Name: Prachi Sadarangani

Date: 10/05/2023

**Who wrote the disputed essays, Hamilton or Madison?**

The aim of this report is to find the authorship of the disputed essays from the Federalist Papers dataset. The dataset contains 51 essays attributed to Hamilton, 15 to Madison, three to both Hamilton and Madison, 5 to John Jay, and 11 essays that are disputed. The goal is to use the Decision Tree Classifier to determine who wrote the essays, Hamilton or Madison.

Data Preparation:

Before starting the data modeling, we need to prepare the data. The data needs to be altered to get the right data frame for training, testing, and prediction. Firstly, the dataset contains the disputed essays. Since we need to predict who wrote those essays, we will filter this out. We don’t want this in our training or testing dataset. A new data frame will contain these essays. That will be the dataset that needs prediction. Now, the remaining essays will be in the dataset that will be used for training and testing. Both the datasets will be divided into two parts X and Y. X is the data frame with only the features (without the target variable which is ‘authors’) whereas Y is a data frame with only the target variable (‘authors’). Now, we will further divide and split the data into training and testing datasets. 20% of the dataset will be our testing data and 80% of the dataset will be the training data. This will be split randomly to avoid bias in the training data, Now, we have training data, testing data, and the disputed essay dataset that needs prediction.

To train the data, we will be using the Decision Tree Classification. This model is a simple model without any hyperparameters being defined as shown below.

A close-up of words

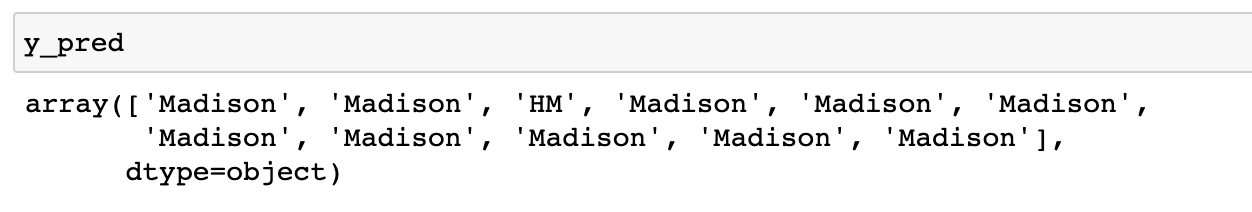
Description automatically generated

The below image shows how the model partitions the data based on features and thresholds and identifies patterns or rules for making predictions or classifications. It shows the nodes and the values are the counts of each of the author.

A diagram of a number of samples

Description automatically generated with medium confidence

The model is then tested on the test dataset. The accuracy of this model is found to be 93.33%, the precision score is 93.93% and the F1 score is 93.02%. All the model evaluation scores tend to be high which means this is a good model. This model will now be used on the disputed essay dataset to predict the authors of the disputed essay. The below image shows the model prediction on the disputed dataset.



As we can see from the image above, 10 out of 11 of the disputed essays are predicted to be written by Madison. Whereas only 1 is predicted to be written by Hamilton and Madison. Based on this we can conclude that the disputed essays are written by Madison. This aligns with the findings in our previous report, where we applied k-means and hierarchical agglomerative clustering (HAC).