Student: Alireza Zarrinmehr

SUID: 542380864

IST707 Applied Machine Learning

HW5: Use Decision Tree to Solve a Mystery in History

This week, we are going to use the decision tree algorithm to solve the disputed essay problem.

Section 1: Data preparation

You will need to separate the original data set to training and testing data for classification experiments. Describe what examples are in your training and what are in your test data.

We don't need any of the observations with the disputed authors in the train data set as in the training step we need the data to be accurate leading to a good model. However, in the test data set, we need the rows with the disputed author as the goal is to predict the authors. We also need some of the rows with the known authors to compare the predictions with the real values.

First, the rows with the disputed author were removed from the original data set, and then the data set was divided into training and test data. Then the rows that were removed from the original dataset earlier were added to the TEST data set. The following pictures show the process of creating the test and training data:

In [3]:
```
import pandas as pd
import sklearn.model_selection
from sklearn.model_selection import train_test_split
df = pd.read_csv("/Users/zarrinmehr/Desktop/11- IST.707.M002.FALL22.Applied Machine Learning 16797.1231/week 5/HW 5/HW4
df.head(12)
```

Out[3]:

	author	filename	a	all	also	an	and	any	are	as	...	was	were	what	when	which	who	will	with	would	your
0	dispt	dispt_fed_49.txt	0.280	0.052	0.009	0.096	0.358	0.026	0.131	0.122	...	0.009	0.017	0.000	0.009	0.175	0.044	0.009	0.087	0.192	0.000
1	dispt	dispt_fed_50.txt	0.177	0.063	0.013	0.038	0.393	0.063	0.051	0.139	...	0.051	0.000	0.000	0.000	0.114	0.038	0.089	0.063	0.139	0.000
2	dispt	dispt_fed_51.txt	0.339	0.090	0.008	0.030	0.301	0.008	0.068	0.203	...	0.008	0.015	0.008	0.000	0.105	0.008	0.173	0.045	0.068	0.000
3	dispt	dispt_fed_52.txt	0.270	0.024	0.016	0.024	0.262	0.056	0.064	0.111	...	0.087	0.079	0.008	0.024	0.167	0.000	0.079	0.079	0.064	0.000
4	dispt	dispt_fed_53.txt	0.303	0.054	0.027	0.034	0.404	0.040	0.128	0.148	...	0.027	0.020	0.020	0.007	0.155	0.027	0.168	0.074	0.040	0.000
5	dispt	dispt_fed_54.txt	0.245	0.059	0.007	0.067	0.282	0.052	0.111	0.252	...	0.007	0.030	0.015	0.037	0.186	0.045	0.111	0.089	0.037	0.000
6	dispt	dispt_fed_55.txt	0.349	0.036	0.007	0.029	0.335	0.058	0.087	0.073	...	0.015	0.029	0.015	0.007	0.211	0.022	0.145	0.073	0.073	0.000
7	dispt	dispt_fed_56.txt	0.414	0.083	0.009	0.018	0.478	0.046	0.110	0.074	...	0.018	0.009	0.009	0.000	0.175	0.018	0.267	0.129	0.037	0.000
8	dispt	dispt_fed_57.txt	0.248	0.040	0.007	0.040	0.356	0.034	0.154	0.161	...	0.027	0.007	0.020	0.020	0.201	0.040	0.154	0.027	0.040	0.000
9	dispt	dispt_fed_62.txt	0.442	0.062	0.006	0.075	0.423	0.037	0.093	0.100	...	0.000	0.000	0.025	0.012	0.199	0.031	0.106	0.081	0.031	0.000
10	dispt	dispt_fed_63.txt	0.276	0.048	0.015	0.082	0.324	0.044	0.058	0.135	...	0.044	0.024	0.005	0.015	0.174	0.015	0.092	0.077	0.053	0.000
11	Hamilton	Hamilton_fed_1.txt	0.213	0.083	0.000	0.083	0.343	0.056	0.111	0.093	...	0.000	0.000	0.000	0.009	0.158	0.074	0.222	0.046	0.019	0.074

12 rows × 72 columns

Figure 1: Loading the library and importing the data

```
In [5]: dfDispt = df[df.author == 'dispt']
        dfDispt
```

Out[5]:

	author	filename	a	all	also	an	and	any	are	as	...	was	were	what	when	which	who	will	with	would	your
0	dispt	dispt_fed_49.txt	0.280	0.052	0.009	0.096	0.358	0.026	0.131	0.122	...	0.009	0.017	0.000	0.009	0.175	0.044	0.009	0.087	0.192	0.0
1	dispt	dispt_fed_50.txt	0.177	0.063	0.013	0.038	0.393	0.063	0.051	0.139	...	0.051	0.000	0.000	0.000	0.114	0.038	0.089	0.063	0.139	0.0
2	dispt	dispt_fed_51.txt	0.339	0.090	0.008	0.030	0.301	0.008	0.068	0.203	...	0.008	0.015	0.008	0.000	0.105	0.008	0.173	0.045	0.068	0.0
3	dispt	dispt_fed_52.txt	0.270	0.024	0.016	0.024	0.262	0.056	0.064	0.111	...	0.087	0.079	0.008	0.024	0.167	0.000	0.079	0.079	0.064	0.0
4	dispt	dispt_fed_53.txt	0.303	0.054	0.027	0.034	0.404	0.040	0.128	0.148	...	0.027	0.020	0.020	0.007	0.155	0.027	0.168	0.074	0.040	0.0
5	dispt	dispt_fed_54.txt	0.245	0.059	0.007	0.067	0.282	0.052	0.111	0.252	...	0.007	0.030	0.015	0.037	0.186	0.045	0.111	0.089	0.037	0.0
6	dispt	dispt_fed_55.txt	0.349	0.036	0.007	0.029	0.335	0.058	0.087	0.073	...	0.015	0.029	0.015	0.007	0.211	0.022	0.145	0.073	0.073	0.0
7	dispt	dispt_fed_56.txt	0.414	0.083	0.009	0.018	0.478	0.046	0.110	0.074	...	0.018	0.009	0.009	0.000	0.175	0.018	0.267	0.129	0.037	0.0
8	dispt	dispt_fed_57.txt	0.248	0.040	0.007	0.040	0.356	0.034	0.154	0.161	...	0.027	0.007	0.020	0.020	0.201	0.040	0.154	0.027	0.040	0.0
9	dispt	dispt_fed_62.txt	0.442	0.062	0.006	0.075	0.423	0.037	0.093	0.100	...	0.000	0.000	0.025	0.012	0.199	0.031	0.106	0.081	0.031	0.0
10	dispt	dispt_fed_63.txt	0.276	0.048	0.015	0.082	0.324	0.044	0.058	0.135	...	0.044	0.024	0.005	0.015	0.174	0.015	0.092	0.077	0.053	0.0

11 rows × 72 columns

Figure 2: Storing the observations with the disputed author as dfDispt

```
In [6]: dfOther = df.drop(range(0,11))
        dfOther
```

Out[6]:

	author	filename	a	all	also	an	and	any	are	as	...	was	were	what	when	which	who	will	with	would	your
11	Hamilton	Hamilton_fed_1.txt	0.213	0.083	0.000	0.083	0.343	0.056	0.111	0.093	...	0.000	0.000	0.000	0.009	0.158	0.074	0.222	0.046	0.019	0.074
12	Hamilton	Hamilton_fed_11.txt	0.369	0.070	0.006	0.076	0.411	0.023	0.053	0.117	...	0.000	0.012	0.012	0.012	0.147	0.029	0.094	0.129	0.270	0.000
13	Hamilton	Hamilton_fed_12.txt	0.305	0.047	0.007	0.068	0.386	0.047	0.102	0.108	...	0.000	0.000	0.007	0.000	0.156	0.007	0.074	0.122	0.149	0.000
14	Hamilton	Hamilton_fed_13.txt	0.391	0.045	0.015	0.030	0.270	0.045	0.060	0.090	...	0.000	0.000	0.000	0.045	0.165	0.045	0.135	0.150	0.210	0.000
15	Hamilton	Hamilton_fed_15.txt	0.327	0.096	0.000	0.086	0.356	0.014	0.086	0.072	...	0.014	0.038	0.014	0.019	0.264	0.029	0.091	0.086	0.062	0.010
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
80	Madison	Madison_fed_45.txt	0.136	0.054	0.014	0.048	0.422	0.027	0.048	0.150	...	0.020	0.027	0.007	0.000	0.116	0.007	0.218	0.102	0.075	0.000
81	Madison	Madison_fed_46.txt	0.212	0.028	0.006	0.050	0.391	0.033	0.073	0.117	...	0.067	0.011	0.022	0.000	0.128	0.028	0.223	0.095	0.162	0.000
82	Madison	Madison_fed_47.txt	0.177	0.052	0.047	0.047	0.436	0.026	0.135	0.083	...	0.021	0.021	0.010	0.010	0.114	0.031	0.016	0.099	0.021	0.000
83	Madison	Madison_fed_48.txt	0.243	0.091	0.008	0.084	0.372	0.008	0.046	0.137	...	0.023	0.023	0.008	0.000	0.213	0.038	0.076	0.061	0.023	0.000
84	Madison	Madison_fed_58.txt	0.347	0.097	0.007	0.056	0.313	0.035	0.049	0.132	...	0.007	0.007	0.000	0.014	0.188	0.035	0.257	0.083	0.083	0.000

74 rows × 72 columns

Figure 3: Removing the observations with the disputed author from the data set

```
In [8]: training_data, testing_data = train_test_split(dfOther, test_size=0.2, random_state=25)

        print(f"No. of training examples: {training_data.shape[0]}")
        print(f"No. of testing examples: {testing_data.shape[0]}")
```

```
No. of training examples: 59
No. of testing examples: 15
```

Figure 4: Dividing the data set to test and train data using scikit-learn

```
In [10]: testing_data_Dispt = pd.concat([testing_data, dfDispt])
         testing_data_Dispt
```

Out[10]:

	author	filename	a	all	also	an	and	any	are	as	...	was	were	what	when	which	who	will	with	would	your
49	Hamilton	Hamilton_fed_75.txt	0.402	0.039	0.008	0.077	0.247	0.015	0.054	0.124	...	0.015	0.000	0.000	0.023	0.131	0.015	0.046	0.108	0.209	0.000
75	Madison	Madison_fed_40.txt	0.250	0.043	0.014	0.058	0.461	0.010	0.062	0.144	...	0.058	0.091	0.019	0.014	0.139	0.043	0.024	0.072	0.029	0.000
11	Hamilton	Hamilton_fed_1.txt	0.213	0.083	0.000	0.083	0.343	0.056	0.111	0.093	...	0.000	0.000	0.000	0.009	0.158	0.074	0.222	0.046	0.019	0.074
44	Hamilton	Hamilton_fed_70.txt	0.298	0.028	0.000	0.085	0.360	0.066	0.099	0.099	...	0.028	0.033	0.028	0.014	0.128	0.076	0.052	0.090	0.052	0.000
40	Hamilton	Hamilton_fed_67.txt	0.203	0.035	0.000	0.071	0.406	0.018	0.062	0.132	...	0.009	0.000	0.000	0.000	0.221	0.053	0.026	0.106	0.044	0.000
17	Hamilton	Hamilton_fed_17.txt	0.261	0.108	0.000	0.072	0.467	0.018	0.045	0.027	...	0.054	0.036	0.009	0.000	0.216	0.018	0.117	0.090	0.108	0.000
66	Jay	Jay_fed_3.txt	0.130	0.040	0.030	0.030	0.601	0.050	0.080	0.221	...	0.000	0.010	0.000	0.010	0.110	0.030	0.241	0.100	0.010	0.000
24	Hamilton	Hamilton_fed_27.txt	0.270	0.040	0.000	0.050	0.320	0.070	0.100	0.080	...	0.010	0.000	0.020	0.010	0.190	0.010	0.340	0.060	0.030	0.000
46	Hamilton	Hamilton_fed_72.txt	0.378	0.029	0.000	0.073	0.364	0.015	0.051	0.109	...	0.029	0.029	0.022	0.065	0.145	0.029	0.058	0.065	0.175	0.000
62	HM	HM_fed_18.txt	0.229	0.040	0.000	0.034	0.532	0.013	0.013	0.081	...	0.189	0.108	0.000	0.020	0.081	0.074	0.007	0.074	0.040	0.000
77	Madison	Madison_fed_42.txt	0.240	0.042	0.000	0.078	0.491	0.057	0.047	0.115	...	0.016	0.016	0.005	0.010	0.141	0.021	0.031	0.083	0.078	0.000
82	Madison	Madison_fed_47.txt	0.177	0.052	0.047	0.047	0.436	0.026	0.135	0.083	...	0.021	0.021	0.010	0.010	0.114	0.031	0.016	0.099	0.021	0.000
23	Hamilton	Hamilton_fed_26.txt	0.329	0.051	0.000	0.133	0.297	0.063	0.038	0.088	...	0.063	0.044	0.006	0.019	0.133	0.019	0.070	0.051	0.101	0.000
80	Madison	Madison_fed_45.txt	0.136	0.054	0.014	0.048	0.422	0.027	0.048	0.150	...	0.020	0.027	0.007	0.000	0.116	0.007	0.218	0.102	0.075	0.000
53	Hamilton	Hamilton_fed_79.txt	0.270	0.028	0.000	0.014	0.327	0.114	0.085	0.057	...	0.014	0.000	0.000	0.014	0.171	0.057	0.085	0.057	0.071	0.000
0	dispt	dispt_fed_49.txt	0.280	0.052	0.009	0.096	0.358	0.026	0.131	0.122	...	0.009	0.017	0.000	0.009	0.175	0.044	0.009	0.087	0.192	0.000
1	dispt	dispt_fed_50.txt	0.177	0.063	0.013	0.038	0.393	0.063	0.051	0.139	...	0.051	0.000	0.000	0.000	0.114	0.038	0.089	0.063	0.139	0.000
2	dispt	dispt_fed_51.txt	0.339	0.090	0.008	0.030	0.301	0.008	0.068	0.203	...	0.008	0.015	0.008	0.000	0.105	0.008	0.173	0.045	0.068	0.000
3	dispt	dispt_fed_52.txt	0.270	0.024	0.016	0.024	0.262	0.056	0.064	0.111	...	0.087	0.079	0.008	0.024	0.167	0.000	0.079	0.079	0.064	0.000

Figure 5: Adding the observation with the disputed author (dfDispt) to the test data set

```
In [11]: training_data.to_csv('Train.csv')
         testing_data_Dispt.to_csv('Test.csv')
```

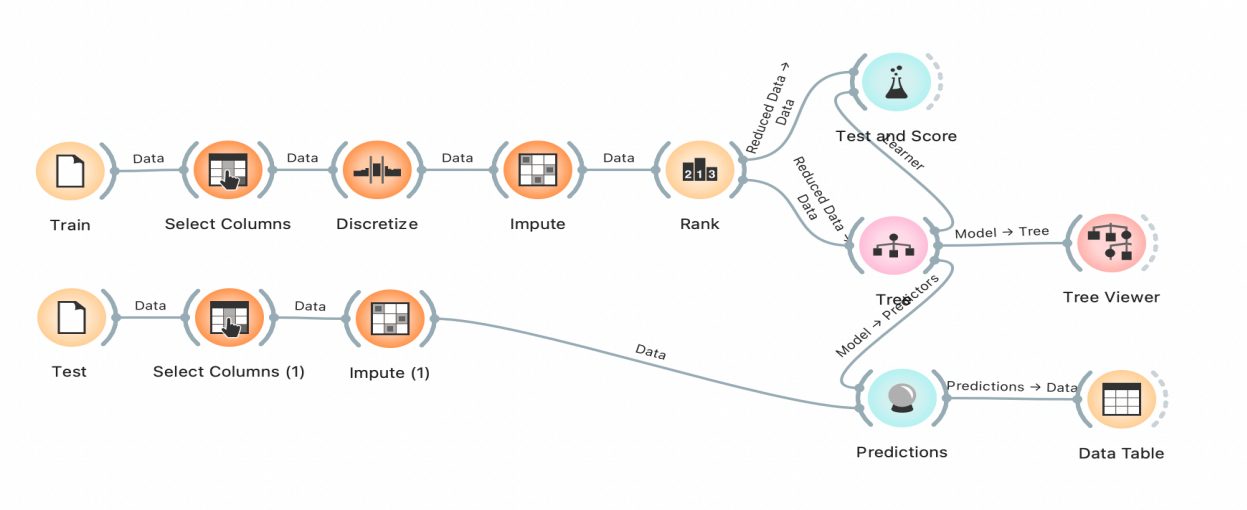
Figure 6: Saving the train and test data set

Section 2: Build and tune decision tree models

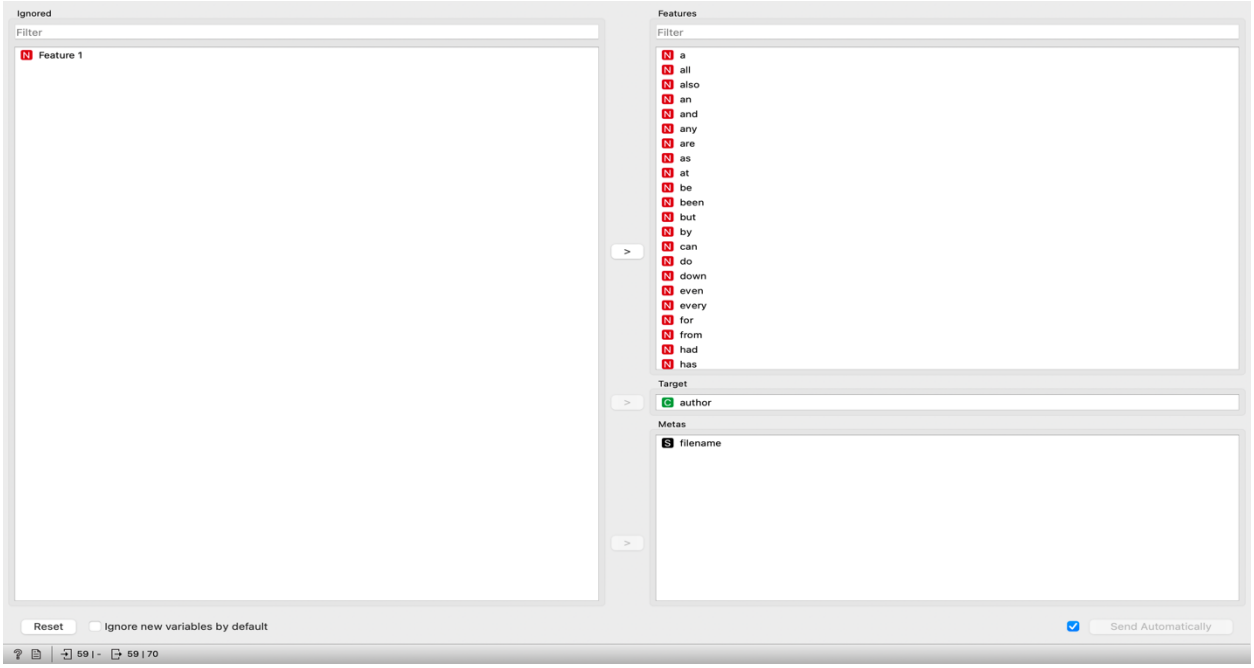
First build a DT model using default setting, and then tune the parameters to see if better model can be generated. Compare these models using appropriate evaluation measures. Describe and compare the patterns learned in these models.

After creating the test and train data set, we use Orange to build the prediction model in the following steps

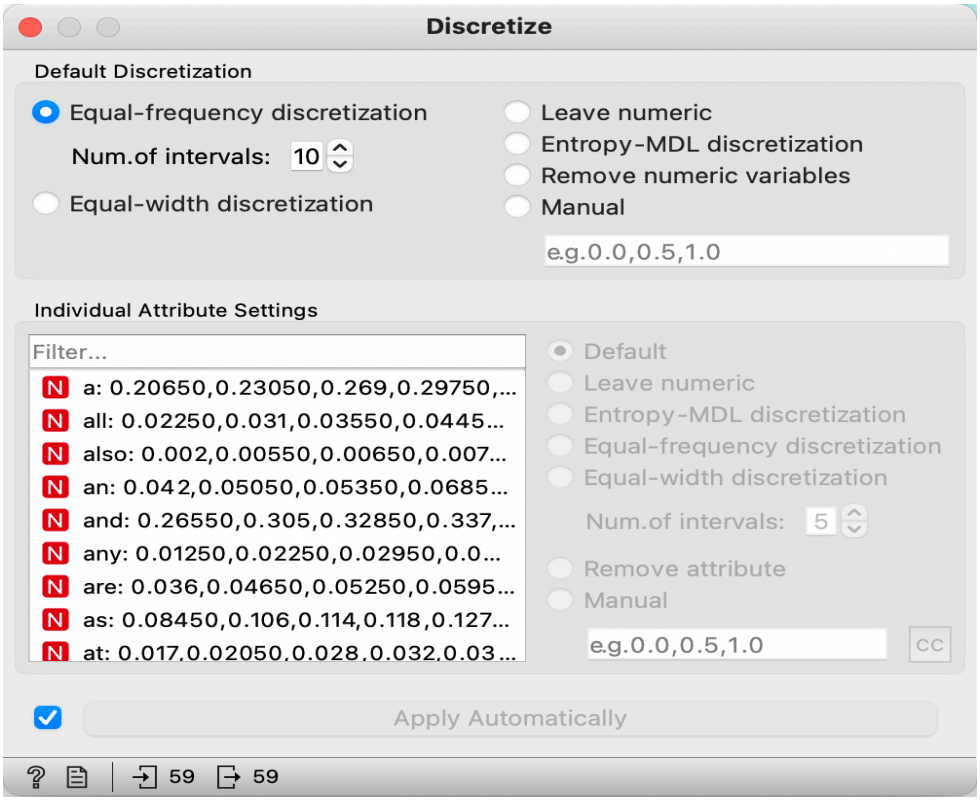
1- Data sets are imported into Orange.



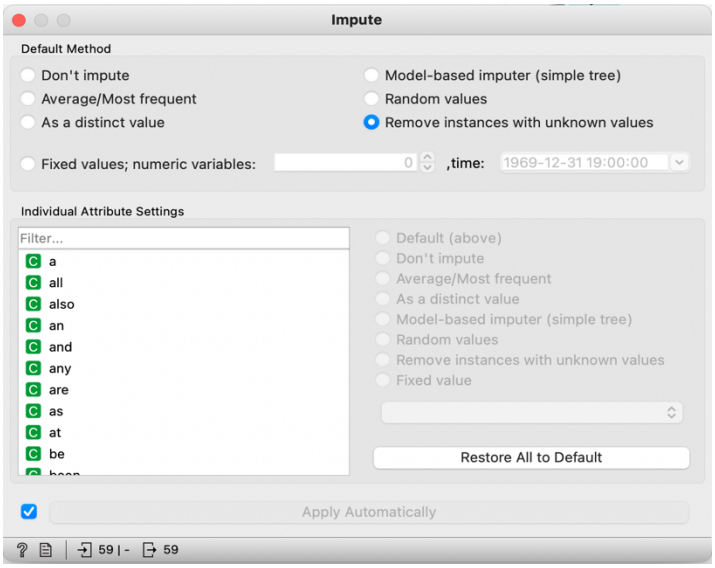
2- The columns are selected. “Feature 1” is the index and will get ignored in the model.



3- The data is discretized in 10 equal frequencies. Generally, we received a better F1 score using equal-frequency discretization instead of equal-interval discretization. Furthermore, we reached the highest F1 score when using 10 intervals.

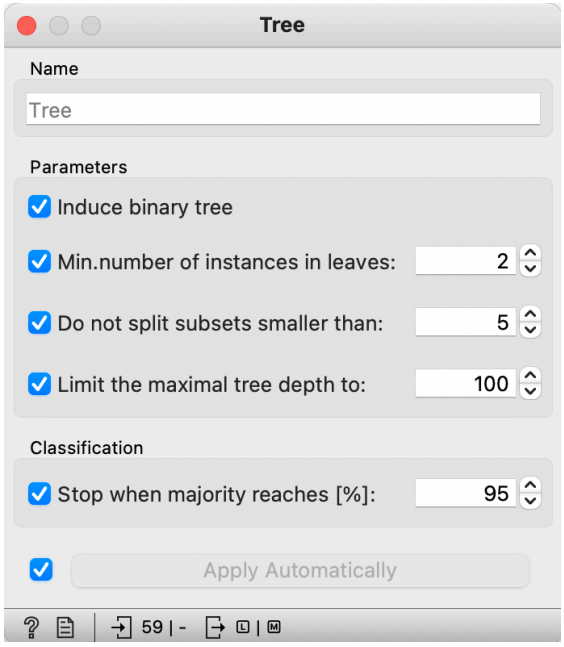


4- We use the impute module to remove the rows with the missing values. For this data set, imputation was not necessary as there was no instance with a missing value.



5- Using the rank module, we pick the top 7 attributes with the highest “Information Gain Ratio” for our model. After trying all the numbers from 1 to 30 and checking the F1 score, the number 7 was chosen as it led to the highest F1 score and lowest misclassified instances.

After implementing various combinations, we figured out that the highest F1 score, and lowest misclassified instances are achieved when we use the following tree setting:



The F1 score of 0.983 and 1 misclassified instance was the best possible outcome

Figure 13: The test and score window indication AUC, CA, F1 score, Precision, and Recall

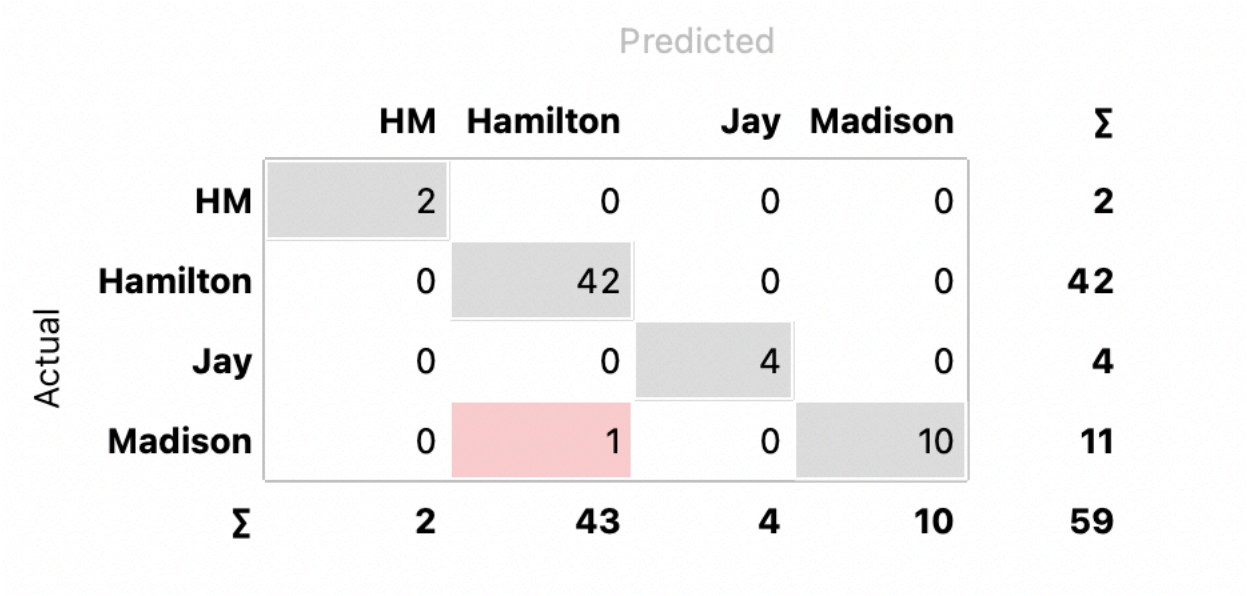


Figure 14: Confusion matrix indicating only 1 misclassification

Section 3: Prediction

After building the classification model, apply it to the disputed papers to find out the authorship. Does the DT model reach the same conclusion as the clustering algorithms did?

After tuning the model to reach the best evaluation outcome, it is time to check the prediction and compare it with the last analysis.

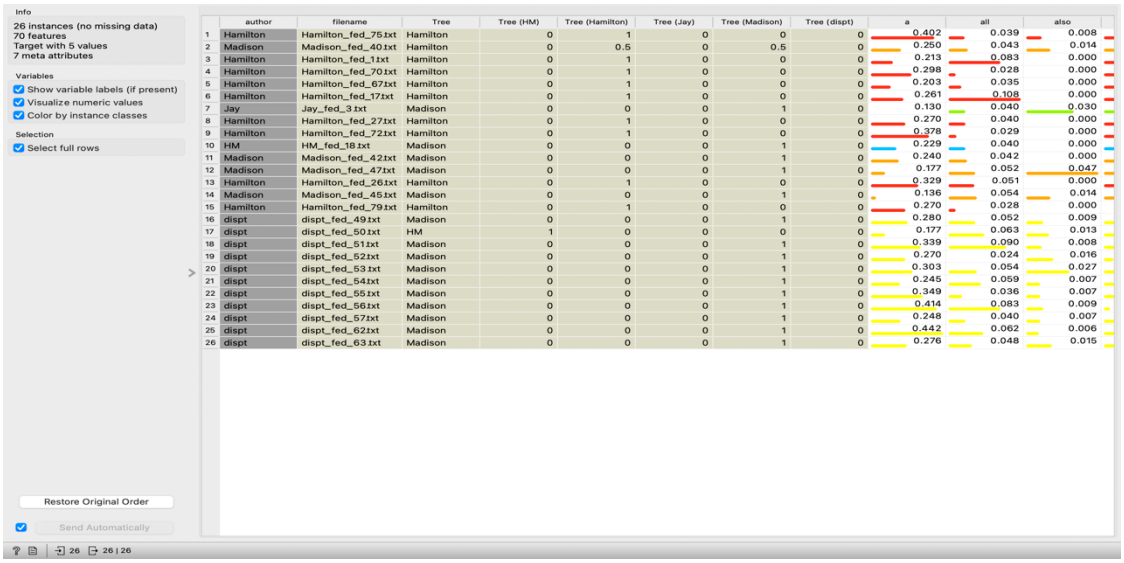
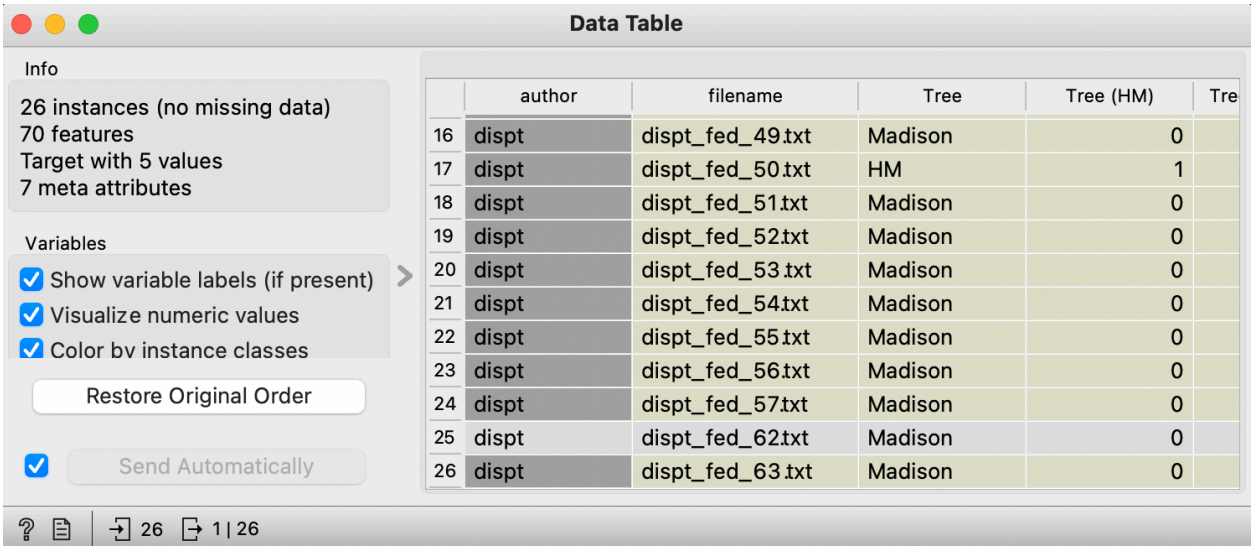


Figure 14: The data table with the prediction (view 1)



	author	filename	Tree	Tree (HM)	Tre
16	dispt	dispt_fed_49.txt	Madison	0	
17	dispt	dispt_fed_50.txt	HM	1	
18	dispt	dispt_fed_51.txt	Madison	0	
19	dispt	dispt_fed_52.txt	Madison	0	
20	dispt	dispt_fed_53.txt	Madison	0	
21	dispt	dispt_fed_54.txt	Madison	0	
22	dispt	dispt_fed_55.txt	Madison	0	
23	dispt	dispt_fed_56.txt	Madison	0	
24	dispt	dispt_fed_57.txt	Madison	0	
25	dispt	dispt_fed_62.txt	Madison	0	
26	dispt	dispt_fed_63.txt	Madison	0	

Figure 15: The data table with the prediction (view 2)

Conclusion: As can be seen in figure 15, our model believes: except for paper number 50 which is believed that has been written by both Hamilton and Madison, all the disputed papers were written by Maddison. However, in the last analysis, the model believed that most of the papers

were written by Maddison except number 56. Therefore, other than paper number 50 and paper number 56 both analyses are in consensus that the papers are written by Maddison. It may be possible that papers 50 and 56 have been written by both authors.