Rachit Kumar
2017CS10364

COL774 Assignment3-a

Q1. Decision Tree Construction:

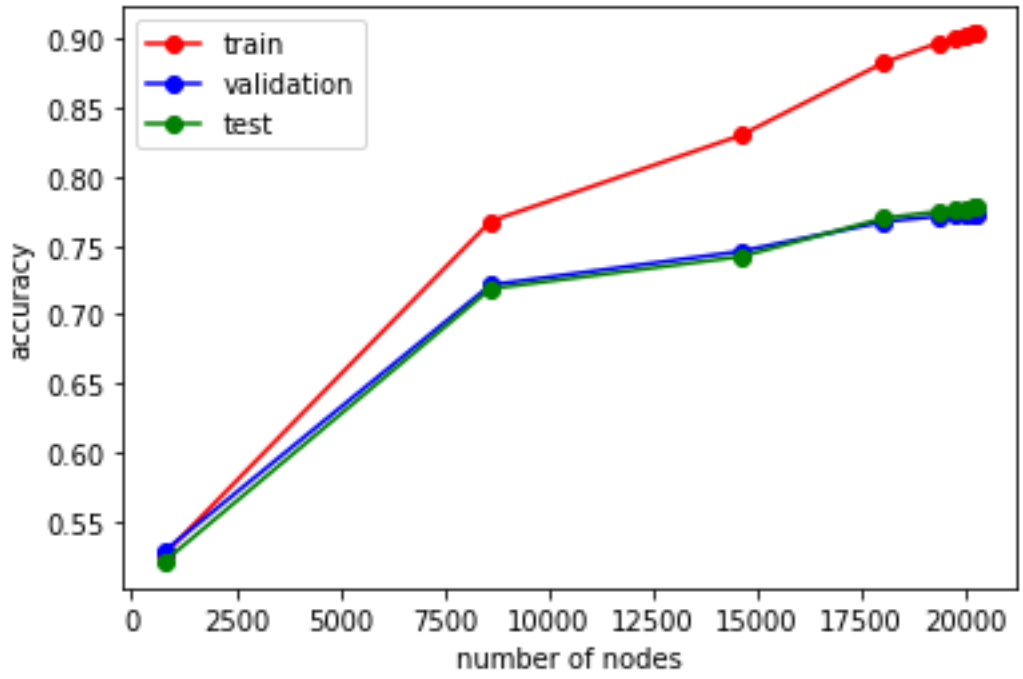
We see that as we increase the number of nodes, both the train and the validation accuracies increase, but only to a certain extent. We see that as the number of nodes increase, training accuracy increases faster than the test and validation accuracies. In general, it can be seen that the tree overfits as the train accuracy is much higher than the test and validation accuracy.

Finally, we had:

train accuracy: 0.9039451114922813

validation accuracy: 0.7723437789727424

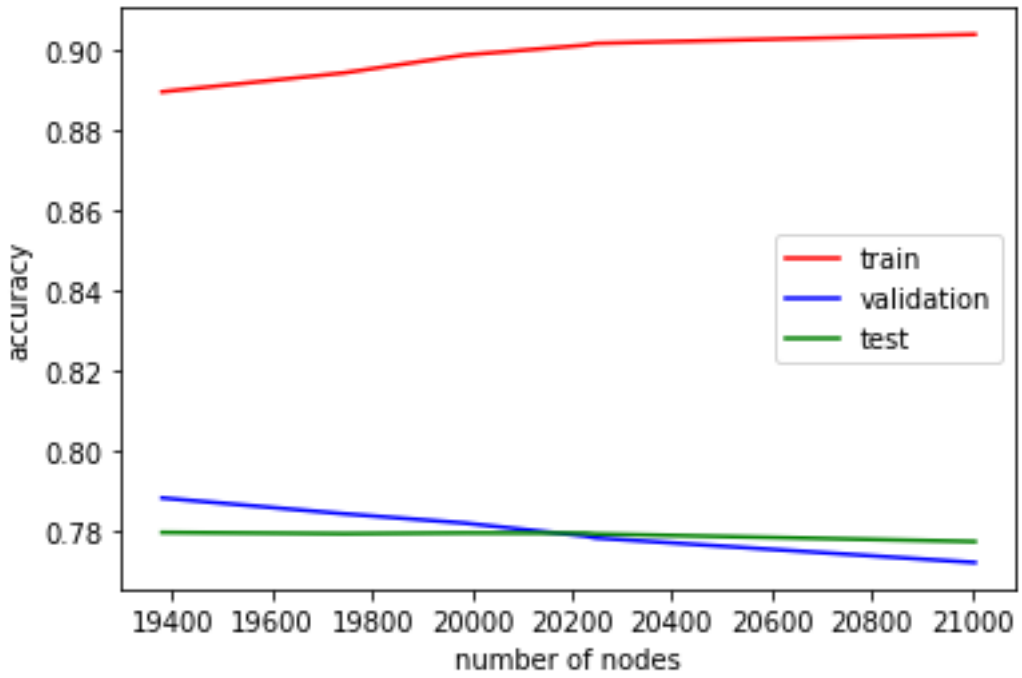
test accuracy: 0.7775717398358908



Q2. Decision Tree Post Pruning:

Pruning reduced overfitting and hence did improve the test and validation accuracies a bit.

But as we pruned more, it started to overfit on validation data.



Q3. Random Forests:

The optimal set of parameters found is:

n_estimators, max_features, min_samples_split= (450, 0.6, 10)

For these set of features:

out-of-bag accuracy: 0.8096518473876965

train accuracy: 0.8898212105759276

validation accuracy: 0.8037734099758946

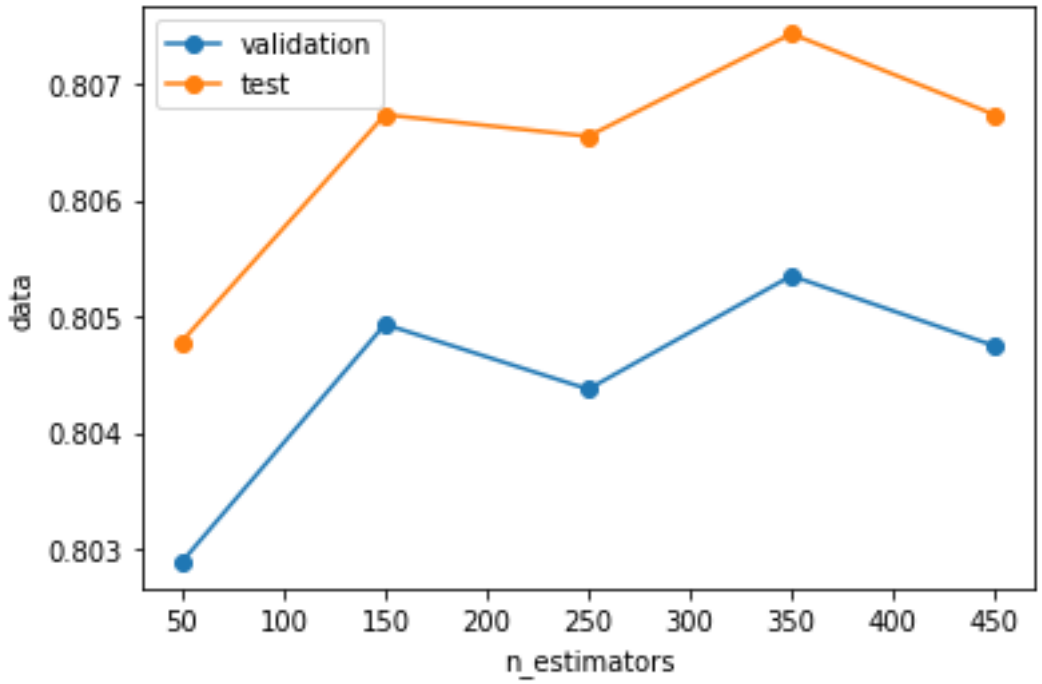
test accuracy: 0.8070094107829957

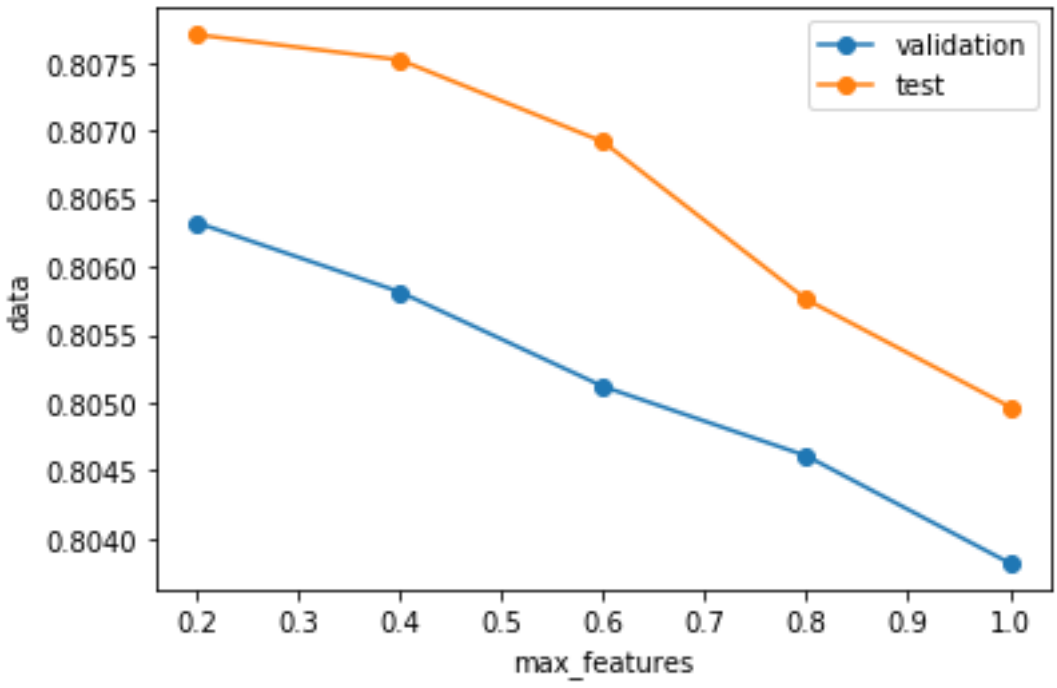
The test and validation accuracy obtained here was little better than what was obtained after pruning and the training accuracy was lower. Hence, overfitting has been reduced.

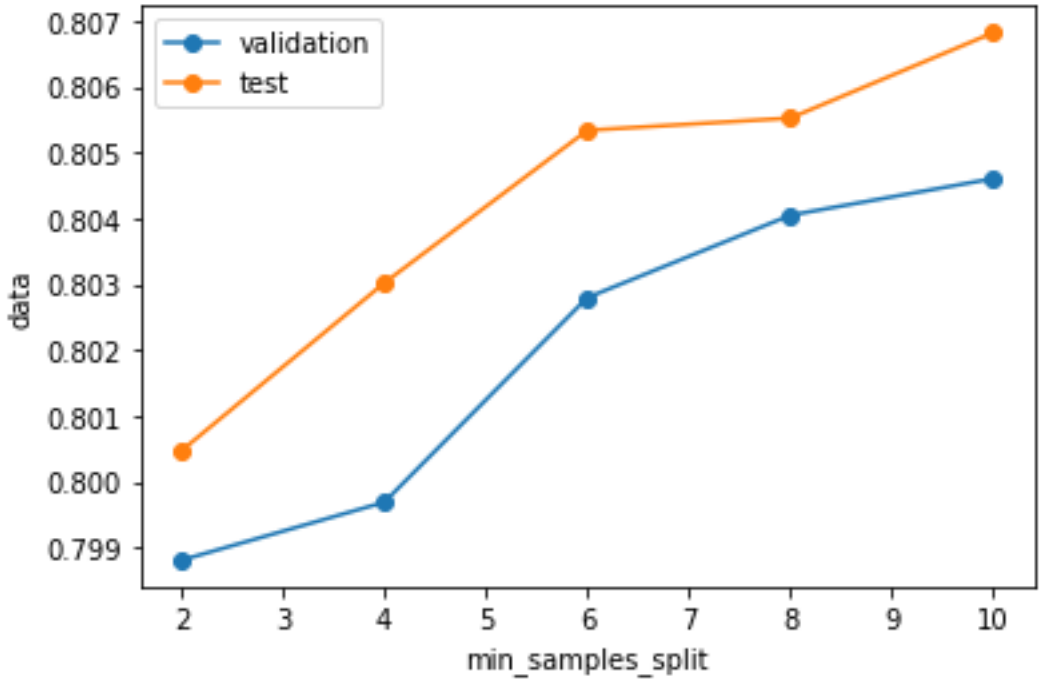
Q4. Random Forests - Parameter Sensitivity Analysis:

Here, we see that as we increase the minimum number of samples required to split an internal node, the accuracy increases. This should be obvious. Also, as we increase the number of features to consider when looking for the best split, the accuracy decreases. As we increase the number of trees in the forest, the train accuracy increases but the test accuracies first increase and then start decreasing due to overfitting. Also, bootstrap=true increased the accuracy by a small amount.

The model is most sensitive to the minimum number of samples required to split an internal node and least to the number of features to consider when looking for the best split.






validation
test
0.807
0.806
0.805
0.804
0.803
0.802
0.801
0.800
0.799
data
2
3
4
5
6
7
8
9
10
min_samples_split