You are also to write a report about the lab. The report should contain:

**A LAB REPORT ON THE MODELING OF TWO VERSIONS OF A NAÏVE BAYES AND A LOGISTIC REGRESSION CLASSIFIER**

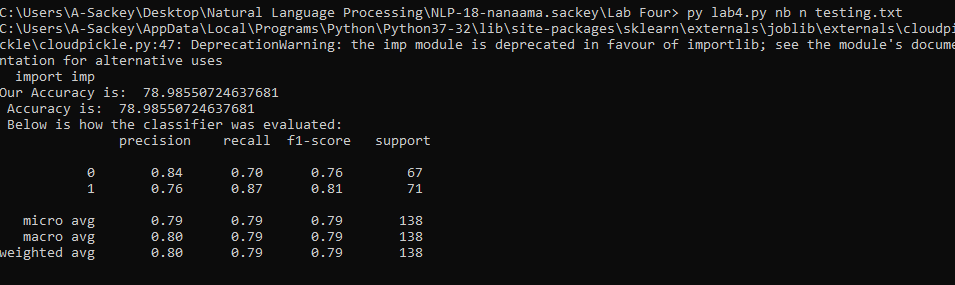
# LIBRARY USED

The libraries that were used in modeling the various classifiers were the Scikit-Learn library, Python Data Analysis Library(pandas), Natural Language Toolkit(nltk) and numpy library. **The pandas library** is a library that makes working with labeled data easy and intuitive. It also helps in handling many data. Data is collected or put in data frames which makes it easier to be able to work with it other than working with lists or dictionaries and looping through various data structures. **The nltk library** was used for text processing such as removing stop words. The numpy was used in the calculation of the accuracy. **The scikit-Learn library** was the main library that was used in the modeling of both the naïve\_bayes and logistic regression classifiers. This is because as compared to other libraries like the nltk, it has many algorithms that make solving natural language problems easy. Also, the documentation of the scikit-learn library was easier to understand and use hence.

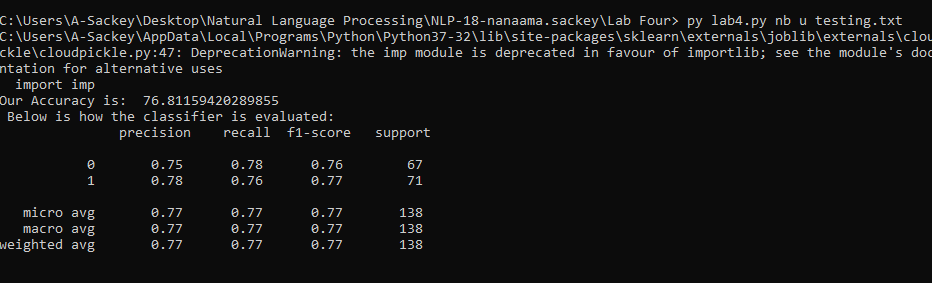
# *EVALUATION OF RESULTS*

On evaluating 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.

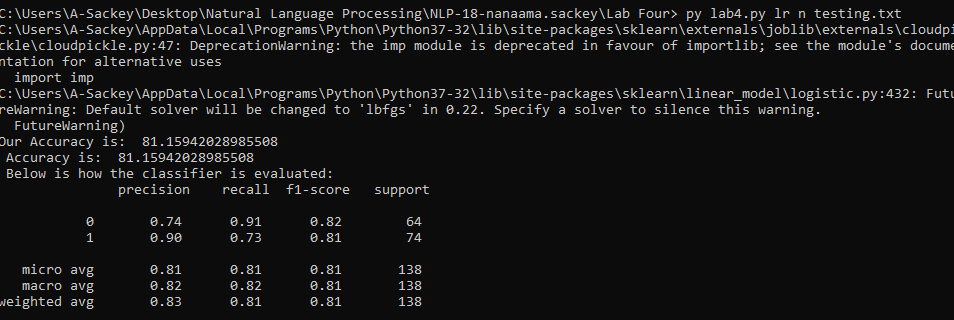
**Result for NB N**



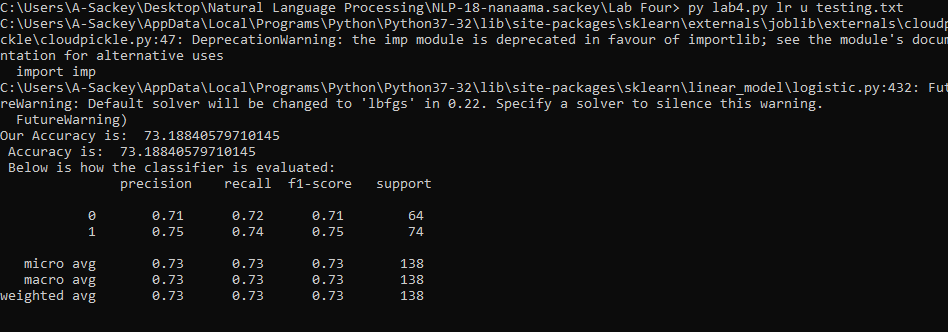
**Result for NB U**



**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**