Sri Sivasubramaniya Nadar College of Engineering, Chennai
(An Autonomous Institution affiliated to Anna University)

Degree & Branch	B.E. Computer Science & Engineering	Semester	VI
Subject Code & Name	UCS2612 – Machine Learning Algorithms Laboratory		
Academic Year	2025–2026 (Even)	Batch: 2023–2027	**Due Date: 27/1/26**

Experiment 4: Spam Email Classification using Logistic Regression and Support Vector Machines

Name: Muralisekar Janissha
Reg. No: 3122235001058
Class: CSE-B

---

1. Aim and Objective

Aim: To classify emails as spam or non-spam using Logistic Regression and Support Vector Machine models and analyze their performance.

Objectives:

- To build a binary classification model for spam detection.
- To study the effect of regularization in Logistic Regression.
- To evaluate different kernel functions in SVM.
- To compare models based on accuracy and generalization ability.

2. Dataset Description

The experiment uses the **Spambase dataset**, which contains numerical features extracted from email messages.

- Number of Features: 57
- Target Variable: Email Class
- Class Labels:
  - 1 – Spam

  - 0 – Non-Spam (Ham)
- Data Type: Fully numerical

The dataset is suitable for supervised binary classification tasks.

3. Preprocessing Steps

- Checked and handled missing values.
- Feature scaling was performed using standardization.
- Dataset was split into training and testing sets.
- Stratified sampling ensured balanced class distribution.

Preprocessing ensures stable model convergence and fair evaluation.

4. Implementation Details

The experiment was implemented using Python with the following libraries:

- NumPy and Pandas for data processing
- Scikit-learn for model building and evaluation
- Matplotlib and Seaborn for analysis and plotting

The following models were implemented:

- Logistic Regression with L1 and L2 regularization
- Support Vector Machine with Linear, Polynomial, RBF, and Sigmoid kernels

Hyperparameter tuning was performed using Grid Search and Cross-Validation.

5. Visualizations

This section presents visual analysis to understand data distribution, feature behavior, model performance, and learning characteristics of Logistic Regression and SVM classifiers.

5.1 Class Distribution

The class distribution plot shows the number of instances belonging to each class. It helps identify class imbalance in the dataset.

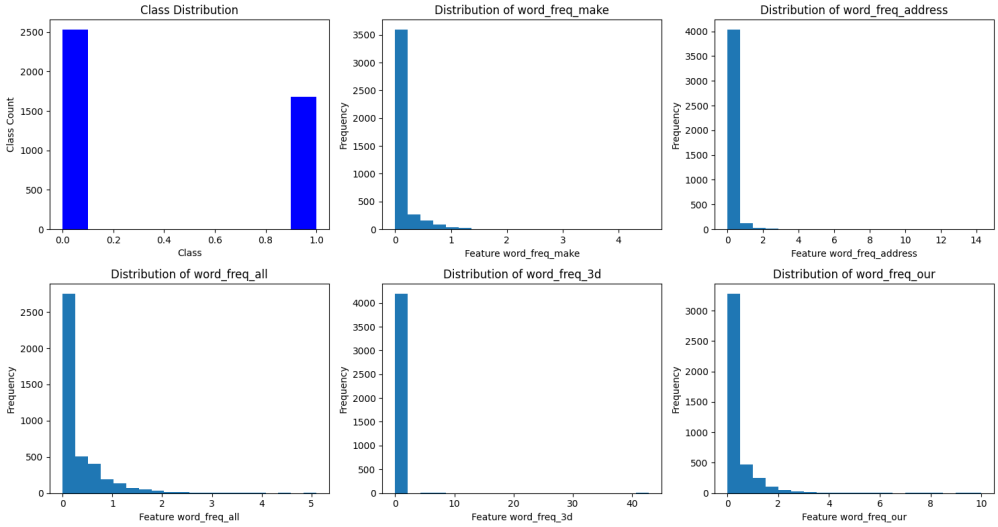

Figure 1: Class Distribution

5.2 Feature Distribution Analysis

Histograms were used to study the distribution of important features. Most features show right-skewed distributions, indicating sparsity in word frequencies.

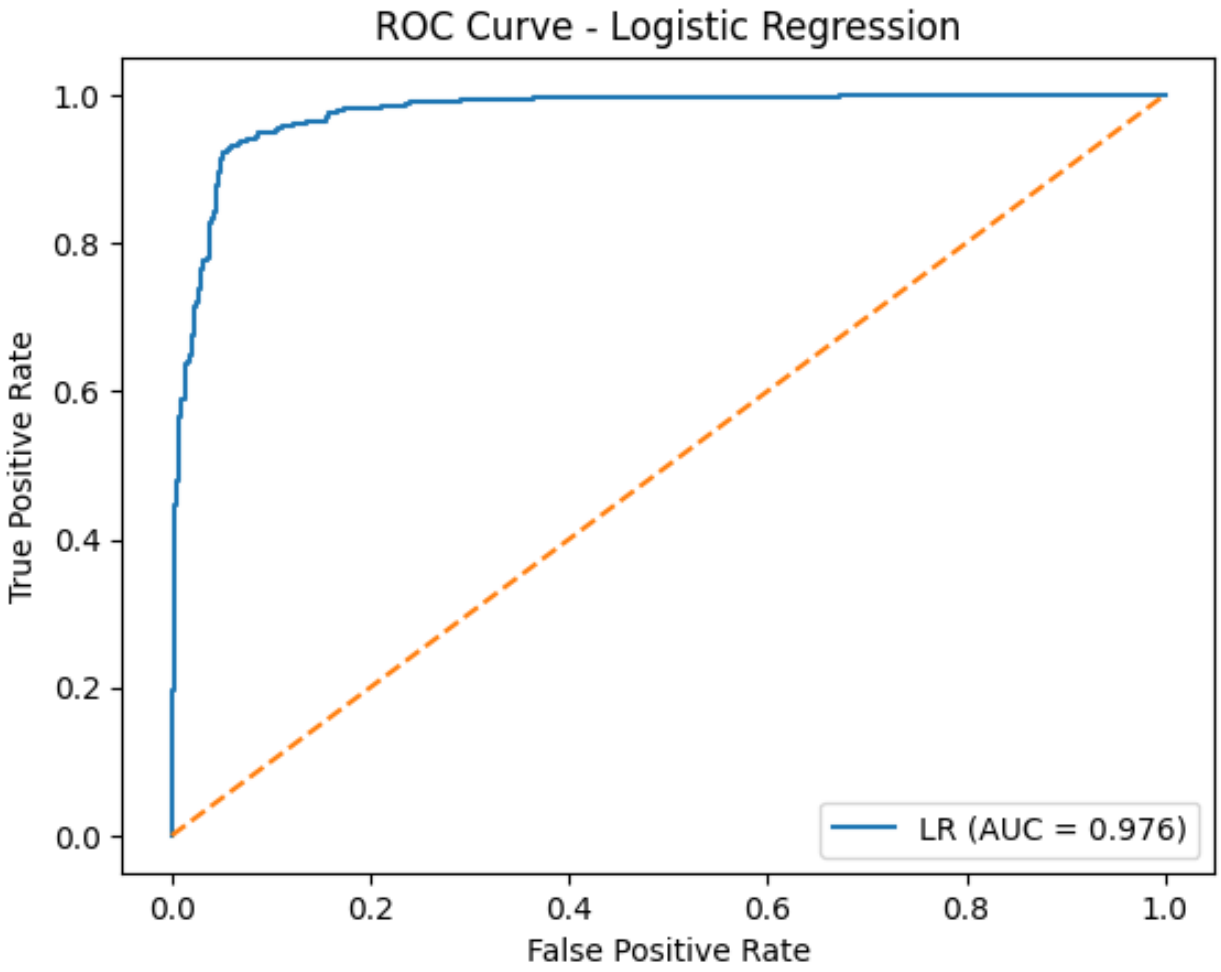

Figure 2: Distribution of Selected Word Frequency Features

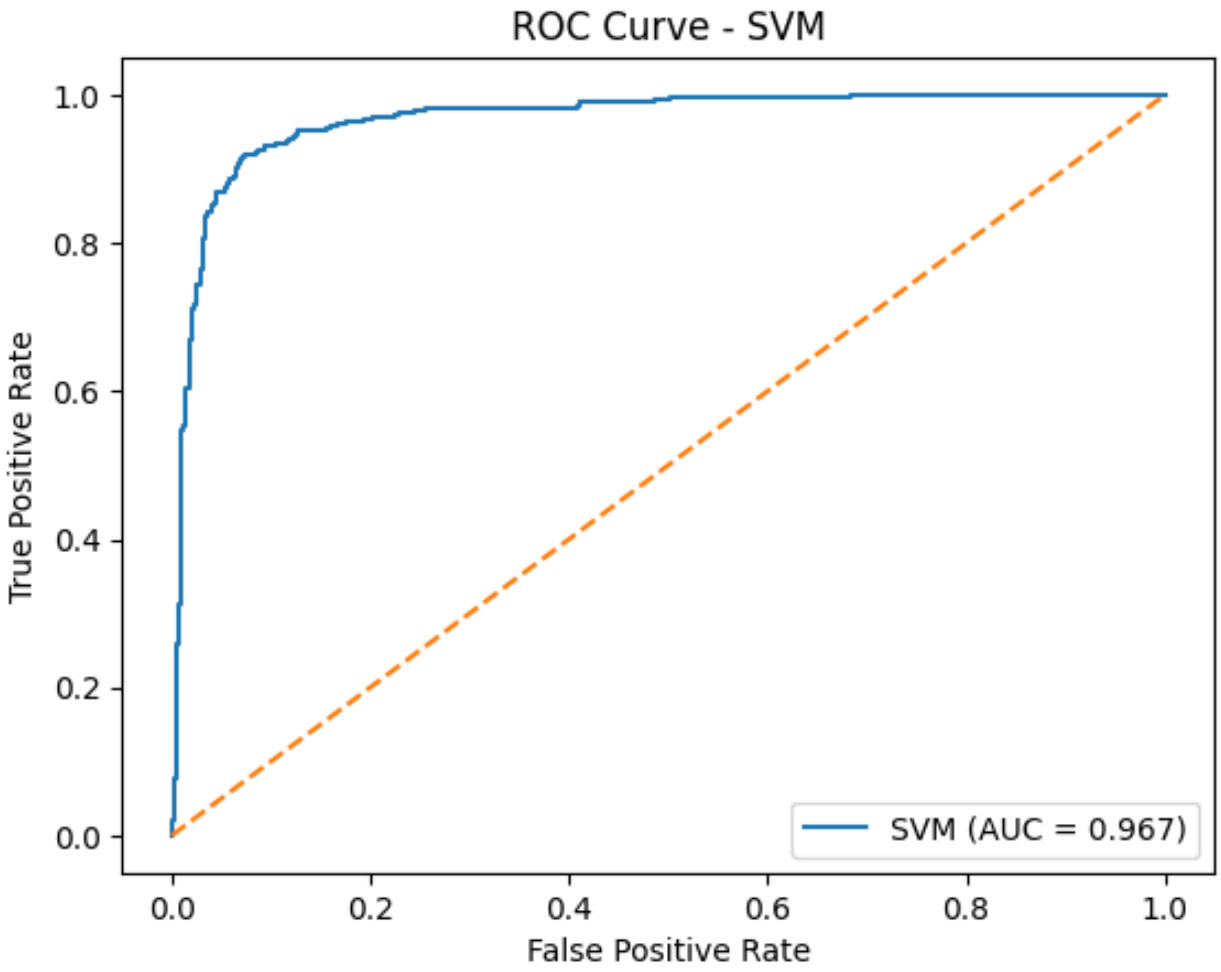

Figure 3: Distribution of Additional Word Frequency Features

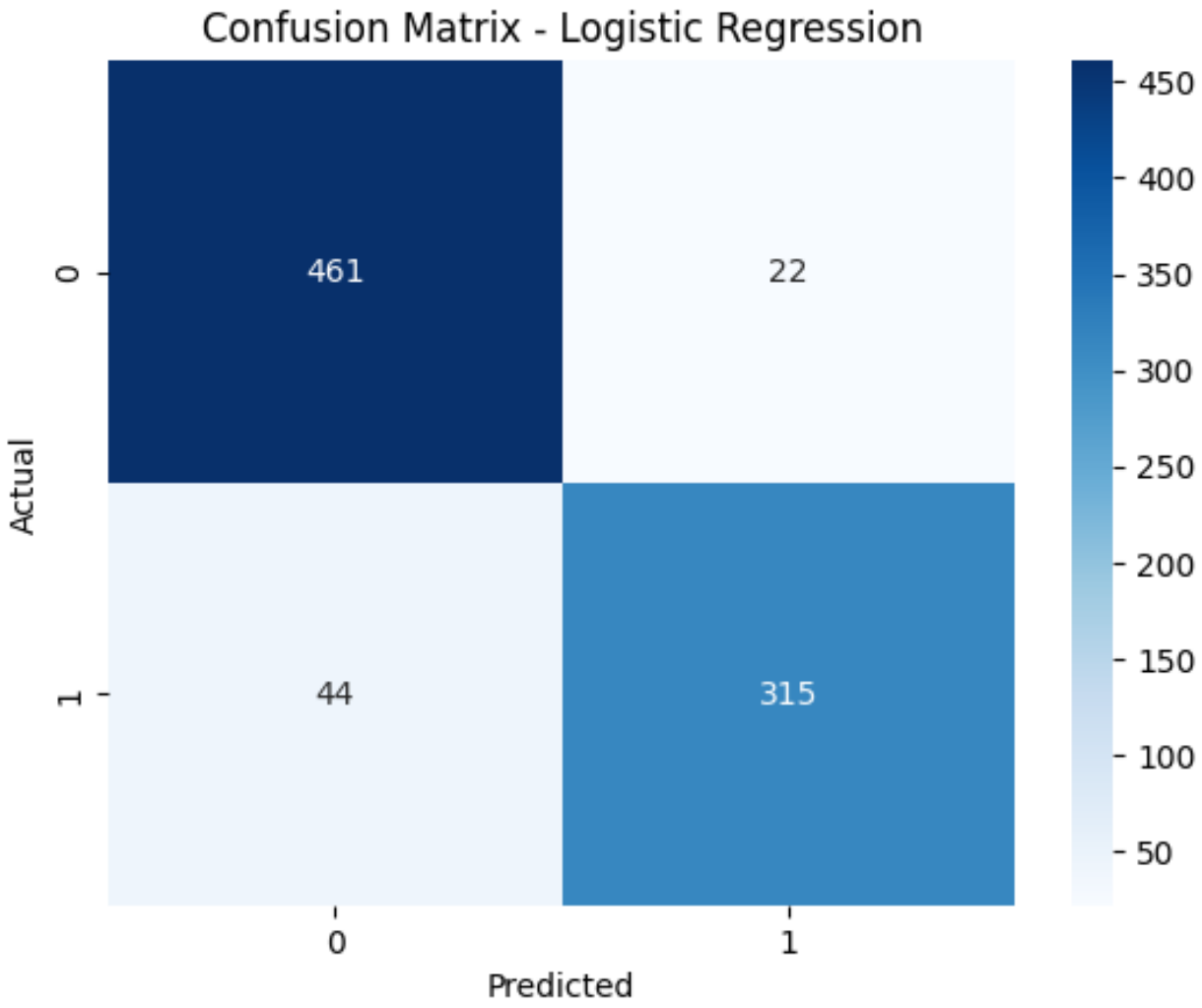

Figure 4: Distribution of word_freq_all

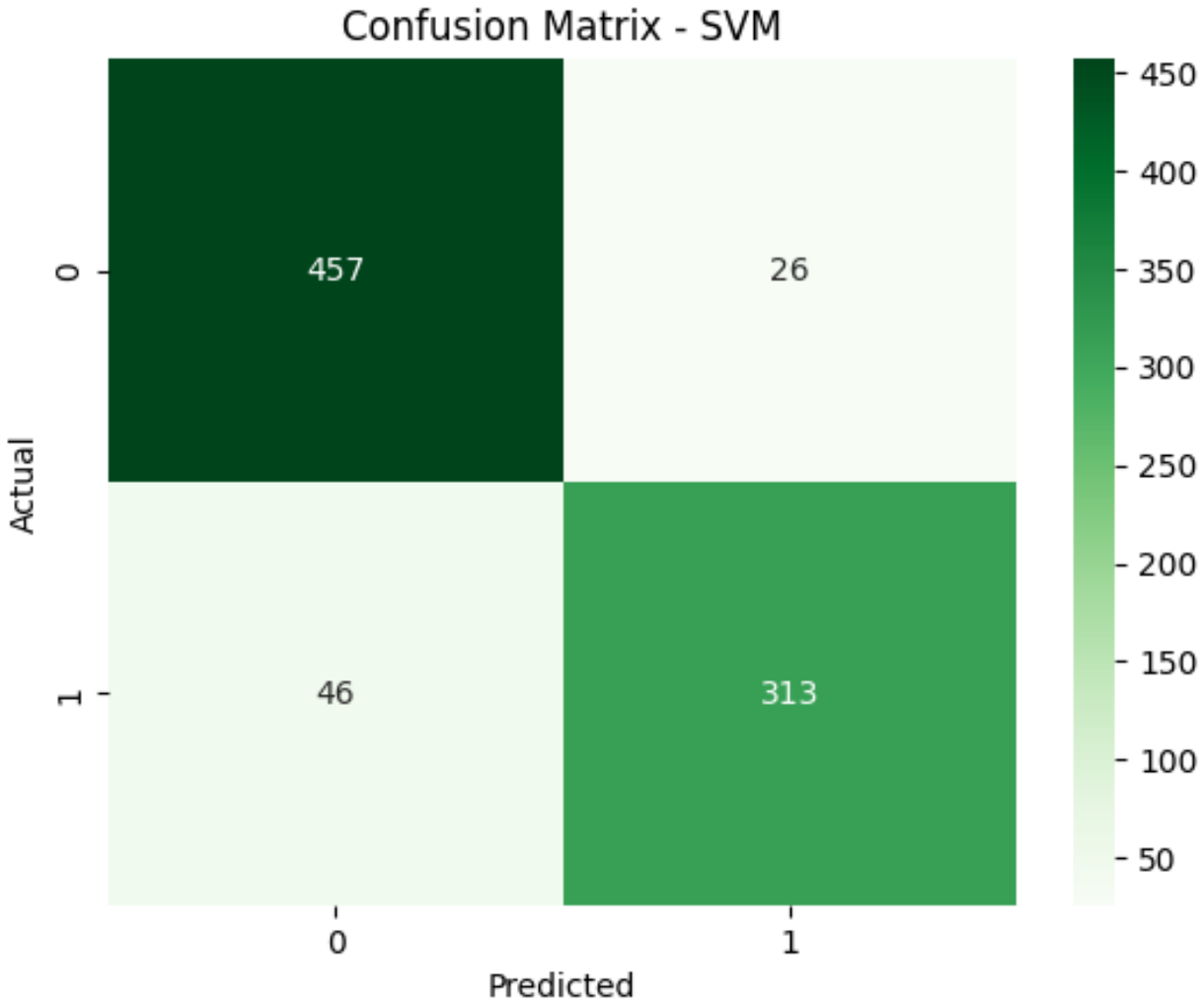

Figure 5: Distribution of word_freq_3d

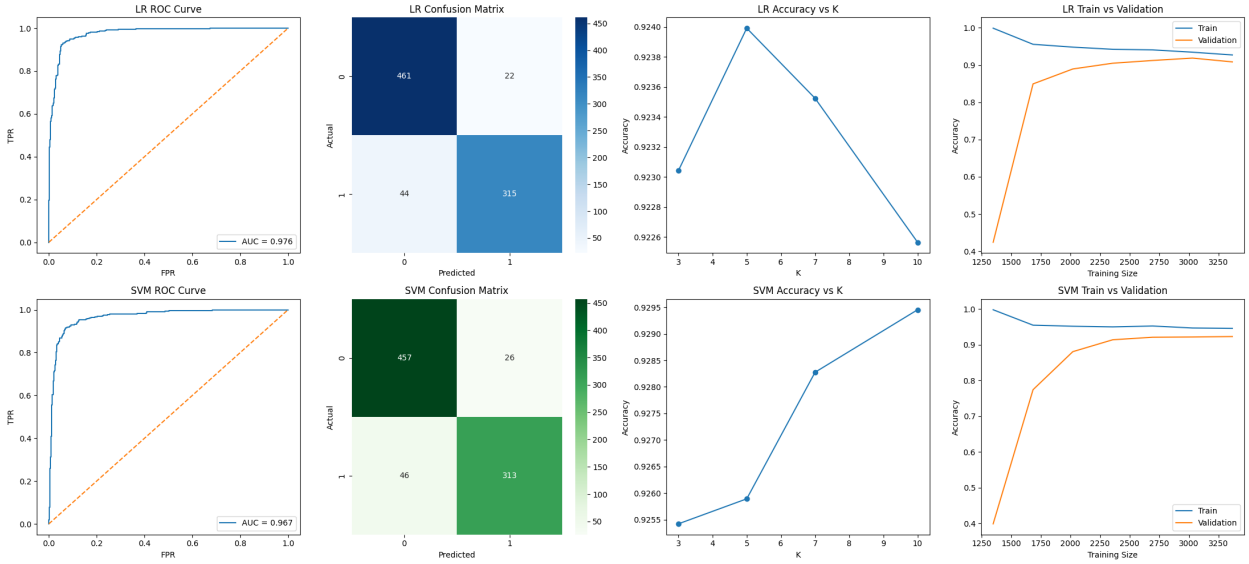

Figure 6: Distribution of word_freq_our

5.3 ROC Curve Analysis

ROC curves were plotted to evaluate the classification capability of both models. Both models achieve high AUC values, indicating strong discriminative performance.

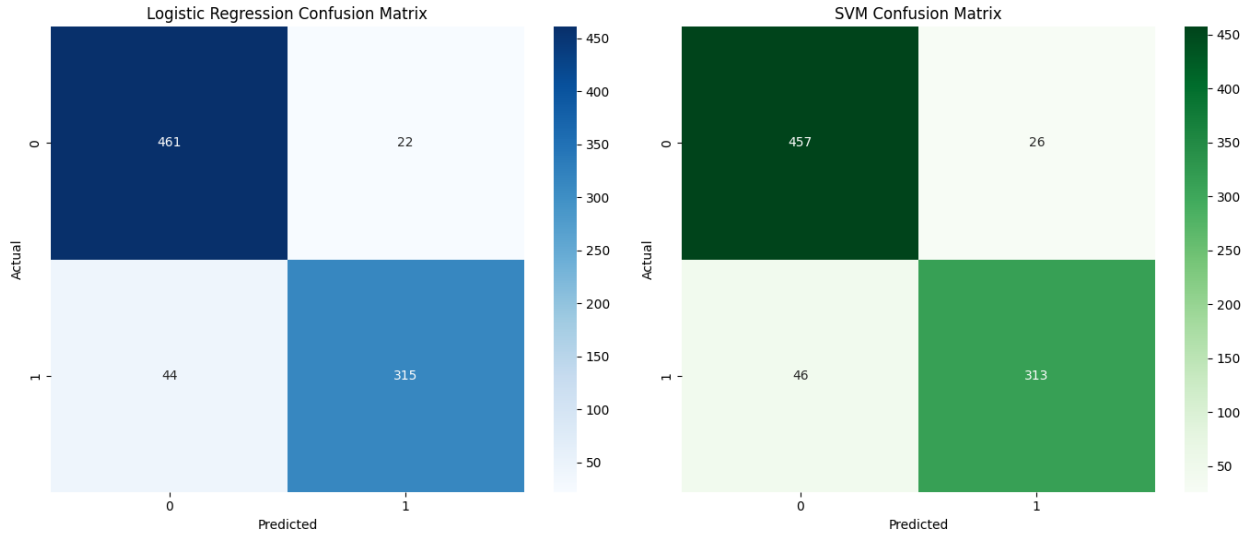

Figure 7: ROC Curve – Logistic Regression

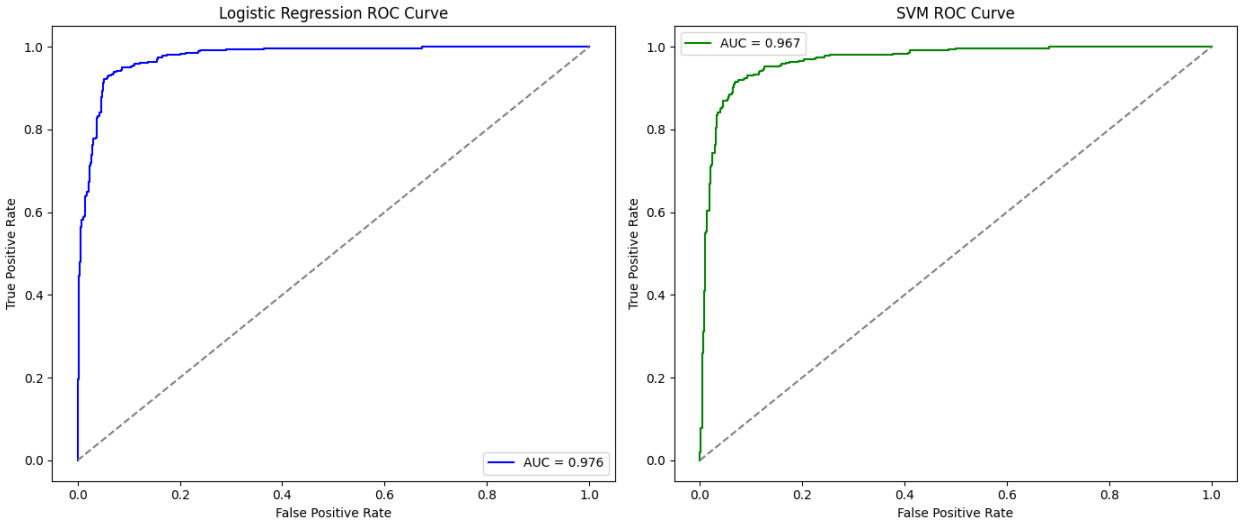

Figure 8: ROC Curve – Support Vector Machine

5.4 Confusion Matrix Analysis

Confusion matrices provide insight into true positives, true negatives, false positives, and false negatives for each classifier.

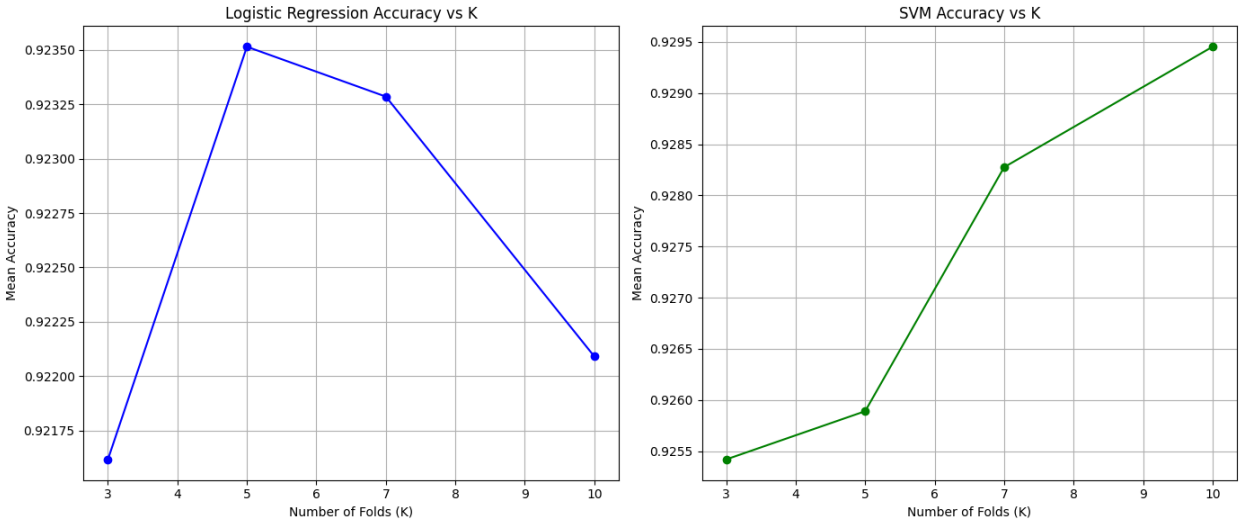

Figure 9: Confusion Matrix – Logistic Regression

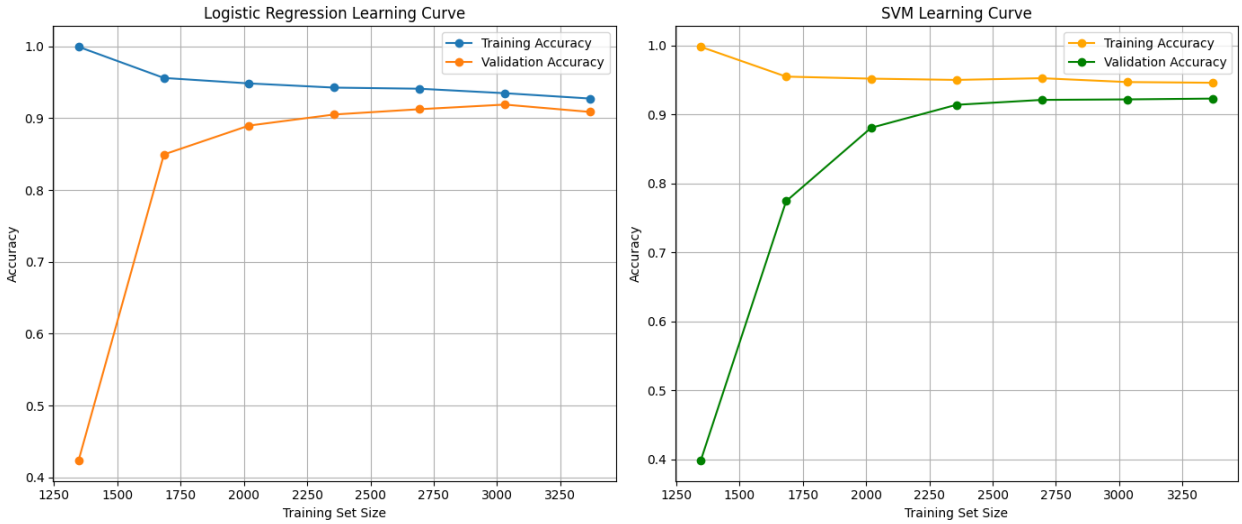

Figure 10: Confusion Matrix – Support Vector Machine

5.5 Learning Curve Analysis

Learning curves illustrate the relationship between training size and model accuracy. They help analyze overfitting and underfitting behavior.

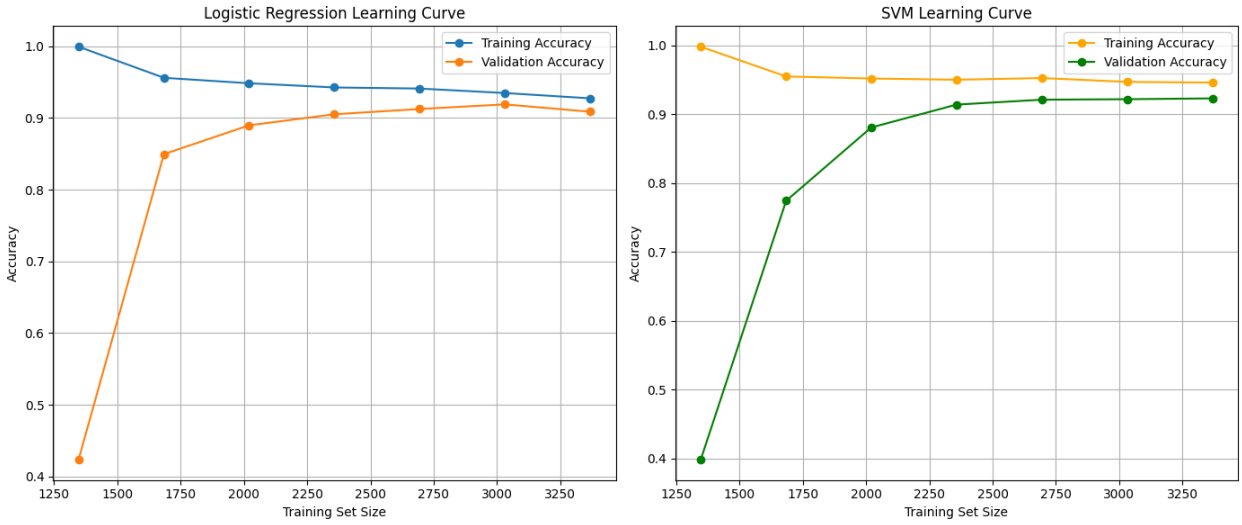

Figure 11: Learning Curves for Logistic Regression and SVM

6. Performance Tables

6.1 Hyperparameter Tuning Results

Model	Search Method	Best Parameters	Best CV
Logistic Regression	Grid Search	C = 10, penalty = L2, solver = liblinear	0.
SVM	Grid Search	C = 10, degree = 2, gamma = scale, kernel = RBF	0.

Table 1: Hyperparameter Tuning Summary

6.2 Logistic Regression Performance

Metric	Value
Accuracy	0.9216
Precision	0.9347
Recall	0.8774
F1 Score	0.9052
Training Time (s)	0.1043

Table 2: Logistic Regression Performance Metrics

6.3 SVM Kernel-wise Performance

Kernel	Accuracy	F1 Score	Training Time (s)
Linear	0.9238	0.8975	2.112
Polynomial	0.7553	0.6128	15.680
RBF	0.9143	0.8968	0.7940
Sigmoid	0.8789	0.8555	0.8132

Table 3: SVM Kernel-wise Performance Comparison

6.4 K-Fold Cross-Validation Results (K = 5)

Fold	Logistic Regression	SVM
Fold 1	0.9211	0.8613
Fold 2	0.9299	0.8789
Fold 3	0.9298	0.8824
Fold 4	0.9235	0.8895
Fold 5	0.9248	0.8824
Average	**0.9246**	**0.8800**

Table 4: K-Fold Cross-Validation Accuracy Comparison

7. Overfitting and Underfitting Analysis

- Logistic Regression shows stable learning behavior with proper regularization.
- SVM models with complex kernels may overfit if hyperparameters are not tuned.
- Cross-validation helps identify optimal model complexity.

8. Bias–Variance Analysis

- Logistic Regression exhibits higher bias due to linear decision boundaries.
- SVM with RBF kernel reduces bias but increases variance.
- Proper kernel selection balances the bias–variance trade-off.

9. Observations and Conclusion

Observations:

- Regularization improves Logistic Regression performance.
- SVM with RBF kernel achieves higher accuracy.
- Feature scaling significantly affects SVM performance.

Conclusion: This experiment demonstrates that both Logistic Regression and SVM are effective for spam classification. While Logistic Regression offers interpretability and stability, SVM provides higher accuracy when optimized with appropriate kernels and parameters.