CS 545 Programming Assignment 2

In the Programming Assignment 2, Naive Bayes Classifier is used to classify the spambase dataset available at https://archive.ics.uci.edu/ml/datasets/spambase into spam or not spam.

In order to design this classfier following steps are followed:

1. The data available is converted and split into training and test data using the train_test_split() function of the sklearn.model_selection.

2. The prior probability of the class 1 and class 0 is calculated based on the count of the spam and non-spam samples in the training data.

3. For each training sample, the mean and standard deviation is calculated for its 57 features and stored in a dictionary.

4. For each of the testing sample, the probability of its 57 features with respect to class 1(positive) and class 0(negative) is calculated using the mean and standard deviation values evaluated in step 3.

5. The Gaussian Naive Bayes Classifier is then used to select the max of the summation of log values of the respective prior probability and its correspoding feature probabilities. If the positive i.e. class 1 probability is more then the test sample is considered to belong to class 1 else it is considered to belong to class 0.

6. Comparing the test predictions and actual test targets, the confusion matrix created is as follows:

```
Confusion matrix:  [[986. 377.]
                    [ 49. 889.]]

 # TP = True Positive, TN = True Negative, FP = False
    Positive, FN = False Negative
 TP = confusion_matrix[1, 1]
 TN = confusion_matrix[0, 0]
 FP = confusion_matrix[0, 1]
```

```
FN = confusion_matrix[1, 0]

accuracy = (TP + TN) / (TP + TN + FP + FN)
precision = TP / (TP + FP)
recall = TP / (TP + FN)

Thus, using the above formulas and confusion matrix
   values:

Accuracy = (889 + 986)/(986 + 377 + 49+ 889) = 0.81
Precision = 889 / (889 + 377) = 0.70
Recall =  889 / (889 + 49) = 0.94
```

Thus, the accuracy of the Gaussian Naive Bayes Classifier is 0.81. Though, decent accuracy is achieved it is based on the assumption that the sample attributes are independent. However, I do not think they are completely independent as first 48 attribute itself represent the percentage of words in an email. Also, there is a possibility of the other attributes not having clear differentiation between them.

Thus, despite of the independence assumption, the Gaussian Naive Bayes classifier does achieve a decent accuracy of 0.81, which could have been improved. As the classifier operates on continuous values rather than discrete, its accuracy is comparatively less. Also, selecting the attributes or features that are more independent i.e. are less correlated, can improve the accuracy of the Naive Bayes classifier.