Group:

Rahul Mallu (50245594)

Likhith Kumar Miryala (50249280)

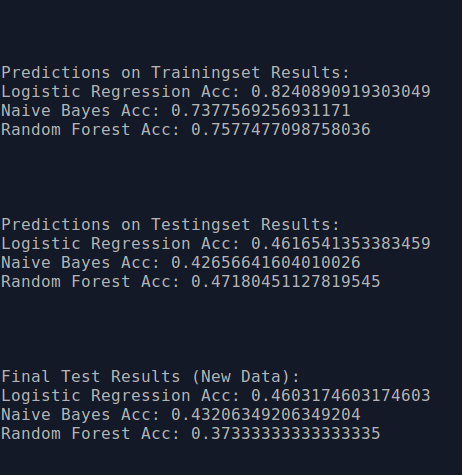
Environment: pyspark Virtual Machine

Command to run our program : spark-submit *filename.py*

Command 1) spark-submit titanic.py

Command 2) spark-submit classific.py

Output:



The accuracies obtained for train, test and random data are as above.

From the above results, we understand that random forest is robust and versatile but for high dimensional sparse data, logistic regression model is preferable

Since there is discrepancy in the accuracy between the training, test and random data and as we have used cross validation we can say that the statistical properties of test, train and random data are different