PROPENSITY SCORE MATCHING IN THE ELECTION FORECASTING BY FOCUSING ON "UNDECIDED" VOTERS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ZELAL TOPRAK KÖSE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER
IN
STATISTICS

AUGUST 2022

Approval of the thesis:

PROPENSITY SCORE MATCHING IN THE ELECTION FORECASTING BY FOCUSING ON "UNDECIDED" VOTERS

submitted by **ZELAL TOPRAK KÖSE** in partial fulfillment of the requirements for the degree of **Master of Science** i**n Statistics Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Özlem İlk Dağ
Head of the Department, **Statistics** _____

Prof. Dr. Özlem İlk Dağ
Supervisor, **Statistics Department, METU** _____

Examining Committee Members:

Assoc. Prof. Dr. Rukiye Dağalp
Statistics Department, Ankara University _____

Prof. Dr. Özlem İlk Dağ
Statistics Department, METU _____

Assist. Prof. Fulya Gökalp Yavuz
Statistics Department, METU _____

Date: ...

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Last name : Zelal Toprak Köse

Signature :

ABSTRACT

PROPENSITY SCORE MATCHING IN THE ELECTION FORECASTING BY FOCUSING ON "UNDECIDED" VOTERS

Köse, Zelal Toprak
M.S., Department of Statistics
Supervisor : Prof. Özlem İlk Dağ

August 2022, 45 pages

How to deal with "undecided" voters is a big worry among those who conduct and analyze pre-election polls. This group makes it difficult to make inferences about elections so it is important to understand the behavior of undecided voters before making election forecasting. Undecided voters may determine the fate of the next elections in Turkey.

This thesis aims to predict the outcome of the next parliamentary election in Turkey using findings of opinion polls based on the undecided voters. In order to analyze the data set, Propensity Score Matching (PSM) is performed. After voters' propensity scores (PS) are calculated, the pairs between decided ("Treated") and undecided ("Control") voters are formed according to their PSs. In order to estimate propensity score, the logistic model, neural network and random forests are conducted. After estimating the propensity score, among the matching methods, the nearest neighbor matching, optimal matching and mahalanobis metric distance are used. With Nearest Neighbor Matching (NNM), for each undecided voter, the

algorithm finds a voter in the control group with the nearest PS. The proposed method of the thesis is the NNM with replacement using Mahalanobis distance because it reaches the most successful matching rate with 77.5%. With this method, while Public Alliance wins the next parliamentary election by 10.6 points according to the 2020 data, the Nation Alliance takes the lead by 4.2 points in 2021. Moreover, the balance in covariate distributions between treatment and control groups is checked by using some diagnostic methods.

Keywords: Election Forecasting, Propensity Score Analysis, Matching Methods, Mahalanobis Distance

ÖZ

SEÇİM TAHMİNLERİNDE "KARARSIZ" SEÇMENLERE ODAKLANARAK EĞİLİM PUANI EŞLEŞTİRME ANALİZİ

Köse, Zelal Toprak
Yüksek Lisans, İstatistik Bölümü
Tez Yöneticisi: Prof. Özlem İlk Dağ

Ağustos 2022, 45 sayfa

"Kararsız" seçmenlerle nasıl başa çıkılacağı, seçim öncesi anketleri yürüten ve analiz edenler arasında büyük bir endişe kaynağı. Kararsız seçmenler, seçimlerle ilgili çıkarımlarda bulunmayı zorlaştırdığı için seçim tahmini yapmadan önce bu grubun davranışlarını anlamak önemlidir. Türkiye'de bir sonraki seçimlerin kaderini kararsız seçmenler belirleyebilir.

Bu tez çalışması, kararsız seçmenlere dayalı kamuoyu yoklamalarının bulgularını kullanarak Türkiye'de bir sonraki milletvekili seçimlerinin sonucunu tahmin etmeyi amaçlamaktadır. Veri setini analiz etmek için Eğilim Puanı Eşleştirmesi - EPE analizi uygulanmıştır. Seçmenlerin eğilim puanları (EP) hesaplandıktan sonra, kararlı ("Vaka") ve kararsız ("Kontrol") seçmenler arasındaki eğilim puanlarına göre eşleştirilmiş çiftler oluşturulur. Eğilim puanını tahmin etmek için lojistik model, sinir ağları ve rastgele ormanlar kullanılmıştır. Eğilim puanını tahmin ettikten sonra ise eşleştirme yöntemlerinden en yakın komşu eşleştirilmesi, optimal eşleştirme ve mahalanobis metrik eşleştirmesi kullanılmıştır. En Yakın Komşu Eşleştirilmesi ile algoritma, kararsız her seçmen için kontrol grubunda en yakın EP'ye sahip bir seçmen bulur. Bu tez çalışmasında önerilen yöntem, %77,5 ile en başarılı eşleştirme

oranına ulaştığı için mahalanobis metriği kullanan ve yerine koyarak uygulanan en yakın komşu eşleştirmesidir. Bu yöntemle 2020 verisine göre Türkiye'de bir sonraki milletvekili seçimini Cumhur İttifakı 10,6 puan fark ile kazanırken, 2021 verisine göre Millet İttifakı 4,2 puan farkla öne geçiyor. Son olarak, vaka ve kontrol grupları arasındaki dengeyi değerlendirmek için bu gruplar arasındaki ortak değişken dağılımları kontrol edilmiştir.

Anahtar Kelimeler: Seçim Tahmini, Eğilim Puanı Analizi, Eşleştirme Metodları, Mahalanobis Metrik

To my beautiful family and friends

ACKNOWLEDGMENTS

First and foremost, I would like to thank my thesis supervisor Prof. Dr. Özlem İlk Dağ, who has supported me both throughout my university life and thesis process, has shared her great knowledge and experience with me, and always patiently has answered my questions during my thesis. This study would not have been possible without her constant support, guidance, and assistance.

I am very thankful to Prof. Dr. Özer Sencar who is the founder of MetroPOLL Strategic and Social Research Center and has accepted my study proposal when I first presented it, and contributed to my study during the data collection process. My very special thanks go to the entire MetroPOLL team, especially Özge Demir.

I want to thank to my friend Sena Ünal in our department for her beautiful friendship and helping me with every confusion. I wish to extend my special thanks to my friends İrem Yakışıklı, Ceren Nur İçöz and Cansu Çetin, who have always provided me with emotional support. Even though we study in different field, when I needed them, they were always there to listen and to share their experiences with me.

Lastly, my family deserves endless gratitude. I would like to thank my dear mummy and father İbrahim Kaan Erten who have supported me throughout my academic life, trusted me unconditionally and supported me in every way.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
AKP	Justice and Development Party
CATI	Computer-Assisted Telephone Interviewing
CHP	Republican People's Party
GBM	Generalized Boosting Model
HDP	Peoples' Democratic Party
M	Mahalanobis Distance
MHP	Nationalist Movement Party
NNM	Nearest Neighbor Matching
OOB	Out of Bag
PS	Propensity Scores
PSA	Propensity Score Analysis
PSM	Propensity Score Matching
TÜİK	Turkish Statistical Institute

CHAPTER 1

INTRODUCTION

The results of the elections are tried to be predicted with the research carried out before each parliamentary election in Turkey, and the results attract the attention of the public and politicians. These estimated results are important in terms of understanding the voting behavior of the voters and guiding the politicians. In this study, firstly, current problems are mentioned while making election predictions, and then it is aimed to deal with these problems. Among the studies in the literature on predicting next parliamentary elections in Turkey, no research has been found in which propensity score analysis is used. Our study is important because propensity score analysis (PSA) will be used for the first time in the field of politics, in Turkey.

1.1 Problem

If there is no early election, a parliamentary election will take place in 2023 in Turkey. However, the high undecided rate in the pre-election polls is a challenging issue for predicting election results. Various methods have been proposed to predict elections, but the most prevalent are those based on public opinion polls. However, using opinion polls has drawbacks, because they ignore the undecided voters (Liu et al., 2021). Politicians try to influence the undecided group in political campaigns because they believe this group is the most important target audience (Domenach, 1995). According to the data we collected for this study, in Turkey, the percentage of undecided voters remained around 20% which also passed the election threshold. This group is even defined as the second largest party in some news or statements.

Therefore, undecided voters are hard to ignore and they constitute a voter group with high analytical value. In this study, the Propensity Score Matching methods have been developed in order to both take into account the undecided voters and create a more practical election forecasting model.

1.2 Research Questions

1. What are the demographic characteristics of undecided voters?
2. Can we predict the voting behavior of the undecided group?
3. Which factors are important in behavior of the undecided voters?
4. How similar are undecided voters to decided voters?

CHAPTER 2

LITERATURE REVIEW

In this section, firstly the studies on voter preferences and election predictions in Turkey, and then some important studies conducted in recent years using PSA are presented.

2.1 Voter Preferences and Election Predictions

Sitembölükbaşı (2004) conducted a survey on voters who live in Isparta in 2004 and the author compared the 1995, 1999 and 2002 general elections in this study. As a result of these studies, it has been determined that the ideology of the party comes first among the situations that most affect the political party preferences of the voters, followed by the party leader and the actions of the parties. Another study aimed at understanding voter behavior was carried out by Canöz (2010) in the sample of Konya. In the findings, it was concluded that the voters care most about the candidates' personal characteristics such as honesty, reliability, positive image and having new projects. Data mining technique was used in the study conducted by Bayır et al. (2016), which is one of the studies on the prediction of political election results. They suggested that the algorithm of C4.5 Decision tree is an important auxiliary tool in determining political trends. In latest study conducted by Gündüz and Kıral (2020), using Markov analysis, it is aimed to estimate the votes of political parties in the next general elections and the level of party loyalty of their voters. This study was carried out only in Adana. According to the results of the analysis, it is predicted that there will be a decrease in the voting rates of the Justice and Development Party (AKP), Good Party and Peoples' Democratic Party (HDP) in the

long term, and an increase in the votes of the Nationalist Movement Party (MHP), Republican People's Party (CHP) and other parties. In this thesis, differently from the studies in the literature, our sample is not drawn from a single province but from 28 provinces of Turkey and the methodology of our study is explained in the next section.

2.2 Propensity Score Analysis

In the literature, it is seen that propensity score analysis can be used in different fields. For instance, it is used to observe the effect of morphine periprocedural usage on death in STEMI patients (Domokos et al., 2021), small school size on mathematics achievement (Wyse, Keesler, and Schneider, 2008), smoking on body mass index (Zhao et al., 2016), teenage alcohol use on education attainment (Staff, Patrick, Loken, and Maggs, 2008) and agri-environment (AE) programmes on input use and farm output of individual farms in Germany (Pufahl and Weiss, 2009). Also, with the help of the data obtained from the PISA 2015 Turkey sample, Yılmaz Koğar (2019) presented an example showing the use of propensity score matching in educational research. After different usage areas like these, we aim to adapt this method to politics.

CHAPTER 3

DATA AND EXPLORATORY ANALYSIS

The surveys are carried out by MetroPOLL Strategic and Social Research Center, using the stratified sampling in 28 provinces based on the 26 regions of Turkey's NUTS-2 system, as determined by Turkish Statistical Institute (TÜİK). All surveys use CATI (computer-assisted telephone interviewing) methodology with a margin error of 2.45 percent at the 95 percent level of confidence.

In order to work on undecided voters and then, predict next parliamentary elections, we compile 1 year data from January to December 2020. In this data, there are 17,626 voters and also these voters' demographic characteristics and political affiliation have been collected. All the estimation studies in this thesis are based on this 2020 data. However, at the end of the thesis, we apply the prediction methods that we worked on the 2020 data to the data collected in 2021, which has the same variables and methodology, and we examine how the results change.

The dataset for 2020 includes four ordered categorical variables: age, educational level, income level, voters' perceptions of Turkey's trajectory; and nine nominal categorical variables: gender, ethnicity, political identity, occupation, regions of Turkey, voters' alliance preferences in the previous and next parliamentary election, approval ratings for president Recep Tayyip Erdoğan and his popularity.

When the voters were asked which party they would vote for if there was a parliamentary election this Sunday, some of them stated that they were "undecided". Others have stated that they will either protest the ballot box or name the party they will vote for, and these voters are called "decided". There are 3,553 undecided and 14,073 decided voters in our dataset.

As seen in Table 3.1, the variables with no missing values are gender, age, educational level, regions of Turkey and alliance preferences in the next

parliamentary election. Other covariates contain missing values. Voters' ethnicity has 5.4% missing values. This rate is measured as 11.2% in political identity, 19.6% in occupation, 14% in income level, 2.5% in Turkey's trajectory, 10% in Erdoğan's approval and popularity, and 10.5% in alliance preferences in the previous parliamentary election. In order to deal with these missing values, missForest which is a random forest imputation algorithm is used, and this algorithm will be explained in Chapter 4.1.

Table 3.1 Missing values in data

Variable	Number of missing values	Percentage of missing values
Ethnicity	950	5.4
Political identity	1978	11.2
Occupation	3449	19.6
Income level	2461	14.0
Voters' perceptions of Turkey's trajectory	444	2.5
Erdoğan's approval	1770	10.0
Erdoğan's popularity	1779	10.1
Alliance preferences in the previous election	1856	10.5

3.1 If There Were An Election On Sunday…

Before any analysis is done on the undecideds, if there is an election on Sunday, the Nation Alliance's (CHP + İYİ Party) vote is 27.8% (see in Table 3.2). The Public Alliance (AKP + MHP) appears to reach 36% and HDP has 5.6%. The 2.9% tend to vote for other parties. While 7.6% of the voters state that they will protest the ballot box, 20.2% are undecided about which party they will vote for.

Table 3.3 shows the distribution of voters' current party preferences by gender, age group and educational level. According to the results of Table 3.3, the Nation and

Public Alliance draw votes from women at a slightly higher rate than the national average while the HDP does clearly better among men. If there is an election on Sunday, it is seen that women are more undecided than men about which alliance they will vote for. The Public Alliance draws less vote than average from 18-24 age group but above average among 35-44 age group. The Nation Alliance is strong among over-55s while HDP does best among the 35-44 age group. Being undecided and tendency to protest the ballot box comes mostly from young people. For the Public Alliance and HDP, votes fall as years of education rise. By contrast, the Nation Alliance see their votes increasing as education rises. Voters with more formal education appear to be more undecided.

Table 3.2 Current party preferences of voters

	Frequency (n)	**Percent (%)**
If there were a parliamentary election this Sunday, which political party would you vote for?		
Nation Alliance	4896	27.8
Public Alliance	6340	36.0
HDP	991	5.6
Other party	511	2.9
Protest vote	1335	7.6
Undecided	3553	20.2

Table 3.3 Distribution of voters' current party preferences by gender, age group and educational level

	Nation	Public	HDP	Other	Protest	Undecided	Total
By gender							
Female	28.3	37.1	3.3	1.5	7.6	22.2	100.0
Male	27.4	35.1	7.4	4.0	7.5	18.6	100.0
By age group							
18-24	26.1	33.8	4.3	2.6	9.3	23.8	100.0
25-34	26.2	35.8	5.8	2.7	8.7	20.8	100.0
35-44	24.7	38.4	7.5	3.2	7.7	18.6	100.0
45-54	28.5	36.3	5.3	3.6	7.0	19.3	100.0
55 and above	33.9	34.3	4.3	2.2	5.5	19.9	100.0
By educational level							
Middle school and below	21.6	43.9	6.2	2.7	6.5	19.1	100.0
High school	29.5	35.8	5.6	2.7	7.5	18.9	100.0
University and above	34.1	25.1	4.8	3.4	9.2	23.3	100.0

3.2 The Undecided

It is obvious that the undecideds have an important role in the upcoming elections. In such a raw form, the data indicates that the Public Alliance will win the elections. So, will this 8-point gap (mentioned in Chapter 3.1) between the Public and the Nation Alliance be closed according to the tendencies of the undecided voters? Which alliance will undecided voters favor? Will HDP be able to increase its votes? Or will these undecideds mostly protest the ballot box? In this study, we will try to find answers to all these questions. First of all, we need to get to know the undecideds closely. We will examine them both by their previous voting behavior and by their demographics.

We said that the undecided make up 20.2%. When we look at the party preferences of undecided voters in the previous parliamentary elections, this group is made up of 6 points from the Public Alliance, 3.5 points from Nation Alliance and 10.1 points from those who protested the ballot box. Almost half of them did not go to the polls in the previous election. Undecided voters seem to be mostly inclined to protest the ballot box. We find 6 points are former AK Party and MHP voters. Will these former government votes make the crossing to join the opposing alliance? If this does not happen by the time of the election, it represents an advantage for the ruling parties and a serious risk for the opposition.

Table 3.4 Undecided voters in terms of the vote cast in the June 24, 2018 parliamentary elections

	20.2% Undecideds Point Distribution	**Percent (%)**
Nation Alliance	3.5	17.5
Public Alliance	6.0	29.4
HDP	0.5	2.7
Other party	0.1	0.7
Protest vote	10.1	49.7

When we look at the demographic characteristics of the undecideds in Table 3.5, the modal class of them are men, 35-44 age group and less formal education. The percentages of undecideds having these demographic characteristics are close to each other. Majority of them are conservative. When we look at which occupational groups the undecideds come from, most of them are employees, housewives, retirees or other occupational groups. Also, they mostly come from voters with earnings under 3000 TL.

According to Table 3.5, 62% of the undecided voters say the nation is moving in the wrong direction. Only 20% of them think Turkey is changing for the better. This means that the undecideds are quite pessimistic about the country. They may have negative views on topics such as the economy, the cost of living, the state of nation

and the future. The 45.8% of undecideds disapprove of the way President Erdoğan is handling his job as president and 43.1% of them do not admire him. That said, half of them seem not to have completely abandoned hope in Erdoğan and the AK Party.

Table 3.5 Distribution of undecided voters by demographic characteristics

	20.2% Undecideds Point Distribution	**Percent (%)**
By gender		
Female	9.7	48.1
Male	10.5	51.9
By age group		
18-24	3.3	16.2
25-34	4.5	22.5
35-44	4.7	23.2
45-54	3.7	18.5
55 and above	4.0	19.6
By educational level		
Middle school and below	7.1	35.2
High school	6.9	33.9
University and above	6.2	30.9
By occupation		
Housewife	4.1	20.5
Employee	3.6	17.8
Retired	2.5	12.6
Student	2.1	10.3
Unemployed	1.7	8.2
Other	6.2	30.6
Income level		
3000 TL and below	11.8	58.4
3001 TL and above	8.4	41.6

Table 3.5 Distribution of undecided voters by demographic characteristics, cont.

	20.2% Undecideds Point Distribution	**Percent (%)**
Political identity		
Conservative	6.9	34.2
Nationalist	5.6	27.8
Secular	4.6	22.9
Other	3.1	15.1
Voters' perceptions of Turkey's trajectory		
Better	4.1	20.1
Worse	12.5	61.9
Neutral	3.6	18.0
Erdoğan's approval		
Approve	10.9	54.2
Disapprove	9.3	45.8
Erdoğan's popularity		
I admire	11.5	56.9
I do not admire	8.7	43.1

CHAPTER 4

METHODS

In this section, we propose an algorithm to deal with missing values, estimate the propensity score using different models, and propose matching methods based on these propensity scores we get.

4.1 Missing Value Analysis

Data has been examined for its completeness using missForest which is a random forest imputation algorithm. Stekhoven and Bühlmann (2012) propose missForest and state that this method performs better on data complex interactions and non-linear relations, can handle any type of data and requires very few assumptions compered to other methods of imputation. This imputation algorithm first imputes all missing data using the mean/mode. Then, for each feature with missing values, missForest applies a random forest model on the observed part and then predicts the missing part.

4.2 Propensity Score Analysis

Random assignment may not be possible in most social and educational studies due to logistical, ethical, political, and economic reasons (Bostian, 2008; Luellen, Shadish and Clark, 2005; Titus, 2007). The propensity score is a conditional probability that is used in situations where random assignment is unlikely or developed to derive causal results from observational data (Rosenbaum and Rubin, 1983, 1984). In order to help researchers cope with such problems, propensity score analysis was developed by Rosenbaum and Rubin. The propensity score can be used to observe matching, stratification, weighting or covariate adjustment in regression

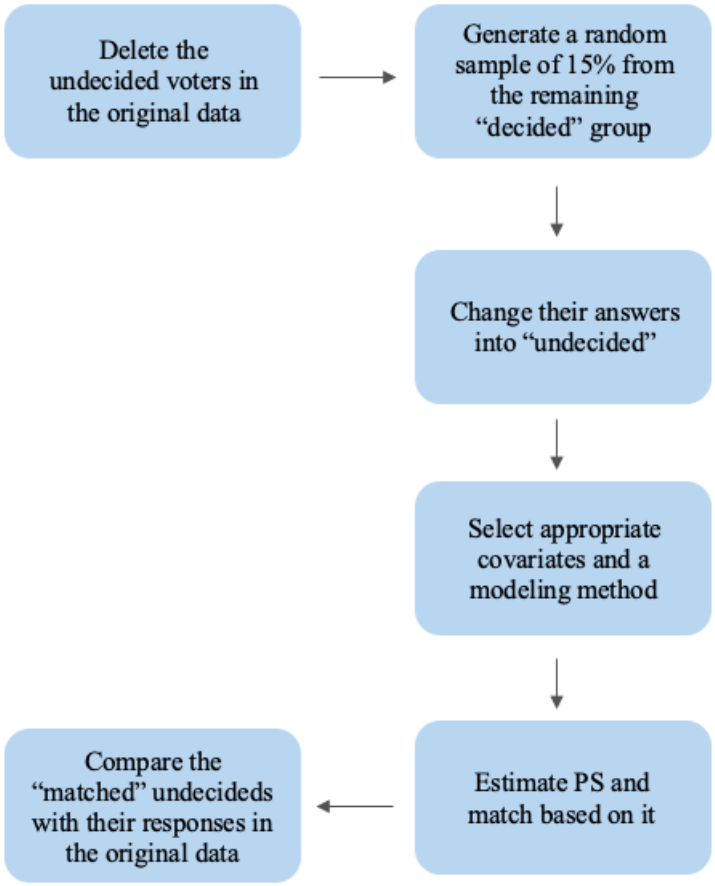
(Leite, 2016). In this paper, we focus on propensity score matching (PSM) and it is later explained in detail.

We mentioned earlier that propensity score analysis is used in different fields. We will use this method, which is used in many fields, by adapting it to politics. In other words, we try to match undecided voters with decided voters based on their propensity scores.

We conduct the propensity score matching (PSM) in 4 steps. The first step is to deal with missing values to be able to prepare data for the analysis. In the second step, a model is specified, such as logistic regression or random forests. In the third step, different matching methods are carried out, such as nearest neighbor or optimal matching. In the last step, by observing balance diagnostics, the balance of covariates between treatment and control groups is checked (Zhao et al., 2016).

In order to understand whether Propensity Score Matching can match correctly, firstly, undecided voters in the original data were deleted. A random sample of 15% from the remaining "decided" group is generated and their answers are changed into "undecided". That is, we actually know the decisions or real responses of the undecided voters in this sample. After selecting appropriate covariates and a modeling method, we estimate PS and match based on it (Zhao et al., 2016). Finally, we compare the "matched" undecideds with their responses in the original data to evaulate the performances of our proposed methods. For better understanding, all these stages are illustrated in Figure 4.1.

Figure 4.1 Steps for Propensity Score Matching



4.3 Propensity Score Estimation

After variable selection, a model is conducted in which the group variable (decided or undecided voters) is considered as the dependent and the co-variables as the independent variables. Logit or probit models can be used in model selection. Since the distribution of logit PS approximates to normal, using the logit of PS to match samples is proposed by Rosenbaum and Rubin (1985). Moreover, in the literature, it is seen that logit models are used more frequently (Fan and Nowell, 2011) due to their nice odds ratio interpretation.

In order to estimate propensity score, the logistic model is said to be advantageous, easily interpretable and applicable using common statistical software packages

(Setoguchi et al., 2008; Westreich et al., 2010). In addition to the logistic model, different methods are also suggested, for example, bagging or boosting (Lee et al., 2010), recursive partitioning algorithms (Lee et al., 2010), the generalized boosting model (GBM) (McCaffrey et al., 2004), random forests (Lee et al., 2010), and artificial neural network (ANN) (Setoguchi et al., 2008). Among these methods, we worked with the logistic model, ANN and random forests. All steps are executed on R and "matchit" function in MatchIt is used to estimate propensity score and then match voters. After Chapter 4.3.1, there is a brief description of the models we use.

The most important argument in deciding which model to use to estimate PS is "distance". The distance between treated and control units is important for matching methods. In order to set the argument of "distance", we include "logit" for logistic regression and "nnet" for neural network model. Another way to calculate distance is to use PS estimated by the researcher. Adding the "mahalanobis" option to matchit() can be given as an example. Mahalanobis distance, which calculates distances in a multidimensional space, is used to conduct Mahalanobis metric distance matching (Guo and Fraser, 2015). Also, we calculate distance using our own estimated PS after building a random forest model (Zhao et al., 2016).

Another argument that can be added to the model is "ratio". In one-to-one (1:1) ratio, each treated voter matches to a single control (Zhao et al., 2016). In many-to-one (M:1) matching on the propensity score, M controls are matched to each treated subject using the propensity score (Austin, 2010).

4.3.1 Logistic model

The estimated propensity score is the predicted probability of getting the treatment evolved from the fitted logistic regression model (Ali et al.,2019). Although this model is frequently used and easy to apply, it may not be advantageous in terms of time management because different polynomial or interaction terms must be need to be added to the model in order to balance the data (Zhao et al., 2016). When these

terms are included, the performance of the logistic regression increases and approaches that of the other machine learning methods (Ali et al., 2018).

Since our dependent variable consists of two categories, decided and undecided voters, we use binary logistic regression to obtain propensity scores. Then, using the different matching methods, which we will describe later, we match voters with similar PSs. The definition of the logistic model is given below:

$$\ln \frac{Pr(Z_i = 1|X_j = x_j)}{1 - Pr(Z_i = 1|X_j = x_j)} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_kx_k \quad (4.1)$$

where:

b_0 is the intercept

b_i are the slope coefficents

X_j, the treatment variables and covariates

x_j, observed value of variables

i=1,…, n; j= 1, …, k

$$Z_i = \begin{cases} 1 & , \text{ treatment} \\ 0 & , \text{ control} \end{cases}$$

In logistic regression, the dependent variable is binary, $Z_i = 1$ is the value for the treatment and the value for the control is $Z_i = 0$.

4.3.2 Neural Networks

A neural network is an algorithm inspired by the structure of the nervous system. A neural network consists of an input layer, a series of "hidden" layers, and an output

layer, each of which has a number of nodes connected to every node in the next layer by directed, weighted edges (Westreich et al., 2010).

When a suitable method cannot be found to achieve data balance or a more practical method is desired instead of adding polynomial or interaction terms, a neural network can be an alternative technique to logistic regression in order to estimate propensity scores (Keller et al., 2013; Westreich et al., 2010; Setoguchi et al., 2008). Considering the disadvantages, it is criticized because of their poor interpretability (Leite W., 2016). There is no simple rule to determine the number of hidden nodes in a network and avoid overfitting. In this study, the best number of hidden layers is determined by trial and error.

4.3.3 Random Forests

Some problems are not considered in the literature when applying propensity score based matching methods, such as nonlinear relationships, misspecified models, categorical variables with more than two levels and missing data. Random forest is proposed considering that it has the ability to solve these problems (Zhao et al., 2016) and it gives more accurate and less model dependent estimate of the propensity score (Caruana et al., 2008). However, if the fraction of treated units is low, random forests cannot estimate the propensity score well enough to remove the selection bias (Goller et al., 2019).

In this study, we also use random forests to rank the importance of variables in our data. First, we observe the matching methods with all the variables. Then, using the important variables we identified, we examine whether there is an improvement in our matching rate.

4.4 Matching Methods

After propensity scores are calculated, a matching process is performed. We use different methods for matching currently available in MatchIt to see which method works best for our study. A brief summary of the methods used is given below:

4.4.1 Nearest Neighbor Matching (NNM)

After voters' PS are calculated, with NNM, the algorithm picks voters, one by one from treatment groups, in a specified order. Then, for each undecided voter, the algorithm finds a voter in the control group with the nearest PS (Zhao et al., 2016).

NNM can be applied in two different ways. The first one is matching without replacement. Using this approach, we match each voter from the control group to at most one voter from the treatment group. In contrast, matching with replacement allows us to match the same voter from the control group to multiple voters from the treatment group (Austin, 2014).

4.4.2 Optimal Matching

The main purpose of this approach is to find the matched samples with the smallest average absolute propensity score distance for all the matched pairs (Gu and Rosenbaum, 1993).In other words, this approach aims to minimize the average of the distances between treatment and control individuals within each set. A very important feature of this approach is that the match between participants can change if it is found that a different match will further minimize the total distance (Olmos and Govindasamy, 2015).

CHAPTER 5

RESULTS

In this chapter, the steps followed for propensity score matching and their results are mentioned respectively. In the first stage, statistical models are conducted using all variables and the matching results are examined. Then, the same steps are followed using only the important features.

5.1 Propensity Score Matching with Logistic Regression

Table 5.1 shows the results of logistic regression established with all variables for each matching method. As mentioned earlier, nearest neighbor matching was applied in two different ways. In matching with replacement, when we compare the "matched" undecideds with their responses in the original data, they match 55% accurately. When we look at the results of matching without replacement, they match 53% correctly. If we change the method to optimal, the successful matching decreases to 37%. On the other hand, if we conduct matching using Mahalanobis distance, the correct match is increased to 75% and this is the method that reaches the highest correct match. For all matching methods, when the ratio is set to anything other than one-to-one, the successful matching rate is observed to decrease. In other words, one-to-one matching gives more successful matching in comparison to other ratios.

Table 5.1 Results of logistic regression for each matching method

Estimate PS by Using Model	Matching Method	Replace	Ratio	Distance	Accurate Matching
Logistic regression	NN	with	1:1	'logit'	54.8
	NN	without	1:1	'logit'	52.7
	Optimal	-	1:1	'logit'	36.6
	NN	with	1:1	M	75.4
	NN	without	1:1	M	74.9
	NN	with	2:1	'logit'	46.0
	NN	without	2:1	'logit'	43.9
	Optimal	-	2:1	'logit'	35.1
	NN	with	2:1	M	72.1
	NN	without	2:1	M	70.1
	NN	with	3:1	'logit'	44.8
	NN	without	3:1	'logit'	41.0
	Optimal	-	3:1	'logit'	35.1
	NN	with	3:1	M	69.6
	NN	without	3:1	M	65.6

NN, Nearest Neighbor Matching. M, Mahalanobis Distance.

5.2 Propensity Score Matching with Neural Network

As mentioned earlier, another model used to estimate PS is the neural network. At this stage, the distance argument is changed to "nnet", and the number of hidden-layer units is set to 16, determined by trial and error. The results of this model are shown in Table 5.2. As with the logistic model, the best results occur in a 1:1 matching. When the optimal matching method is applied, the accurate matching rate is 39%, while it increases to about 50% in the nearest neighbor matching. With the nearest neighbor matching with replacement in the 1:1 ratio, the correct match reaches its highest level of 57%.

Table 5.2 Results of neural network for each matching method

Estimate PS by Using Model	Matching Method	Replace	Ratio	Distance	Accurate Matching
Neural network	NN	with	1:1	‘nnet’	57.4
	NN	without	1:1	‘nnet’	52.2
	Optimal	-	1:1	‘nnet’	38.6
	NN	with	2:1	‘nnet’	47.5
	NN	without	2:1	‘nnet’	42.6
	Optimal	-	2:1	‘nnet’	33.5
	NN	with	3:1	‘nnet’	49.0
	NN	without	3:1	‘nnet’	36.3
	Optimal	-	3:1	‘nnet’	34.0

NN, Nearest Neighbor Matching.

5.3 Propensity Score Matching with Random Forests

So far we have used the built-in functions included in the "MatchIt", but we can also estimate PS ourselves. Before estimating our own PS, we conduct a random forest model and observe a confusion matrix to examine the performance of this model (see Table 5.3). According to Table 5.3, the number of trees is 500 in the model and the number of variables tried at each split are 3. Classification error in Nation is 0.105, i.e., 11%, Public is 6%, HDP is 23%, other is 95% and protest is 72%. Classification error is quite high in the categories of other and protest vote. This can be explained by the insufficient sample size (Janitza and Hornung R., 2018). Accuracy for train data is 95%. That indicates 95% of the values classified correctly while test data accuracy is measured as 82%.

After we conduct random forests, we use the treatment probabilities to get propensity scores. We now enter our own estimated PS as a user defined option “eps” (in Table 5.4) into the distance argument in the matching formula (Zhao et al., 2016). In this scenario, with the nearest neighbor matching without replacement in the 3:1 ratio, the accurate matching reaches its highest level which is 45%. However, we observe that matching with random forests gives lower matching rates when compared to the

results of other models. Also, in the logistic and neural network models, the nearest neighbor matching method with replacement gives the higher results compared to the other matching methods, while in the random forest model, the nearest neighbor matching method without replacement is higher.

Table 5.3 Summary of the random forest model and performance of this model

Arguments	**Outputs**
Number of trees	500
No. of variables tried at each split	3
OOB estimate of error rate	17.94%
Classification error in Nation	0.105
Classification error in Public	0.055
Classification error in HDP	0.225
Classification error in Other	0.950
Classification error in Protest vote	0.718
Accuracy for train data	0.947
Accuracy for test data	0.820

Table 5.4 Results of random forests for each matching method

Estimate PS by Using Model	**Matching Method**	**Replace**	**Ratio**	**Distance**	**Accurate Matching**
Random forest	NN	with	1:1	eps	37.4
	NN	without	1:1	eps	44.2
	Optimal	-	1:1	eps	39.5
	NN	with	2:1	eps	34.9
	NN	without	2:1	eps	44.3
	Optimal	-	2:1	eps	42.0
	NN	with	3:1	eps	38.2
	NN	without	3:1	eps	44.9
	Optimal	-	3:1	eps	39.9

NN, Nearest Neighbor Matching. eps, our own estimated PS by using random forests.

5.4 Distribution of the undecided

After observing the accurate matching rates of the methods, we distribute the undecideds according to these methods. Since we also know the actual answers of these undecided voters, we compare the future election preferences of these voters in the original data with the results after distributing according to the methods. In Table 5.5, the first column shows the future election preferences in the original data where we deleted the real undecideds. The remaining columns show the results after distributing the undecideds by matching methods.

When we look at the logistic and neural network models in Table 5.5, we see that we get results close to the real results for all matching methods. Only 1% deviation is observed for the Public Alliance in the optimal method for logistic regression. On the other hand, for random forest model, the nearest neighbor matching with replacement has close to 5% deviation for Nation and Public Alliance and this may be a significant deviation for the election forecasting. There are differences in the correct matching rates of the methods, but when we distribute the undecideds into the overall percentages, we get results that are quite close to the real ones. Therefore, in the final stage, we can distribute the real undecided voters according to these methods and make our election predictions.

Table 5.5 Results of preferences in the future election after distributing the undecideds according to the matching methods

		Logistic Regression					**Neural Network**			**Random Forests**		
	Original	**NNM with rep**	**NNM w/o rep**	**Opt**	**NNM with rep (M)**	**NNM w/o rep (M)**	**NNM with rep**	**NNM w/o rep**	**Opt**	**NNM with rep.**	**NNM w/o rep**	**Opt**
*Preferences In The Future Election**												
Nation Alliance	34.8	34.6	34.6	34.8	34.7	34.7	34.7	34.8	34.4	40.4	34.2	34.7
Public Alliance	45.1	45.6	45.5	44.0	45.6	45.5	45.3	45.0	45.6	40.3	46.1	45.2
HDP	7.0	7.0	7.0	7.1	7.0	7.0	7.2	7.2	7.1	6.9	6.9	6.7
Other party	3.6	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.5	3.7	3.6
Protest vote	9.5	9.2	9.2	10.1	9.1	9.1	9.1	9.2	9.2	8.9	9.1	9.5

NNM with rep, Nearest Neighbor Matching with replacement. NNM w/o rep, Nearest Neighbor Matching without replacement. Opt, Optimal Matching. M, Mahalanobis Distance.

* Voters were asked which party they would vote for if there was a parliamentary election this Sunday.

5.5 Variable Importance

In order to determine the order of importance of the variables, we observe the mean decrease in Gini coefficient. Table 5.6 clearly shows that the party preferences in the previous elections is the most important feature followed by Erdoğan's approval and his popularity. These features, which are at the top of the order of importance, can also be factors that affect the voting behavior of voters. From this point of view, it can also make it easier to understand the behavior of undecided voters. In the next part of the thesis, we will update all matching methods by using the important features that we have identified and see whether there is an improvement in the methods.

Table 5.6 Mean decrease in Gini coefficient

Variables	**MeanDecreaseGini**
Alliance preferences in the previous election	1572.70985
Erdoğan's approval	808.78557
Erdoğan's popularity	785.56020
Regions of Turkey	483.99869
Occupation	392.85602
Political identity	392.62980
Voters' perceptions of Turkey's trajectory	316.56181
Ethnicity	270.61637
Age	234.29498
Educational level	164.83018
Income level	103.55332
Gender	91.12856

5.6 Continue with important variables

We reapply all the matching methods mentioned so far by selecting the top 7 most important variables ranked according to the mean decrease in Gini coefficient. These variables are voters' alliance preferences in the previous and next parliamentary election, approval ratings for president Recep Tayyip Erdoğan and his popularity, regions of Turkey, occupation, political identity and voters' perceptions of Turkey's trajectory. By doing this, we observe an improvement in methods. Since many-to-one (M:1) matching results gives more unsuccessful matching rates, only the results of the one-to-one (1:1) matching are shown in Table 5.7. According to the table, if we continue the analysis with important variables, for all matching methods, we can see accurate matching rate increases. In this scenario, if the nearest neighbor matching with replacement using Mahalanobis distance is conducted, it results in the most successful match with 77.5%.

Table 5.7 Results of 1:1 matching methods used only with important variables

Estimate PS by Using Model	**Matching Method**	**Replace**	**Ratio**	**Distance**	**Accurate Matching**
Logistic regression	NN	with	1:1	'logit'	74.7
	NN	without	1:1	'logit'	73.2
	Optimal	-	1:1	'logit'	38.0
	NN	with	1:1	M	77.5
	NN	without	1:1	M	77.1
Neural network	NN	with	1:1	'nnet'	71.5
	NN	without	1:1	'nnet'	66.9
	Optimal	-	1:1	'nnet'	46.2
Random forest	NN	with	1:1	eps	38.2
	NN	without	1:1	eps	45.2
	Optimal	-	1:1	eps	39.6

NN, Nearest Neighbor Matching. M, Mahalanobis Distance. eps, our own estimated PS by using random forests.

5.7 Balance Diagnostics

A new dataset is created after matching based on PS is applied; however, data balance may not be achieved. Therefore, we calculate the standardized differences to check the balance in covariate distributions between treatment and control groups, and for this, we use the data which is created after using the nearest neighbor matching with replacement using Mahalanobis distance (mentioned in Chapter 5.6). Since all the variables in our data are categorical, the standardized difference is defined as

$$d = \frac{(\hat{p}_{treatment} - \hat{p}_{control})}{\sqrt{\frac{\hat{p}_{treatment}(1-\hat{p}_{treatment}) + \hat{p}_{control}(1-\hat{p}_{control})}{2}}} \qquad (5.1)$$

where $\hat{p}_{treatment}$ and $\hat{p}_{control}$ state the prevalences of the dichotomous variable in treatment and control groups, respectively (Austin P. C., 2009).

In order to avoid confusion, instead of giving the standardized difference values for all matching methods, we give the results of the method that achieves highest accurate matching rate in Table 5.8. According to the results in Table 5.8, in our matched sample, the largest absolute standardized difference with 0.253 is seen in the voters who did not go to the polls in the last elections. This is followed by those who support the Public Alliance in the previous elections with 0.155. All these results seem pretty close to zero. Therefore, we can say that our inferences after using PS-based matching are trustworthy because in our matched sample, treatment and control groups have similar distributions of measured baseline covariates (Austin P. C., 2009).

Table 5.8 Standardized difference between treatment and control subjects in the matched sample.

Variable	**Undecided (N=3553)**	**Decided (N=1643)**	**Standardized difference**
Political Identity			
Conservative	1215 (34.2%)	513 (31.2%)	0.063
Nationalist	988 (27.8%)	509 (31.0%)	-0.070
Secular	815 (22.9%)	350 (21.3%)	0.039
Other	535 (15.1%)	271 (16.5%)	-0.039
Turkey's Trajectory			
Better	714 (20.1%)	361 (22.0%)	-0.046
Worse	2198 (61.9%)	1025 (62.4%)	-0.011
Neutral	641 (18.0%)	257 (15.6%)	0.064
Erdoğan's Approval			
Approve	1927 (54.2%)	868 (52.8%)	0.028
Disapprove	1626 (45.8%)	775 (47.2%)	-0.028
Erdoğan's Popularity Ratings			
I admire	2022 (56.9%)	928 (56.5%)	0.009
I do not admire	1531 (43.1%)	715 (43.5%)	-0.009
Party preferences in the last election			
Nation Alliance	621 (17.5%)	352 (21.4%)	-0.100
Public Alliance	1046 (29.4%)	603 (36.7%)	-0.155
HDP	95 (2.7%)	55 (3.4%)	-0.039
Other	24 (0.7%)	20 (1.2%)	-0.056
Protest vote	1767 (49.7%)	613 (37.3%)	0.253

5.8 Distribution of the undecided in the final data

We mentioned that the size of our original data was 17,626 and 20.2% consisted of undecided voters. From the methods we have worked with so far, we have followed a path on how to distribute the undecided voters, and in this chapter, we distribute the "real undecideds" with these methods. In Table 5.9, the first column shows the future election preferences in the original data and the second column shows the results in which the undecideds are proportionally distributed. The remaining columns give the results after distributing the undecideds by matching methods.

According to Table 5.9, after distributing the undecided voters with the nearest neighbor matching with replacement using Mahalanobis distance, the Public Alliance has 44.2%, the Nation Alliance 33.6%, the HDP 6.6% and other parties 3.9% in the upcoming parliamentary election. The rate of protesting the ballot box is 11.7%. When we compare these results with the results in the 2nd column of Table 5.9, a decrease of approximately 1 point has been observed in both the Public Alliance and the Nation Alliance. There is an increase of 2.2 points in the protest votes. Similar alliance preferences in the next election are observed in other matching methods, except for the random forest model with the nearest neighbor matching with replacement.

Table 5.9 Results of alliance preferences in the future election after distributing the undecideds according to the matching methods in the original data

			Logistic Regression					**Neural Network**			**Random Forests**		
	Orig.	**Prop. Dist.**	**NNM with rep**	**NNM w/o rep**	**Opt**	**NNM with rep (M)**	**NNM w/o rep (M)**	**NNM with rep**	**NNM w/o rep**	**Opt**	**NNM with rep.**	**NNM w/o rep**	**Opt**
*Preferences In The Future Election**													
Nation Alliance	27.8	34.8	33.5	33.9	34.0	33.6	33.9	33.4	33.8	33.8	44.0	35.6	35.7
Public Alliance	36.0	45.1	43.6	43.6	43.1	44.2	44.6	44.2	43.9	43.6	37.1	44.9	43.2
HDP	5.6	7.0	6.7	6.7	6.7	6.6	6.6	6.6	6.6	6.6	6.4	6.4	6.6
Other party	2.9	3.6	3.9	3.8	4.0	3.9	3.8	3.8	3.9	4.1	3.5	3.5	3.9
Protest vote	7.6	9.5	12.3	12.0	12.2	11.7	11.2	12.1	11.8	11.9	9.0	9.5	10.6
Undecided	20.2												

NNM with rep, Nearest Neighbor Matching with replacement. NNM w/o rep, Nearest Neighbor Matching without replacement. Opt, Optimal Matching. M, Mahalanobis Distance.

* Voters were asked which party they would vote for if there was a parliamentary election this Sunday.

5.9 How do the results change in the 2021 data?

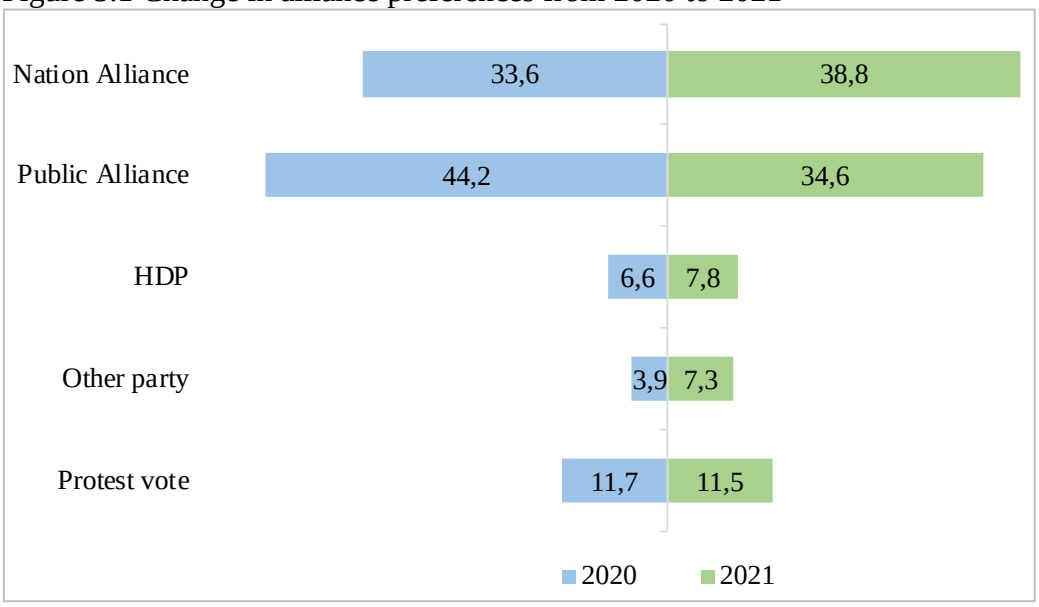
In this study, we have worked on the 2020 data to create our prediction model, but we also want to see how the results would change if we apply the same model to the 2021 data with the same variables. In 2021 data, there are 18,451 voters and the data includes voters' demographic characteristics and political affiliation as in 2020 data. There is no change in the methodology either. 2021 data have been collected by stratified sampling based on Turkey's NUTS-2 system and CATI is used.

When we examine the important features of the 2021 data by observing the mean decrease in Gini coefficient, we see that the most important feature is again the party preferences in the previous elections. While in 2020 data, the second most important feature is Erdoğan's approval, in 2021 data it is regions of Turkey. Although there are minor changes in the order of importance of the variables in the 2020 and 2021 data, when we look at the 7 most important variables, we observe the same variables (mentioned in Chapter 5.6). So, how do the election prediction results change when we apply the same model to the data of 2021 with the same variables?

Figure 5.1 shows the alliance preferences in the future elections and compares the 2020 and 2021 data. Nearest neighbor matching with replacement using Mahalanobis distance is applied to both data. According to Figure 5.1, from 2020 to 2021, while the Nation Alliance increased its votes by 5.2 points, the votes of the Public Alliance decreased by 9.6 points. HDP's support increased by only 1.2 points, while other small parties increased by 3.4 points.

It seems that all balances are changing in 2021 data compared to 2020. The Nation Alliance seems to win the election with 4.2 points. On the side of the Public Alliance, the gains recorded in 2020 have gone. The economic crisis, 'negative' perceptions of the state of the nation, the events in foreign policy may have caused the Public Alliance to lose 10 points. The increase in the Nation Alliance votes can be interpreted as the voters' desire for government change.

Figure 5.1 Change in alliance preferences from 2020 to 2021



CHAPTER 6

CONCLUSION

If they are held on time, the presidential and parliamentary elections are a little over a year away. An early election is still not out of the question. In any case, it is a question of what will be the result of the election. Although different methods are tried in the literature for election forecasting, mostly undecided voters are not taken into account. Considering that the rate of undecided voters is around 20% in our data, it is important to understand the voting behavior of this population. In this study, in order to apply the election forecasting methods, we have worked on the data collected by Metropoll Research Center, using the stratified sampling in the 26 regions of Turkey's NUTS-2 system throughout the year 2020. Also, we have examined the demographic structure of the undecided, their voting behavior, and their perceptions of the state of the nation and the president.

According to 2020 data, it has been observed that mostly men, the age group of 35-44, voters with less formal education, conservatives, housewives, employees, retirees and voters with an income of less than 3000 TL are undecided. Pessimism reaches 62% among the undecided and almost half of them disapprove of the way President Erdoğan is handling his job. Moreover, half of the undecideds did not go to the polls in the past parliamentary elections and 29.4% supported the Public Alliance.

Based on these information about the undecideds, we developed the method of matching the undecided voters with the decided ones. With propensity score matching, undecided voters with similar characteristics were matched with decided voters. In this study, the distribution of undecideds to alliance preferences is distributed in this way, instead of proportional distribution. After determining our model such as logistic regression, random forests or neural networks, we have tried

different matching methods. In order to find the accurate matching rate, we deleted the undecided voters from the main data and then changed the 20% of the remaining data to undecided. In the end, when we use the most important 7 variables, the nearest neighbor matching with replacement using Mahalanobis distance reaches the most successful match with 77.5%. In other words, comparing the "matched" undecideds with their responses in the original data, it matches 77.5% accurately. Also, with this method, a data balance is achieved. When 20.2% of the real undecideds in the original data are distributed to the alliance preferences according to this method, the Public Alliance wins the election by 10.6 points difference to the Nation Alliance. If the HDP stands alone, its vote is measured at 6.6% and other parties at 3.9%. 11.7% of the voters protest the ballot box.

When we look at the alliance preferences in the next election after distributing the undecideds to the alliances based on all other matching methods, similar results are observed, except for the random forest model with the nearest neighbor matching with replacement. However, in the literature, we could not find a valid reason why the random forest model with NNM with replacement gives different results from other matching methods. From the accurate matching rates of the methods, we also see that this method is unsuccessful, therefore we do not recommend this method for our data.

After studying the 2020 data and observing the election predictions, we use the same methods on the 2021 data contains the same variables. In this scenario, the Nation Alliance wins the election by 4.2 points with 38.8% of votes and it has gained 5.2 points from the previous year. The Public Alliance's vote is 34.6% and this is a fall of 9.6 points from 2020. In 2021, the HDP has %7.8 and other parties have 7.2%. There is no significant change in the voter turnout rate from 2020 to 2021.

REFERENCES

Ali, M., Khalid, S., Collins, G., & Prieto-Alhambra, D. (2018). The comparative performance of logistic regression and random forest in propensity score methods: A simulation study. 33rd International Conference on Pharmacoepidemiology and Therapeutic Risk Management(ICPE 2017), 26(S2), 489.

Ali, M. S., Prieto-Alhambra, D., Lopes, L. C., Ramos, D., Bispo, N., Ichihara, M. Y., Pescarini, J. M., Williamson, E., Fiaccone, R. L., Barreto, M. L., & Smeeth, L. (2019). Propensity Score Methods in Health Technology Assessment: Principles, Extended Applications, and Recent Advances. Frontiers in pharmacology, 10, 973. https://doi.org/10.3389/fphar.2019.00973

Austin P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statistics in medicine, 28(25), 3083–3107. https://doi.org/10.1002/sim.3697

Austin P. C. (2010). Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. American journal of epidemiology, 172(9), 1092–1097. https://doi.org/10.1093/aje/kwq224

Austin P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. Statistics in medicine, 33(6), 1057–1069. https://doi.org/10.1002/sim.6004

Bayır A., Özdemir Ş., Gülseçen S. (2016), Türkiye'deki Seçmen Eğilimlerinin C4.5 Karar Ağacı Algoritması İle Belirlenmesi, Yönetim Bilişim Sistemleri Dergisi, 3(1).

Bostian, B. E. (2008). Avoiding remedial education: Academic effects on college transfer students. (Doctoral dissertation). ProQuest Dissertations and Theses. (Accession Order No. AAT 3320959)

Canöz, K. (2010). Seçmen Tercihinde Aday İmajının Rolü: 29 Mart 2009 Yerel Seçimleri Öncesinde Konya Seçmeni Üzerine Bir Araştırma, Selçuk Üniversitesi İletişim Fakültesi Akademik Dergisi, 6(2): 95-114.

Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. In: Proceedings of the 25th International Conference on Machine learning, 2008, pp. 96–103. ACM.

Domenach J M (1995) Politika ve Propaganda, Tahsin Yücel (çev), Varlık Yayınları, İstanbul.

Domokos, D., Szabo, A., Banhegyi, G., Major, L., Kiss, R. G., Becker, D., Edes, I. F., Ruzsa, Z., Merkely, B., & Hizoh, I. (2021). Impact of periprocedural morphine use on mortality in STEMI patients treated with primary PCI. PloS one, 16(1), e0245433.

Fan, X., & Nowell, D. L. (2011). Using propensity score matching in educational research. Gifted Child Quarterly, 55(1), 74-79.

Goller, Daniel; Lechner, Michael; Moczall, Andreas; Wolff, Joachim (2019) : Does the Estimation of the Propensity Score by Machine Learning Improve Matching Estimation? The Case of Germany's Programmes for Long Term Unemployed, IZA Discussion Papers, No. 12526, Institute of Labor Economics (IZA), Bonn

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. Journal of Computational and Graphical Statistics, 2(4), 405-420.

Guo, X. S., & Fraser, M. W. (2015). Propensity score analysis: Statistical methods and applications (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Gündüz, S., Kıral, G. (2020), "Markov Zincirleri Kullanılarak Seçmen Tercihlerinin Tahmin Edilmesi", Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, cilt 29, sayı 1, s. 270-281.

Janitza S, Hornung R (2018) On the overestimation of random forest's out-of-bag error. PLoS ONE 13(8): e0201904. https://doi.org/10.1371/journal.pone.0201904

Keller, B., Kim, J., & Steiner, P. (2013). Abstract: Data mining alternatives to logistic regression for propensity score estimation: Neural networks and support vector machines. Multivariate Behavioral Research, 48(1), 164-164. doi:10.1080/00273171.2013.752263

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. Statistics in medicine, 29(3), 337–346. https://doi.org/10.1002/sim.3782

Leite, W. (2016). Practical Propensity Score Methods Using R. Sage Publications.

Liu, Y., Ye, C., Sun, J., Jiang, Y., & Wang, H. (2021). Modeling undecided voters to forecast elections: From bandwagon behavior and the spiral of silence perspective. International Journal of Forecasting.

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: An introduction and experimental test. Evaluation Review, 29(6), 530-558.

McCaffrey, Daniel F., Greg Ridgeway, and Andrew R. Morral (2004). "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies". In: Psychological Methods 9.4, pp. 403–425.

Olmos, A., & Govindasamy, P. (2015). Propensity Scores: A Practical Introduction Using R. Journal of MultiDisciplinary Evaluation, 11(25), 68–88. Retrieved from https://journals.sfu.ca/jmde/index.php/jmde_1/article/view/431

Pufahl A, Weiss C (2009) Evaluating the effects of farm programmes: results from propensity score matching. Eur Rev Agric Econ 36(1):79–101

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. Journal of the American statistical Association, 79(387), 516-524.

Rosenbaum, P.R. and Rubin, D. (1985) Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. The American Statistician, 39, 33-38.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiology and drug safety, 17(6), 546–555. https://doi.org/10.1002/pds.1555

Sitembölükbaşı, Ş. (2004). Isparta'da seçmenlerin parti tercih nedenleri üzerine bir araştırma: 1995, 1999 ve 2002 genel seçimleri karşılaştırması. Akdeniz Üniversitesi İİ BF Dergisi, 8, 156-176.

Staff, J., Patrick, M. E., Loken, E., & Maggs, J. L. (2008). Teenage alcohol use and educational attainment. Journal of Studies on Alcohol and Drugs, 69(6), 848–858.

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics (Oxford, England), 28(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Titus, M. A. (2007). Detecting selection bias, using propensity score matching, and estimating treatment effects: An application to the private returns to a master's degree. Research in Higher Education, 48(4), 487-521.

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. Journal of Clinical Epidemiology, 63(8), 826-833. https://doi.org/10.1016/j.jclinepi.2009.11.020

Wyse, A. E., Keesler, V., & Schneider, B. (2008). Assessing the effects of small school size on mathematics achievement: A propensity score-matching approach. Teachers College Record, 110(9), 1879-1900.

Yılmaz Koğar, E. (2019). Eğilim puanı eşleştirme analizinin eğitim araştırmalarında kullanılması. Amasya Üniversitesi Eğitim Fakültesi Dergisi, 8(1), 24-61.

Zhao, P., Su, X., Ge, T., & Fan, J. (2016). Propensity score and proximity matching using random forest. Contemporary clinical trials, 47, 85–92.

APPENDICES

A. R Codes For Preparing Data For Analysis

```
#Deal with missing values

deneme %>% is.na() %>% colSums()

imp <- missForest(deneme)

imputed_X <- imp$ximp

deneme <- imputed_X

#After deleting the undecided voters from the data, create a treat group of 20% in the remaining data

smp <- sample(nrow(deneme), 0.20*nrow(deneme), replace = FALSE)

treat <- deneme[smp,]

control <- deneme[-smp,]

#Change the answers of this 20% group into “undecided”

treat$alliance<-recode(treat$alliance, HDP = "Undecided", Nation = "Undecided", Other = "Undecided", Public = "Undecided",Protest = "Undecided")

#create "sample" column

treat$Sample<-as.factor('treat')

control$Sample<-as.factor('control')

#combine treat and control groups

mydata <- rbind(treat, control)

#create binary variable

mydata$Group <- as.logical(mydata$Sample == 'treat')
```

B. R Codes for Propensity Score Analysis

```
###Use Logistic Regression###

#Nearest neighbor matching is used as follows and the ratio and replace arguments can be changed according to the user’s request.
```

```
m.out <- matchit(Group ~  . , data = mydata, method="nearest", distance=”logit”, replace = …., ratio = …)

#Optimal matching

m.out <- matchit(Group ~  ., data = mydata, method="optimal", distance=”logit”, ratio=…, replace = ….)

#Use the mahalanobis distance

ps <- glm(Group ~ . , data = mydata, family = binomial())

psvalue <- predict(ps, type = "response")

mydata2 <- cbind(mydata,psvalue)

m.out <- matchit(Group ~ . + psvalue, data = mydata2, distance = "mahalanobis" , replace = …, ratio = …)

###Use Neural Network###

#For a neural network model, the distance argument should be 'nnet' and for different matching methods, the “method” argument can be changed to “nearest” or “optimal”. The number of hidden-layer units is set to 16.

m.out <- matchit(Group ~ . , data = mydata, distance = 'nnet', method = ‘…’, distance.options = list(size = 16), replace = … , ratio = …)

###Use Random Forests###

#First, create a random forest model

model.rf <- randomForest(alliance ~ ., data = mydata)

#Then, estimate the propensity scores by using random forest model

eps <- model.rf$votes[,2]

#Observe matching algorithms

m.out <- matchit(Group ~ D7 + D8 + NUTS1 + S1 + S7 + S8 + S9 , data = mydata, distance = eps, discard = 'both', method = …, replace = … , ratio = …,)

### Observe the final matched data and matched pairs

final_data <- match.data(m.out)

head(m.out$match.matrix)

# Assign the undecided voter in the treat group to the voter they matched in the control group and apply this rule all data

for(i in 1:nrow(treat)){
```

```
final_data[row.names.data.frame(final_data)==row.names(m.out$match.matrix)[i],
]$alliance=final_data[row.names.data.frame(final_data)==m.out$match.matrix[i,1]
,]$alliance

}

#Observe the accurate matching rate

newdata <- final_data[1:nrow(treat),]

count<-0

for(i in 1:2110){

if(newdata[row.names.data.frame(newdata)[i],]$alliance==treat1[row.names.data.fr
ame(treat1)[i],]$alliance){

count=count+1

  } }

accurate_matching_rate <- count*100/nrow(treat)
```