**LAB FOUR REPORT – NAÏVE BAYES AND LOGISTIC REGRESSION CLASSIFIERS**

**Introduction**

In this lab, the task was to develop a naïve Bayes and a logistic regression classifier. There was supposed to be an option for word normalization on the training data given or not. For these classifiers, I employed the use of libraries to help classify the data.

The libraries I used were mainly the sklearn library, nltk library, pandas library, and sys. The sklearn library is the main library from which I derived the naïve Bayes classifier and the logistic regression classifier functions. The nltk library is the source of the various functions used for word normalization. The panda's library is the source for all functions used for organizing the data, and the sys library is used for reading from the terminal.

**Methodology**

The training data was first imported into the program and put into a data frame with the help of the panda's library. According to the user’s input, the data would be normalized or not. After that stage, the training data was vectorized and transformed to get the features extracted. That is, the words that are to be checked. After feature extraction, the training data is split into two – one for training and the other for testing. By my program, the testing data was 10% of the training data. The model was trained by the naïve Bayes or logistic regression classifier as per the user’s input. For testing, the user had to provide a text document with sample sentences. The text document was put into a data frame and vectorized before it was tested with the model.

**Testing and Evaluation**

Included in the sklearn library is a function known as accuracy\_score that gives the percentage accuracy of the classifier in classifying the test data. This was how I evaluated each classifier.

I first tested the classifier with the test data extracted from the training data. For the naïve Bayes classifier, the accuracy was 83.27%, and for the logistic regression, the accuracy was 84%.

From the results, it can be observed that the logistic regression classifier achieved a better result than the naïve Bayes classifier. This could be because of the type of classifier it is. While logistic regression is a discriminative classifier, the naïve Bayes is a generative one, meaning the naïve Bayes generates its features while the logistic regression learns them and so it could be that more time is spent in generating features than classifying them. Also, the naïve Bayes uses the Bayes rule to expand the formula for calculating the probability of a sentence given a class while the logistic regression computes that probability directly. Moreover, the logistic regression uses more advanced calculations than the naïve Bayes resulting in higher accuracy.

**Conclusion**

Through this lab exercise, I learned about more complex libraries and how they are used to solve problems. I also gained knowledge of vectorizing and how after the training data has been fit transformed, the test data is only transformed and not fitted. That was the source of my main error – refitting the test data. The importance of the jupyter environment became more real to me as well when I was debugging because I could run just a section of the code to see the results being given.