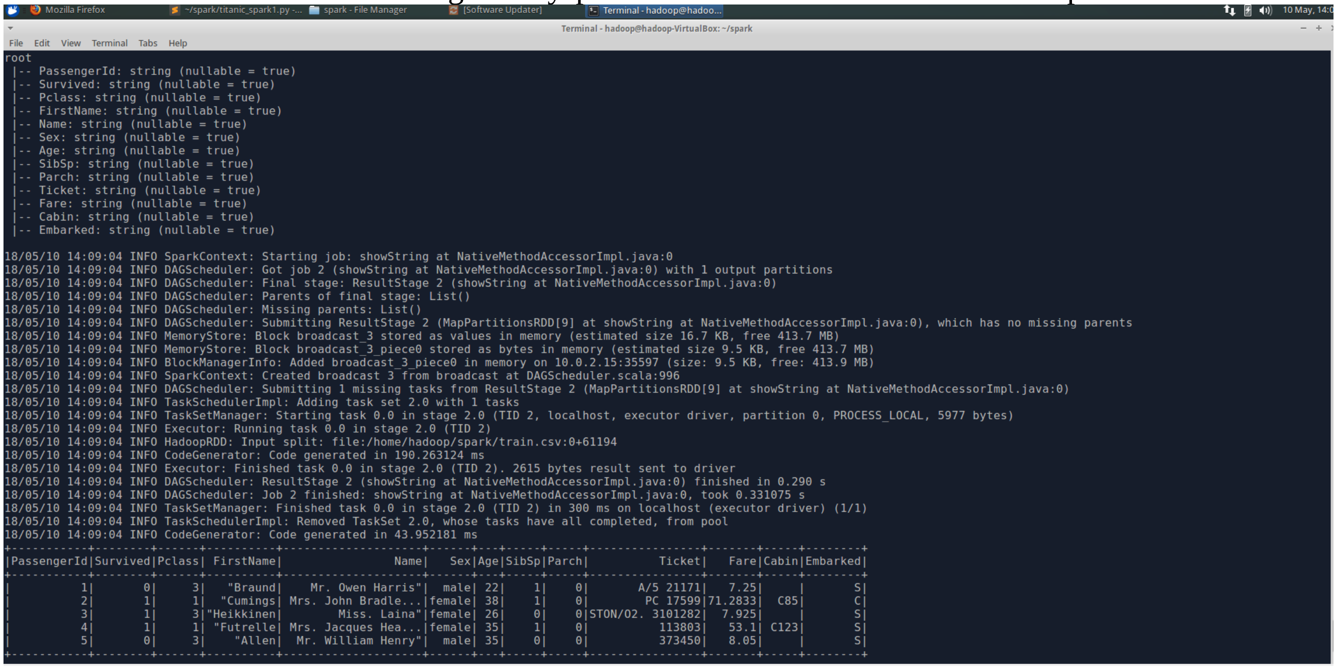
Sai Kiran Putta
Hari Krishna Rangineeni

PART – I : Titanic DataSet Exploration in PySpark
PART – II : Text Classification using PySpark

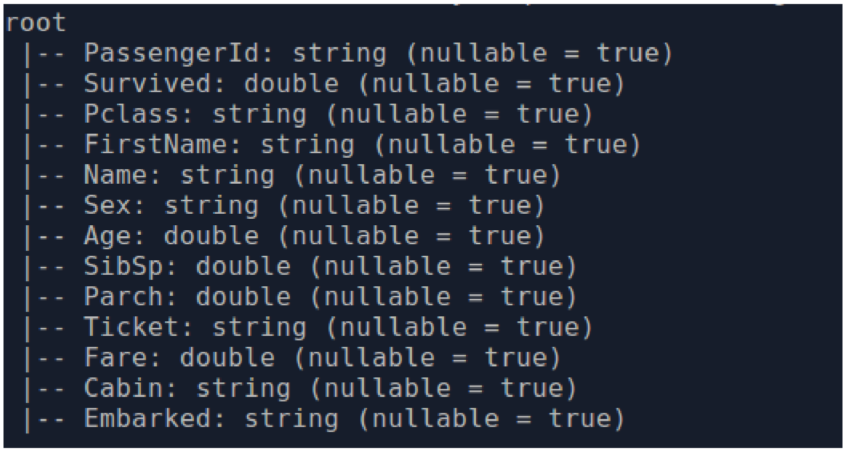
PART – I :

- Titanic Dataset after loading into Pyspark has a schema like follows with top 5 rows as below:


Mozilla Firefox
~/spark/titanic_spark1.py -...
spark - File Manager
[Software Updater]
Terminal - hadoop@hadoo...
Terminal - hadoop@hadoop-VirtualBox: ~/spark
File
Edit
View
Terminal
Tabs
Help

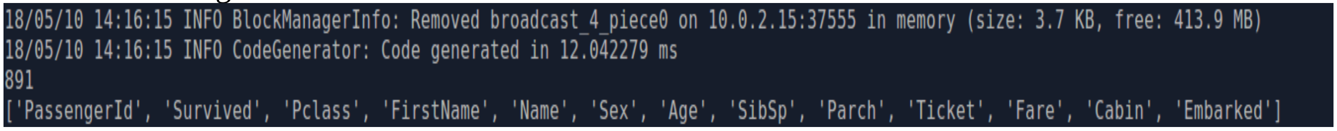
```
root
 |-- PassengerId: string (nullable = true)
 |-- Survived: string (nullable = true)
 |-- Pclass: string (nullable = true)
 |-- FirstName: string (nullable = true)
 |-- Name: string (nullable = true)
 |-- Sex: string (nullable = true)
 |-- Age: string (nullable = true)
 |-- SibSp: string (nullable = true)
 |-- Parch: string (nullable = true)
 |-- Ticket: string (nullable = true)
 |-- Fare: string (nullable = true)
 |-- Cabin: string (nullable = true)
 |-- Embarked: string (nullable = true)

18/05/10 14:09:04 INFO SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:0
18/05/10 14:09:04 INFO DAGScheduler: Got job 2 (showString at NativeMethodAccessorImpl.java:0) with 1 output partitions
18/05/10 14:09:04 INFO DAGScheduler: Final stage: ResultStage 2 (showString at NativeMethodAccessorImpl.java:0)
18/05/10 14:09:04 INFO DAGScheduler: Parents of final stage: List()
18/05/10 14:09:04 INFO DAGScheduler: Missing parents: List()
18/05/10 14:09:04 INFO DAGScheduler: Submitting ResultStage 2 (MapPartitionsRDD[9] at showString at NativeMethodAccessorImpl.java:0), which has no missing parents
18/05/10 14:09:04 INFO MemoryStore: Block broadcast_3 stored as values in memory (estimated size 16.7 KB, free 413.7 MB)
18/05/10 14:09:04 INFO MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size 9.5 KB, free 413.7 MB)
18/05/10 14:09:04 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on 10.0.2.15:35597 (size: 9.5 KB, free: 413.9 MB)
18/05/10 14:09:04 INFO SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:996
18/05/10 14:09:04 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 2 (MapPartitionsRDD[9] at showString at NativeMethodAccessorImpl.java:0)
18/05/10 14:09:04 INFO TaskSchedulerImpl: Adding task set 2.0 with 1 tasks
18/05/10 14:09:04 INFO TaskSetManager: Starting task 0.0 in stage 2.0 (TID 2, localhost, executor driver, partition 0, PROCESS_LOCAL, 5977 bytes)
18/05/10 14:09:04 INFO Executor: Running task 0.0 in stage 2.0 (TID 2)
18/05/10 14:09:04 INFO HadoopRDD: Input split: file:/home/hadoop/spark/train.csv:0+61194
18/05/10 14:09:04 INFO CodeGenerator: Code generated in 190.263124 ms
18/05/10 14:09:04 INFO Executor: Finished task 0.0 in stage 2.0 (TID 2). 2615 bytes result sent to driver
18/05/10 14:09:04 INFO DAGScheduler: ResultStage 2 (showString at NativeMethodAccessorImpl.java:0) finished in 0.290 s
18/05/10 14:09:04 INFO DAGScheduler: Job 2 finished: showString at NativeMethodAccessorImpl.java:0, took 0.331075 s
18/05/10 14:09:04 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 300 ms on localhost (executor driver) (1/1)
18/05/10 14:09:04 INFO TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
18/05/10 14:09:04 INFO CodeGenerator: Code generated in 43.952181 ms
+-----------+--------+------+----------+-------------------+------+---+-----+-----+----------------+-------+-----+--------+
|PassengerId|Survived|Pclass| FirstName|               Name|   Sex|Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|
+-----------+--------+------+----------+-------------------+------+---+-----+-----+----------------+-------+-----+--------+
|          1|       0|     3|   "Braund|   Mr. Owen Harris"|  male| 22|    1|    0|       A/5 21171|   7.25|     |       S|
|          2|       1|     1|  "Cumings|Mrs. John Bradle...|female| 38|    1|    0|        PC 17599|71.2833|  C85|       C|
|          3|       1|     3|"Heikkinen|       Miss. Laina"|female| 26|    0|    0|STON/O2. 3101282|  7.925|     |       S|
|          4|       1|     1| "Futrelle|Mrs. Jacques Hea...|female| 35|    1|    0|          113803|   53.1| C123|       S|
|          5|       0|     3|    "Allen| Mr. William Henry"|  male| 35|    0|    0|          373450|   8.05|     |       S|
+-----------+--------+------+----------+-------------------+------+---+-----+-----+----------------+-------+-----+--------+
```

- We can observe that all the numeric features too are in String Format. After changing the datatype of features and checking the schema gives the following.



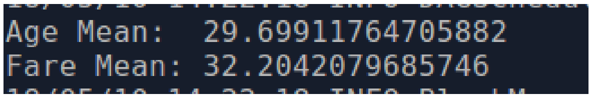
```
root
 |-- PassengerId: string (nullable = true)
 |-- Survived: double (nullable = true)
 |-- Pclass: string (nullable = true)
 |-- FirstName: string (nullable = true)
 |-- Name: string (nullable = true)
 |-- Sex: string (nullable = true)
 |-- Age: double (nullable = true)
 |-- SibSp: double (nullable = true)
 |-- Parch: double (nullable = true)
 |-- Ticket: string (nullable = true)
 |-- Fare: double (nullable = true)
 |-- Cabin: string (nullable = true)
 |-- Embarked: string (nullable = true)
```

- Following are total number of observations in the dataset and columns in the dataset as follow.



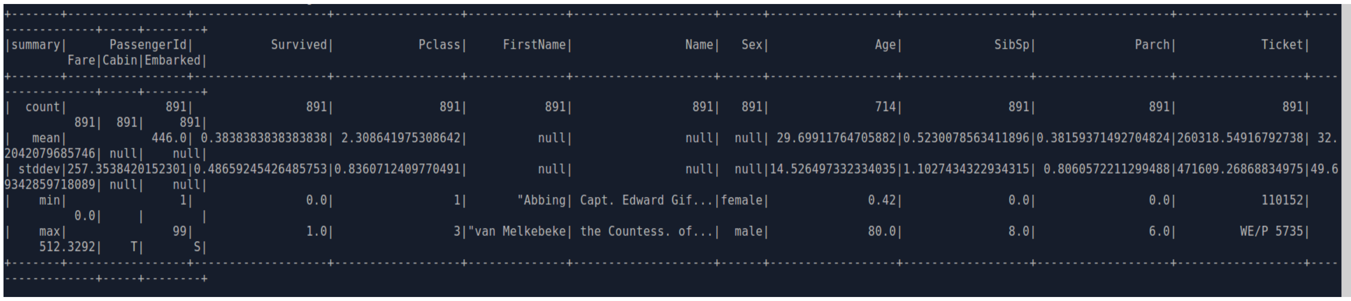
```
18/05/10 14:16:15 INFO BlockManagerInfo: Removed broadcast_4_piece0 on 10.0.2.15:37555 in memory (size: 3.7 KB, free: 413.9 MB)
18/05/10 14:16:15 INFO CodeGenerator: Code generated in 12.042279 ms
891
['PassengerId', 'Survived', 'Pclass', 'FirstName', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked']
```

- Features, Age and Fare have few observations missing. Imputing those features with the mean of respective columns. Following are the means.



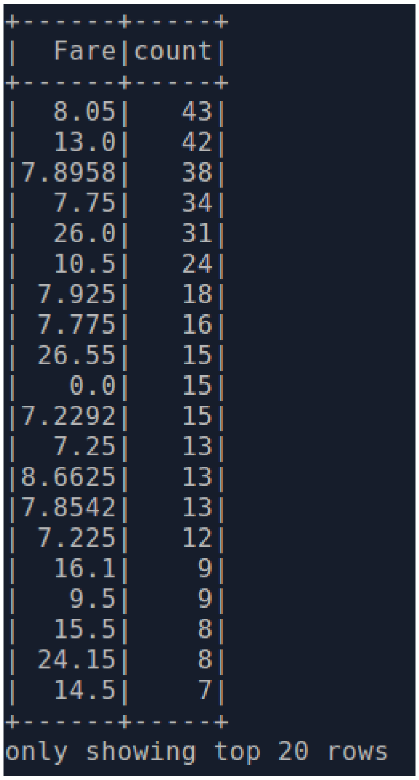
```
Age Mean:  29.69911764705882
Fare Mean: 32.2042079685746
```

- The statistics of each feature is as follows.



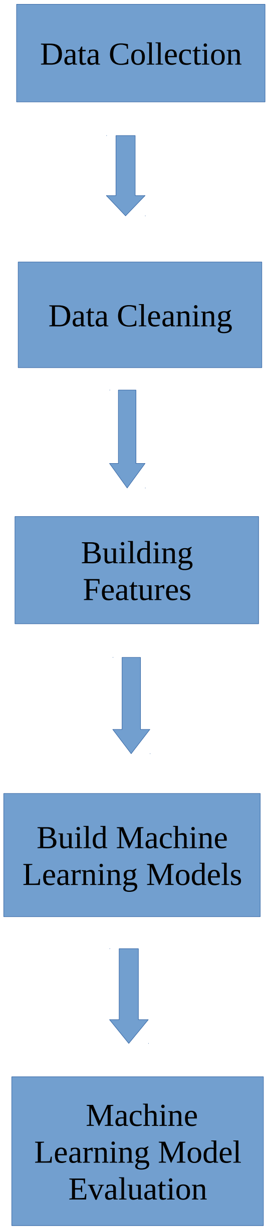
```
+-------+-----------------+-------------------+------------------+--------------+--------------------+------+------------------+------------------+-------------------+------------------+-----
------------+-----+--------+
|summary|      PassengerId|           Survived|            Pclass|     FirstName|                Name|   Sex|               Age|             SibSp|              Parch|            Ticket|
        Fare|Cabin|Embarked|
+-------+-----------------+-------------------+------------------+--------------+--------------------+------+------------------+------------------+-------------------+------------------+-----
------------+-----+--------+
|  count|              891|                891|               891|           891|                 891|   891|               714|               891|                891|               891|
         891|  891|     891|
|   mean|            446.0| 0.3838383838383838| 2.308641975308642|          null|                null|  null| 29.69911764705882|0.5230078563411896|0.38159371492704824|260318.54916792738| 32.
2042079685746| null|    null|
| stddev|257.3538420152301|0.48659245426485753|0.8360712409770491|          null|                null|  null|14.526497332334035|1.1027434322934315| 0.8060572211299488|471609.26868834975|49.6
9342859718089| null|    null|
|    min|                1|                0.0|                 1|       "Abbing| Capt. Edward Gif...|female|              0.42|               0.0|                0.0|            110152|
         0.0|     |        |
|    max|               99|                1.0|                 3|"van Melkebeke| the Countess. of...|  male|              80.0|               8.0|                6.0|         WE/P 5735|
    512.3292|    T|       S|
+-------+-----------------+-------------------+------------------+--------------+--------------------+------+------------------+------------------+-------------------+------------------+-----
------------+-----+--------+
```

- Grouping the dataset by Fare, following are top 20 counts.



```
+------+-----+
|  Fare|count|
+------+-----+
|  8.05|   43|
|  13.0|   42|
|7.8958|   38|
|  7.75|   34|
|  26.0|   31|
|  10.5|   24|
| 7.925|   18|
| 7.775|   16|
| 26.55|   15|
|   0.0|   15|
|7.2292|   15|
|  7.25|   13|
|8.6625|   13|
|7.8542|   13|
| 7.225|   12|
|  16.1|    9|
|   9.5|    9|
|  15.5|    8|
| 24.15|    8|
|  14.5|    7|
+------+-----+
only showing top 20 rows
```

PART – II:

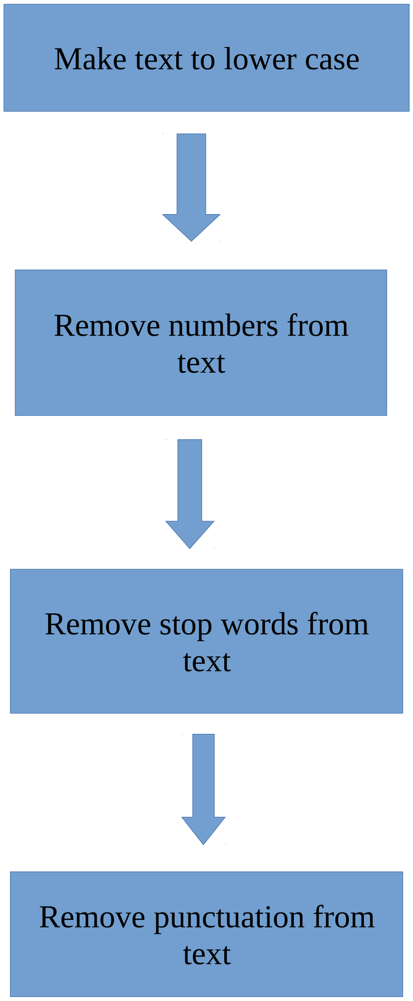

Data Collection
Data Cleaning
Building Features
Build Machine Learning Models
Machine Learning Model Evaluation

Data Collection :

Using the API, NYTimesArticle different posts on Business, Technology, Sports and Politics are collected. The data collection code can be found in the ipynb notebook attached as part of submission.

In total we collected around 4MB of data for all the different classes.

Data Cleaning:



Building Features:

We experimented with different way we can build features. Such as,

- Term Frequency
- TF – IDF
- Term Frequency with N-grams
- TF – IDF with N-grams

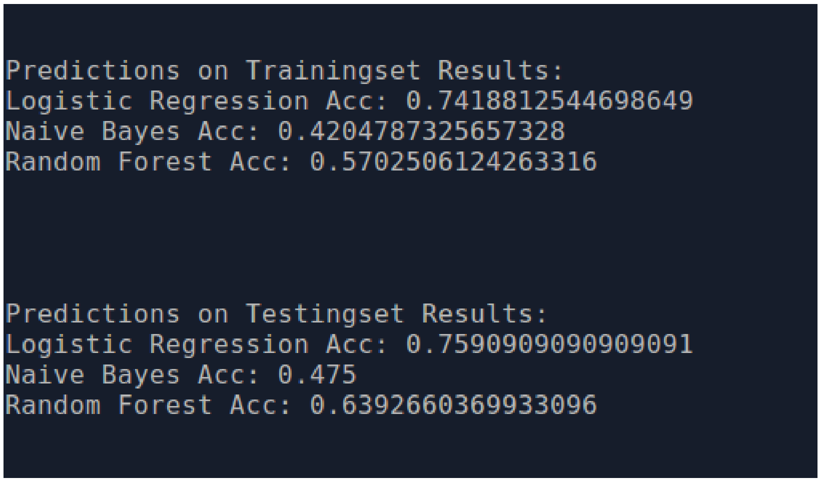
Out of all the combinations we observed that Term Frequency was giving best results while doing validation later on while building Machine Learning models.

Building Machine Learning Models:

We tried three different Machine Learning models as follows,

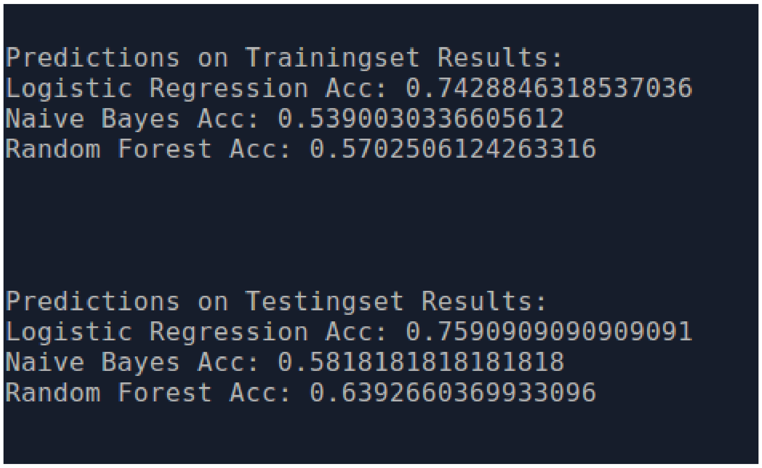
- Logistic Regression
- Naive Bayes
- Random Forest

While the best results were obtained using Term Frequency, the second best were obtained while using TF-IDF. The results are as below.



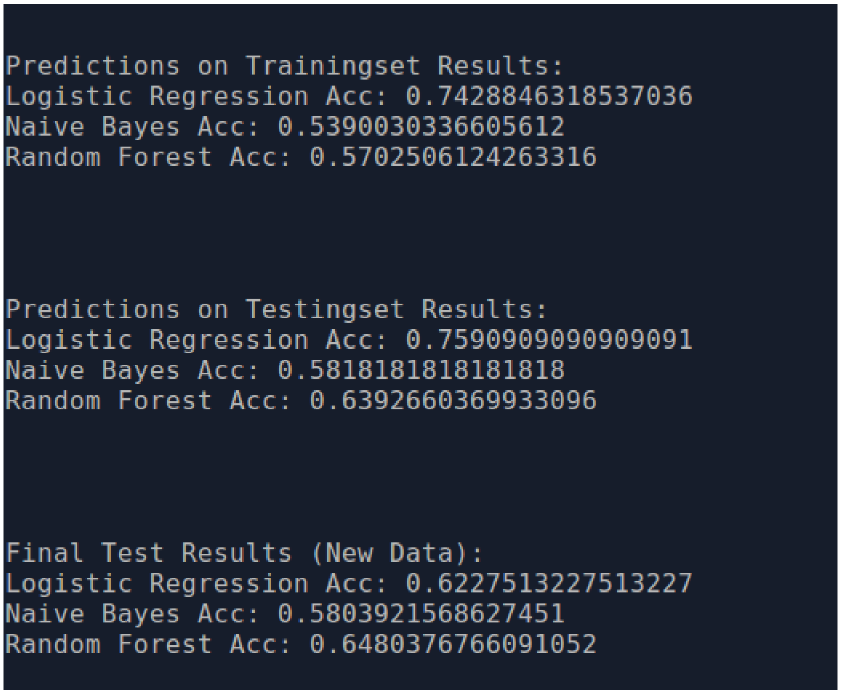
```
Predictions on Trainingset Results:
Logistic Regression Acc: 0.7418812544698649
Naive Bayes Acc: 0.4204787325657328
Random Forest Acc: 0.5702506124263316



Predictions on Testingset Results:
Logistic Regression Acc: 0.7590909090909091
Naive Bayes Acc: 0.475
Random Forest Acc: 0.6392660369933096
```

The best results were obtained while using Term Frequency. The results are like below.



```
Predictions on Trainingset Results:
Logistic Regression Acc: 0.7428846318537036
Naive Bayes Acc: 0.5390030336605612
Random Forest Acc: 0.5702506124263316



Predictions on Testingset Results:
Logistic Regression Acc: 0.7590909090909091
Naive Bayes Acc: 0.5818181818181818
Random Forest Acc: 0.6392660369933096
```

Taking new data and running the classifiers gave following results.



```
Predictions on Trainingset Results:
Logistic Regression Acc: 0.7428846318537036
Naive Bayes Acc: 0.5390030336605612
Random Forest Acc: 0.5702506124263316



Predictions on Testingset Results:
Logistic Regression Acc: 0.7590909090909091
Naive Bayes Acc: 0.5818181818181818
Random Forest Acc: 0.6392660369933096



Final Test Results (New Data):
Logistic Regression Acc: 0.6227513227513227
Naive Bayes Acc: 0.5803921568627451
Random Forest Acc: 0.6480376766091052
```

Inference : Getting similar results for the final testing set same as testing and training set implies that the models have been performing consistently.

Final Inference: Looking at all the figures the order of precedence of best classifier for this dataset is as follows,

- Logistic Regression
- Random Forest
- Naive Bayes