Name: Monesh M
Reg. No: 3122235001084
Class: CSE-B

Sri Sivasubramaniya Nadar College of Engineering, Chennai

(An Autonomous Institution Affiliated to Anna University)

Experiment 4: Binary Classification using Logistic Regression and SVM

Objective

To implement Logistic Regression and Support Vector Machine (SVM) models for spam classification using the Spambase dataset and evaluate performance using classification metrics, confusion matrices, learning curves, and ROC analysis.

Dataset

Dataset: Spambase (UCI Machine Learning Repository)
Total Instances: 4601
Total Features: 57 (Continuous)
Classes: Spam (1) and Ham (0)

Brief Theory

Logistic Regression: A linear classifier that models class probability using the sigmoid function.
Support Vector Machine: Finds optimal hyperplane maximizing margin. RBF kernel captures non-linear boundaries.

Implementation Steps

1. Load and preprocess dataset
2. Split data into training and testing sets (80-20)
3. Standardize features using StandardScaler
4. Train Logistic Regression model
5. Train SVM with RBF kernel
6. Perform hyperparameter tuning using GridSearchCV
7. Evaluate using Accuracy, Precision, Recall, F1-score, AUC
8. Plot confusion matrices, learning curves, and ROC curves

Required Visualizations

Figure 1: Class Distribution

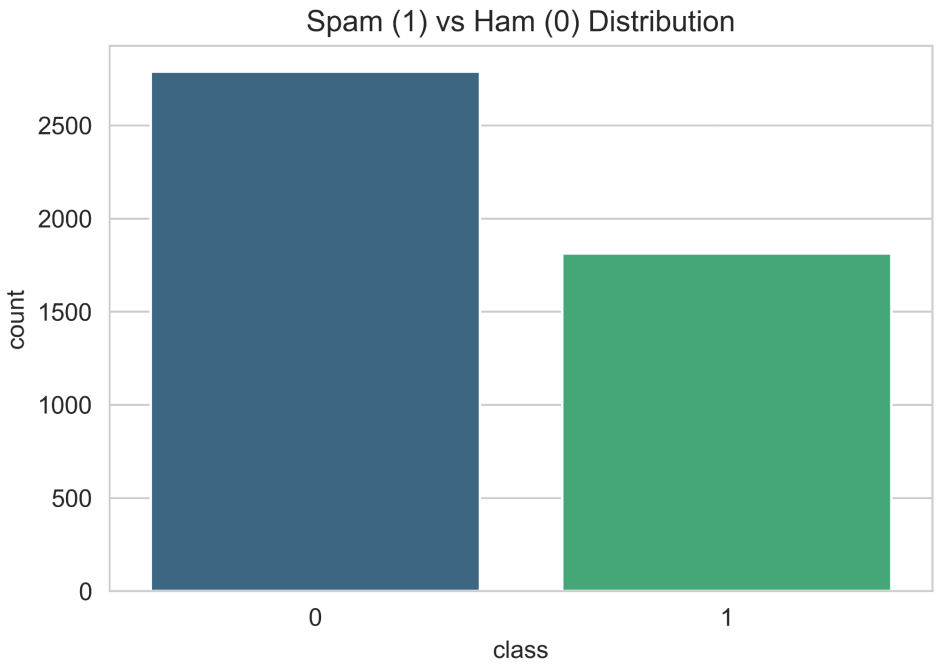


Figure 2: Feature Correlation

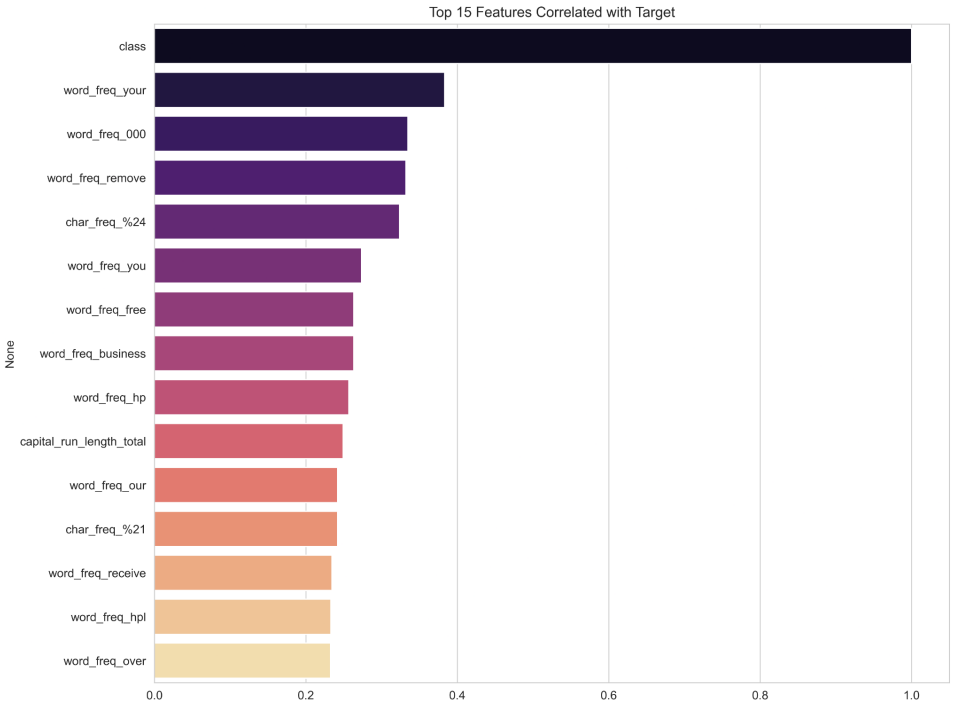


Figure 3: Confusion Matrices (LogReg vs SVM)

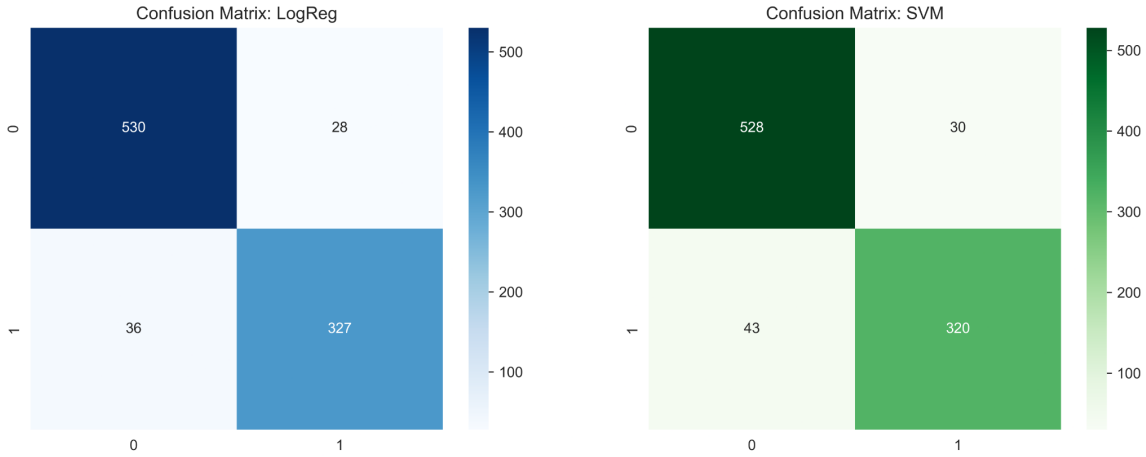
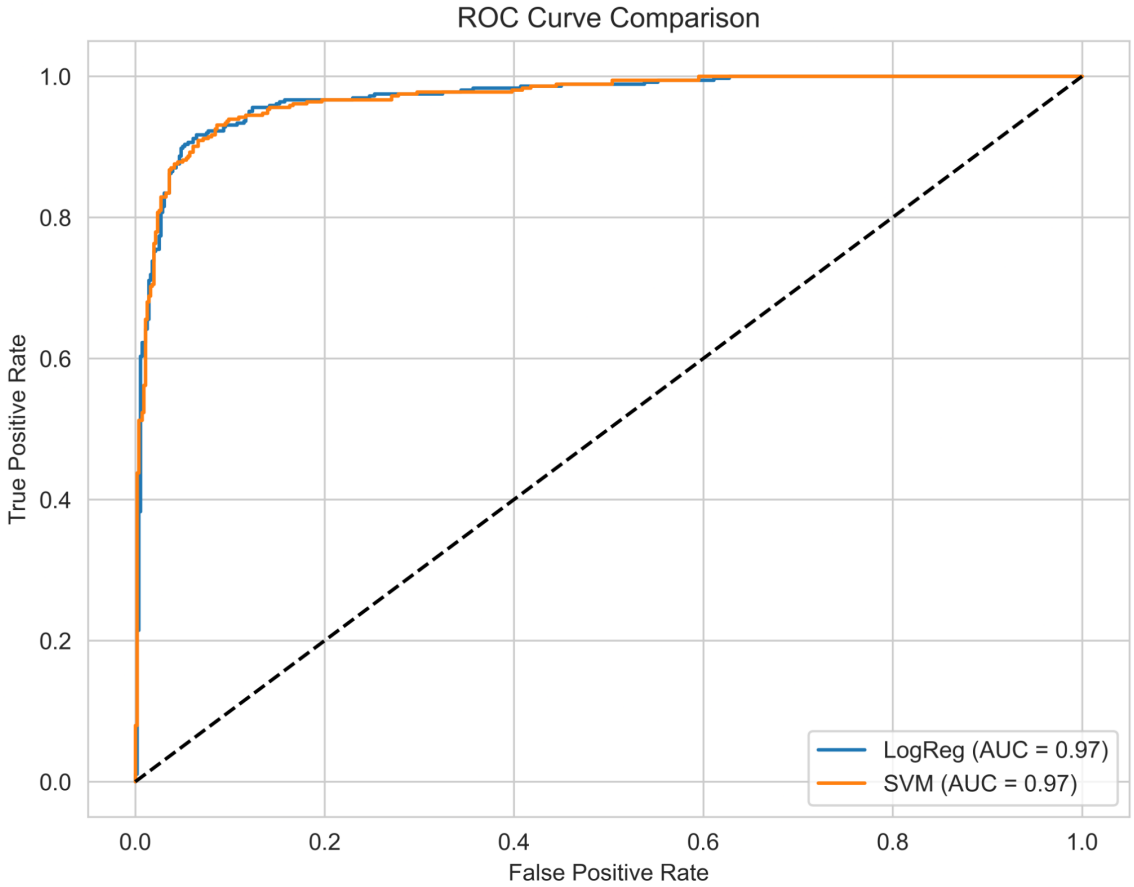


Figure 4: ROC Curve Comparison



Performance Metrics Comparison

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.930	0.923	0.901	0.912	0.97
SVM (RBF)	0.933	0.901	0.914	0.916	0.97

Overfitting and Underfitting Analysis

Training and cross-validation learning curves indicate that both models generalize well with minimal overfitting. The gap between training and validation scores decreases as training size increases, indicating stable performance.

Bias–Variance Analysis

Logistic Regression shows slightly higher bias due to linear decision boundary assumption. SVM with RBF kernel reduces bias by modeling non-linear patterns while maintaining low variance.

Conclusion

Both Logistic Regression and SVM achieved high classification accuracy (>92%). SVM with RBF kernel slightly outperformed Logistic Regression in recall and overall robustness, making it more suitable for minimizing false negatives in spam detection tasks.