Spam Email Detection using Logistic Regression — Project Report

Project: Spam-Email-Detection
Link: Mayurroro/Student-Performance-Classification
Author: Mayuresh Thorve

1. Introduction

Email spam is one of the most common forms of unsolicited digital communication. Spam emails not only waste user time but can also pose security risks through phishing and malicious links. In this project, a machine learning model using Logistic Regression is developed to classify emails as spam or not spam based on their textual content. The objective is to demonstrate a simple and effective binary text classification system using Natural Language Processing techniques.

2. Dataset Summary

Kaggle Dataset: https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset

DataSet File: spam.csv

Dataset Source: SMS Spam Collection Dataset. The dataset consists of labeled text messages categorized as Spam or Ham (Not Spam). Text preprocessing includes lowercasing, punctuation removal, tokenization, stop-word removal, and TF-IDF vectorization.

Attribute	Description
Number of samples	Labeled text messages
Feature type	TF-IDF numerical text vectors
Target variable	Spam (1) / Not Spam (0)
Missing values	No

Libraries Used

The following Python libraries were used in the implementation:

- **pandas** – for data loading and manipulation
- **numpy** – for numerical operations
- **nltk** – for text preprocessing (stopwords removal, stemming)
- **string** – for handling punctuation removal
- **scikit-learn (sklearn)**:
  - CountVectorizer – to convert text into numerical feature vectors
  - LogisticRegression – to build the classification model
  - train_test_split – to split data into training and testing sets
  - accuracy_score, precision_score, recall_score, f1_score
  - confusion_matrix, classification_report – for model evaluation
- **matplotlib & seaborn** – for data visualization and confusion matrix plotting

3. Model Performance

The Logistic Regression model was trained using an 80-20 train-test split. TF-IDF vectors were used as input features. The model achieves high accuracy and balanced precision and recall, making it suitable for spam email detection tasks.

Model	Accuracy (%)	F1-score
Logistic Regression	95	0.95

Evaluation Summary

The Logistic Regression model was evaluated on the test dataset using standard classification metrics.

Evaluation Metrics Used:

- **Accuracy**
- **Precision**
- **Recall**
- **F1 Score**
- **Confusion Matrix**

Observations:

- The model achieved **high accuracy**, indicating effective overall classification.
- **High recall for spam class** ensures that most spam messages are correctly identified.
- **F1 Score** demonstrates a good balance between precision and recall, which is crucial in spam detection tasks.
- The **confusion matrix** visualization helps understand false positives and false negatives

4. Feature Importance

In Logistic Regression, feature importance is determined by the learned coefficients. Words commonly associated with spam such as 'free', 'win', 'offer', and 'click' have strong positive influence on spam prediction. TF-IDF helps reduce bias from frequently occurring but less informative words.

5. Conclusion

This project demonstrates that spam email detection can be effectively implemented using Logistic Regression combined with TF-IDF text representation. The model is simple, interpretable, and achieves strong performance. Future enhancements may include n-gram features, ensemble models, or deep learning approache.