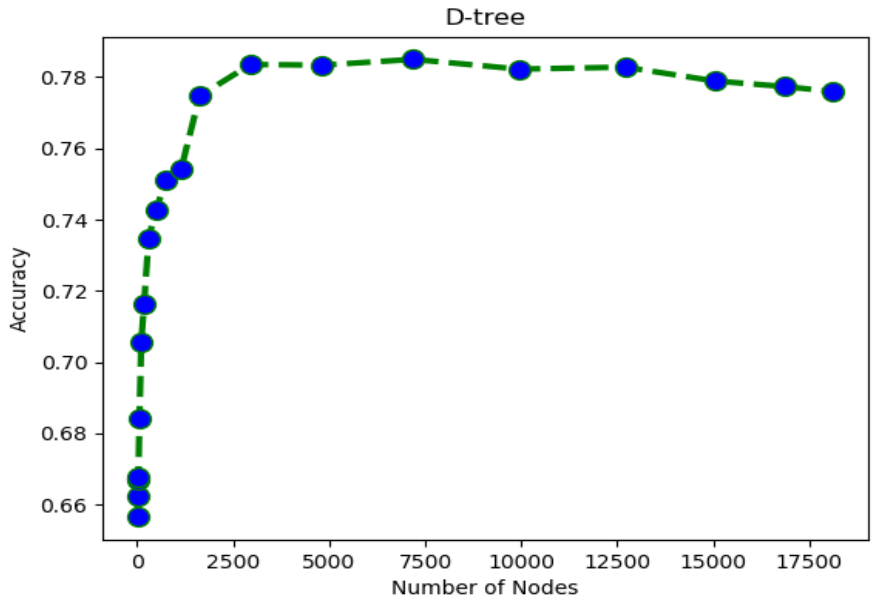
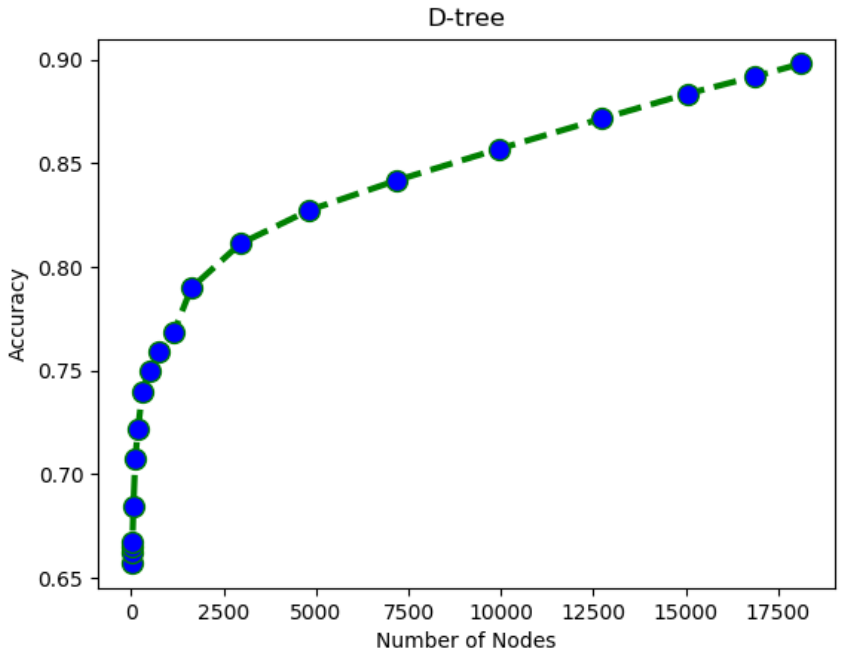
COL 774: Assignment 3 (Part A)

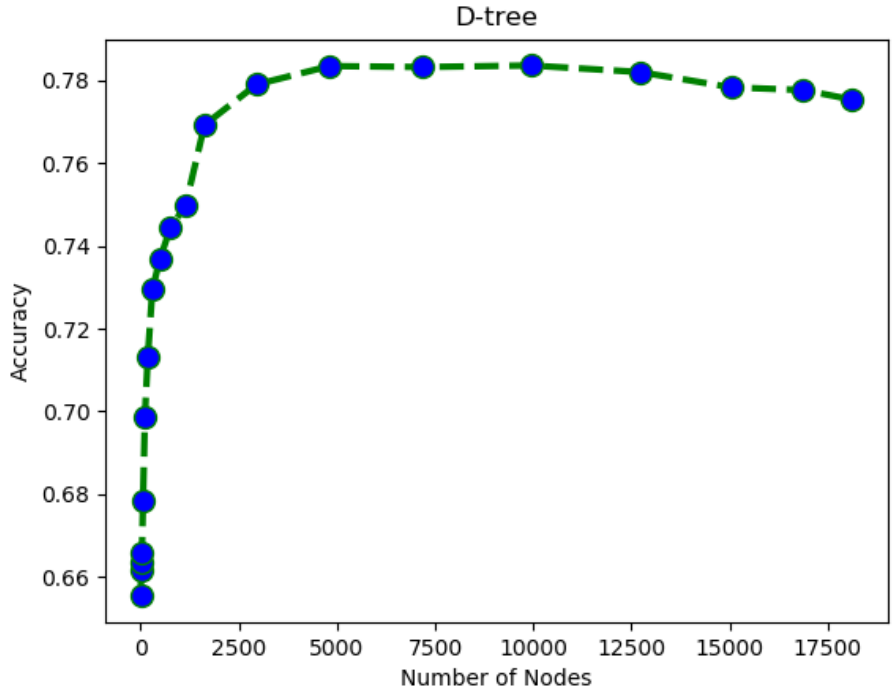
a.



Test Data - Accuracy = 77.57



Train Data - Accuracy = 90.27



Validation Data – Accuracy = 77.74

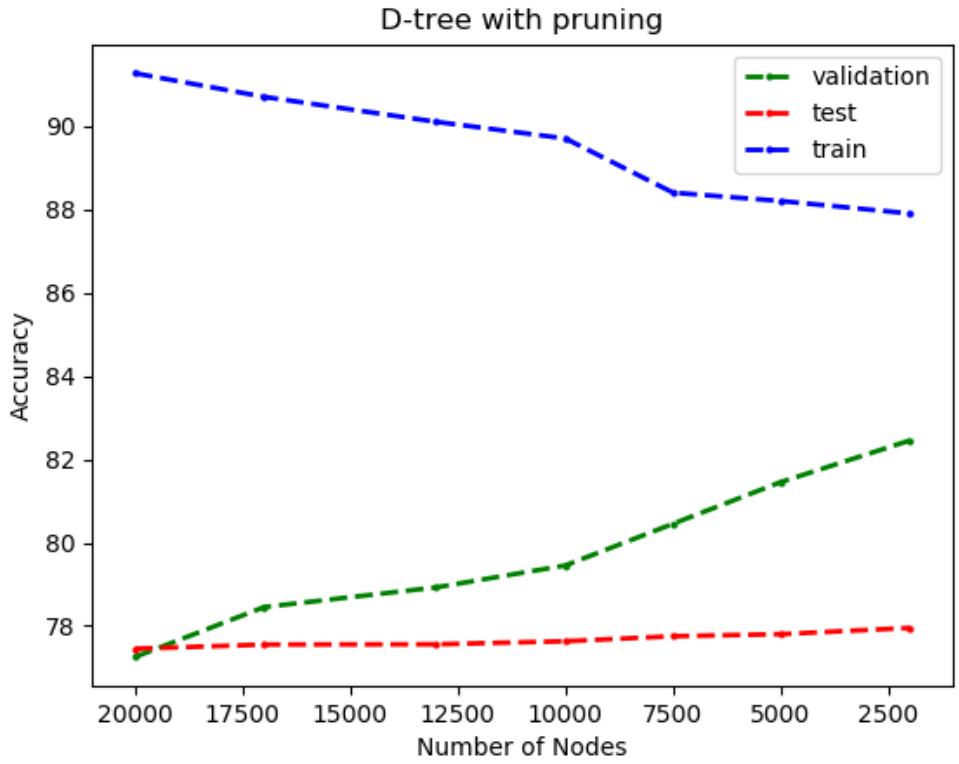
we can observe that the training accuracy increases with increase in the number of nodes. Also, the validation, test accuracies increase until a point and then start decreasing, the decision tree is overfitting on the training data. The decision tree is not able to completely fit the data because there are instances in the data where two samples having same feature values have different label.

Total number of Nodes – 21907

(Full Grown Tree with height of 52)

(10544 internal leaf nodes)

b.



From the graph we can see that the overfitting is decreasing because of post pruning. The training accuracy is decreasing while the validation, test accuracies are increasing.

Test Accuracy – 78.65
Train Accuracy – 87.45
Validation Accuracy – 82.12
Total number of Nodes – 2322

c.

optimal set of parameters obtained:

n_estimators: 350

min_sample_split: 10

max_features: 0.1

Train accuracy – 87.37

Test accuracy – 80.85

Validation accuracy – 80.67

Out of bag accuracy – 80.88

Accuracy obtained are pretty close those obtained using pruning but the values are better of using the **sklearn Library** implementation of random forest.

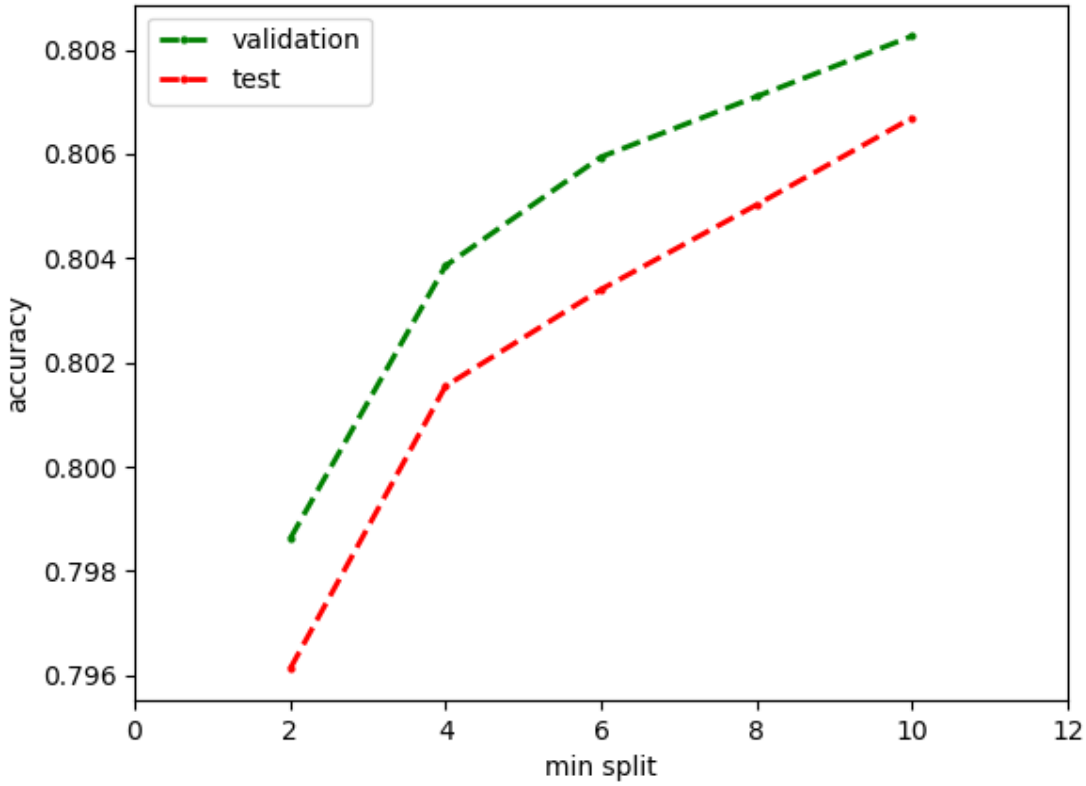
d.

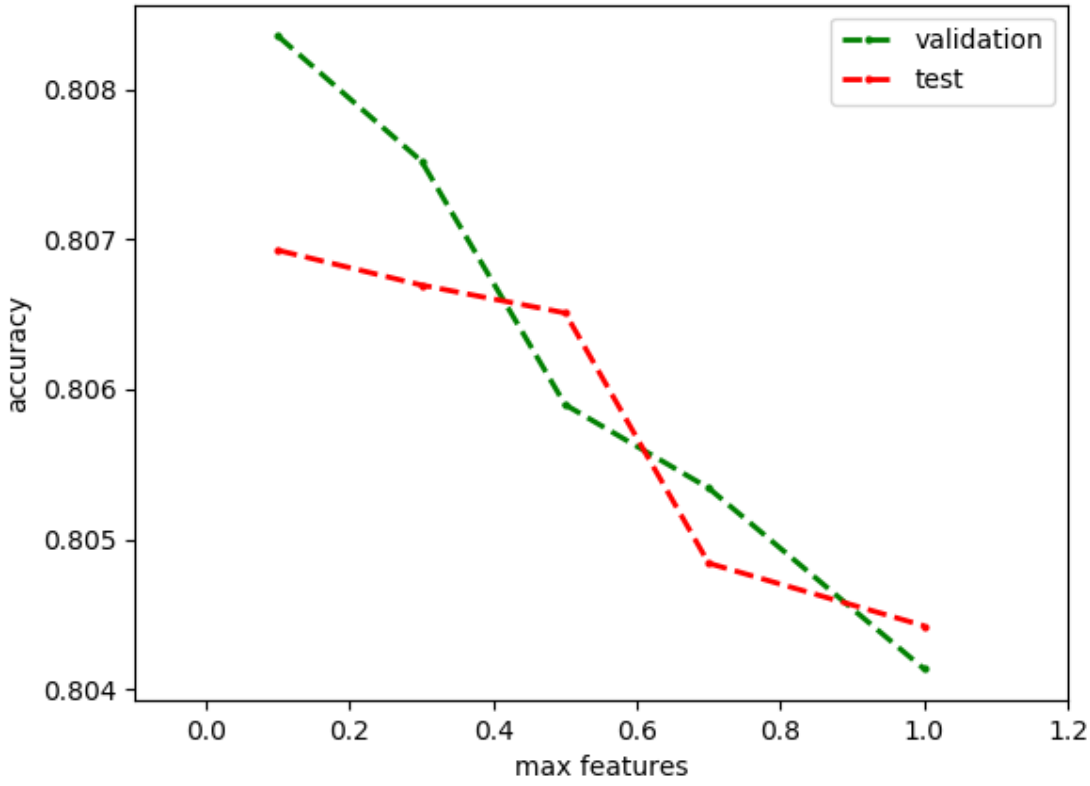
min_sample_split restricts the splitting of small nodes.

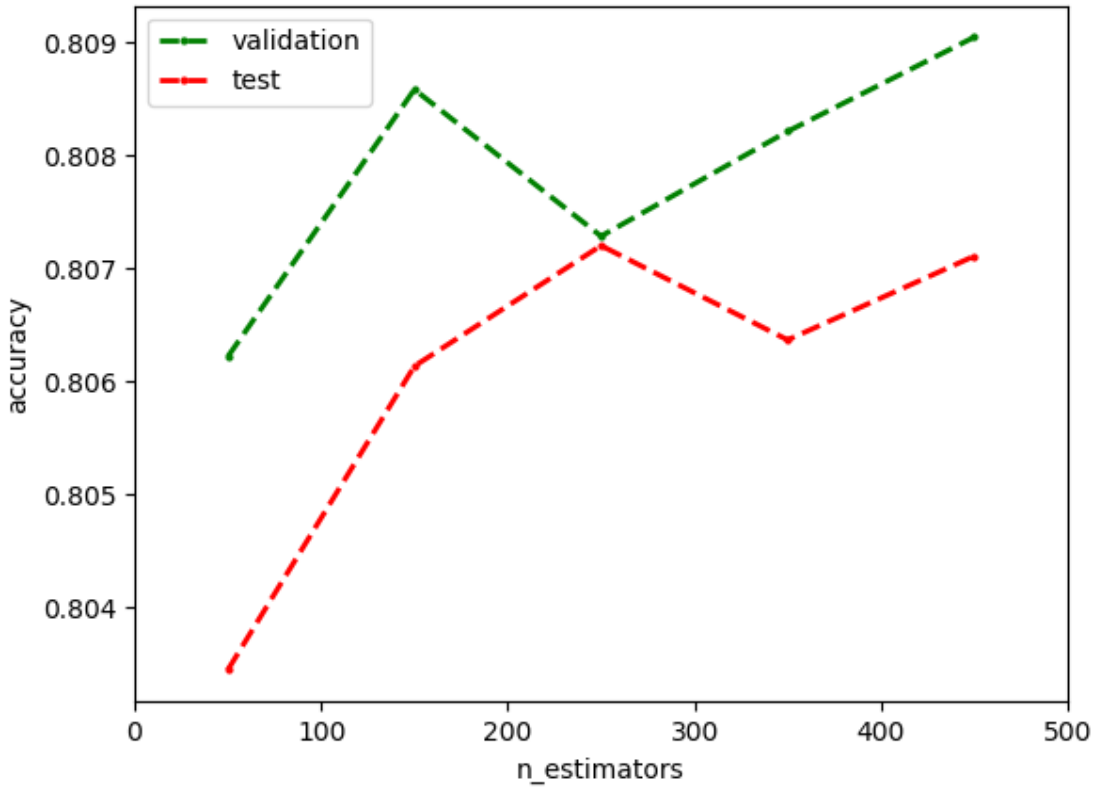
n_estimators specifies the number of decision trees in the random forest.

max_features restricts the number of features searched over while choosing the split.

It can be observed that as we deviate away from optimal params the accuracy decreases.


validation
test
0.808
0.806
0.804
0.802
0.800
0.798
0.796
accuracy
0
2
4
6
8
10
12
min split


validation
test
0.808
0.807
0.806
0.805
0.804
accuracy
0.0
0.2
0.4
0.6
0.8
1.0
1.2
max features



Report By:
Shivam Jadhav
2017CS10378