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
Name	Nithin Josh Albert	Register No.	3122235001090
Due Date	26.01.2026		

Experiment 4: Binary Classification using Logistic and Kernel-Based Models

Objective

To classify emails as spam or ham using Logistic Regression and Support Vector Machine (SVM) classifiers and to analyze the effect of hyperparameter tuning on classification performance.

Dataset

The **Spambase** dataset contains numerical features extracted from email content and a binary label indicating spam or non-spam (ham).

Dataset Links (for reference):

- Kaggle: https://www.kaggle.com/datasets/somesh24/spambase

Preprocessing Steps

The following preprocessing steps were applied.

1 Missing Value Check

The dataset was examined for missing or null values. No missing values were detected.

2 Feature Standardization

All numerical features were standardized using `StandardScaler` to achieve a mean of 0 and a standard deviation of 1.

3 Train–Test Split

The dataset was partitioned into training and testing subsets using an 80:20 split.

Implementation Details

The models were implemented using the `scikit-learn` library with the following configurations and tuning strategies.

1 Logistic Regression

Logistic Regression models were trained using multiple solvers, including `liblinear` and `saga`. Both $L1$ (Lasso) and $L2$ (Ridge) regularization techniques were evaluated. The inverse regularization strength parameter C was tuned over a predefined range to identify the optimal balance between bias and variance.

2 Support Vector Machine (SVM)

Support Vector Machine classifiers were evaluated using four different kernel functions: Linear, Polynomial, Radial Basis Function (RBF), and Sigmoid. Hyperparameters tuned during model selection included the regularization parameter C (ranging from 0.1 to 100), the kernel coefficient `gamma` (with values `scale` and `auto`), and the polynomial degree for polynomial kernels.

3 Validation Strategy

A 5-Fold Cross-Validation strategy was employed during hyperparameter tuning to ensure stability and robustness of the experimental results.

Visualizations

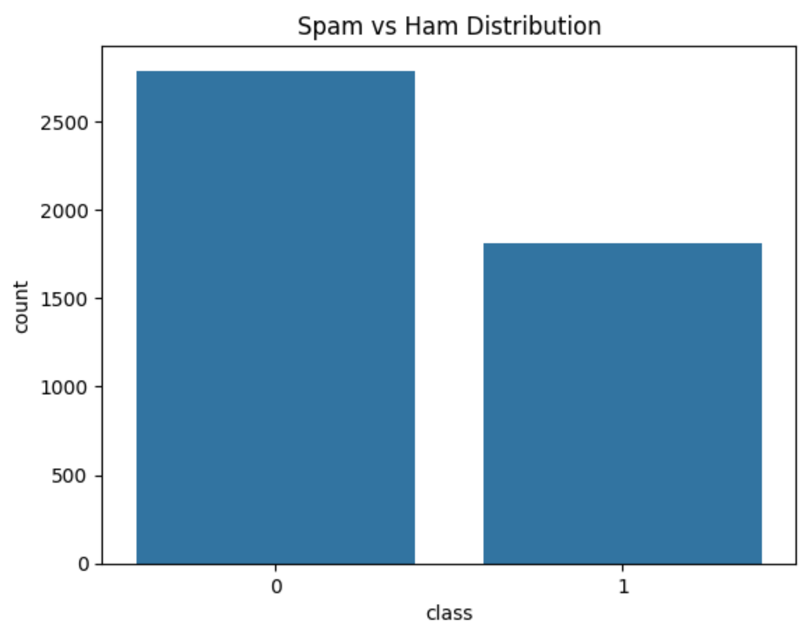

Figure 1: class Distribution

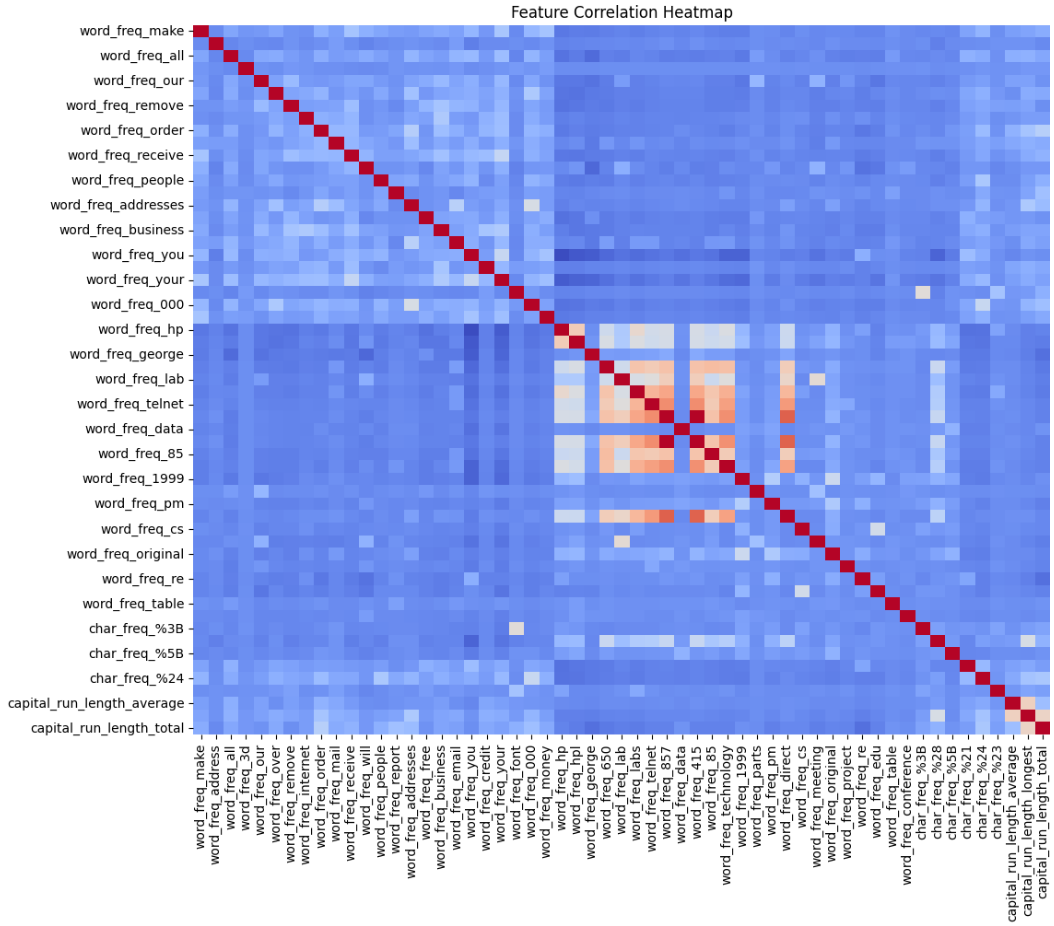

Figure 2: Correlation

Hyperparameter Tuning Results

Model	Search	Best Parameters	Best CV Accuracy
Logistic Regression	Grid	C=100,Penalty=L1,Solver=liblinear	0.9239
SVM	Grid	C=1, Gamma=scale, Kernel=RBF	0.9339

Logistic Regression Performance

Metric	Value
Accuracy	0.9294
Precision	0.9209
Recall	0.8980
F1 Score	0.9093
Training Time (s)	0.0893

SVM Kernel-wise Performance

Kernel	Accuracy	F1 Score	Training Time (s)
Linear	0.930510	0.910615	2.684362
Polynomial	0.779587	0.621974	2.672709
RBF	0.927253	0.905501	1.799805
Sigmoid	0.883822	0.852414	1.052342

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

Fold	Logistic Regression	SVM
Fold 1	0.919653	0.93268
Fold 2	0.931522	0.933696
Fold 3	0.895652	0.95
Fold 4	0.95108696	0.948913
Fold 5	0.82282609	0.85
Average	0.904148	0.923058

Comparative Analysis

Criterion	Logistic Regression	SVM
Accuracy	90.41%	92.30%
Model Complexity	Low	High
Training Time	Low	High
Interpretability	High	Low

Observations:

Best Performing Classifier: The Support Vector Machine (SVM) with the Radial Basis Function (RBF) kernel demonstrated the best performance in this experiment. It achieved a test accuracy of 92.30% and an F1 score of 0.9206, outperforming the tuned Logistic Regression model, which recorded an accuracy of 90.41% and an F1 score of 0.8979. This indicates that the margin-based optimization of SVM was more effective than the linear probabilistic decision boundary of Logistic Regression for this dataset.

Impact of Regularization: Regularization significantly influenced model performance. Logistic Regression achieved optimal results with L1 regularization and an inverse regularization strength of C = 100. L1 regularization enabled feature selection by suppressing less informative features, while the relatively high C value indicates weak regularization, allowing the model to better capture important patterns in the data.

Kernel behavior in SVM: Kernel selection played a critical role in SVM performance. The RBF kernel achieved the highest accuracy, confirming the presence of non-linear decision boundaries between spam and non-spam emails. The linear kernel also performed well, with an accuracy of 91.75%, suggesting that the data is largely linearly separable. In contrast, the polynomial kernel performed poorly, achieving only 76.44% accuracy, indicating an unsuitable feature mapping for this task.

From a bias–variance perspective, Logistic Regression exhibited higher bias due to its linear nature. The RBF-based SVM achieved a better balance, with high accuracy and close agreement between cross-validation (92.77%) and test performance (93.49%), indicating good generalization. The selected regularization parameter C = 1 provided sufficient flexibility without overfitting.

Learning Outcomes

- Understand probabilistic and margin-based classifiers.
- Apply hyperparameter tuning.
- Evaluate classification models.
- Interpret experimental results.

References

- Scikit-learn: Logistic Regression
- Scikit-learn: Support Vector Machines
- Scikit-learn: Hyperparameter Optimization
- Spambase Dataset – Kaggle
- UCI ML Repository – Spambase