**PROJECT REPORT**

ANIL ARSLAN - 1900119007

**Abstract**

The aim of this project is trying to predict the election results in Turkey using different socio-economic indicators. The data collection for that study is done by querying birth rates, marriage ages for men and women, Gross Domestic Product (GDP), total export values, Starbucks outlets, and school rates by cities in Turkey. These factors were selected consciously in such a way that they could affect social and economic welfare. The data collected was cleaned properly, after which normalization through MinMax scaling was done to retain comparability in different metrics. Thereafter, a composite socioeconomic score was obtained for each city; this scored represented the overall measure of development and welfare status. The k-Nearest Neighbors (kNN) algorithm, efficient for pattern recognition and clustering, was used to predict the dominant political party in each city based on these socio-economic scores. The methodology further involved strict steps of data cleaning, normalization, and merging with the implementation of cross-validation techniques to enhance the predictive performance of the model. The results demonstrated a positive trend in the relationship between the calculated socio-economic score and the election results. Cities with high scores had more of a leaning toward the party CHP, while the cities with lower scores were going toward the parties AKP and YSP. Precision and recall rates produced by kNN model, especially for forecasting AKP control, were quite high. This showed how strong sensitivities of socio-economic conditions were with voting outcomes. This project concludes that the considered socio-economic factors could really and effectively forecast election results, giving political analysts and policymakers a lot of information. By doing so, this work adds to the literature by marrying diverse data sources with a machine learning approach to elucidate the intricate relationship between socio-economic factors and electoral outcomes. The results of the study underline the centrality of socio-economic development in shaping political landscapes and provide a stable framework for further research and practical applications in socio-political strategy and planning.

**Introduction**

Elections form the cornerstone of a democratic society and mirror the public's will and choices. Understanding the factors influencing voting behavior is crucial for political analysts, policymakers, and researchers. Common hypotheses suggest that socio-economic conditions such as economic prosperity, education levels, and social well-being significantly impact election outcomes. Research indicates that socio-economic factors strongly affect political preferences. For instance, studies show that people with higher income levels tend to favor conservative parties, while those with higher education and social well-being often lean towards liberal parties (Rodden, 2019; Solt, 2008).

This project aims to explore Turkish elections through the research question: Can socio-economic indicators predict election results in Turkey? To address this question, a detailed analysis was conducted in several key steps. The first step involved data collection from reliable sources, including TUIK (Turkish Statistical Institute) and the Ministry of National Education (MEB). Additionally, web scraping was used to gather relevant data on factors such as birth rates, marriage ages, Gross Domestic Product (GDP), total export values, the number of Starbucks outlets, and high school education rates. The raw data was cleaned and normalized using Min-Max scaling to ensure uniformity and comparability across different indicators. This process included managing NA values, standardizing city names, and merging multiple disjointed datasets into one integrated dataset. These steps are crucial for ensuring the analysis's reliability and follow best practices in data science (Provost & Fawcett, 2013).

In the next step, an overall socio-economic score for each city was generated from the standardized indicators to show the general socio-economic profile and level of development. The k-Nearest Neighbors (kNN) algorithm was then used to classify the dominant political party in these cities based on the predicted socio-economic score. Cross-validation methods were employed to validate the model and optimize its performance. The use of machine learning algorithms in political science research is well-documented, with studies showing their effectiveness in predicting political outcomes based on socio-economic data (Bishop, 2006; Hastie, Tibshirani, & Friedman, 2009).

**Data**

This study uses data from the most reliable sources in order to avoid discrepancies. The main sources of the study are Turkish Statistical Institute (TUIK), Ministry of National Education (MEB), and other supplementary data extracted through web scraping, elections results obtained from the CNN Turk API. It aimed at recent data that would reflect socio-economic conditions pertaining to the 2023 elections in Turkey. The data collection process was multistep. The first step in the procedure was that of acquiring data from TUIK on birth rates, marriage ages for both men and women, GDP, and total export values. These datasets enabled understanding the various economic and social factors that prevailed within the different cities in Turkey over the last ten years. To ensure the analysis was based on the most up-to-date information, it kept only the data for the most recent years and converted this to CSV files for ease of use. Besides these datasets, I have also recognized that Starbucks count may act as an indicator of economic and social welfare; that is due to the apparent strategic siting that Starbucks has been known to do, relative to their economic viability, which makes them an apt proxy for economic development. However, there wasn't an available dataset. Thus, I collected this data by scraping from the web pages of the Alshaya Group, which is the holder of all the Starbucks outlets in Turkey, by using Python libraries such as requests and BeautifulSoup. To collect data about high school education rates for each province, I used a dataset provided by MEB. I gathered this data, in fact, from a long PDF file by using the PyPDF2 library. Relevant data was searched for and scraped in terms of pages, followed by parsing, which was done to collect high school education rates that were converted to a CSV format. This also included the election results for the latest election by scraping the deputy and party results using the CNN Turk API. The API was designed to have an endpoint so that the deep data of the elections could be taken rather simply, and then after this, it can be parsed as well as transformed into a CSV file and be analyzed. This data is indeed absolutely essential if one is going to compare any of the diverse socio-economic factors with an actual voting outcome. Now all these various assorted data sets having been collected, the final process involved in merging them into one integrated dataset. Since the city names were not uniformly represented in all datasets and even while the Starbucks data used English city names that then have to be converted to its Turkish equivalents, any errors in the city names were manually addressed for accurate merging. Normalization was the most important step taken in the preparation process of the data. All the variables were min-max normalized to a common scale, generally ranging from 0 to 1. This procedure facilitated comparison and amalgamation of different indices regarding the socio-economic status. Normalised data was used in scores that determine the development and welfare status of each city in the composite socio-economic status.

**Methodology**

The first step of the project was to gather multi-data sources in order to get a wide array of socio-economic indicators. Main data sources included the Turkish Statistical Institute (TUIK), having the data on birth rates, marriage ages of men and women, Gross Domestic Product (GDP), and total export values by provinces. Other data were taken from the Ministry of National Education (MEB), which provided the rate of high school education for each province. I also scraped the web in order to collect information about the number of Starbucks stores in each city; an index of economic and social welfare. More interestingly, the most recent election outcomes along with the deputy and party results for each province, were taken from the CNN Turk API. After collecting data, it was prepared quite rigorously in a bid to ensure it was consistent and useful. In essence, raw data were checked for missing values and inconsistency from observations which were then treated as necessary. Scaling all the variables to a common scale using Min-Max scaling so that all variables contribute equally in the further steps of analysis was done. The different data sets were then merged to one cohesive data set. This necessitated that all city names in the various datasets be standardized, in this case, changing the English city names to Turkish city names and manually rectifying any differences. The score for measuring a city's socioeconomic status was aggregated out of indicators. The following is a list of indicators that were used in the calculation of the score: birth rates, marriage ages, GDP, total exports, Starbucks counts, high school education rates. Because the value of each indicator is incomparable, a MinMax scaling was conducted on each indicator to have them as comparable:. Next, a weighted sum of the normalized indicators was calculated to obtain a composite socio-economic score for each city; the weights were calculated based on the relative importance of an indicator as inferred from the data. Most of the analysis hinged on predicting election outcomes using the k-Nearest Neighbors (kNN) algorithm. I have chosen this algorithm because of its simplicity and good effectiveness in pattern recognition and clustering tasks: it predicts the class of a data point by classes of points k nearest to it in the feature space. The dataset is split into training and testing sets, where 80% of the data is used for training, while the remaining 20% is used for testing. The kNN model was trained by the training set to indicate the municipality's politically dominant party according to a socio-economic score. To ensure the robustness of the model, cross-validation was performed. Thus, k partitions were prepared for the training set. The model was then trained and validated k times—in each iteration, using different partitions as a validation set—therefore avoiding overfitting and allowing us to evaluate the performance of the model. Parameter tuning was carried out to maximize the performance of the model with respect to bias-variance trade-off. Experimentation for different values of k gave the best value which maximized the performance of the model. Different metrics have been applied to describe the performance of the kNN model. The confusion matrix uncovers how many correct and incorrect predictions have been made within each class. For measuring how accurate the model is in making predictions with regard to each of these political parties, precision and recall measures were developed. Precision defines the number of true positive predictions divided by all the positive predictions; recall determines the number of true positive predictions divided by all the real positives. The F1 score gives a single measure of the balance between precision and recall because it is the harmonic mean of these two measures.

**Findings**

The analysis yielded significant insights into the relationship between socio-economic factors and election outcomes in Turkey. Several key findings emerged from the data exploration, visualization, and machine learning predictions.

The first step involved examining the correlation between the socio-economic indicators and the composite score. The correlation matrix before calculating coefficients (Figure 1) shows the relationships between variables such as birth rate, marriage ages, GDP, total export values, Starbucks count, school rates, and the composite score. The analysis revealed that GDP, total exports, and school rates positively correlate with the composite score, indicating that higher values in these indicators are associated with higher socio-economic development.

A blue squares with white text

Description automatically generated

Figure 1 :Correlation Matrix Before Coefficients Calculation

Conversely, birth rate showed a negative correlation, suggesting that higher birth rates are associated with lower socio-economic scores. After calculating the coefficients for the score calculation, the refined correlation matrix (Figure 2) provided a clearer picture of the influence of each factor on the composite score. The refined correlations highlighted the significant impact of GDP, export values, and school rates on the overall socio-economic status of the cities.

A blue squares with white text

Description automatically generated

Figure 2 : Correlation Matrix After Coefficients Calculation

A detailed analysis of individual features provided further insights. The chart comparing the correlation between features (Figure 3) illustrates the varying degrees of correlation between GDP, total export values, Starbucks count, school rates, and birth rates across different cities. This visualization emphasized the importance of economic indicators such as GDP and exports in determining the socio-economic status.

A graph of colored lines

Description automatically generated

**Figure 3:** Correlation Between Features

The analysis also explored the relationship between socio-economic scores and political party dominance. The chart depicting features and scores according to parties (Figure 4) shows that cities with higher scores tend to favor the CHP party, while those with lower scores are more likely to support the AKP and YSP parties. This correlation aligns with the hypothesis that socio-economic development influences voting behavior.

A graph with different colored bars

Description automatically generated

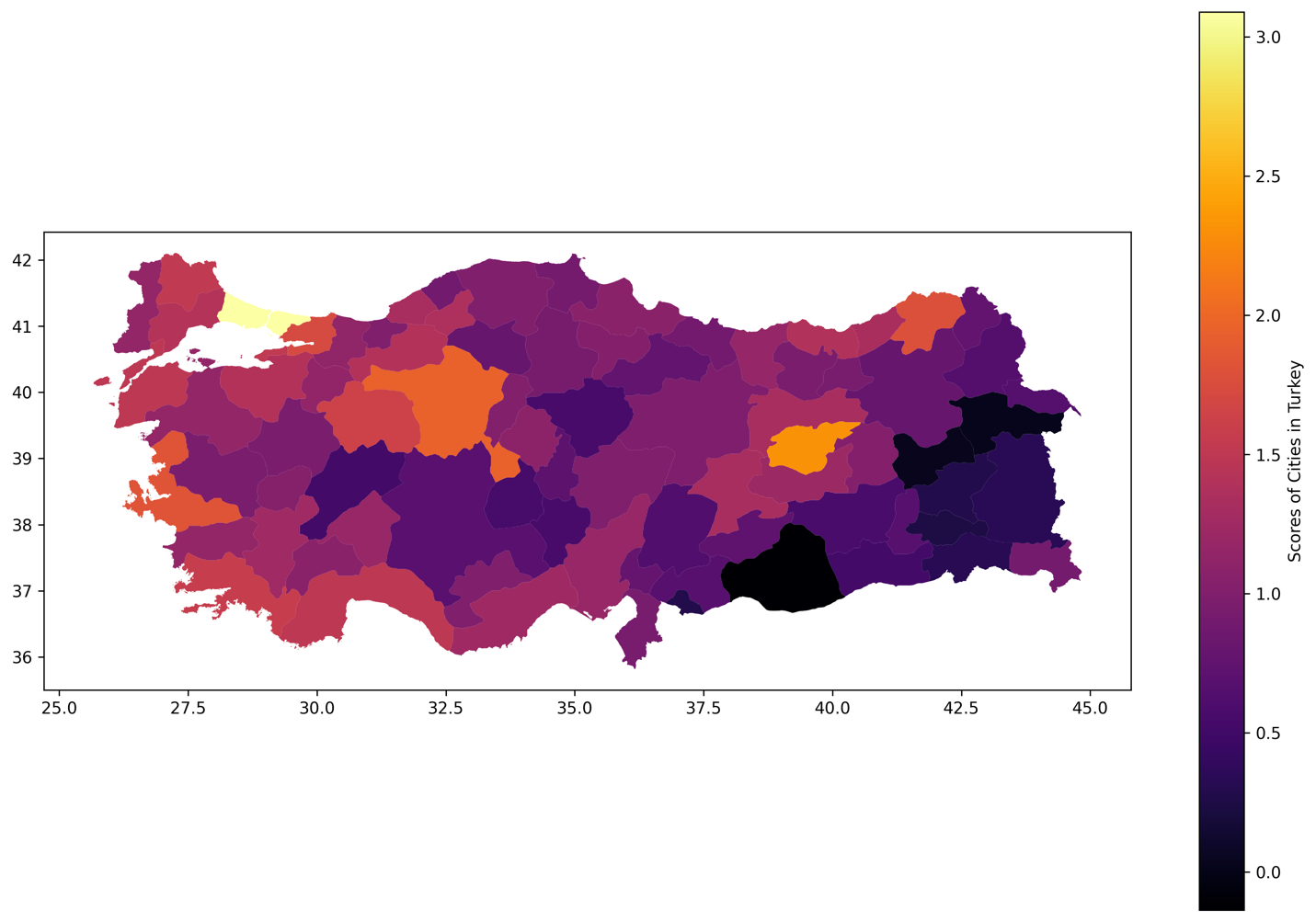
**Figure 4:** Features and Scores According to Parties

The comparison between GDP and socio-economic scores (Figure 5) highlighted the significant role of GDP in determining the overall development of cities. Cities like Istanbul, with high GDP values, also had the highest socio-economic scores, whereas cities with lower GDP values had correspondingly lower scores. This underscores the critical impact of economic prosperity on socio-economic status.

A graph with blue and orange lines

Description automatically generated

**Figure 5:** Comparison of GDP and Socio-Economic Scores

The geographical distribution of socio-economic scores across Turkish cities (Figure 6) provided a visual representation of regional disparities. The map shows that cities in the western and central regions generally have higher scores, indicating better socio-economic conditions. In contrast, cities in the eastern regions tend to have lower scores, reflecting socio-economic challenges.

**Figure 6:** Geographical Distribution of Socio-Economic Scores in Turkey

The k-Nearest Neighbors (kNN) algorithm was used to predict election outcomes based on socio-economic scores. The model demonstrated a mean cross-validation score of 0.7654, indicating a strong predictive capability. The confusion matrix (Figure 7) and classification report (Figure 8) provided insights into the model's performance. The model showed high precision, recall, and F1 scores, particularly for predicting AKP dominance. The precision for AKP was 1.00, with a recall of 0.93 and an F1 score of 0.97. For the Yeşil Sol party, the precision was 0.67, with a recall of 1.00 and an F1 score of 0.80. The overall accuracy of the model was 0.94, with macro and weighted average F1 scores of 0.88 and 0.95, respectively.

A screenshot of a computer

Description automatically generated

Figure 7: Classification Report

A black background with white numbers

Description automatically generated

Figure 8: Confusion Matrix

The findings from this analysis confirm the significant impact of socio-economic indicators on election outcomes in Turkey. Economic factors such as GDP and total exports, along with social indicators like school rates, play a crucial role in shaping voting behavior. The geographical disparities in socio-economic scores highlight the need for targeted socio-economic policies to address regional inequalities. The machine learning model provides a robust tool for predicting election results based on socio-economic data, offering valuable insights for political analysts and policymakers.

**Conclusion**

The aim of this project was to find out the potential of various socio-economic indicators in using them for forecasting elections results in Turkey. It is also evidenced from this analysis, when a large number of diverse datasets are normalized and machine learning approaches are applied herein, that voting behavior is also affected by the existing socio-economic conditions. The findings have shown that both economic factors like GDP and total exports, as well as some social indicators for school rates, contribute to electoral preferences. Cities with high socio-economic scores tended to support the CHP party, whereas those with low scores leaned towards AKP and YSP, indicating the role of socio-economic development in creating political landscapes.  
  
There were substantial geographical differences in the socio-economic scores: cities located in the western and central regions generally had better socio-economic conditions than their eastern counterparts. The geographical differences within the socio-economic scores of the cities clearly indicate that regional disparities exist and, as such, targeted policies would need to be applied in order to reduce or bridge them in a way that will bring balanced socio-economic development.  
  
The kNN model turned out to be great, having high precision, recall, and F1-scores. More specifically, in the prediction of AKP dominance, it was evidently shown that the model had a high level of precision.  
  
All in all, such a project provides the model framework for understanding and forecasting election outcomes using socio-economic indicators. These results may also benefit political analysts and policymakers to provide insight into voting behavior and guide policy strategies that can alleviate some of these socio-economic challenges. Future research might further the scope of these findings by including more socio-economic factors, other machine learning algorithms, or stretching it to different regions or countries for an additional level of validation.

**References**

**Datasets**

* Household income by province
  + <https://data.tuik.gov.tr/Bulten/Index?p=Il-Bazinda-Gayrisafi-Yurt-Ici-Hasila-2021-45619>
* High School Number / Population by province
  + <https://sgb.meb.gov.tr/meb_iys_dosyalar/2021_09/10141326_meb_istatistikleri_orgun_egitim_2020_2021.pdf>
* Marriage age women/man by province
  + <https://data.tuik.gov.tr/Bulten/DownloadIstatistikselTablo?p=B8eRA2vkujeYknMy0TAYbqte6CyyXEFl9DfO61apy3dCfypG2ZTEHtM17/l9I8ve>
* The number of marriages and the number and rate of consanguineous marriages by province
  + <https://data.tuik.gov.tr/Bulten/DownloadIstatistikselTablo?p=OJSOcu/MYAYma1RQLPf5OrGYbFAFruwQI5Z6cWeIkiapDnJIvnjoGvT8dn1sHi6Q>
* Birth rate by province
  + <https://data.tuik.gov.tr/Bulten/DownloadIstatistikselTablo?p=cfo3rGQVEwQ15XGmRT/6HP2cQDCaed1/H6qeucMSqczQ2zFpoxw2hHbxDByqkRA9>
* Export by provinces
  + <https://data.tuik.gov.tr/Bulten/DownloadIstatistikselTablo?p=cr5fsUeZBI3JUeQpNsczZEYDUpmcg708g3YsdGeKrBwlEQaqy6azjRUDPR47QMTi>
* Starbucks data by province
  + <https://www.shaya.com.tr/tr/magazalarimiz/?topZone=32>
* Parliamentary elections data by province
  + <https://secim2023.cnnturk.com/api/home>

**Articles**

* Rodden, J. (2019). Why Cities Lose: The Deep Roots of the Urban-Rural Political Divide. Basic Books.
* Solt, F. (2008). Economic Inequality and Democratic Political Engagement. American Journal of Political Science, 52(1), 48-60.
* Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media.
* Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
* Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.