جامعة أم القرى
UMM AL-QURA UNIVERSITY

Data Analysis 2

Task2

Student Name	ID
Mawaddah Alharthi	443004953
Bayan Alqarni	443002589

Naive Bayes (spambase)

1. Dataset Overview

The dataset used is the spambase dataset, which helps classify emails as either spam (1) or not spam (0). Each row represents an email, and the columns contain features like word frequencies, special character counts, and other metrics. The last column (57) is the target column that tells whether the email is spam or not.

2. Data Preprocessing

We performed the following steps to prepare the data for the model:

- - Split Features and Target: The last column was used as the target (y), and the rest of the columns were used as features (X).
- - Scaling Features: We used MinMaxScaler to scale the features to values between 0 and 1. This helps the model handle different ranges of values.
- - Train-Test Split: The data was split into 80% for training and 20% for testing, so we could check the model's performance on unseen data.

3. Model Choice: Multinomial Naive Bayes (MNB)

We used the Multinomial Naive Bayes model because it works well with text-based data, like word counts or frequencies, which is common in spam detection. The model assumes that the features are independent, and even though this assumption isn't always true, Naive Bayes often gives good results for classification problems like this.

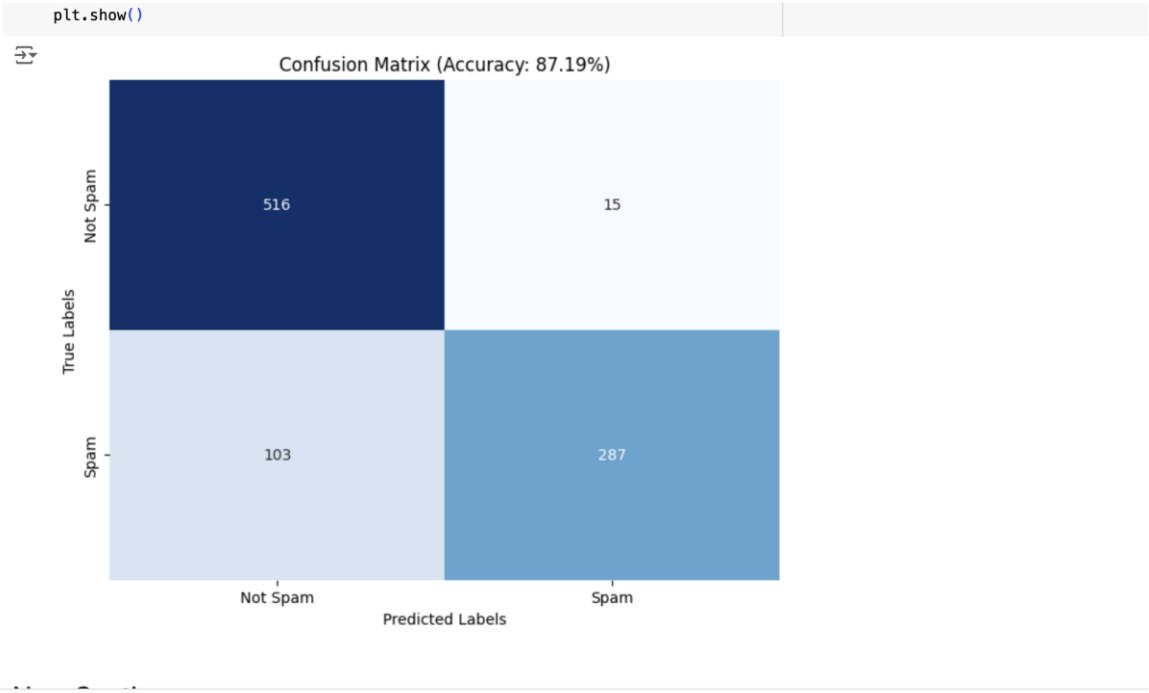
4. Model Performance

We evaluated the model with different metrics:

- - Accuracy: The model achieved 87.19% accuracy, meaning it correctly classified most emails.
- - Classification Report: The report shows precision, recall, and F1-score for both spam and not-spam emails. All metrics are high, indicating the model performs well.

	precision	*recall*	*f1-score*	*support*
Not Spam	*0.95*	*0.93*	*0.94*	*590*
Spam	*0.93*	*0.96*	*0.94*	*608*

- - Confusion Matrix: The confusion matrix shows how many emails were correctly or incorrectly classified.



- True Positives (TP): Spam emails correctly identified as spam.
- True Negatives (TN): Non-spam emails correctly identified as not spam.
- False Positives (FP): Non-spam emails wrongly classified as spam.
- False Negatives (FN): Spam emails wrongly classified as not spam.

5. Insights

The model performed very well with an accuracy of 87.19% . Balanced precision and recall mean the model avoids many false positives and false negatives. Multinomial Naive Bayes is a good choice for this problem, as it handles frequency-based features efficiently. The results show that the model can reliably detect spam, but further improvements can be made by testing other models or tuning the parameters.