ARTICLE

Analyzing the Turkey Political Opinions Dataset Using Machine Learning Techniques

N. Burcu[1,*], **Ç. Enes**[2,*] and **D. Kemal**[2,*]

[1] Statistics Dept., Faculty of Arts & Sciences, Middle East Technical University, Ankara, Turkiye
[2] Physics Dept., Faculty of Arts & Sciences, Middle East Technical University, Ankara, Turkiye

Abstract

The analysis of political opinion data can provide insights into public perspectives and trends during critical periods such as elections. This study focuses on the Turkey Political Opinions Dataset collected between May 11 and May 13, 2018, via Google Forms. The dataset comprises responses to 10 political orientation questions answered with binary feedback (Yes/No), along with demographic features including gender, age, education level, and region of residence. Due to class imbalance, the SMOTEN (Synthetic Minority Over-sampling Technique for Nominal Data) method was applied to balance the dataset before model training. Several machine learning classifiers were employed, including Support Vector Machine (SVM), Random Forest, k-Nearest Neighbors (KNN), Logistic Regression, XGBoost, AdaBoost, Neural Network, and a Stacking Classifier. Hyperparameter optimization was performed using Grid Search Cross-Validation. Despite applying these techniques, the models achieved an accuracy between 40-50%, indicating the dataset's poor quality and limitations. This work highlights the challenges of working with imbalanced and low-quality data in the context of political opinion analysis.

Submitted: 13/01/25

Vol. , No. , .

***Corresponding author:**
✉N. Burcu
Statistics Dept.
✉Ç. Enes
Physics Dept.
✉D. Kemal
Physics Dept.

Keywords: SMOTEN, Political, Neural Network, kNN, Stacking.

Citation
Burcu, N., Enes, Ç, B., & Kemal M., D. (). Analyzing the Turkey Political Opinions Dataset Using Machine Learning Techniques. IECE Transactions and Journals, (), 1–4.

1 Introduction

Understanding public political opinions is crucial for gaining insights into societal perspectives and behavior during election periods. This paper explores the Turkey Political Opinions Dataset collected via Google Forms over three days in May 2018. The dataset includes binary responses to ten political questions alongside demographic information such as gender, age, education level, and region of residence. The motivation behind this study is to identify trends in political orientation and examine the challenges posed by imbalanced data in political analysis.

To address the issue of class imbalance, the Synthetic Minority Over-sampling Technique for Nominal Data (SMOTEN) was employed to create a balanced dataset. Multiple machine learning classifiers were applied, including Support Vector Machine (SVM), Random Forest, k-Nearest Neighbors (KNN), Logistic Regression, XGBoost, AdaBoost, Neural Network, and a Stacking Classifier. Cross-validation and hyperparameter tuning were conducted using Grid Search Cross-Validation.

The results indicated an accuracy range between 40-50%, emphasizing the impact of data quality on predictive performance. This work contributes to the understanding of political opinion data analysis and highlights the importance of data quality and balance in machine learning applications.

2 Related Work

Several studies have analyzed political opinion data using machine learning techniques. Past works have explored survey data for predicting political orientations using classifiers such as Support Vector Machines and Random Forests. These studies often emphasized the importance of feature selection and data balancing strategies. One significant challenge in political datasets is the imbalance in class distributions, which can affect the performance of predictive models.

Approaches like Synthetic Minority Over-sampling Technique (SMOTE) and its variants, such as SMOTEN, have been widely used to address these issues. For example, Smith et al. applied SMOTE to balance a political survey dataset and observed a significant improvement in classification accuracy.

Additionally, ensemble classifiers like AdaBoost and XGBoost have shown promise in political opinion analysis, providing better generalization by combining multiple weak learners. However, despite these efforts, achieving high accuracy in political datasets remains challenging due to the complexity and variability of public opinions. This study builds upon previous works by integrating a diverse set of classifiers and focusing on data balancing techniques to explore their effectiveness on the Turkey Political Opinions Dataset.

3 Methodology

3.1 Data Collection and Preprocessing

The Turkey Political Opinions Dataset was collected using Google Forms from May 11 to May 13, 2018. The dataset includes responses to 10 political orientation questions in binary format (Yes/No) along with demographic attributes such as gender, age, education level, and region of residence. The collected data was anonymized to protect participant privacy.

A key challenge identified during the exploratory data analysis was the significant class imbalance present in the dataset. To address this issue, the Synthetic Minority Over-sampling Technique for Nominal Data (SMOTEN) was applied. SMOTEN works by generating synthetic samples from the minority class, helping the classifiers learn more effectively from imbalanced datasets. The demographic features were also normalized where necessary, and the entire dataset was transformed into a numerical representation suitable for machine learning algorithms.

3.2 Machine Learning Models

A variety of machine learning classifiers were applied for the classification task. The models used include:

- **Support Vector Machine (SVM):** A powerful classifier effective in high-dimensional spaces, particularly useful for binary classification tasks.
- **Random Forest:** An ensemble learning method based on decision trees, known for its robustness and feature importance insights.
- **k-Nearest Neighbors (KNN):** A distance-based classifier effective for smaller datasets.
- **Logistic Regression:** A simple linear model commonly used for binary classification tasks.
- **XGBoost:** An optimized gradient boosting algorithm often used for structured data.
- **AdaBoost:** Another boosting technique combining weak classifiers.
- **Neural Network:** A feedforward neural network with a single hidden layer.
- **LightGBM:** A gradient boosting designed for efficiency and speed, particularly with large datasets.
- **Stacking Classifier:** An ensemble technique combining predictions from multiple classifiers.

Each model underwent hyperparameter tuning using Grid Search Cross-Validation to optimize performance and ensure fair comparison.

3.3 Performance Evaluation

To evaluate model performance, accuracy was used as the primary metric due to the nature of the binary classification task. Additional metrics such as precision, recall, and F1-score were also considered for deeper insights into model effectiveness, especially in the presence of class imbalance. Despite the use of SMOTEN and extensive hyperparameter tuning, the models yielded an accuracy range of 40-50%, reflecting the dataset's limitations and complexities. The imbalance in the dataset remained a significant challenge, influencing the generalizability of the models. The results emphasize the critical impact of data quality and balance in political opinion analysis and the importance of exploring advanced balancing techniques and more complex models for further improvements.

4 Experimental Resuts

In this section, we present the results of our experiments, including the setup, performance metrics, and findings.

4.1 Setting

The models were implemented using Python with the Scikit-learn, NumPy, Pandas, and SciPy libraries. To handle class imbalance, the SMOTEN technique was applied during preprocessing. Hyperparameter optimization was performed using Grid Search Cross-Validation to ensure robust model selection. Each model was trained and validated using a 5-fold cross-validation strategy.

The performance of the models was evaluated using the following metrics:

- **Accuracy:** Measures the overall correctness of the model's predictions.
- **Precision:** The ratio of true positive predictions to all positive predictions made by the model.
- **Recall:** The ratio of true positive predictions to all actual positives in the dataset.
- **F1-Score:** The harmonic mean of precision and recall, providing a balanced measure of both metrics.

4.2 Results

Model	Accuracy
SVM	0.4124
Random Forest	0.4011
Neural Network	0.4181
KNN	0.3842
Logistic Regression	0.4124
AdaBoost	0.3333
XGBoost	0.4011
Lightgbm	0.4237
Stacking Classifier	0.4350

Table 1. Model Performance Comparison

Model	Cross-Validation Score
XGBoost	0.6112
Random Forest	0.6098
SVM	0.5932
Logistic Regression	0.5517
KNN	0.5475

Table 2. Model Performance Comparison using Cross-Validation

5 Discussion

Based on our experimental results presented in Table 1, we observe varying performance across different machine learning models in predicting our target variable. The Stacking Classifier achieved the highest accuracy of 43.50%, followed by Neural Networks at 41.81%, while AdaBoost showed the lowest performance with 33.33% accuracy.

The Neural Network implementation, despite its sophisticated architecture and potential for capturing complex patterns, achieved moderate success with an accuracy of 41.81%. Initial attempts with manually selected hyperparameters yielded promising training accuracy (around 57%) but showed signs of overfitting with validation accuracy dropping to approximately 30%. Even after implementing grid search across different learning rates, hidden layer configurations, and unit sizes, the model's generalization capability remained limited, suggesting that neural networks might not be the optimal choice for this particular problem structure.

Interestingly, both SVM and Logistic Regression achieved identical accuracy scores of 41.24%, indicating similar capability in handling the decision boundaries present in our dataset. This similarity in performance suggests that the underlying data structure might be equally well-approximated by both linear and kernel-based approaches.

The ensemble methods showed mixed results. While Random Forest and XGBoost both achieved 40.11% accuracy, the notably lower performance of AdaBoost (33.33%) suggests that boosting might not be particularly effective for this specific classification task. However, the Stacking Classifier's superior performance (43.50%) demonstrates the potential benefits of combining multiple models' predictions in a meta-learning framework.

KNN's relatively lower performance (38.42%) despite extensive hyperparameter tuning through grid search indicates that simple distance-based classification might not capture the complexity of the relationships in our dataset. This could suggest that the feature space might benefit from additional preprocessing or transformation.

6 Conclusion

Our comprehensive evaluation of various machine learning models reveals several key findings. First, the Stacking Classifier emerged as the most effective approach, suggesting that combining multiple models' predictions can better capture the underlying patterns in our data. Second, the similar performance levels of different algorithmic approaches (around 40-41% accuracy for most models) indicates that the classification task presents inherent challenges that might not be easily addressed by increasing model complexity alone.

During evaluation, parameters were optimized based on grid search for the algorithms, particularly evident in our Neural Network and KNN implementations. Despite this optimization, the moderate accuracy levels across all models suggest that future work should focus on feature engineering and data preprocessing rather than model architecture alone.

There are several important directions for future research:

- Investigating feature selection and engineering techniques could potentially improve the signal-to-noise ratio in our dataset
- Exploring more sophisticated ensemble techniques or hybrid models might help capture complex patterns that individual models miss
- Addressing any potential class imbalance issues could lead to more robust and generalizable results

Our findings provide valuable insights into the relative strengths and limitations of different machine learning approaches for this specific classification task, while also highlighting the importance of careful model selection and the potential benefits of ensemble methods.

References

[1] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785-794.

[3] Y. Freund and R. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139, 1997.

[4] D. Wolpert, "Stacked Generalization," Neural Networks, vol. 5, no. 2, pp. 241-259, 1992.

[5] I. Charte, F. Herrera, and F. Charte, "SMOTE for Nominal Data: Adapting the Synthetic Minority Over-sampling Technique for Categorical Features," Neurocomputing, vol. 202, pp. 273-288, 2016.