1. **Problem Statement:** Detect Spam Mail using machine learning model.
2. **Methodology:**

   Data Collection: Use datasets like SpamAssassin or Enron Emails.

   Preprocessing:

   a. Clean text (remove punctuation, lowercase, etc.).
   b. Tokenize and remove stop words.
   c. Convert text to numeric form (TF-IDF, embeddings).

   Feature Engineering: Extract patterns like keywords, email length, etc.

   Model Selection: Train a model using algorithms like Naive Bayes, Logistic Regression, or Transformers (e.g., BERT).

   Evaluation: Use metrics like accuracy, precision, recall, and F1-score.

   Deployment: Integrate the model for real-time email classification.

   Continuous Improvement: Update the model regularly to adapt to new spam trends.

3. **Algorithm:**

```
Random Forest Classifier
Confusion Matrix:
[[964   1]
 [ 29 121]]
Accuracy:  0.9730941704035875
--------------------------------
Decision Tree Classifier
Confusion Matrix:
[[959   6]
 [ 22 128]]
Accuracy:  0.9748878923766816
--------------------------------
Multinomial Naïve Bayes
Confusion Matrix:
[[955  10]
 [ 10 140]]
Accuracy:  0.9820627802690582
```

4. **Opinion:**
   1.It is the best for spam detection due to its simplicity and efficiency with text data
   2. High efficiency with sparse data
   3. Works well with Imbalanced data