**Project completed by SHUBHAM SHAILESH PANDEY, UBID - spandey4**

The environment used was Jupyter.

**Where is the output –**

Since, the platform chosen was Jupyter, you can see the output in the python notebook itself. The notebooks are also converted to markdown files inside the ‘Markdown files’ folder.

Screenshots are available in Screenshots folder.

Accuracies obtained for Test dataset were –

* Naive Bayes – 97.69%
* Logisitic Regression – 93.02%

Accuracies obtained for Unknown dataset were –

* Naive Bayes – 92.82%
* Logisitic Regression – 87.06%

**How to run the program –**

* First install python packages pyspark and BeautifulSoup for the Jupyter environment.
* For Titanic Data Analysis – You just need train.csv and TitanicPySpark.ipynb in the same folder. Running TitanicPySpark.ipynb in Jupyter will produce similar output as seen in the notebook.
* For News Article Data Analysis – You need folders named ‘Data’ and ‘Unknown’ and NewsClassification.ipynb. Running NewsClassification.ipynb in Jupyter will produce same output as seen now.
* There is also DataCollect.ipynb that was used to collect articles into txt files inside the ‘Data’ and ‘Unknown’ folder.

**Data pipeline Flowchart –**

