MACHINE LEARNING PROJECT (PHASE 1)

GROUP MEMBERS:

SALMAN SHAH

(17L4449)

ANEEEQ ASGHAR

(17L4525)

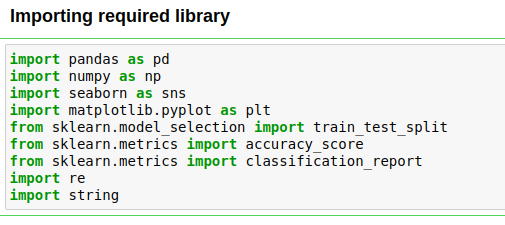
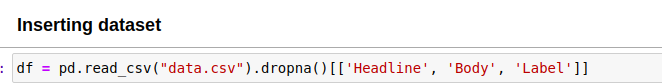
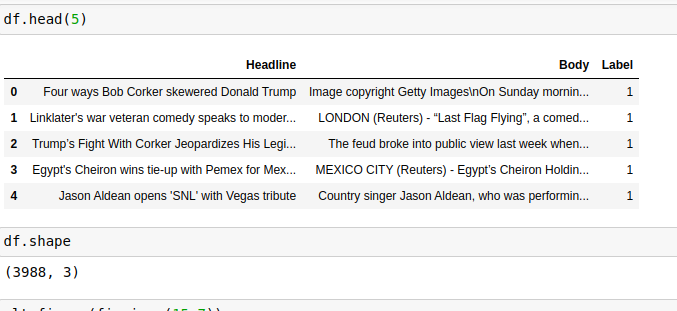
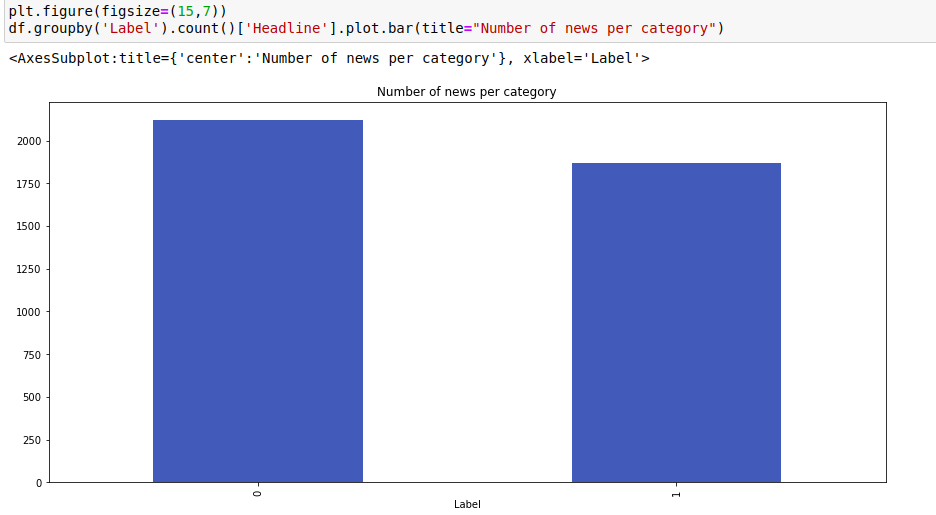
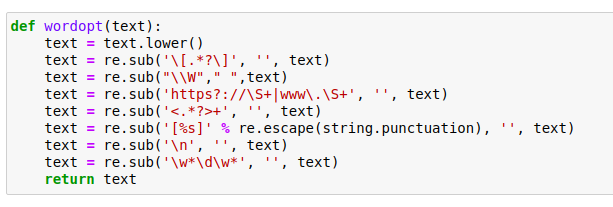
**SUBJECT:**

**Fake news detection using deep learning**

**SOFTWARES:**

\* JUPYTER NOTEBOOK

**PROCEDURE:**

* Firstly, we import all libraries required for phase 1 on JUPYTER notebook as shown below:
* 
* We downloaded the CSV file from Kaggle.com.
* Now we inserted the dataset.
* 
* We selected only the Headline, and Label to predict the Label of the dataset.
* 
* Now we visualized the data using matplotlib library and classified the data according to the news whether it is fake or true.
* 
* Next, we cleaned the dataset by removing all kinds of punctuations, end line characters using the **re** library. Here, a function of **wordopt** is used to clean the whole dataset’s body.
* 





Here, I vectorized the body of the dataset using the Tfid Vectorizer by importing it from Sklearn library as shown below:

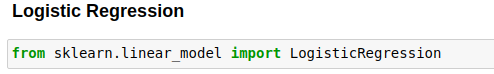
 Now, I split the dataset into five different sizes of training and test set respectively.

1. 75% training set and 25% test set.
2. 85% training set and 15% test set.
3. 60% training set and 40% test set.
4. 90% training set and 10% test set.
5. 70% training set and 30% test set.

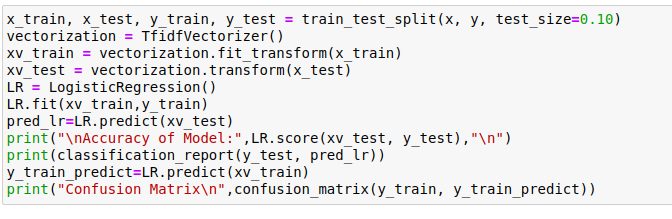
The code for splitting the dataset into 90% train dataset and 10% test dataset by setting the test\_size to 0.1. Vectorization is done through TfifVectorizer.

The Classification Model used in this experiment is the Logistic Regression for all of the five splits.

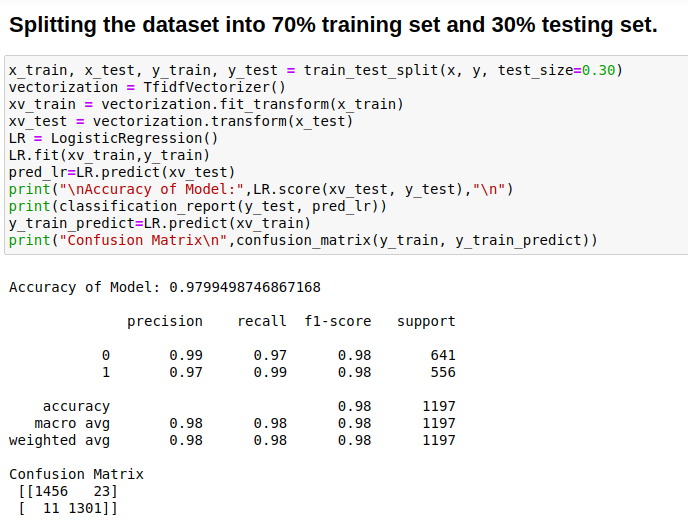
Logistic Regression is applied using the inbuilt Logistic Regression from sklearn.linear\_model library.



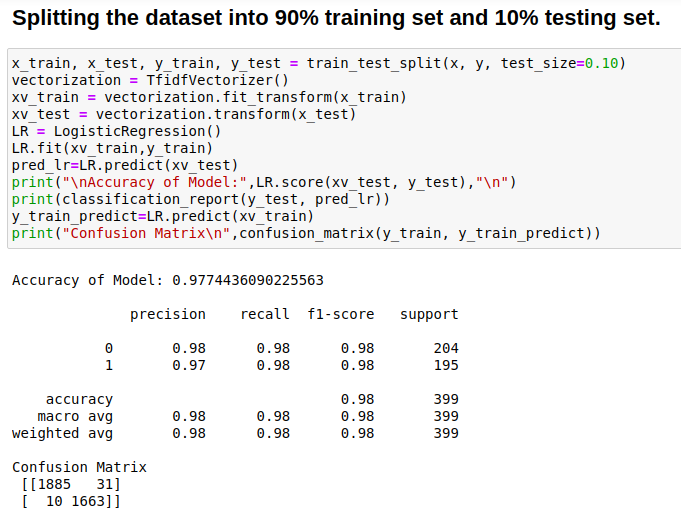
Furthermore, the accuracy and confusion matrix for each of the splits is calculated using the code given below.



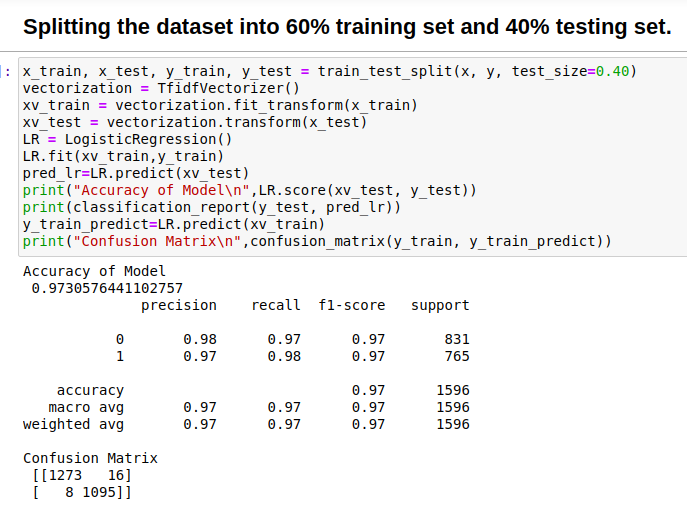
The outputs for each of the split is shown below:



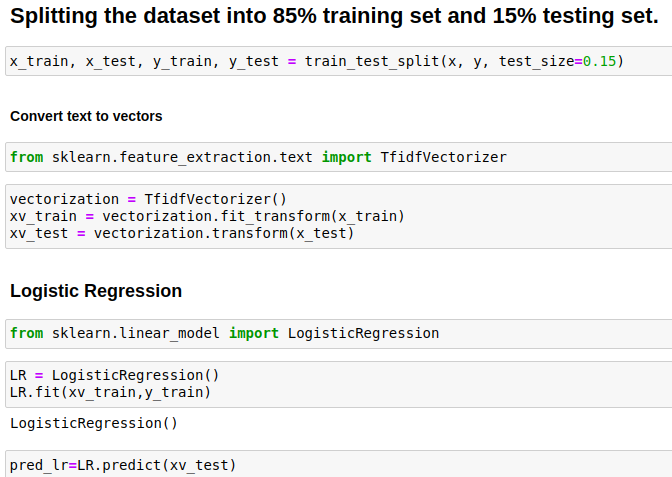
90% training set and 10% test set.

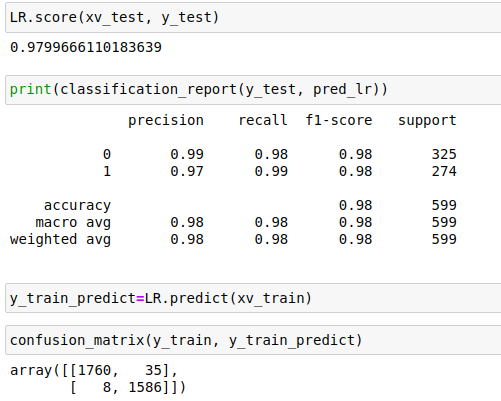


60% training set and 40% test set.

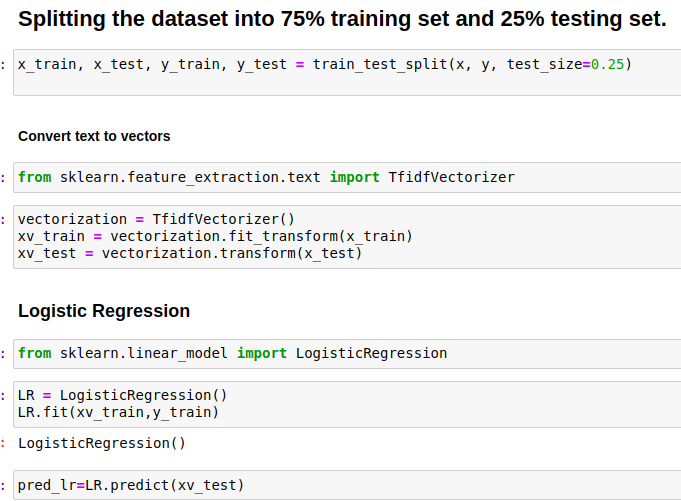


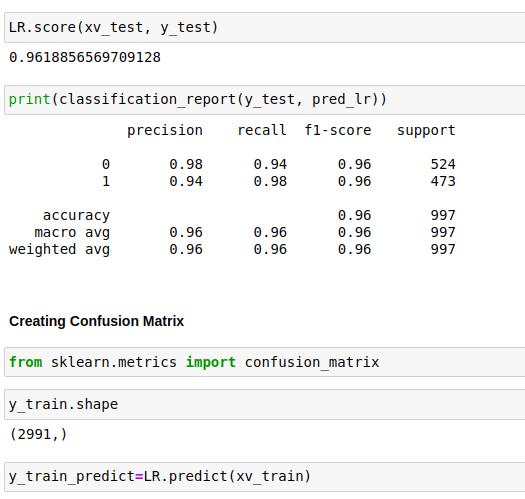
85% training set and 15% test set.

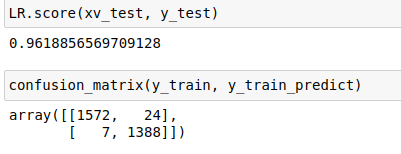




75% training set and 25% test set.







As we have seen the accuracy of each of the above-mentioned splits. We have selected the train-test split with the maximum accuracy which is 85% training set and 15% test set. It has a maximum accuracy of 97.9966%. The confusion matrix provides us with True Positives and False Positive, False Negative, and True Negative.

