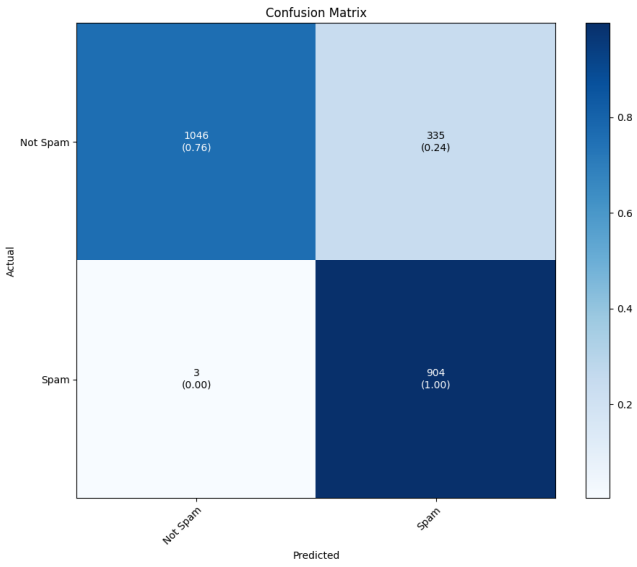
In this experiment, we aim to classify emails from the Spambase dataset as either spam or non-spam using a Gaussian Naïve Bayes classifier. The experiment was set up as follows:

Data Processing: The dataset was split into a training set and a test set, maintaining the class distribution of approximately 39.4% spam and 60.6% non-spam in both sets. Around 2,300 instances were selected for each set.

Gaussian Naïve Bayes Classification: A Gaussian Naïve Bayes classifier was trained on the training set, assuming independence among features given the class label. The classifier learned the distribution of features for each class (spam and non-spam) using Gaussian probability density functions.

Evaluation and Metrics: The trained model was then evaluated on the test set. Accuracy, precision, and recall metrics were calculated to assess the classifier's performance in correctly classifying instances as spam or non-spam. Confusion matrices were used to analyze the distribution of true positive, true negative, false positive, and false negative predictions. A total of 3 tests were conducted.

Test 1

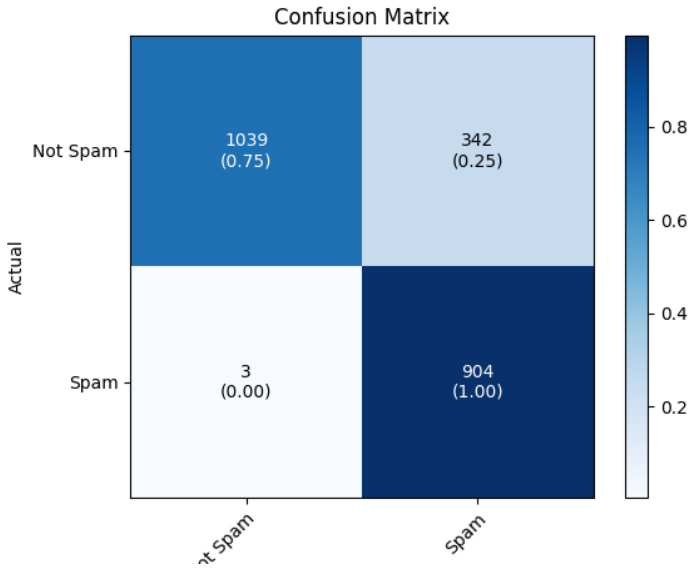


Accuracy: 0.8522727272727273

Precision: 0.7296206618240516

Recall: 0.9966923925027563

Test 2

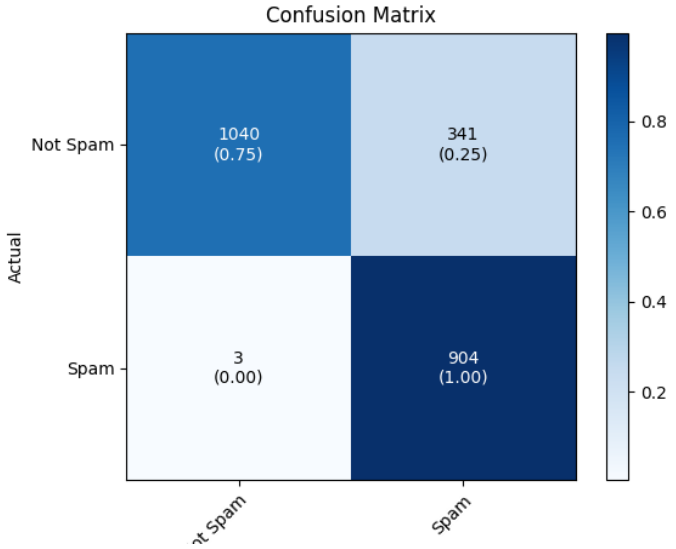


Accuracy: 0.8492132867132867

Precision: 0.7255216693418941

Recall: 0.9966923925027563

Test 3



Accuracy: 0.8496503496503497

Precision: 0.7261044176706827

Recall: 0.9966923925027563

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

Based on the provided test results, the Gaussian Naïve Bayes classifier achieved reasonably good accuracy, precision, and recall scores on the Spambase dataset. In all three tests, the accuracy was around 85%, indicating that the model is performing well in classifying spam and non-spam emails. The model was precise about 72~73% of the time when predicting an instance as spam. The recall values were consistently high around 99.7%, indicating that the model was able to correctly identify the vast majority of actual spam instances.

However, the precision-recall trade-off is evident in these results. While the model demonstrates high recall (few actual spam instances were missed), the precision is relatively moderate (some instances predicted as spam were not actually spam).

Regarding the assumption of attribute independence in Naïve Bayes, it's important to note that Naïve Bayes assumes that features are conditionally independent given the class label. In reality, this assumption might not hold perfectly for all datasets. Despite this assumption, Naïve Bayes can often perform surprisingly well, as seen in the results.

Q. Does Naïve Bayes do well on this problem in spite of the independence assumption? Speculate on other reasons Naïve Bayes might do well or poorly on this problem.

The confusion matrix values also reveal certain patterns. Notably, there are about 24~25% non-spam being predicted as spam in each test (false positives). This suggests that the assumption of feature independence made by the Gaussian Naïve Bayes classifier might not hold perfectly for this dataset.

Despite the independence assumption, Naïve Bayes seems to perform well on the problem of predicting spam correctly as spam (true positives). It's worth noting that Naïve Bayes can still perform reasonably well when the independence assumption isn't met, as long as the conditional dependencies between features are not strong enough to significantly affect the classification decision. The relatively high accuracy and precision suggest that the classifier is making accurate decisions for a majority of instances.

Possible reasons for the observed performance could include:

Feature correlations: Although Naïve Bayes assumes feature independence, certain features might still have correlations that align well with the class distribution, leading to accurate predictions.

Dominant features: If certain features have a significant impact on the classification decision and are indicative of the class, their influence might overshadow any inter-feature correlations.

Class separability: The features might exhibit clear separation between spam and non-spam instances in the data space, allowing Naïve Bayes to make effective decisions.

On the other hand, some reasons for potential challenges in performance could include:

Strong Feature Dependencies: If certain features are strongly dependent on each other, violating the independence assumption, Naïve Bayes might struggle to capture such relationships accurately.

Outliers: Outliers or extreme values could disproportionately affect the Gaussian distribution assumptions, leading to misclassifications.