Machine Learning - Assignment 3

Meraj Ahmed

2019MCS2565

1 Decision Tree Construction

1.1 Introduction

We have to implement decision tree. We have built the decision tree best first search manner.
We split the nodes of decision tree with respect to best attribute.
Best attribute is determined by choosing that attribute which gives maximum information gain. We split the data at each node about its median, median value data goes to left child of node.

1.2 Parameters

No. of nodes in decision tree = 19913
No. of leaves in decision tree = 9957
Tree depth = 51
Time to build = 585 sec
Train accuracy = 90.4485 %
Test accuracy = 77.8777 %
Validation accuracy = 77.6284 %

1.3 Plots and figures

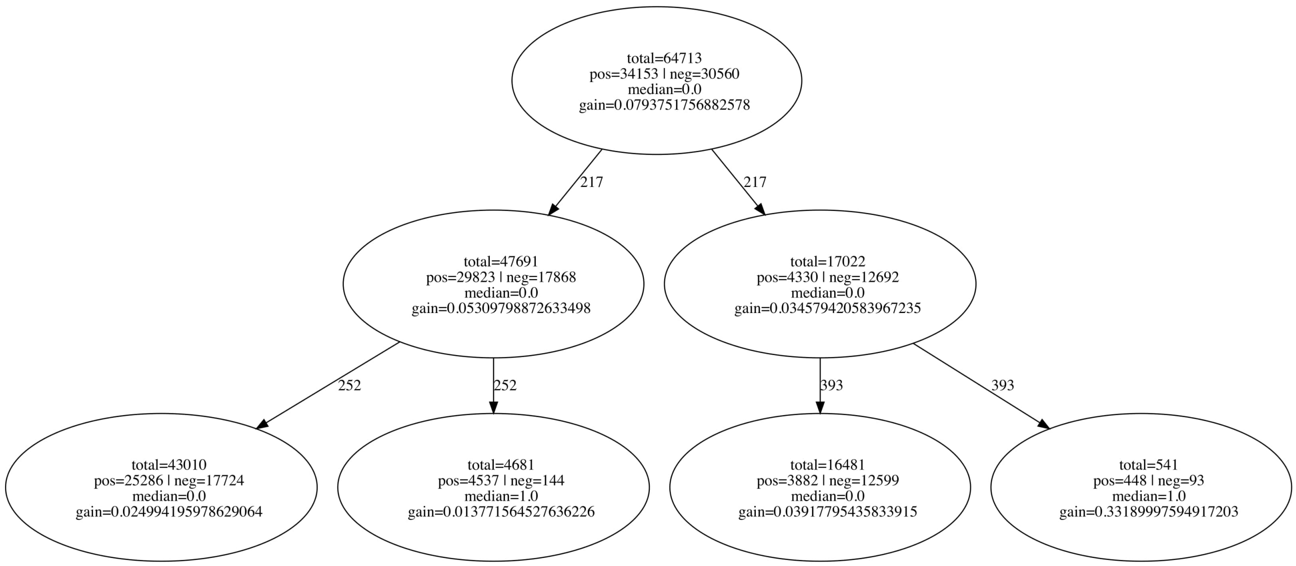

Figure 1: Partially built decision tree

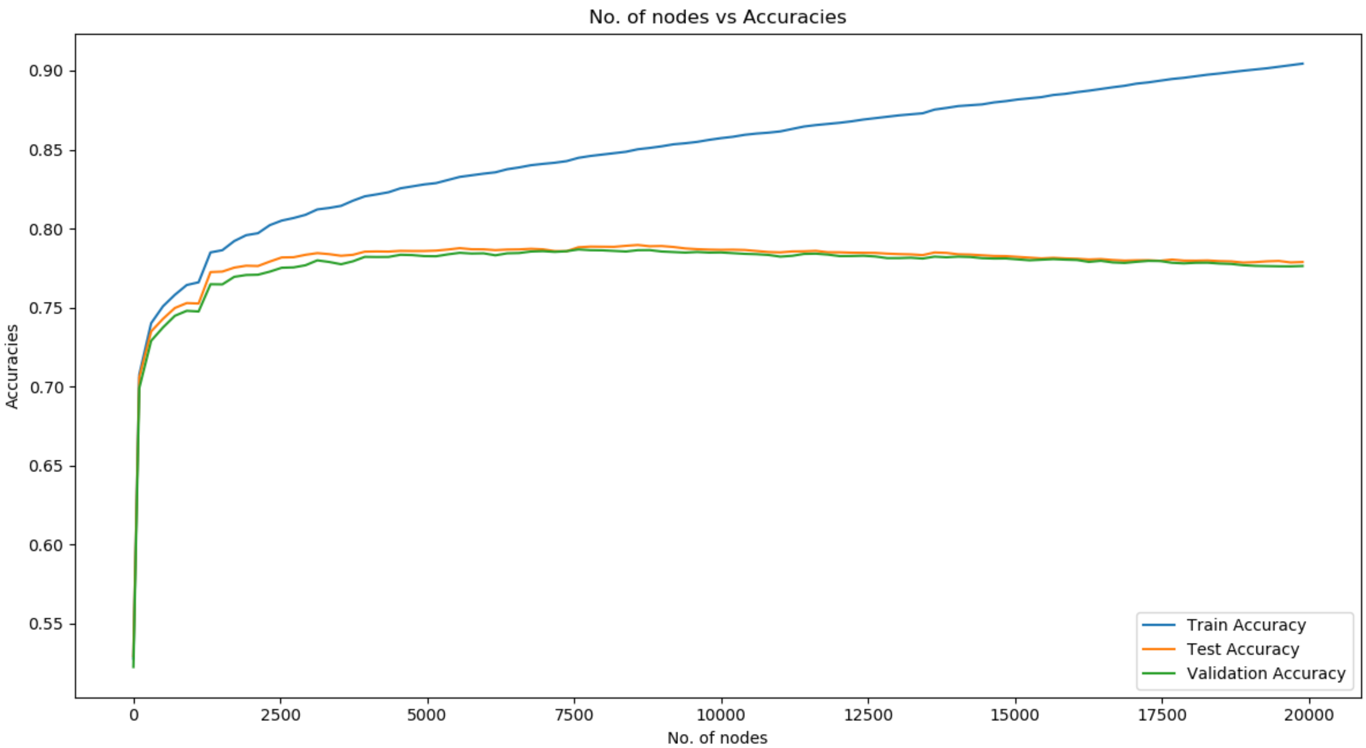

Figure 2: Accuracy vs no. of nodes

1.4 Comments

Initially with very less nodes in decision tree, train/test/val accuracy is about 0.5.

But after adding some more nodes, all accuracies ramp up quickly as we can see from the plot.

After adding some more nodes train accuracy is increased continuously but test/val acccuracies increases very slowly

After adding some more nodes test/val accuracies seems to decrease slightly.

Overall, conclusion is that, our decision tree is over-fit to training set and its generalization accuracy doesn't increases even after adding many nodes in the tree.

It is happening because our decision tree is biased towards train set, so it may also try to fit the noise in the data.

One solution to get rid of this to prune the decision tree on the validation set.

2 Decision Tree Post Pruning

2.1 Introduction

Decision tree's generalization accuracy can be improved by pruning the tree and removing some of the sub-trees.

We post prune the tree on validation set. We have tried two approach:

(i) In first approach, we traverse the tree in post-order traversal and for each non leaf nodes, we calculate the validation accuracy before and after pruning that node. If validation accuracy increases then we prune that node and continue in post order.

(ii) In Second approach, we tries to minimizes the no. of misclassification of sub-tree on validation set. If sub-trees misclassification is more than if that sub-trees were pruned than we prune that node.

2.2 Parameters

2.2.1 First Approach

No. of nodes in decision tree = 14207

No. of leaves in decision tree = 7104

Tree depth = 51

Train accuracy = 86.9516 %

Validation accuracy = 82.3660 %

Test accuracy = 79.0969 %

2.2.2 Second Approach

No. of nodes in decision tree = 4697

No. of leaves in decision tree = 2349

Tree depth = 50

Train accuracy = 81.0378 %

Validation accuracy = 83.9328 %

Test accuracy = 78.4479 %

2.3 Plots

2.3.1 First Approach

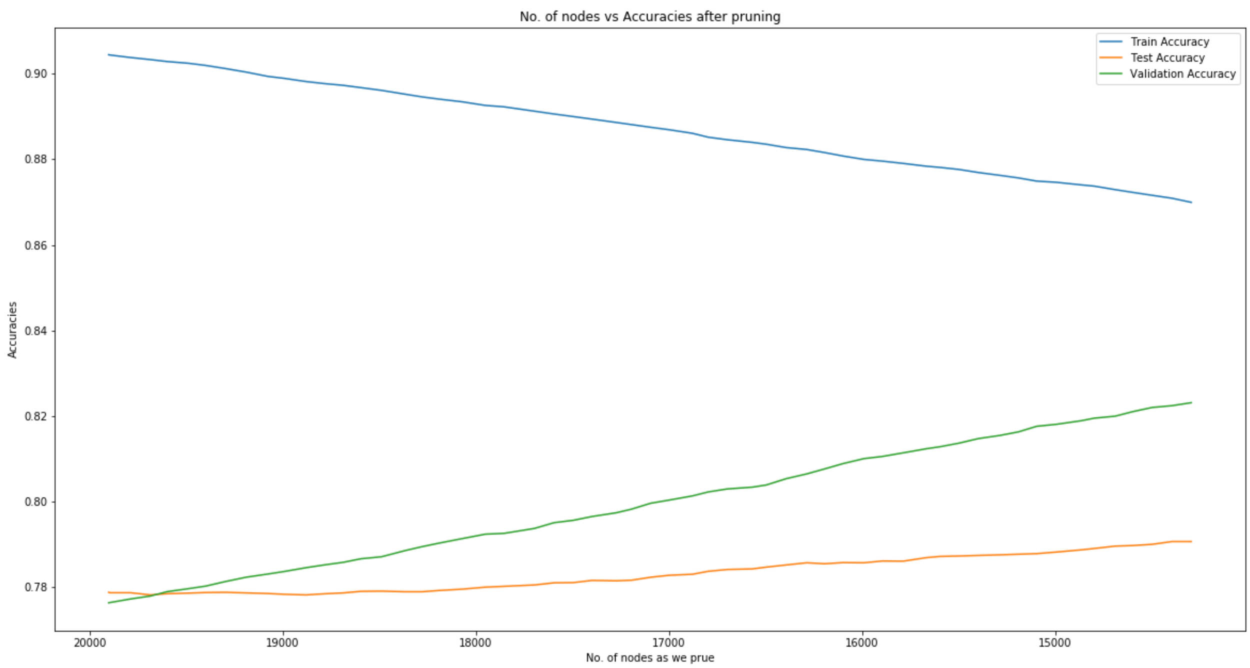

Figure 3: Accuracy vs no. of nodes

2.3.2 Second Approach

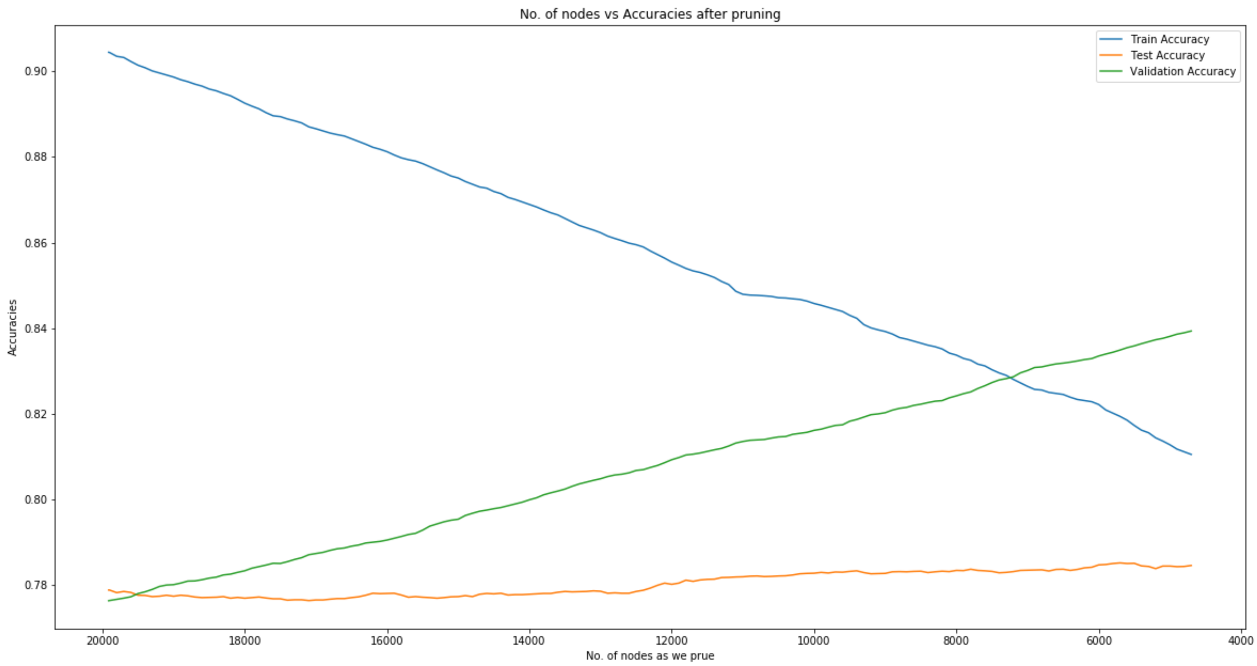

Figure 4: Accuracy vs no. of nodes

2.4 Comments

We can observe from the plot that the as we prune the nodes training accuracy decreases but validation accuracy increases.

Test accuracy is also increased slightly after pruning of decision tree.

It is happening because the original tree was more biased towards the training set. Now we prune the tree on validation set. So generalization accuracy increases.

Advantage of pruning is that the now pruned tree is more robust to new data set than unpruned tree.

Also testing time of test set is also decreased because now we have to test on shorter tree.

3 Random Forests

3.1 Introduction

Random Forests are extensions are decision trees, where we grow multiple decision trees in parallel on bootstrapped samples constructed from the original training data

We have to find the set of parameters by grid search which minimizes the oob error.

3.2 Parameters

Optimal parameters

n_estimators = 350

max_features = 0.3

min_samples_split = 10

Out-of-bag accuracy = 81.0609

Train accuracy = 88.1213

Test accuracy = 80.8863

Validation accuracy = 80.6183

3.3 Comments

Training accuracy using random forest is better than decision tree after pruning. Because in decision tree we pruned the sub-tree on validation set

Validation accuracy is lesser than decision tree after pruning because we prune the decision tree on validation set, so tree becomes more accurate on validation set

Test accuracy is somewhat similar in both of the case.

4 Parameter Sensitivity Analysis

4.1 Fix n_estimator and max_features

n_estimator = 350

max_features = 0.3

min_sample_split	Train accuracy	Test accuracy	Validation accuracy
2	91.2722	79.8897	79.8164
4	90.4408	80.3857	80.1966
6	89.4688	80.5897	80.3959
8	88.6838	80.738	80.5257
10	88.1213	80.738	80.6184

Table 1: Accuracy vs varying min_sample_split

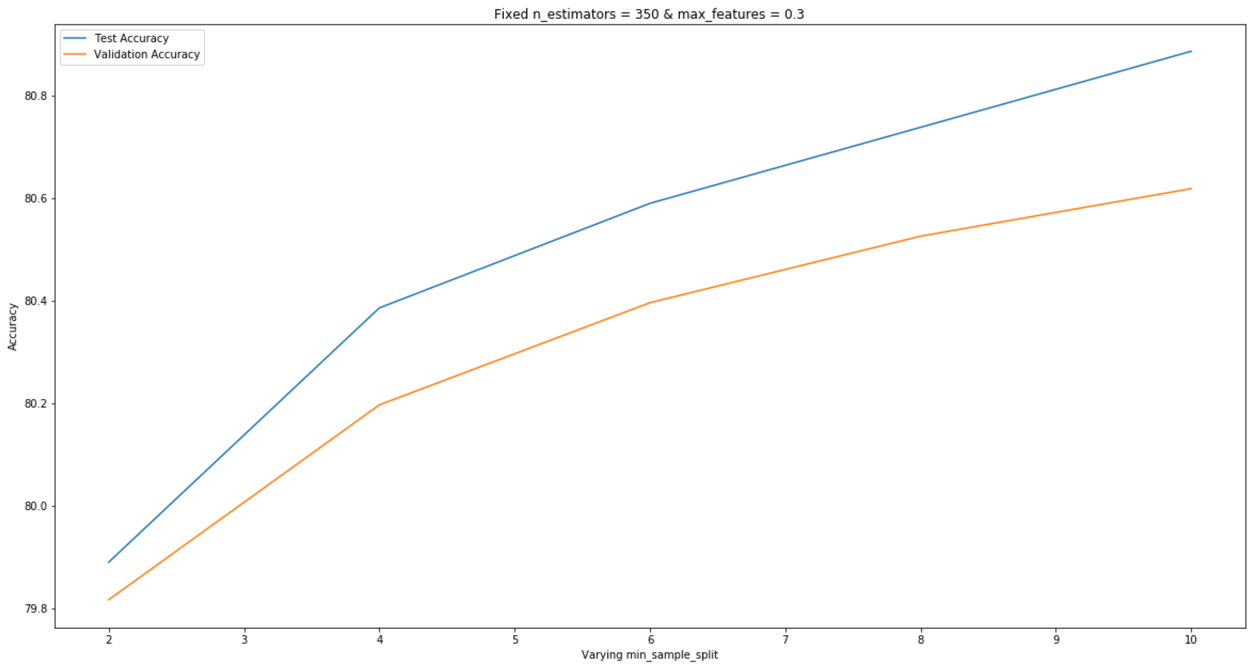

Figure 5: Accuracy on varying min_sample_split

4.2 Comments

By varying min_sample_split by fixing other two, test accuracy's max-min is 0.8483. Validation accuracy's max-min is 0.802.
So accuracy is more sensitive to max_features parameter

Also we can observe that increasing min_sample_split(min sample require to split node) result in increase in accuracy of test/validation set but decrease in training accuracy. Because it generalizes better for test/validation set.

4.3 Fix n_estimator and min_sample_split

n_estimator = 350

min_sample_split = 10

max_features	**Train accuracy**	**Test accuracy**	**Validation accuracy**
0.1	87.3797	80.8539	80.7296
0.3	88.1214	80.8864	80.6184
0.5	88.4397	80.7427	80.6462
0.7	88.6159	80.6407	80.4701
0.9	88.721	80.4691	80.3681

Table 2: Accuracy vs varying max_features

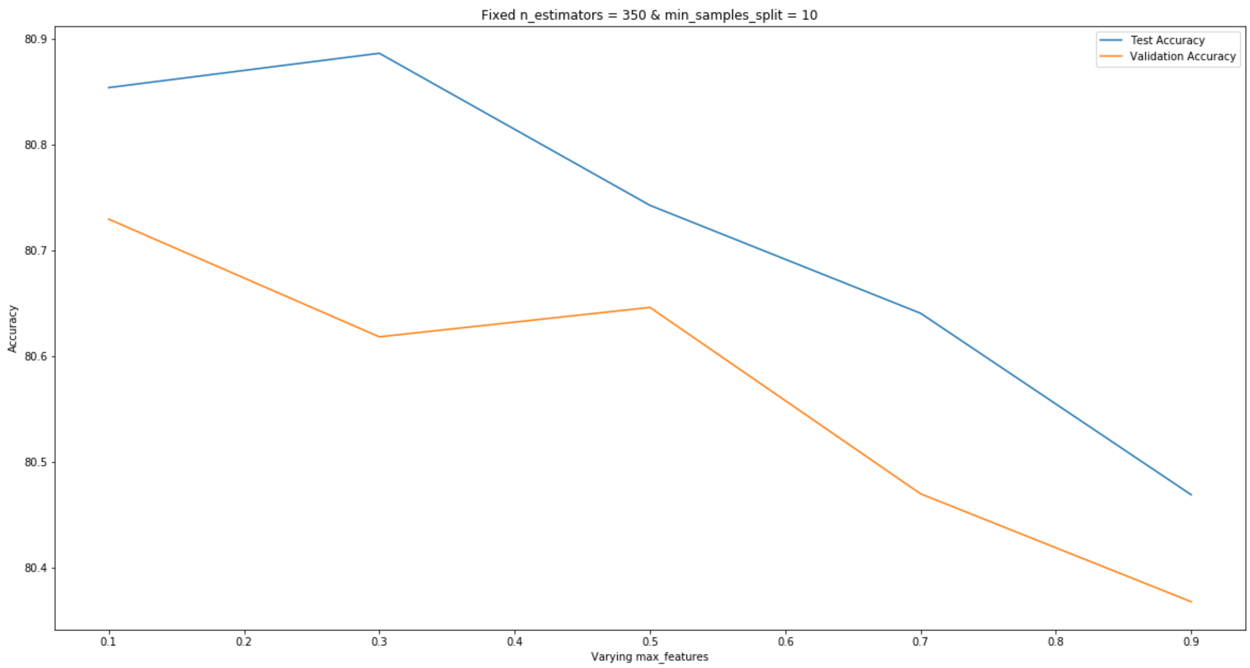

Figure 6: Accuracy on varying max_features

4.4 Comments

By varying max_features by fixing other two, test accuracy's max-min is 0.4173.
Validation accuracy's max-min is 0.3615
So accuracy is moderately sensitive to max_features parameter.
Also we can observe that increasing max_features result in decreases in accuracy.
Because by taking more features, our model is more biased towards training set.
We can see from the table that at max_features=0.9, training accuracy is larger but test and validation is smaller.

4.5 Fix max_features and min_sample_split

max_features = 0.3

min_sample_split = 10

n_estimator	Train accuracy	Test accuracy	Validation accuracy
50	87.9993	80.5804	80.4561
150	88.0889	80.8076	80.5952
250	88.1245	80.738	80.4979
350	88.1214	80.8864	80.6184
450	88.1029	80.8354	80.7018

Table 3: Accuracy vs varying n_estimator

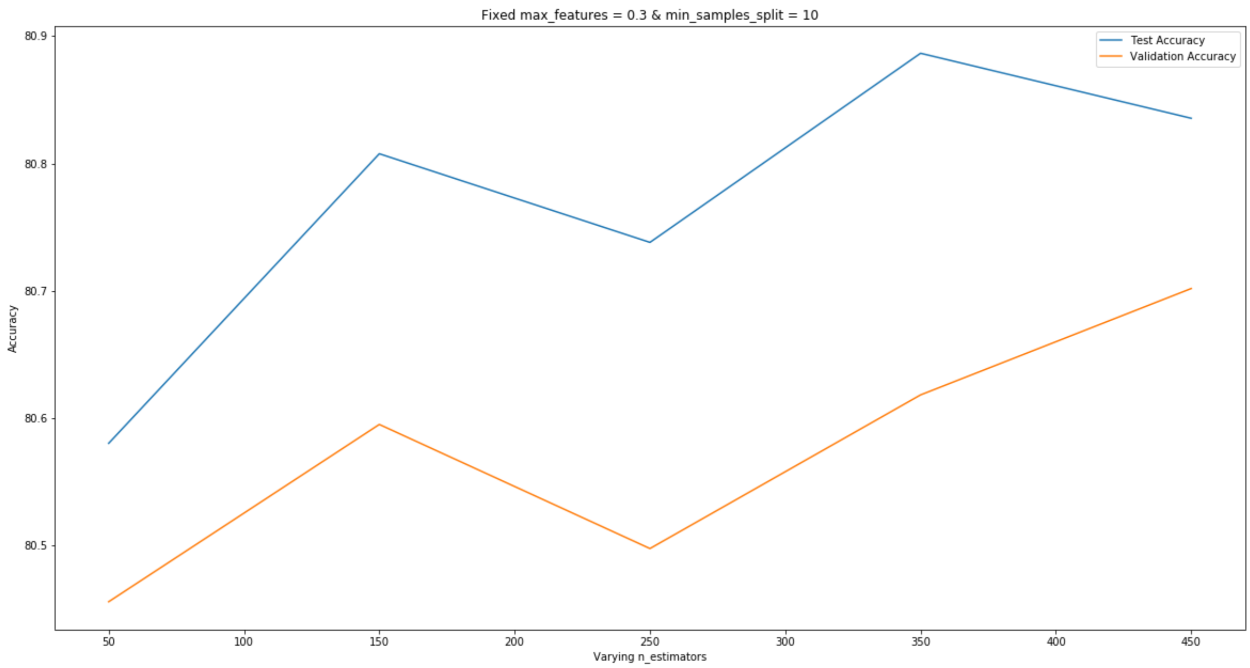

Figure 7: Accuracy on varying n_estimator

4.6 Comments

By varying n_estimator by fixing other two, test accuracy's max-min is 0.3064.

Validation accuracy's max-min is 0.2457

So accuracy is less sensitive to n_estimator parameter

Also we can observe that by increasing n_estimator parameter, broader trend in accuracy also increases

Overall we can see that accuracy is more sensitive to max_features parameter and less sensitive to n_estimators parameter.