# Machine Learning-Assignment 3 (Part A)

Kartik Jain-2019MCS2563

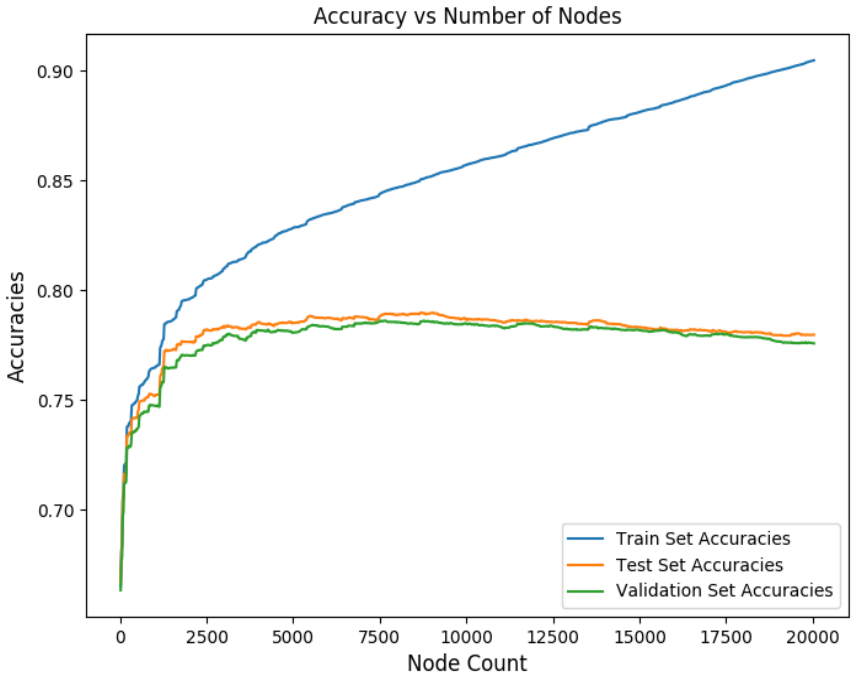
## 1 Decision Trees and Random Forests

### 1.1 Decision Tree Construction

The accuracy on the datasets generated after training the tree is as follows:

| | |
|---|---|
| Training Set | 90.47 |
| Test Set | 77.97 |
| Validation Set | 77.59 |

The plot generated is as follows:



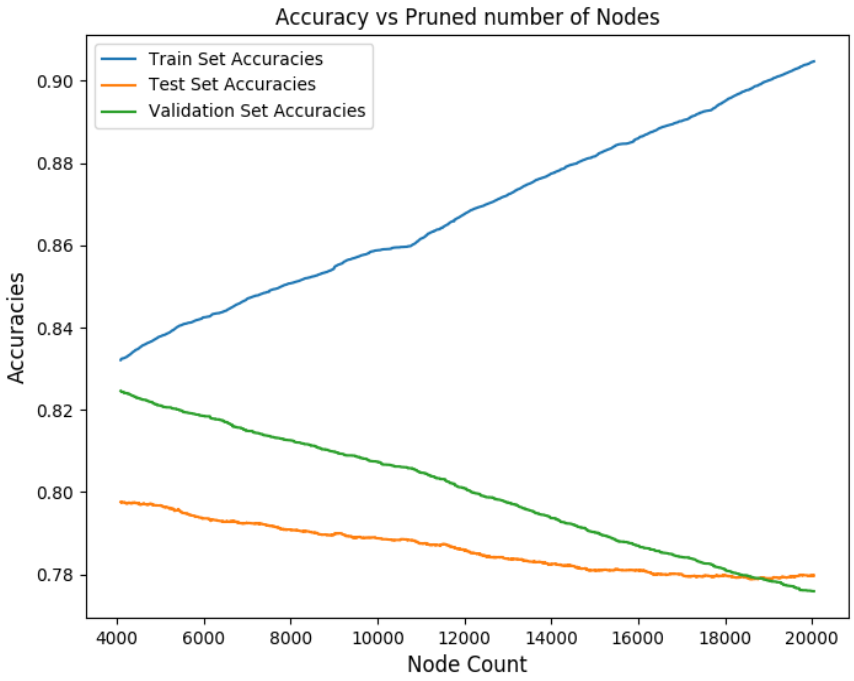
We observed that as we increased the number of nodes the accuracy on all the data sets increased upto a certain number of nodes and after that the tree started to overfit on the training data and the accuracy on train data increased while the accuracy on the other data sets started reducing slowly.

## 1.2 Decision Tree Post Pruning

The accuracy on the datasets generated after pruning the tree is as follows:

| | |
|---|---|
| Training Set | 83.21 |
| Test Set | 79.76 |
| Validation Set | 82.45 |

The plot generated is as follows:




Post pruning is used to reduce the overfitting of a fully grown tree. Here in the graph as we pruned the subtrees based on the validation dataset we are reducing the overfitting on the train set and thus the accuracy of train set originally was reduced and the accuracy on other data sets including testing set was increased. Now the tree generated is not overfitting the train dataset.

### 1.3 Random Forests

I have used grid search with cross validation equal to 3 over the given set of parameters and found that the optimal set of parameters are:

| | |
|---|---|
| n_estimators | 450 |
| max_features | 0.1 |
| min_samples_split | 10 |

And the accuracies obtained using these parameters are as follows:

| | |
|---|---|
| Training Set Accuracy | 87.62 |
| Out of Bag Accuracy | 81.01 |
| Testing set Accuracy | 80.83 |
| Validation set Accuracy | 80.63 |

And by using grid search without cross validation over the given set of parameters we found the optimal set of parameters to be:

| | |
|---|---|
| n_estimators | 350 |
| max_features | 0.1 |
| min_samples_split | 10 |

And the accuracies obtained using these parameters are as follows:

| | |
|---|---|
| Training Set Accuracy | 87.61 |
| Out of Bag Accuracy | 81.00 |
| Testing set Accuracy | 80.81 |
| Validation set Accuracy | 80.64 |

These accuracies are almost the same with very little differnece between them, and as random forests use bootstrapped samples generated from the original dataset randomly, thus these accuracies also keep changing but only with a very little amount.

The accuracies obtained above are slightly greater than the ones obtained after pruning in the above part as shown by the table below:

| | **With random forest** | **After Pruning** |
|---|---|---|
| Training Set Accuracy | 87.61 | 83.21 |
| Testing set Accuracy | 80.81 | 79.76 |
| Validation set Accuracy | 80.64 | 82.45 |

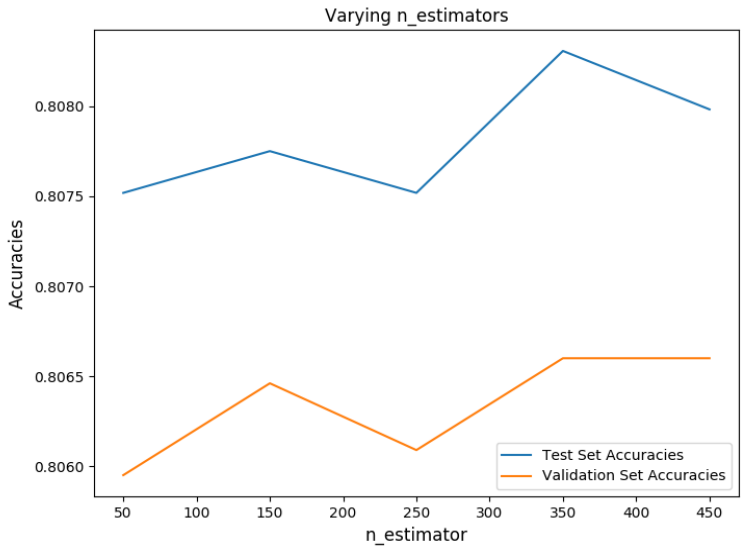
These accuracies might be better due to the ensemble model that the random forests uses which involves including creating multiple decision trees to make the prediction and thus the random forests are able to generalize the given dataset better which results in a higher accuracy (by about **1 percent**) on the test set.

## 1.4 Random Forests-Parameter Sensitivity Analysis

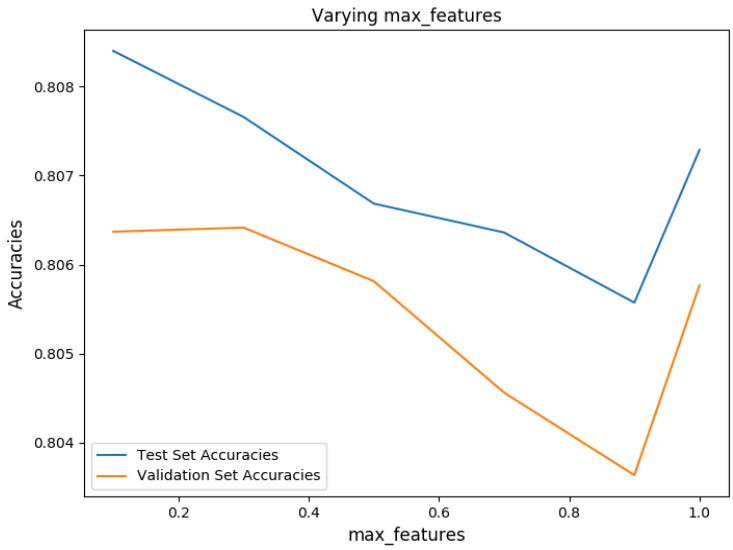
The variation plots obtained by varying each of the parameters obtained keeping the other two parameters are fixed are shown below.

For grid search **without cross validation**:

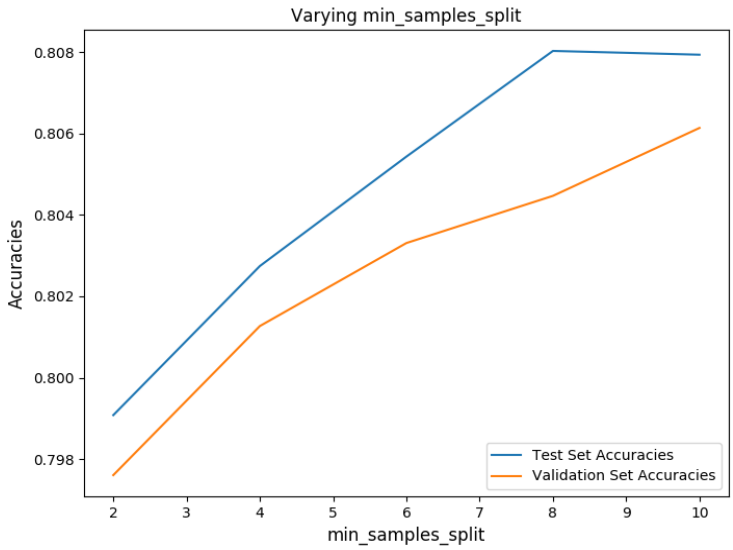
1. Varying n_estimators keeping max_features and min_samples_split to be 0.1 and 10 respectively.




2. Varying max_features keeping n_estimators and min_samples_split to be 350 and 10 respectively.

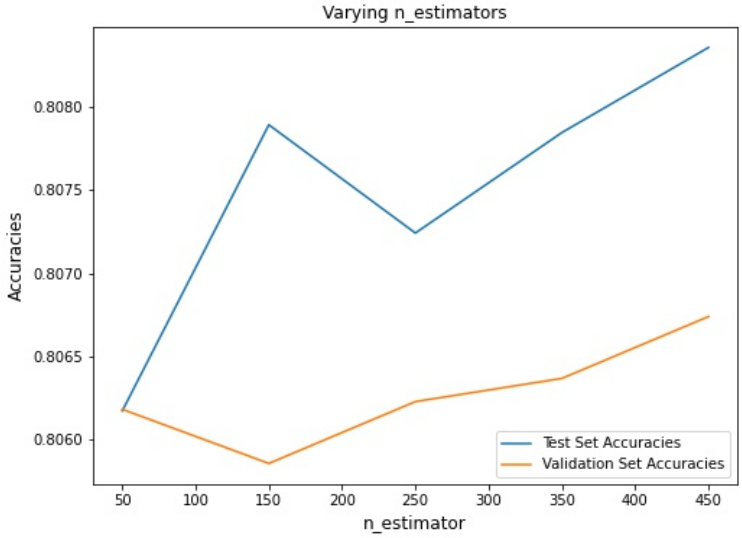



3. Varying min_samples_split keeping n_estimators and max_features to be 350 and 0.1 respectively.

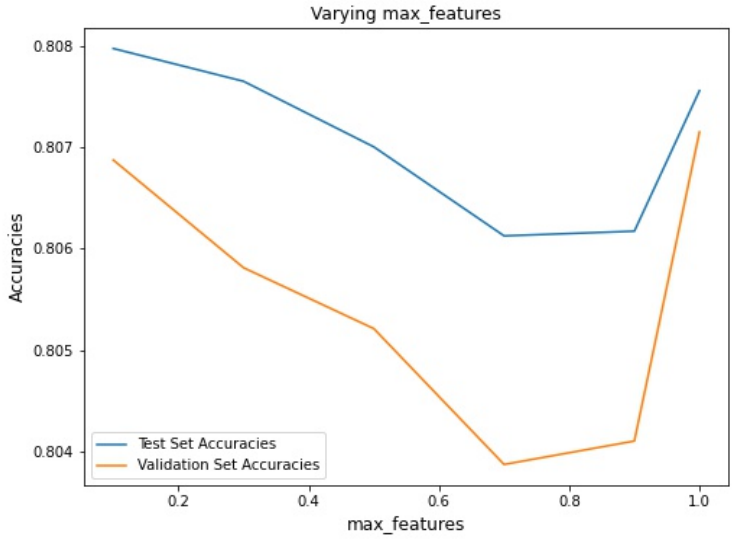



For grid search **with cross validation equal to 3**:

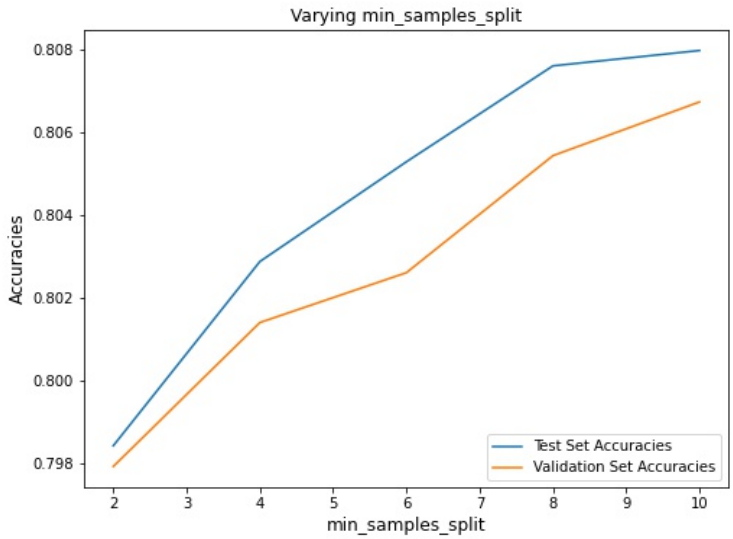
1. Varying n_estimators keeping max_features and min_samples_split to be 0.1 and 10 respectively.




2. Varying max_features keeping n_estimators and min_samples_split to be 450 and 10 respectively.




3. Varying min_samples_split keeping n_estimators and max_features to be 450 and 0.1 respectively.




As it is also clear from the above graphs there is very little difference between the accuracies obtained from using n_estimators to be 450 and 350 which was also clear by running the code multiple times as sometimes it gave higher accuracy on 450 and sometimes on 350.
While it is very clear from the other two graphs that higher accuracy was achieved on 0.1 and 10 for max_features and min_samples_split.

Also, the random forest model is fairly sensitive to **max_features** which selects the number of features to use when training the model as higher the value of max_features the more features we will be using thus it helped in dimensionality reduction.
Varying **n_estimatores** does not resulted in much difference in accuracy after a certain number of features as the increase/decrease in accuracy is not much for the values in range [250,450] which is clear from the graph. n_estimators in the number of trees in the forest thus after a certain number of trees are present in the forest then further increasing the number of trees in the forest does not result in much incerement of accuracy.
Varying **min_samples_split** upto 10 resulted in continuous increment of accuracy,since the parameter was to split a node only if the samples at that node exceeded a particular number. By varying this parameter we reduced the overfitting of the model on the train set since now there will be splitting only if the number of samples at a node exceed a particular limit otherwise in the worst case the nodes will be splitted for each data point untill only 1 sample remained at the leaf and thus overfitting the model on the train set.