Assignment 3(PRML)

Name- Ravi Vishwakarma

Roll No. CS22M070

Task:

To build a spam classifier from scratch.

It’s accuracy on test data is -

```
Accuracy on test Data: 97.032323
```

I implemented it with naive bayes.

STEP1:- I used a spam.csv dataset which was available on kaggle. I extracted features such as labels and the text from the dataset since they were necessary. Named the label column as target and message as text.

STEP2:- Dropped the duplicates from the dataset.

STEP3:- Then I removed the stop words and punctuation since they are not very useful in deciding whether the mail is spam or ham.

STEP4:- Then divide the dataset into two lists one contains the spam mails and another one contains ham mails.

STEP5:- a vocabulary of unique words which occur in known spam mails and a different vocabulary of unique words which occur in known ham mails.

STEP6:- Calculated the spamicity of dataset which was 0.1266485647788983 Calculated the hamicity of dataset which was 0.8733514352211016

STEP7:- Now fetch the test data from the test folder of the current directory and remove the punctuation from test mail.

STEP8:- Consider the word of test data only if it is present in spam or ham in training data otherwise simply drop it.

STEP9:- Then call the Bayes function which then computes the probability of whether the mail is spam or not given its words.