LAB REPORT-3

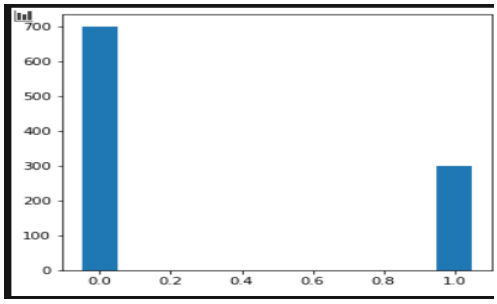
MAYANK RAJ (B19CSE053)

BTECH CSE

PROBLEM 1

1) PREPROCESSING

1. Getting the distribution on bar plot.



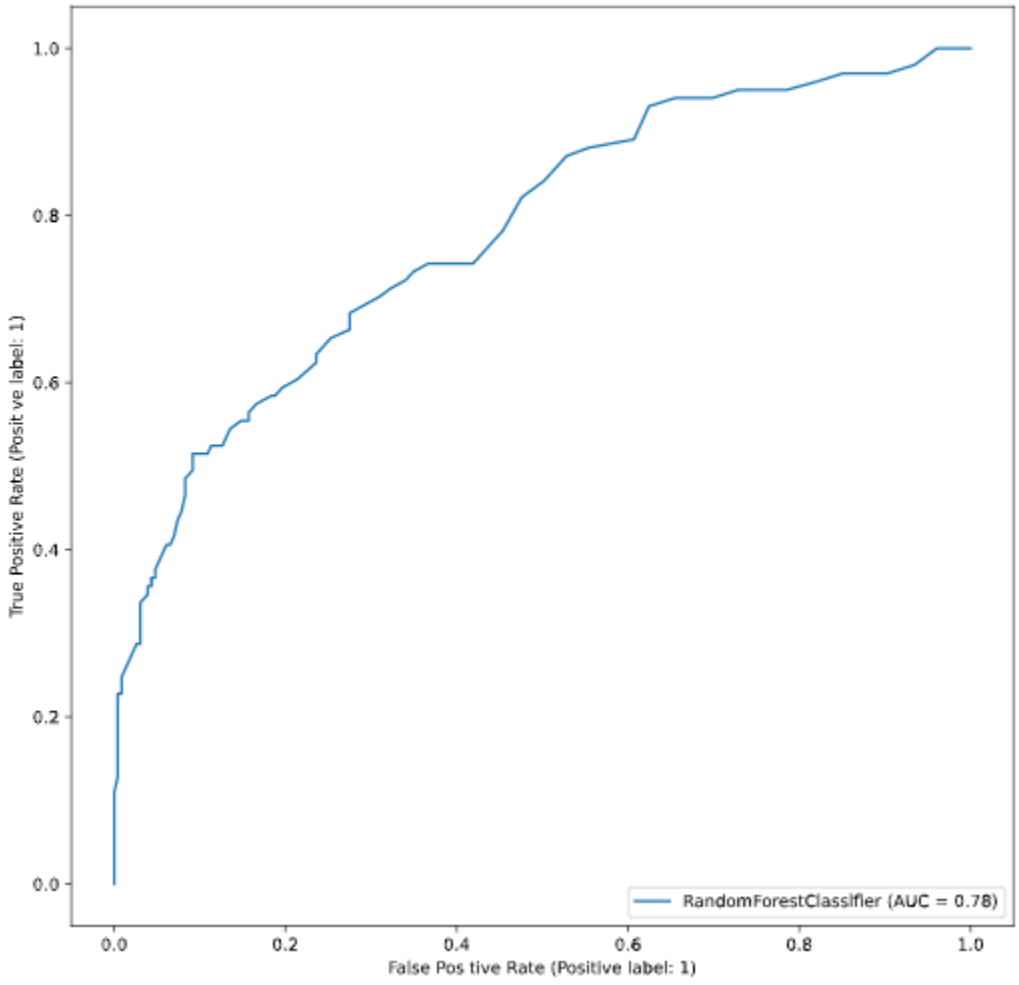
2. NaN values were replaced with mean of a particular column.
3. Distribution for each feature was plotted against dependent variables.

2) Trained against Random Forest classifier

- Data was splitted into training and testing sample and then the model was fit on Random Forest

3)Performed Cross Validation

- Stratified k-fold cross validation was done and Grid-search cv was also done and best score was obtained as 0.815
- Roc Curve



4)Least Impact feature was calculated

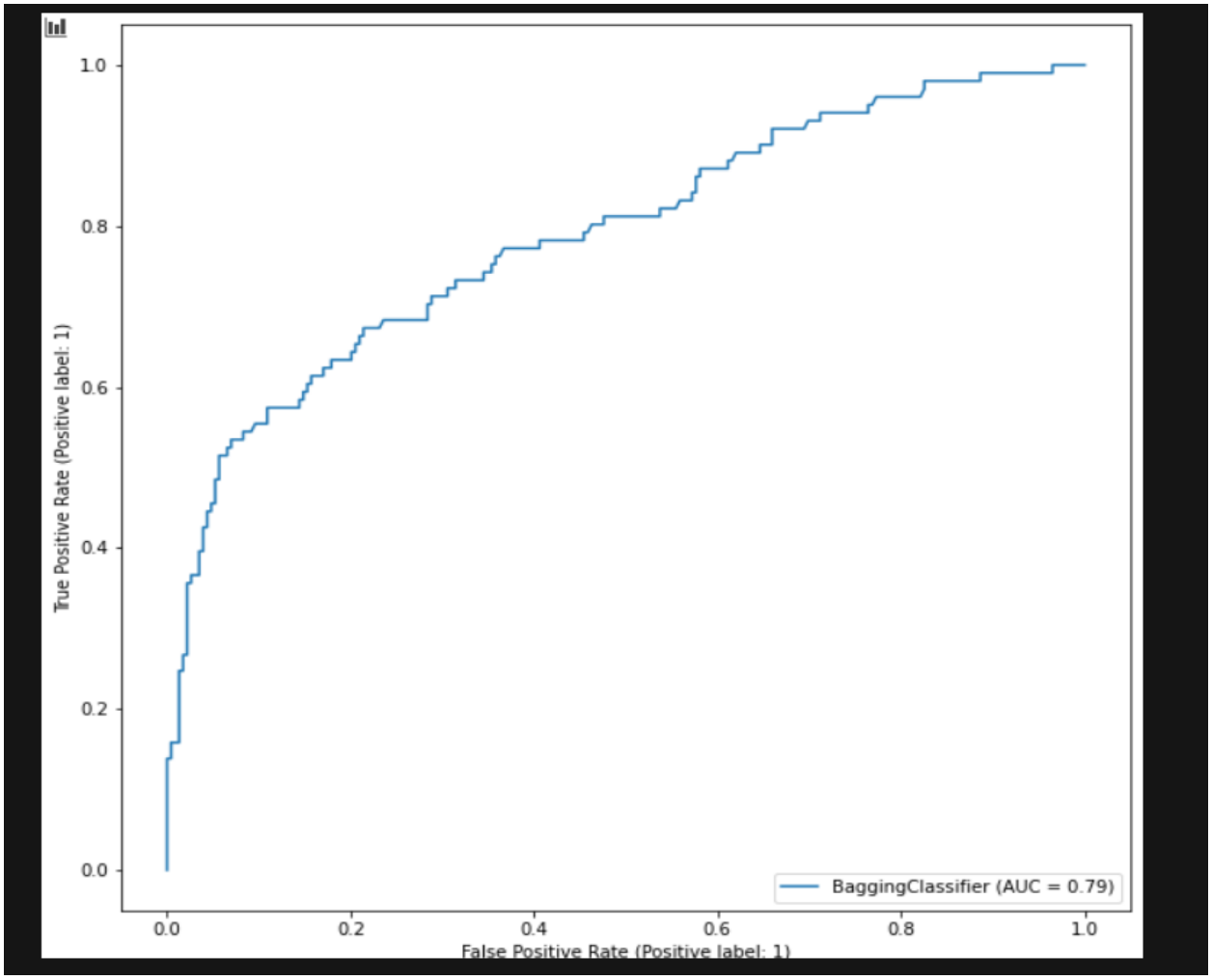
- Least impact feature was calculated, “Dependent” feature was obtained as least

impact feature.

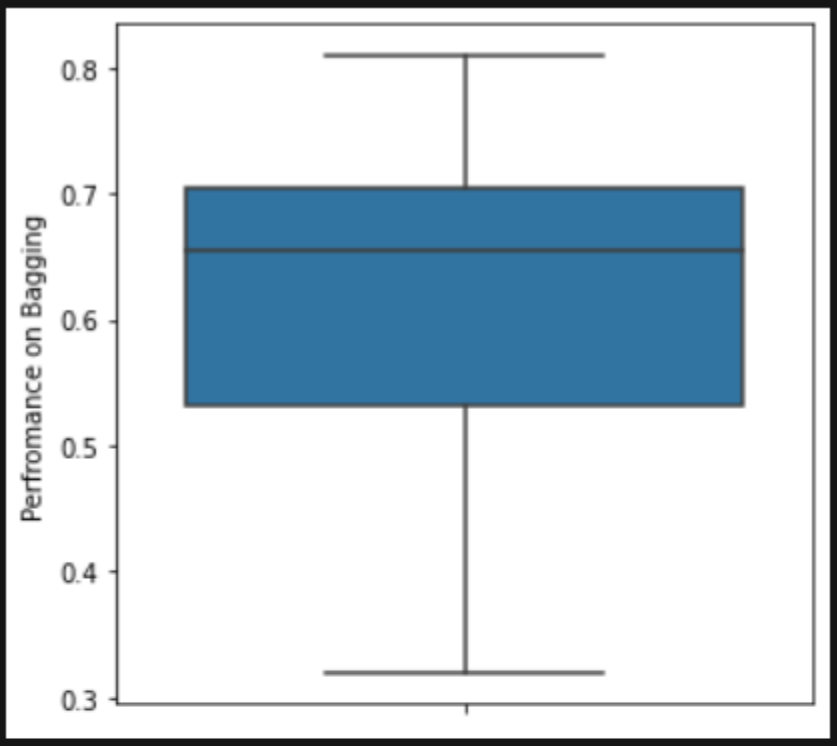
2)PROBLEM 2

1. Bagging classifier using decision tree as base classifier

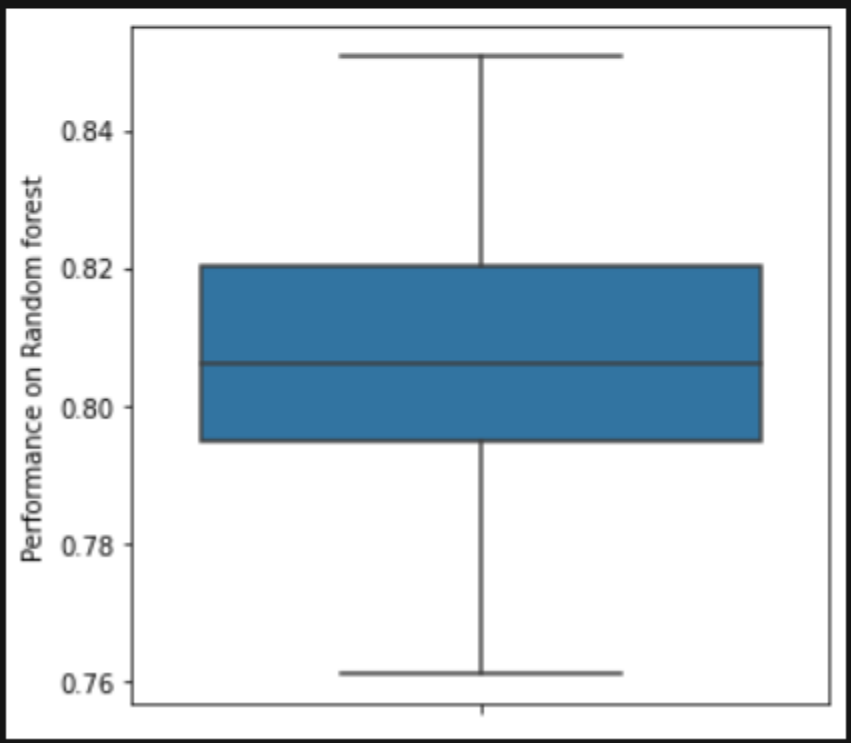
- Testing accuracy on same data set was obtained as 0.736,
- 5 fold cross validation was done and a list of cross validation scores for different classifiers were obtained.
- ROC



2. Comparison was done using box-plot




```
Text(0, 0.5, 'Performance on Random forest')
```



INFERENCE:

- As we can see in the box plot, the median of Random Forest is higher than the median of the box-plot of the decision tree with bagging ,hence in this case the boosting algorithm in Random Forests performs better than the bagging in decision tree with base classifier.