KnE Social Sciences

EBEEC 2019
Economies of the Balkan and Eastern European Countries
Volume 2019

Knowledge E
enriching | engaging | empowering

Conference Paper

Exploring Decision Rules for Election Results by Classification Trees

İpek Deveci Kocakoç and İstem Keser

Dokuz Eylul University, Econometrics Dept.

Abstract

This study explores the most important socio-economic variables determining the voting decisions of the provinces in Municipality Elections by using classification trees. We collected data on many potential variables that may affect voting decisions in favor of a political party. Each province's economic, geographic and demographic data is taken into consideration as independent variables. The dependent variable is the winner party in 2014 Municipality Elections. Data set consists of 81 provinces' data on 69 variables. The aim of the study is to find which variables affect voting decision the most and try to find a pattern that may lead political campaigns. Amongst many classification algorithms, we used C5.0 algorithm coded in R. It helps us explore the structure of a set of data, while developing easy to visualize decision rules for predicting a categorical (classification tree) or continuous (regression tree) outcome. The C5.0 algorithm determines the separation criterion with the greatest information gain in each decision node and performs optimal separation.

Since our data size is small, we used k=1000 trials (estimations) and then summarized them to provide more robust results. By choosing C5.0 algorithm's sub-trial size as 5, 5000 trees are formed and the mean of all importance scores of all trees formed are calculated and interpreted. The most important independent variables discriminating the voting decision are found to be the result of the previous elections, mean household population, proportion of population between ages 15 and 19, electricity consumption per person, and proportion of population between ages 55 and 64.

Keywords: classification trees, voting decision, C5.0 algorithm, decision trees

jel CLASSIFICATION codes: C02, C44, D72

Corresponding Author:
İpek Deveci Kocakoç
ipek.deveci@deu.edu.tr

Received: 17 November 2019
Accepted: 6 January 2019
Published: 12 January 2020

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the EBEEC Conference Committee.

1. Introduction

The scientific study of voting behavior is marked by three major research schools: the sociological model (often identified as School of Columbia) focuses on the influences of social factors; The psychosocial model (identified as School of Michigan) assumes that party identification is the main factor behind the behavior of voters; and rational choice theory (referred to as a model of economic voting, or even as School of Rochester) puts emphasis on variables such as rationality, choice, uncertainty and information [1, 2].

It can be said that people in Turkey vote based on various reasons. The ideology of the voters is primarily effective on the voting behavior. Also stakeholder and media support for the party, the desire to maintain stability, the electorate's feeling of attachment to the party leader (particularly the leader's ``countrymen''), the lure of promises by the party leader (together with the feasibility and likelihood of the delivery of the promised outcomes); track record of the fulfilment of popular demands through services provided can be listed among the factors that could have a bearing on electoral choices.

Moreover, the rallies and election campaigns of the parties during the election period are aimed at influencing the voting preferences and receiving the support of the voter. Campaigns for election are a means of introducing candidates to the voters in the local elections and allowing them to recognize the candidates they will vote for. This campaign period is very important for the voters who are defined as undecided. The undecided electorate determines the voting behavior in the ballot box based on the behaviors that the parties exhibit or cannot exhibit in this period.

There is a wide range of literature on the factors that influence voter preferences. However, the factors that influence Turkish voters' choices can be summarized as the candidate, ideology, leader, agenda, party commitment, socio-economic conditions, social environment, probability of winning the party, media coverage, and political advertising [3-5].

The aim of this study is to determine the factors that affect the voter preference on the provincial basis, rather than determining the factors that affect the individual preferences of the voters. In the local elections, the economic and social development factors of the provinces are expected to be highly effective on voting behavior. For this purpose, social and economic variables are taken into consideration.

The social and economic factors of the 81 provinces have been ranked and the political parties have been compared with the distribution of votes in the 2014 local elections in these provinces. Similar studies have been conducted in the United States to determine the variables affecting the distribution of votes [6].

In this study which covers 81 provinces, data from the statistical bulletins of 2014 were obtained and a decision tree was applied to variables selected from social and economic indicators. 2014 local election results were examined to find out which variables were more effective. The variables used in this study are obtained from government web databases.

The differences in development resulting from the failure to achieve balanced development are also observed in Turkey. Variability of economic and social development

levels can be seen between different regions of the country. Therefore, development levels are reflected in our geographical regions and provinces.

Turkey sets its agenda on issues such as economy, unemployment, education, health, income distribution, poverty, terrorism, lack of investment and services, internal-external debt, lack of infrastructure, illegal construction, environmental problems and so on [7]. These variables cause differences in the level of development between provinces and regions.

Turkey is among the developing countries, but it cannot be said that each province shows the same level of development. Socio-economic differences are observed between provinces. Socio-economic variables and socio-economic differences are expected to affect local election results. For this purpose, the impact of 69 of the socio-economic variables on the distribution of votes was examined.

Table 1: Some statistics for 30 March 2014 local elections.

#Polling Stations	# Registered Voter	#Votes	#Valid Votes
172,958	48,905,743	43,609,960	41,766,549

Although there were 48,905,743 registered voters in the local elections held on 30 March 2014 in Turkey, 43,609,960 voted and 41,766,549 were valid. Vote distribution is given in Figure 1. Winning party for each city can be seen in Figure 2. It is seen that the AKP was elected as the first party in Turkey. CHP, MHP and BDP follow the AKP.

In this study, the result variable, which shows which party has received majority of votes in the city, was investigated by classification trees. It has been calculated which of the variables in 81 cities are the most important in the election result variable.

2. Classification Trees

Using the models created on a certain dataset with the help of class labels, the process of estimating which classes the new samples will belong to is a classification problem. The aim is to assign objects (test set) that are not in the learning set to the correct classes as best as possible. In order to solve these problems encountered in many fields, new methods are being studied in different disciplines. Decision trees, artificial neural networks, Bayesian classifiers, Bayesian networks are some data mining methods for classification [8:169-186, 9]. The most popular tools in artificial intelligence applications are "decision trees" and "artificial neural networks". The application and interpretation of decision trees is much easier than artificial neural networks [10].

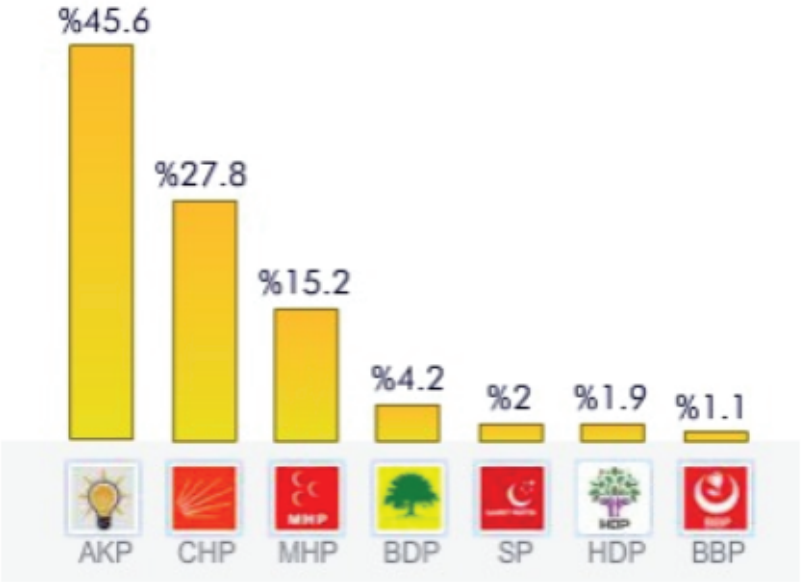

Figure 1: Votes of the parties throughout Turkey in 30 March 2014 local elections.

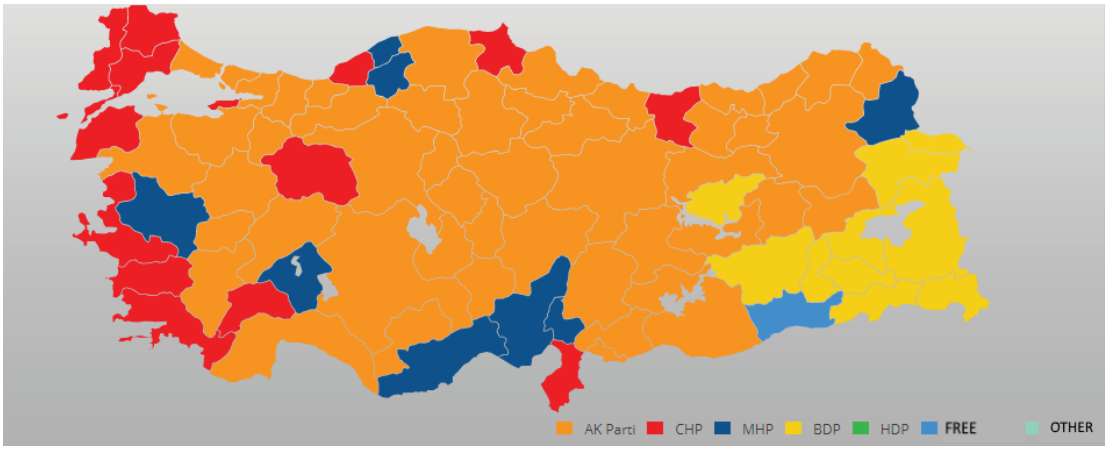

Figure 2: Winning party for each city.

A decision tree can be used to classify a case by starting at the root of the tree and moving through it until a leaf is encountered. At each nonleaf decision node, the case's outcome for the test at the node is determined and attention shifts to the root of the subtree corresponding this outcome. When this process finally leads to a leaf, the class of the case is predicted to be that recorded at the leaf [11]. A sample decision tree can be found in Figure3.

In decision tree learning, the cluster on which the training is performed is divided into sub-sets according to various characteristics, this process is repeated recursively and continues until the repetition has no effect on the estimation. This process is called recursive partitioning. There are many different algorithms for decision tree

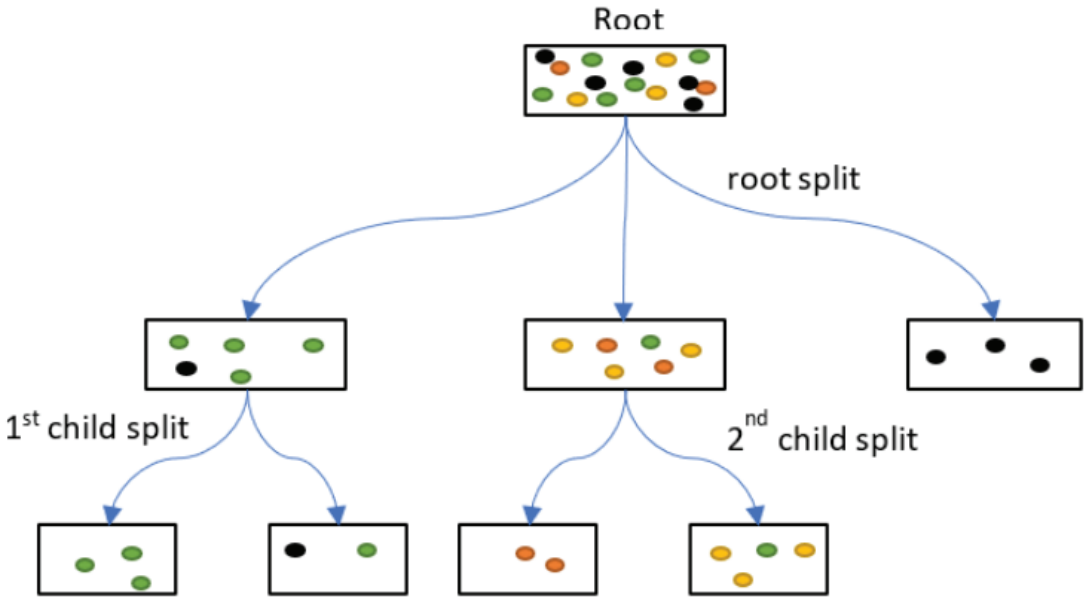

Figure 3: A sample decision tree.

classification such as Random Forest, Boosted Trees, Rotation Forest, ID3, C4.5, C5.0, Classification and Regression Tree (CART), QUEST, CRUISE, Chi-Square Automatic Interaction Detector (CHAID), and MARS [12, 13].

Wu et al. [14] examines top 10 data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM-2006). Zhang et al. [15] make an up-to-date comparison of state-of-the-art classification algorithms and give a detailed summary of related work on classifier comparison in chronological order. Ali and Smith [16] examine 8 algorithms/classifiers with 100 different classification problems and evaluate the algorithms' performance in terms of a variety of accuracy and complexity measures.

As can be seen from the studies mentioned above, there are numerous implementations of decision trees, but one of the most wellknown is the C5.0 algorithm. This algorithm was developed by computer scientist J. Ross Quinlan as an improved version of his prior algorithm, C4.5, which itself is an improvement over his ID3 (Iterative Dichotomiser 3) algorithm [17:124].

The C5.0 algorithm has become the industry standard for producing decision trees, because it does well for most types of problems directly out of the box. Compared to more advanced and sophisticated machine learning models (e.g. Neural Networks and Support Vector Machines), the decision trees under the C5.0 algorithm generally perform nearly as well but are much easier to understand and deploy [17:124].

The C5.0 algorithm determines the separation criterion with the greatest information gain in each decision node and performs optimal separation. Since it can also accommodate multi-class problems, this algorithm is used in this study.

3. Analyses and Results

In order to find out which variables were more effective on the distribution of votes, the impact of socio-economic and other variables is analyzed by C5.0 algorithm coded in R. Data set consists of 81 cities' data on 69 variables (one election result variable, 67 socio-economic variables plus one variable defining the result of the previous (2009) local elections). One of the limitations of the study is the small size of the data set. Most of the decision tree algorithms are designed to be used for big data. We used k=1000 trials (estimations) and then summarized them to provide more robust results.

The algorithm forms many trees as we desire in the parameters and calculates importance scores for all variables. The variable importance reflects the contribution each variable makes in classifying or predicting the target variable. The importance score measures a variable's ability to perform in a specific tree of a specific size. Since there are many trees formed, we run the codes for 1000 times and calculated the mean of all importance scores of all trees formed. For each trial, we chose C5.0 algorithm's sub-trial size as 5. This makes 5x1000=5000 trees. Since it is impossible to report all, here we will only give some samples of these trial results and report and interpret their summaries.

For each trial, 70 percent of cases (56 provinces) is selected randomly as training data and the rest (25 provinces) is selected as validation data. The best model is selected as the one with the highest validation accuracy.

The best model (tree) among 5000 trees has a classification rate of 0.98 for training data and a classification rate of 0.96 for validation data. The output for this model is in Figure 4. For the best model, 3 out of 5 sub-trials has the same minimum error rate. Only one case is misclassified for training set. Decision trees for these three sub-trials can be seen in Figure 5. Since the algorithm chooses a different variable to begin the split in every sub-trial, trees are different from each other. By using this model, we can predict what happens if some changes in socio-economic variables occur.

Top 10 variables according to split and usage importances are given in Table 2. The most important variables discriminating the votes in both type importances are the result of the previous elections, mean household population, proportion of population between ages 15 and 19, electricity consumption per person and proportion of

```
Evaluation on training data (56 cases):

Trial        Decision Tree
-----      ----------------
           Size      Errors

   0         6    8(14.3%)
   1         4   13(23.2%)
   2         7    8(14.3%)
   3         7    8(14.3%)
   4         7   10(17.9%)
boost              1( 1.8%)   <<
```

Error rates of sub-trials

```
  (a)   (b)   (c)   (d)    <-classified as
 ----  ----  ----  ----
   33                      (a): class AKP
         9                 (b): class CHP
    1          6           (c): class MHP
                     7     (d): class HDP
```

Classification matrix

Figure 4: Output for the best model.

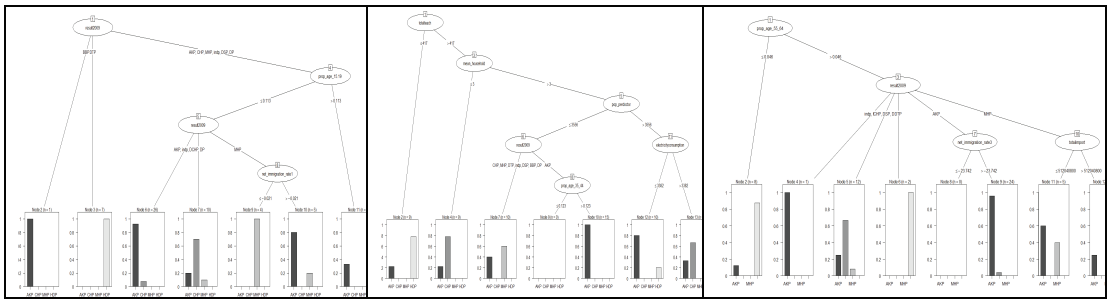

Figure 5: Best classification trees for the best model.

population between ages 55 and 64. Higher electricity consumption per person shows that the province is big and industrialized. Eastern provinces, where mean household population is higher, have a specific voting pattern.

TABLE 2: Variable importance scores.

Variable	Split_ importance	Variable	Usage_ importance
result2009	20.15	result2009	97.35
mean_ household	9.44	mean_ household	76.22
prop_ age_ 15.19	7.43	prop_ age_ 15.19	65.67
electricity consumption	6.91	electricity consumption	64.56
prop_ age_ 55_ 64	4.16	prop_ age_ 55_ 64	42.72
region	2.79	prop_ age_ 45_ 54	33.65
prop_ age_ 45_ 54	2.68	elementary_ and_ other	28.16
elementary_ and_ other	2.53	region	25.90
net_ immigration_ rate3	2.41	totalteach	24.48
young_ agedependency	2.09	young_ agedependency	22.83

4. Conclusion

In the 2014 elections, it was observed that the mean household population and age factor had a significant effect on the distribution of votes. This may give political parties

a chance to change their campaign, promises and even candidates to accommodate this information.

It was seen that in 2014 local elections, individuals voted more according to their political affiliation rather than socio-economic factors. Since complete data for 2019 local elections were not published at the time of the analysis yet, this study analyzes 2014 data. Especially the effects of economic regression and crisis in Turkey in 2018 may change the results of the study. A comparative study with new data will be enlightening. If data on smaller local areas (sub-provinces) such as towns or villages could be gathered, the investigation would also give insights about the effect of the size of the local unit. In further researches, previous election results may be excluded from the analysis to identify the sole effect of socio-economic variables.

In determining the level of development of the provinces, usually, a single criterion is evaluated. But it would be wrong to say that evaluations based on a single criterion can lead to a general conclusion. Both social and economic criteria should be considered as complementary to obtain meaningful and consistent results.

References

[1] Antunes, R. (2010). Theoretical models of voting behavior, *Escola Superior de Educação - Instituto Politécnico de Coimbra*, http://www.exedrajournal.com/docs/N4/10C_RuiAntunes_pp_145-170.pdf.

[2] Harrop, M., Miller, W.L. (1987). Psychological, Economic and Sociological Models of Voting. In: *Elections and Voters*. London: Palgrave.

[3] Çakır, H., and Biçer, A. (2014). Türkiye'de Yerel Seçimlerinde Seçmen Tercihlerini Etkileyen Kriterler: 30 Mart Yerel Seçimleri Kayseri Örneği (in Turkish, Voter Preference Criteria Influencing the Local Elections in Turkey: March 30 local elections in Kayseri Example), *Erciyes İletişim Dergisi*, vol.4, no.1, pp. 98-112.

[4] Teyyare, E., and Avcı, M. (2016). Yerel Seçimlerde Seçmen Davranışları: 2014 Yerel Seçimleri ve Zonguldak İli Örneği (in Turkish; Voter Behavior in Local Elections: 2014 Local Elections and Zonguldak Province Example). *Siyaset, Ekonomi ve Yönetim Araştırmaları Dergisi*, vol. 4, no. 1, pp. 51-76.

[5] Miş, N. (2018). Seçmen Tercihini Belirleyecek Faktörler (in Turkish, Factors Determining Voter Preference), *Kriter*, vol.3, no.28, https://kriterdergi.com/siyaset/secmen-tercihini-belirleyecek-faktorler (Last access date: 11.09.2019).

[6] Chibanda, K., and Greeff, J.F. (2017). Advanced Predictive Modeling Workshop Tree-based Methods, 2017 *Ratemaking and Product Management Seminar & Workshops*,

https://www.casact.org/education/rpm/2017/presentations/Workshop-6_1.pdf (Last access date: 28.03.2019).

[7] Miş, N. (2018). Yerel Seçimin Belirleyici Unsurları (in Turkish; Determinants of Local Selection), https://www.setav.org/yerel-secimin-belirleyici-unsurlari/ (Last access date: 11.09.2019)

[8] Nisbet, R., Miner, G., and Yale, K. (Editors), (2018). *Handbook of Statistical Analysis and Data Mining Applications* (Second Edition), Academic Press, https://doi.org/10.1016/B978-0-12-416632-5.00009-8.

[9] Narula, G. (2019). Machine Learning Algorithms for Business Applications -- Complete Guide. https://emerj.com/ai-sector-overviews/machine-learning-algorithms-for-business-applications-complete-guide/ (Last access date: 11.09.2019).

[10] Varghese, D. (2018). Comparative Study on Classic Machine learning Algorithms. https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222 (Last access date: 11.09.2019).

[11] Quinlan, R., (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

[12] Gupta, B., Rawat, A., Jain, A., Arora, and A., Dhami, N. (2017). Analysis of Various Decision Tree Algorithms for Classification in Data Mining. *International Journal of Computer Applications*, vol. 163, no. 8, pp. 15-19.

[13] Dai, Q.Y., Zhang, C.P., and Wu,H. (2016). Research of Decision Tree Classification Algorithm in Data Mining, *International Journal of Database Theory and Application,* vol.9, no.5, pp.1-8.

[14] Wu, X., Kumar, V., Quinlan, J.R., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.H., Steinbach, M., Hand, D.J., and Steinberg, D. (2008). Top 10 Algorithms in Data Mining, *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1-37, https://doi.org/10.1007/s10115-007-0114-2.

[15] Zhang, C., Liu, C., Zhang, X., and Almpanidis, G. (2017). An up-to-date comparison of state-of-the-art classification algorithms, *Expert Systems with Applications*, vol.82, pp. 128-150, https://doi.org/10.1016/j.eswa.2017.04.003.

[16] Ali, S., and Smith, K.A. (2006). On learning algorithm selection for classification, *Applied Soft Computing*, vol.6, no.2, pp. 119-138, https://doi.org/10.1016/j.asoc.2004.12.002.

[17] Lantz, B. (2013). *Machine Learning with R*. Packt Publishing.