Name – Raj Uday Mahajan

Roll Number - CS22M067

PRML Assignment 3 – Report

- I have used **SMSSpamCollection** data set for this assignment. I have attached it in CS22M067.zip file.

- This dataset consists around 5.5 thousand messages which are tagged for spam and ham.

- I have divided dataset into training dataset (80%) and testing dataset (20%) randomly so that ham and spam messages should spread evenly throughout dataset.

- I have preprocessed data by converting upper case letters to lower case letters and removing the punctuation marks.

- I have used **Naïve Bayes** algorithm to build the program.

- Basically, we calculate conditional probability like given all words in mail what is probability that it is Ham and similarly given all words in mail what is probability that it is Spam.

- We calculate the probability of message being spam or ham. Whichever is greater probability we mark message accordingly.

- **Observations :-**
  I got **98.7432 %** accuracy for testing dataset (20%) when trained on training dataset (80%).