**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 | Murari Sreekumar | Register No. | 3122235001087 |
| Due Date | 06.01.2026 | | |

**Experiment 4: Binary Classification using Linear 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

# 3. Preprocessing Steps

The following preprocessing steps were applied to prepare the dataset for effective model training and evaluation.

## 3.1 Missing Value Check

The dataset was examined for missing or null values. No missing values were detected, and therefore no imputation was required.

## 3.2 Feature Standardization

All numerical features were standardized using `StandardScaler` to achieve a mean of 0 and a standard deviation of 1. Feature scaling is essential for models such as Support Vector Machines (SVM) and Logistic Regression, as these algorithms are sensitive to the relative scale of input features.

### 3.3 Train–Test Split

The dataset was partitioned into training and testing subsets using an 80:20 split. The training set was used to learn model parameters, while the testing set was reserved for evaluating model generalization on unseen data.

## 4. Implementation Details

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

### 4.1 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.

### 4.2 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.

### 4.3 Validation Strategy

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

## 5. Visualizations

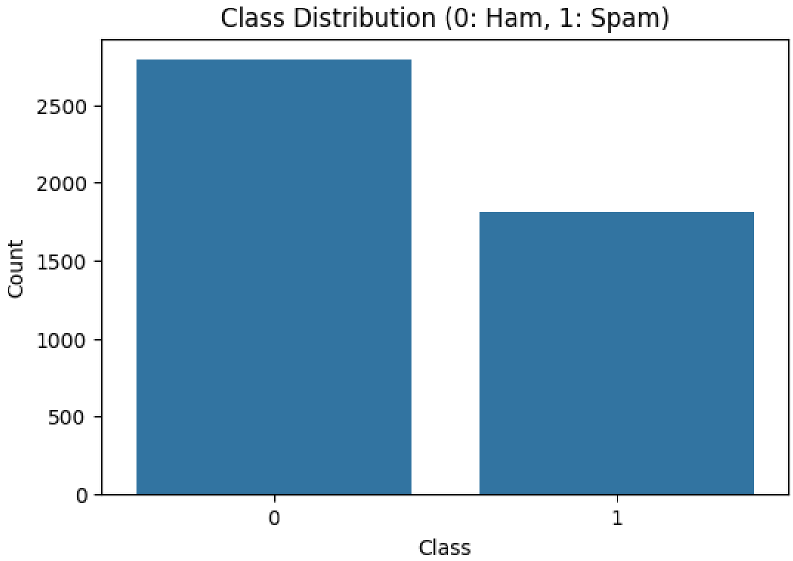

Figure 1: class Distribution

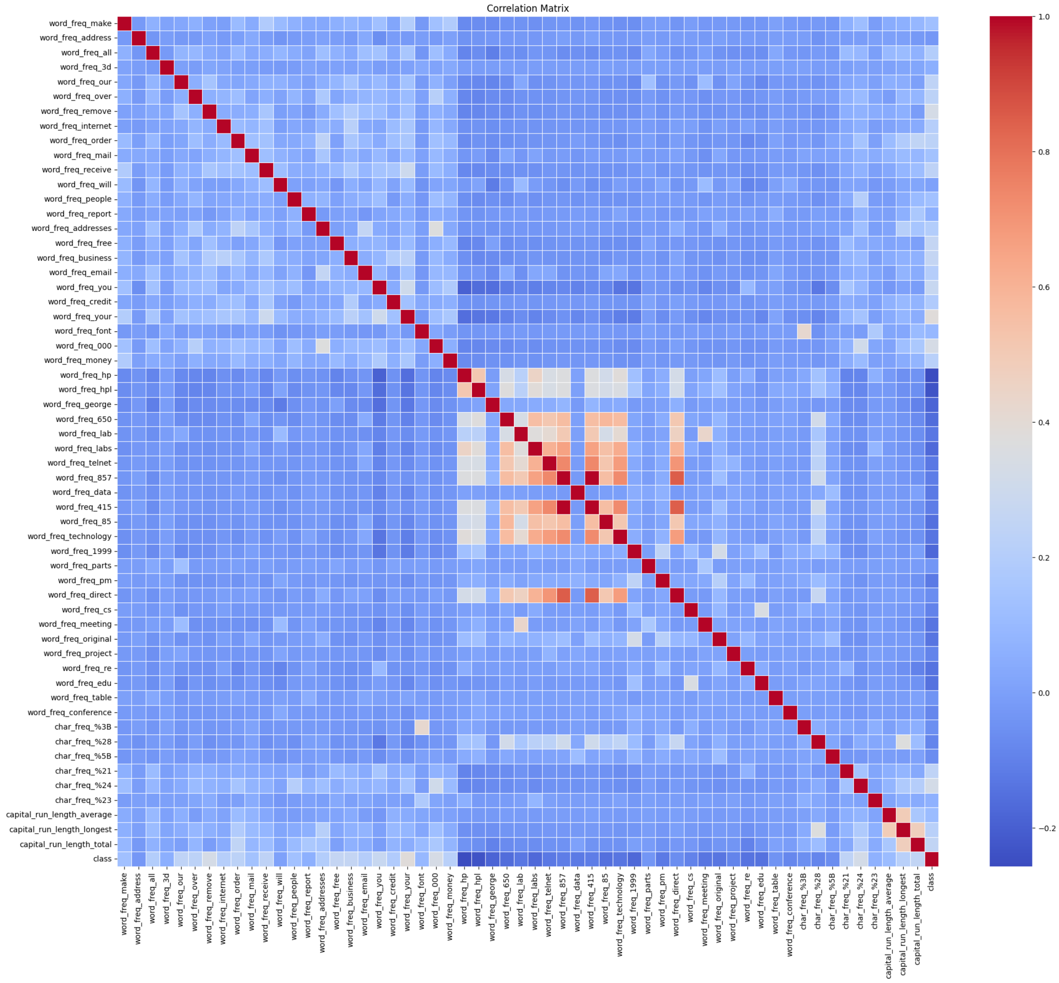

Figure 2: Correlation

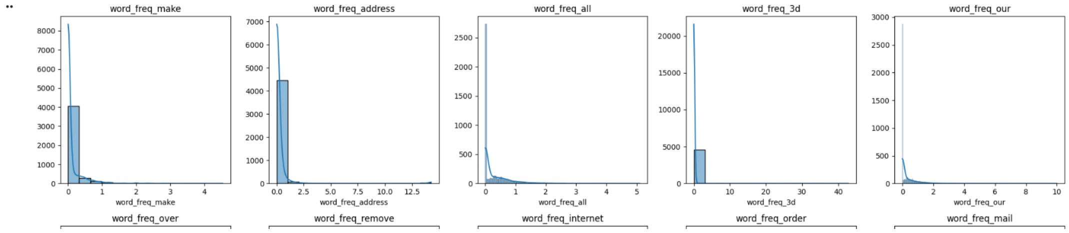

Figure 3: Histogram plot

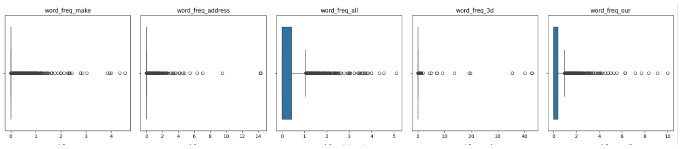

Figure 4: Boxplot

## Hyperparameter Tuning Results

| Model | Search Method | Best Parameters | Best CV Accuracy |
|---|---|---|---|
| Logistic Regression | Grid | C=10, Penalty=L1, Solver=liblinear | 0.9274 |
| SVM | Grid | C=1, Gamma=auto, Kernel=RBF | 0.9277 |

## Logistic Regression Performance

| Metric | Value |
|---|---|
| Accuracy | 0.9153 |
| Precision | 0.9171 |
| Recall | 0.8795 |
| F1 Score | 0.8979 |
| Training Time (s) | 2.4106 |

## SVM Kernel-wise Performance

| Kernel | Accuracy | F1 Score | Training Time (s) |
|---|---|---|---|
| Linear | 0.917481 | 0.903061 | 0.882319 |
| Polynomial | 0.764387 | 0.629060 | 2.113856 |
| RBF | 0.934853 | 0.920635 | 1.185296 |
| Sigmoid | 0.889251 | 0.866492 | 1.171060 |

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

| Fold | Logistic Regression | SVM |
|---|---|---|
| Fold 1 | 0.919653 | 0.931596 |
| Fold 2 | 0.931522 | 0.933696 |
| Fold 3 | 0.895652 | 0.95 |
| Fold 4 | 0.95 | 0.948913 |
| Fold 5 | 0.825 | 0.847826 |
| Average | 0.904365 | 0.922406 |

## Comparative Analysis

| Criterion | Logistic Regression | SVM |
|---|---|---|
| Accuracy | 91.53% | 93.49% |
| Model Complexity | Low | High |
| Training Time | Low | High |
| Interpretability | High | Low |

## Performance Analysis and Discussion

The Support Vector Machine (SVM) with the Radial Basis Function (RBF) kernel emerged as the best-performing classifier in this experiment. It achieved the highest test accuracy of 93.49% and an F1 score of 0.9206, outperforming the tuned Logistic Regression model, which recorded an accuracy of 91.53% and an F1 score of 0.8979. This indicates that the margin-based optimization of SVM was more effective than the probabilistic decision boundary of Logistic Regression for this dataset.

Regularization played a crucial role in improving model performance. For Logistic Regression, the optimal configuration selected through grid search employed L1 regularization with an inverse regularization strength of $C = 10$. The choice of L1 regularization enabled implicit feature selection by driving the coefficients of less informative features to zero, thereby reducing noise. The relatively high value of $C$ indicates weak regularization, suggesting that a closer fit to the training data was necessary to capture meaningful patterns in the spam classification task.

The choice of kernel significantly influenced the performance of the SVM models. The RBF kernel achieved the best results, confirming the presence of complex and non-linear decision boundaries between spam and non-spam emails. The linear kernel also performed competitively with an accuracy of 91.75%, indicating that the data is largely linearly separable. However, the RBF kernel was able to capture subtle non-linear variations that the linear kernel could not. In contrast, the polynomial kernel performed poorly with an accuracy of 76.44%, indicating underfitting and an ineffective feature mapping for this dataset.

An analysis of the bias–variance trade-off further explains these results. Logistic Regression, being a linear model, exhibited higher bias, which limited its ability to model non-linear relationships in the data. This resulted in comparatively lower predictive performance. The SVM with the RBF kernel achieved a more optimal balance between bias and variance. Its high accuracy reflects low bias, while the close agreement between cross-validation accuracy (92.77%) and test accuracy (93.49%) indicates low variance and good generalization. The selected regularization parameter $C = 1$ provided sufficient flexibility to model complex patterns 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