



RESEARCH ARTICLE

Comparative Analysis of Machine Learning Algorithms Based on Variable Importance Evaluation

^{1,*} Hasan YILDIRIM

^{1,*} Karamanoğlu Mehmetbey University, Kamil Özdağ Science Faculty, Mathematics, Karaman, Turkey,
hasanyildirim@kmu.edu.tr, Orcid: 0000-0003-4582-9018.

HIGHLIGHTS

- Machine learning algorithms may provide useful insights on variable importance via effective tuning.
- The cubist algorithm can present predictive results on property value assessments.
- The cascade of machine learning and statistical methods are powerful tools for the property valuation.

Keywords:

- Cubist
- Random Forest
- Machine Learning
- MARS
- Variable Importance

GRAPHICAL ABSTRACT

One of the main goals in machine learning studies is to determine the most significant variables on a specific research problem. Various algorithms have been developed to achieve this goal. Random forest, Cubist, and MARS algorithms are the most common ones among these algorithms. Although classical statistical algorithms have been useful to obtain the importance level of the effective variables on the output in a certain amount, the machine learning algorithms may provide clearer and more precise results. In this study, the estimation results of Random Forest, Cubist, and MARS algorithms have been presented comparatively in terms of some performance criteria like mean squares error, the coefficient of determination, and mean absolute error by using a real data set. The results show that the performances of Random Forest and Cubist are similar amongst themselves but better than MARS. Additionally, the rank of the most important variables varies according to the type of algorithm. The concordance between algorithms is investigated from a statistical perspective and found satisfactory. Consequently, Random Forest, Cubist, and MARS can be considered effective and reasonable algorithms for both estimation performance and variable importance evaluation.

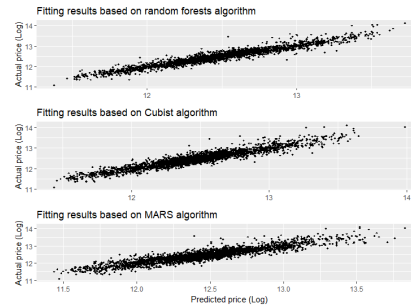


Figure A. Fitting results of whole data set based on each algorithm

Article Info:

Received : 2021-08-30
Accepted : 2021-10-06
Published : 2021-12-21

DOI:

10.53525/jster.988672

*Correspondence:

Hasan YILDIRIM
hasanyildirim@kmu.edu.tr
0 (338) 226 2000 - 3766.

Aim of Article: The main goals are (1) to compare the most effective machine learning algorithms by using a real and original estate data set and (2) to present the variable importance abilities.

Theory and Methodology: Based on a solid experimental setting and hyper-parameter tuning, including k-fold cross-validation, Cubist, Mars, and Random Forest algorithms have been compared to each other on real estate data set. Also, the Kendall W coefficient as a statistical significance test has been added to check the concordance of all algorithms.


Findings and Results: The main results show that all algorithms present reasonable results, but the Cubist is found as the best algorithm. The concordance between algorithms is significant, according to the Kendall W coefficient.

Conclusion: Machine learning algorithms may provide better insights in terms of variable importance evaluation than conventional methods like correlational or exploratory methods. With careful hyper-parameter tuning, each algorithm has both generalizability and the ability of variable importance determination.



RESEARCH ARTICLE

Comparative Analysis of Machine Learning Algorithms Based on Variable Importance Evaluation

1,*  Hasan YILDIRIM

1, * Karamanoğlu Mehmetbey University, Kamil Özdağ Science Faculty, Mathematics, Karaman, Turkey,
hasanyildirim@kmu.edu.tr, Orcid: 0000-0003-4582-9018.

Citation:

Yıldırım, H. (2021). *Comparative Analysis of Machine Learning Algorithms Based on Variable Importance Evaluation*, Journal of Scientific Technology and Engineering Research, 2(2): 46-53. DOI: 10.53525/jster.988672

HIGHLIGHTS

- Machine learning algorithms may provide reasonable insights on variable importance via effective tuning.
- The cubist algorithm can present predictive results on property value assessments.
- The cascade of machine learning and statistical methods are powerful tools for the property valuation.

Article Info

Received : 2021-08-30

Accepted : 2021-10-06

Published : 2021-12-21

DOI:

10.53525/jster.988672

*Corresponding Author:

Hasan YILDIRIM

hasanyildirim@kmu.edu.tr

0 (338) 226 2000 - 3766

ABSTRACT

One of the main goals in machine learning studies is to determine the most significant variables on a specific research problem. Various algorithms have been developed to achieve this goal. Random forest, Cubist, and MARS algorithms are the most common ones among these algorithms. Although classical statistical algorithms have been useful to obtain the importance level of the effective variables on the output in a certain amount, the machine learning algorithms may provide clearer and more precise results. In this study, the estimation results of Random Forest, Cubist, and MARS algorithms have been presented comparatively in terms of some performance criteria like mean squares error, the coefficient of determination, and mean absolute error by using a real data set. The results show that the performances of Random Forest and Cubist are similar amongst themselves but better than MARS. Additionally, the rank of the most important variables varies according to the type of algorithm. The concordance between algorithms is investigated from a statistical perspective and found satisfactory. Consequently, Random Forest, Cubist, and MARS can be considered effective and reasonable algorithms for both estimation performance and variable importance evaluation.

Keywords: Cubist, Random Forest, Machine Learning, MARS, Variable Importation

I. INTRODUCTION

The amount of data to be considered in research has been increased exponentially with rapid advances in technology. Besides the possibility of getting new insights by using big data, there have been some challenges including processing and storing data effectively. The necessity of reducing processing time increased the interest of researchers from different disciplines [1],[2] in this field. Although various analysis techniques have been proposed for especially high dimensional data set, it may be wise to reduce the size of the data set. There are two common ways for reducing the dimension of a data set: (1) to use some

dimensional reduction techniques (such as principal component analysis, locally linear embedding, t-SNE, etc.) and (2) to determine the most important attributes by using some techniques including intrinsic measurements (such as Random forest, partial least squares, etc.). In the first case, the idea is to obtain new attributes representing the actual ones to be explained the data set as much as possible. Thus, the dimension is reduced in a preprocessing step. The idea in the second case is to omit the least important attributes from the data set. During this process, the redundant or irrelevant attributes are determined and excluded. The decision of exclusion is researcher-dependent. The purpose of this process is to help the researcher to filter or determine the most important determinants in



measures and take a decision wisely. Herewith feature extraction provides some advantages such as better learning performance, lower computational cost, smaller memory necessity, and models with better generalization performance [3].

There have been many studies on feature selection, feature exclusion, and feature extraction. Most studies focus on a particular task, such as regression, classification, or clustering. Although the main idea is to reach the advantages mentioned in Li et al. [3], different approaches may be adopted. Hall and Smith [4] proposed a filter approach based on correlation and used it as a wrapper feature selection technique. Guyon and Elisseeff [5] carried out a review study on feature selection and feature extraction in a broad context, including feature construction, feature ranking, multivariate feature selection, efficient search algorithms, and feature validity assessment algorithms. The filter, wrapper, and embedded algorithms for feature selection were summarized in Saeys et al. [6] study. Alelyani et al. [7] focused on feature selection algorithms for clustering tasks only. In Chandrashekar and Şahin's [8] study, the comparative results using well-known benchmark data sets to conduct variable elimination processes were presented. Tang et al. [9] summarized the feature selection algorithms for classification tasks only. Li et al. [3] suggested a data perspective process on feature selection and gave a comprehensive overview of recent advances in this field. From the perspective of big data analytics, El-Hasnony [10] proposed some heuristic algorithms to improve the feature selection performance of machine learning algorithms. Karasu et al. [11] developed a wrapper-based method based on particle swarm optimization to achieve a more predictive regression model. More detailed explanations related to heuristic feature selection algorithms can be found in the study of Sharma and Kaur [12]. Feature selection methods can be classified into three parts as in most of the literature: (1) Filter algorithms, (2) Wrapper algorithms, and (3) Embedded algorithms. The embedded methods correspond to the algorithms having intrinsic measurements for determining the most important features.

In this study, we focused on some embedded algorithms, including, Random forest, multivariate adaptive regression splines, and the cubist, and compared them in terms of the performance on variable selection. The following sections are organized as follows: In Section 2, the details of algorithms have

been presented. The results of the performance comparison have been given in Section 3. The conclusions have been summarized in Section 4.

II. ALGORITHMS

In this section, the details of algorithms used in this study were given in brief.

A. *Random forest*

Random forest is a very well-known machine learning algorithm and proposed by Leo Breiman [13]. The idea behind the Random forest algorithm is to create decision trees as independently as possible and combined them to obtain a single and strong learner. Random forest is a very powerful and computationally efficient algorithm in consequence of considering only a fraction of feature space for each split in a decision tree. This property makes Random forest superior because of the reduction in the variance of the estimates. The main reason for the high variance in estimation results is to have some correlated results which commonly occur during the usage of all possible features for building trees and the possibility of being dominant for some important features in the whole process. As a single tree, i.e., base learner, CART, or conditional inference trees are used commonly. For more detailed theoretical explanations about this algorithm, James et al. [14] and Hastie et al. [15] are suggested to the reader.

B. *Multivariate Adaptive Regression Splines*

Multivariate adaptive regression splines (MARS) is proposed by Friedman [16]. MARS algorithm is a non-parametric regression and highly effective algorithm for especially high dimensional settings. Unlike classical regression algorithms, it does not need any assumption for the underlying distribution of data sets.

The main idea behind MARS is to model the data by using a set of surrogate features instead of the original measurements. These new features are the hinge functions of the original data. A hinge function whose cutting (or threshold) value is equal to c , is usually expressed as $h(x - c)$ and $h(c - x)$. Based on the value of the feature, different parameter estimation for each possible hinge condition is obtained and added to the main regression equation. In other words, it creates a piecewise linear model by determining cutting points for a feature and getting the parameter estimates between the feature and a dependent variable.



The cutting point is found in each feature and the new model is created depending on this point. The model error based on each new model is calculated and the model and corresponding cutting point with the lowest error is used for the rest of the estimation process. The MARS algorithm has some superiority such as interpretability, being fast, suitable for the high dimensional task, conducting automatic feature selection process, and robustness to outliers [17]. For more detailed theoretical explanations about this algorithm, James et al. [14] and Hastie et al. [15] are suggested to the reader.

C. Cubist

Cubist is a rule-based and complement algorithm of C5.0 that is used for classification. It is based on different approaches proposed in Quinlan [18-20]. Cubist has some properties unlike other opponents: (1) different types of pruning, smoothing and creating rules process, (2) an optional boosting procedure, and (3) adjustable estimation with the possibility of choosing nearby units for the training data set [17]. The process of building a tree is similar to other decision algorithms, but Cubist carried out a different pruning by considering a weighted linear combination of two trees including an actual tree and the parent of it. The weights of each tree are calculated using a criterion based on the covariance of the tree residuals and the variance of the difference between the residuals. The

model with lower error has a larger weight compared to the other one. After determining the weights of each model, the adjusted error rate is calculated by removing each rule from the rule-set. If the adjusted error rate is increased when a rule is deleted, that rule is omitted from the set. For more detailed theoretical explanations about of this algorithm, James et al. [14] and Hastie et al. [15] are suggested to the reader.

III. EXPERIMENT

A. Description of Data Set

The data set belongs to a common task in machine learning field [21] and has been retrieved from a popular real estate website [22] by using entries in January & February 2018. The size of the data set is 3102 and the number of variables is 11. Seyhan, Çukurova, Yüreğir, and Sarıçam which are the most developed and crowded districts have been considered in the study. The name variables in the data set are location, age, credit status, size, distance to the city center, type of heating system, floor location, dues, number of rooms, number of bathrooms, the number of floors. The output (i.e., dependent) variable is the price of a house. The descriptive statistics for qualitative and quantitative variables are given in Table I and Table II, respectively. The number of houses and their percentages corresponding to each category of variables are presented in Table I.

Table I. Descriptive Statistics (Qualitative Variables)

	Categories	N	Percent
Location	Çukurova	1371	44.2
	Seyhan	1261	40.7
	Sarıçam	395	12.7
	Yüreğir	75	2.4
Credit status	Yes	2951	95.1
	No	151	4.9
Age	0-5	2079	67.0
	6-10	263	8.5
	11-15	377	12.2
	16-20	306	9.9
	21-25	57	1.8
	26-30	14	0.5
	31-35	3	0.1
	36-40	3	0.1
Heating system	Combi boilers	1801	58.1
	Air conditioning	974	31.4
	Central heating	302	9.7
	Stove	19	0.6
Dues	Floor heating	6	0.2
	Yes	3021	97.4
	No	81	2.6
	Total	3102	100.0



Table II. Descriptive Statistics (Quantitative Variables)

	Min	Max	Mean	SD
Size	30	400	160.99	45.730
# Rooms	1	8	3.81	0.797
# Bathrooms	0	6	1.56	0.521
# Floors	1	20	10.65	3.474
Floor location	1	20	5.27	3.623
Distance (km)	1.7	27.2	10.855	3.6785
Rental price (TL)	215	3480	1120	370
Price (TL)	65000	135000	282412	125684.89
Price (Log)	11.0822	14.1157	12.5511	11.7416

B. Experimental Settings

The data set is randomly split into training and testing sets with ratios of %70 and %30, respectively. The tuning parameters are determined by using a 10-fold cross-validation approach based on only training data set are listed as follows:

Random forest: The number of selected predictors for each split(m).

-(m): {2, 3, ... ,23}

Cubist: The number of committees (C) and the number of neighbors(n).

-(C): {10, 20, ... ,100} and (n): {2, 3, ... ,9}

MARS: The level of degree(d) and the number of prune (p)

-(d): {1,2} and (p): {2, 3, ... ,20}

The results are obtained via R software packages including caret[23], earth [24], Cubist [25], randomForest [26]. The natural logarithm of the output variable is modeled in this study. The basic CART learner is used as a base learner in the Random forest algorithm. RMSE, R^2 and MAE are used as the performance criteria.

Table III.

Comparison of training and testing performance

Algorithm	Split	Size	RMSE	R^2	MAE
Random forest	Train	2171	0.1612	0.8347	0.1190
	Test	931	0.1554	0.8639	0.1158
Cubist	Train	2171	0.1611	0.8352	0.1211
	Test	931	0.1552	0.8626	0.1167
MARS	Train	2171	0.1708	0.8148	0.1295
	Test	931	0.1668	0.8409	0.1278

C. Performance Comparison

In this part, the performance of Random forest, Cubist, and MARS algorithms for training and testing data is presented and given in Table III. According to the results of Table III, the performance of Random forest and Cubist algorithms are quite similar for both training and testing data. Based on RMSE and R^2 performance criteria, it is shown that Cubist is slightly better than random forest. On the other side, the MARS algorithm produces poorer results than these algorithms for all criteria.

The fitting results using best tuning parameters for Random forest, Cubist, and MARS algorithms are given in Fig.1. The relative superiority of Random forest and Cubist against MARS can be seen visually. The approximation performance for the higher and lower house prices via the MARS algorithm is more unstable than the opponents. Also, the linearity corresponding to the closeness between the estimated and actual prices of houses for Random forest and Cubist is more explicit than the MARS algorithm. Random forest, Cubist and MARS algorithms have intrinsic measurements for determining the most important features (i.e., variables). After fitting each algorithm using the best tuning parameters which are found via cross validation approach, the importance level and rank of each feature have been obtained. In Table IV, the results are given separately for each algorithm.

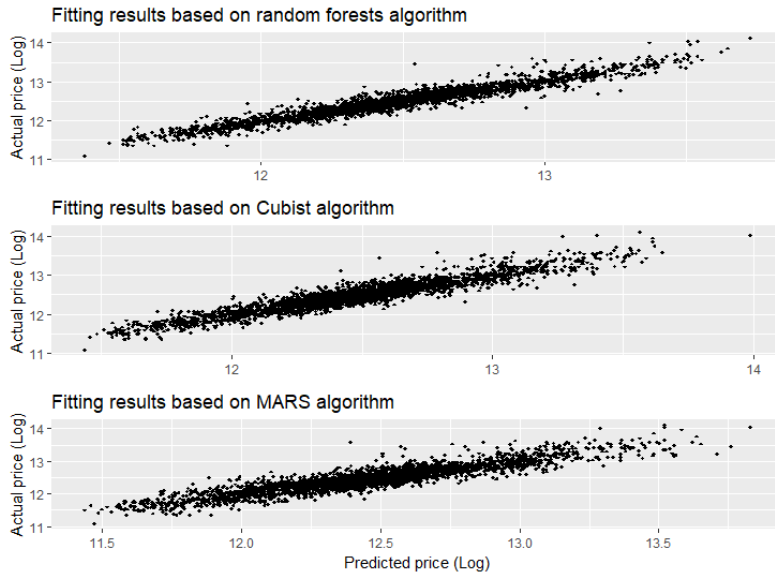


Figure 1. Fitting results of the whole data set based on each algorithm

According to Table IV, rental price, distance to the city center, the number of baths, the number of floors are commonly found the most important features for the price of a house by all algorithms. District (2) and district (4) have better rank values. These districts correspond to Seyhan and Yüreğir, respectively. Seyhan is the biggest district in Adana. On the other side, the economic level of Yüreğir is the lowest compared to the other districts. The reason for higher ranks may be the demand for Seyhan and the

purchasing power for Yüreğir. Heating (3) has the highest importance value among other categories. This corresponds the central heating. The top ten most important features for each algorithm are quite similar. It is shown that the MARS algorithm is more parsimonious on scoring importance compared to the Random forest and Cubist. Most dummy variables created for categorical features have no importance value according to the MARS algorithm.

Table IV. The results of variable importance evaluation based on each algorithm

Random Forest			Cubist			Mars		
Value	Variable	Rank	Value	Variable	Rank	Value	Variable	Rank
100	Distance	1	100	Rental price	1	100	Rental price	1
92,7	Baths	2	94,4	Distance	2	40,1	Baths	2
88,4	Number of Floors	3	83,9	Size	3	35,5	District(4)	3
87	Rental price	4	70,6	Baths	4	30,9	Rooms	4
76,4	District(2)	5	69,9	Number of floors	5	28,8	Floor	5
73,5	Size	6	58	Rooms	6	26,6	Distance	6
72,8	District(4)	7	57,3	District(2)	7	22,6	District(3)	7
63,8	Rooms	8	50,3	District(4)	8	18,5	Number of floors	8
52	Heating(3)	9	39,2	Age(4)	9	16,4	District(2)	9
46,9	Floor	10	37,1	Heating(3)	10	14,3	Heating(3)	10
44,8	District(3)	11	36,4	Floor	11	10	Size	11
32,6	Age(4)	12	32,9	Age(3)	12	0	Credit(2)	12
30,1	Age(3)	13	26,6	Age(2)	13	0	Age(2)	13
20,4	Credit(2)	14	25,2	District(3)	14	0	Age(3)	14
20,1	Age(2)	15	11,9	Age(5)	15	0	Age(4)	15
11,5	Heating(2)	16	7,7	Heating(2)	16	0	Age(5)	16
9,2	Heating(5)	17	2,8	Credit(2)	17	0	Age(6)	17
9	Dues(2)	18	2,1	Age(6)	18	0	Age(7)	18
8,8	Age(6)	19	0	Age(7)	19	0	Age(8)	19
6,8	Age(7)	20	0	Age(8)	20	0	Heating(2)	20
5,3	Heating(4)	21	0	Heating(4)	21	0	Heating(4)	21
2,7	Age(5)	22	0	Heating(5)	22	0	Heating(5)	22
0	Age(8)	23	0	Dues(2)	23	0	Dues(2)	23



Although Table IV provides very insightful and clear results on determining a house price, it may be useful to investigate the concordance between algorithms. The importance value and corresponding ranks are more similar between Random forest and Cubist, unlike MARS. The statistical examination of concordance is carried out by using Kendall's W coefficient of statistics. According to Kendall W statistics, Random forest, Cubist, and MARS algorithms statistically significantly matched in scoring ranks of important features ($W = .913, p < 0.0001$).

IV. CONCLUSION

In this study, we investigated the performance of Random forest, Cubist, and MARS algorithms in terms of accuracy and variable importance evaluation. The tuning parameters for each algorithm have been obtained via 10-fold cross-validation. Furthermore, the concordance between estimated ranks of features has been examined using Kendall's W coefficient. The results show that Random forest and Cubist present similar performance with each other but better than MARS. Based on overall performance, Cubist is found as the best among these algorithms. On the other hand, the top ten most important features are more common for all algorithms. The agreement of three algorithms on ranking features is found satisfactory.

Consequently, Random forest, Cubist and MARS algorithms produce not only good estimation performance, but also clear insights for variable importance evaluation. These kinds of algorithms can provide some useful insights from the inferential perspective and can be seen as functional tools from the pre-processing steps of all information-based applications in any area. The main advantage of this approach is to improve computational cost with a better understanding for particularly expert systems.

V. LIMITATIONS AND FUTURE STUDIES

One of the main limitations of this study is to be based on only non-heuristic feature selection algorithms. Because of that, a wide range of machine learning algorithms has feature selection properties, a comprehensive comparison study can be conducted. In future studies, we consider developing some novel algorithms that combined both wrapper-like and heuristic approaches via variable dimensions of data and tasks.

CONFLICTS OF INTEREST

No conflict of interest or common interest has been declared by the authors.

RESEARCH AND PUBLICATION ETHICS

In the studies carried out within the scope of this article, the rules of research and publication ethics were followed.

REFERENCES

- [1] Ertoý, U. & Akçay, M. (2021). Covid-19 Virüsü Salgını İle Mücadelede Büyük Veri Çalışmaları: Çin Örneği . *Journal of Scientific, Technology and Engineering Research* , 2 (2) , 4-14 . DOI: 10.5281/zenodo.4718425.
- [2] Pazar, Ş. , Bulut, M. & Uysal, C. (2020). Yapay Zeka Tabanlı Araç Algılama Sistemi Geliştirilmesi . *Journal of Scientific, Technology and Engineering Research* , 1 (1) , 31-37 . DOI: 10.5281/zenodo.3922425
- [3] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6), 1-45. DOI: 10.1145/3136625.
- [4] Hall, M. A., & Smith, L. A. (1999, May). Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. In *FLAIRS conference (Vol. 1999, pp. 235-239)*. DOI: 10.5555/646812.707499.
- [5] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar), 1157-1182. DOI: 10.5555/944919.944968.
- [6] Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19), 2507-2517. DOI: 10.1093/bioinformatics/btm344.
- [7] Alelyani, S., Tang, J., & Liu, H. (2018). Feature selection for clustering: A review. *Data Clustering*, 29-60. DOI: 10.1201/9781315373515-2.
- [8] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), 16-28. DOI: 10.1016/j.compeleceng.2013.11.024.
- [9] Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*, 37. DOI: 10.1201/b17320.
- [10] El-Hasnony, I. M., Barakat, S. I., Elhoseny, M., & Mostafa, R. R. (2020). Improved feature selection model for big data analytics. *IEEE Access*, 8, 66989-67004. DOI: 10.1109/ACCESS.2020.2986232.
- [11] Karasu, S., Altan, A., Bekiros, S., & Ahmad, W. (2020). A new forecasting model with wrapper-based feature selection approach using multi-objective optimization technique for chaotic crude oil time series. *Energy*, 212, 118750. DOI: 10.1016/j.energy.2020.118750
- [12] Sharma, M., & Kaur, P. (2021). A Comprehensive Analysis of Nature-Inspired Meta-Heuristic Techniques for Feature Selection Problem. *Archives of*



- Computational Methods in Engineering, 28(3). DOI: 10.1007/s11831-020-09412-6.
- [13] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32. DOI: 10.1023/A:1010933404324.
- [14] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *Statistical learning*. In *An introduction to statistical learning* (pp. 15-57). Springer, New York, NY. DOI: 10.1007/978-1-0716-1418-1_2.
- [15] Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer. DOI: 10.1007/978-0-387-84858-7.
- [16] Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, 1-67. DOI: 10.1214/aos/1176347963.
- [17] Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26, p. 13). New York: Springer. DOI: 10.1007/978-1-4614-6849-3.
- [18] Quinlan, J. R. (1987). Simplifying decision trees. *International journal of man-machine studies*, 27(3), 221-234. DOI: 10.1016/S0020-7373(87)80053-6.
- [19] Quinlan, J. R. (1992, November). Learning with continuous classes. In *5th Australian joint conference on artificial intelligence* (Vol. 92, pp. 343-348). DOI: 10.1142/9789814536271
- [20] Quinlan, J. R. (1993, June). Combining instance-based and model-based learning. In *Proceedings of the tenth international conference on machine learning* (pp. 236-243). DOI: 10.5555/3091529.3091560.
- [21] Yıldırım, H. (2019). Property value assessment using artificial neural networks, hedonic regression and nearest neighbors regression methods. *Selcuk University Journal of Engineering, Science and Technology*, 7(2), 387-404. DOI: 10.15317/Scitech.2019.207.
- [22] Zingat. An online real estate website. <https://www.zingat.com>. (Last Access: March, 2018).
- [23] Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of statistical software*, 28(1), 1-26. DOI: 10.18637/jss.v028.i05.
- [24] Milborrow, S. (2019). earth: Multivariate Adaptive Regression Splines. R package version 5.1.1.
- [25] Kuhn, M., Weston, S., Keefer, C., Coulter, N., & Quinlan, R. (2014). Cubist: Rule-and instance-based regression modeling, R package version 0.0. 18.
- [26] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3), 18-22. DOI: 10.1021/ci034160g.