ADDIS ABABA SCIENCE AND TECHNOLOGY UNIVERSITY (AASTU)

COLLEGE OF ELECTRICAL AND MECHANICAL ENGINEERING

DEPARTMENT OF SOFTWARE ENGINEERING

INTRODUCTION TO MACHINE LEARNING (SWEG 4112)

EMAIL SPAM DETECTION REPORT

SECTION: E

	Group Members	ID Number
1.	Samuel Kendie	ETS 1413/14
2.	Samuel Tarekegn	ETS 1420/14
3.	Selamawit Elias	ETS 1452/14
4.	Semira Hussien	ETS 1465/14
5.	Sifen Mekonnen	ETS 1474/14
6.	Yasub Demissie	ETS 1632/14
7.	Yeabtsega Tesfaye	ETS 1660/14

SUBMITTED TO: Instructor Tesfaye

SUBMISSION DATE: Tuesday, April 15, 2025

Email Spam Detection Report

Introduction

This report outlines the implementation and evaluation of three machine learning models- Decision Tree, Random Forest, and Logistic Regression- for detecting spam emails. Using the "spam.csv" dataset, the workflow includes data preprocessing, feature engineering, model training, evaluation, and visualization of results.

Data Overview

The dataset used for this analysis contains labeled email messages categorized as "ham" (non-spam) or "spam." Key columns include:

- **label**: Indicates whether the message is spam (1) or ham (0).
- **message**: The text content of the email.
- **feature_1, feature_2, feature_3**: Additional features extracted from the dataset.

Basic preprocessing steps include:

- Handling missing values by replacing them with empty strings.
- Combining text features into a single column for vectorization.

Data Processing

1. Feature Engineering:
    - Text features (message + additional columns) are combined into a single column called combined_features.
    - CountVectorizer is used to convert text data into numerical feature vectors.
2. Train-Test Split:
    - The dataset is split into training (80%) and testing (20%) sets to evaluate model performance.

Model Selection

Three classification models were employed:

1. Decision Tree Classifier:
    - A simple tree-based model that splits data based on feature thresholds.
    - Trained using default parameters with a random state for reproducibility.

2. Random Forest Classifier:
    - An ensemble method using 100 decision trees for improved accuracy and reduced overfitting.
    - Features are aggregated using majority voting.
3. Logistic Regression:
    - A linear model that predicts probabilities for binary classification tasks.
    - Suitable for datasets where relationships are linear.

Model Training and Evaluation

1. Training:
    - Each model was trained on the training set using its respective algorithm.
2. Evaluation Metrics:
    - **Accuracy**: Measures overall prediction correctness.
    - **Classification Report**: Provides precision (how many predicted spam emails were spam), recall (how many actual spam emails were correctly identified) , F1-score (balances precision and recall using harmonic mean) for each class.
    - **Confusion Matrix**: Visualizes true positives, false positives, true negatives, and false negatives.

Results Comparison

Classification Model	Accuracy	Precision (Spam)	Recall (Spam)	F1-Score (Spam)
Decision Tree	96.77%	91%	85%	88%
Random Forest	97.40%	100%	81%	89%
Logistic Regression	97.85%	99%	85%	91%

Visualization

Confusion matrices were plotted using Seaborn's heatmap functionality to highlight model performance in classifying ham and spam emails.

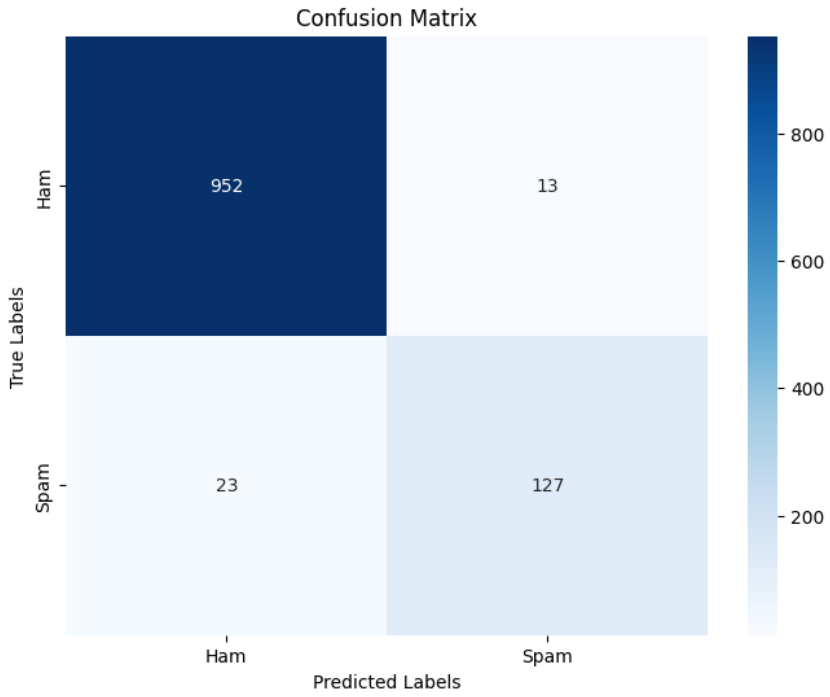

Figure 1: Decision Tree Confusion Matrix

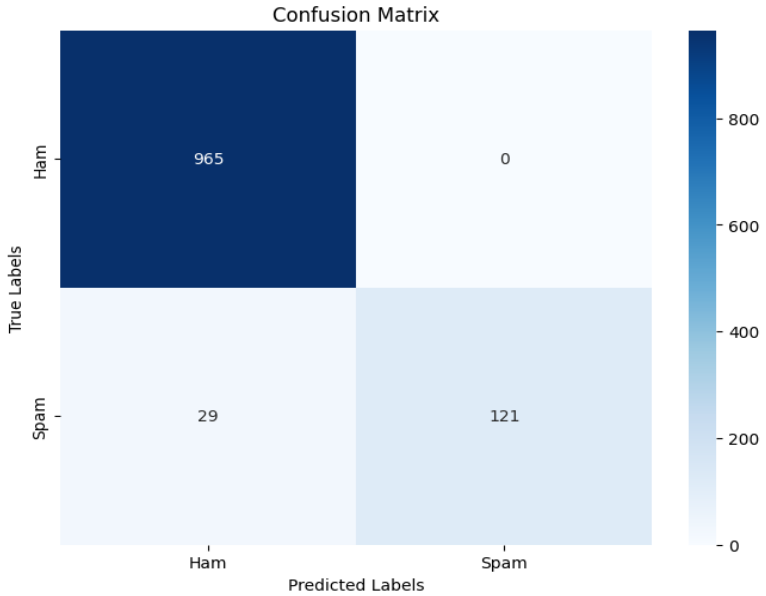

Figure 2: Random Forest Confusion Matrix

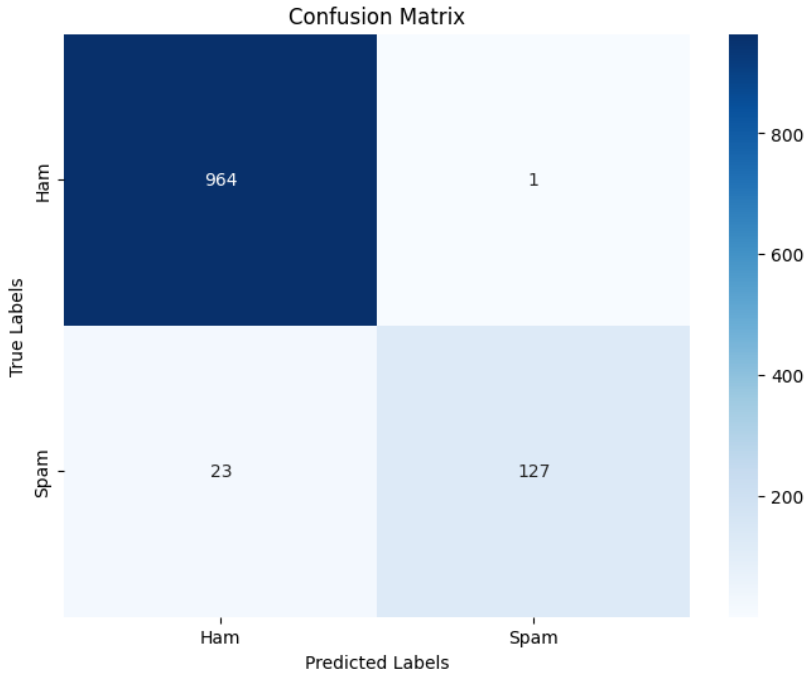

Figure 3: Logistic Regression Confusion Matrix

Conclusion

As observed from the comparison table, the Logistic Regression model performed the best with an accuracy of 97.85% and F1-score of 91%. As for the precision – the number of predicted spam emails and those that were spam, Random Forest achieved a 100% precision making it a better model than Logistic Regression. Overall, ensemble methods like Random Forest generally outperform individual classifiers due to reduced variance.

Reference

[1] “Email Spam Detection Dataset (classification),” www.kaggle.com. https://www.kaggle.com/datasets/shantanudhakadd/email-spam-detection-dataset-classification