ML-2 PROJECT CODED_PROBLEM-1

KUMAR ANKIT

Kr.ankit7896@gmail.com

PROBLEM-1

CNBE, a prominent news channel, is gearing up to provide insightful coverage of recent elections, recognizing the importance of data-driven analysis. A comprehensive survey has been conducted, capturing the perspectives of 1525 voters across various demographic and socio-economic factors. This dataset encompasses 9 variables, offering a rich source of information regarding voters' characteristics and preferences.

Objective

The primary objective is to leverage machine learning to build a predictive model capable of forecasting which political party a voter is likely to support. This predictive model, developed based on the provided information, will serve as the foundation for creating an exit poll. The exit poll aims to contribute to the accurate prediction of the overall election outcomes, including determining which party is likely to secure the majority of seats.

Data Description

1. **vote**: Party choice: Conservative or Labour
2. **age**: in years
3. **economic.cond.national**: Assessment of current national economic conditions, 1 to 5.
4. **economic.cond.household**: Assessment of current household economic conditions, 1 to 5.
5. **Blair**: Assessment of the Labour leader, 1 to 5.
6. **Hague**: Assessment of the Conservative leader, 1 to 5.
7. **Europe**: an 11-point scale that measures respondents' attitudes toward European integration. High scores represent 'Eurosceptic' sentiment.
8. **political.knowledge**: Knowledge of parties' positions on European integration, 0 to 3.
9. **gender:** female or male.

The objective of this analysis was to develop predictive models to classify voter behaviour using a dataset titled "Election_Data.xlsx." The dataset contains information on various demographic and socioeconomic features of voters, along with their voting behaviour categorized into two classes: 'Yes' and 'No'.

We employed several machine learning algorithms including K-Nearest Neighbours (KNN), Naive Bayes, Bagging, AdaBoost, Decision Trees, Gradient Boosting, and Random Forest classifiers to build predictive models. The models were evaluated based on their accuracy, confusion matrix, classification report, and Receiver Operating Characteristic (ROC) curve.

Data Preprocessing:

- The dataset consisted of 1525 entries and 9 variables.
- Missing values were checked, and it was ensured that the dataset contained no null values.

```
vote                        0
age                         0
economic.cond.national      0
economic.cond.household     0
Blair                       0
Hague                       0
Europe                      0
political.knowledge         0
gender                      0
```

- Categorical variables were encoded using label encoding.
- The data was split into training and testing sets.
- Standardization was applied to scale the features.

Model Building and Evaluation:

- K-Nearest Neighbors (KNN):

Achieved training accuracy of 87% and testing accuracy of 77%.

```
Training Accuracy: 0.8772258669165885
Testing Accuracy: 0.7751091703056768
```

- Navie Bayes:

  Achieved training accuracy of 84% and testing accuracy of 82%.

```
Training Accuracy: 0.8397375820056232
Testing Accuracy: 0.8187772925764192
```

- AdaBoost:

  Achieved training accuracy of 86% and testing accuracy of 80%.

```
Training Accuracy: 0.85941893158388
Testing Accuracy: 0.8034934497816594
```

- Bagging:

  Achieved training accuracy of 98% and testing accuracy of 80%.

```
Training Accuracy: 0.9840674789128397
Testing Accuracy: 0.7969432314410481
```

- Decision Tree:

  Achieved training accuracy of 81% and testing accuracy of 78%.

- Gradient Boosting:

  Achieved training accuracy of 81%.

- Random Forest:

  Achieved training accuracy of 81% and testing accuracy of 81%.

Model Performance evaluation:

Model: KNN

```
Training Accuracy: 0.8772258669165885
Testing Accuracy: 0.7751091703056768

Confusion Matrix (Test Data):
 [[ 84  49]
 [ 54 271]]

Classification Report (Test Data):
              precision    recall  f1-score   support

           0       0.61      0.63      0.62       133
           1       0.85      0.83      0.84       325

    accuracy                           0.78       458
   macro avg       0.73      0.73      0.73       458
weighted avg       0.78      0.78      0.78       458
```

Model: Naive Bayes

```
Training Accuracy: 0.8397375820056232
Testing Accuracy: 0.8187772925764192

Confusion Matrix (Test Data):
 [[ 87  46]
 [ 37 288]]

Classification Report (Test Data):
              precision    recall  f1-score   support

           0       0.70      0.65      0.68       133
           1       0.86      0.89      0.87       325

    accuracy                           0.82       458
   macro avg       0.78      0.77      0.78       458
weighted avg       0.82      0.82      0.82       458
```

Model: Bagging

```
Training Accuracy: 0.9840674789128397
Testing Accuracy: 0.7969432314410481

Confusion Matrix (Test Data):
 [[ 92  41]
 [ 52 273]]

Classification Report (Test Data):
              precision    recall  f1-score   support

           0       0.64      0.69      0.66       133
           1       0.87      0.84      0.85       325

    accuracy                           0.80       458
   macro avg       0.75      0.77      0.76       458
```

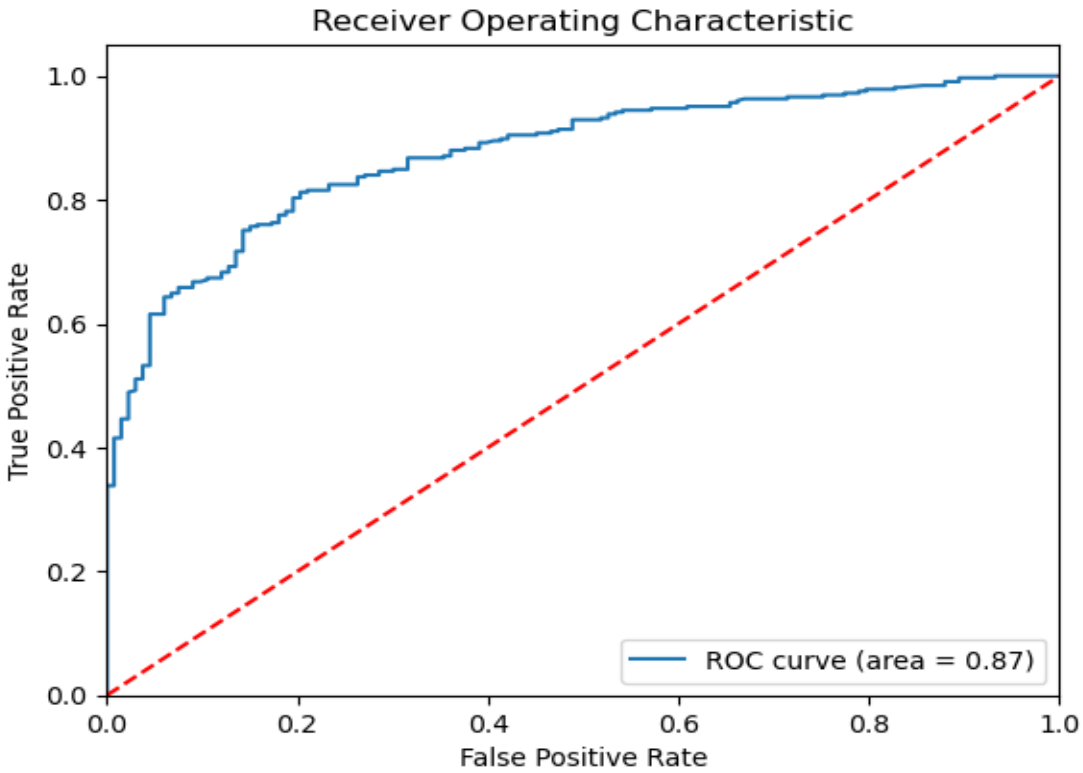
```
weighted avg       0.80      0.80      0.80       458
```

Model: Boosting

```
Training Accuracy: 0.85941893158388
Testing Accuracy: 0.8034934497816594

Confusion Matrix (Test Data):
 [[ 85  48]
 [ 42 283]]

Classification Report (Test Data):
              precision    recall  f1-score   support

           0       0.67      0.64      0.65       133
           1       0.85      0.87      0.86       325

    accuracy                           0.80       458
   macro avg       0.76      0.75      0.76       458
weighted avg       0.80      0.80      0.80       458
```

ROC-AUC Score:

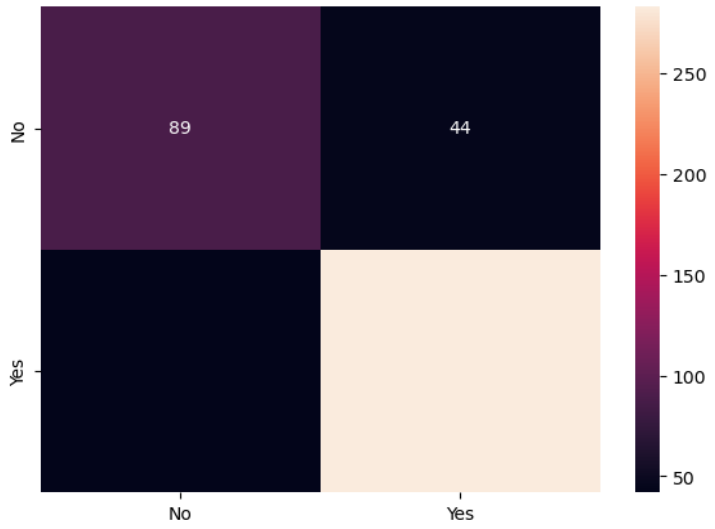
```
ROC-AUC Score: 0.8746211683053788
```



Model Performance improvement:

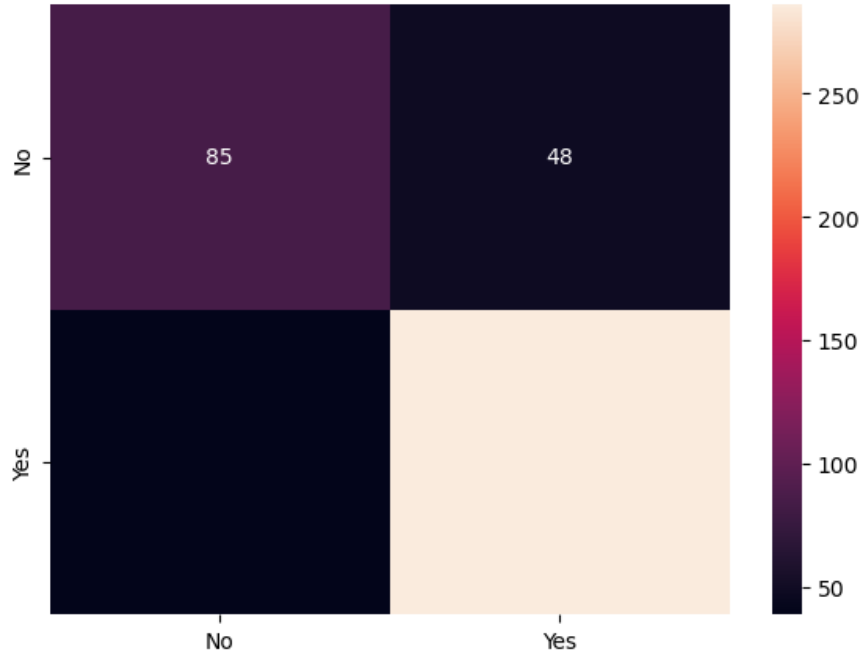
Bagging:-

0.8122270742358079



Boosting:

0.8100436681222707



Model Comparison:

```
Model Comparison:
KNN Accuracy: 0.7751091703056768
Naive Bayes Accuracy: 0.8187772925764192
Bagging Accuracy: 0.7969432314410481
Boosting Accuracy: 0.8034934497816594
```

Model Selection:

Among all the models tested, the Naïve Bayes model performed the best on the testing data with an accuracy of 82%.

The most important features influencing the model's predictions were identified as follows.

	Imp
age	0.069192
economic.cond.national	0.000000
economic.cond.household	0.000000
Blair	0.319860
Hague	0.516214
Europe	0.094734
political.knowledge	0.000000
gender	0.000000

Conclusion:

In conclusion, the predictive models developed in this analysis demonstrate the feasibility of using machine learning techniques to predict voter behaviour based on demographic and socioeconomic features. The models can potentially assist political parties and election campaigns in targeting spe cific voter segments more effectively.

Recommendations:

- Further refinement of models and feature engineering could potentially enhance prediction accuracy.
- Continuous monitoring and updating of models with new data could improve their performa nce over time.
- Deployment of the best-performing model for real-world applications, such as targeted cam paigning and resource allocation.

Limitations:

- The analysis is based on historical data and may not fully capture dynamic changes in voter behavior.
- The performance of models may vary depending on the quality and representativeness of the data.
- Ethical considerations regarding data privacy and bias need to be addressed when deploying predictive models in sensitive domains like elections.

Future Directions:

- Exploration of ensemble methods and advanced deep learning techniques for further improving model accuracy.
- Integration of additional data sources such as social media activity and sentiment analysis to enrich the predictive models.
- Collaboration with domain experts and policymakers to leverage predictive insights for informed decision-making in electoral processes.