
now	0.00000
than	0.00000
a	0.00000
can	0.00000
are	0.00000
at	0.00000
be	0.00000
down	0.00000
may	0.00000
will	0.00000
shall	0.00000
then	0.00000
had	0.00000
not	0.00000
would	0.00000

Inference:

From all the models tested, Considering the optimal models for both cases of data, the final predictions were the same(all articles were classified as Madison). The difference was in their accuracy measures. For the data that included HM and Jay articles/papers, the accuracy was only 97% where as for the model that excluded these cases, the accuracy was 98% with AUC 99%. The last model(model 4) turned out to be the best model. When we don't parameter tune the model, the default models seemed better but using them is not a good idea as they usually tend to overfit. Parameter tuning by giving complexity parameter(cp) values ensures that while post pruning, best results are achieved. It is the trade-off between the size of a tree and the error rate, which has to be less. So we considered smaller values to get better accuracy. The sensitivity for optimal models in the 2 cases, for Hamilton were the same(98%). We can consider all of our results and make inferences because the p-value for the models are statistically significant. For the optimal models, the words/attributes that were prioritised or which were used to classify the articles were the same. For Hamilton, in both cases, "used" was the root node with highest value.

In comparison to clustering, the results are much easier to interpret. But k-means and HAC clustering also classified the articles to be belonging to Madison, though it was a little difficult to interpret HAC output.