Name:- MOHAMMED SHAZI UL ISLAM
SRN:- PES2UG23CS348
ML LAB

Purpose of the Lab

The primary goal of this lab was to explore and apply text classification techniques using Machine Learning, with a focus on the **Multinomial Naive Bayes (MNB)** classifier and an approximation of the **Bayes Optimal Classifier (BOC)**. This experiment aimed to enhance understanding of **probabilistic classification**, **feature extraction through TF–IDF**, and **model performance evaluation** using metrics such as *accuracy*, *F1-score*, and the *confusion matrix*.

Summary of Tasks Performed

The key steps executed in this lab included:

1. **Data Preprocessing and Sampling** – The dataset was cleaned, prepared, and sampled to facilitate efficient and faster experimentation.

2. **Custom Naive Bayes Implementation** – A Naive Bayes model was built from the ground up to gain insights into its internal computation and logic.

3. **Model Building with Multiple Algorithms** – Various classifiers, such as Naive Bayes, Logistic Regression, Random Forest, Decision Tree, and K-Nearest Neighbors, were developed using a unified TF–IDF pipeline.

4. **BOC Approximation via Ensemble Learning** – The classifiers were combined using a VotingClassifier with both *hard* and *soft* voting to approximate the Bayes Optimal Classifier.

5. **Model Evaluation and Visualization** – Each model's performance on unseen data was assessed and compared through metrics like accuracy and F1-score, supported by confusion matrix visualizations.

Methodology

1. **Multinomial Naive Bayes (MNB)**
   • The text data was converted into numerical form using a **TF–IDF Vectorizer**, which

assigns importance weights to words based on their frequency and relevance across documents.

- The **MNB model** was trained on the resulting TF–IDF feature representation.
- Assuming conditional independence of words given their class, MNB is particularly effective for **document classification** tasks.
- Predictions were generated on the test dataset, and the model's effectiveness was assessed using **accuracy**, **macro-averaged F1-score**, and a **confusion matrix** for detailed performance interpretation.

2. The **Bayes Optimal Classifier (BOC)** was approximated using an **ensemble learning** strategy implemented through a *VotingClassifier* that combined multiple models, including Multinomial Naive Bayes, Logistic Regression, Random Forest, Decision Tree, and K-Nearest Neighbors. Each of these base models was independently trained on the same **TF–IDF-transformed dataset** to ensure consistency in feature representation. In the **soft voting** mechanism, the ensemble aggregated the predicted probabilities from all classifiers and took their average, while in **hard voting**, it selected the class label that appeared most frequently among the models' predictions. This ensemble design aimed to exploit the complementary strengths of different algorithms, thereby minimizing individual model bias and improving overall generalization. Finally, the ensemble's performance was compared against the individual classifiers to evaluate how closely it approximated the **Bayes Optimal decision boundary**.

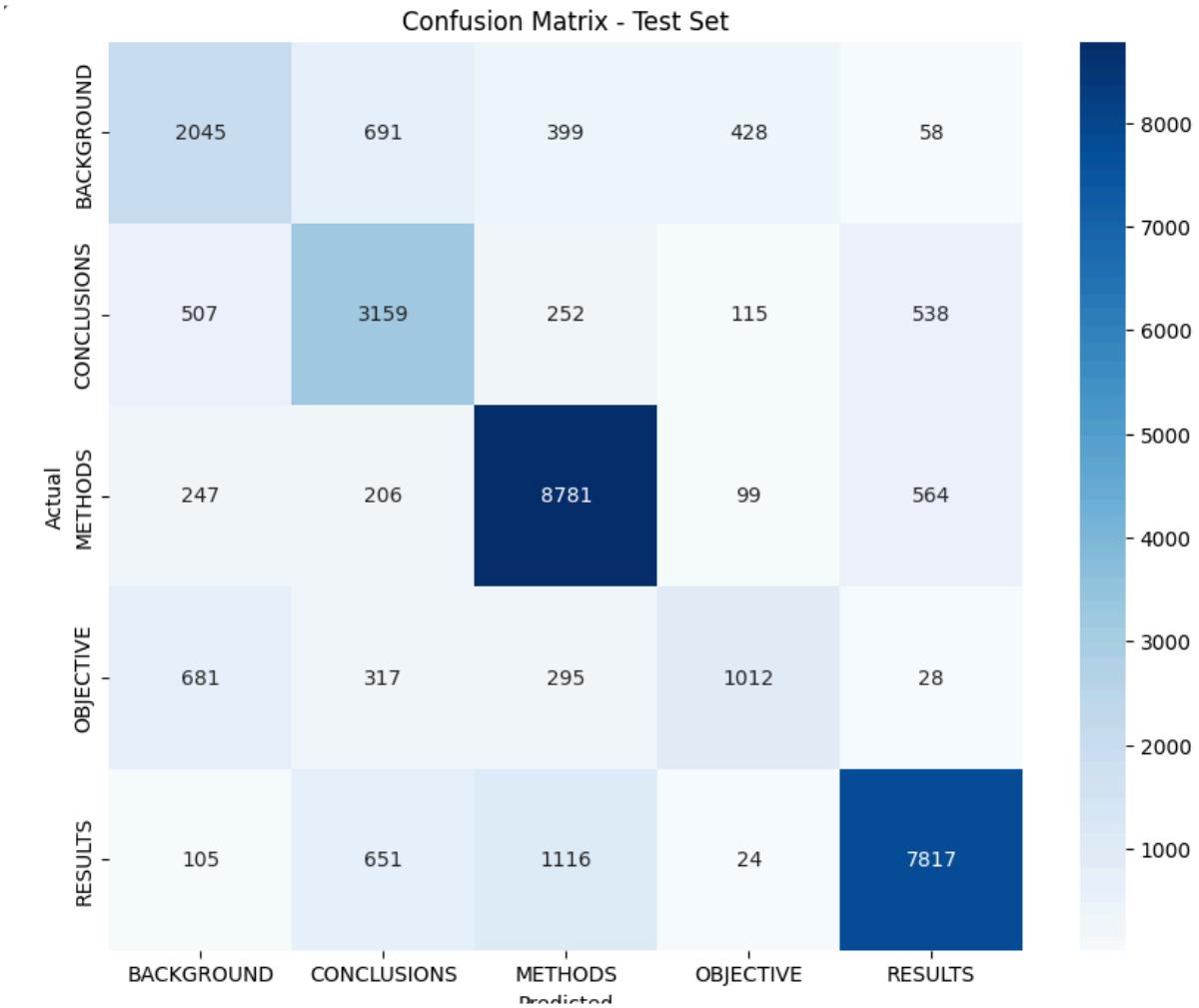
PART-A

```
=== Test Set Evaluation (Custom Count-Based Naive Bayes) ===
Accuracy: 0.7571
              precision    recall  f1-score   support

  BACKGROUND       0.57      0.56      0.57      3621
 CONCLUSIONS       0.63      0.69      0.66      4571
     METHODS       0.81      0.89      0.85      9897
   OBJECTIVE       0.60      0.43      0.50      2333
     RESULTS       0.87      0.80      0.84      9713

    accuracy                           0.76     30135
   macro avg       0.70      0.68      0.68     30135
weighted avg       0.76      0.76      0.75     30135

Macro-averaged F1 score: 0.6825
```


Confusion Matrix - Test Set
BACKGROUND
CONCLUSIONS
METHODS
OBJECTIVE
RESULTS
Actual
2045
691
399
428
58
507
3159
252
115
538
247
206
8781
99
564
681
317
295
1012
28
105
651
1116
24
7817
BACKGROUND
CONCLUSIONS
METHODS
OBJECTIVE
RESULTS
8000
7000
6000
5000
4000
3000
2000
1000

PART-B

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

=== Test Set Evaluation (Initial Sklearn Model) ===
Accuracy: 0.7266
              precision    recall  f1-score   support

  BACKGROUND       0.64      0.43      0.51      3621
 CONCLUSIONS       0.62      0.61      0.62      4571
     METHODS       0.72      0.90      0.80      9897
   OBJECTIVE       0.73      0.10      0.18      2333
     RESULTS       0.80      0.87      0.83      9713

    accuracy                           0.73     30135
   macro avg       0.70      0.58      0.59     30135
weighted avg       0.72      0.73      0.70     30135

Macro-averaged F1 score: 0.5877

Starting Hyperparameter Tuning on Development Set...
Fitting 2 folds for each of 4 candidates, totalling 8 fits
Grid search complete.

Best parameters: {'nb__alpha': 0.5, 'tfidf__ngram_range': (1, 1)}
Best cross-validation score: 0.5384
```

Part C

PART C

```
Please enter your full SRN (e.g., PES1UG22CS345): PES2UG23CS348
My SRN isPES2UG23CS348
Using dynamic sample size: 10348
Actual sampled training set size used: 10348
Using 10348 samples for training base models.

=== Training Base Models (H1–H5) ===
Training NaiveBayes...
NaiveBayes trained successfully.
Training LogisticRegression...
/usr/local/lib/python3.12/dist-packages/sklearn/linear_model/_logistic.py:1247: FutureWarr
  warnings.warn(
LogisticRegression trained successfully.
Training RandomForest...
RandomForest trained successfully.
Training DecisionTree...
DecisionTree trained successfully.
Training KNN...
KNN trained successfully.

=== Evaluation of Individual Hypotheses on Test Set ===
NaiveBayes           | Accuracy: 0.6999 | F1 (macro): 0.6203
LogisticRegression   | Accuracy: 0.7089 | F1 (macro): 0.6147
RandomForest         | Accuracy: 0.5213 | F1 (macro): 0.2726
DecisionTree         | Accuracy: 0.4686 | F1 (macro): 0.2444
KNN                  | Accuracy: 0.1656 | F1 (macro): 0.1033

=== Training Voting Classifier (Bayes Optimal Approximation) ===
Voting Classifier trained successfully.
```

```
=== Final Evaluation: Bayes Optimal Classifier Approximation ===
Accuracy: 0.6445
              precision    recall  f1-score   support

  BACKGROUND       0.52      0.42      0.46      3621
 CONCLUSIONS       0.64      0.50      0.56      4571
     METHODS       0.57      0.94      0.71      9897
   OBJECTIVE       0.63      0.02      0.04      2333
     RESULTS       0.87      0.65      0.74      9713

    accuracy                           0.64     30135
   macro avg       0.65      0.50      0.50     30135
weighted avg       0.67      0.64      0.62     30135

Macro F1 Score: 0.5024
```

Macro F1 Score: 0.5024