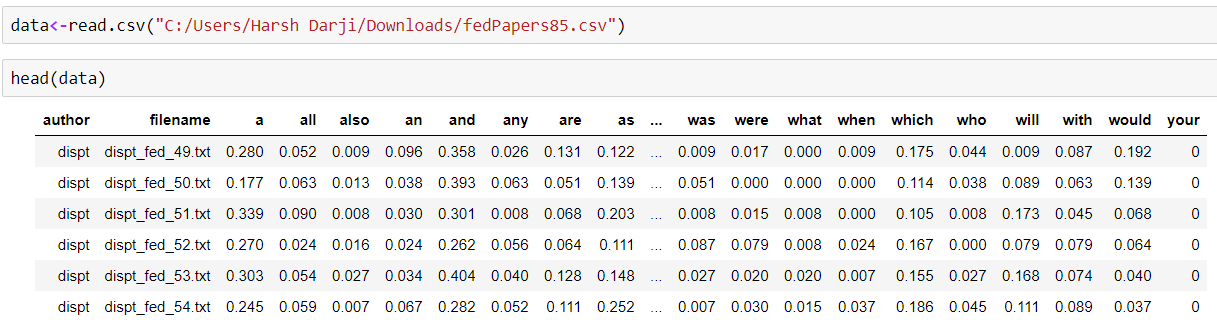
1. **Data Preparation**

First, we read the FedPapers csv file.



Now, as per the instruction we want our data to be divided into 3 parts, Training Data, Testing Data and Verification Data (to verify disputed data)

So, we create a data frame newdata which excludes disputed files.

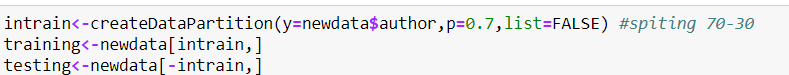


And my\_dispt\_data which contains only disputed files.



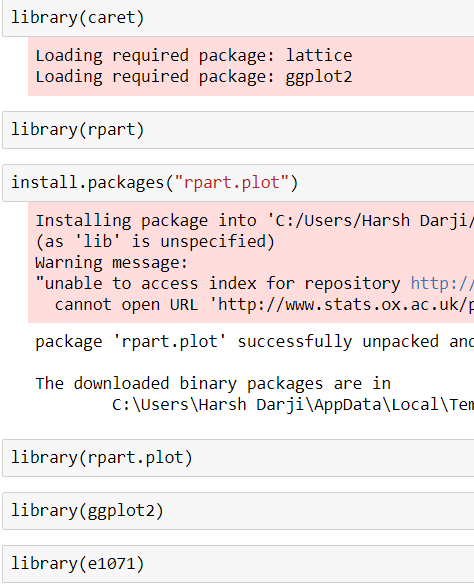
Now, we split our newdata data frame into training and testing dataset. Our training data consist of 70% of

our actual data and and testing data consist of 30% of our actual data.



1. **Build and tune decision tree models**

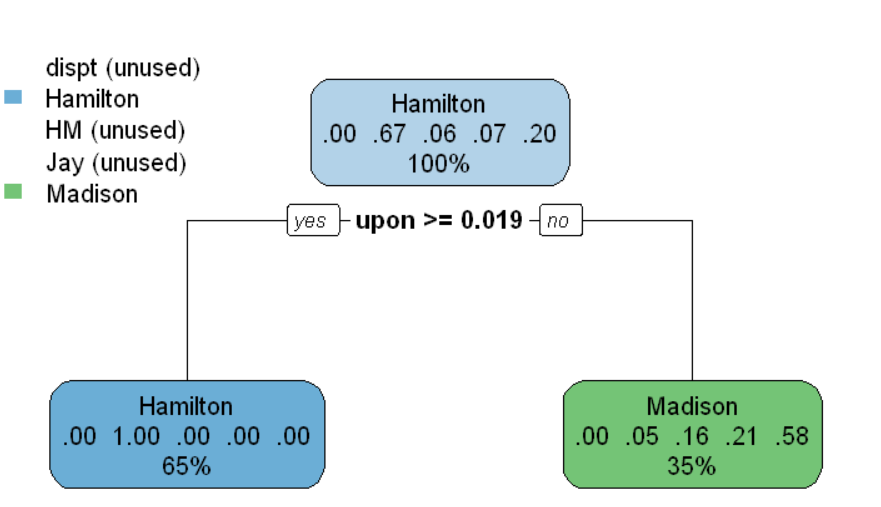
Loading the necessary packages to run decision tree algorithm on our dataset.



Now, let’s make a decision tree.



**Visualization:**



**Summary:**



Call:

rpart(formula = author ~ . - filename, data = training, method = "class",

control = rpart.control(cp = 0))

n= 54

CP nsplit rel error xerror xstd

1 0.5555556 0 1.0000000 1.0000000 0.1924501

2 0.0000000 1 0.4444444 0.5555556 0.1585831

Variable importance

upon there by on to and

25 17 16 16 14 12

Node number 1: 54 observations, complexity param=0.5555556

predicted class=Hamilton expected loss=0.3333333 P(node) =1

class counts: 0 36 3 4 11

probabilities: 0.000 0.667 0.056 0.074 0.204

left son=2 (35 obs) right son=3 (19 obs)

Primary splits:

upon < 0.019 to the right, improve=16.033140, (0 missing)

on < 0.0825 to the left, improve=10.574870, (0 missing)

there < 0.016 to the right, improve= 9.485770, (0 missing)

to < 0.554 to the right, improve= 8.703704, (0 missing)

by < 0.1385 to the left, improve= 8.608061, (0 missing)

Surrogate splits:

there < 0.016 to the right, agree=0.889, adj=0.684, (0 split)

by < 0.1385 to the left, agree=0.870, adj=0.632, (0 split)

on < 0.0745 to the left, agree=0.870, adj=0.632, (0 split)

to < 0.4745 to the right, agree=0.852, adj=0.579, (0 split)

and < 0.421 to the left, agree=0.815, adj=0.474, (0 split)

Node number 2: 35 observations

predicted class=Hamilton expected loss=0 P(node) =0.6481481

class counts: 0 35 0 0 0

probabilities: 0.000 1.000 0.000 0.000 0.000

Node number 3: 19 observations

predicted class=Madison expected loss=0.4210526 P(node) =0.3518519

class counts: 0 1 3 4 11

probabilities: 0.000 0.053 0.158 0.211 0.579



A screenshot of a cell phone

Description generated with high confidence



Gives a visual representation of the cross-validation results in an rpart object.

A screenshot of a cell phone

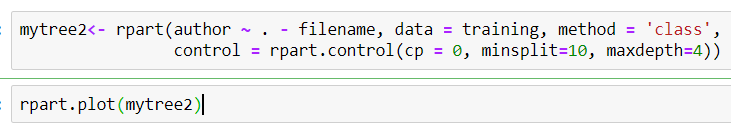
Description generated with very high confidence

**Inference:**

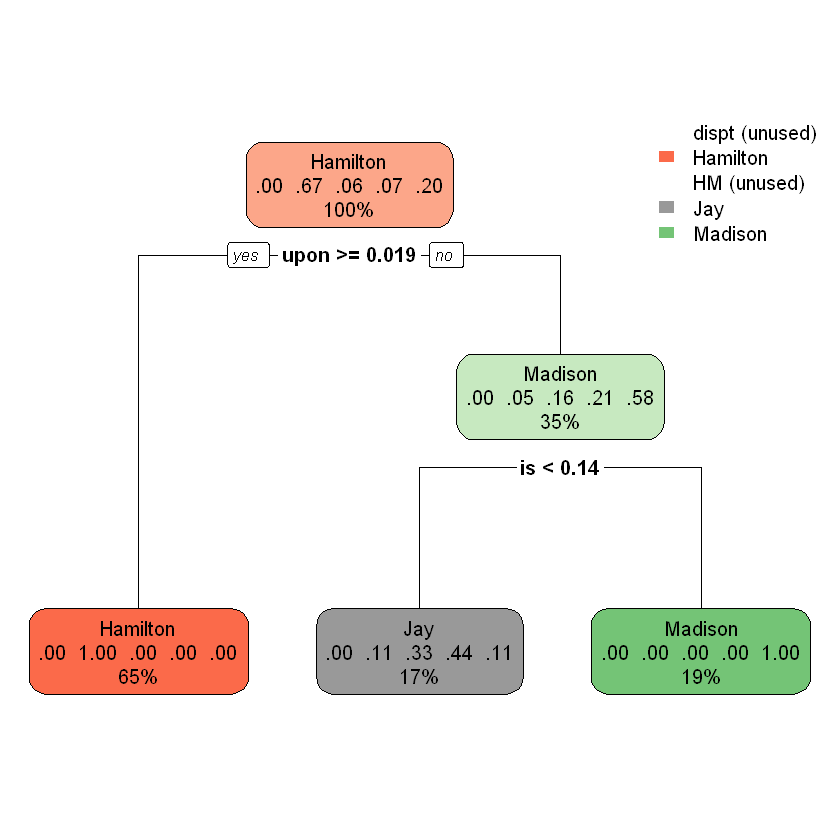
From there, the model asks whether the word ‘upon’ is used above 1.9 percent. If the ‘upon’ appears more frequently, then the model asigns the paper to Hamilton. But, if ‘upon’ is used less than 1.9 percent, then the model immediately ascribes the text to Madison. It is interesting that - at this point in the model - no papers are assigned to Jay or to the dual Hamilton-Madison papers. The model will have to be tuned to fix this. Furthermore, when checking the model summary, we find that at two splits, the relative error of the model is 0.4444. (Relative error =1-Rsquare) This is not desirable as we want a more accurate model.

**Model 2:**

In this model we change our minimum split to 10 papers in the bucket and max depth of 4 leaf nodes.



**Visualization:**



This is a model with a **multi-class response**. Each node shows

* the predicted class (author name),
* the predicted probability of each class,
* the percentage of observations in the node

**Summary:**



Call:

rpart(formula = author ~ . - filename, data = training, method = "class",

control = rpart.control(cp = 0, minsplit = 10, maxdepth = 4))

n= 54

CP nsplit rel error xerror xstd

1 0.5555556 0 1.0000000 1.0000000 0.1924501

2 0.1666667 1 0.4444444 0.5555556 0.1585831

3 0.0000000 2 0.2777778 0.7222222 0.1745397

Variable importance

upon and there by on to is when been every no

18 13 13 12 12 11 6 5 4 4 4

Node number 1: 54 observations, complexity param=0.5555556

predicted class=Hamilton expected loss=0.3333333 P(node) =1

class counts: 0 36 3 4 11

probabilities: 0.000 0.667 0.056 0.074 0.204

left son=2 (35 obs) right son=3 (19 obs)

Primary splits:

upon < 0.019 to the right, improve=16.033140, (0 missing)

on < 0.0825 to the left, improve=10.574870, (0 missing)

there < 0.016 to the right, improve= 9.485770, (0 missing)

to < 0.554 to the right, improve= 8.703704, (0 missing)

by < 0.1385 to the left, improve= 8.608061, (0 missing)

Surrogate splits:

there < 0.016 to the right, agree=0.889, adj=0.684, (0 split)

by < 0.1385 to the left, agree=0.870, adj=0.632, (0 split)

on < 0.0745 to the left, agree=0.870, adj=0.632, (0 split)

to < 0.4745 to the right, agree=0.852, adj=0.579, (0 split)

and < 0.421 to the left, agree=0.815, adj=0.474, (0 split)

Node number 2: 35 observations

predicted class=Hamilton expected loss=0 P(node) =0.6481481

class counts: 0 35 0 0 0

probabilities: 0.000 1.000 0.000 0.000 0.000

Node number 3: 19 observations, complexity param=0.1666667

predicted class=Madison expected loss=0.4210526 P(node) =0.3518519

class counts: 0 1 3 4 11

probabilities: 0.000 0.053 0.158 0.211 0.579

left son=6 (9 obs) right son=7 (10 obs)

Primary splits:

is < 0.1405 to the left, improve=5.263158, (0 missing)

of < 0.7305 to the left, improve=4.996491, (0 missing)

the < 1.0785 to the left, improve=4.996491, (0 missing)

and < 0.5955 to the right, improve=4.996491, (0 missing)

which < 0.113 to the left, improve=4.878543, (0 missing)

Surrogate splits:

when < 0.0135 to the right, agree=0.895, adj=0.778, (0 split)

and < 0.4925 to the right, agree=0.842, adj=0.667, (0 split)

been < 0.0345 to the left, agree=0.842, adj=0.667, (0 split)

every < 0.019 to the left, agree=0.842, adj=0.667, (0 split)

no < 0.0365 to the left, agree=0.842, adj=0.667, (0 split)

Node number 6: 9 observations

predicted class=Jay expected loss=0.5555556 P(node) =0.1666667

class counts: 0 1 3 4 1

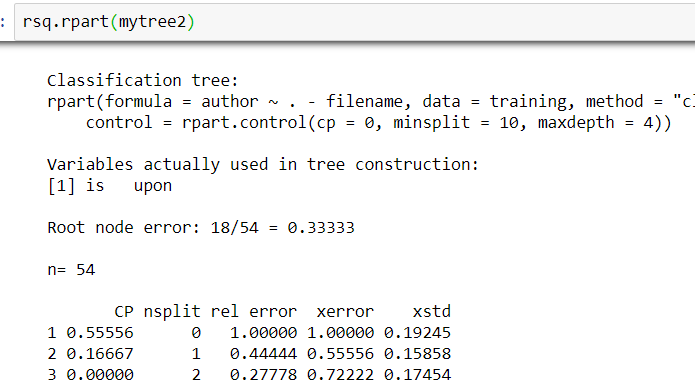
probabilities: 0.000 0.111 0.333 0.444 0.111

Node number 7: 10 observations

predicted class=Madison expected loss=0 P(node) =0.1851852

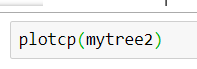
class counts: 0 0 0 0 10

probabilities: 0.000 0.000 0.000 0.000 1.000



A close up of a map

Description generated with high confidence

A close up of a map

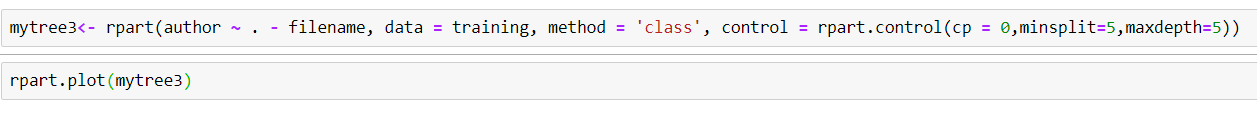
Description generated with very high confidence

**Inference:**

This tree is much more accurate, producing a relative error of 0.27 at three splits. We also see that the tree further elaborates beyond the Madison leaf node, going so far as to identify the Jay papers. Returning to the Madison leaf node, if the author uses ‘is’ less than 14% percent of the time, then it is a Jay paper. Anything more is still attributed to Madison. Unfortunately, the model is still failing to determine the Hamilton-Madison papers.

**Model 3:**

In this model we set min split to 5 and max depth to 5 as well.

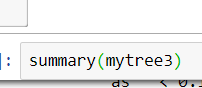


**Visualization:**

A screenshot of a cell phone

Description generated with high confidence

**Summary:**



Call:

rpart(formula = author ~ . - filename, data = training, method = "class",

control = rpart.control(cp = 0, minsplit = 5, maxdepth = 5))

n= 54

CP nsplit rel error xerror xstd

1 0.55555556 0 1.00000000 1.0000000 0.1924501

2 0.16666667 1 0.44444444 0.5555556 0.1585831

3 0.05555556 3 0.11111111 0.5555556 0.1585831

4 0.00000000 4 0.05555556 0.5555556 0.1585831

Variable importance

and upon there by on to been is when every no it of

14 14 9 9 9 8 5 4 3 3 3 3 3

only so the a as be can

3 3 3 2 2 2 2

Node number 1: 54 observations, complexity param=0.5555556

predicted class=Hamilton expected loss=0.3333333 P(node) =1

class counts: 0 36 3 4 11

probabilities: 0.000 0.667 0.056 0.074 0.204

left son=2 (35 obs) right son=3 (19 obs)

Primary splits:

upon < 0.019 to the right, improve=16.033140, (0 missing)

on < 0.0825 to the left, improve=10.574870, (0 missing)

there < 0.016 to the right, improve= 9.485770, (0 missing)

to < 0.554 to the right, improve= 8.703704, (0 missing)

by < 0.1385 to the left, improve= 8.608061, (0 missing)

Surrogate splits:

there < 0.016 to the right, agree=0.889, adj=0.684, (0 split)

by < 0.1385 to the left, agree=0.870, adj=0.632, (0 split)

on < 0.0745 to the left, agree=0.870, adj=0.632, (0 split)

to < 0.4745 to the right, agree=0.852, adj=0.579, (0 split)

and < 0.421 to the left, agree=0.815, adj=0.474, (0 split)

Node number 2: 35 observations

predicted class=Hamilton expected loss=0 P(node) =0.6481481

class counts: 0 35 0 0 0

probabilities: 0.000 1.000 0.000 0.000 0.000

Node number 3: 19 observations, complexity param=0.1666667

predicted class=Madison expected loss=0.4210526 P(node) =0.3518519

class counts: 0 1 3 4 11

probabilities: 0.000 0.053 0.158 0.211 0.579

left son=6 (9 obs) right son=7 (10 obs)

Primary splits:

is < 0.1405 to the left, improve=5.263158, (0 missing)

of < 0.7305 to the left, improve=4.996491, (0 missing)

the < 1.0785 to the left, improve=4.996491, (0 missing)

and < 0.5955 to the right, improve=4.996491, (0 missing)

which < 0.113 to the left, improve=4.878543, (0 missing)

Surrogate splits:

when < 0.0135 to the right, agree=0.895, adj=0.778, (0 split)

and < 0.4925 to the right, agree=0.842, adj=0.667, (0 split)

been < 0.0345 to the left, agree=0.842, adj=0.667, (0 split)

every < 0.019 to the left, agree=0.842, adj=0.667, (0 split)

no < 0.0365 to the left, agree=0.842, adj=0.667, (0 split)

Node number 6: 9 observations, complexity param=0.1666667

predicted class=Jay expected loss=0.5555556 P(node) =0.1666667

class counts: 0 1 3 4 1

probabilities: 0.000 0.111 0.333 0.444 0.111

left son=12 (5 obs) right son=13 (4 obs)

Primary splits:

and < 0.5835 to the left, improve=3.2, (0 missing)

it < 0.153 to the left, improve=3.2, (0 missing)

of < 0.7305 to the right, improve=3.2, (0 missing)

only < 0.0355 to the left, improve=3.2, (0 missing)

so < 0.0255 to the left, improve=3.2, (0 missing)

Surrogate splits:

it < 0.153 to the left, agree=1, adj=1, (0 split)

of < 0.7305 to the right, agree=1, adj=1, (0 split)

only < 0.0355 to the left, agree=1, adj=1, (0 split)

so < 0.0255 to the left, agree=1, adj=1, (0 split)

the < 1.0785 to the right, agree=1, adj=1, (0 split)

Node number 7: 10 observations

predicted class=Madison expected loss=0 P(node) =0.1851852

class counts: 0 0 0 0 10

probabilities: 0.000 0.000 0.000 0.000 1.000

Node number 12: 5 observations, complexity param=0.05555556

predicted class=HM expected loss=0.4 P(node) =0.09259259

class counts: 0 1 3 0 1

probabilities: 0.000 0.200 0.600 0.000 0.200

left son=24 (3 obs) right son=25 (2 obs)

Primary splits:

a < 0.279 to the left, improve=1.8, (0 missing)

and < 0.4115 to the right, improve=1.8, (0 missing)

as < 0.111 to the left, improve=1.8, (0 missing)

be < 0.229 to the left, improve=1.8, (0 missing)

been < 0.0365 to the left, improve=1.8, (0 missing)

Surrogate splits:

and < 0.4115 to the right, agree=1, adj=1, (0 split)

as < 0.111 to the left, agree=1, adj=1, (0 split)

be < 0.229 to the left, agree=1, adj=1, (0 split)

been < 0.0365 to the left, agree=1, adj=1, (0 split)

can < 0.0155 to the left, agree=1, adj=1, (0 split)

Node number 13: 4 observations

predicted class=Jay expected loss=0 P(node) =0.07407407

class counts: 0 0 0 4 0

probabilities: 0.000 0.000 0.000 1.000 0.000

Node number 24: 3 observations

predicted class=HM expected loss=0 P(node) =0.05555556

class counts: 0 0 3 0 0

probabilities: 0.000 0.000 1.000 0.000 0.000

Node number 25: 2 observations

predicted class=Hamilton expected loss=0.5 P(node) =0.03703704

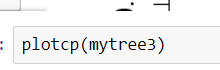
class counts: 0 1 0 0 1

probabilities: 0.000 0.500 0.000 0.000 0.500



A close up of a map

Description generated with very high confidence



A close up of a map

Description generated with very high confidence

**Inference:**

This model gives us a relative error of 0.055, which is amazing. Most of the papers are being assigned correctly to Hamilton, Jay, and Madison. In the jay leaf node if ‘and’ is used less than 58% time then it’s assigned to HM, otherwise to Jay. We can assume that both Hamilton’s and the Hamilton Madison papers are being classified here.

**Testing:**



**Output:**

Confusion Matrix and Statistics

Reference

Prediction dispt Hamilton HM Jay Madison

dispt 0 0 0 0 0

Hamilton 0 15 0 0 0

HM 0 0 0 0 2

Jay 0 0 0 1 0

Madison 0 0 0 0 2

Overall Statistics

Accuracy : 0.9

95% CI : (0.683, 0.9877)

No Information Rate : 0.75

P-Value [Acc > NIR] : 0.09126

Kappa : 0.759

Mcnemar's Test P-Value : NA

Statistics by Class:

Class: dispt Class: Hamilton Class: HM Class: Jay

Sensitivity NA 1.00 NA 1.00

Specificity 1 1.00 0.9 1.00

Pos Pred Value NA 1.00 NA 1.00

Neg Pred Value NA 1.00 NA 1.00

Prevalence 0 0.75 0.0 0.05

Detection Rate 0 0.75 0.0 0.05

Detection Prevalence 0 0.75 0.1 0.05

Balanced Accuracy NA 1.00 NA 1.00

Class: Madison

Sensitivity 0.5000

Specificity 1.0000

Pos Pred Value 1.0000

Neg Pred Value 0.8889

Prevalence 0.2000

Detection Rate 0.1000

Detection Prevalence 0.1000

Balanced Accuracy 0.7500

**Inference:**

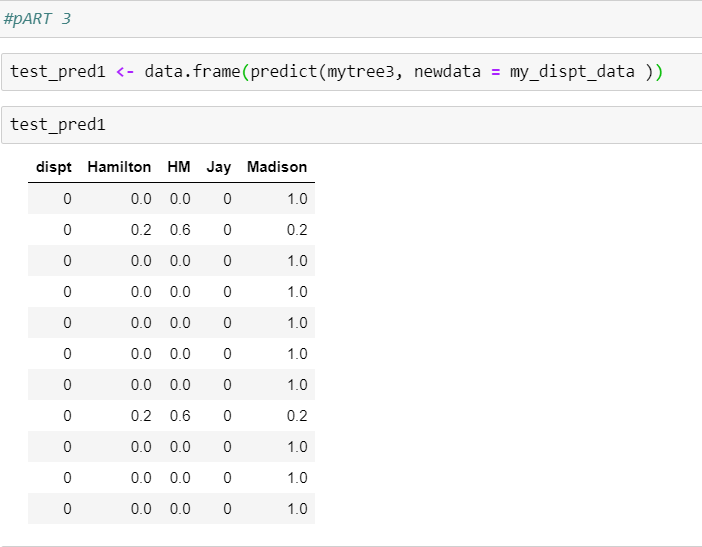
Our model has a predicting capability of 90%. By checking the confusion matrix, we find that the model

correctly assigns the Hamilton, Jay and Madison papers to each author. However, there seem to be a

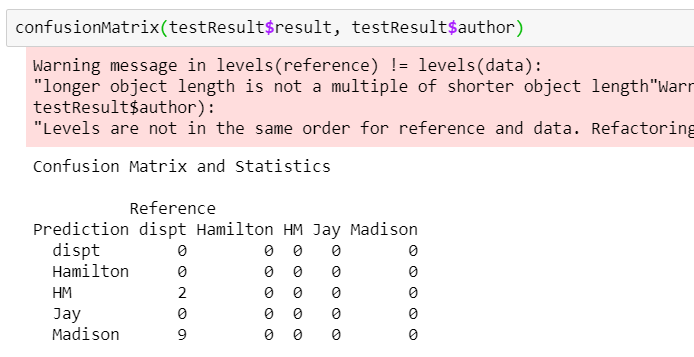
problem with HM papers, here 2 HM papers are assigned to Madison, Which s the only drawback of the model.

1. **Prediction**

**Model:**



**Ouput:**



**Inference:**

Out of 11 disputed papers, according to our decision tree model 2 are either written by Hamilton or Madison or collaboration and 9 of them are written by Madison. However, when we applied clustering technique, we had enough evidence to say that 8 papers were written by Madison and 3 by Hamilton.