Question 2: Decision Tree Classifier

We are going to use Training_set.csv (cleaned version) as the file for the dataset. The dataset is turned to 2-class n-feature classification.

ID has been removed from the data set while using plotting decision tree classifier, as the column is not a good feature to predict alues on.

A total of 10 line graph are plotted, for each Max_Depth.

```
sahil@sahilvb:~/Assignment2/Q2-DECISIONTREE$ cat training_set.csv |python DTree.
py
Let's do decision tree!!
Total Rows in the data are- 4662
Total Rows are 4662
Training Set has 4195 rows, which is about 90.0 % of the total rows
Test Set has 467 rows, which is about 10.0 % of the total rows
Total Values of Very Late Adopters- 1127
Total Values of Other Adopters- 3535
Total Rows in the new created table are- 2255
IN WHICH
Very Late Adopters are 1127 which is about 49.0 % of the total rows
Others' are 1127 which is about 49 % of the total rows
Training Set has 2029 rows, which is about 90.0 % of the total rows
Test Set has 226 rows, which is about 10.0 % of the total rows
Hope you Enjoyed the graphs!!!!!
Goodbye World!!!
```

Leaf Node. In a tree data structure, the node which does not have a child is called as LEAF Node. In simple words, a leaf is a node with n child. In a tree data structure, the leaf nodes are also called as External Nodes. External node is also a node with no child. In a tree, leaf node is also called as **'Terminal' node**.

Max Depth. Controls the **maximum depth** of the **tree** that will be created. It can also be described as the length of the longest path from the **tree** root to a leaf. The root node is considered to have a **depth** of 0

Machine learning is a problem of trade-offs. The classic issue is *overfitting* versus *underfitting*

When we increase the number of Max_Leaf_Nodes, we see a particular trend on almost all the plots where the accuracy tends to decrease on the test set. One possible reason to explain such decrease in accuracy could be attributed to Overfitting of the data.

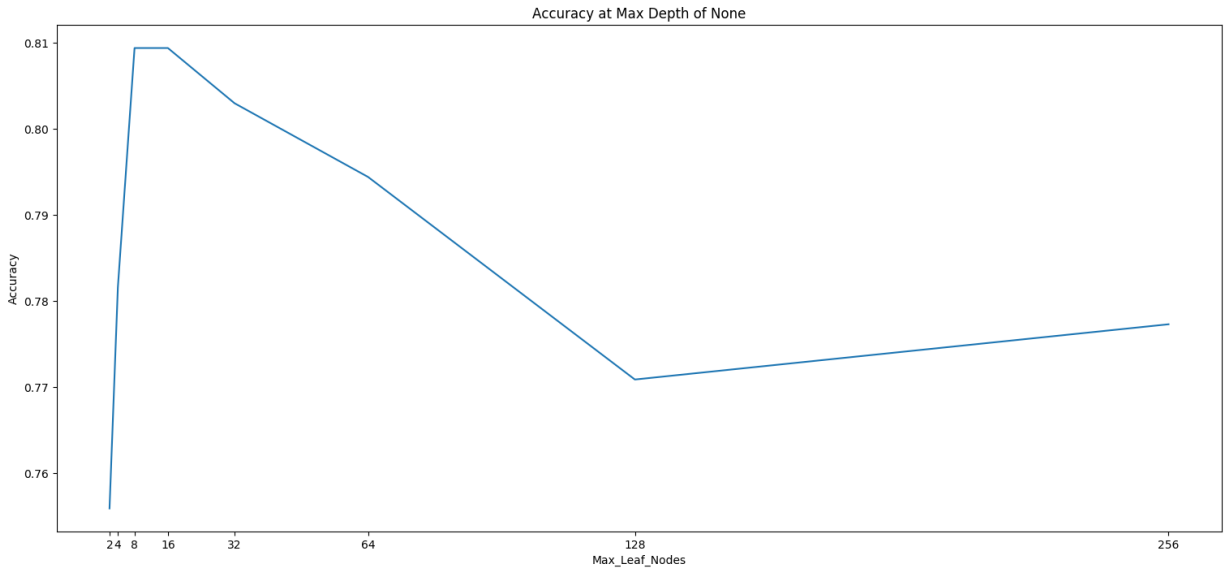
Over-fitting is the phenomenon in which the learning system tightly fits the given training data so much that it would be inaccurate in predicting the outcomes of the untrained data. In decision trees, over-fitting occurs when the tree is designed so as to perfectly fit all samples in the training data set. Thus it ends up with branches with strict rules of sparse data. Thus this effects the accuracy when predicting samples that are not part of the training set.

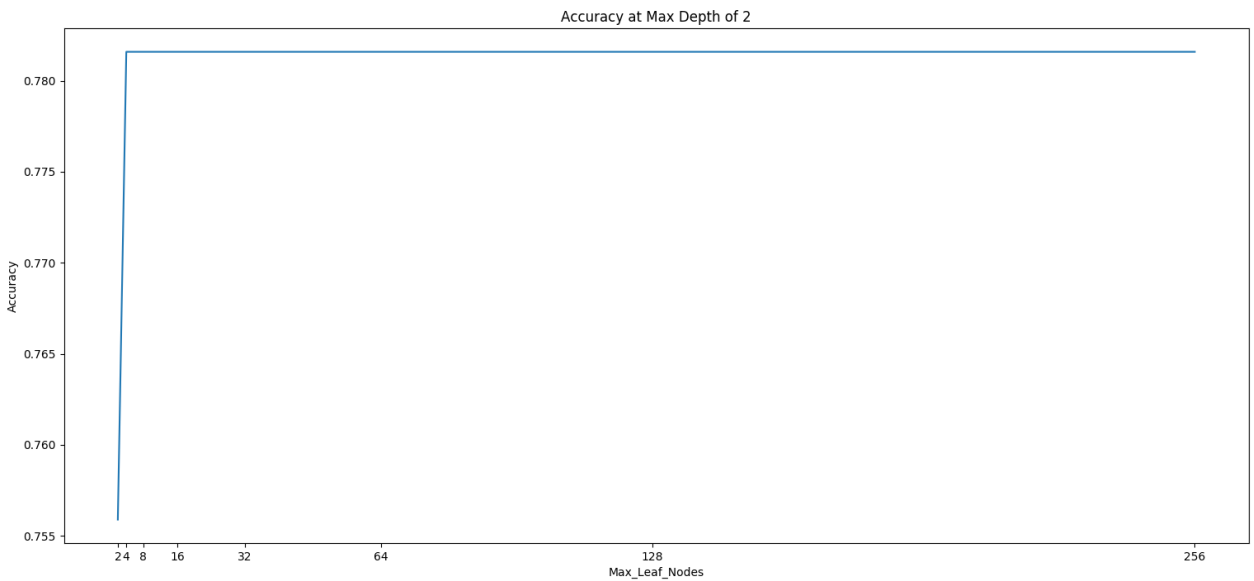
Hence increasing the Max_leaf_nodes leads to lower accuracy, hence model tends to work good for training set, but not so much for the Test set.

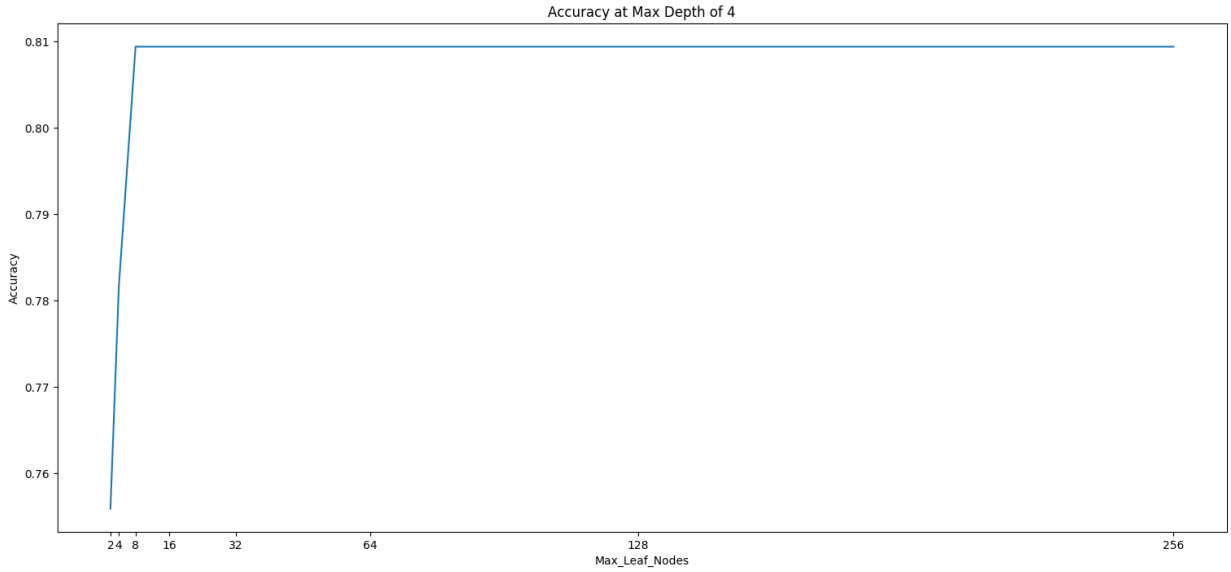
Thus, we can choose the best model by taking into account both the Max_Depth and Max_Leaf_nodes, so that the model does not overfit the data, and gives out maximum accuracy on the Test set.

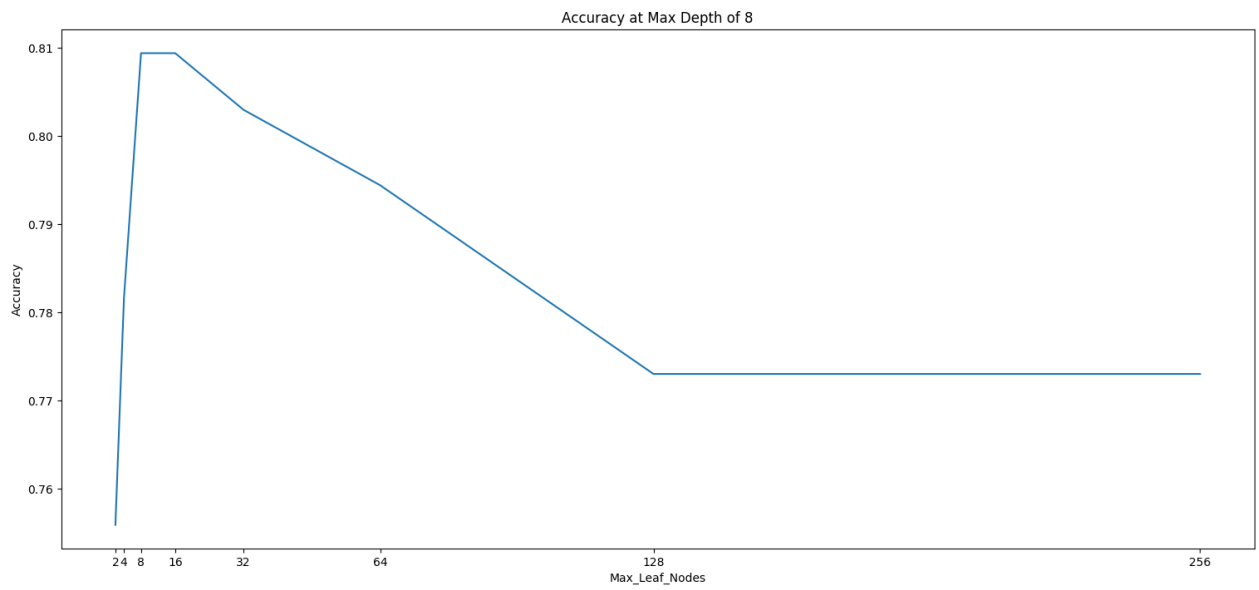
A. Dataset is divided into Train and Test in the ratio 90:10 respectively

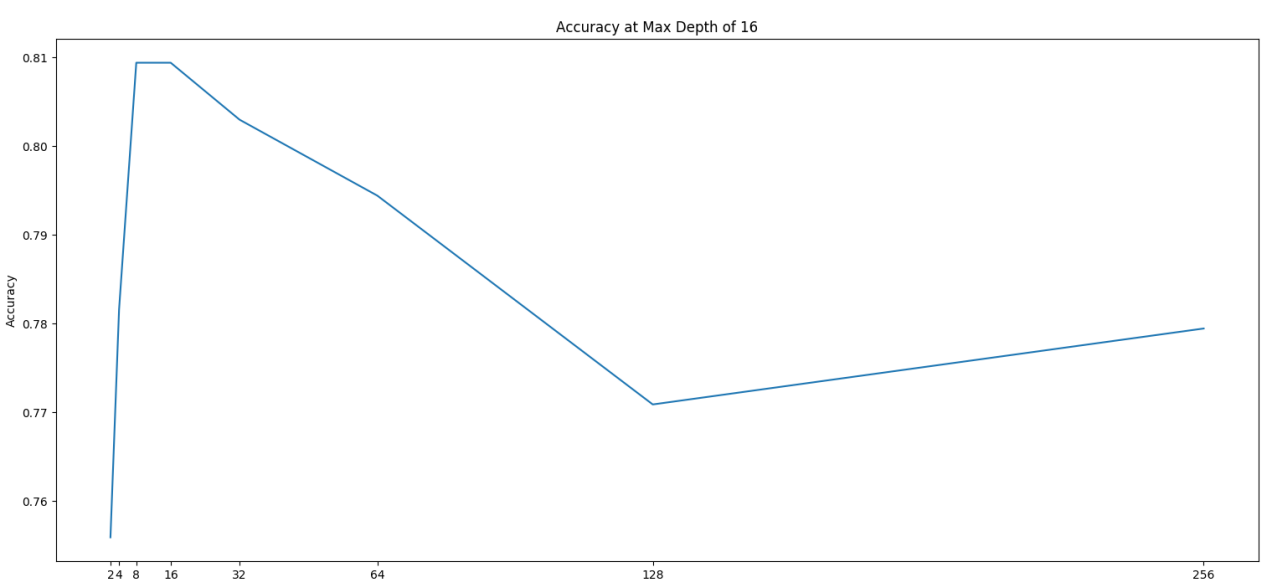
By looking at these 5 graphs, the sweet spot for the decision tree can be at Depth=4 and Max leaf nodes=8,16,32. We can chose 8 as it gives the same accuracy for lowest tree nodes, hence avoiding overfitting of the data.






Accuracy at Max Depth of 4
Accuracy
0.81
0.80
0.79
0.78
0.77
0.76
2 4 8 16 32 64 128 256
Max_Leaf_Nodes

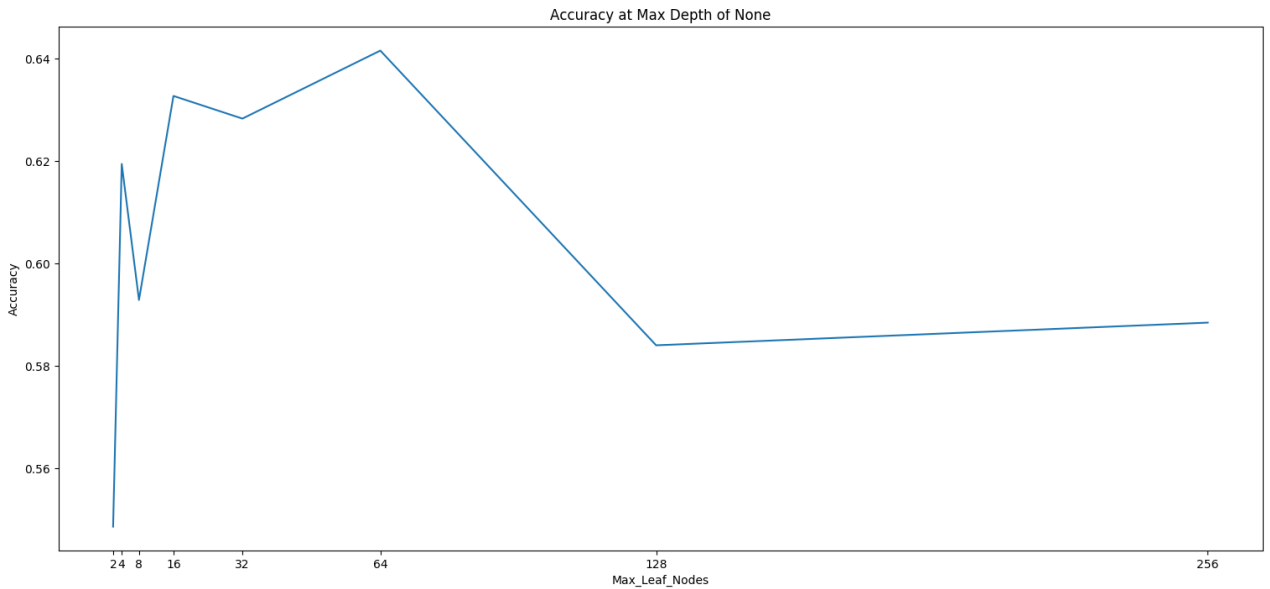

Accuracy at Max Depth of 8
Accuracy
0.81
0.80
0.79
0.78
0.77
0.76
2 4 8 16 32 64 128 256
Max_Leaf_Nodes

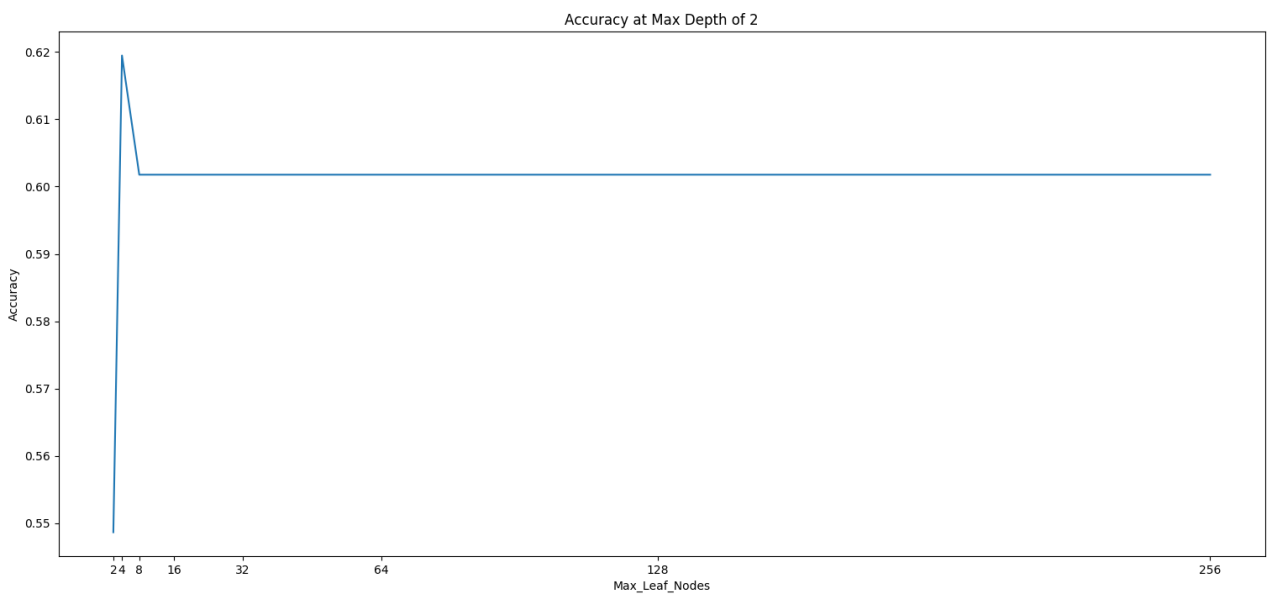

Accuracy at Max Depth of 16
Accuracy
0.81
0.80
0.79
0.78
0.77
0.76
2 4 8 16 32 64 128 256
Max_Leaf_Nodes

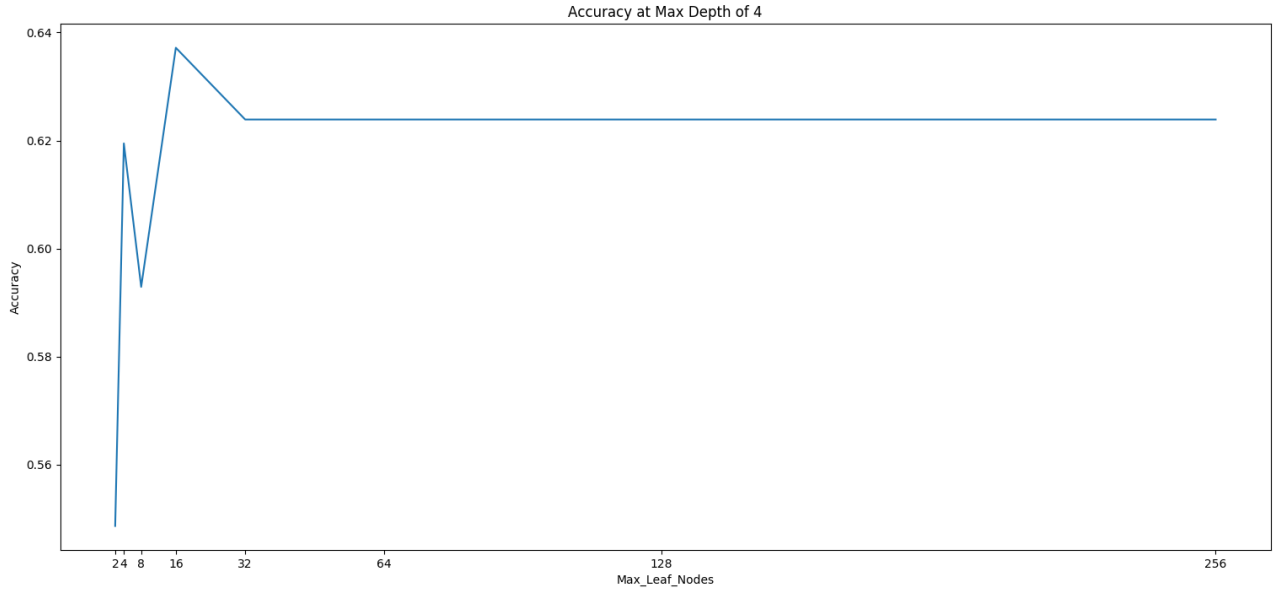
B. 50:50 Dataset

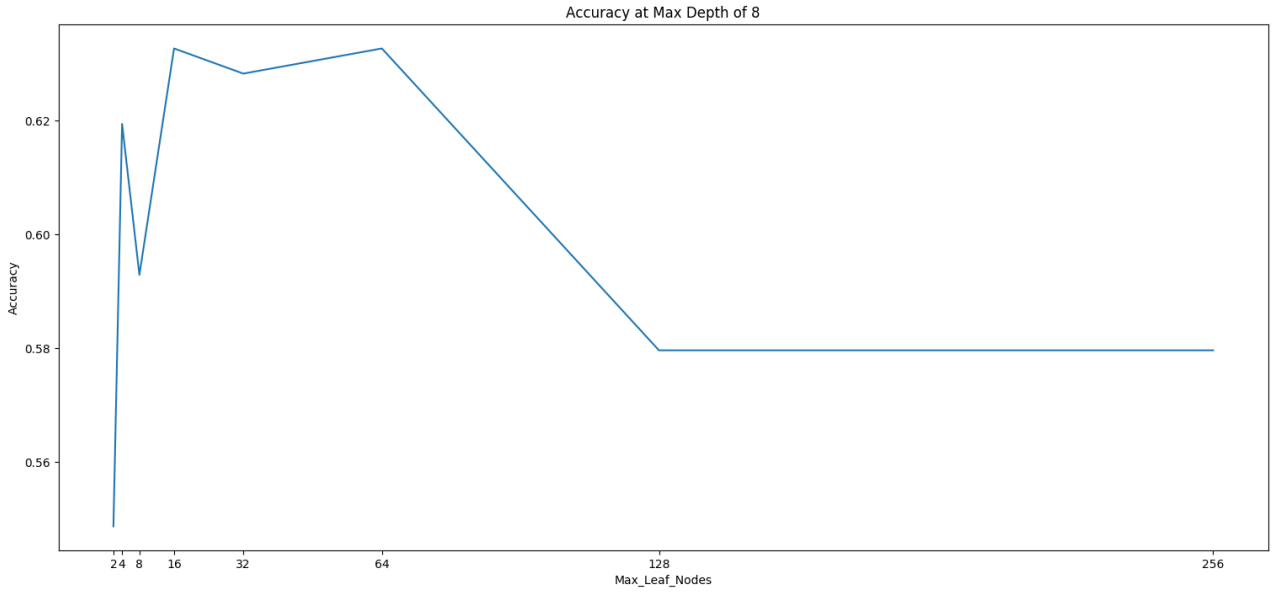
1) New Dataset is derived from the original dataset so that the number of Late Adopters are equal to the number of Other' type of Adopters (50:50 ratio)
2) New Dataset is divided into Train and Test in the ratio 90:10

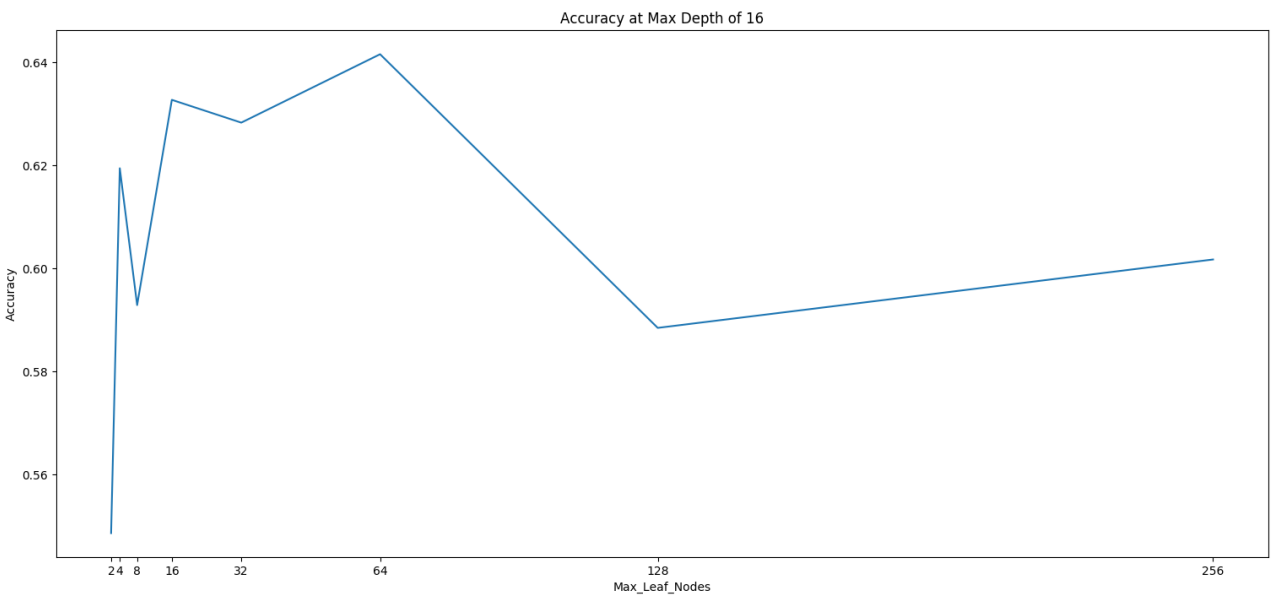
After pruned the data, the accuracy is less stable, and drops to the range of 56%-62.5%.
The accuracy drops significantly when the model is too complex, (depth=16 and none).
The sweet spot for this data can be Depth=8 and Max leaf nodes=16.






Accuracy at Max Depth of 4
0.64
0.62
0.60
0.58
0.56
Accuracy
2 4 8 16 32 64 128 256
Max_Leaf_Nodes


Accuracy at Max Depth of 8
0.62
0.60
0.58
0.56
Accuracy
2 4 8 16 32 64 128 256
Max_Leaf_Nodes


Accuracy at Max Depth of 16
0.64
0.62
0.60
0.58
0.56
Accuracy
2 4 8 16 32 64 128 256
Max_Leaf_Nodes