



# Mass Appraisal With A Machine Learning Algorithm: Random Forest Regression

*Araştırma Makalesi/Research Article*

 Sibel CANAZ SEVGEN\*,  Yeşim ALİEFENDİOĞLU TANRIVERMİŞ

Gayrimenkul Geliştirme ve Yönetimi Bölümü, Ankara Üniversitesi, Ankara, Türkiye

[ssevgen@ankara.edu.tr](mailto:ssevgen@ankara.edu.tr), [aliefendioglu@ankara.edu.tr](mailto:aliefendioglu@ankara.edu.tr)

(Geliş/Received:18.04.2019; Kabul/Accepted:25.06.2020)

DOI: 10.17671/gazibtd.555784

**Abstract**—Traditional methods in many areas have been replaced by modern methods known as machine learning with the rapidly developing technology and innovations in science. One of these areas is real estate valuation (appraisal) area. Real estate appraisal can be conducted on a single real estate as well as appraisal of more than one real estate together, which is called as mass appraisal, is possible. In this study, a mass appraisal is performed by a Random Forest Regression method, and the results were evaluated. For this purpose, data of 189 flats expected real value and their 13 variables were collected in Yenimahalle, Ankara. 75% of these data were used as training data and 25% as test data. According to the results, a difference of at minimum 600 TL, maximum 60.000 TL and averagely 25.000 TL were observed between the predicted value by the Random Forest regression and the expected real value. According to these results, random forest regression is a successful method in mass appraisal, and it is observed that valuation with different machine learning methods such as random forest regression has a positive effect on time and labor force comparing with valuation of real estate by traditional methods individually.

**Keywords**— Machine Learning, Random Forest Regression, Real Estate Valuation, Mass Appraisal

## Kitlesel Değerlemede Makine Öğrenme: Rasgele Orman Regresyonu

**Özet**— Hızla gelişen teknoloji ve bilimde bulunan yenilikler ile birçok alanda geleneksel yöntemlerin yerini makine öğrenme diye anılan modern yöntemleri almıştır. Bu alanlardan biri ise gayrimenkul değerlendirme alanıdır. Gayrimenkuller tek başına değerlendirilmesi yapılabileceği gibi kitlel olarak ta birçok gayrimenkulün bir arada değerlendirilmesinin yapılması mümkündür. Bu çalışmada, popüler bir makine öğrenme tekniği olan Random Forest (Rasgele Orman) Regresyonu yöntemi seçilerek gayrimenkullerin kitlel değerlendirilmesi yapılmış ve sonuçların gerçek değere yakınlığı incelenmiştir. Bu amaçla, Ankara İli Yenimahalle İlçesinde 189 adet apartman dairesine ait değer ve bu gayrimenkullere ait 13 adet değişken verisi toplanmıştır. Bu verinin, %75'i eğitim verisi ve % 25'i ise test verisi olarak kullanılmıştır. Elde edilen sonuçlara göre, tahmin edilen değer ile olması beklenen değer arasında en az 600 TL, en fazla 60.000 TL ve ortalama 25.000 TL fark gözlemlenmiştir. Bu sonuçlara göre rasgele orman regresyonunun kitlel değerlendirilmede başarılı olduğu, geleneksel yöntemlerle gayrimenkul değerlemek yerine rasgele orman regresyonu gibi farklı makine öğrenme yöntemleriyle değerlendirilmesinin zaman ve insan gücü tasarrufu açısından pozitif etkilerinin olacağı ortaya konmuştur.

**Anahtar Kelimeler**— Makine Öğrenme, Rasgele Orman Regresyonu, Gayrimenkul Değerleme, Kitlel Değerleme

### 1. INTRODUCTION

In recent years, the studies carried out with traditional methods in many disciplines have been replaced by methods called Machine Learning. Specifically, as the size of the data grows, a new research branch called Big Data

has emerged and the effects such as speed and performance have been increased by using machine learning methods. The application areas of machine learning methods are very wide. Many different methods of machine learning in medicine, engineering, finance, sociology etc. are used in

many different applications. The usage of machine learning algorithms in areas are very wide, and for instance, machine learning algorithms were used by researchers for electricity energy needs forecast, intrusion detection, and firm failure prediction [1-3]. Simple definition of machine learning is to teach the computer by giving both output & input and the variables that affect the result so that the result of the remaining big data automatically is obtained by using one of the different machine learning algorithms.

One of the areas where machine learning methods are used in recent years is the real estate appraisal area [4]. The machine learning method usually deals with large data groups in real estate appraisal processes. In real estate science appraisal can be performed individually or together. Valuation of many real estate together is called as mass appraisal. In other words, in a certain date range, the valuation of much similar type of real estate together instead of individually is a mass appraisal process.

There are mass appraisal studies in the literature using different machine learning techniques. In the literature, it the first time for the real estate appraisal using machine learning algorithms was conducted by [5]. A mass appraisal study was carried out with the Artificial Neural Networks (ANN) algorithm. After this study, researches that carried out mass appraisal studies using ANN algorithm increased. In many studies, the researchers observed that ANN yielded good results in real estate valuation [5-17]. On the other hand, some researchers compared ANN according to some classical statistical methods (hedonic, regression) and observed that ANN did not contribute much [18-20]. As can be understood from the literature, the researchers who observed that ANN performed is good in mass appraisal are in the majority of researchers.

A researcher [21] conducted a mass appraisal study using ANN machine learning method. They collected a large number of data samples (16,366) and 18 different variables of these samples in Louisville, Kentucky, the United States. According to their results, they produced different scenarios with different variables and observed that distribution of non-homogeneous variables gave better results comparing with homogeneous distribution. Another study carried out with 100 apartment flats and their 12 variables such as area, number of rooms, building age, existing of an elevator, location etc. in Istanbul province, Çekmeköy District using ANN method, and they acquired reliable results on the mass appraisal [22].

After the success of ANN in mass appraisal, researcher have started to try other machine learning algorithm for mass appraisal studies. For instance, [23] used Random Forest (RF), Support Vector Machines (SVM), ANN and multiple regression methods in the real estate appraisal collected from the United States (USA) real estate sales sites, and observed that the RF algorithm yielded better results compared to SVM and ANN. Using 100 property and 12 variables (area, age, type of heating, sewerage

system, and the material of the house) a mass appraisal study was performed by [4]. The linear regression model was compared with the Support Vector Machine (SVM), and it is found out that the linear regression results are better than SVM results. [24] produced land value maps as a result of nominal valuation using fuzzy logic technique. According to their results, the authors observed that fuzzy logic technique is suitable for mass appraisal studies. 16,601 samples for 2006–2017 from the transaction records for apartments with 26 parameters and RF approach, and it was found that the RF method may be a useful complement to the hedonic models [25]. [26] made a literature reviews on ANN techniques for mass appraisal. The researchers pointed out that in 16 out of the 21 studies, the ANN technique was found to either outperform or be a good alternative to Hedonic Pricing Model (HPM), while only a few studies observed not predictive accuracy of the ANN model.

There are also many studies that compare different machine learning techniques especially with RF algorithms to other machine learning techniques in mass appraisal. For example, RF and Recurrent Neural Networks (RNN) were compared in Brazil for a mass appraisal, and it was found out that RF worked well with numeric features better than RNN [27]. 57,974 real estate data and their 44 property specific features were used to conduct a comparison study in Sweden [28]. Three different machine learning techniques; RF, SVM and ANN were compared for a mass appraisal study. The researchers observed that RF algorithms gave better results [28]. [29] conducted a study with 29,680 real estate data and 55 variables in Pune city using RF methods to prove that RF gives reliable results for mass appraisal techniques.

It is observed from the literature that RF algorithm gives the best performance for mass appraisal studies. Therefore, in this study, a mass appraisal is aimed to perform with RF Regression algorithm. Furthermore, in many mass appraisal studies that conducted by RF and other machine learning algorithms, the data is generally acquired from official institutions and generally with more real estate data and its variables. The aim of this study is differ from other in this way. The data in this study were collected from a Turkish real estate website with 13 variables to figure out the RF performance with comparably less data from the literature.

In the Turkish Civil Code (Article 704) real estate is defined in 3 categories; Land, independent and permanent rights registered in a separate page in the Land Registry Records, and independent units registered in the Condominium Registry Records. The type of real estate, which is the subject of this study and used for the mass appraisal is the flats, which are independent units of apartments (flats). Using flat data and 13 dependent variables of the flats from Yenimahalle district of Ankara province (Capital of Turkey), a popular machine learning algorithm, RF regression was performed for the mass appraisal study, and the results were visualized by ArcGIS version 10.1. The mass appraisal results by RF regression

were also evaluated in the view of the importance of the variables affecting the results.

## 2. RELATIONS BETWEEN MASS APPRAISAL AND MACHINE LEARNING

Mass appraisal is the valuation of several real estates together with the same methods. It is expected that the real estates to be used in mass appraisal will have similar features in order to be valued and that these features should exist in many of these real estates. Mass appraisal is often used to calculate the property tax [4].

In contrast to the mass appraisal, real estates also can be valued individually. Generally, in the single valuation all the different individual features of the real estate are used while common features of the real estates are sought in the mass valuation. Mass appraisal methods are generally based on the method of comparative sales analysis [4]. In other words, the value of a real estate is determined by the comparative analysis of other real estates, and computers and programs carry out whole process.

There are different methods in real estate valuation and these methods can be divided into three categories: traditional, statistical and modern methods [30]. Widespread traditional methods are comparative sales analysis, income approach and cost method approach. Statistical methods are more modern method than traditional methods and the popular ones are multiple regression and hedonic valuation methods. Modern methods can be considered as machine learning methods and types of these methods. Fuzzy logic, ANN, SVM, RF are some of the popular algorithms used in modern valuation methods. Especially in mass appraisal studies, the use of statistical or modern methods will contribute to valuation in terms of speed and accuracy. Therefore, in this study, the RF algorithm which can be used in modern valuation is tested as mass appraisal and the results are analyzed.

Solving problems with large amounts of data one by one is a waste of time. Instead, it is more appropriate to create a model by using some amount of data and to estimate the value for the remaining data. In these cases, machine learning can be discussed. Machine learning is the branch of artificial intelligence that determines the connection between an input data and its results within an algorithm. Machine learning, solving a problem, using the data related to that problem, modeling the problem with different machine learning algorithms and solving the problem for the rest of the data. In other words, machine learning can be defined as using the input and output of data for a particular problem, estimating the model of the problem and estimating the result for the remaining inputs (Figure 1).

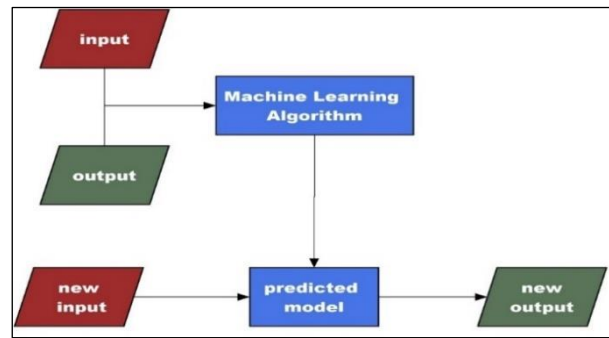


Figure 1. Principles of machine learning algorithms

If the input and output are known, the machine learning algorithms try to determine the function between input and output. Machine learning types are generally divided into 3 main classes, and Figure 2 summarizes the general classification of machine learning. Machine learning is generally classified as Supervised Learning, Unsupervised Learning and Reinforcement Learning. Supervised learning is divided into two sub-groups as classification and regression. Supervised learning is generally is task-based (task driven) model, and it depends on the application, while unsupervised learning is based on data (data-driven).

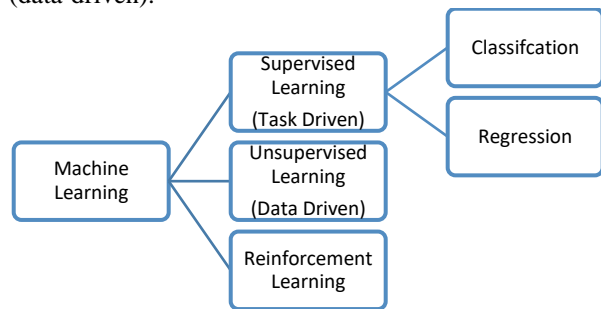


Figure 2. General classification of machine learning

Table 1 shows the popular machine learning algorithms according to the machine learning categories. Some machine learning algorithms can be located in different type of machine learning categories, in which example neural networks and decision trees can be used in both regression type and classification type supervised machine learning categories. In supervised learning, a model is estimated using data. In this type of learning, both input and output are given to the algorithm. The most challenging part of this type of learning is the creation of a training data. Because the problem will be solved using the training data, the reliability of the training data will directly affect the result. Whereas, unsupervised learning, estimation is made according to the data. Training data are not used in this kind of learning. In unsupervised learning, the data is grouped in such a way that the model is created and the other data are assigned to the most appropriate group. The machine learns its own and categorizes the data according to its characteristics. In this kind of learning, the separation of similar features can be difficult and they can achieve more erroneous results than supervised learning. Lastly, reinforcement learning method is noteworthy as a learning type similar to human learning. The machine learns by making observations and concludes. This group

is a new type of learning compared to other types of machine learning and is often used in gaming algorithms.

Table 1. Popular machine learning algorithms in 3 categories

Machine Learning Algorithms			
Supervised		Unsupervised	Reinforcement
Classification	Regression		
Support Vector Machines	Linear Regression	K-Means	Artificial Neural Networks
Neural Networks	Neural Networks	Neural Networks	Markov Decision Process
Decision Trees	Decision Trees	Fuzzy Logic	Q-Learning
Random Forest	Random Forest	Gaussian Mixture	
Nearest Neighbor	Ordinary Least Squares Regression	Hidden Markov Model	
Naive Bayes		DBSCAN	

### 3. RANDOM FOREST REGRESSION AND TRAINING DATA

RF is a decision tree machine learning algorithm. RF is used in many different areas such as finance, health sciences, map, electronics, mechanical engineering, physics, and biology. RF machine learning algorithm was found by [31] and it was introduced into the literature as an improved version of bagging method known [32]. RF's accuracy is considered to be one of the highest algorithms in machine learning algorithms [33]. RF is the result of producing more than one decision tree and that's why it is called as "forest". In the RF algorithm, the number of trees (N) and the number of variables (m) to be used in each node are asked to the user. Each tree is generated by randomly selected variables from the training data and this is the reason why algorithm is called as "random". Figure 3 shows an example of a decision tree structure. At the bottom of the N trees, the results are obtained by the RF algorithm and the result is calculated by taking the weights of the variables of the data. In the RF regression machine learning type, the result will be obtained by taking the average results from each tree. The RF model then used to get the results for the rest of the data.

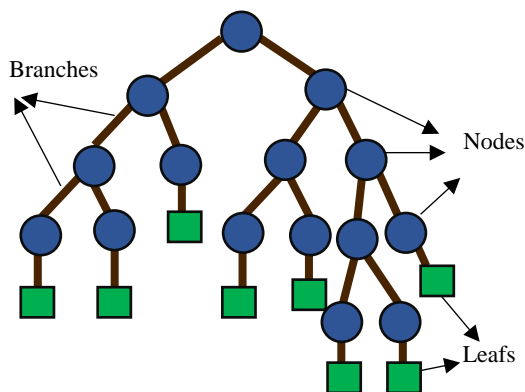


Figure 3. An example of decision tree structure

The first step of the RF algorithm is the creation of training data, as in supervised learning. The RF training data to calculate the model and the remaining test data's results are obtained using this model. Learning, as can be seen in the name, teaches the machine. That is, the process of providing the system to learn. In this process, the training data is the data presented to the algorithm and the variables of this data for the algorithm to form the model. The test data are data that are not used as training data in the model. Using the test data, the quality of the results obtained from the algorithm is checked, and the test data, which is excluded from the training data, is inserted into the model of the RF algorithm and the results from each tree are examined.

The most important criteria in the valuation process are location and when location is mentioned than the most important science is Geographic Information Systems (GIS). GIS is a popular branch of science that is used in many different areas to collect, store, process and present spatial data. GIS software was used to create data as explained in the next subsections. Furthermore, GIS-aided value maps can be generated, and the resultant product can be presented. In this study, value maps also will be generated to visually interpret the results using a GIS software.

#### 3.1. Digitizing Process

The first step of the method is to collect data from the study area. At this stage, the location of each apartment was digitized manually as polygon feature data by using the ArcMap 10.1 software by looking at the location information from the website of real estate trading [34]. In this process, ESRI online base map was used as a map base [35].

Digitizing is a process that creates new vector data from a raster. The vector data can be point, line, and polygon. In this step, the vector data type is polygon since the apartments flat's boundaries from the raster, which is an image in this case, can be extracted as polygon.

#### 3.2 Distance Calculation

Real Estate appraisal can be done using with many variables of the real estate. The variables depend on the professional values. In this study, distances from apartment flats to hospital, school and subway station data were calculated.

Firstly, hospitals, subway stations, and schools are digitized as point features from the map base. Schools, one hospital, and one subway stations were found in the study area. The school, hospital and subway station locations were shown in Figure 4. The distance from these features were calculated by ArcMap using the *Proximity, Near tool*. This tool calculates distance from a point feature (i.e. hospital) to any other features such as point, line and polygon. In this case, points to polygon distances were

calculated. The distances take the closest part of the polygon from the point. This distance calculation process was performed automatically.

Consequently, in this study, a total of 189 flats were digitized from the base map and their some of the variables were acquired from the real estate website. Moreover, distance parameters for the flats were calculated from the GIS software automatically. Summary of the real estate data and their variables are shown below (Table 2).



Figure 4. School, hospital and subway station

Table 2. Example of the data

ID	Barbaros School Distance (m)	Halide School Distance (m)	Fatih School Distance (m)	Hospital Distance (m)	Subway Distance (m)	Value (TL)	# of Rooms	Gross Area (m <sup>2</sup> )	Net Area (m <sup>2</sup> )	Age	# of Bath room	# of floor	Floor	Rent Value
1	1038,12	325,91	654,26	363,84	82,37	360000	4	120	110	30	1	4	1	1100
2	984,20	311,23	623,57	398,79	52,78	450000	4	115	105	35	1	3	1	1250
3	846,77	452,88	680,41	615,16	247,81	360000	4	120	105	34	1	4	2	1200
4	929,10	493,43	749,05	612,82	257,11	299000	4	125	115	40	1	4	1	1000
5	829,00	502,73	713,37	670,84	303,49	280000	4	120	115	35	1	12	5	1100
6	670,57	436,04	94,37	712,35	647,25	380000	7	180	165	17	2	4	4	1200
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
189	639,32	438,87	106,90	720,72	643,38	350000	4	120	110	12	1	4	2	1650

### 3.3 Programming Random Forest Regression Algorithm

The code of RF algorithm was written in Python, which has a written code of RF in its library. For Python codes, Anaconda 2018.12 version and Python 3.7 version are used. The written codes firstly read the data, call the RF run and print the results. The results are obtained and printed on a separate file.

## 4. RESULTS and DISCUSSION

### 4.1 Study Area and Data

The subject real estate of the mass appraisal in this study is the flats which belong to the condominium class. Yenimahalle district, which is one of the oldest settlements of Ankara Province, was chosen as the study area. Neighborhoods were selected as the central location of the district of Yenimahalle, which is part of the district of Ragıp Tüzün, Çarşı and Yunus Emre, Yeniçağ and Gayret Districts (Figure 5). Another reason to select this area as a study area is that presence of schools, hospitals and subway around it. Also it will be observed whether the absence of

these buildings have an impact on the value of real estates or not.



Figure 5. Study area, Ankara, Yenimahalle district (yellow boundary)

In this study, flats and different variables of the flats were collected from the website [34]. In this study, the data of the flats' 13 variables are shown in Table 3. These variables are gross area, net area, building age, number of rooms, floor, rent return, subway, hospital and proximity to schools. Gross area, net area, building age, number of room, number of bedroom, rent value, total floor number, flat's floor number were extracted from Turkish real estate website [34], while distances to schools, hospital and subway station were automatically generated using GIS software.

Table 3. 13 Variables of flats used in this study

Output	P <sub>1</sub>	P <sub>2</sub>
Value	Gross Area	Net Area
P <sub>3</sub>	P <sub>4</sub>	P <sub>5</sub>
Building Age	Number of room	Number of bathroom
P <sub>6</sub>	P <sub>7</sub>	P <sub>8</sub>
Distance to Subway	Total floor number	Floor number
P <sub>9</sub>	P <sub>10</sub>	P <sub>11</sub>
Rent Value	Distance to School 1	Distance to School 2
P <sub>12</sub>	P <sub>13</sub>	
Distance to School 3	Distance to Hospital	

#### 4.2 Experimental Results

The RF regression was written and executed in Python. The two important parameters for RF are N and m as previously mentioned. Using these parameters in different ways, the algorithm has been tested. Choose of training and test data percentage can also affect the results. In the literature training data size changes between 60%-80%, and generally it is taken as 75%, while test data range changes 20%-40%. Training data are used to educate the machine learning algorithms, while test data are used for validation of the results. In this study, 75% of all data was used as training data and 25% as test data. In other words, total of 189 apartment data, 152 of which were used as training data and 47 of them were used as test data to examine the results of the algorithm. Table 4 shows the combination of various N and m parameters and the real value and estimated value of 47 test data. While the lowest difference was 664 TL, the highest difference was observed as 59.494 TL. Although the best results were found to be N = 150, the change of the N (number of trees) parameter did not show much effect in this study. In other words, increasing or decreasing the number of trees in the RF algorithm did not make big effect on the results for this study.

Table 4. The estimated results of the test data by the RF regression

ID	Real Value (TL)	Estimated Value (TL) N=100, m=3	Estimated Value (TL) N=150, m=3	Estimated Value (TL) N=200, m=3	ID	Real Value (TL)	Estimated Value (TL) N=100, m=3	Estimated Value (TL) N=150, m=3	Estimated Value (TL) N=200, m=3
1	295000	284972	292446	289016	25	185000	194963	191585	189764
2	470000	500386	473324	490692	26	385000	332720	340423	335188
3	275000	216111	226592	223907	27	520000	549058	559791	560149
4	220000	191068	194868	197681	28	459000	484294	461762	467431
5	170000	203353	211063	208712	29	250000	254020	254594	258069
6	240000	196677	197821	200657	30	180000	169556	170044	171995
7	350000	289383	300161	304841	31	220000	192610	193015	193703
8	650000	616907	600630	608878	32	250000	216730	223156	225442
9	230000	223987	231961	230664	33	620000	583331	579771	582960
10	190000	195434	202766	203638	34	325000	348411	354898	338208
11	158000	176924	172255	169483	35	390000	379052	373441	374610
12	240000	196831	199657	203481	36	340000	346013	355149	349859
13	440000	402896	410206	402577	37	230000	202294	201390	200227
14	220000	185612	191198	192195	38	260000	204809	209086	205732
15	213000	231153	229256	234000	39	327000	340906	338866	328121
16	160000	186520	183132	187904	40	500000	446454	440506	455383
17	330000	289581	298561	294992	41	180000	194243	188960	187429
18	215000	220588	220735	224012	42	210000	241070	248454	245287
19	350000	334301	336604	334422	43	213000	223796	218754	215876
20	455000	442203	453168	438730	44	430000	438440	469434	457645
21	455000	431798	442287	428165	45	380000	381006	416830	406568
22	160000	181537	177637	182380	46	230000	210130	212821	208223
23	245000	208328	216352	216255	47	425000	428117	436260	442554
24	320000	321887	333968	327452					

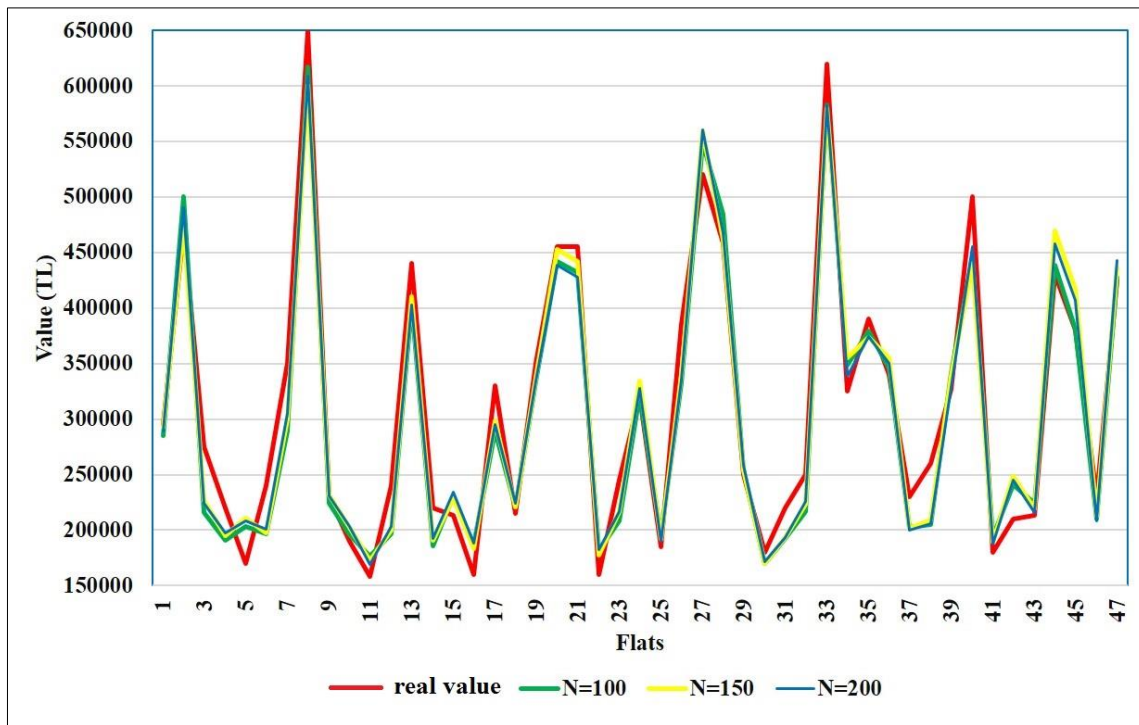


Figure 6. Real value and estimated values by RF algorithm

Figure 6 shows how the values of flats calculated by RF regression vary according to the number of different trees and how they change according to the real value. In Figure 7, the difference between the real value and the estimated values are obtained as a result of different tree numbers are taken and shown graphically. These differences can range from 600 TL to 60,000 TL. Excessive difference is seen as

related with reliable data of real estate variables. When the 13 variables are examined, it is noteworthy that the variables that have the most effect on the value are the rent value, area, distance to school, hospital, subway, building age, number of floors and floor of the flats. The number of rooms and the number of baths affect the model very less.

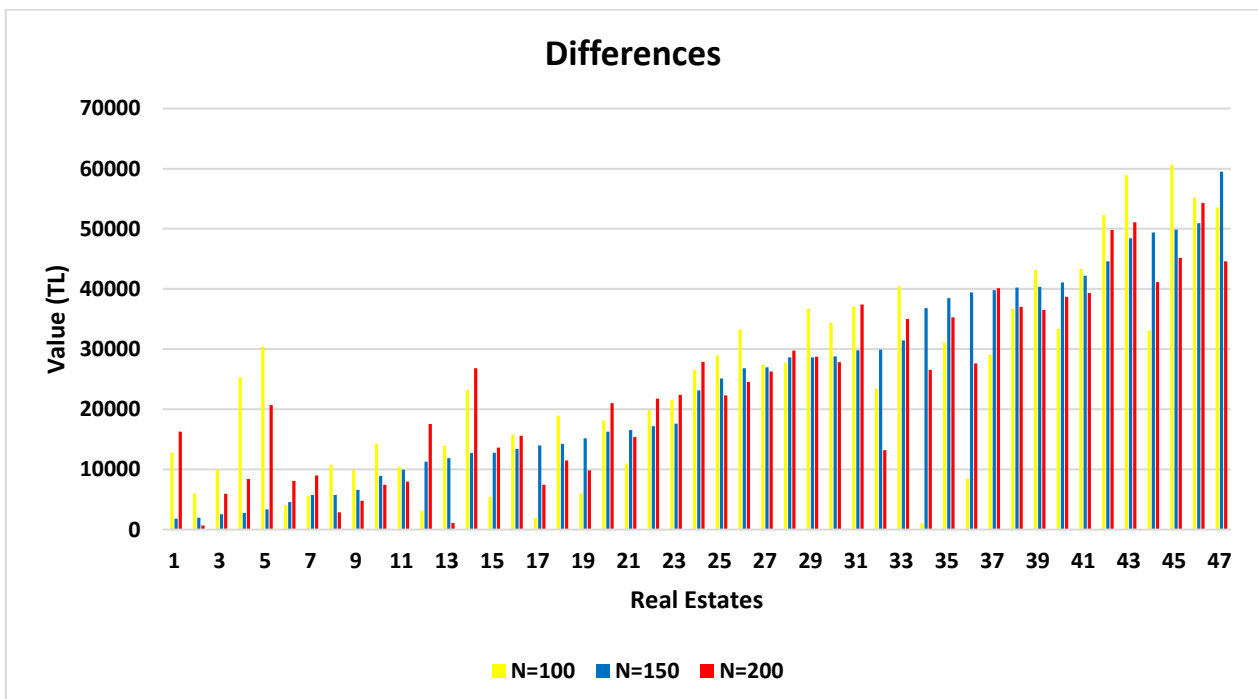


Figure 7. The difference between the estimated values of real values

Moreover, RMSE of the results were also calculated for RF algorithm. Eq (1)

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (P_i - Q_i)^2}{n}} \quad (1)$$

RMSEs were calculated for N=100, N=150 and N=200 and it was found as, 29309.90, 28735,35, 27747,77 respectively.

To evaluate RF results, an ANN algorithm (Multi Layer Perceptron with two hidden layers, 0.01 learning rate, sigmoid activation function and Adam optimizer) is also performed with the same dataset for comparison. Python 3.7 in Anaconda 2018.12 with scikit-learn machine learning library used for the execution of the ANN

algorithm. %75 of the data was used for training and, and % 25 of the data was test data. The training and testing data partitions are identical to RF experiment.

Unlikely to RF, a data scaling preprocessing step is required in ANN. Therefore, the data is scaled from 0 to 1 for stability and convergence. Table 5 Shows the results and the difference between real value and RF (N=150) and ANN (same configuration above) results. According to the results, it can be said that RF gives better results for mass appraisal. ANN predictions and real value difference ranges between 3199 and 238017. In addition, RMSE of the ANN also was calculated and found as 69664. The RMSE is almost triple of RMSEs of RF results. Overall, this result concludes that RF gives closer results with respect to the real values than ANN algorithm.

Table 5. The estimated results of the test data by the RF regression

ID	Real Value (TL)	RF Est. Value (TL) N=150, m=3	ANN Est. Value (TL)	Abs (Difference Real vs. RF)	Abs (Difference Real vs. ANN)	ID	Real Value (TL)	RF Est. Value (TL) N=150, m=3	ANN Est. Value (TL)	Abs (Difference Real vs. RF)	Abs (Difference Real vs. ANN)
1	295000	292446	209187	2554	85813	25	185000	191585	204556	6585	19556
2	470000	473324	552081	3324	82081	26	385000	340423	428810	44577	43810
3	275000	226592	387283	48408	112283	27	520000	559791	671596	39791	151596
4	220000	194868	206056	25132	13944	28	459000	461762	462199	2762	3199
5	170000	211063	242574	41063	72574	29	250000	254594	488017	4594	238017
6	240000	197821	209825	42179	30175	30	180000	170044	160450	9956	19550
7	350000	300161	241965	49839	108035	31	220000	193015	196065	26985	23935
8	650000	600630	634798	49370	15202	32	250000	223156	273044	26844	23044
9	230000	231961	280564	1961	50564	33	620000	579771	469199	40229	150801
10	190000	202766	170676	12766	19324	34	325000	354898	408234	29898	83234
11	158000	172255	170026	14255	12026	35	390000	373441	382974	16559	7026
12	240000	199657	227782	40343	12218	36	340000	355149	447852	15149	107852
13	440000	410206	391967	29794	48033	37	230000	201390	296927	28610	66927
14	220000	191198	181785	28802	38215	38	260000	209086	223471	50914	36529
15	213000	229256	251551	16256	38551	39	327000	338866	336408	11866	9408
16	160000	183132	155536	23132	4464	40	500000	440506	455145	59494	44855
17	330000	298561	312065	31439	17935	41	180000	188960	221764	8960	41764
18	215000	220735	209789	5735	5211	42	210000	248454	236430	38454	26430
19	350000	336604	357533	13396	7533	43	213000	218754	223223	5754	10223
20	455000	453168	480741	1832	25741	44	430000	469434	518844	39434	88844
21	455000	442287	496083	12713	41083	45	380000	416830	337605	36830	42395
22	160000	177637	184659	17637	24659	46	230000	212821	246357	17179	16357
23	245000	216352	201589	28648	43411	47	425000	436260	547535	11260	122535
24	320000	333968	415681	13968	95681						

In the last step, the results for the test data were reflected on the map and a value map was created for the area by taking the interpolation of the estimated results (Figure 8). In addition, a map of the known real values is also created

(Figure 9). When these two maps were compared, it was observed that there was almost no difference between the actual and the RF algorithm's estimated value



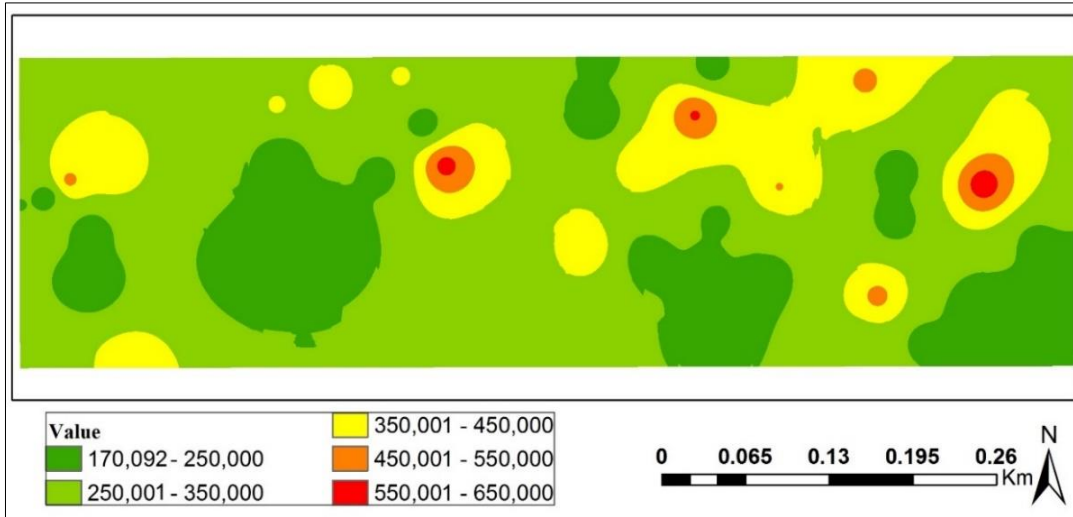


Figure 8. Value map created by using estimated value of the test data by RF regression

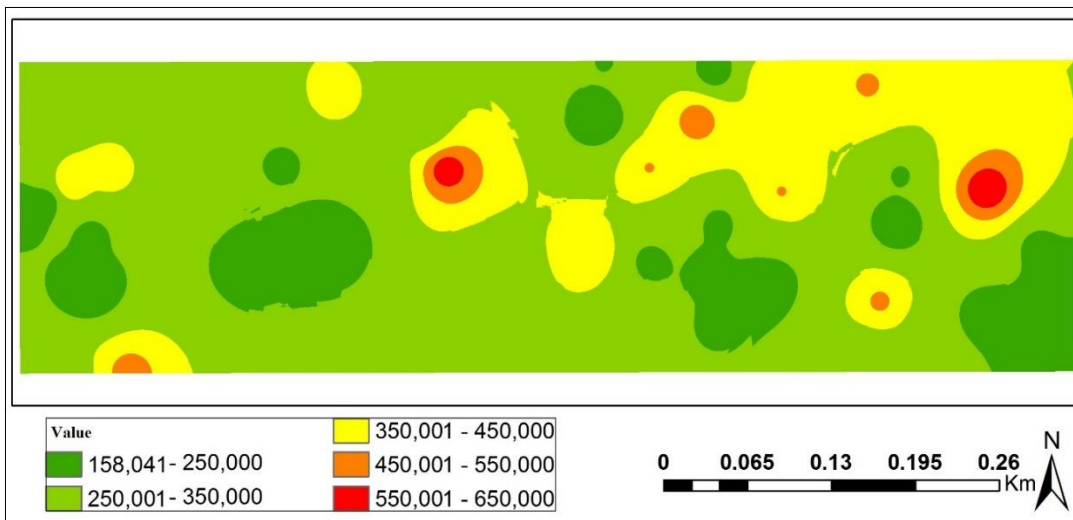


Figure 9. Value map created by using real value of the test data

## 5. CONCLUSIONS

When the real estate appraisal is conducted in mass appraisal form, it is seen that time and labor force to estimate value is less than valuing the real estates individually with traditional methods. Instead, modern valuation methods have been used in mass valuations. In modern methods, machine learning comes to the forefront in real estate appraisal as in every field. The machine learning algorithms are a computer aided method which uses the variables and results of a data to make a model estimation of the problem and produce results for the remaining data using this model. Although there are many different algorithms in machine learning, one of the most popular is the RF algorithm which is the decision tree structure and the RF algorithm is based on both classification and regression.

In this study, using the RF regression method, the sales values of flats and some of the variables were collected neighborhoods Yenimahalle District of Ankara Province and mass appraisal process was performed. The variables

of real estate are the distance to 3 different schools, the distance to the subway, the distance to the hospital, the number of rooms, the number of bathrooms, the rent, the age of the building, the floor of the building and the floor on which the real estate is located and the net and gross area values. A total of 189 real estate data were collected and 75% (142) of these data were training data, and 25% (47) were the data collected for the purpose of analyzing the result of the algorithm. The real sales value of the 47 data used during the RF algorithm to test result. The estimated price difference between range between 600 TL-60.000 TL and the average value has been found to be 25.000 TL difference. When the excess errors are examined, it is assumed that this may happened due to some variable values which can be considered as not very reliable. Moreover, the results are compared with the ANN algorithm as well. IT is found out that RF gives better results compared with ANN for mass appraisal.

Finally, for visual quality control, the value maps were generated and compared for the test data by using the values obtained from both the real value obtained during

the data collection phase and the results of the RF algorithm. From the results obtained, it was observed that the appraisal result was mostly affected by the rental income and area dimensions of the real estate as well as the building age, the number of bathrooms and the number of rooms had little effect.

The study show us that using with less variables of the apartments flats can also generated reliable results by RF algorithm. In the literature, the real estate data and their valuables are huger than this study's data. Also, in Turkey it is a very big issue that the real transaction data cannot be reached from the former institutions since people generally present non real values of the real estate while they sell/buy real estate properties.

Furthermore, in future studies, it is aimed to conduct a study on mass appraisal by RF algorithm with a data set which is more reliable, larger and have more variable. Moreover, the RF algorithm, different machine learning algorithms such as ANN, SVM will be also tested in mass appraisal and these algorithms will be compared.

## REFERENCES

- [1] S. Özden, A. Öztürk, "Yapay Sinir Ağları ve Zaman Serileri Yöntemi ile Bir Endüstri Alanının (İvedik OSB) Elektrik Enerjisi İhtiyaç Tahmini.", *Bilişim Teknolojileri Dergisi*, 11 (3), 255-261, 2018.
- [2] O. Kaynar H. Arslan, Y. Görmez., "Işık, Makine Öğrenmesi ve Öznitelik Seçim Yöntemleriyle Saldırı Tespiti", *Bilişim Teknolojileri Dergisi*, 11(2), 175-185, 2018.
- [3] H. Erdal, T. Yapraklı, "Firma Başarısızlığı Tahminlemesi: Makine Öğrenmesine Dayalı Bir Uygulama". *Bilişim Teknolojileri Dergisi* 9(1), 2016.
- [4] V. Kontrimas, A. Verikas, "The mass appraisal of the real estate by computational intelligence", *Applied Soft Computing*, 11, 443-448, 2011.
- [5] R.A. Borst, "Artificial neural networks: the next modelling/calibration technology for the assessment community", *Property Tax Journal*, 10(1), 69-94. 1991.
- [6] D. P. Tay., D. K. Ho, 1992. "Artificial intelligence and the mass appraisal of residential apartments." *Journal of Property Valuation and Investment*. 1992.
- [7] W. McCluskey, "Predictive accuracy of machine learning models for the mass appraisal of residential property", *New Zealand Valuers Journal*, 16(1), 41-47. 1996.
- [8] I.D. Wilson, S.D. Paris, J.A. Ware, D.H. Jenkins, "Residential property price time series forecasting with neural networks", *Knowledge-Based Systems*, 15(5), 335-341. 2002.
- [9] X.J. Ge, G. Runeson, "Modeling property prices using neural network model for Hong Kong", *International Real Estate Review*, 7(1), 121-138. 2004.
- [10] G. Özkan, Ş. Yalpir, Ş. O. Uygunol, "An investigation on the price estimation of residable real-estates by using artificial neural network and regression methods", **paper presented at the 12th Applied Stochastic Models and Data Analysis International conference (ASMDA)**, Crete, May 29-June 1. 2007
- [11] K. C. Lam, C. Y. Yu, K. Y. Lam, "An artificial neural network and entropy model for residential property price forecasting in Hong Kong", *Journal of Property Research*, 25(4), 321-342. 2008.
- [12] H. Selim, "Determinants of house prices in Turkey: hedonic regression versus artificial neural network", *Expert Systems with Applications*, 36(2), 2843-2852. 2009
- [13] A. G. Musa, O. Daramola, E. A. Owoloko, O. O. Olugbara, "A neural-CBR system for real property valuation". *Journal of Emerging Trends in Computing and Information Sciences*, 4(8), 611-622. 2013.
- [14] J.N.M. Tabales, C.J.M Ocerin, F.J.R., Carmona, "Artificial neural networks for predicting real estate prices", *Revista de Metodos Cuantitativos para la Economia y la Empresa*, 15, 29-44. 2013.
- [15] S. Ahmed, M. Rahman, S. Islam, "House rent estimation in Dhaka city by multilayer perceptions neural network", *International Journal of U-and E-Service, Science and Technology*, 7(4), 287-300. 2014.
- [16] P. Morano, F. Tajani, C. M. Torre, "Artificial intelligence in property valuations: an application of artificial neural networks to housing appraisal", paper presented at the **11th International Conference on Energy, Environment, Ecosystems and Sustainable Development (EEESD '15)**, Canary Islands, 10.12.2015.
- [17] A. Varma, A. Sarma, S. Doshi, R. Nair, "House Price Prediction Using Machine Learning And Neural Networks". **In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT, 1936-1939)**. IEEE. 2018.
- [18] A. Worzala, M. Lenk, A. Silva, "An exploration of neural networks and its application to real estate valuation", *Journal of Real Estate Research*, 10(2), 185-201. 1995.
- [19] M. M. Lenk, E. M. Worzala, A. Silva, "High-tech valuation: should artificial neural networks bypass the human valuer?", *Journal of Property Valuation and Investment*, 15(1), 8-26. 1997.
- [20] W. J. McCluskey, M. McCord, P. Davis, M. Haran, D. McIlhatton, "Prediction accuracy in mass appraisal: a comparison of modern approaches", *Journal of Property Research*, 30(4), 239-265. 2013.
- [21] J. Zurada, A. Levitan, J. Guan. "A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context", *Journal of Real Estate Research*, 33(3), 349- 87, 2011.
- [22] E. Saraç, **Yapay Sinir Ağları Metodu İle Gayrimenkul Değerleme**, Yüksek Lisans Tezi, İstanbul Kültür Üniversitesi, 2012.
- [23] A. S. Ravikumar, Real Estate Price Prediction Using Machine Learning. *National College of Ireland*, 2016.
- [24] M. A. Derinpinar, A. Ç. Aydınoglu, "Bulanık Mantık ile Coğrafi Bilgi Teknolojilerini Kullanarak Taşınmaz Değerlemesi", 15. Türkiye Harita Bilimsel ve Teknik Kurultayı, Ankara, 25-28 Mart 2015.

- [25] J. Hong, H. Choi, W. Kim, "A House Price Valuation Based On The Random Forest Approach: The Mass Appraisal Of Residential Property In South Korea" *International Journal of Strategic Property Management*, 24(3), 140-152, 2020.
- [26] R. B., Abidoye, A. P. C. Chan. "Artificial neural network in property valuation: Application framework and research trend", *Property Management*, 35(5), 554-571, 2017.
- [27] B. K. A. Afonso, L. C. Melo, W. D. G. Oliveira, S. B. S. Sousa, L. Berton. "Housing Prices Prediction with a Deep Learning and Random Forest Ensemble", 2019.
- [28] M. Dellstad, Comparing three machine learning algorithms in the task of appraising commercial real estate Degree Project in Computer Science and Engineering. 2018.
- [29] R. Sawant, Y. Jangid, T. Tiwari, S. Jain, A. Gupta, "Comprehensive Analysis of Housing Price Prediction in Pune Using Multi-Featured Random Forest Approach," **2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)**, 1-5, Pune, India, 2018.
- [30] N. Erdem. "Toplu (Küme) Değerleme Uygulama Örnekleri ve Ülkemiz İçin Öneriler", TMMOB Harita ve Kadastro Mühendisleri Odası, 16. Türkiye Harita Bilimsel ve Teknik Kurultayı, Ankara, 3-6 Mayıs 2017.
- [31] L. Breiman, "Random forests", *Machine learning*, 45(1), 5-32, 2001.
- [32] L. Breiman, "Bagging predictors", *Machine Learning*, 26(2), 123-140, 1996.
- [33] İnternet: L. Breiman, A. Cutler, "Random Forest", [http://www.stat.berkeley.edu/~breiman/RandomForests/cc\\_home.htm](http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm), 2005.
- [34] İnternet: Hürriyet Emlak. <https://www.hurriyetemlak.com>. 15.12.2018.
- [35] Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community. 2018