Homework 5

Shivangi Mundhra | SUID – 842548148 | 10/12/2022

Introduction

In this homework, I am going to try to determine the author of the disputed Federalist essays using decision tree algorithm. I am going to walk through all the aspects of this assignment including what the Federalist papers are, what the assignment is and what approach I am using. In the analysis itself, I will try to explain all reasoning and thought process behind each significant step. The intent is that any person, especially one without a technical knowledge and that is not familiar with this problem statement, should be able to read this report and follow along.

About the Federalist Papers –

The Federalist Papers were a series of eighty-five essays written by Alexander Hamilton, James Madison, and John Jay urging the citizens of New York to ratify the new United States Constitution. These papers are considered one of the most important sources for interpreting and understanding the original intent of the Constitution.

About the disputed authorship –

The data for the assignment comes from the file named HW4-data-fedPapers85.csv uploaded on Blackboard. In the author column in this file, we find 74 essays with identified authors: 51 essays written by Hamilton, 15 by Madison, 3 by Hamilton and Madison, 5 by Jay. The remaining 11 essays are the famous essays with disputed authorship. Hamilton wrote to claim the authorship before he was killed in a duel. Later, Madison also claimed authorship.

About the assignment –

The task at hand, in this assignment, is to determine who the actual author was for these 11 disputed papers – Hamilton or Madison, using decision tree algorithm. In the last assignment, homework 4, I ventured to answer the same question but using a different technique – the clustering methods.

About the Decision Tree algorithm –

Decision tree is a machine learning model used for classification purposes. The idea is to create a model that predicts the value of a target variable (in this case, author) by learning simple decision rules inferred from the data features. In this assignment, we want to create a model that predicts the author of papers by analyzing the accompanying data for those papers. The data consists of filename, author, and features - function words like ‘upon’, ‘by’, ‘which’, etc., and their accompanying feature values which is the percentage of the word occurrence in an essay.

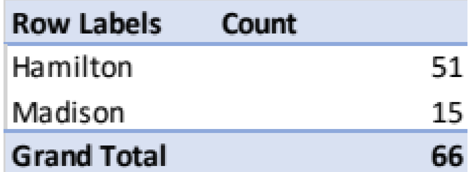
Analysis

I used scikit-learn machine learning libraries to create a decision tree model in python. For ease of understanding, I have divided my analysis in 3 sections –

1. Data preparation,
2. Build and tune decision tree model, and,
3. Prediction

Section 1 – Data preparation

1. I start with importing different packages like sklearn, pandas, numpy, etc. Some of these, like sklearn, help us with creation and implementation of the decision tree model, some, like matplotlib and graphviz help in creating visualizations and others just help in data manipulation.
2. I store the contents of the data from HW4-data-fedPapers85.csv into a dataframe. After making sure that the data does not include any missing values in any columns, I create a subset of the data that would include records belonging only to Hamilton and Madison. This is because, we want to train our model on these two author's data because the disputed papers belonged to one of these two authors. Other data is not required. I also drop the column called 'filename' as this is not one of the features that we want to train our model on. The resulting subset data has the below number of records by author.

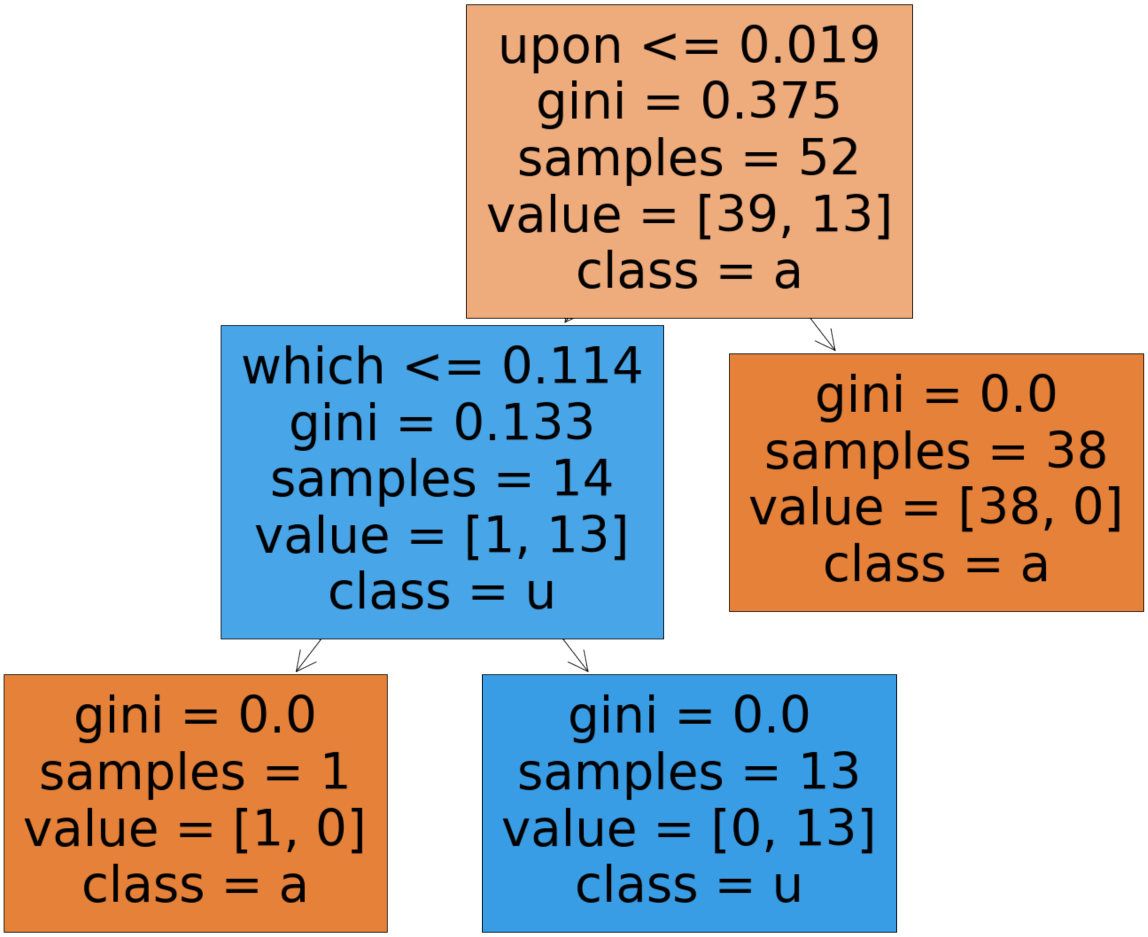


Row Labels	Count
Hamilton	51
Madison	15
Grand Total	**66**

3. After creating the subset data, I split the data in training dataset and testing dataset. X represents the features (columns) that we want to train our data on, and y represents the target value that we want to predict, in this case – author. We split the data in training and testing so that we can experiment with the model(s) and land on the best model. I chose to split the data so that 80% of it gets stored as training and 20% as testing. We can be sure that the train and test data set will also show the skewness towards Hamilton data because the number of records is skewed toward Hamilton in the main data set.
    a. The training data set consists of below number of records by author
    ```
    author
    Hamilton    39
    Madison     13
    ```
    b. The test data set consists of below number of records by author
    ```
    author
    Hamilton    12
    Madison      2
    ```

Section 2 – Build and tune decision tree models

1. First, we create a decision tree model with default settings using the training data. We use this model to make predictions on the test data set and calculate its accuracy and AUC.
    a. Accuracy is (number of correctly predicted samples/total number of samples).
    b. AUC – Area under the curve is a metric that considers the relationship between false-positives and true-positives
2. After testing the default decision tree model against the test data set, we receive an accuracy score of 92.86% and an AUC of 0.75. From these metrics, we can deduce that there is room for improvement in this model.
3. We plot the decision tree that's being used in this model and obtain below.



    a. As we can see, the decision tree model works on the 52 samples in training data set. This model uses the word "upon" in the 1st node.
    b. If the feature value for the word "upon" is greater than 0.019, then it will separate 38 files from the total 52 sample. This step classifies 38 of the 39 samples that belong to Hamilton in training data set.
    c. Next, it looks at the word "which". If feature value of "which" is greater than 0.114, then it classifies the document as belonging to Madison, else it classifies the document as belonging to Hamilton. This step classifies the remaining 14 samples, 13 of which are Madison's and 1 Hamilton's.

d. Gini impurity is a function that determines how well a decision tree was split. Basically, it helps us to determine which splitter is best so that we can build a pure decision tree. Gini impurity ranges values from 0 to 0.5, 0.5 being the worst.

e. According to the above model, the gini impurity at node “upon” = 0.375 and the same at “which” = 0.133. This suggests that we could improve the accuracy of model if we are able to remove these impurities.

4. To find a better decision tree model, we create different decision tree models and compare their accuracy. Ideally, we would want our model’s accuracy to be 100%.

   a. We compare accuracies for decision tree models with max_depth from 1 to 5 and min_samples_leaf from 1 to 5.

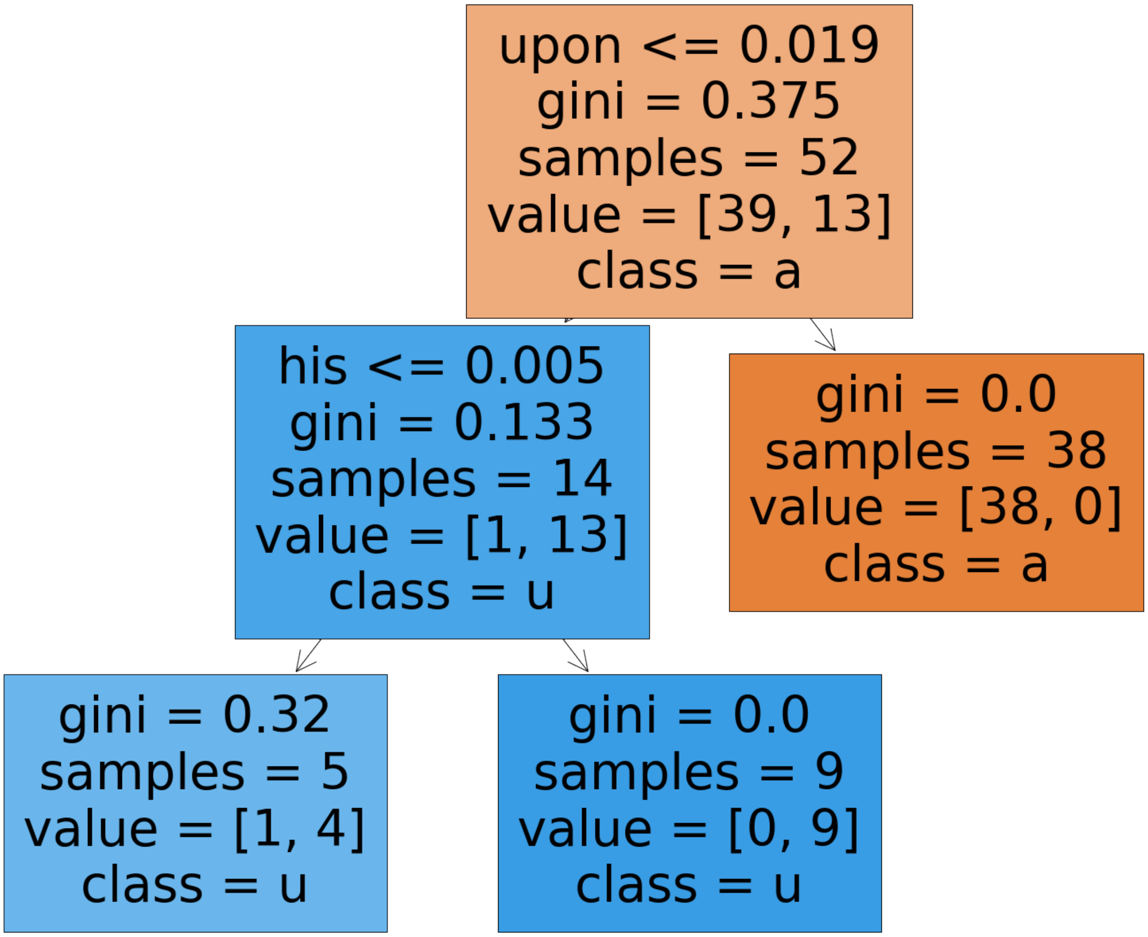
   b. Max_depth - The maximum depth of the tree. If None (default), then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

   c. Min_samples_leaf – default is 1. The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min_samples_leaf training samples in each of the left and right branches. This may have the effect of smoothing the model, especially in regression.

   d. We get 100% accuracy in for all models that have a max_depth >= 3 and min_samples_leaf >= 3.

   e. Hence, we can pick any of the models that give us a 100% accuracy and plot their decision tree and use it to predict our disputed papers.

5. To fine tune the default model, we set the max_depth (= 3), min_samples_leaf (= 5) variable in the next decision tree model.
6. We test the accuracy and AUC of this fine-tuned model on the test data again and find that the accuracy for this model is 100% and AUC is 1. This indicates that we landed on a very good model.
7. We plot its decision tree and obtain the below.



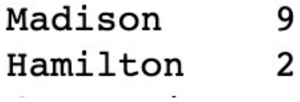
a. As we can see, the decision tree model works on the 52 samples in training data set. This model uses the word “upon” in the 1st node.
b. If the feature value for the word “upon” is greater than 0.019, then it will separate 38 files from the total 52 sample. This step classifies 38 of the 39 samples that belong to Hamilton in training data set.
c. Next, it looks at the word “his”. If feature value of “his” is greater than 0.005, then it classifies the document as belonging to Madison. The remaining 5, 1 by Hamilton and 1 by Madison are left to be further classified.
d. Gini impurity is a function that determines how well a decision tree was split. Basically, it helps us to determine which splitter is best so that we can build a pure decision tree. Gini impurity ranges values from 0 to 0.5, 0.5 being the worst.
e. According to the above model, the gini impurity at node “upon” = 0.375 and the same at “his” = 0.133.

8. Because the second model, the tuned model has an accuracy of 100% and AUC of 1, we will choose that for our final prediction.

Section 3 – Prediction

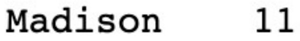
We apply both the models to the disputed papers.

1. The default decision tree model predicted that of the 11 disputed files, 9 belonged to Madison and 2 belonged to Hamilton as shown below.



```
Madison      9
Hamilton     2
```

2. The fine-tuned decision tree model predicted that all 11 disputed files belonged to Madison as below.



```
Madison     11
```

Since the accuracy of the second model is 100%, we can say with certainty that the 11 disputed papers were written by Madison. This is the same conclusion that we obtained using my clustering analysis in the last homework.

Conclusion

We successfully used a decision tree classification model to find the author of the disputed Federalist papers. The final model shows that the disputed papers were written by Madison, a conclusion that matches with what we found using our clustering analysis as well.