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

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

Experiment 3: Email Spam or Ham Classification using Naive Bayes, KNN, and SVM

1 Aim:

To design and implement classification models using Naive Bayes variants and K-Nearest Neighbors (KNN) algorithms to accurately classify emails as spam or ham. Additionally, to evaluate and compare their effectiveness using multiple performance metrics.

2 Libraries used:

- Numpy
- Pandas
- Matplotlib
- Scikit-learn
- Seaborn

3 Objective:

- To preprocess the email dataset by cleaning text data, vectorizing features, and splitting the data for training and testing.
- To implement Naive Bayes classifiers (Bernoulli, Multinomial, Gaussian) and KNN classifiers, tuning parameters such as k-value.
- To measure and compare model performance using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC, enabling informed model selection.

4 Naive Bayes Code:

width=!,height=!,pages=-

5 KNN Code:

width=!,height=!,pages=-

6 Comparision Tables

Table 1: Performance Comparison of Naïve Bayes Variants

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
BernoulliNB	0.8899	0.8843	0.8290	0.8558
MultinomialNB	0.8950	0.9424	0.7813	0.8543
GaussianNB	0.8197	0.7001	0.9485	0.8056

Table 2: KNN Performance for Different k Values

k	**Accuracy**	**Precision**	**Recall**	**F1 Score**
1	0.8899	0.8551	0.8676	0.8613
3	0.8892	0.8710	0.8438	0.8571
5	0.8993	0.8828	0.8585	0.8705
7	0.8950	0.8815	0.8474	0.8641

Table 3: KNN Comparison: KDTree vs BallTree

Metric	**KDTree**	**BallTree**
Accuracy	0.8899	0.8899
Precision	0.8551	0.8551
Recall	0.8676	0.8676
F1 Score	0.8613	0.8613
Training Time (s)	0.4470	0.4058

Table 4: Cross-Validation Scores for Each Model (K = 5)

Fold	Naïve Bayes Accuracy	KNN Accuracy (k=1)	SVM Accuracy
Fold 1	0.8719	0.9034	—
Fold 2	0.8935	0.9076	—
Fold 3	0.8891	0.9152	—
Fold 4	0.8913	0.9000	—
Fold 5	0.8859	0.8935	—
Average	**0.8863**	**0.9039**	—

7 Observation:

- KNN with k=1 achieved the highest accuracy consistently across all folds, indicating strong capability in classifying email spam versus ham.
- Naive Bayes classifiers, particularly MultinomialNB, provided stable and competitive results, showing robustness in handling text data with varying feature distributions.
- Although KNN attained better peak performance, Naive Bayes models required less training time and are more scalable for larger datasets.
- The choice between KNN and Naive Bayes depends on the trade-off between accuracy and computational efficiency for the specific application scenario.

GitHub Repository: https://github.com/Thamizhmathibharathi/project.git

8 Conclusion:

- The experiment demonstrated that KNN (k=1) outperforms Naive Bayes variants in terms of accuracy on the email classification task.
- Naive Bayes remains valuable due to its simplicity, fast training, and effectiveness on high-dimensional, sparse data typical in text classification.
- For deployment, Naive Bayes may be preferable when quick predictions are required, whereas KNN suits applications where highest accuracy is critical and computational resources are sufficient.