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: 27/1/26**

Experiment 2: Binary Classification using Naïve Bayes and K-Nearest Neighbors

Name: Muralisekar Janissha
Reg. No: 3122235001058
Class: CSE-B

1. Aim and Objective

Aim: To implement Naïve Bayes and K-Nearest Neighbors (KNN) classifiers for a binary classification problem.

Objectives:

- To evaluate models using multiple performance metrics.
- To tune KNN hyperparameters using cross-validation.
- To compare KDTree and BallTree neighbor search strategies.
- To analyze overfitting, underfitting, and bias–variance characteristics.

2. Dataset Description

A benchmark binary classification dataset with numerical features is used.

- Dataset: Spambase Dataset
- Source: Kaggle
- Classes: Spam and Ham
- Features: Numerical attributes

3. Preprocessing Steps

- Handling missing values
- Feature scaling and normalization
- Exploratory Data Analysis (EDA)
- Splitting dataset into training and testing sets

Feature scaling is essential for KNN due to distance-based computation.

4. Implementation Details

The experiment was implemented using Python with NumPy, Pandas, Scikit-learn, and Matplotlib.

- Naïve Bayes variants: Gaussian, Multinomial, Bernoulli
- Baseline KNN classifier
- Hyperparameter tuning using GridSearchCV and RandomizedSearchCV
- KNN with KDTree and BallTree

5-Fold Cross-Validation was used during tuning.

5. Visualizations

The following visualizations were generated to understand the dataset characteristics and analyze the performance of Naïve Bayes and KNN classifiers.

- Class distribution plot
- Confusion matrices for classifiers
- ROC curves for classifiers
- Accuracy vs. k plot for KNN

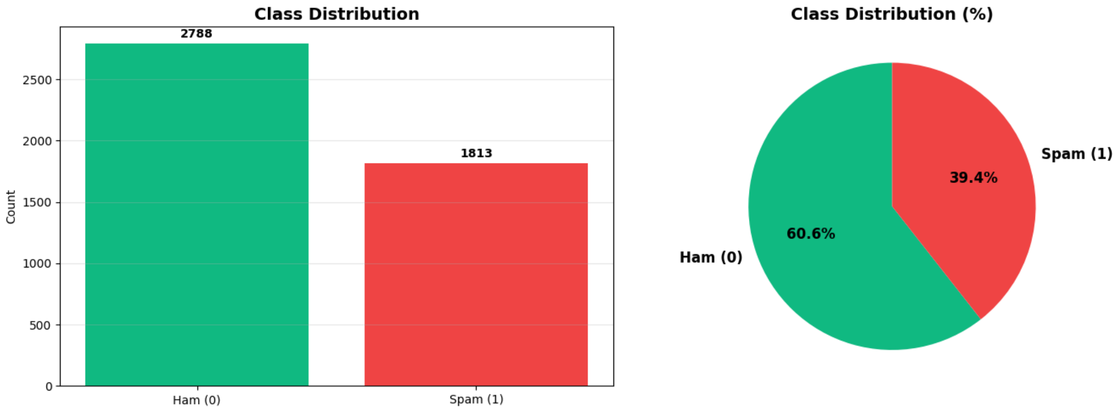

Figure 1: Class Distribution of Spam and Ham Emails

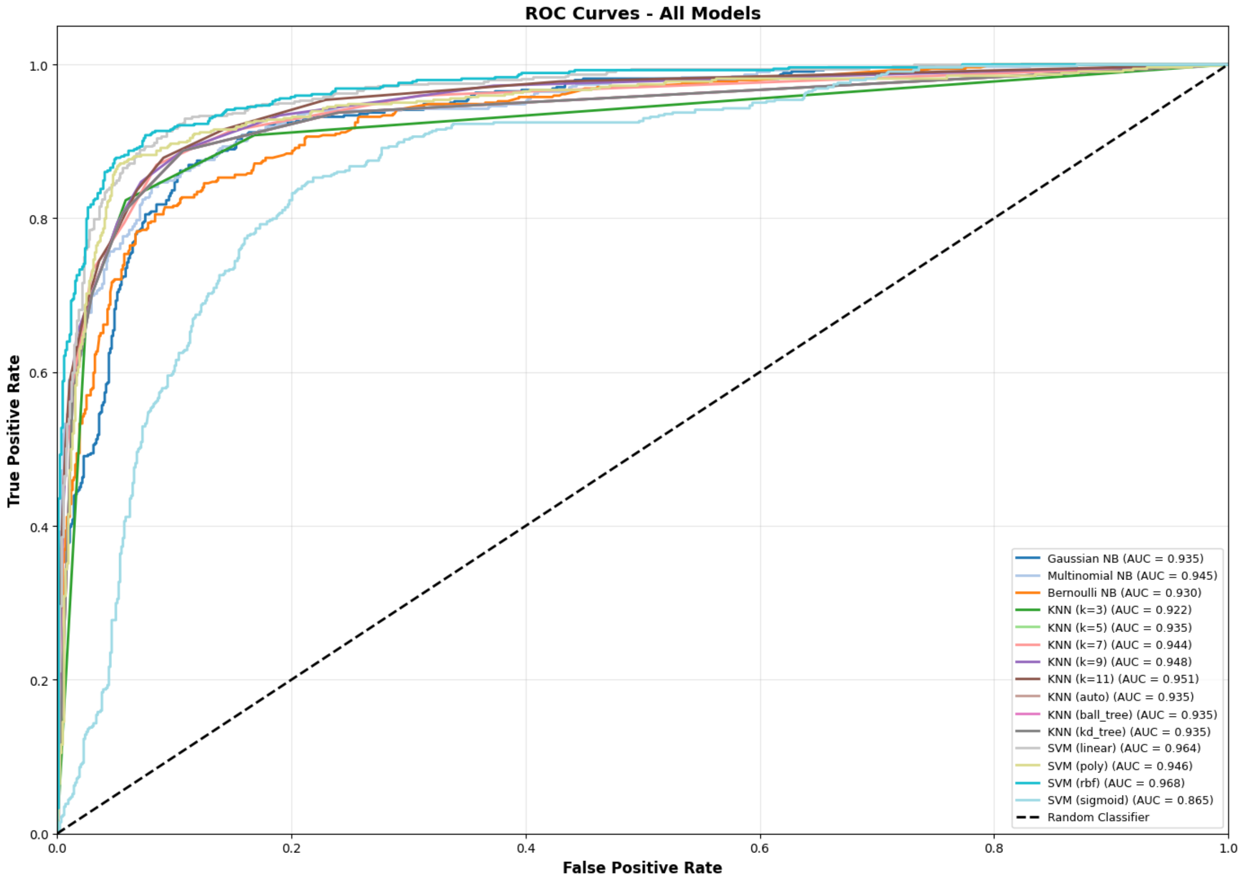

Figure 2: Confusion Matrices for Naïve Bayes and KNN Models

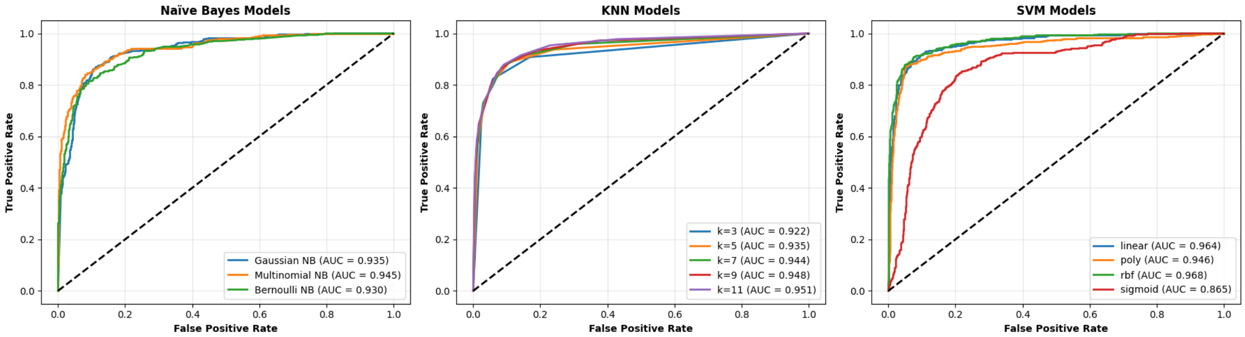

Figure 3: ROC Curves for Naïve Bayes and KNN Models

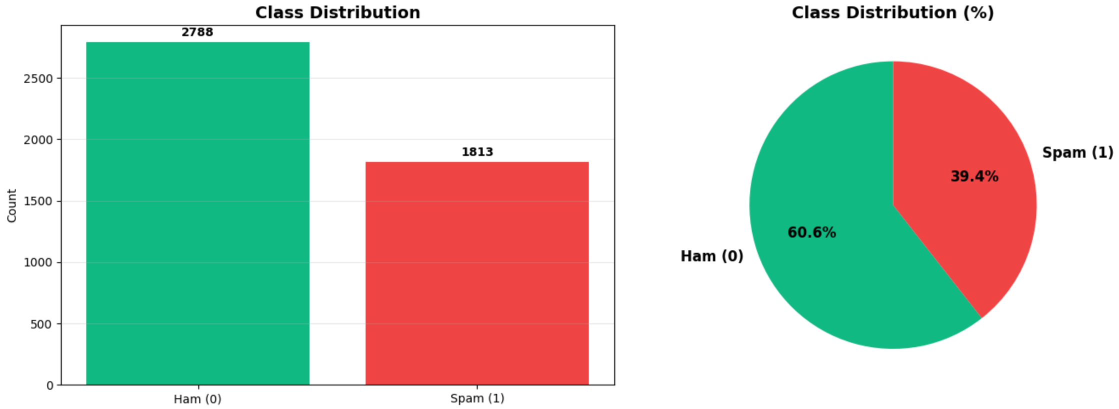

Figure 4: Accuracy vs. k for KNN Classifier

6. Performance Tables

The performance of Naïve Bayes and KNN classifiers was evaluated using multiple metrics.

Table 1: Naïve Bayes Performance Metrics

Metric	**Gaussian NB**	**Multinomial NB**	**Bernoulli NB**
Accuracy	0.8639	0.7306	0.8718
Precision	0.8886	0.9943	0.8605
Recall	0.7482	0.3180	0.8051
F1 Score	0.8124	0.4819	0.8319
Specificity	High	Very High	High
Training Time (s)	Low	Low	Low

Table 2: KNN Hyperparameter Tuning Results

Search Method	**Best k**	**Best Accuracy**	**Best Parameters**
Grid Search	3	0.8950	k=3

Table 3: KNN Performance using KDTree

Metric	**Value**
Optimal k	5
Accuracy	0.8899
Precision	0.9033
Recall	0.8070
F1 Score	0.8524
Training Time (s)	Low
Prediction Time (s)	Fast

Table 4: KNN Performance using BallTree

Metric	**Value**
Optimal k	5
Accuracy	0.8899
Precision	0.9033
Recall	0.8070
F1 Score	0.8524
Training Time (s)	Medium
Prediction Time (s)	Fast

Table 5: Comparison of Neighbor Search Algorithms

Criterion	**KDTree**	**BallTree**
Accuracy	0.8899	0.8899
Training Time	Low	Medium
Prediction Time	Fast	Fast
Memory Usage	Low / Medium	Medium / High

7. Overfitting and Underfitting Analysis

Small values of k cause overfitting, while large values lead to underfitting. Training and validation accuracy trends observed during cross-validation confirm this behavior.

8. Bias–Variance Analysis

Naïve Bayes has higher bias due to independence assumptions, whereas KNN shows higher variance. Hyperparameter tuning balances the bias–variance trade-off.

9. Observations and Conclusion

Observations:

- Bernoulli Naïve Bayes achieved the best balance among NB variants.
- KNN achieved highest accuracy at k = 3.
- KDTree and BallTree improved prediction efficiency.
- Very small k caused overfitting, while large k caused underfitting.

Conclusion: This experiment validates the effectiveness of Naïve Bayes and KNN for binary classification with proper tuning and evaluation.