Name: Vaibhav Agrawal
Roll Number: 15CE10057
Assignment 2: Decision Tree (Part 2)
Machine Learning (CS60050)

Part 2

(Description) In this part, you will implement the decision tree algorithm to learn a classifier that can assign a topic (science, sports, atheism etc.) to any news article.

(Dataset) Train and test your algorithms with a subset of the 20 newsgroup dataset . More precisely, you will use the documents on alt.atheism and comp.graphics newsgroup. To simplify your implementation, these articles have been pre-processed and converted to the bag of words model. Each article is converted to a vector of binary values such that each entry indicates whether the document contains a specific word or not.
Download the training set (traindata.txt) and test set (testdata.txt) of articles with their correct newsgroup label (trainlabel.txt, testlabel.txt) here.
Each line of the files traindata.txt and testdata.txt are formatted "docId wordId" which indicates that word *wordId* is present in document *docId* . The files trainlabel.txt and testlabel.txt indicate the category (1=alt.atheism or 2=comp.graphics) for each document (docId = line number). The file *words.txt* i ndicates which word corresponds to each wordId (denoted by the line number).

(Tasks)
(A) Implement the decision tree learning algorithm. Here, each decision node corresponds to a word feature, which is selected by maximizing the information gain.
Design your algorithm to take as input a maximum depth. If a branch of the decision tree reaches this specified depth, it should not be grown further.
Experiment with your algorithm by building trees with increasing maximum depth until a full tree is obtained. Report the training and testing accuracy (i.e., percentage of correctly classified articles) of each tree by producing a graph with two curves - one curve for training accuracy and one curve for testing accuracy - as a function of the maximum depth of the tree.
Report also the tree that achieved the highest testing accuracy.
(B) Use the decision tree algorithm implemented in scikit learn , using Information Gain, to classify the same data. Report the accuracies on test data.

(Deliverables) Your report should contain :

1. A graph showing the training and testing accuracy as the maximum depth increases.

2. Does overfitting occur? If yes, after what maximum depth does overfitting occur?

3. A brief discussion of the word features selected by the decision tree that achieved the highest testing accuracy. In your opinion, did all the word features selected make sense?

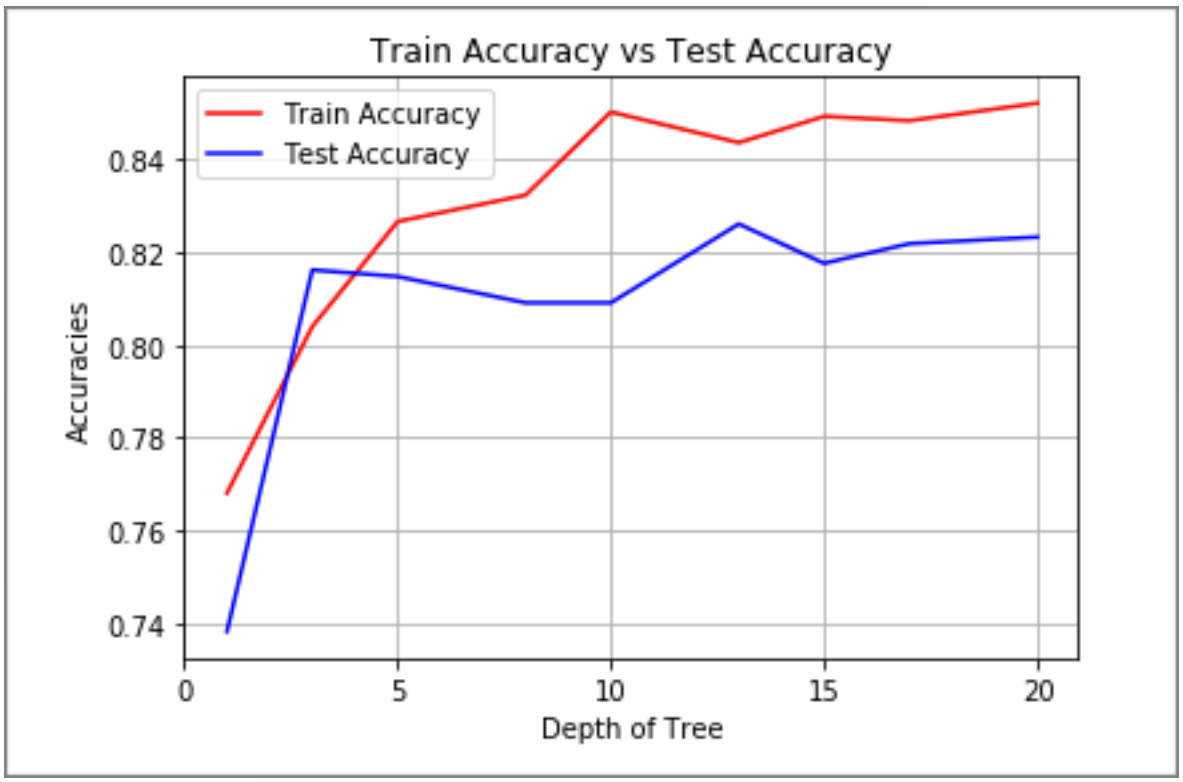
Accuracy of Scikit Learn model:

Following is the accuracy of the scikit model with (information gain) on the part 2 of the problem:
Training accuracy: 84.4467 %
Test accuracy: 82.4467 %

The tree that achieved maximum accuracy is the tree with depth = 13

1. **A graph showing the training and testing accuracy as the maximum depth increases.**

Ans. 1



2. **Does overfitting occur? If yes, after what maximum depth does overfitting occur?**

Ans. 2 Yes, overfitting occurs. As can be seen from the above plot that with the increasing depth of tree the training accuracy and testing accuracy both increases, But after the depth of 5 there is a decrease in the test accuracy whereas the training accuracy decreases for a short while and again starts to increase. We can conclude that the the overfitting is occurred after larger depths. ie. for depth > 5.

3. A brief discussion of the word features selected by the decision tree that achieved the highest testing accuracy. In your opinion, did all the word features selected make sense?

Ans. 3. Some of the word feature have more influence on the determination of the the newsgroup of the data. These features are often known as the critical or the key factors affecting the the newsgroup. Determination of the critical factors is one of the major task in the field of data mining and related fields. It can provide a much needed insight to the problem. In my opinion all the word selected did not made sense, as common English words like: as, the, and etc. won't have much influence on the determination of the newsgroup. Whereas, some specific words can alone tell a detailed story about the document type. For eg. a particular documents consisting of more words features like cricket, football, goal, sixes, teamwork etc. can tell us that the tendency of the newsgroup being from sports category is much higher.