**A LAB REPORT ON THE COMPARISON IN PERFORMANCE BETWEEN NAÏVE BAYES AND LOGISTIC REGRESSION**

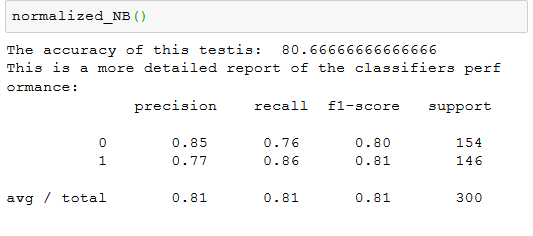
# LIBRARY USED

In order to conduct this study I needed to use various libraries in python. The libraries I took advantage of when modeling the various classifiers were the scikit-learn library, Natural Language Toolkit(NLTK), and re. I used **re** which is a regex library to help preprocess my data and by that I mean clean all punctuation so I was only left with characters. **The NLTK library** was also used for text processing such as stemming and lemmatization in an attempt to improve the performance of my classifier. **The scikit-learn Library** was the main library that was used to model both my naïve\_bayes and logistic regression classifiers. It is very user friendly, including many preprogrammed algorithms that make machine learning easy especially for novices. In addition, there was a lot of well written documentation for this particular library which made it far easier to understand and therefore use.

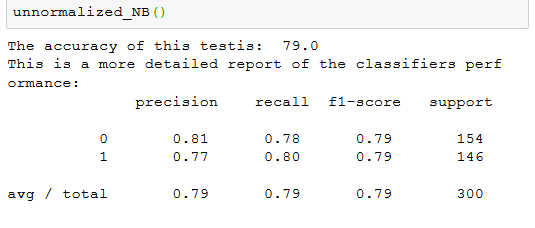
# *EVALUATION OF RESULTS*

In order to evaluate the various classifiers, accuracy, precision, recall, F-meausre were calculated. However, the weghted averages of the various classifers was also calculated with the help of the sklearn classification library. My classifier was evaluated mainly using precision, recall and F-measure. However, the accuracy was calculated to help distinguish the difference between the results gained when evaluated using accuracy and the results gained when using either precision, recall or f-measure. The Accuracy was not used because it is said be an inefficient way of evaluation when the data is unbalanced or bias.

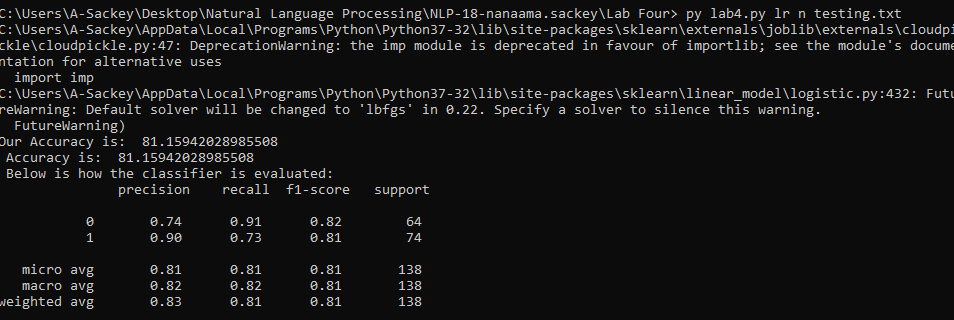
**Result for Normalized Naïve Bayes**



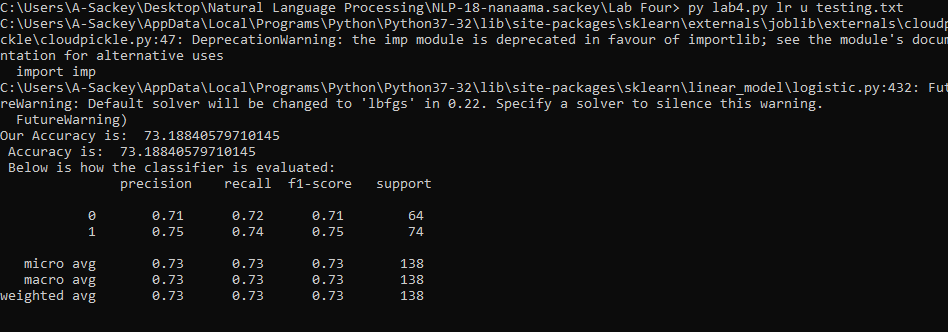
**Result for Unormalized Naïve Bayes**



**Result for LR N**



**Result for LR U**



# Discussion of and Comparison of Results.

# DISCUSSION OF RESULTS

Looking at the result that was gotten from the various classifier, the normalized classifiers as compared to the unnormalized classifiers for both the Naïve bayes and logistic regression had better performance based on all the measures used in evaluating the classifier (accuracy, recall, precision, and F1- Measure. This is so because normalization in any form applied to a classifier is said to reduce empirical error and improve upon the performance of the classifier.

**Comparison of Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier | Accuracy(%) | Precision | Recall | F1 measure |
| Normalised Naïve bayes | 78 | 0: 0.84  1: 0.76 | 0: 0.70  1: 0.87 | 0: 0.76  1: 0.81 |
| UnNormalised Naïve bayes | 76 | 0: 0.75  1: 0.78 | 0: 0.78  1: 0.76 | 0: 0.76  1: 0.77 |
| Normalised LR | 81 | 0: 0.74  1: 0.90 | 0: 0.91  1: 0.73 | 0: 0.82  1: 0.81 |
| UnNormalised LR | 73 | 0: 0.71  1: 0.75 | 0: 0.72  1: 0.74 | 0: 0.71  1: 0.75 |

Nb: 0 and 1 are the system labels

However, looking at the weighted averages the classifiers had the following percentages.

**Normalised Naïve bayes = 0.79**

**UnNormalised Naïve bayes = 0.77**

**Normalised LR = 0.81**

**UnNormalised LR =0.73**