ML Assignment-4
Report

1. Bank loan prediction using Decision Tree and Random Forest:

a. The problem is solved by using both Decision tree and Random forest. We can solve it by using criterion as ‘gini’ or ‘cross entropy’. I used ‘gini’ as criterion.I tried different parameters like max_depth, min_samples_leaf, min_samples_split for parameter tuning and took the best parameters by analysing and observing the accuracy,f1_score plots.
Among these two techniques, Random forest is giving better results than Decision tree. This is because Random forest is an Ensemble technique i.e., it combines the different trees by taking different parameters and gives best results.

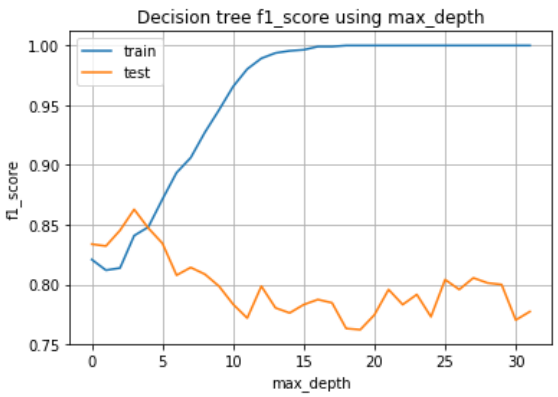
Decision tree has high variance, we can use Bagging technique to reduce the variance of this. On the other hand, Random forest will create the several decision trees and combines them and gives results.
Due to this method, Random forest give better results than Decision tree.

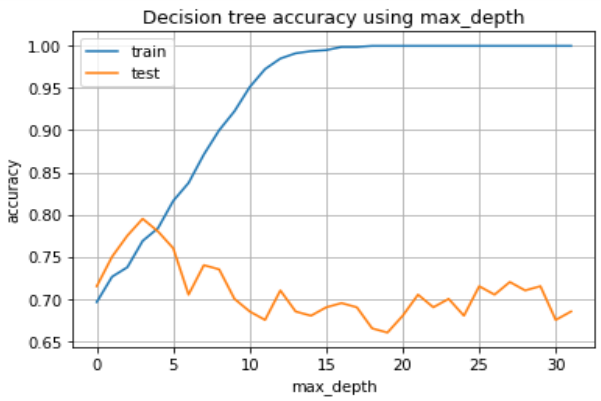
b,c. best parameters used for Decision tree and Random forest are

	criterion	max_depth	min_samples_leaf	Other parameters
Decision Tree	gini	5	0.1	min_samples_splits=0.2
Random Forest	gini	8	3	max_features=11 ; n_estimators=11

And I used some default parameters which inbuilt defaults for this in function.

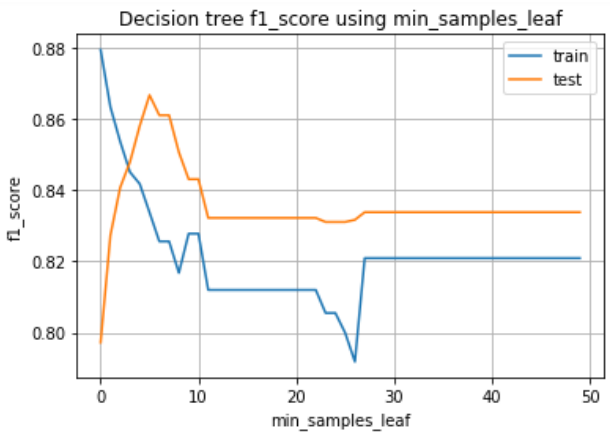
Plots for Decision tree:
I plotted the graph for accuracy, f1_score for all these parameters and selected the best parameter.

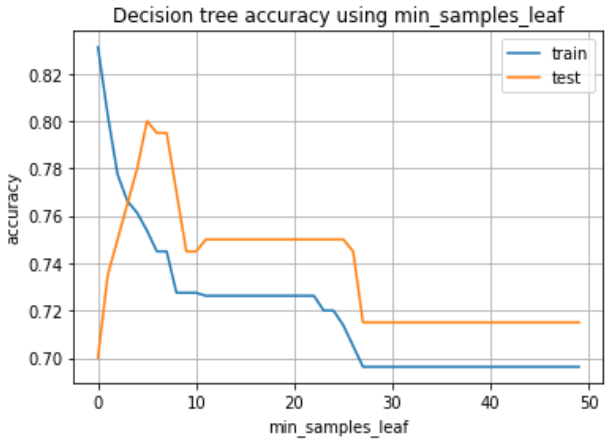




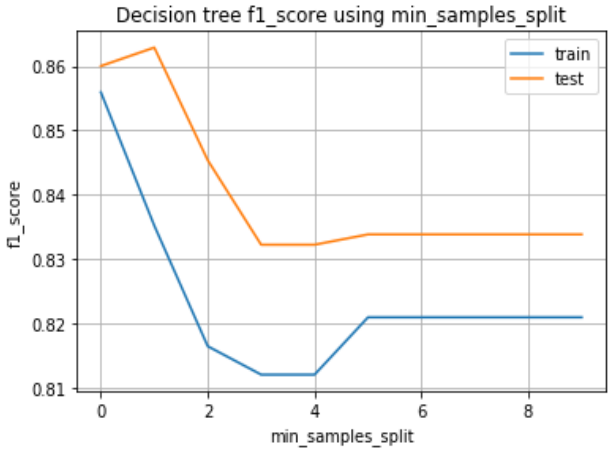
Here I tunned for max_depth parameter and plotted f1_score and accuracy and I selected value of max_depth=5.

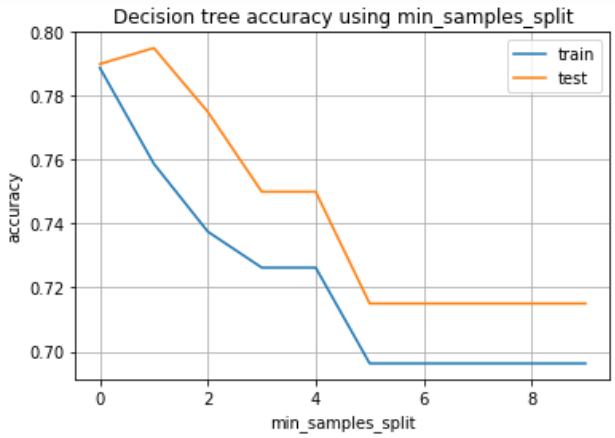
min_samples_leaf plots: here I selected value of min_samples_leaf = 10 the number samples considered in leaf nodes.



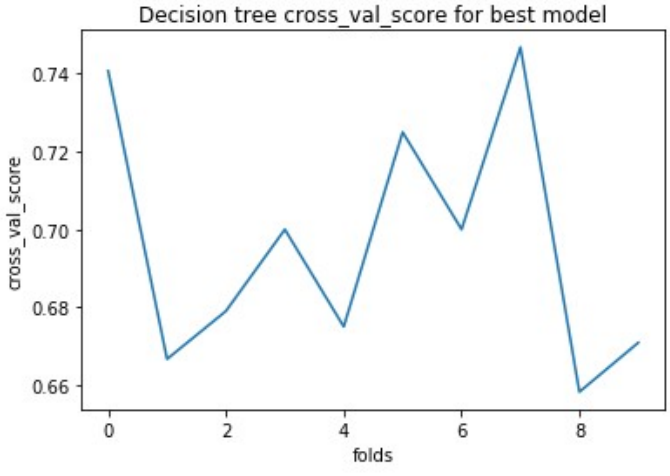


plots for min_samples_split: here I selected the value of min_samples_split=0.2 it is perfoming better.



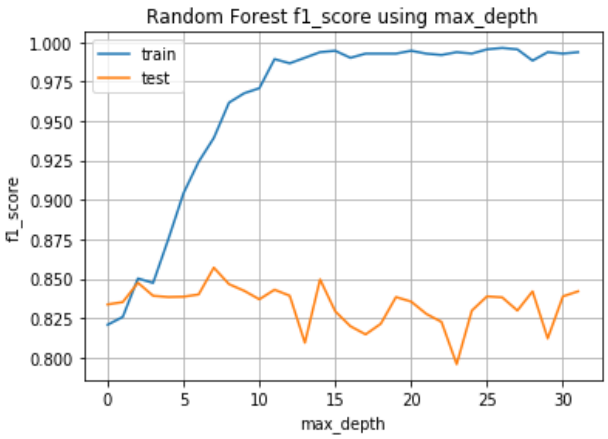


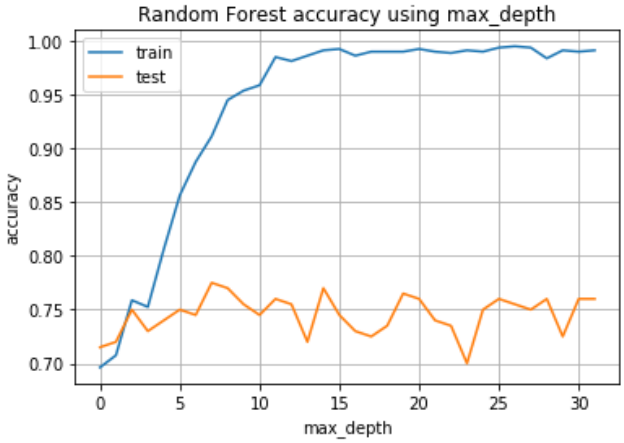
after selecting the best hyperparameters I tried decision tree with them and used cross validation of 10 and plotted their accuracies.



Plots for Random forest:

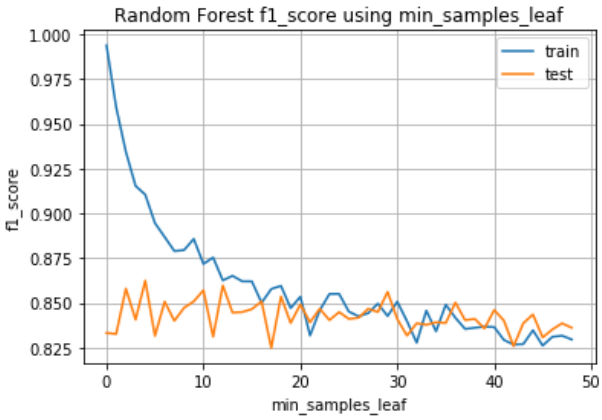
I plotted the graph for accuracy, f1_score for all these parameters and selected the best parameter.

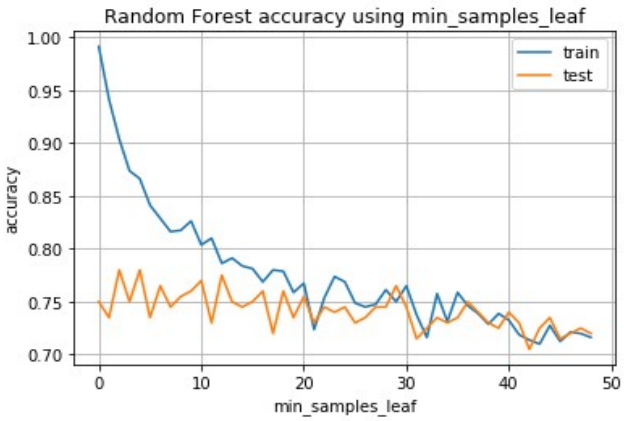




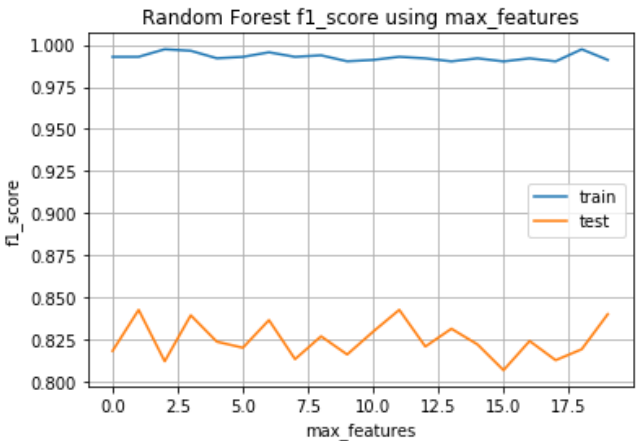
Here I tunned for max_depth parameter and plotted f1_score and accuracy and I selected value of max_depth=8.

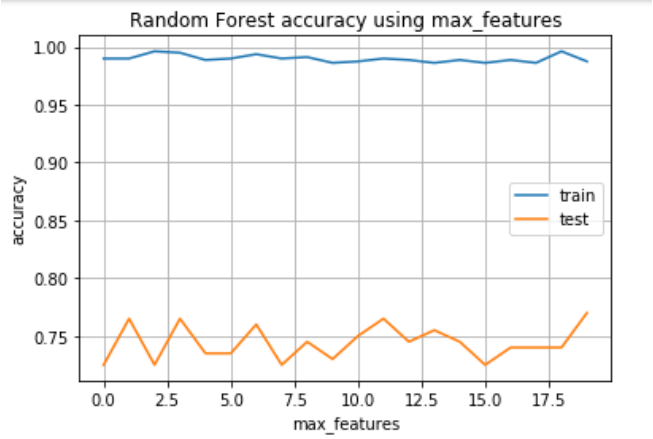
min_samples_leaf plots: here I selected value of min_samples_leaf = 3



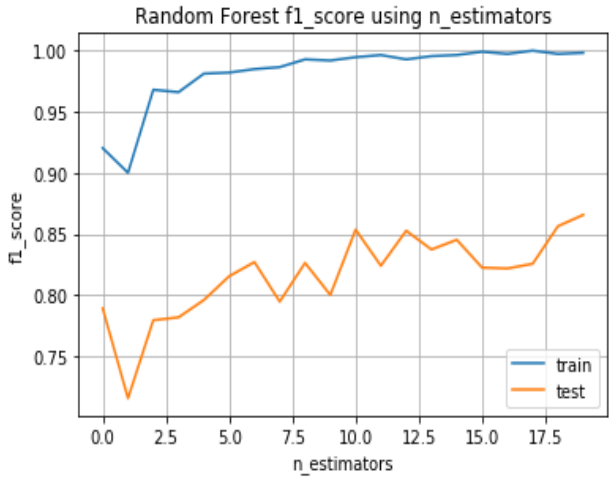


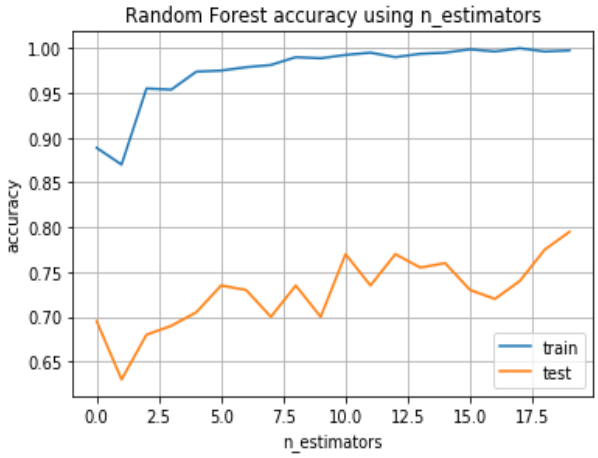
max_features plots: here I selected value of max_features = 11



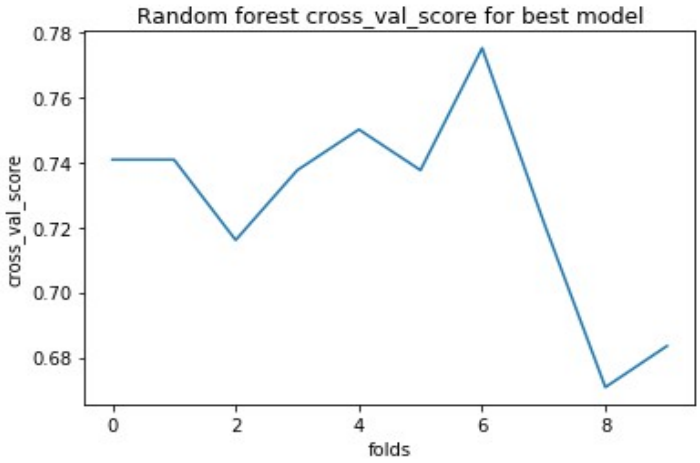


n_estimator plots: here I selected value of n_estimator = 11





after selecting the best hyperparameters I tried decision tree with them and used cross validation of 10 and plotted their accuracies.



Initially with default hyparameters both models are overfitting and giving a normal accuracy for test data,but after picking tuned parameter my models are not overfitting nor underfitting and giving accuracies of 0.81 and 0.84 respectively.

d.

I perfomed cross_val_score validtion and Kfold validation for bestmodel and used fold of size=5 and verified my results.

The variance of Decision tree is always more than the random forest because decision trees are more sensitive to training data, a slight change in data will impact the model resulting in high variance when compared to random forest.

On the other hand, Random forest has low variance , because it uses multiple decision tree to give its results.
So Decision tree has high variance than the Random forest.

```
print(np.var(dec_score),np.var(rand_score))
np.var(dec_score)>np.var(rand_score)
```

```
0.002290624999999999 0.0011249999999999997
```

```
True
```

e. after tuning hyperparameters, I saved the best models using sklearn's joblib function and predict the test data.