**Project Report**

**Spam Email Classifier**

# Introduction

**The Spam Email Classifier project aims to automate the detection of spam emails based on email content using machine learning techniques. By analyzing patterns in text data, the system can distinguish between legitimate (not spam) and spam messages.**

## Objectives

**- Develop a machine learning model to classify emails.  
- Use text preprocessing and feature extraction.  
- Evaluate model accuracy with real-world-like data.  
- Enable prediction on custom input.**

# Tools and Technologies

**- Python 3.x  
- Pandas  
- scikit-learn  
- Jupyter Notebook  
- Matplotlib  
- Seaborn  
- VS Code**

# Methodology

**- Data Collection:  
 A sample dataset of 100 synthetic emails labeled as spam or not spam.  
  
- Preprocessing:  
 Using TF-IDF vectorization to convert text into numerical features.  
  
- Modeling:  
 A Naive Bayes classifier was trained on 80% of the data and tested on 20%.  
  
- Evaluation:  
 Classification report and confusion matrix used to assess performance.  
  
- Prediction:  
 A CLI script allows prediction of custom email text.**

# Dataset Description

**- Total Records: 100  
- Classes: Spam (1), Not spam (0)  
- Features: Email text content**

# Results

**The classifier demonstrated:  
- High accuracy in distinguishing spam.  
- Precise detection of common spam phrases like "Congratulations," "free iPhone," "click here," etc.  
  
Sample output:  
Accuracy: 1.00  
Classification Report:  
 precision recall f1-score support  
  
 Not spam 1.00 1.00 1.00 10  
 Spam 1.00 1.00 1.00 10  
  
 accuracy 1.00 20**

# Conclusion

**The Spam Email Classifier effectively identifies spam messages using Naive Bayes. It is a scalable solution that can be integrated with email clients or expanded with larger datasets.**

# Future Scope

**- Connect to live email servers via IMAP.  
- Train on real-world email corpora.  
- Deploy a REST API for real-time predictions.  
- Explore advanced models (SVM, BERT).**

**Submitted By: Abhishek Singh**

**Company: Codec Technologies**