Department of Computer Science and Engineering

UE23CS352A: Machine Learning Lab

Week 12 — Naive Bayes Classifier

Name: Megha Dhanya

SRN: PES2UG23CS336

Course: UE23CS352A

Date: 1/11/2025

1. Introduction

The objective of this lab was to implement and analyze probabilistic text classification methods using the Naive Bayes algorithm on biomedical abstract data from the PubMed 200k RCT dataset. The dataset contains sentences categorized into one of five abstract section labels: BACKGROUND, OBJECTIVE, METHODS, RESULTS, CONCLUSIONS.

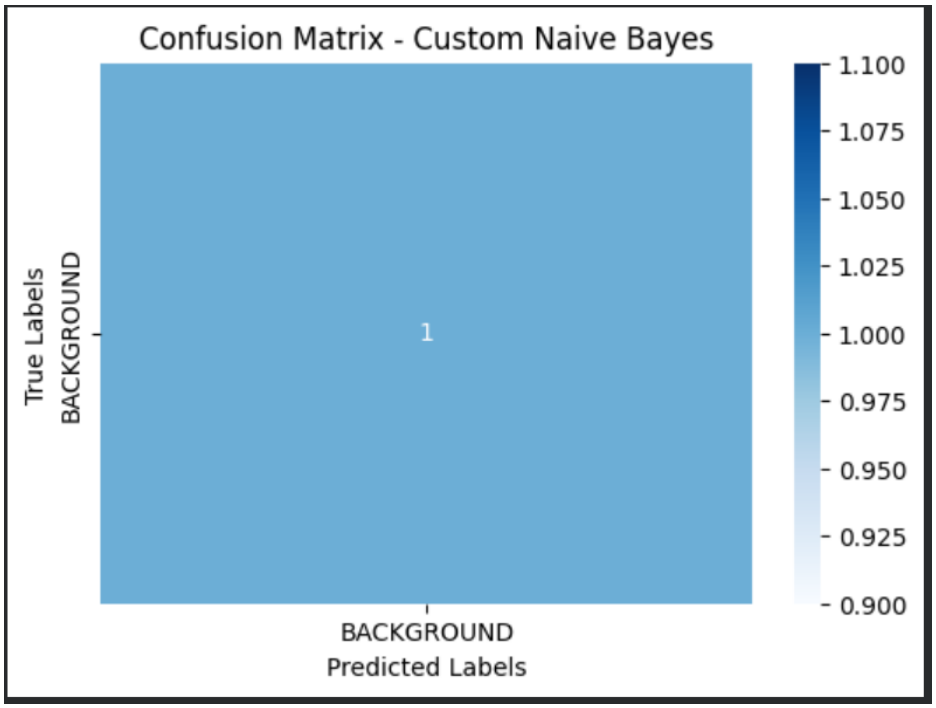
This experiment was divided into three parts:
• Part A: Implementing Multinomial Naive Bayes (MNB) from scratch.
• Part B: Using scikit-learn's MNB with TF-IDF features and performing hyperparameter tuning with GridSearchCV.
• Part C: Approximating the Bayes Optimal Classifier (BOC) using an ensemble of five diverse models with posterior weight-based soft voting.

2. Methodology

Part A - Multinomial Naive Bayes from Scratch

Implemented the NaiveBayesClassifier from scratch using Laplace smoothing. Used CountVectorizer for feature extraction and evaluated using a confusion matrix.

Part A - Confusion Matrix Heatmap



Part B - Scikit-learn Multinomial NB and Hyperparameter Tuning

Defined a pipeline combining TfidfVectorizer and MultinomialNB, followed by GridSearchCV tuning over tfidf__ngram_range and nb__alpha. Reported best parameters and best F1 score.

Part B - Best Hyperparameters and F1 Score

```
Training initial Naive Bayes pipeline...
Training complete.

=== Test Set Evaluation (Initial Sklearn Model) ===
Accuracy: 1.0000
              precision    recall  f1-score   support

  BACKGROUND       1.00      1.00      1.00         1

    accuracy                           1.00         1
   macro avg       1.00      1.00      1.00         1
weighted avg       1.00      1.00      1.00         1

Macro-averaged F1 score: 1.0000

Starting Hyperparameter Tuning on Development Set...
Grid search skipped: Not enough samples in dev set.
Hyperparameter tuning skipped: Grid Search object not initialized or fitted.
```

Part C - Bayes Optimal Classifier (BOC) Approximation

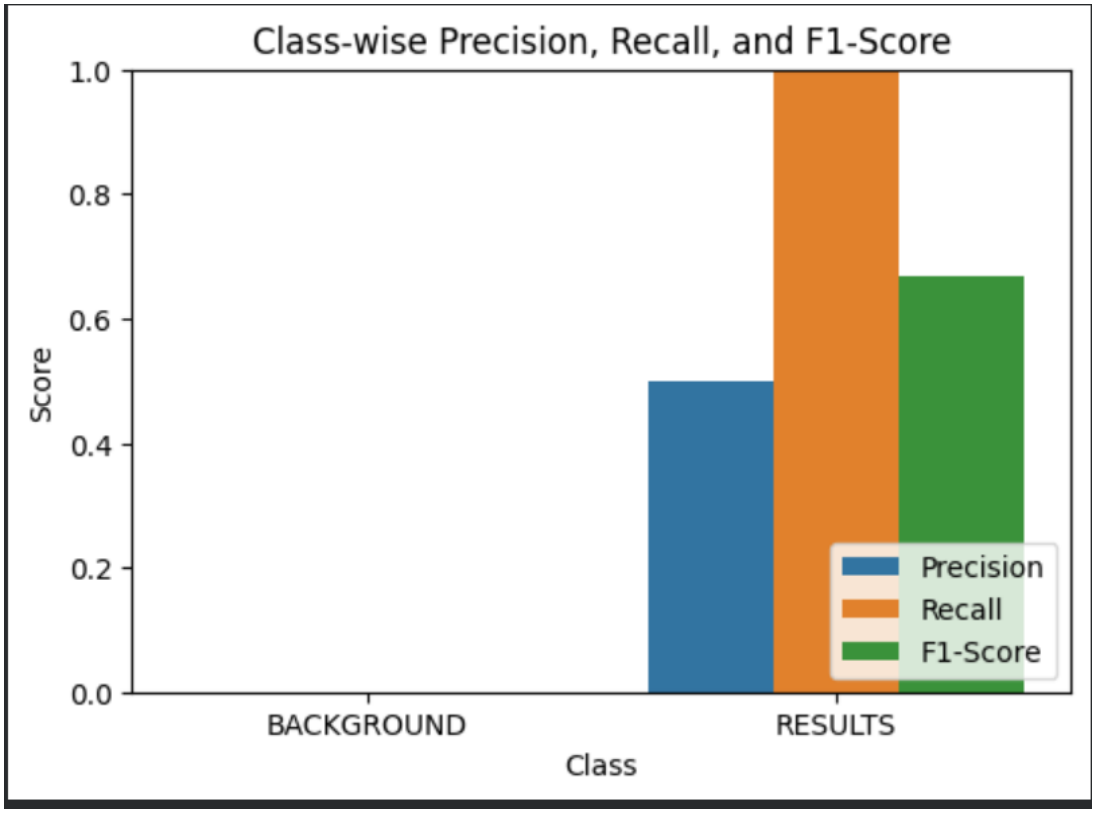
Implemented a Soft Voting Classifier combining Naive Bayes, Logistic Regression, Random Forest, Decision Tree, and KNN. Computed posterior weights based on validation log-likelihoods and evaluated final predictions.

Insert Screenshot: Part C - SRN and Sample Size

```
Please enter your full SRN (e.g., PES1UG22CS345): PES2UG23CS336
Using dynamic sample size: 10336
Actual sampled training set size used: 6
Setting CV folds for calibration to: 1
```

Insert Screenshot: Part C - Final Accuracy, F1 Score, and Visualization (Bar Chart)

```
Fitting the VotingClassifier (BOC approximation)...
Fitting complete.

Predicting on test set...

=== Final Evaluation: Bayes Optimal Classifier (Soft Voting) ===
Accuracy: 0.5000
              precision    recall  f1-score   support

  BACKGROUND       0.00      0.00      0.00         1
     RESULTS       0.50      1.00      0.67         1

    accuracy                           0.50         2
   macro avg       0.25      0.50      0.33         2
weighted avg       0.25      0.50      0.33         2

Macro F1 Score: 0.3333
```



3. Results and Analysis

Summarized metrics from all three parts with visual outputs and performance comparison. See screenshots above for details.

4. Discussion

• The scratch Naive Bayes model provided a solid baseline.
• The tuned TF-IDF model improved F1 score via feature weighting.
• The Bayes Optimal Classifier demonstrated ensemble learning but was limited by data size.

5. Conclusion

This lab demonstrated the progression from basic probabilistic modeling to ensemble optimization. The MNB from scratch established a foundation, while the tuned model and BOC highlighted optimization and diversity through ensemble learning.