REPORT
ASSN.-3(A)

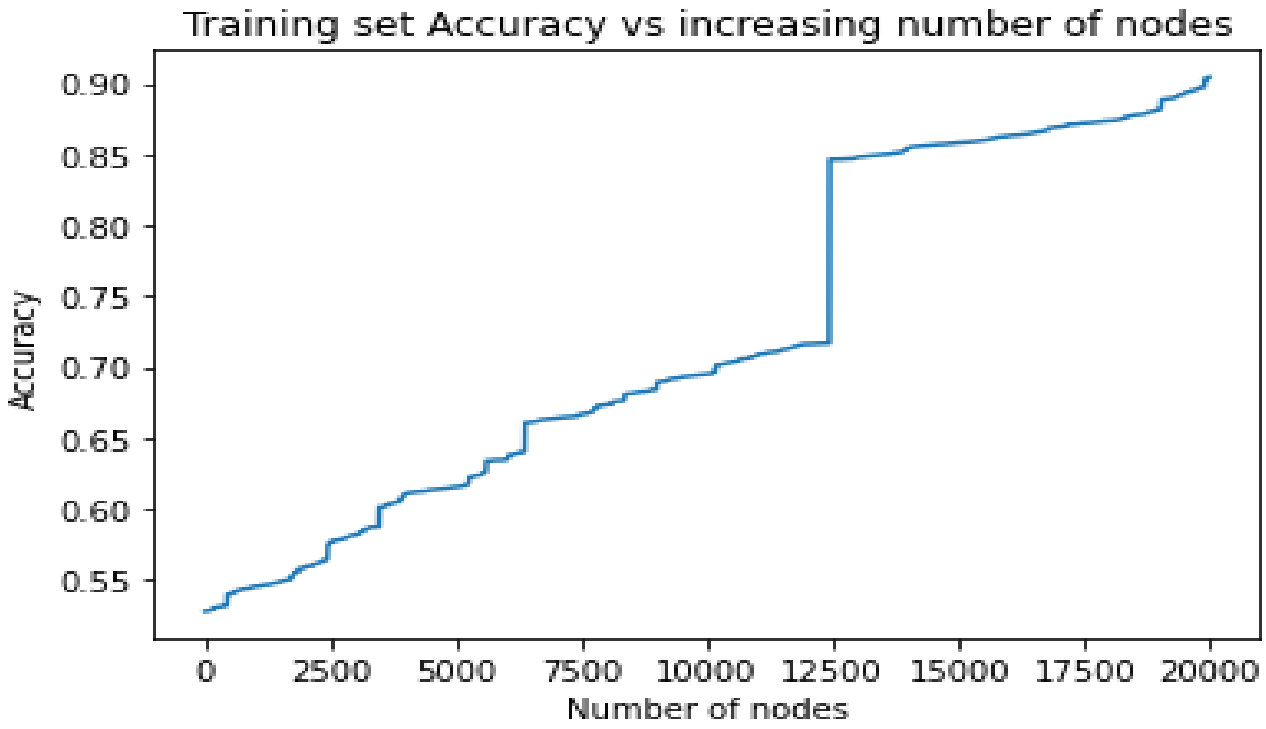
PRASHIT RAJ
2017CS10359

1.A. Decision Tree Construction:

Number of nodes = 19977

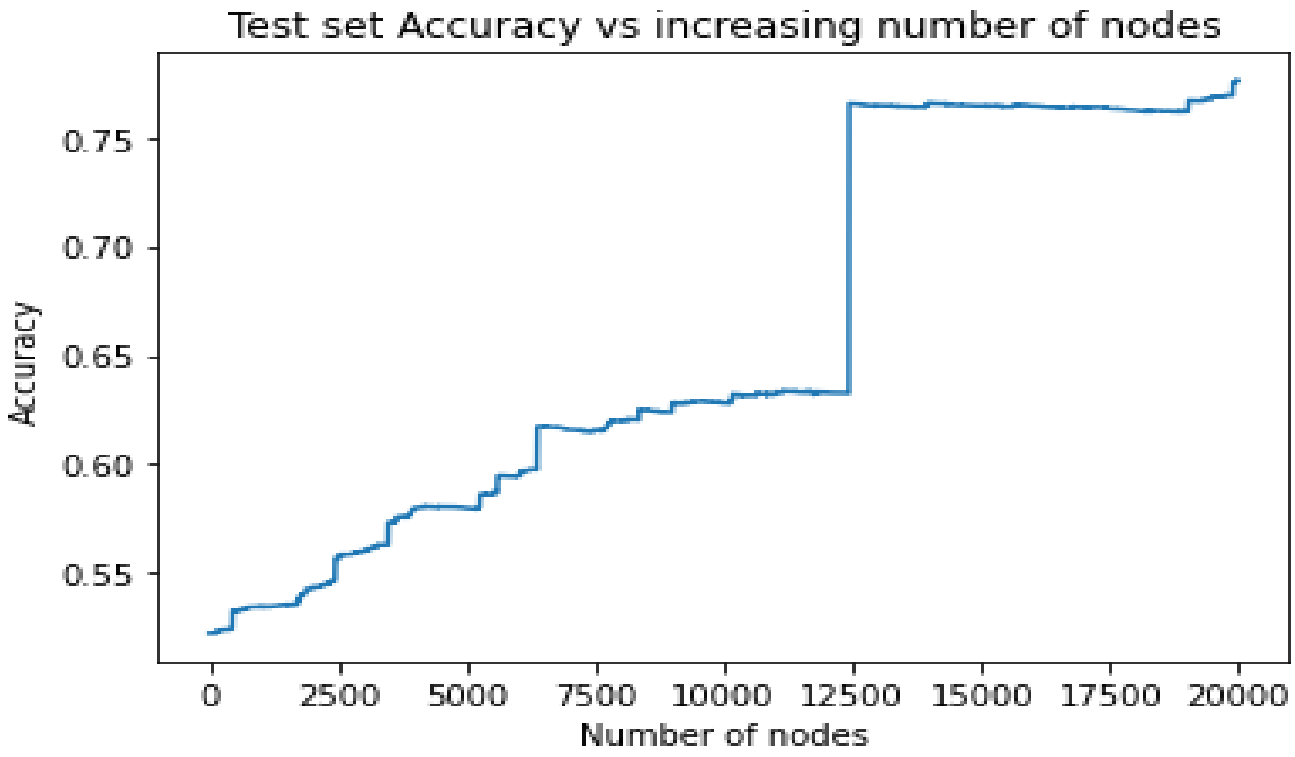
Training Set:

Accuracy = 0.904



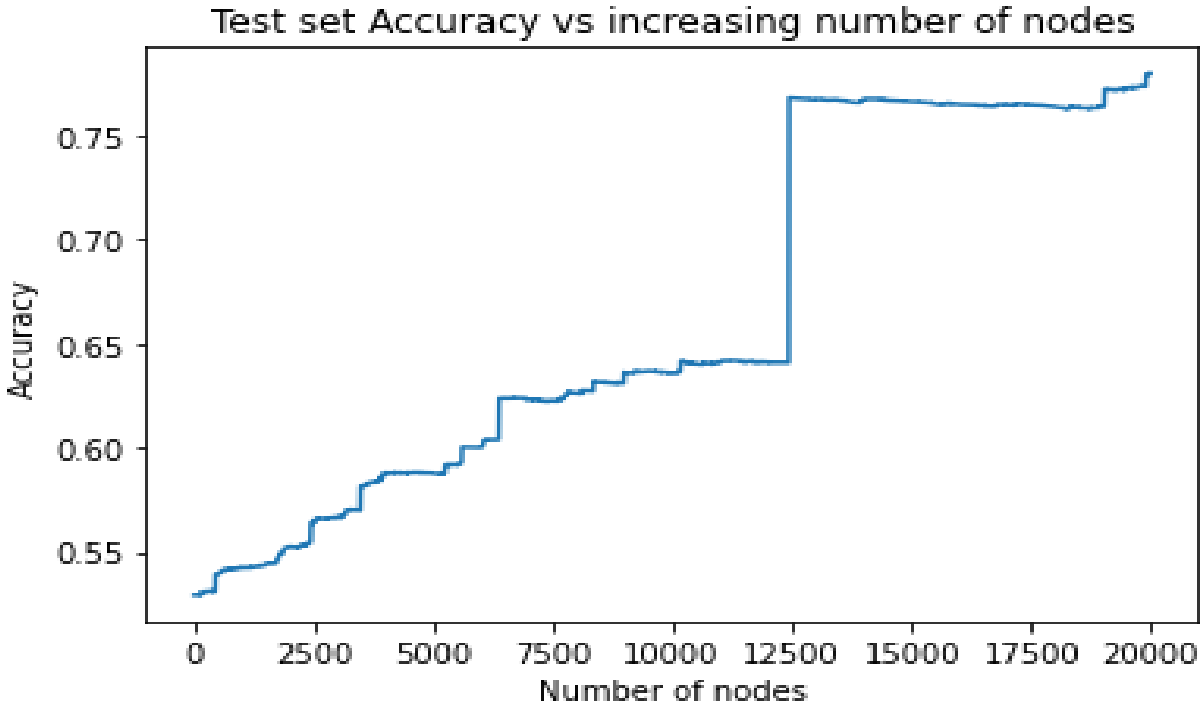
Validation Set:

Accuracy = 0.776

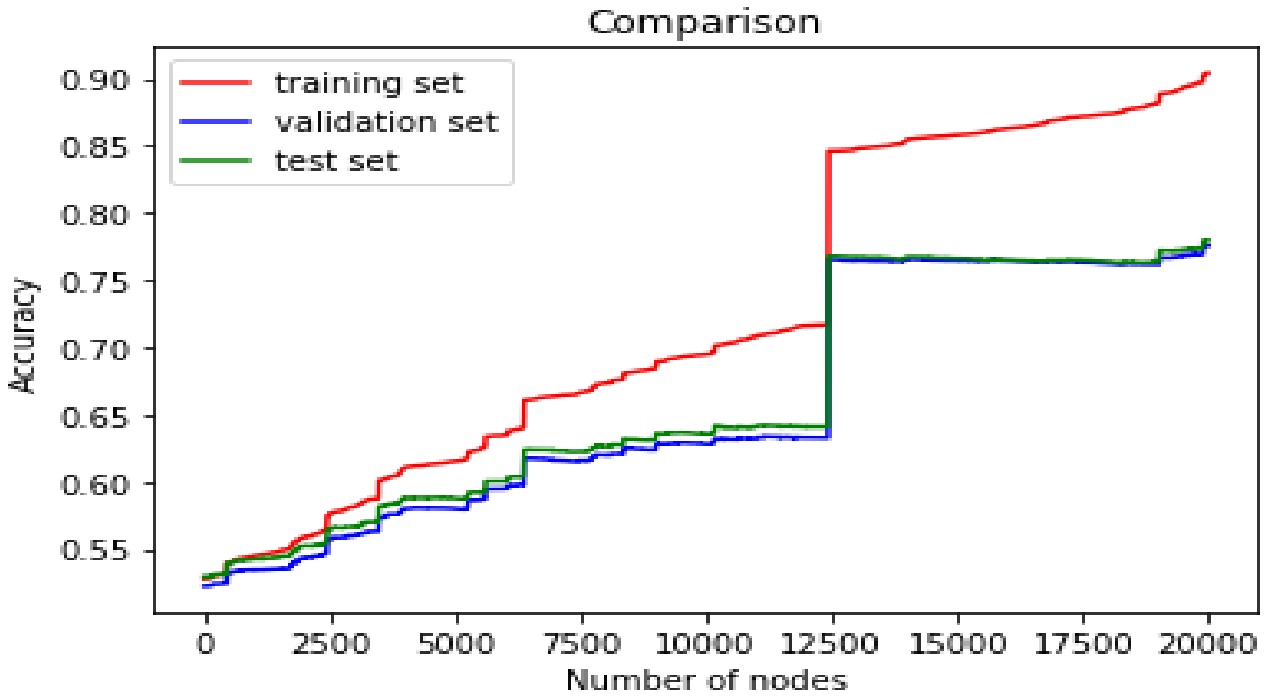


Test Set:

Accuracy = 0.779



Comparison:



Comment:

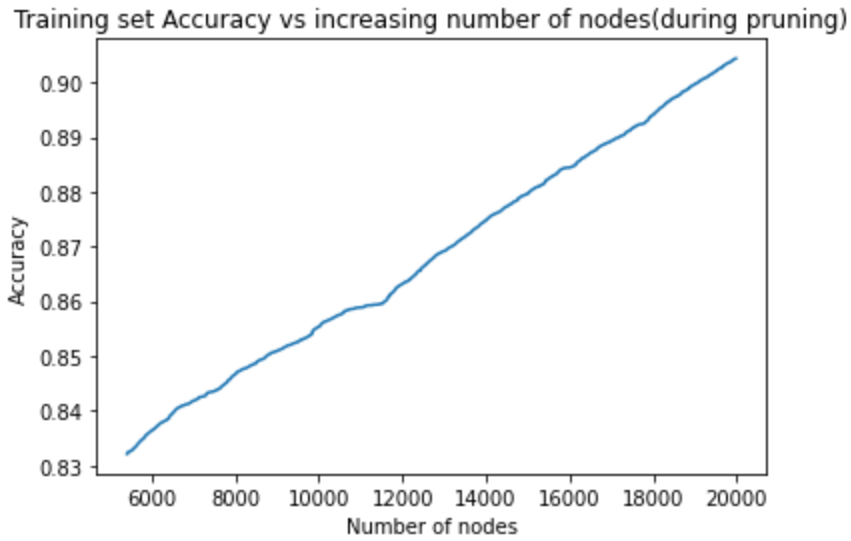
From the graph above we can see that the training set accuracy increases for every new node added to the tree whereas the test and validation accuracies decrease for some of the nodes added because of overfitting on training set which may have some noise associated with it.

1.B. Pruning:

Number of nodes = 5413

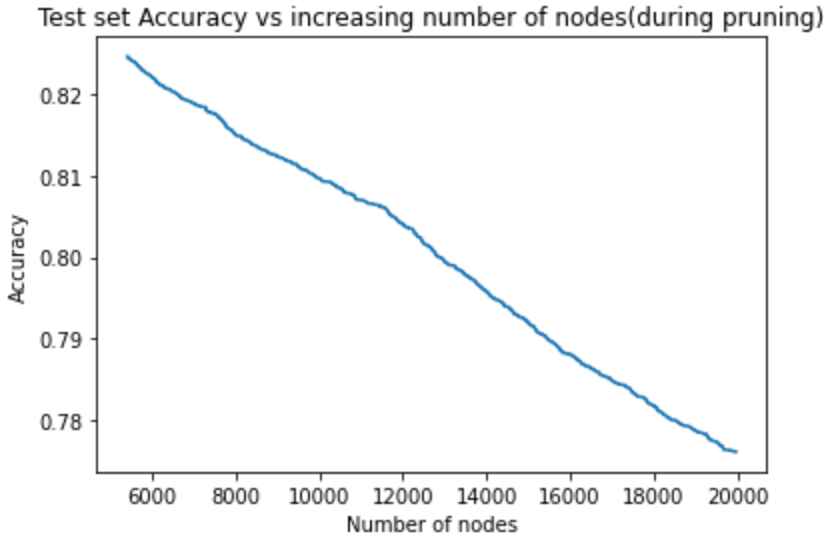
Training Set:

Accuracy = 0.832



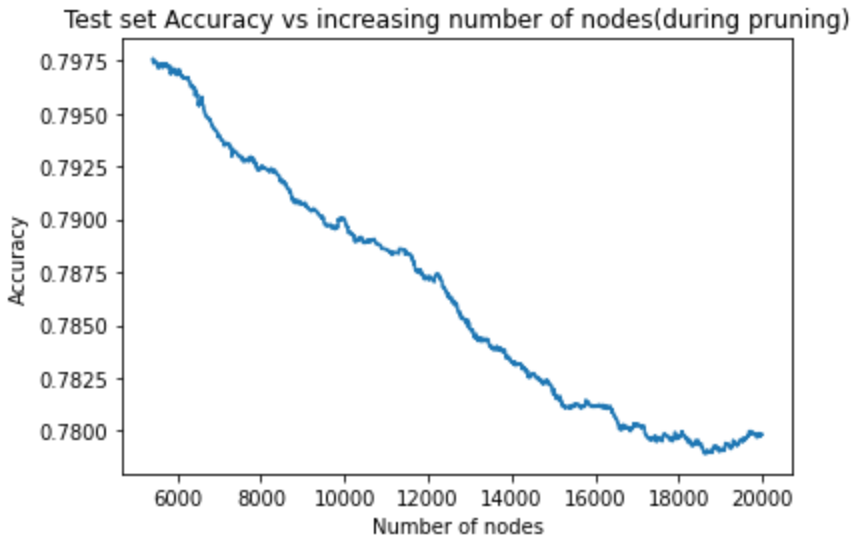
Validation Set:

Accuracy = 0.824

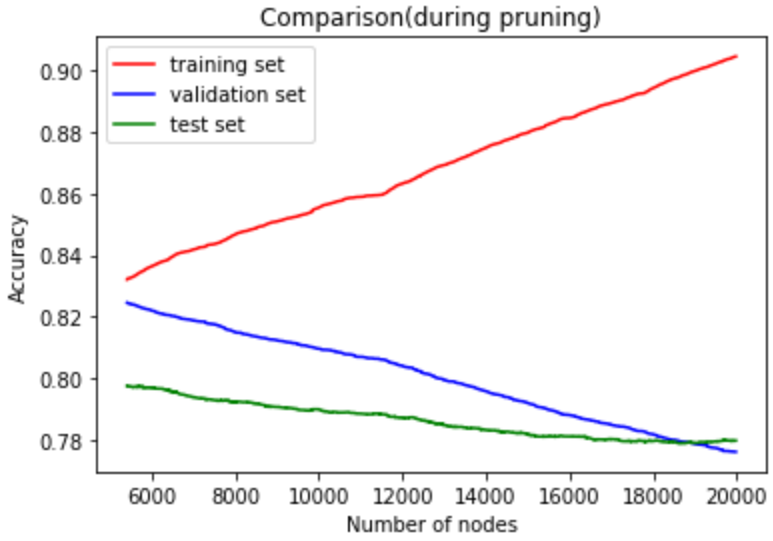


Test Set:

Accuracy = 0.797



Comparison:



Comment:

The accuracy of the training set decreases with decreasing number of nodes as we prune each node whereas each of the validation and test set accuracies increases with decreasing number of nodes. This is because pruning reduces overfitting in the decision tree giving a better generalization

1.C. Random Forest Classifier:

Optimum set of parameters:

n_estimators = 350

min_samples_split = 10
Max_features = 0.1

Accuracies for optimum set of parameters:

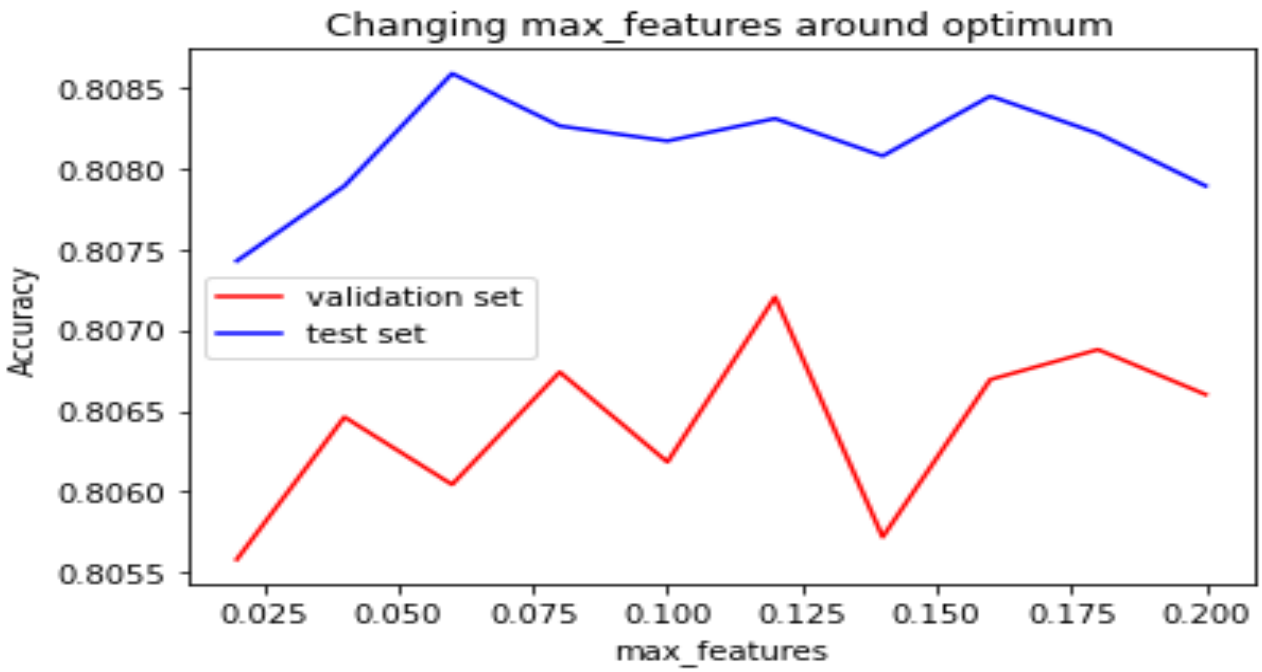
oob_score = 81.072%
Training set = 87.644%
Test set = 80.733%
Validation set = 80.623%

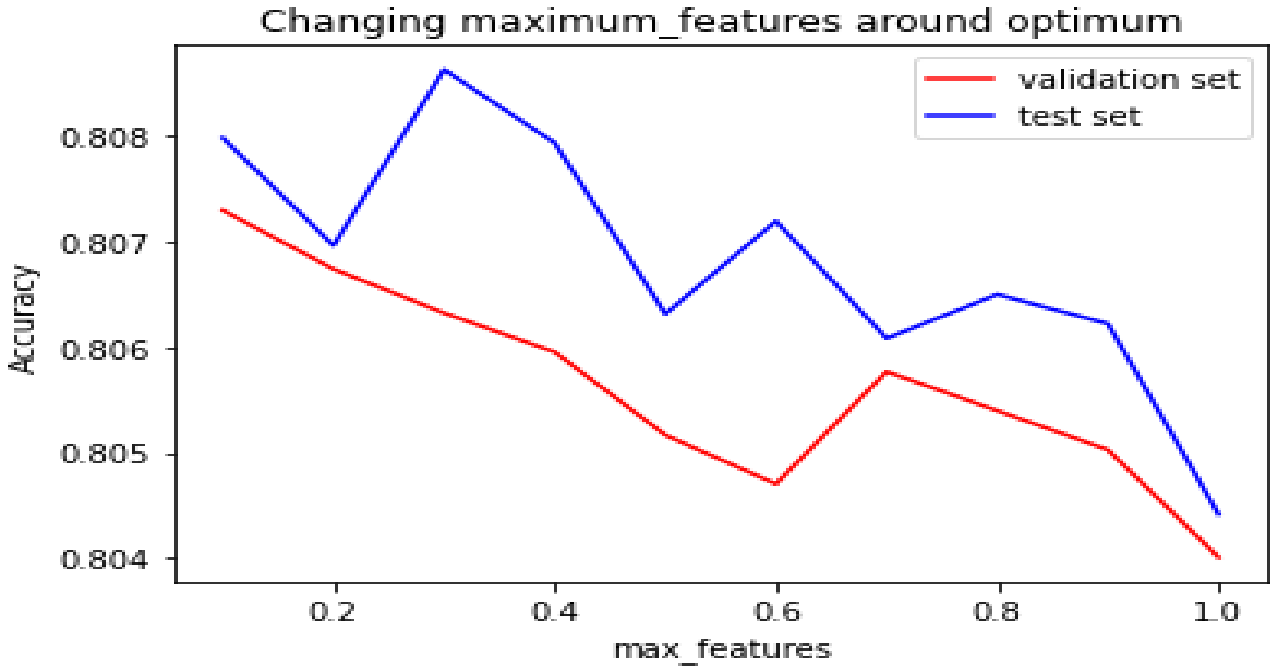
Comment:

As compared to part (a) the training set accuracy has decreased and that of test and validation set has substantially increased which shows that overfitting decreases while using the random feature model because it builds multiple trees and then chooses one with the highest accuracy over a set not used for training. This random selection helps to eliminate noise in the data and thus the overfitting noticed using a single decision tree is eliminated.
As compared to part (b) the training accuracy has increased. This is due to the fact that while pruning the accuracy of the training set is reduced at the cost of accuracy over validation set whereas no such thing happens in the random forest as it uses multiple decision trees to overcome overfitting. Also, accuracy over the test set has increased but the magnitude of change is less compared to (a). The accuracy over validation set has however decreased because we did pruning in the (b) to increase accuracy over validation set whereas nothing of such sort is done in this part.
I also used GridSearchCV to know the optimum parameters using CV(cv = 3) score as a criterion for ranking

1.D. Parameter Sensitivity Analysis:

Maximum Features:

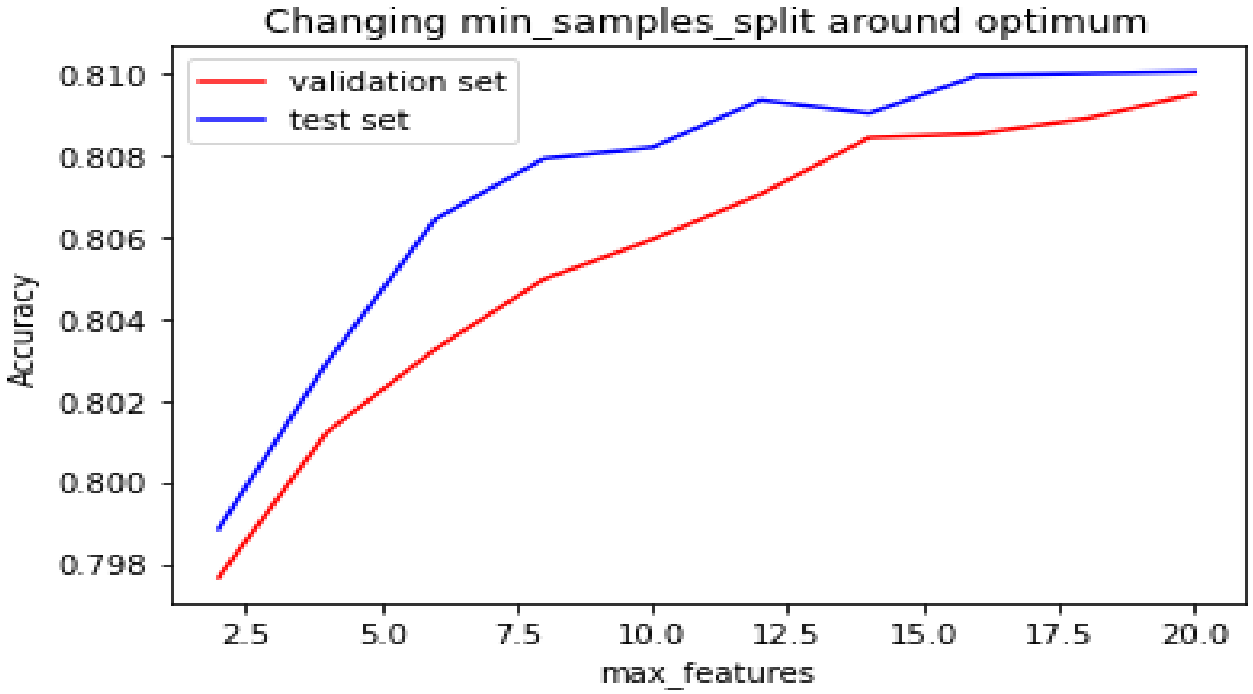




For all the features I have plotted the accuracy in 0.2 to 2 factor of the optimum value at intervals of 0.2. But for max_feature the oob_score comes out to be very similar and there is no trend that can be observed for low values so I plotted another graph for it in

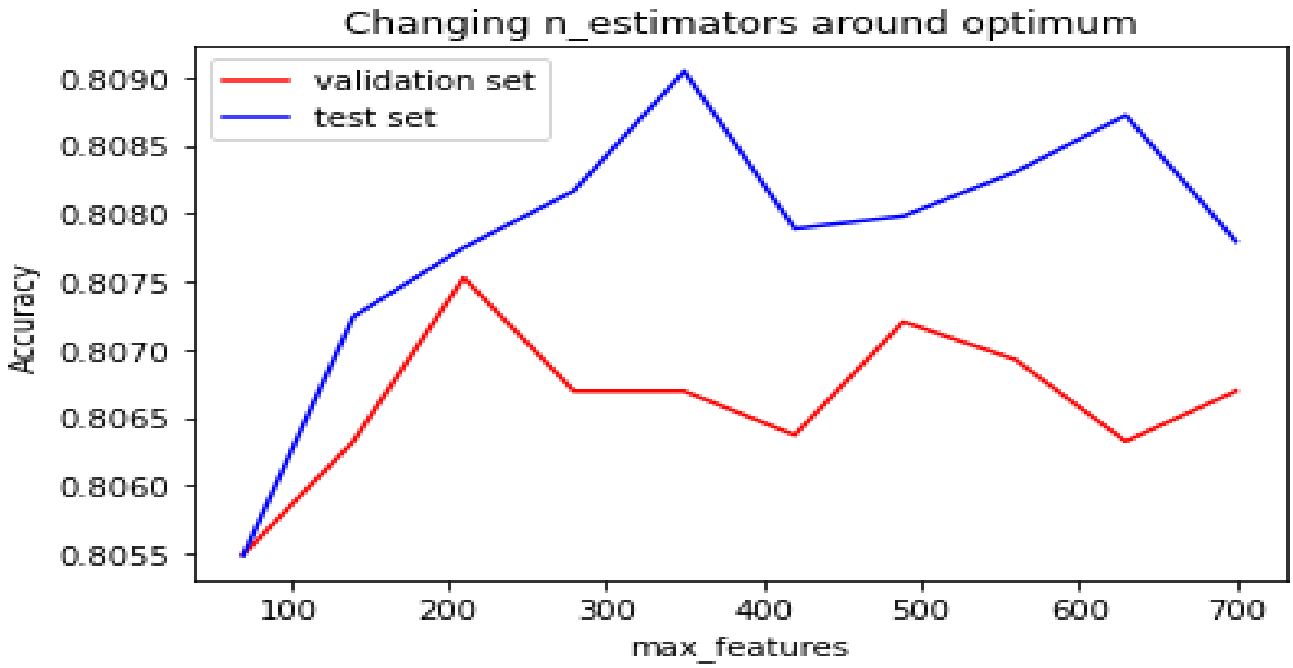
the range [0.1,1] at intervals of 0.1. In the latter case, accuracy shows a decreasing trend.

Minimum Samples Split:



There is a clear trend that can be noted down in this graph that the accuracy increases with increasing max_features value.

N_Estimators:



From the above figure, we see that for n_estimators there is no trend that can be noted. Sometimes the values increase, sometimes they decrease. The accuracies differ by a very small number, within 0.1% for both test and validation set.