# **HW 05: Use Decision Tree to Solve a Mystery in History**

**Section 1: Data preparation**

**About the dataset**

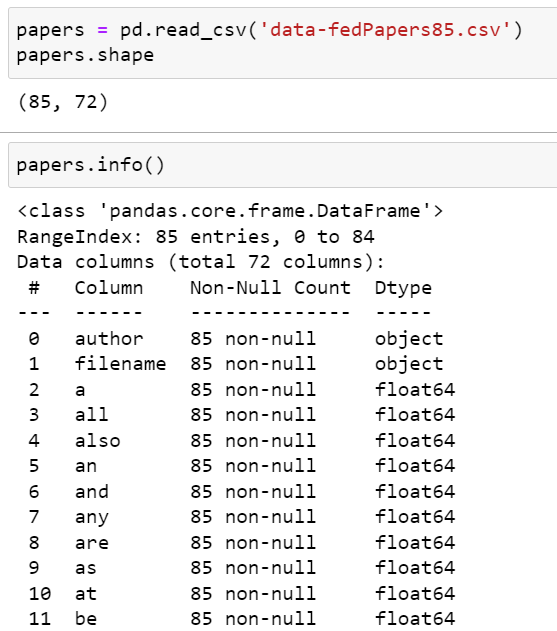
The Federalist Papers were a series of eighty-five essays urging the citizens of New York to ratify the new United States Constitution. Written by Alexander Hamilton, James Madison, and John Jay. The essays originally appeared anonymously in New York newspapers in 1787 and 1788 under the pen name “Publius”. The Federalist Papers are considered one of the most important sources for interpreting and understanding the original intent of the Constitution.

In this dataset, all 85 papers are titled based on their author, Hamilton, Madison, or Jay. Some essays have dual authorship. Additionally, there are 11 essays with historically disputed authorship.

Additionally, a Term Document Matrix is included which has normalized functional word counts per essay.

**Classification Problem**

While the authorship of 73 of The Federalist essays is fairly certain, the identities of those who wrote the twelve remaining essays are disputed by some scholars. We are trying to solve this mystery using the decision tree algorithm.

**Data Pre-Processing**

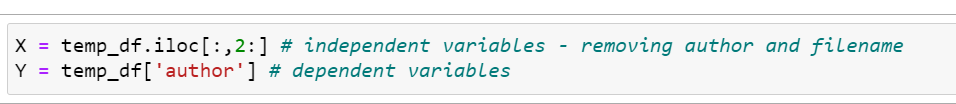
Data Preparation first required loading the .csv file as “papers”. Next, the file was generally assessed using papers.info().

Next, we start preparing our data frames for the experiment. First, we established and created a disputed authors data frame that will be the prediction target for the decision tree model. We then created a not disputed authors data frame by excluding all rows where the author is considered disputed. This not disputed data frame was then used to create our training and test datasets.

**Text

Description automatically generated**

We then removed the first two columns (author and filename) in order to leave the data frame as just function words and feature values.

****

These data sets were split using the train\_test\_split() function where 40% of the data was split into the test file and 60% was split into the training file.

****

Application

Description automatically generated with medium confidence

So, 44 essays are used for training the decision tree classifier and 30 essays for testing the predictions.

Text

Description automatically generated with medium confidence

Ensuring training and testing essays have the essays of all the authors – Hamilton, Madison, Jay, and Hamilton & Madison.

**Section 2: Build and tune decision tree models**

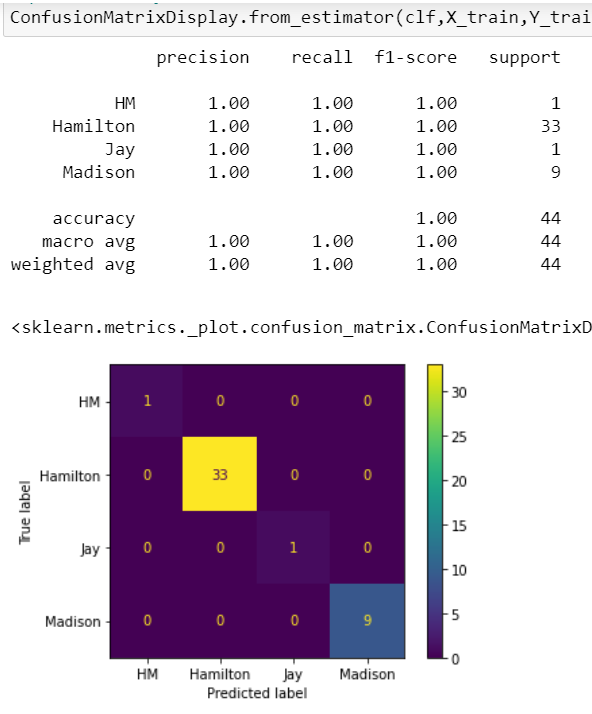
We started by creating a general decision tree model using API DecisionTreeClassifier(). We have used ‘entropy’ to measure the purity of the split and to obtain the information gained in every split. We had used ‘Gini impurity’ as well however, there was no significant improvement in the accuracy.

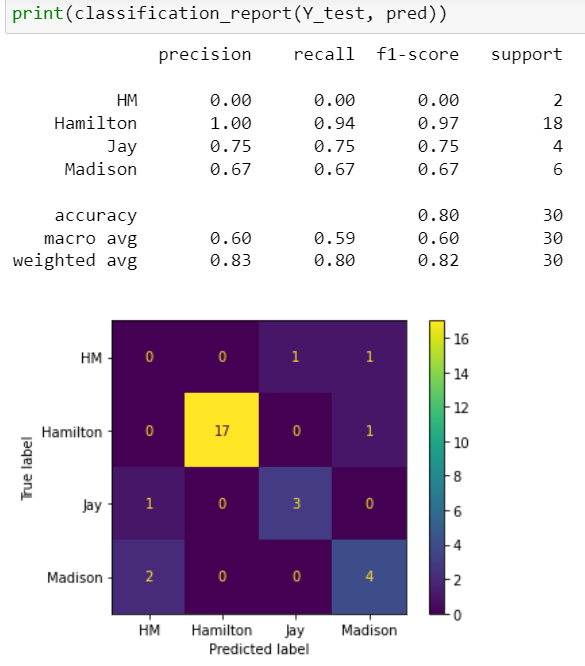
**Text

Description automatically generated**

**Model Evaluation**

From the below classification report on training and testing datasets, We can notice that the accuracy of the model on training data is 100% and on testing data is 80%. We can suspect that our model is overfitting. Also, we have only 74 essays (excluding disputed) to train and test our model. This can be one of the reasons for overfitting. However, We can perform a few post pruning techniques to resolve the problem of overfitting.





**Performing Post Pruning to avoid overfitting**

Pruning technique is parameterized by the cost complexity parameter, ’ccp\_alpha’. Greater values of ccp\_alpha increase the number of nodes pruned. It is necessary to choose the right ‘ccp\_alpha’ to cut down the branches of the decision tree.

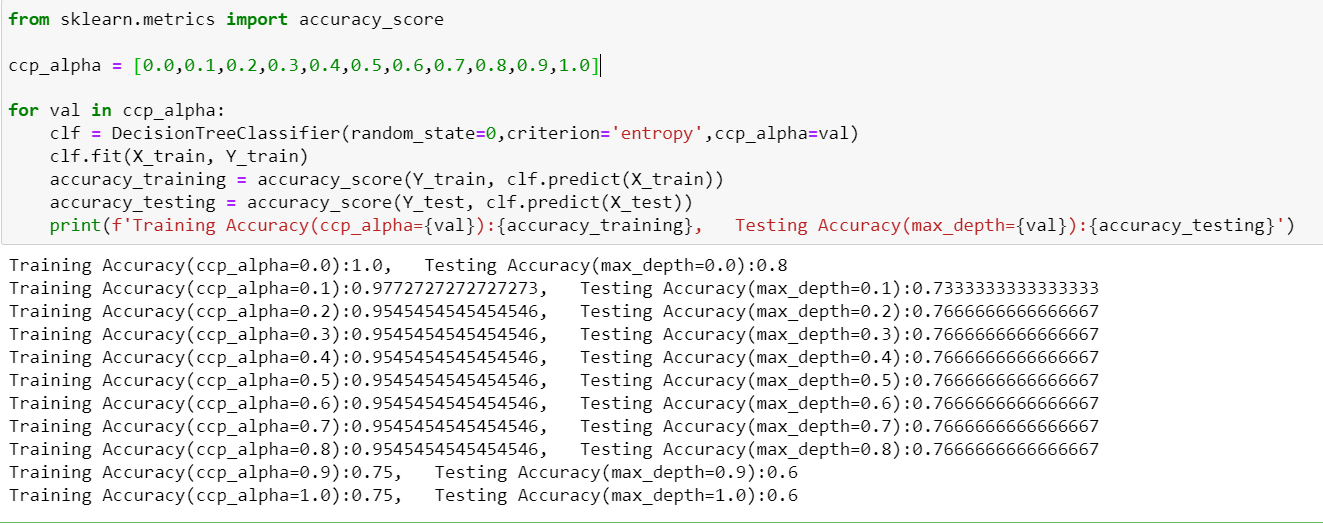
Based on different ‘ccp\_alpha’ values found from the training data, accuracy was plotted for training and testing data sets.

Graphical user interface, text, application

Description automatically generated

-> code showing different ccp\_alpha values

For different ccp\_alpha values, we evaluated the performance of the model on training and testing data. ccp\_alpha value between 0.2 - 0.8 resulted in better testing accuracy and training accuracy.



**A picture containing text

Description automatically generated**

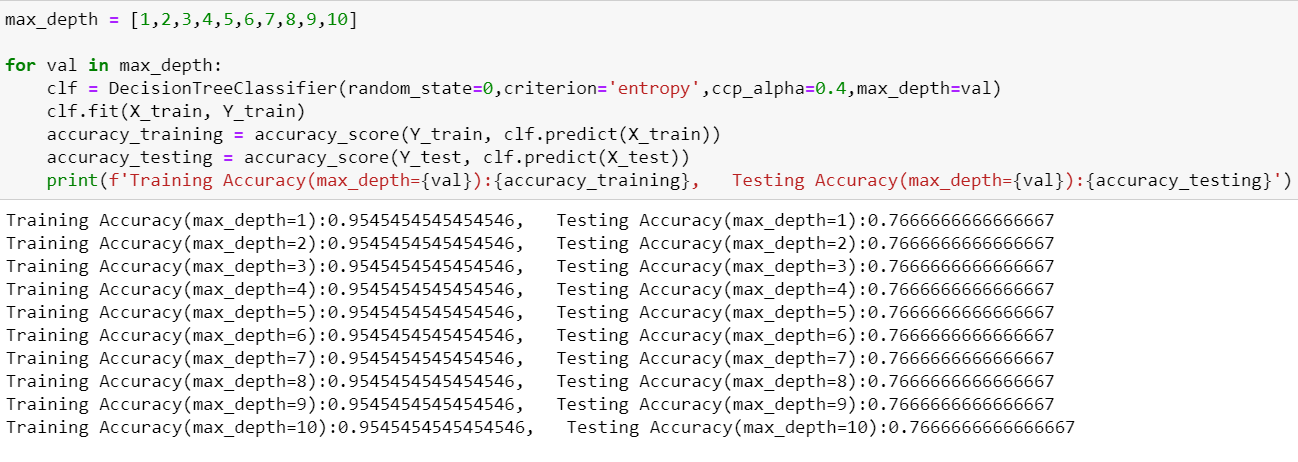
From the above plot, cost complexity value between 0.2 to 0.8 seems to be stable with accuracy on training and testing data sets. We decided to proceed with a minimum cost complexity parameter of 0.4 for the model.

**Hyperparameter Tuning**

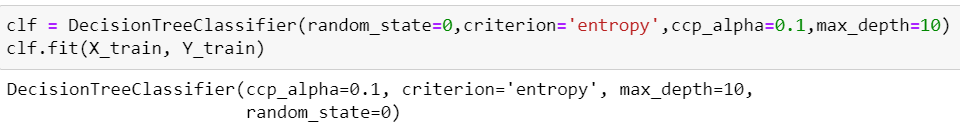
Here, we are going to be tuning based on ‘max\_depth’. We will try with max depth starting from 1 to 10 and depending on the final ‘accuracy’ score we are going to choose the value of max\_depth.

But, surprisingly there was no increase or decrease in the accuracy of the model for any max\_depth values. The accuracy remained constant for both training and testing data. The accuracy of training data is 0.95 and testing data is 0.77. So, any value between 1-10 can be used for building the classifier.

The below snapshot shows the code of hyperparameter tuning.

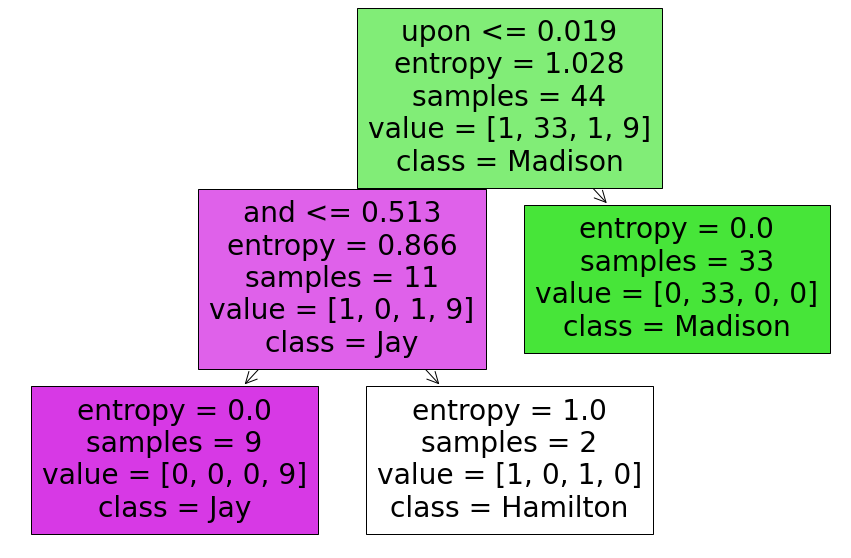


For, our final model (after applying the pruning technique and hyperparameter tuning), we decide to go with minimum cost complexity of 0.4 and the maximum depth of the tree as 10. We trained the model again on the same training dataset.



We then used tree.plot\_tree() to visualize the tree and tree.export\_text() to obtain all the rules in the string format. We then inspected our first general tree.

* Feature ‘**upon**’ is having highest information gain of 0.02 and is considered to be the first feature in classifying the authors.

Textual Representaion of the Decision Tree

|--- upon <= 0.02

| |--- and <= 0.51

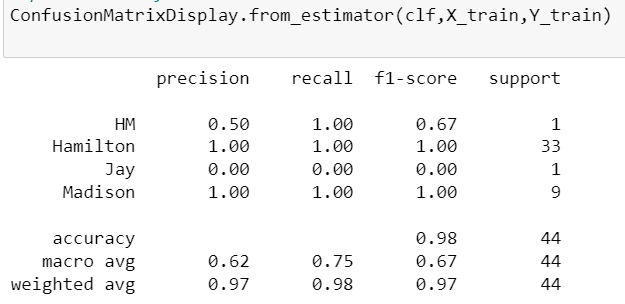
| | |--- class: Madison

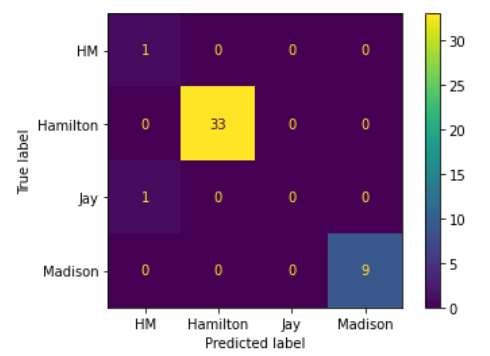
| |--- and > 0.51

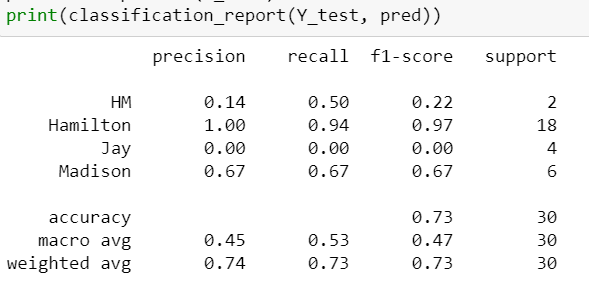
| | |--- class: HM

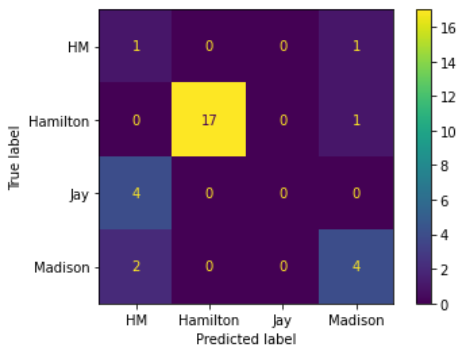
|--- upon > 0.02

| |--- class: Hamilton

**Training data accuracy**



**Testing data accuracy**

****

**Section 3: Prediction**

Federalist papers remain very important documents related to the history of the US constitution. Understanding who is the author of 11 disputed papers is still a mystery. While all the models that we have trained show a stronger likelihood that the author was Madison. However, the results are still inconclusive. Due to the small dataset, models can be overfitting resulting in inaccurate predictions. There are many essays still being classified as HM indicating Hamilton and Madison are helping each other in writing the federal papers.

According to the prediction by the decision tree classifier, the disputed papers are likely written by Madison.

**Table

Description automatically generated**