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: 06.01.2026**

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

(Logistic Regression vs. Support Vector Machines)

Name: Rahul V S
Reg. No: 3122235001104
Section: CSE B

---

1. Aim:
To implement and compare the performance of **Logistic Regression** (a probabilistic linear classifier) and **Support Vector Machines (SVM)** (a margin-based classifier) on the Spambase dataset. The experiment specifically analyzes the impact of regularization ($L1/L2$) and the effectiveness of different kernel functions (Linear, RBF, Polynomial, Sigmoid) in handling non-linear data distributions.

2. Libraries & Environment Setup:
The following Python libraries were utilized for data processing and model building:

- **Data Manipulation:** `Pandas`, `Numpy`.
- **Visualization:** `Matplotlib.pyplot`, `Seaborn` (Heatmaps, Boxplots).
- **Machine Learning (Scikit-Learn):**
  - *Classifiers:* `LogisticRegression`, `SVC` (Support Vector Classifier).
  - *Preprocessing:* `StandardScaler` (Z-score normalization).
  - *Model Selection:* `train_test_split`, `GridSearchCV`, `cross_val_score`.
  - *Metrics:* `accuracy_score`, `classification_report`, `confusion_matrix`.

3. Methodology and Mathematical Concepts:

1. **Data Preprocessing:** Features were standardized using `StandardScaler` to ensure that the distance-based calculations in SVM (margin maximization) were not biased by features with larger magnitudes.

$$x_{scaled} = \frac{x - \mu}{\sigma}$$

2. **Logistic Regression (Probabilistic):** Models the probability of the positive class ($y = 1$) using the logistic sigmoid function.

   [Image of sigmoid function graph]

   $$P(y = 1|X) = \frac{1}{1 + e^{-(w^T x + b)}}$$

   To prevent overfitting, Regularization parameters (C, penalty type) were tuned using Grid Search.

3. **Support Vector Machine (Margin-Based):** Constructs a hyperplane that maximizes the margin separating the two classes.

   $$\min_{w,b} \frac{1}{2}||w||^2 \quad \text{subject to} \quad y_i(w^T x_i + b) \geq 1$$

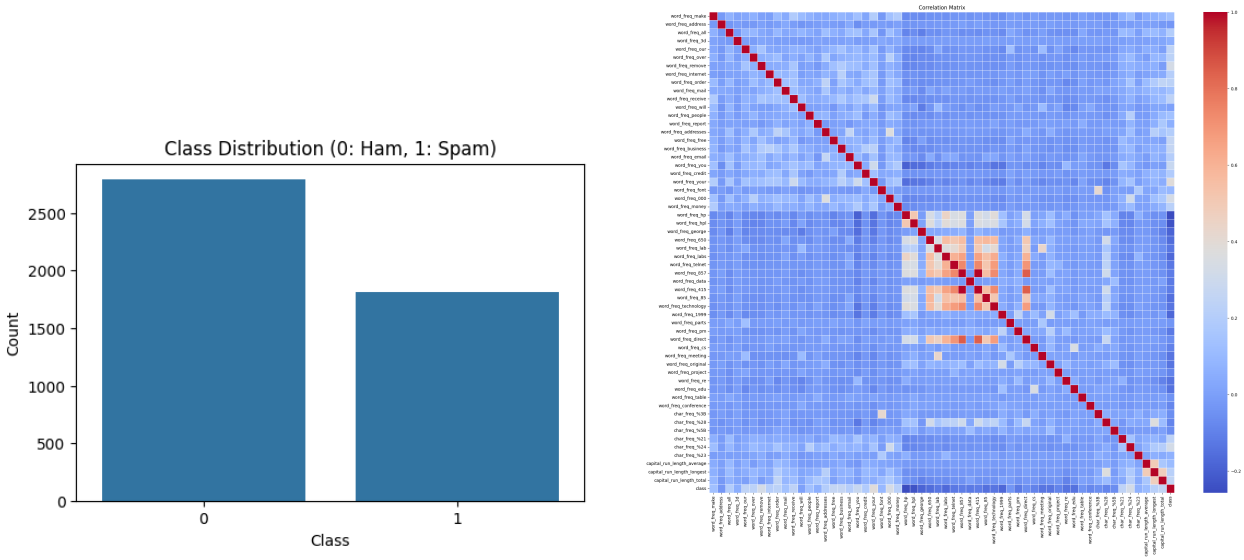
4. **Kernel Trick:** To handle non-linear separability, input vectors were mapped to high-dimensional spaces:

   - **RBF Kernel:** $K(x, x') = \exp(-\gamma||x - x'||^2)$ - Captures local proximity.
   - **Polynomial Kernel:** $K(x, x') = (\gamma\langle x, x'\rangle + r)^d$ - Models feature interactions.

4. Results and Discussions:

0.1 Exploratory Data Analysis (EDA)

The correlation matrix and histograms were analyzed to understand feature relationships and distributions. The class distribution plot confirmed the need for robust evaluation metrics like F1-Score due to potential imbalance.



(a) Class Distribution (b) Feature Correlation Matrix

Figure 1: Data Exploration

0.2 Logistic Regression Optimization

Grid Search identified the optimal hyperparameters: **C=10**, **Penalty=L1**, **Solver=liblinear**. The use of L1 regularization indicates that feature selection (sparsity) improved the model by removing irrelevant noise.

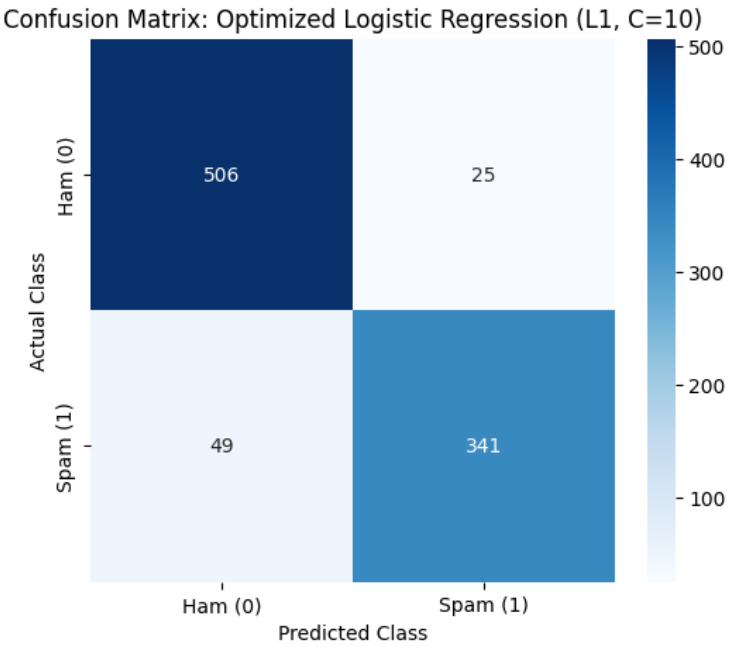


Figure 2: Confusion Matrix for Optimized Logistic Regression

0.3 SVM Kernel Performance Analysis

A comprehensive comparison of four kernel functions revealed distinct performance characteristics:

- **RBF Kernel:** Achieved the highest accuracy (**93.49%**), proving that the decision boundary is highly non-linear.
- **Linear Kernel:** Performed robustly (91.75%), suggesting the data is partially linearly separable.
- **Polynomial Kernel:** Significantly underperformed (76.44%), likely due to overfitting high-degree interactions.

Table 1: Impact of Kernel Functions on SVM Accuracy

Kernel	**Accuracy**	**F1 Score**	**Training Time (s)**
Linear	0.9175	0.9031	0.88
Polynomial	0.7644	0.6291	2.11
RBF	**0.9349**	**0.9206**	**1.18**
Sigmoid	0.8893	0.8665	1.17

0.4 Cross-Validation Robustness

5-Fold Cross-Validation was performed to assess generalization capability.

Table 2: 5-Fold Cross-Validation Scores

Fold	Logistic Regression	SVM (RBF)
Fold 1	0.9197	0.9316
Fold 2	0.9315	0.9337
Fold 3	0.8957	0.9500
Fold 4	0.9500	0.9489
Fold 5	0.8250	0.8478
Mean Accuracy	**0.9044**	**0.9224**

5. Inference and Conclusion:

The comparative analysis leads to the following conclusions:

- **Model Superiority:** The **SVM with RBF kernel** consistently outperformed Logistic Regression and other SVM kernels. This confirms that the relationship between email features (word frequencies) and the target (Spam/Ham) is complex and best modeled in a higher-dimensional space.
- **Trade-offs:**
  - *Logistic Regression:* Offers high interpretability (feature weights) and faster training, suitable for real-time applications where explaining the decision is crucial.
  - *SVM:* Provides superior accuracy but at the cost of higher training time (quadratic complexity) and lower interpretability (black-box model).
- **Regularization Effect:** The preference for L1 regularization in Logistic Regression suggests that the Spambase dataset contains redundant features that do not contribute to the classification, which L1 successfully filtered out.