# MACHINE LEARNING PROGRAMMING-2 REPORT

Submitted by: - Parth Parashar PSU ID: - 923928157

## Description:

In this experiment we have used “spambase” dataset which has rows having information about the spam mails (1 class) and non-spam mails (0 class).

We have calculated standard deviation, mean on the training data (50% of the data with 40% spam and 60 % nor spam) and then have tested 50% of data using the calculated standard deviation and mean of the training data. For testing we have calculated the probabilities and based on those probabilities we have classified the classes of the data (using logarithmic sums of probabilities) and then finally calculated accuracy, recall, precision, and confusion matrix on the test data.

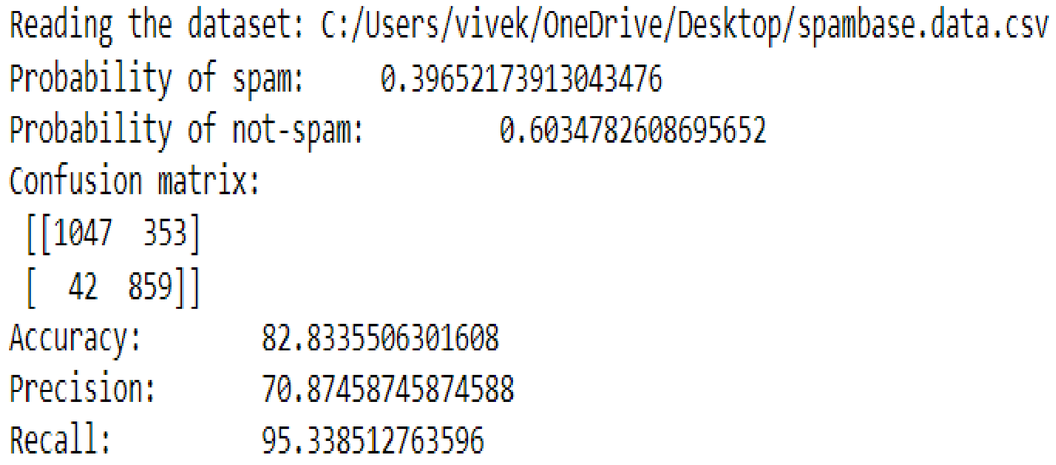
## Creating Training and Testing dataset:

Training and testing dataset have been created by splitting the total instances into half with both datasets containing 40% spam and 60% non-spam.

## Probability Model:

For the probability model, the mean and standard deviation for each of the 57 feature is calculated. In case, any of the feature results in a 0-standard deviation, then it is scaled to 0.0001 to avoid the divide-by-zero error.

## Results:



```
Reading the dataset: C:/Users/vivek/OneDrive/Desktop/spambase.data.csv
Probability of spam:     0.39652173913043476
Probability of not-spam:        0.6034782608695652
Confusion matrix:
 [[1047  353]
 [  42  859]]
Accuracy:       82.8335506301608
Precision:      70.87458745874588
Recall:         95.338512763596
```

## About Result:

Upon successful compilation and consequent run of the program, an accuracy of 82.83% is obtained along with a precision and recall of 70.87% and 95.33% respectively.

The accuracy of the naïve bayes classifier is not that good when compared to the other classifiers taught in class as well as other experiments performed in previous programming assignments. Also, it can be sensed from the results that there is room for improvement in terms of recall and precision as well.

The parameters used in the results are explained below: -

The confusion matrix gives all the data which is classified correctly and incorrectly

Precision and recall provide us with the relevance and sensitivity values of data.

## Do you think the attributes here are independent, as assumed by Naïve Bayes?

We have assumed that the attributes are independent but there might be the presence of one word in the spam mail implying that the other word is present most of the times (say 90% of times) that can lead to some dependence (Ex: Say spam mail is about some jackpot then there is high probability of having word money in it similarly for project, discussion, and meeting).

In broader terms, “spam database” has attributes indicating how frequently words are present/displayed in the email. Considering each independent frequency of words, the actual context of the message might be ignored which can also lead to false results. These false results can point towards some form of dependencies in the data.

So, it may not be that independent (depending on the words and their frequencies in the dataset).

## Does Naïve Bayes do well on this problem despite the independence assumption?

No, Naïve Bayes does not do well on this problem. This is evident from the results we are obtaining as the accuracy is on the lower side when compared to our previous programming assignments.

Other parameters obtained are also not great meaning that there is a room for improvement which might be done using other algorithms such as SVM(s).

## Speculate on other reasons Naïve Bayes might do well or poorly on this problem?

Naïve Bayes may perform better if we consider only the important features. We can use feature selection methods (like tf-idf) and can observe what attributes are statistically meaningful and have a dense presence over the entire data and then can use only those to classify which would probably increase the performance.