**Email spam Detection with Machine Learning**

**Project Report**

**Email spam**, are also called as junk emails, are unsolicited messages sent in bulk by email (spamming).

In this **Data Science**Project I will show you how to detect email spam using **Machine Learning** technique called  **Natural Language Processing**and **Python**.

### **So this program will detect if an email is spam (1) or not (0)**

**Import the libraries:**

****

FIG 1

**Now let’s explore the data and get the number of rows & columns :**



FIG 2

**To check for duplicates and remove them :**

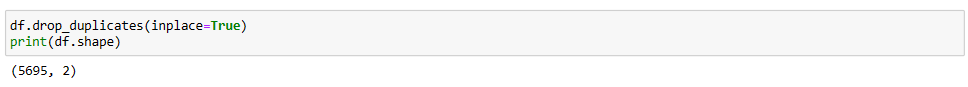
****

FIG 3

**To see the number of missing data for each column :**

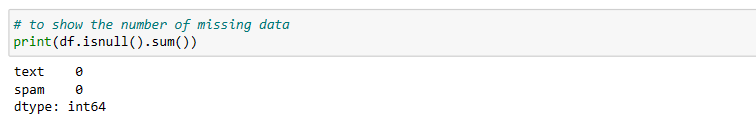
****

FIG 4

### Now Download the stop words

Stop words in natural language processing, are useless words (data).

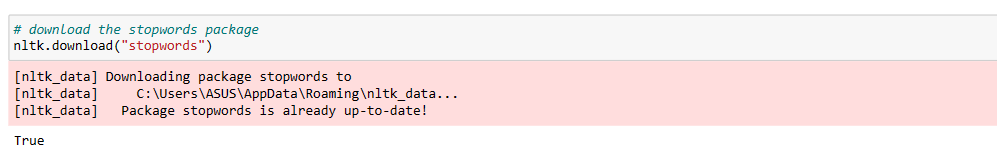


FIG 5

Now Create a function to clean the text and return the tokens. The cleaning of the text can be done by first removing punctuation and then removing the useless words also known as **stop words.**

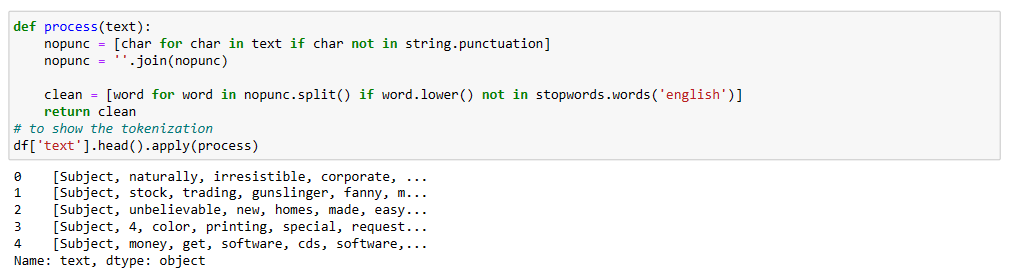


FIG 6

**Now convert the text into a matrix of token counts :**

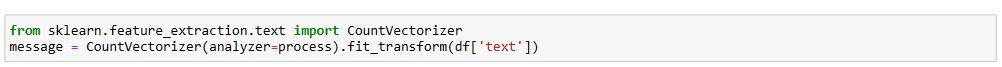
****

FIG 7

Now we need to split the data into training and testing sets, and then we will use this one row of data for testing to make our prediction later on and test to see if the prediction matches with the actual value.



FIG 8

Now we need to create and train the**Multinomial Naive Bayes** classifier which is suitable for classification with discrete features.



FIG 9

**To see the classifiers prediction and actual values on the data set :**

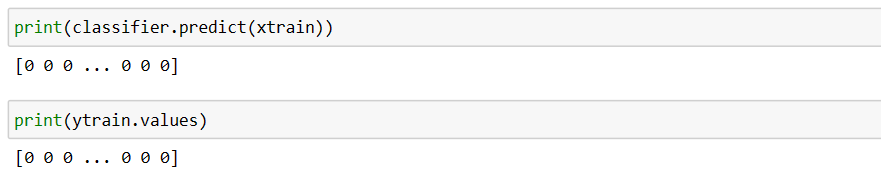
****

FIG 10

Now let’s see how well our model performed by evaluating the **Naive Bayes classifier**and the **report, confusion matrix & accuracy score.**

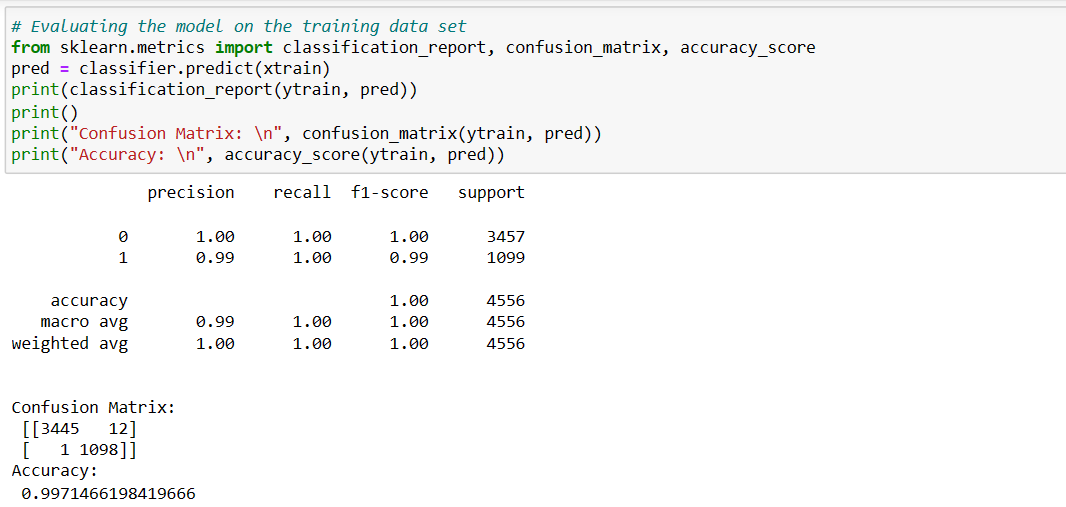


FIG 11

It looks like the model used is 99.71% accurate. Let’s test the model on the test data set (xtest &  ytest) by printing the predicted value, and the actual value to see if the model can accurately classify the email text.

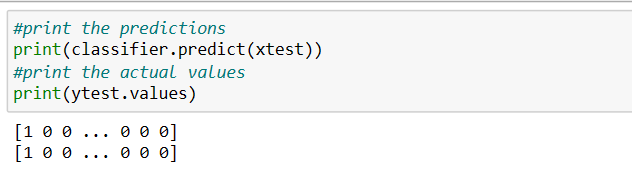


FIG 12

**Now let’s evaluate the model on the test data set:**

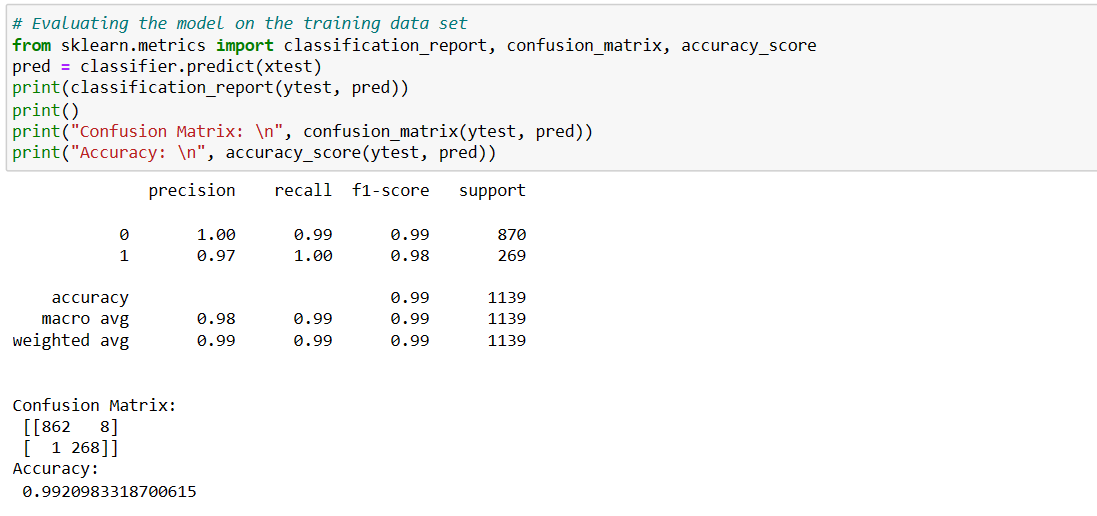
****

FIG 13

The classifier accurately identified the email messages as spam or not spam with 99.2 % accuracy on the test data.