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

Degree & Branch	M. Tech (Integrated) Computer Science & Engineering	Semester	V
Subject Code & Name	ICS1512 & Machine Learning Algorithms Laboratory		
Academic year	2025-2026 (Odd)	Batch:2023-2028	**Due date:**

Experiment 2: Email Spam or Ham Classification using Naïve Bayes, KNN, and SVM

Objective

To classify emails as spam or ham using three classification algorithms—Naïve Bayes, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM)—and evaluate their performance using accuracy metrics and K-Fold cross-validation.

Dataset

Download from:
Spambase –Kaggle
This dataset includes extracted features from emails, labeled as spam or ham.

Task Description

Develop models using Naïve Bayes, KNN, and SVM to classify email data. Evaluate their performance using a test split and K-Fold cross-validation, and interpret results with visualizations.

Implementation Steps

1. Load and preprocess the dataset (missing values, normalization).
2. Perform EDA (class balance, feature distributions).
3. Split into train and test sets.
4. Train the following models:
   - Naïve Bayes (Gaussian, Multinomial, Bernoulli)
   - K-Nearest Neighbors (vary k, KDTree, BallTree)
   - Support Vector Machine (Linear, Polynomial, RBF, Sigmoid kernels)
5. Evaluate using:
   - Accuracy, Precision, Recall, F1-score
   - Confusion Matrix

- ROC Curve

6. Perform K-Fold Cross Validation ($K = 5$).
7. Compare and record observations.

Naïve Bayes Variant Comparison

Table 1: Performance Comparison of Naïve Bayes Variants

Metric	**Gaussian NB**	**Multinomial NB**	**Bernoulli NB**
Accuracy			
Precision			
Recall			
F1 Score			

KNN: Varying k Values

Table 2: KNN Performance for Different k Values

k	**Accuracy**	**Precision**	**Recall**	**F1 Score**
1				
3				
5				
7				

KNN: KDTree vs BallTree

Table 3: KNN Comparison: KDTree vs BallTree

Metric	**KDTree**	**BallTree**
Accuracy		
Precision		
Recall		
F1 Score		
Training Time (s)		

Table 4: SVM Performance with Different Kernels and Parameters

Kernel	Hyperparameters	Accuracy	F1 Score	Training Time
Linear	C =			
Polynomial	C = , degree = , gamma =			
RBF	C = , gamma =			
Sigmoid	C = , gamma =			

SVM Kernel-wise Results

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

Table 5: Cross-Validation Scores for Each Model

Fold	Naïve Bayes Accuracy	KNN Accuracy	SVM Accuracy
Fold 1			
Fold 2			
Fold 3			
Fold 4			
Fold 5			
Average			

Observation Notes

- Which classifier had the best average accuracy?
- Which Naïve Bayes variant worked best?
- How did KNN accuracy vary with k and tree type?
- Which SVM kernel was most effective?
- How did hyperparameters influence performance?

Report Checklist

- Aim and Objective
- Libraries Used
- Code for All Variants and Models
- Confusion Matrix and ROC for Each
- All Comparison Tables
- Observations and Conclusions

References

- scikit-learn: Naïve Bayes
- scikit-learn: KNN
- scikit-learn: SVM
- Spambase Dataset