Shri Guru Gobind Singhji Institute of Engineering & Technology
Nanded - 431606, India.
(An Autonomous Institute of Govt. of Maharashtra)

Participants:
1. Shweta Anil Dalal 2017BEC023
2. Laxmikant Vasudev Suryavanshi 2017BEC094

Title: Email Spam Detection

Classifier:

Naive Bayes Classifier is a probabilistic classifier based on Bayes theorem assuming independence between every pair of features.
An advantage of naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification.
For this project, we specifically used *multinomial naive Bayes classifier*. With a multinomial event model, feature vectors represent the frequencies with which certain events have been generated by a multinomial (p_1, p_2, ... ,p_n) where p_i is the probability that event i occurs. A feature vector x = (x_1, x_2, ... ,x_n) is then a histogram, with x_i counting the number of times event i was observed in a particular instance. This is the event model typically used for document classification, with events representing the occurrence of a word in a single document.

Result:

The model is trained on a dataset with 5695 unique values and train-test split is 80% and 20% respectively. The results are given in the table below.

Evaluation of model on train dataset

	precision	recall	f1-score	support
0	1.00	1.00	1.00	3457
1	0.99	1.00	0.99	1099
accuracy			1.00	4556
macro avg	0.99	1.00	1.00	4556
weighted avg	1.00	1.00	1.00	4556

Confusion Matrix:
```
 [[3445   12]
 [   1 1098]]
```
Accuracy: 0.9971466198419666

```
Evaluation of model on test dataset
              precision    recall  f1-score   support

           0       1.00      0.99      0.99       870
           1       0.97      1.00      0.98       269

    accuracy                           0.99      1139
   macro avg       0.98      0.99      0.99      1139
weighted avg       0.99      0.99      0.99      1139

Confusion Matrix:
 [[862   8]
 [  1 268]]
Accuracy: 0.9920983318700615
```

Resources:

Dataset and source code can be found here.
Reference: Spam-or-Ham-Email-Classification.