Election Data Analysis and Voter Prediction - Documentation

# Problem Statement

CNBE, a leading news channel, conducted a survey on 1525 voters to understand their voting preferences between the Labour and Conservative parties. The objective is to analyze voter behavior based on demographic and opinion-based variables and build a classification model that can predict which party a voter will vote for. This model is used to simulate an exit poll.

# Dataset Description

The dataset includes 1525 records and 9 features including:  
1. vote: Target variable (Labour or Conservative)  
2. age: Voter age  
3. economic.cond.national: Perception of national economic conditions (1-5)  
4. economic.cond.household: Household economic perception (1-5)  
5. Blair: Labour leader rating (1-5)  
6. Hague: Conservative leader rating (1-5)  
7. Europe: Sentiment toward European integration (1-11)  
8. political.knowledge: Political awareness (0-3)  
9. gender: male/female

# Steps Performed

1. Data Loading and Initial Exploration

2. Dropping Irrelevant Column (Serial Number)

3. Checking for Duplicates and Removing Partial Duplicates

4. Exploratory Data Analysis (EDA): Generated Automated EDA Report using ydata-profiling

5. Outlier Detection via Boxplots

6. Encoding Categorical Variables (Vote and Gender)

7. Feature Scaling for Numerical Columns

8. Data Splitting (70:30 with Stratification)

9. Model Building and Evaluation:

* Logistic Regression
* Linear Discriminant Analysis (LDA)
* k-Nearest Neighbors (kNN)
* Naive Bayes
* Random Forest (Bagging)
* General Bagging
* Boosting (Adaptive Boosting)

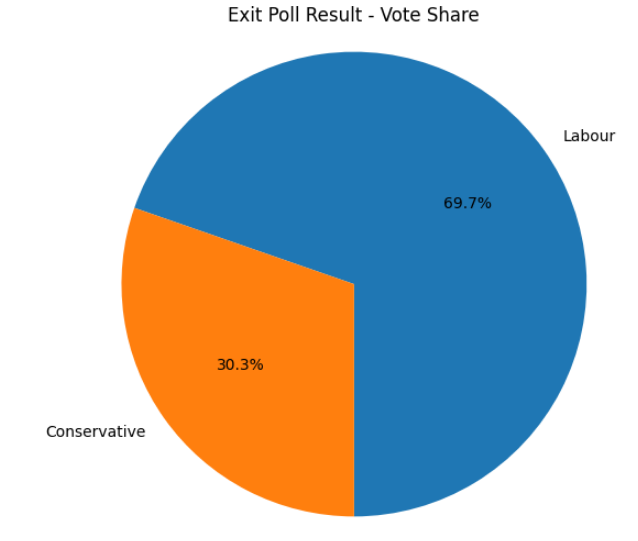
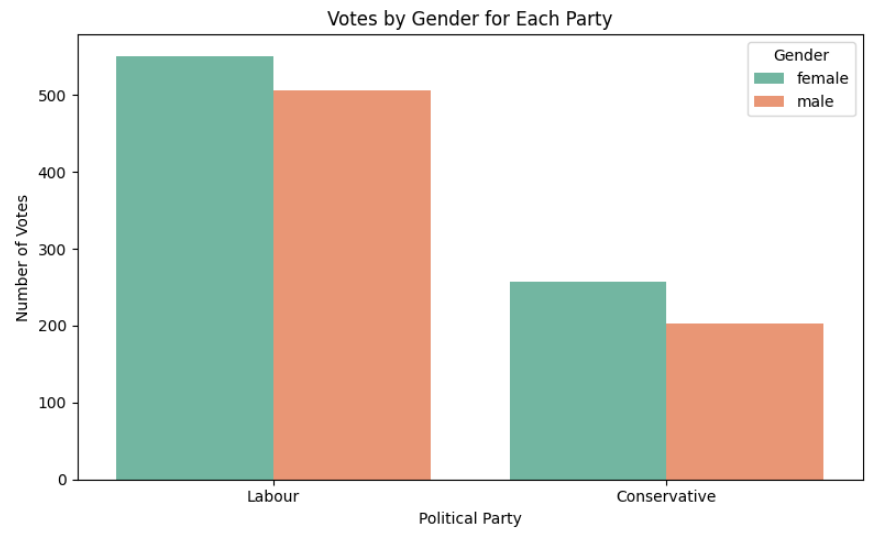
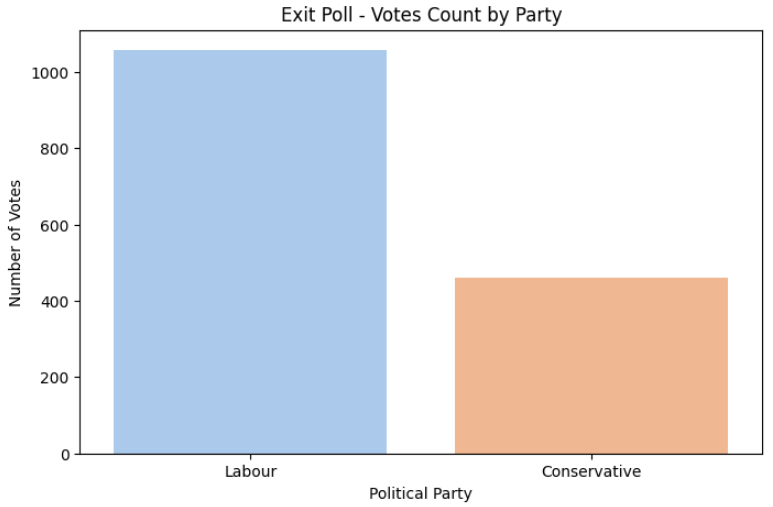
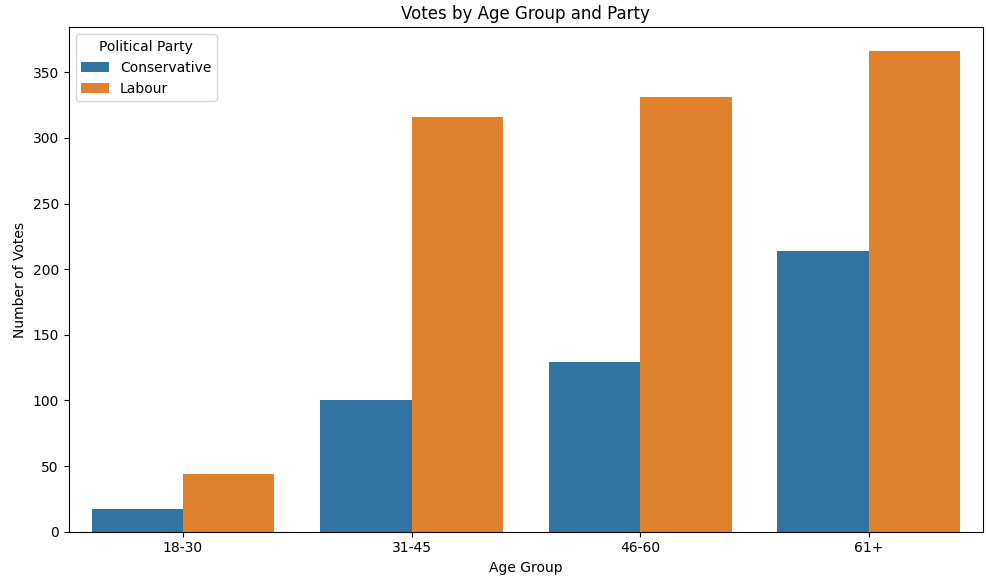
10. Evaluation Metrics: Accuracy, Confusion Matrix, Classification Report

11. Final Model Selection and Conclusion

# Exploratory Data Analysis (EDA)

Univariate and bivariate analyses revealed the following insights:  
- Most voters showed high support for Blair and low for Hague.  
- Labour received a majority of votes across both genders, especially from females.  
- Younger voters leaned slightly more towards Labour.  
- Political knowledge correlated positively with Labour support.

# Exit Poll Insights

- Labour has strong support across both genders, especially females.  
- Female voters are more active and politically engaged in this dataset.  
- Age-wise visualization showed younger and middle-aged groups favor Labour more.  
- These trends suggest the Labour party's campaign had wider appeal and engagement.   

# Model Accuracy Comparison

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Training Accuracy** | **Test Accuracy** | **Comments** |
| Logistic Regression | 0.836 | 0.840 | Solid baseline; balanced performance |
| LDA | 0.838 | 0.842 | Slightly better than logistic regression |
| kNN | 0.852 | 0.836 | High training accuracy; possible overfitting |
| Naive Bayes | 0.825 | 0.844 | Lower training, good generalization |
| Random Forest | - | 0.827 | Stable but slightly underperforming |
| Bagging | - | 0.827 | Consistent but not top-performing |
| Boosting | - | **0.849** | Best performing model overall |

# Model Comparison and Inference

Among all models, Boosting delivered the highest test accuracy (84.9%), indicating that it is the best performing model for this dataset. Naive Bayes also showed strong generalization performance despite its lower training accuracy. LDA and Logistic Regression offered consistent and reliable results. kNN had the highest training accuracy but may suffer from overfitting. Random Forest and Bagging were stable but didn't outperform Boosting.  
Hence, Boosting is recommended as the final model with possible hyperparameter tuning for further optimization.

**Inference**

* Adaptive Boosting outperformed all other models in terms of test accuracy (84.9%), making it the best overall. It achieved a good balance of precision and recall.
* Naive Bayes surprised with a strong test performance despite a lower training score — indicating excellent generalization.
* LDA and Logistic Regression were also strong contenders with consistent and high accuracy, showing they are still competitive for linear separability.
* kNN had the highest training accuracy, suggesting it might be overfitting slightly — as its test accuracy dropped a bit.
* Random Forest and Bagging had solid but slightly lower performance, possibly due to parameter defaults or redundancy in tree-based ensembles for this dataset.

Conclusion

* Best Performing Model: Boosting **— Adaptive Boosting** — due to its superior accuracy and f1-score.
* Caution: Models like kNN and Random Forest can overfit if hyperparameters aren’t tuned.
* Recommended Final Model: Boosting, with potential tuning for learning\_rate and n\_estimators=1000.

# Streamlit Deployment

A user-friendly Streamlit web application was created to allow users to input voter features and get real-time predictions using the trained AdaBoost model (***voter\_model.pkl***).