**BOHOHOYT PROJECT SUMMARY**

**Eren Altın – Atakan Öz – Batuhan Işıldak – Anıl Arslan – Burak Ersöz – Deren Doğan**

Hello, we are team Bohohoyt. The aim of this project is to predict the election results according to the social and economic features of the cities in Turkey. We calculated a score through the data we obtained according to social and economic factors such as birth rate, marriage age of men and women, GDP, total export, Starbucks count and school rate in cities in Turkey. First of all, we divided our project into the following sections: data finding and preparing, data merging, data visualization and score explaining, machine learning, and finally conclusion.

**Data Finding and Preparation**

By utilizing the CNN Turk API, we successfully obtained the deputy and party results for the election. The API's data type, accessible through the provided endpoint, facilitated the retrieval of necessary information. The extracted data was parsed into a CSV file, enabling efficient analysis. Furthermore, the creation of the score attribute allowed for the evaluation of each province's performance based on various factors.

**Online Datasets**

First of all, Let’s take a look at our datasets and how we collected data. We got most of our data from TUIK. Most of the data were including approximately 10 years data. In order to reach the most up-to-date data, we cleaned the raw data to take the most recent years and then converted the cleaned data into csv file. We aimed to reach the most up-to-date data because the results we used belonged to the last 2023 elections. We have 2021 GDP data by cities, we have marriage age of men and women separately in 2022, we have export data by provinces and finally we have birth rate by province which are the data that negatively affect the score, but we will talk about the score in the following sections.

**Web Scrapping**

Especially in Turkey, Starbucks is one of the signs for welfare and productivity. But Starbucks is a for-profit company, and they are great at their job. So, they carefully select where they will focus on. Therefore, we thought that Starbucks count is a great indication of economic and social welfare. Unfortunately, we could not find current and actual dataset for Starbucks count. So, we decided to get our dataset from owner of all Starbucks in Turkey, Alshaya group. When we try to scrape their website, it was impossible since they are using lazy loading according to country. We thought Selenium is our best choice but before that search and found turkey branch of Alshaya group, Shaya. They have same data-table in their website. We used requests, and BeautifulSoup to get and scrape data after that convert it to csv.

**PDF Parsing**

For getting the high school education rates of each province, we used the data which is provided by MEB here is the data of high school education rates for each province. To parse and extract this data, we used PyPDF2 library and PDFReader function. Then as we showed our data it is a long PDF and our data is in certain place so we run this code for every page our wanted data exists and index of page changes

**APIs**

**Data Merging into A One Giant Dataset**

First, we kept our MEB data in txt after we wrote a txt converter to turn that information to csv. We used Turkish names for our cities, but the Starbucks data was English. We turned the English city names into Turkish names and while transforming them we encountered with an error with the letter “İ” in Turkish. We handle it for required cities with for loop and changed it to Turkish versions. We need that all city names should be same because we will merge them according to that.

We had a lot of data from different sources. In order for us to process them, we had to combine them under a single data frame. We ran into some problems when combining them, as we got them from different sources. First of all, we noticed that city names are written differently from each other in some data. This prevented us from combining them all under one data frame. However, since these errors are very rare and we have limited data, we did not need a parser and chose to correct them manually. Another problem is that in our Starbucks count data, the number of cities did not match because there were no Starbucks in every city. However, this was not a problem for us because we know that there are no Starbucks in the missing cities, so we filled them with a value of zero.

**Score Calculation**

**MinMax Scaling and Column Normalization**

We try to find a score for every province by looking into these columns: birth\_rate, men\_marriage\_age, women\_marriage\_age, gdp, total\_export\_$, starbucks count, school\_rate

To make this task, we need normalization.

Normalization is a process of transforming data to a common scale, usually between 0 and 1. Min-max scaling is a type of normalization where the values of a variable are transformed proportionally to fit within a specified range, typically between 0 and 1.

After doing the normalization what we need to do is getting the correlations between features and score. To do this, we created a heat map and checked the score field’s coefficients of other features.

**Calculating Coefficients for Score Calculation**

By getting these coefficients we can get the desired coefficients of other features. So, with all of these we can find the score for each city which is their socioeconomic level and their development rate. Coefficients are in order -0.377, 0.716, 0.772, 0.727, 0.636, 0.669, 0.601.

**Data Visualization**

**GDP and Score Correlation**

GDP (Gross Domestic Product) is one of the most important indicator of economic improvement of the city. From the graph it is understood that GDP and our score variable is correlated. Additionally, when, we look at the Kocaeli, it has the nearly same GDP with Istanbul. However, Istanbul’s score is nearly doubles Kocaeli’s score which shows that GDP is not the only parameter that effects our score’s value. Also, Tunceli’ s GDP is very low related to other cities and its score is too high from others which shows score could be high without high GDP.

**Political Party and Score Correlation**

In this graph, we created scores for CHP, AKP, and YSP which are the top 3 party of last election, according to data of the social and economic factors. We can observe that our score decreases much more in AKP and YSP parties, as the places where the birth rate is high have a negative contribution to the score and welfare level. In CHP, the score is higher when the birth rate is low. Again, we observe that the age of marriage for men and women is younger in AKP and YSP regions. In GDP ratio, we observe that GDP is higher in CHP and this contributes positively to the score and level of development. Again, we observe that CHP's score is higher than AKP and YSP in places where the rate of schools and Starbucks is high.

**Cities and Score Correlation**

Here is the score scale that we calculated through the data and the map where it is distributed to the cities. If we look at Istanbul, it appears as the city with the highest score level. We can interpret this result in direct proportion to the data that increases the level of welfare and development. On the contrary, we observe that the score rate is very low in cities such as Şanlıurfa, which is also influenced by factors that negatively affect the level of development such as low exports, low schooling rates and young marriage age.

**Machine Learning**

**Why kNN and Unsupervised Learning?**

K-Nearest Neighbors (kNN) is an unsupervised machine learning algorithm used for pattern recognition and clustering tasks. It is a non-parametric method that determines the class or cluster of a given data point by comparing it to its k nearest neighbors in the feature space.

We will predict the parties based on scores. As we concluded with our visuals, Cities with similar scores tend to vote for same parties. Since our score calculations based on different features, we concluded a balanced distribution. That canalized us to make our predictions based on kNN method.

Since our data, AKP dominant, we select a high number of neighbors, so our model is using more cities within similar range. But we had to limit it with 5 to prevent overfitting to our train data.

We used randomization and %20 of our data as our test data. We train our model with rest %80 percent.

For default, kNeighborsClassifier method uses uniform distribution and Minkowski distance calculation.

**Cross Validation and Testing Our Model**

Cross validation is used for better calculations for model trainings scores. The cross validation is important because it test our model more than one time and gets their means for average score. It divides the training set into k subsets, and it train the model with k-1 subset of k for k times. We divide our training data into five subsets, and we trained our model 5 times and tried 5 times. We calculate results mean and get our cross-validation score.

**Conclusion**

Finally, kNN algorithm predicts test values then we printed confusion matrix of the results, when we examine this matrix, or AKP its precision rate is 1 and 0.93 as recall value. Also, we calculated the f1 score of the labels and all of them is relatively high from these facts it can be concluded that predictions have high matching rates.

As a conclusion, our null hypothesis stated that there is a not correlation between election results and citizens level of development. With our graphs and correlation maps, it can be seen that there is a relation between these factors since their correlation rates is higher than the random 2 parameters. We can reject the null hypothesis. Finally, our hypothesis becomes true, and we could make predictions about the next result of given province with examining their socioeconomic conditions.

**BIBLIOGRAPHY**

* **Household income by province**

**Retrieved from**

* <https://data.tuik.gov.tr/Bulten/Index?p=Il-Bazinda-Gayrisafi-Yurt-Ici-Hasila-2021-45619> **at 10.06.2023**
* **High School Number  / Population by province**

**Retrieved from**

* <https://sgb.meb.gov.tr/meb_iys_dosyalar/2021_09/10141326_meb_istatistikleri_orgun_egitim_2020_2021.pdf> **at 10.06.2023**
* **Marriage age women/man by province**
  + [**https://data.tuik.gov.tr/Bulten/DownloadIstatistikselTablo?p=B8eRA2vkujeYknMy0TAYbqte6CyyXEFl9DfO61apy3dCfypG2ZTEHtM17/l9I8ve**](https://data.tuik.gov.tr/Bulten/DownloadIstatistikselTablo?p=B8eRA2vkujeYknMy0TAYbqte6CyyXEFl9DfO61apy3dCfypG2ZTEHtM17/l9I8ve)

**Retrieved from** [**https://data.tuik.gov.tr/Search/Search?text=%C4%B0llere%20G%C3%B6re%20Ortalama%20%C4%B0lk%20Evlenme%20Ya%C5%9F%C4%B1**](https://data.tuik.gov.tr/Search/Search?text=%C4%B0llere%20G%C3%B6re%20Ortalama%20%C4%B0lk%20Evlenme%20Ya%C5%9F%C4%B1) **at 10.06.2023**

* **Birth rate by province**
  + [**https://data.tuik.gov.tr/Bulten/DownloadIstatistikselTablo?p=cfo3rGQVEwQ15XGmRT/6HP2cQDCaed1/H6qeucMSqczQ2zFpoxw2hHbxDByqkRA9**](https://data.tuik.gov.tr/Bulten/DownloadIstatistikselTablo?p=cfo3rGQVEwQ15XGmRT/6HP2cQDCaed1/H6qeucMSqczQ2zFpoxw2hHbxDByqkRA9) **Retrieved from** [**https://data.tuik.gov.tr/Search/Search?text=%C4%B0llere%20g%C3%B6re%20toplam%20do%C4%9Furganl%C4%B1k%20h%C4%B1z%C4%B1**](https://data.tuik.gov.tr/Search/Search?text=%C4%B0llere%20g%C3%B6re%20toplam%20do%C4%9Furganl%C4%B1k%20h%C4%B1z%C4%B1) **at 10.06.2023**
* **Export by provinces**
  + [**https://data.tuik.gov.tr/Bulten/DownloadIstatistikselTablo?p=cr5fsUeZBI3JUeQpNsczZEYDUpmcg708g3YsdGeKrBwlEQaqy6azjRUDPR47QMTi**](https://data.tuik.gov.tr/Bulten/DownloadIstatistikselTablo?p=cr5fsUeZBI3JUeQpNsczZEYDUpmcg708g3YsdGeKrBwlEQaqy6azjRUDPR47QMTi)

**Retrieved from**

[**https://data.tuik.gov.tr/Search/Search?text=%C4%B0llere%20G%C3%B6re%20%C4%B0hracat**](https://data.tuik.gov.tr/Search/Search?text=%C4%B0llere%20G%C3%B6re%20%C4%B0hracat) **at 10.06.2023**

* **Starbucks data by province**
  + [**https://www.shaya.com.tr/tr/magazalarimiz/?topZone=32**](https://www.shaya.com.tr/tr/magazalarimiz/?topZone=32)

**Retrieved from** [**https://www.shaya.com.tr/tr/magazalarimiz/?topZone=32**](https://www.shaya.com.tr/tr/magazalarimiz/?topZone=32)

[**https://www.alshaya.com/en/locations/turkey/starbucks/**](https://www.alshaya.com/en/locations/turkey/starbucks/) **at 10.06.2023**

* **Parliamentary elections data by province**
  + [**https://secim2023.cnnturk.com/api/home**](https://secim2023.cnnturk.com/api/home)