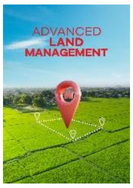
Advanced Land Management

https://publish.mersin.edu.tr/index.php/alm

e-ISSN 2822-7050



Comparison of modern methods using the python programming language in mass housing valuation

Gültekin Büyük*[1] , Fatma Bünyan Ünel[2]

[1]*Mersin University, Remote Sensing and Geographic Information Systems, Türkiye, gultekinbuyuk33@gmail.com*
[2]*Mersin University, Department of Geomatics Engineering, Türkiye, fatmabunel@mersin.edu.tr*

Cite this study: Büyük, G., & Ünel, F. B. (2021). Comparison of modern methods using the python programming language in mass housing valuation. Advanced Land Management, 1 (1), 23-31.

Keywords
Real Estate Valuation
Python Programming Language
Multiple Linear Regression
Decision Tree Regression

Research Article
Received: 15.10.2021
Revised: 14.11.2021
Accepted: 21.11.2021
Published: 15.12.2021

Abstract

Multiple methods are used in the mass housing valuation. With the developing technology, modern methods have gained speed. In this study, a systematic study was conducted with machine learning to estimate the final price of housing property. The studied dataset contains 73 samples and 18 arguments. In this study with the Python programming language, NumPy, Pandas, Scikit–learn, Matplotlib and Seaborn, which are the basic libraries of Python, were used. Multiple linear regression (MLR) and decision tree regression method were used to perform the study. The adjusted determination coefficient (r^2) was used to measure the performance of the estimation accuracy of applications. As a result of applications, the multiple linear regression model showed better results than the decision tree model.

1. Introduction

Immovable property; the name given to the real estate, such as land, buildings, apartments, business places [1]. Valuation; market conditions by analyzing the change of the business is process of finding it in value of the property's real estates in the face of economic developments [2]. The real estate business valuation; it can be defined as the estimation of the probable value of an immovable property, real estate rights and benefits attached to the project or based on a valuation day, independent, impartial and objective criteria [3].

Real estate valuation appreciation or evaluation only if a property is not limited to, at the same time, the price or value of the property using historical data about the variables that affect the price of various predicts. A specific prediction of action includes the removal of useful information from raw data. Then comparisons are made with current sales prices. Estimates often using traditional assessment methods is cumbersome, however, artificial neural networks, machine learning, algorithms such as computers and the emergence of explore the correlations between the variables that affect the price of real estate, and multi-dimensional variables facilitated the creation of patterns in [4-5].

Real estate valuation methods; traditional, statistical, spatial analysis, modern, hybrid, separate valuation method as it is possible. Traditional methods in institutions often are the preferred method. Other methods can be divided into modern methods. For maps of real estate valuation Esri software (ArcGIS, ArcMap, ArcPAD, ArcCatalog, Bottom 4 image Mapper), immovable bulk value for processing GLASS (computer assisted Mass Appraisal), MATLAB fuzzy logic, decision tree analysis for Precision Tree, and program SPSS for neuro solutions neural networks, etc. software is used.

1.1. Importance of Real Estate Valuation

The city planning and economic development, reconstruction plans real estate appropriate valuation methods are possible. New residential areas to the selection of the conditions for consolidating fragmented parcels and their organization in real estate valuation in urban areas is of great importance [6].

City Planning, Public Investments (investment evaluations (pre and post), the selection of a residential area, parcel, regulations, and making transparent and reliable for the valuation of real estate in the real estate market is very important. In order to achieve the best results in the evaluation of the scientific method to select the most suitable one for the purpose of valuation are required [7].

1.2. Literature Review

Many modern methods for valuation of immovable properties are used in literature studies and in particular those who made the decision tree method with the application examined.

Random forest regression method had been used with 2695 samplings in the Russian city of St. Petersburg. A prediction value is close 95% accuracy to the actual value of the immovable property [4].

Installation work of 125 PV power plants in Turkey for the purpose of multiple linear regression by using the method for different locations have estimated the PV power. 96% accuracy obtained in the study [8].

In Mamak district of Ankara, the independent variables consisting of 96 by the method of regression and decision tree CDR dataset is studied. The decision tree method yielded higher than predictive value [9].

Singapore's private residential property market had been investigated from 1995 until 2017 and made an application for a dataset with more than 300,000 real estate transaction. The tree-based method, they have concluded that the assumptions of multiple regression analysis showed better performance than the traditional [10].

It was carried out a study by using the Random Forest community algorithm in Pendik. prediction performance of random forest regression method with the engine thematic compatible datasets in Turkish National Geographic Information System (TUCBS) was evaluated. The result obtained 85% with an accuracy rate [11].

The subject of valuation of immovable properties that are outside the heating value of solid fuels with higher sample argument CDR and 21st 185 examined by regression method and decision tree was seen in %76% 75% gave better results indicated that the performance of the decision tree method by a small margin has been [12].

The aim of this study using machine learning thanks to artificial intelligence, Python programming language, multiple linear regression and decision tree regression analysis to compare. Criteria were made to estimate the market value of the middle-ranked houses in Mersin Yenişehir district by classifying them.

2. Material and Method

The material of this study consists of 73 samplings in Mersin, Yenişehir, Çiftlikköy District. The samplings prepared with sale price and features of housing properties constitute a dataset.

2.1. The study area

Yenişehir, is a district of Mersin province in the south and the Mediterranean, the North Freeway, in the East, the Mufti's Creek, the area within the boundaries of the area in the west of Mezitli municipality 3681 hectares. According to the 2019 census, the population of the province 266,117 person.

In north of the study area, Mersin University Çiftlikköy Campus established on an area of 400 hectares is located. Approximately 20 thousand students on campus and educated in medicine, dentistry such as to be open to the public health organizations has a positive impact on housing prices. The study area is in Figure 1.

2.2. Data

The value of immovable housing type is collected. Because it is very close to campus 1+1, 2+1, 3+1 and 4+1 illustrates the diversity of rooms. The criteria that affect the value of housing: area (gross-m^2), area (net-m^2), housing type, number of rooms, number of bathrooms, number of floors, floor, facades, parking, security, swimming pool, type of heating, balcony, the age of the building, road frontage-shaped housing characteristics have been taken as independent variables. As housing prices have benefited from the dependent variable.

In preparing the data input as a float, integer data, and the normalization of the observed data in Table 1 are reviewed.

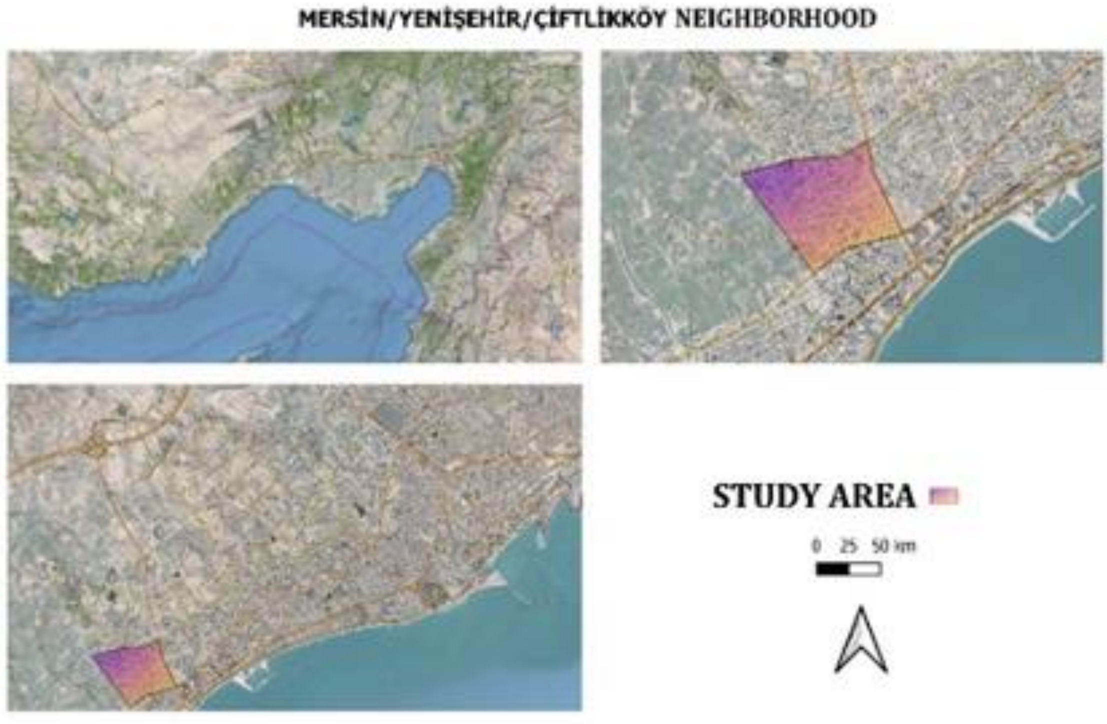


Figure 1. Study area

Table 1. The numeric equivalents of the data

Criteria	Description	Value
Type of housing	Apartment	1
	Site	0
Parking	There	1
	No	0
Security	There	1
	No	0
Pool	There	1
	No	0
Fronts	North	1
	South	2
	West	3
	East	4
Heating Type	Air conditioning	1
	Centre	2
	Private (Combi)	3

2.3. Method

2.3.1. Valuation Methods

Regression analysis between them cause-and-effect relationship in order to make predictions or extrapolations about the relationship between two or more variables called the regression model, statistical analysis is a technique that is characterized by a mathematical model [13].

2.3.1.1. Multiple linear regression (MDR)

The best relationship between a dependent variable and several independent variables of a dataset CDR can be used to predict under applied regression and machine learning is one of the models. The MDR model which minimizes the sum of squares of differences of observed and predicted values based on the method of Least Squares [8].

The MDR model, the dependent variable used in the calculation of the equality can be expressed as Equation 1.

$$Housing\ Price = x_0 + x_1y_1 + x_2y_2 + x_ny_n + \varepsilon \tag{1}$$

n: Number of samples,
$x_1, x_2, \ldots. x_n$: Weight values,
$y_1, y_2, \ldots y_n$: Independent variable, and
ε: Error term.

MDR analysis on the basis of the logic of the method of Least Squares operate. The connection between the one or more variable is used to detect [14].

2.3.1.2. *Decision Tree Regression*

A decision tree classifier or a regression analysis that is used to obtain tree-type data structure. Subsets is handled by dividing the dataset to be used every time. Consists of decision nodes and leaves. Decision nodes each time depending on the property, it is divided into two or more sub. The end nodes of the tree the leaf node represents the decision was taken [14].

This method is known as the space segment for a set of Predictor uses a decision tree splitting rule. The decision tree, the simplest interpretation is one of the methods which is easy to machine learning [10].

More specifically, the decision tree consisting of nodes and branches algorithmic structures. Each node is judged whether it is higher or lower than a value of a variable [15].

2.3.2. Performance

2.3.2.1. Adjusted determination coefficient ($\overline{R^2}$)

Coefficient of determination, in addition to measuring the success of the regression equation, the equation that reflects the power prediction of a statistic [16-17].

The function can be expressed with Equation 2.

$$\overline{R^2} = 1 - (1 - R^2)\,\frac{(n-1)}{(n-p-1)} \tag{2}$$

$$\mathbf{0 < R^2 < 1}$$

n: record number,

p: the number of independent variables

3. Results

MDR methods and random forest in Python the program was carried out. Python libraries that are used in the program; NumPy, Pandas, Scikit–learn Seaborn Matplotlib and consists of. Dataset are converted to the shape and ready for analysis more useful. The output from the python script is as shown in Table 2.

Table 2. Pandas MDR command output

	Gross Area	Certain Area	Type of Housing	Number of Rooms	Number of Bathrooms	Number of Floors	Current Floor
0	70	55	1	1+1	1	13	9
1	65	50	1	1+1	1	11	6
2	80	60	1	1+1	1	13	13
3	70	60	1	1+1	1	10	8
4	65	55	0	1+1	1	9	2

3.1. Multiple Linear Regression Model

The first libraries to be used in the application are included in the program. Pandas and missing data were achieved with the library entry for the control of dataset have been found.

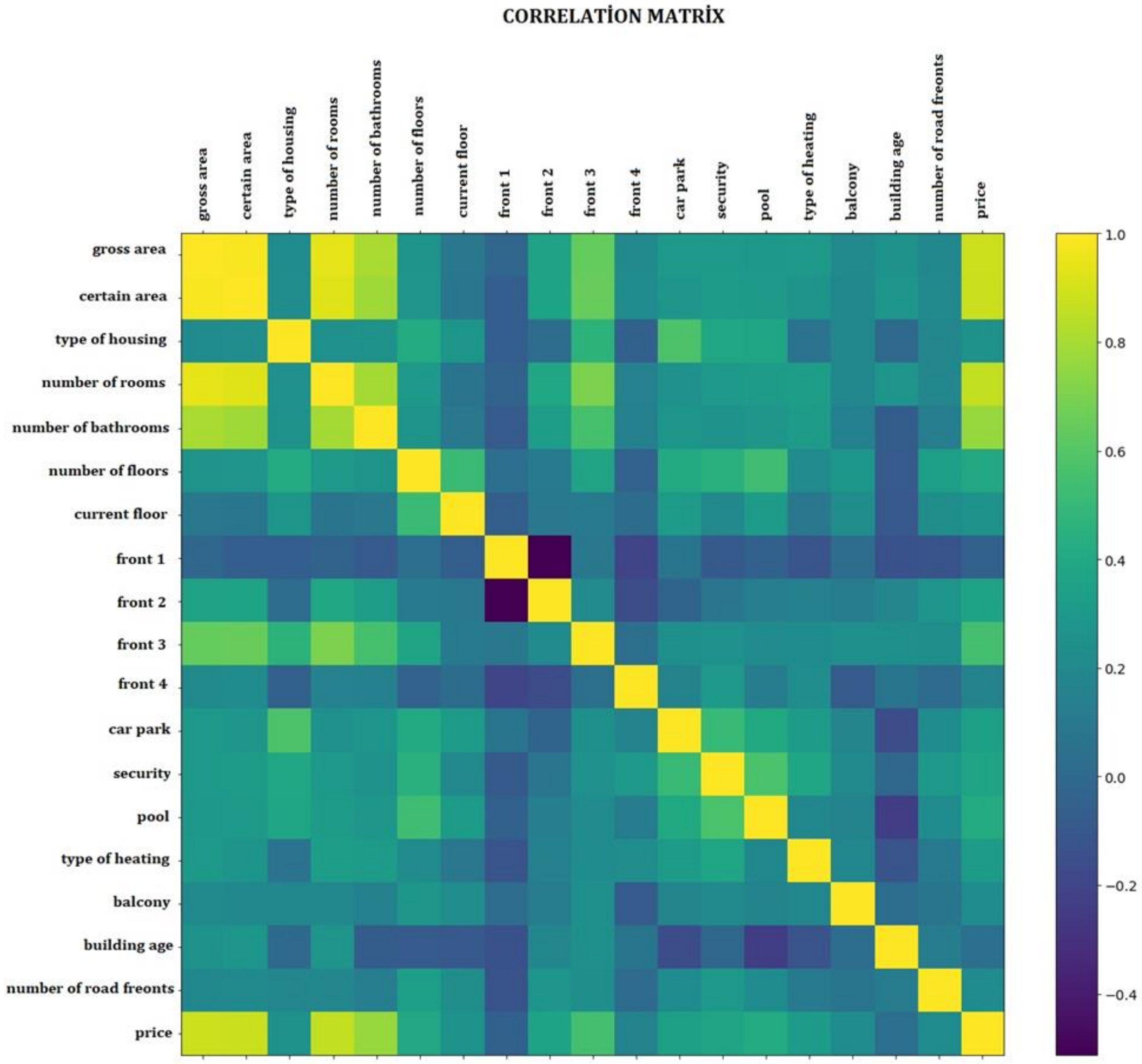
As the data in Table 1, the correlation relationship between the independent variables and the remaining made the normalization process Thanks to the library Seaborn heat map (heatmap) was investigated by creating.

As for the correlation, probability theory and statistics, random variables, direction and strength of the linear relationship between two or more shows. Multiple correlation of a variable with two or more variable; its relationship with other variables fixed partial correlation techniques with any of these variables is calculated. The correlation coefficient "r" and is indicated by -1 and +1 values between takes. $r = -1$ There was a linear relationship negative. $r = +1$ there is a positive linear relationship full. $r = 0$ there is a relationship between the two variables.

In this context, $r =$ the results are interpreted according to the following ranges for the relationship in the following way: 0.00 relationship-0.01 – 0.29 low level of the relationship, 0.30– 0.70 moderate relationship 0.71 – 0.99 a high level of iliski1.00 means perfect relationship.

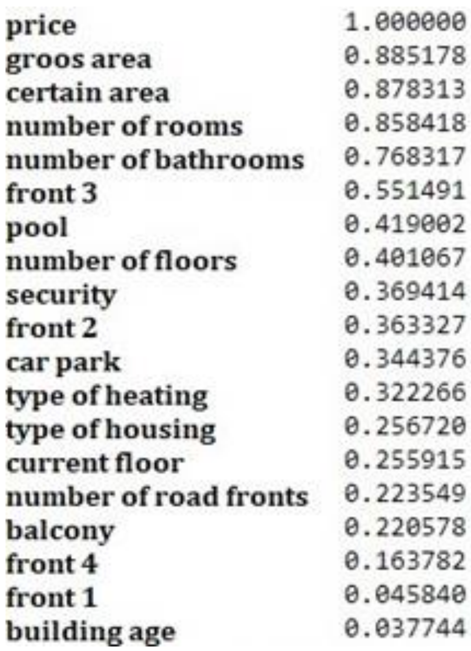
The heat map as shown in Table 3, when the price of housing other variables held constant, the relationship between the level of were visualized.

Table 3. Correlation Relationship



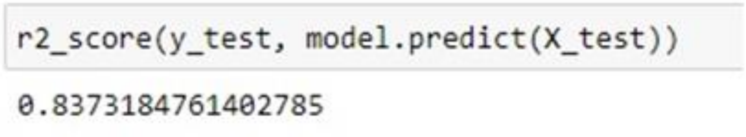
Depending on the heat map, housing prices in Figure 2 of the most important variables that affect the output from the python script as a sequence of gross area, Net area and number of rooms. Other criteria like are shown in Figure 3.

Establish the MDR model r^2 Python script to calculate the score are written in the output. "Linear Regression" command by running the program has been called and were obtained from the model, and r^2 0,84 is. What this means is that the MDR of the model to 84% with an accuracy rate of reflect the criteria (Figure 3).



```
price                      1.000000
groos area                 0.885178
certain area               0.878313
number of rooms            0.858418
number of bathrooms        0.768317
front 3                    0.551491
pool                       0.419002
number of floors           0.401067
security                   0.369414
front 2                    0.363327
car park                   0.344376
type of heating            0.322266
type of housing            0.256720
current floor              0.255915
number of road fronts      0.223549
balcony                    0.220578
front 4                    0.163782
front 1                    0.045840
building age               0.037744
```

Figure 2. Factors That Affect Price



```
r2_score(y_test, model.predict(X_test))

0.8373184761402785
```

Figure 3. Account Accuracy

3.2. Decision *Tree Regression*

As in the first application dataset been summoned and is put into operation. In the model, the target variable, ‘price’ for density-separated from the rest of the dataset looking at the distribution and price. The output from the python script is in the attachment (Figure 4).

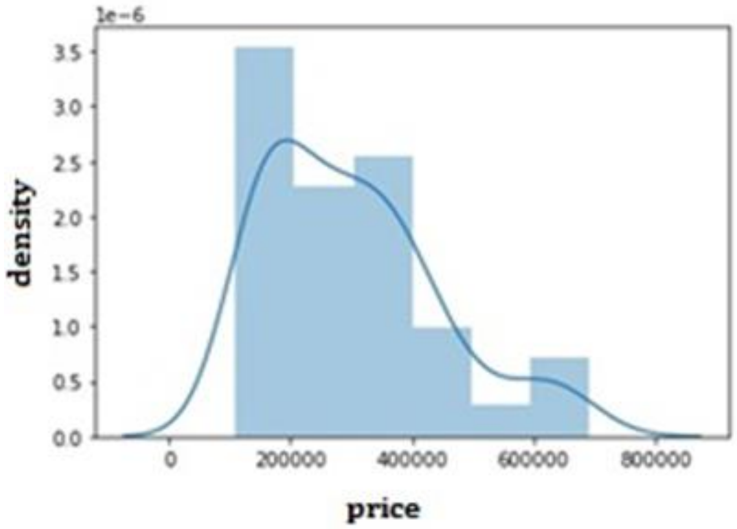
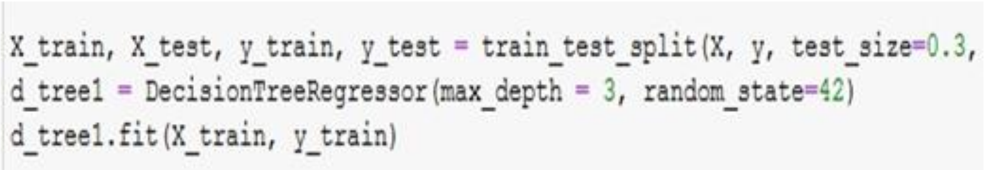


Figure 4. Density - Distribution Price

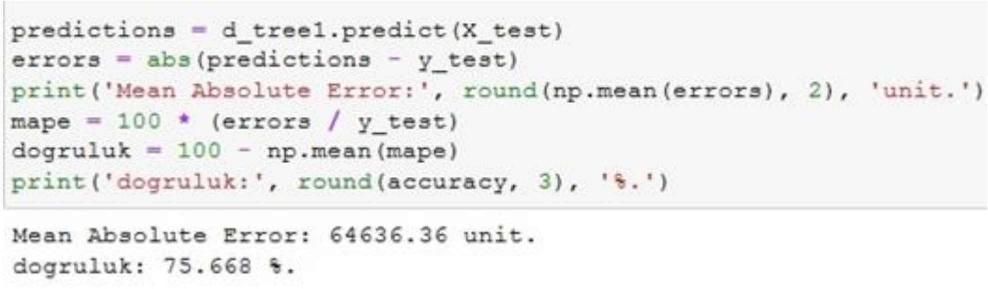
Ready to build the application model has become. Dataset the training and test sub-sets of data, allocation, create a decision tree model with the training data and the model ‘fit’ don't be performed. The output from the python script is in the attachment (Figure 5).

Based on the similarity between the actual value predicted with the accuracy of the calculation was made in the process (Figure 6).



```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
d_tree1 = DecisionTreeRegressor(max_depth = 3, random_state=42)
d_tree1.fit(X_train, y_train)
```

Figure 5. Model building



```
predictions = d_tree1.predict(X_test)
errors = abs(predictions - y_test)
print('Mean Absolute Error:', round(np.mean(errors), 2), 'unit.')
mape = 100 * (errors / y_test)
dogruluk = 100 - np.mean(mape)
print('dogruluk:', round(accuracy, 3), '%.')
```

```
Mean Absolute Error: 64636.36 unit.
dogruluk: 75.668 %.
```

Figure 6. Model building

76% regression decision tree with an accuracy on the order gave no results. While establishing the ranking of the importance attribute to a classification model, made visualization was performed (Figure 7).

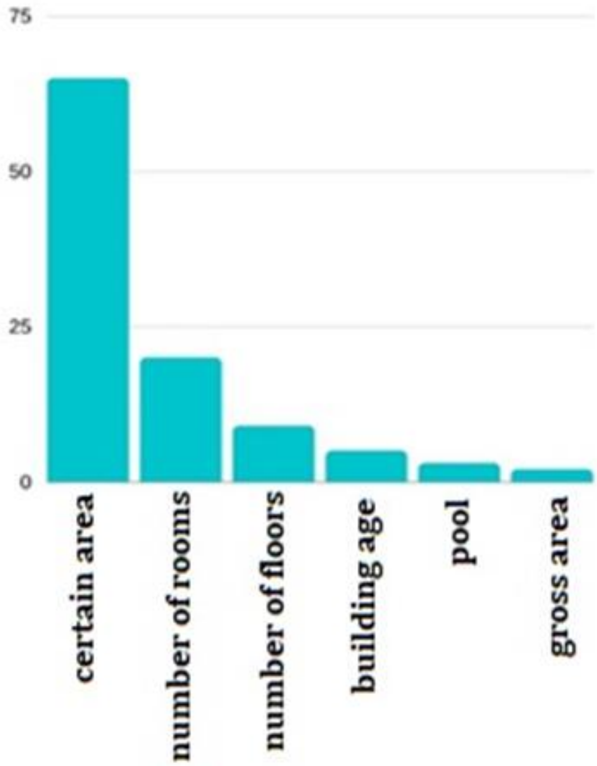


Figure 7. Attribute Importance Ranking

Visualization can be understood as the three most important variables that affect the price of the net this time house area, number of rooms and the number of layers as it seems.

4. Conclusion

As a result, the two methods applied to very accurate results were obtained. The variables that affect the estimate of the two applications was different. Multiple linear regression model accuracy is 84%, and the accuracy of Decision Tree regression model to be 76% was calculated.

Mass Real Estate Appraisal, which is the largest share of measuring the economic value of real property in the country has become one of the indicators of development. Gross domestic profitability of the real estate market in important economic dimensions. Various public and private institutions in the areas of unbiased, objective and scientific approach should be performed.

On the other hand, created a revolution in the industry business technology scientific machine learning. Many real estate site, Real Estate began using machine learning to predict the values of each of the technology. The need for a fast and economic evaluation of real estate and the greater availability of current information that can be accessed over the internet, big data and machine learning techniques to perform led to the implementation of real estate valuation.

They will buy or sell their property without the help of people's assessors. However, the economic crisis, pandemic, natural disaster or extraordinary situations such as the possibility of the presence of continuous development and immovable properties and their features because of the changes in the environment and realtors Valuation Professionals will always be needed.

Funding

This research received no external funding.

Author contributions

Gültekin Büyük: Investigation, Methodology, Software, Application, Validation, **Fatma Bünyan Ünel:** Conceptualization, Reviewed and Edited.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Demirel, B., Yelek, A., Alağaş, H. M., & Eren, T. (2018). Determination of Criteria for Immovable Valuation and Calculation Weights of Criteria with Multicriteria Decision Making Method. Kırıkkale University Journal of Social Sciences (KUJSS), 8(2), 665–682.
2. Yomralıoğlu, T., Nişancı, R., Çete, M., Candaş, E. (2012). Dünya’da ve Türkiye’de Taşınmaz Değerlemesi, Türkiye’de Taşınmaz Değerlemesi: II. Arazi Yönetimi Çalıştay 21-22 May 2012, İTÜ, İstanbul.
3. Yalpır, Ş., (2007). Bulanık mantık metodolojisi ile taşınmaz değerleme modelinin geliştirilmesi ve uygulaması: Konya örneği, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, Konya.
4. Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. Expert Systems with Applications, 39(2), 1772–1778.
5. Peter, N. J., Okagbue, H. I., Obas, E. C. M., & Akinola, A. O. (2020). Review on the application of artificial neural networks in real estate valuation. International Journal of Advanced Trends in Computer Science and Engineering, 9(3), 2918–2925.
6. Taktak, F., & Temiz, M. S. (2021). Pioneering Institutions in Sector on Real Estate Appraisal. Turkish Journal of Engineering, 5(3), 123–133.
7. Khamrabaeva, L. (2020). Real Estate Valuation Methods: An Application of Hedonic Price Model. Master Thesis, Bursa Uludag University, Social Sciences Institute, 167 p.
8. Uğuz, S., Çağlayan, N., & Oral, O. (2019). Estimation of Energy to be Obtained from PV Power Plants Using Machine Learning Methods. International Journal of Engineering Research and Development, 11(3), 769–779.
9. Yilmazer, S., & Kocaman, S. (2020). A mass appraisal assessment study using machine learning based on multiple regression and random forest. Land Use Policy, 99(September 2019), 104889.
10. Sing, T. F., Yang, J., & Yu, S-M. (2020). Decision Tree and Boosting Techniques in Artificial Intelligence Based Automated Valuation Models (AI-AVM). SSRN Electronic Journal, December, 1–36.
11. Aydinoglu, A. C., Bovkir, R., & Çölkesen, I. (2020). Implementing a mass valuation application on interoperable land valuation data model designed as an extension of the national GDI. Survey Review, 0(0), 1–17.
12. Elmaz, F., Yücel, Ö., & Mutlu, A. Y. (2020). Machine Learning Based Approach for Predicting of Higher Heating Values of Solid Fuels Using Proximity and Ultimate Analysis. International Journal of Advances in Engineering and Pure Sciences, 32(2), 145–151.
13. Şahinler, S. (2000). Basic Principles of Creating a Linear Regression Model with the Least Squares Method Mustafa Kemal University, 5, 57–73.
14. Gülağız, F. K., & Ekinci, E. (2020). Estimation of Housing Price by Using Different Regression Analysis Methods. International Symposium on Industry 4.0 and Applications, Karabük, 12-14 October 2017.

15. Dimopoulos, T., Tyralis, H., Bakas, N. P., & Hadjimitsis, D. (2018). Accuracy measurement of Random Forests and Linear Regression for mass appraisal models that estimate the prices of residential apartments in Nicosia, Cyprus. Advances in Geosciences, 45, 377–382.
16. Günel, A. (2003). Regresyon Denkleminin Başarısını Ölçmede Kullanılan Belirleme Katsayısı ve Kritiği (Some Critics on the Use of Coefficient of Determination as a Significance Test Criterion for Regression Equation), Doğu Üniversitesi Dergisi, 4 (2), 133-140.
17. Takma, Ç., Atıl, H., & Aksakal, V. (2012). Çoklu doğrusal regresyon ve yapay sinir ağı modellerinin laktasyon süt verimlerine uyum yeteneklerinin karşılaştırılması. Kafkas Üniversitesi Veteriner Fakültesi Dergisi, 18(6), 941-944.