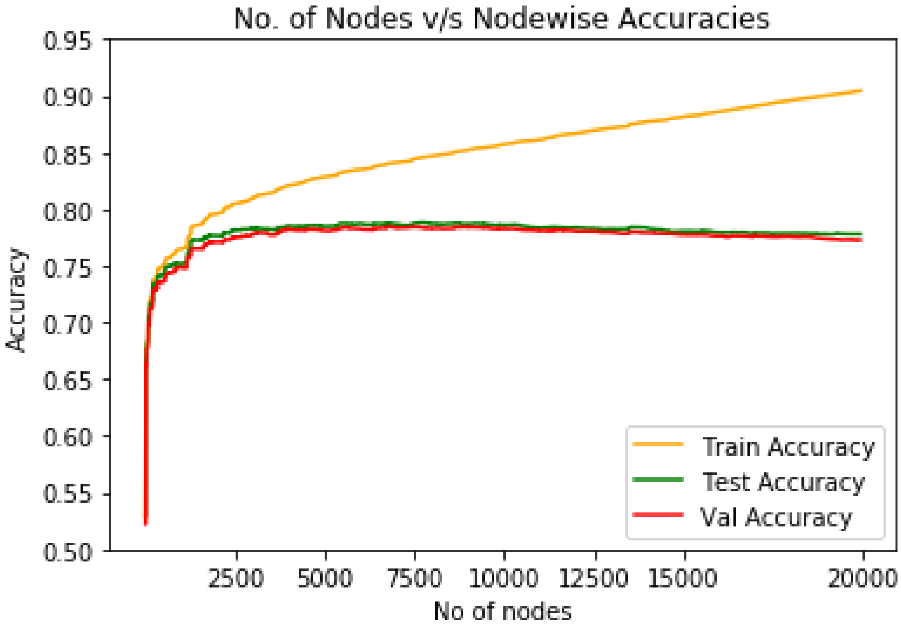
Que_1(A)

As the number of nodes increases in the decision tree the accuracies for the training, test and validation data increases, but sooner the accuracies for the test and validation data decreases and the accuracy for the training data increases, it is due to the overfitting of the tree. An overfitted model does exceptionally well on the training data as the model is trained on the same data but may not perform well during predictions on some unseen data.

```
Training Set Accuracy   : 0.904377791170244
Testing Set Accuracy    : 0.7780353252051365
Validation Set Accuracy : 0.7727146300760245
```

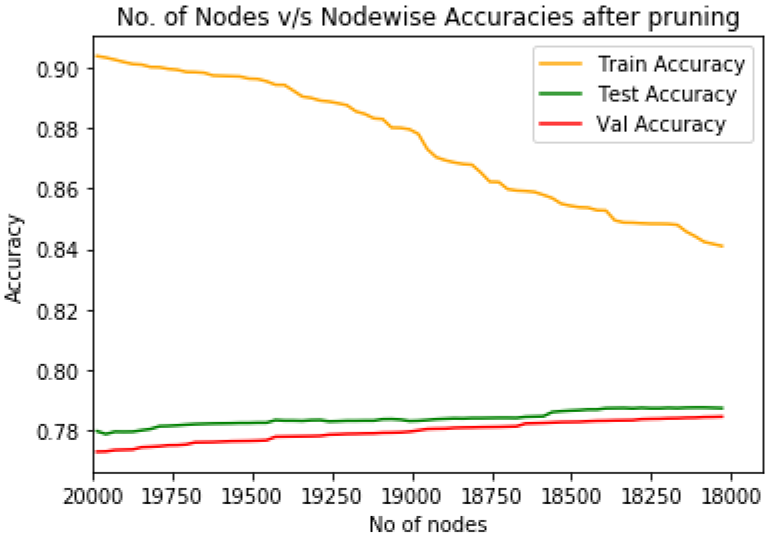


Que_1(B):

Followings are the accuracies after performing Pruning on the tree:

a. Training Accuracy after Pruning : 0.8400939533015005
b. Test Accuracy after Pruning : 0.7874461082008252
c. Validation Accuracy after Pruning : 0.7846282217689597

Here, we can not notice the accuracy on the training data has been reduced while the accuracy on the validation and testing data has been increased.



Que_1 (C)

1. Best parameters observed after GridSearchCV:
   n_estimators: 450, **max_features**: 0.3, **min_samples_split**: 10, **oob_score:**True
2. RF Model Accuracies with default parameters:
   Default parameters:-
   n_estimators: 100, **max_features**: ‘auto’, **min_samples_split**: 2, **oob_score:**False
   Accuracies:-
   a. Accuracy on training set : 0.9127068749710259
   b. Accuracy on testing set : 0.7970886838811367
   c. Accuracy on validation set : 0.7964954570739848

3. RF Model Accuracies with best parameters with oob_score as **True**:
   Best parameters:-
   n_estimators: 250, **max_features**: 0.1, **min_samples_split**: 10, **oob_score:**True
   Accuracies:-
   a. Accuracy on training set : 0.8733175714307789
   b. Accuracy on testing set : 0.8077047888368644
   c. Accuracy on validation set : 0.8060448729834971
   d. Out of bag (oob) accuracy : 0.809667300233338

Comparison of accuracies with the ones recieved in the part (b):

a. Train Accuracy after Pruning : 0.8400939533015005
b. Test Accuracy after Pruning : 0.7874461082008252
c. Validation Accuracy after Pruning : 0.7846282217689597

Here, we can notice that Random forest with tuned parameters has performed better than the decision tree after pruning.

Conclusion: We can notice with default parameters accuracy on training is more than 91% but accuracy on test and validation set is not that good.
With the best parameters, accuracy has slightly been decreased on the **training** data set and increased on the **test** and the **validation** set. When the oob_score flag is set to true then we get oob_score also which is mentioned above.

Que_1(d)

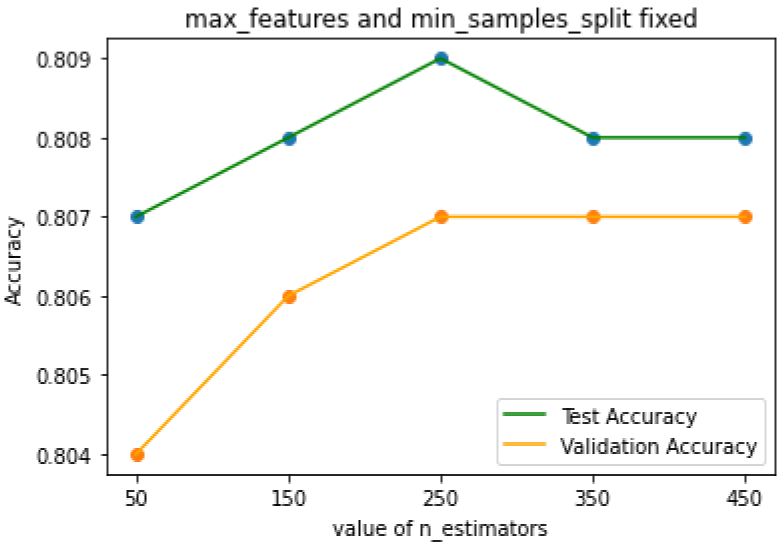
1. max_features and min_samples_split are fixed n_estimators varies for [50,150,250,350,450]

    a. **Test Data Accuracy** : [0.807, 0.808, 0.809, 0.808, 0.808]
    b. **Validation Data Accuracy** : [0.804, 0.806, 0.807, 0.807, 0.807]

    The difference between the min and max accuracies:

    a. Test Data accuracy : 0.809 - 0.807 = 0.002
    b. Validation Data accuracy : 0.807 - 0.804 = 0.003

Here we can notice that there are negligible differences between the accuracies when the value of n_estimators changes. Hence it's **not a sensitive parameter**.



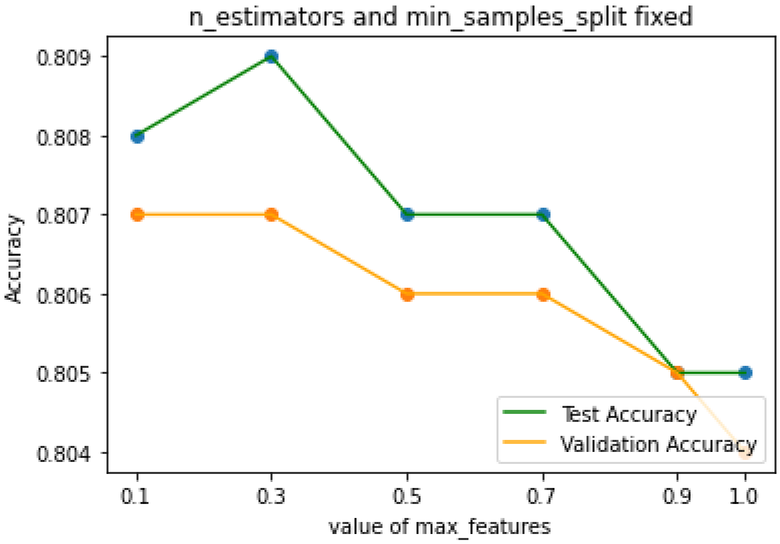
2. n_estimators and min_samples_split are fixed max_features varies for [0.1,0.3,0.5,0.7,0.9,1.0]

   Test Data Accuracy : [0.808, 0.809, 0.807, 0.807, 0.805, 0.805]
   Validation Data Accuracy : [0.807, 0.807, 0.806, 0.806, 0.805, 0.804]

   The difference between the min and max accuracies:

   c. Test Data accuracy : 0.809 - 0.805 = 0.004
   d. Validation Data accuracy : 0.807 - 0.804 = 0.003

Here we can notice that there are negligible differences between the accuracies when the value of max_features changes. Hence it's **not a sensitive parameter**.



We notice here that when the number of features are increasing then the accuracy getting decreased, hence features selection plays an important role.

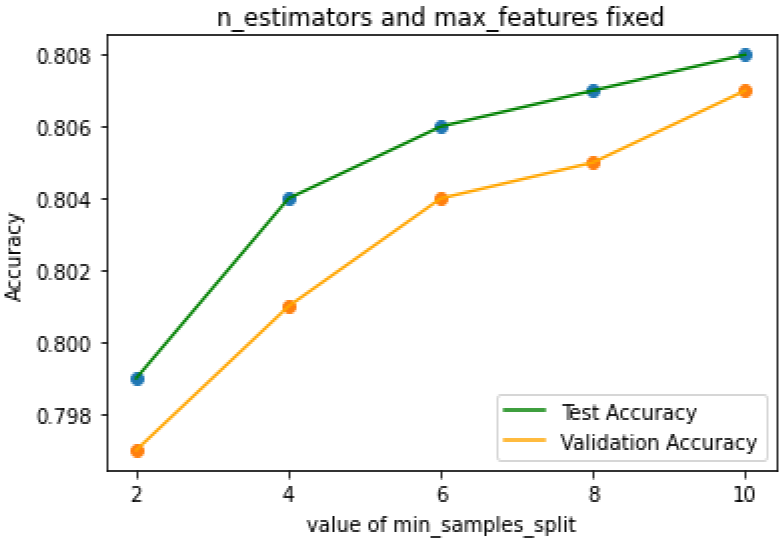
3. n_estimators and max_features are fixed min_samples_split varies for [2,4,6,8,10]

   Test Data Accuracy : [0.799, 0.804, 0.806, 0.807, 0.808]
   Validation Data Accuracy : [0.797, 0.801, 0.804, 0.805, 0.807]

   The difference between the min and max accuracies:

   e. Test Data accuracy : 0.808 - 0.799 = 0.009
   f. Validation Data accuracy : 0.807 - 0.797 = 0.01

Here we can notice that there are bigger differences between the accuracies in comparison to the previous ones when the value of max_features changes. Hence it's **a sensitive parameter**. The same can be noticed from the graphs also.


n_estimators and max_features fixed
Accuracy
value of min_samples_split
Test Accuracy
Validation Accuracy