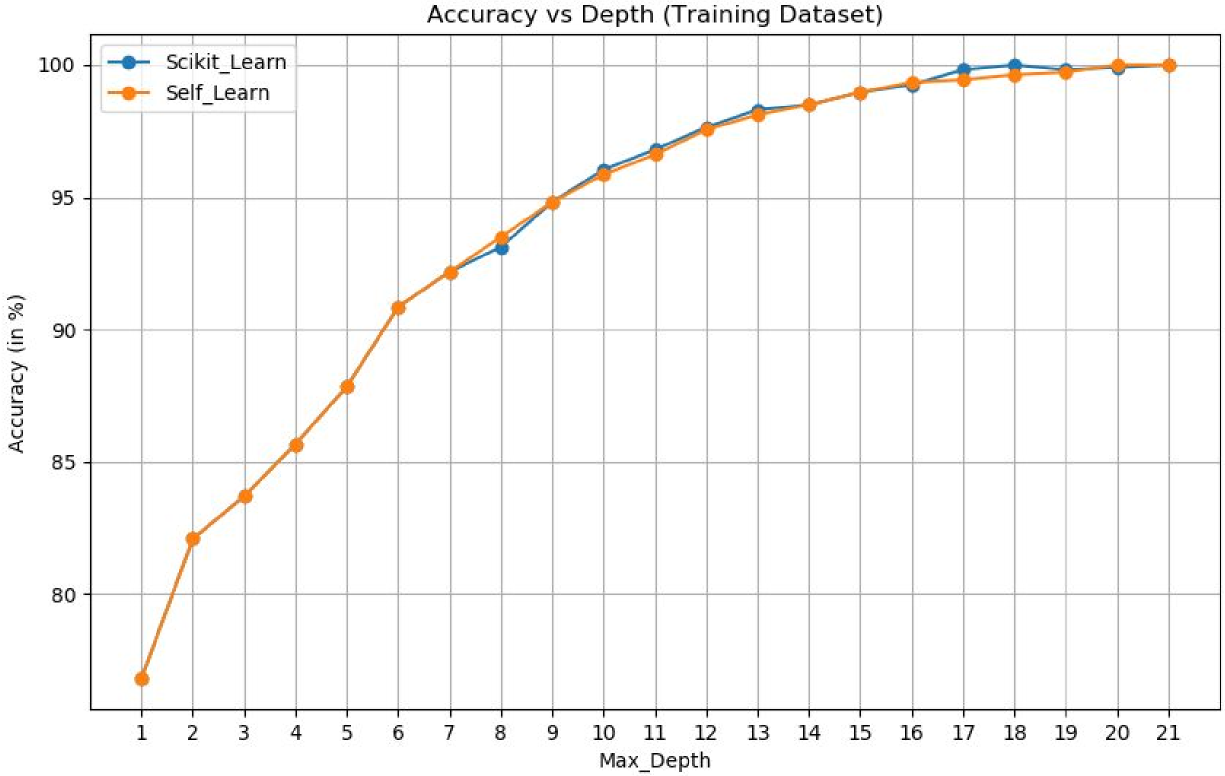
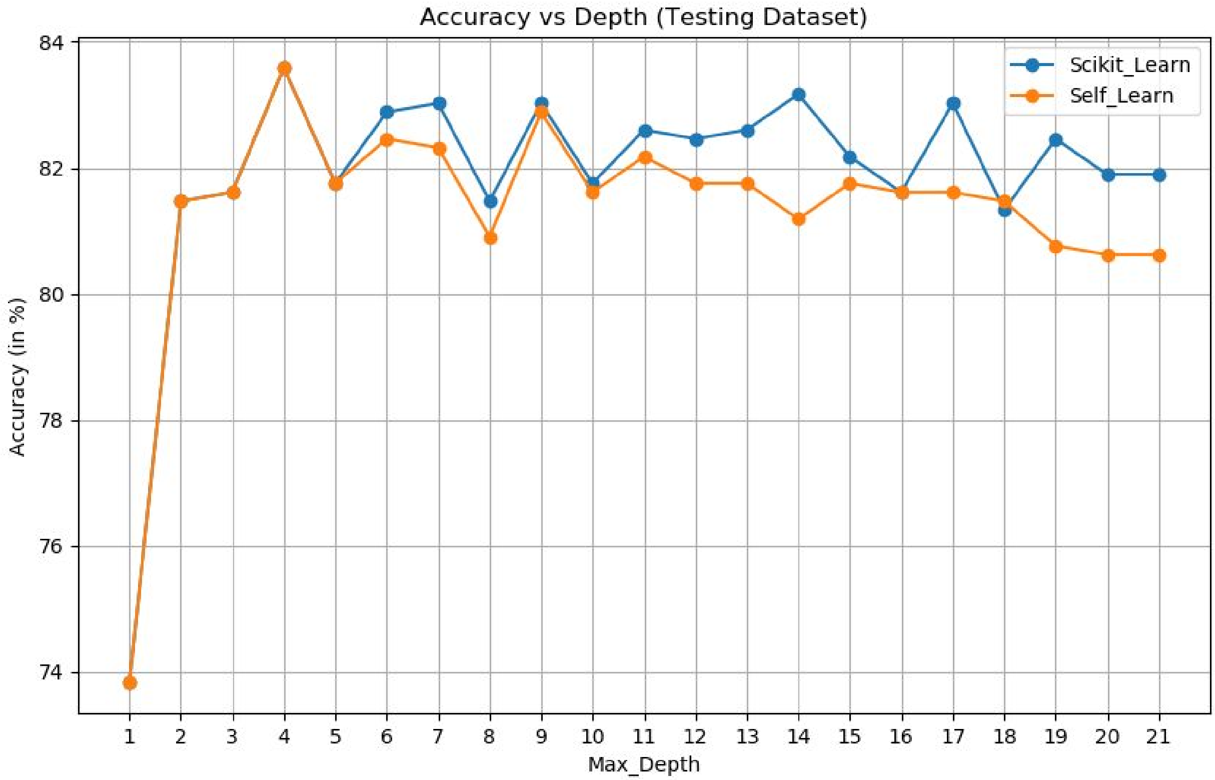
CS60050 - MACHINE LEARNING ASSIGNMENT 2
DECISION TREES - PART 2

NAME: **Himanshu Mundhra**
ROLL No. : **16CS10057**

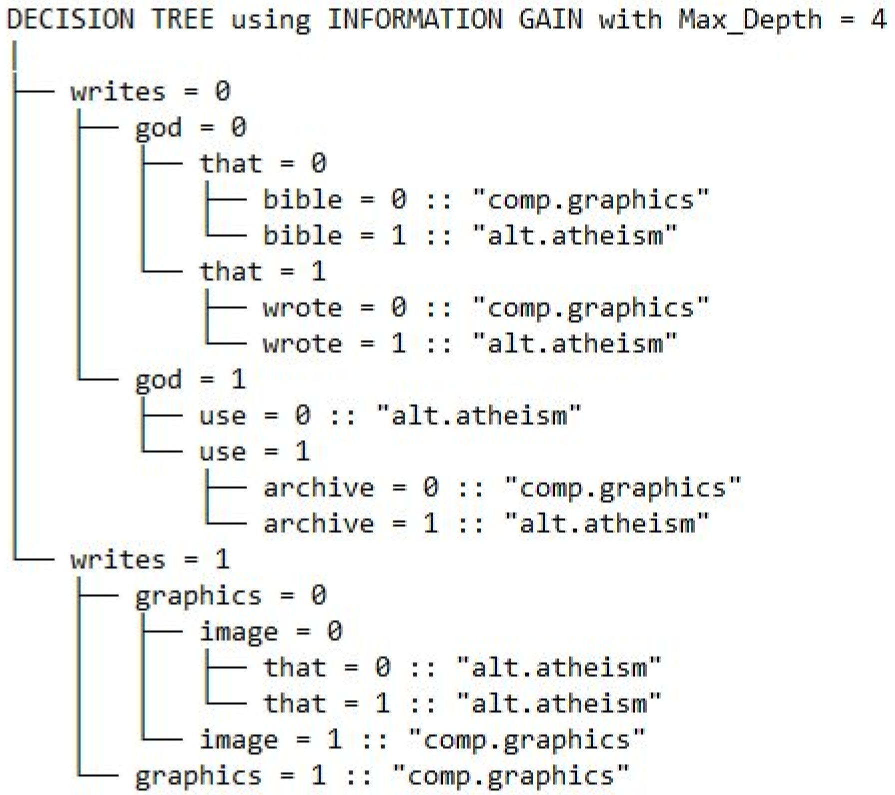
In this assignment, we build a Decision Tree on a DataSet of the Bag of Words Model.
We have 3566 words taken from articles related to **alt.atheism** and **comp.graphics** which are assigned numeric keys in 'words.txt'. These act as features in our dataset. We assign a numeric value of ***1 for alt.atheism*** and ***2 for comp.graphics*** in our labelling set to the different news articles.
We use the criterion of ***entropy*** to grow trees progressively from a Depth of 1 till when the Depth Saturates (i.e. Depth of two consecutive trees is the same) and then use the testing data to report the tree with the **Highest Testing Accuracy**.
We then **compare** our model results with **Scikit-Learn's models**.



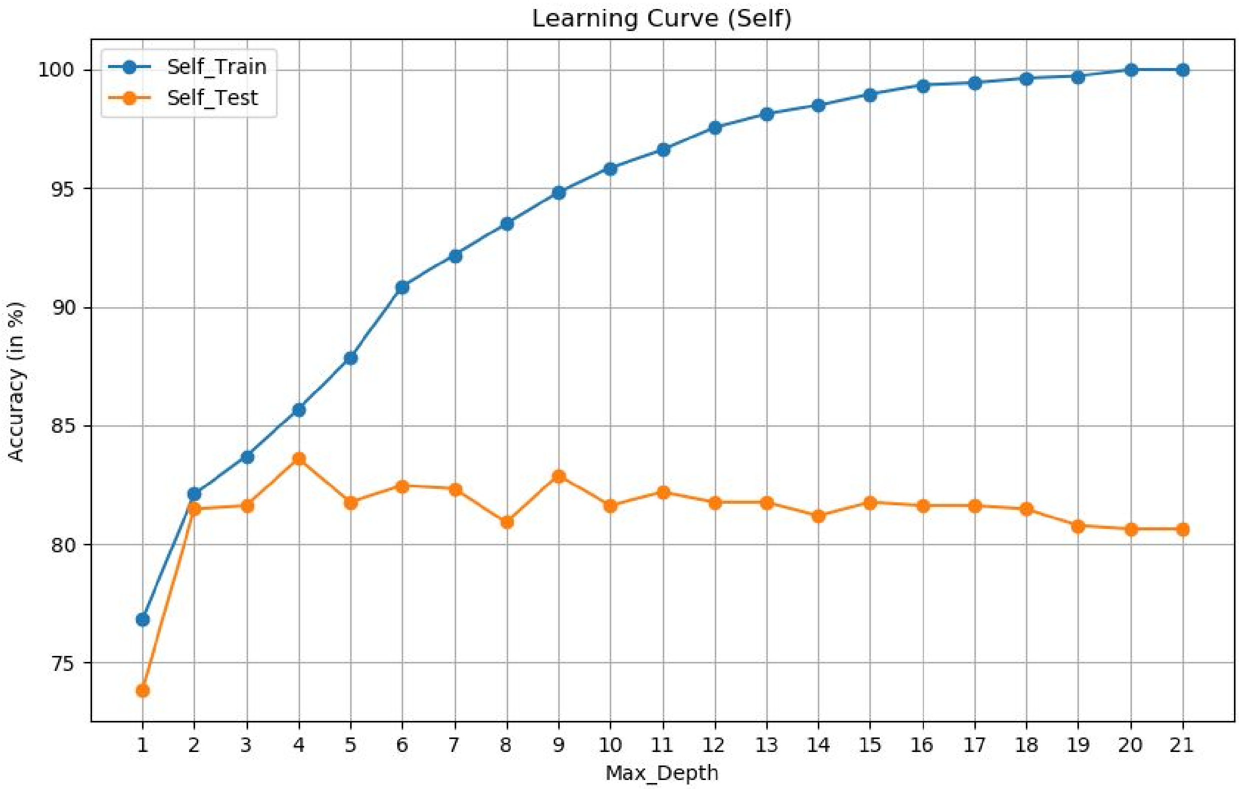
Plot of Training Accuracy of Trees Learnt with Splitting Criterion as Information Gain vs Max_Depth As we can see, the Accuracy reached around 100% for Max_Depth 20-21 and then saturates at that value. This means that perfect classification has taken place and the Tree can grow no further.



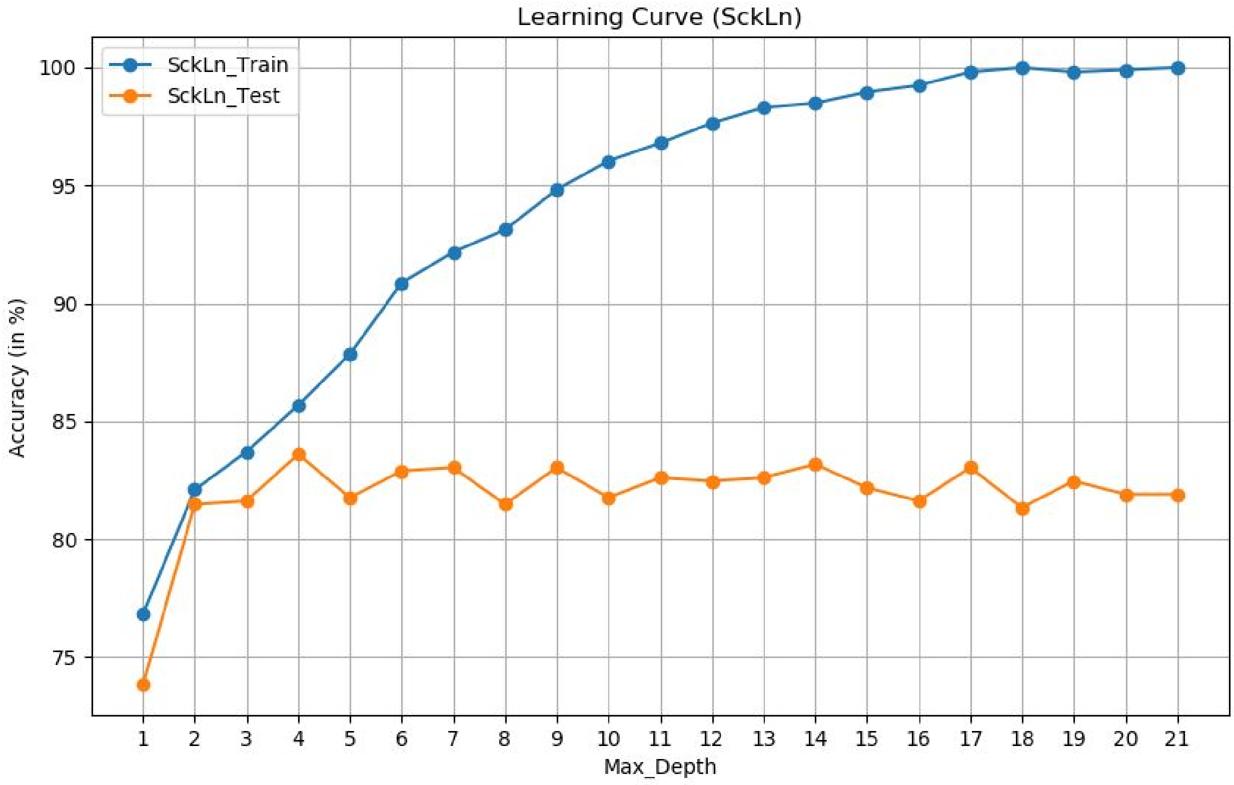
Max Test Accuracy Occurs at a Max_Depth of 4 where the Value of Accuracy = 83.59% (entropy)



```
DECISION TREE using INFORMATION GAIN with Max_Depth = 4
│
├── writes = 0
│   ├── god = 0
│   │   ├── that = 0
│   │   │   ├── bible = 0 :: "comp.graphics"
│   │   │   └── bible = 1 :: "alt.atheism"
│   │   └── that = 1
│   │       ├── wrote = 0 :: "comp.graphics"
│   │       └── wrote = 1 :: "alt.atheism"
│   └── god = 1
│       ├── use = 0 :: "alt.atheism"
│       └── use = 1
│           ├── archive = 0 :: "comp.graphics"
│           └── archive = 1 :: "alt.atheism"
└── writes = 1
    ├── graphics = 0
    │   ├── image = 0
    │   │   ├── that = 0 :: "alt.atheism"
    │   │   └── that = 1 :: "alt.atheism"
    │   └── image = 1 :: "comp.graphics"
    └── graphics = 1 :: "comp.graphics"
```



A Plot between Training Accuracy and Test Accuracy when Splitting Criteria is Information Gain
The Self and Scikit Test Accuracy Plots both have maxima at Max_Depth = 4

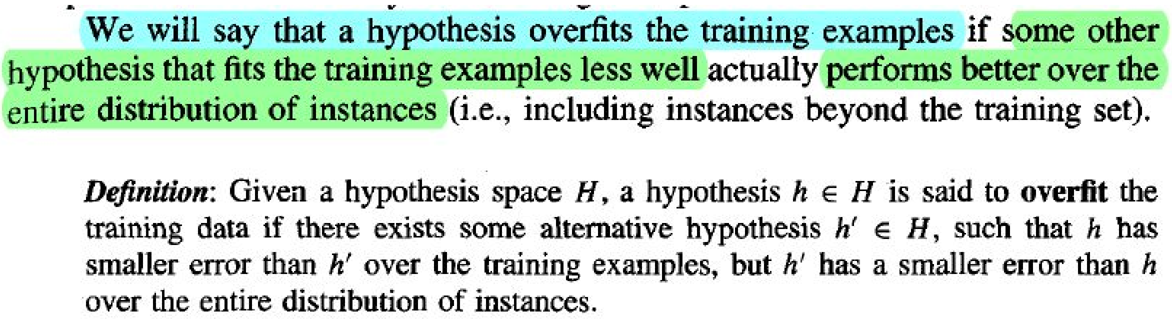


THE PROBLEM OF OVERFITTING

The decrease of testing accuracy with an increase in depth can be attributed to the fact that overfitting is occurring in the tree. As the model tries to fit the Training Data better, it starts incorporating the noise and possible outliers in the training data and hence leads to a poorer performance when tested on the “Test Data”.

From the Learning Curve of the Self-Learnt Tree, we can infer that after a Max_Depth of 4 the Accuracy goes down. However, at certain depths, there is an increase in the accuracy from the previous values but this value is still less than the Maximum Value at 4. After oscillating for certain points, there is a steady decrease in the Test Accuracy after a Max_Depth = 15.

If we consider this formal definition of Overfitting, we can say Overfitting occurs after a Max_Depth = 4, which is the tree which has the best accuracy.



We will say that a hypothesis overfits the training examples if some other hypothesis that fits the training examples less well actually performs better over the entire distribution of instances (i.e., including instances beyond the training set).

Definition: Given a hypothesis space H, a hypothesis $h \in H$ is said to **overfit** the training data if there exists some alternative hypothesis $h' \in H$, such that h has smaller error than h' over the training examples, but h' has a smaller error than h over the entire distribution of instances.

Machine Learning - Tom M. Mitchell : Chapter 3 - Decision Trees Page 67

A Discussion on Selected Word Features

Through our model, we aim to predict whether, given an article marked with the presence of the features in contention, it belongs to **“alt.atheism”** or **“comp.graphics”**. These two topics are pretty disjoint in terms of the matter of discussion pertaining to it, which is perhaps one reason why a smaller tree as this comes out to be the best tree. However certain features of the tree can be said to be a weird combination.

The branch with **[writes = 0 , god = 1, use = 1, archive = 0]** has an prediction of **“comp.graphics”**, which is absurd in the sense that the presence of **‘god’** as a feature should more or less establish that the topic is **“alt.atheism”** unless people talking about “**comp.graphics**” these days are really pious.

Moreover, really generic word features like **‘writes’**,**’that’**,**’use’** and **‘wrote’** are very common words which can appear in any of the two mentioned topics. Being used for the classification of news articles into **“alt.atheism”** or **“comp.graphics”** is absurd. This has happened perhaps because of the writing style of the authors whose articles we have used as our Training Dataset. It can be a mere coincidence that these features come out to be numerically superior to other more relevant ones such as **‘god’**, **‘graphics’** and ‘**bible’**, which is why there is quite some room for improvement even in the maximum test accuracy.

Other words like **‘atheism’**, **‘religion’**, **‘christian’**, **‘biblibal’**,**‘church’**, **’catholic’**, **‘gui’**, **‘cdrom’**, **‘xview’**, **‘coreldraw’**, **‘imagemagick’**, **‘xloadimage’** and many more words which make much more sense as classifiers are not even chosen. Its the bias of the dataset towards certain words and against such clear cut classifiers which is the cause of this absurdity in my honest opinion.