sciendo
www.degruyter.com/view/j/remav

# COMPARISON OF MACHINE LEARNING ALGORITHMS FOR MASS APPRAISAL OF REAL ESTATE DATA

**Sibel Canaz Sevgen[1]*, Yeşim Tanrivermiş[2]**

[1]*Department of Real Estate Development and Management, Ankara University, Emniyet, Dögol Cd., 0600 Yenimahalle/Ankara, Turkey, e-mail: ssevgen@ankara.edu.tr, ORCID 0000-0001-5552-6067*

[2]*Department of Real Estate Development and Management, Ankara University, Emniyet, Dögol Cd., 0600 Yenimahalle/Ankara, Turkey, e-mail: aliefendioglu@ankara.edu.tr, ORCID 0000-0002-0859-7150*

** Corresponding author*

ARTICLE INFO

*Keywords:*

mass appraisal, machine learning algorithms, random forest, artificial neural network, real estate valuation map

*JEL Classification:*

R39

ABSTRACT

In recent years, machine learning algorithms have been used in the mass appraisal of real estate. In this study, 5 machine learning algorithms are used for residential type real estate. Machine learning algorithms used for mass appraisal in this study are Artificial Neural Networks (ANN), Random Forest (RO), Multiple Regression Analysis (MRA), K-Nearest Neighborhood (k-nn), Support Vector Regression (SVR). To test the study, real estate data collected from the central districts of Ankara, were used. The main purpose of this study is to find out which machine learning algorithm gives the best results for the mass appraisal of real estates and to reveal the most important variables that affect the prices of real estate. According to the results obtained for the city of Ankara, it was observed that the best algorithm for mass appraisal is RF in residential-type real estates, followed by the ANN, k-nn, and linear regression algorithms, respectively. According to the results obtained from the residential real estate, it was concluded that heating and distances to places of importance had the greatest effect on the value.

*Citation:* Canaz Sevgen, S., & Tanrivermiş, Y. (2024). Comparison of machine learning algorithms for mass appraisal of real estate data. *Real Estate Management and Valuation, 32*(2), 100-111. https://doi.org/10.2478/remav-2024-0019

## 1. Introduction

In recent years, research carried out with traditional methods have been replaced by methods called Machine Learning (ML) in many branches of science. In particular, as the data size grows, a new research branch called big data has emerged and the effects, such as speed and performance degradation, observed with traditional methods in applications related to large data have been increased by using ML algorithms. The usage areas of ML algorithms are quite wide. Many different algorithms of ML are used in many different applications such as medicine, engineering, finance, sociology, etc. One of the areas where ML algorithms have been used in recent years is the real estate valuation area (Kontrimas & Verikas, 2007; Yılmaz & Bostancı, 2023). ML algorithms generally deal with large data groups, and in real estate valuation processes, the valuation process can be done individually as well as in mass form. When many properties are valued collectively, it is called mass valuation, or more commonly, mass appraisal. Also, mass appraisal with the ML algorithm are crucial for sustainably, and sustainably is the awareness around this concept is raised as it is a living, changing and growing concept (Gültekin et al., 2017). For instance, Unel and Yalpir (2023) explored sustainable mass appraisal systems for taxation, while Sisman et al. (2023) conducted a study on the development of a novel hybrid model for mass appraisal in real estates, specifically in the context of sustainable land management.

The mass appraisal of real estate study was carried out using one of the ML algorithms - Artificial Neural Networks (ANN) - for the first time by Borst (1991). After this study, the number of mass appraisal studies using the ANN algorithm has increased. Researchers observed, in their scientific publications, that the ANN

algorithm gives successful results in the mass appraisal of real estate (Ahmed S et al., 2014; Bilgilioğlu & Yılmaz, 2023; Lam et al., 2008; McCluskey, 1996; Morano & Tajani, 2013; Musa et al., 2013; Özkan et al., 2007; Sampathkumar et al., 2015; Saraç, 2012; Selim, 2009; Tabales et al., 2013; Tay & Ho, 1992; Torres-Pruñonosa et al., 2021; Valier, 2020; Varma et al., 2018; Wilson et al., 2002; Xin & Runeson, 2004). On the other hand, some researchers compared the ANN algorithm with the regression method and observed that the ANN algorithm did not contribute much to the mass appraisal result (Lenk et al., 1997; McCluskey, 1996; Worzala et al., 1995). As can be understood from the literature, although the majority of researchers who observed that the ANN algorithm gave good results in mass appraisal, there were also researchers who observed that ANN did not contribute too much.

Some researchers, who have found that the ANN algorithm generally performs better in mass appraisal, have started to try ML algorithms other than ANN in recent years. For example, 16,601 real estate data from purchase and sale transaction records for flats with 26 variables between 2006 and 2017 were used for mass appraisal, and it was found that the Random Forest (RF) algorithm (Breiman, 2001) could be a useful complement to hedonic models (Dambon et al., 2022; Hong et al., 2020). Sawant et al. (2018) conducted a mass appraisal study using the RF algorithm with a high number of data (29,680 sales and 55 variables) and observed that the RF algorithm gave very good results in valuation processes. Iban (2022) compared RF and the tree-based algorithm with regression, and observed that ML algorithms perform better than classic regression methods. Ravikumar (2017), on the other hand, used RO, Support Vector Machines (SVM), ANN, and multiple regression (MR) methods for the mass appraisal of 49,980 real estates collected from real estate sales websites in the United States (USA) and observed that the RF algorithm gave better results than SVM and ANN. Some authors collected 12,223,582 ads from Brazilian real estate websites from 2015 to 2018 and compared 24 variables of these real estates with the RF and Recurrent Neural Network (RNN) algorithms. Dellstad (2018) used RF, ANN, and SVM algorithms in mass appraisal, and the researcher collected 57,974 samples with 44 variables in Switzerland, with the observation that the RF algorithm gave the best results. Hong et al. (2020) also observed that the RF method was more acceptable than the hedonic models for real estate mass appraisal studies. The SVR, RF, XGBoost, LightGBM, and CatBoost algorithms were combined and tested with data on 57,000 apartments in Seoul by Hong and Kim (2022). Gnat (2021) claimed that machine learning methods such as Knn and XGBosst gives more accurate result than multiple regression models.

As can be seen from the literature, the ML algorithms commonly used in mass appraisal studies are RF, ANN, and SVR. In this study, the aim is to compare the frequently used ML algorithms, which are RF, ANN, SVR, k-nn, and MR analysis for Turkish residential real estate data. The central districts of Ankara (capital of Turkey) province were chosen as the study area and the variables and values of the real estate were obtained from a reliable real estate Internet data portal. A raster-based value map was produced using the interpolation method as the output product of the study.

## 2. Machine Learning Algorithms

Machine learning, artificial intelligence, and deep learning have been frequently used in many fields for the last 30 years. Although these three terms are similar to each other, there are differences between them. Artificial intelligence is in a higher position to be more inclusive; it can be defined as a technique that mimics human behavior. Machine learning, on the other hand, as the name suggests, is making predictions as a result of a learning process. Deep learning, on the other hand, can be defined as a technique that uses neural networks to reveal characteristic features in the data. Artificial intelligence is a broader term that encompasses machine learning and deep learning.

Solving problems with large amounts of data manually, one by one, is seen as a waste of time. Instead, it seems more appropriate to create a model by using a certain amount of data and to make a value estimation for the remaining data. Machine learning is a branch of artificial intelligence that determines the connection between input data and its results within an algorithm.

### 2.1. Random Forest Regression

RF is a machine learning algorithm, and is a type of decision tree. It is used in many different fields such as finance, health sciences, thematic maps, electronics, mechanical engineering, physics, and biology. RF was developed by Breiman in 2001, and an improved version of the bagging method invented by Breiman (1996). It is considered one of the algorithms with the highest accuracy in ML (Breiman & Cutler, 2005). RF achieves the result by generating more than one


eISSN: 2300-5289 | Received 2023-09-19 | Revised 2023-12-07 | Accepted 2024-02-16

decision tree, and is called a "forest" algorithm because of the use of many decision trees. In the RF algorithm, the user is asked how many trees (N) there will be and the number of variables (m) to be used in each node. Each tree is created with randomly selected variables from the training data.

The results are obtained with the RF algorithm, and the model is created by calculating the weights of the variables of the data at the end of the created N trees. When looking at the results of each tree, the result in the RF classification machine learning type is obtained with the highest probability; on the other hand, in the RF regression machine learning type, the result is obtained by taking the average from each tree. Unlike pixel-based RF classification, RF regression is an object-based process. In the regression algorithm, the results and the variables are given to the machine, and the machine learns from these data and tries to find the result for the remaining test data. For regression tasks, the final prediction is often the average of the predictions from all the trees. Mathematically, if the predictions are denoted of individual trees as $y_i$, the prediction of the Random Forest ensemble ($\hat{y}$) can be expressed as the average for regression:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{n} y_i \quad (1)$$

### 2.2. Artificial Neural Network Algorithm

ANN is an ML algorithm derived from the structure of the biological human brain. As in many ML algorithms, the learning process is modeled mathematically. ANN starts with the modeling of neurons, which are the biological units of the human brain. It is one of the most widely used ML types in many fields today. It is an ML algorithm that is frequently encountered in many areas, such as data mining, optical character recognition, direction determination in robots, communication, internet of things, diagnosis of diseases, classification of trees and the valuation of real estates.

To better understand the ANN algorithm, it is first of all necessary to examine the biological brain nerve cell. A biological nerve cell consists of a body, an axon, dendrites, and many nerve cells. The extensions between the nerve cell and the nerve end are called axons. Dendrites transmit incoming signals to the nucleus. The task of the nucleus is to collect all incoming signals and transmit them to the axon. These collected signals are processed by the axon and sent to the synapses. Synapses also transmit newly produced signals to other nerve cells. The axon processes these signals and presents them as output. ANN is formed as a result of simulating real biological cells.

The ANN algorithm consists of 3 main layers, i.e. the input layer, hidden layer(s), and output layer. The number of hidden layers is chosen as 2 in many studies. Weighting is performed between the input layer and the hidden layers and all variables are associated with each other. The number of neurons in the input layer is equal to the number of inputs. Each neuron belongs to an input. The inputs are transmitted to the hidden layer without any processing. The hidden layer, on the other hand, processes the information it receives from the input layer and transmits it to the next layer. The number of neurons in each hidden layer varies. The number of neurons in this layer is higher than the number of neurons in the input and output layers. The weights form the intelligence of the ANN algorithm. The learning ability of the ANN algorithm is directly related to the weights. ANN can be calculated as in the following formulas.

Input layer; x is the input feature vector.

$$a^{(0)} = x \quad (2)$$

Hidden Layers; For each hidden layer $l$, compute the weighted sum of inputs:

$$z^{(l)} = W^{(l)}a^{(l-1)} + b^{(l)} \quad (3)$$

Finally an activation function (σ) is applied.

$$a^{(l)} = \sigma(z^{(l)}) \quad (4)$$

### 2.3. Support Vector Regression Algorithm

The Support Vector Machine (SVM) algorithm is a popular ML algorithm. Although the SVM algorithm was used for classification at first, it started to be used in regression problems over time and was widely used as an SVR algorithm. It can be said that the SVR algorithm is the regression of the SVM algorithm.

SVMs solve binary classification problems by formulating them as convex optimization problems (Vapnik, 1998). The SVM algorithm is often used in pattern recognition, object recognition and text recognition. SVM is generally an algorithm developed to separate two different classes. The SVM algorithm tries to find the optimal hyperplane to distinguish between the two classes. It is a method based on estimating the best-fit function for separating classes. Functions used to separate classes can be linear or non-linear. Linear solutions are easier and simpler than non-linear ones.

Although the SVM algorithm is generally used for

classification problems, it can also be used for regression problems; this is achieved by giving an alternative loss function (Gunn, 1998). In the SVR algorithm, a hyperplane function is used as in the classification problems. In addition, ($\mathcal{E}$) is the amount of deviation from the target (Figure 1).

$$y = f(x) = w.x + b \quad (5)$$

The error value is added to both sides on the hyperplane as follows.

$$y_i - wx_i - b \leq \varepsilon \; ise \quad (6)$$

$$-y_i + wx_i + b \leq \varepsilon \quad (7)$$

For the solution, it is necessary to minimize the weight w.

$$min_w = \frac{1}{2}\|w\|^2 \quad (8)$$

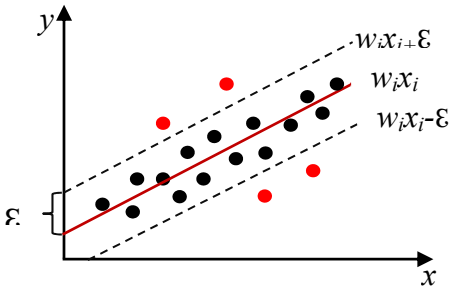


**Fig. 1.** Support Vector Regression and Error Parameter Representation. *Source*: own study.

### 2.4. K Nearest-Neighbors Algorithm

Although the k nearest neighbor algorithm, briefly k-nn algorithm, is generally used for classification problems, it is also frequently used in the literature for regression problems. The k-nn algorithm uses "feature similarity" to estimate the value of any new data point. For this new point with an unknown value, a value is assigned based on how close it is to the points in the training set.

The K-nn regression algorithm consists of 3 main steps in the stage; these are;

1. Calculation of the distance between the point to be estimated and each training point.
2. Selection of the K closest points by distance.
3. Estimation of the new point by averaging the selected data points.

One of the biggest advantages of the k-nn algorithm is that it operates with fewer parameters compared to other ML algorithms. In the k-nn algorithm, the k-value and a distance metric are sufficient. The biggest disadvantage is that it needs more memory and storage space.

Some frequently used distance calculation methods in the k-nn algorithm are as follows:

– *Euclidean distance*

It is the most widely used measure of distance. It is calculated as the square root of the sum of the squared differences between a point ($x$) and an existing point ($y$) using the following formula.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} \quad (9)$$

– *Manhattan distance*

The Manhattan distance is another popular measure of distance that measures the absolute value between two points. It is also called taxi distance, or city block distance as it is commonly visualized with a grid showing how a person can get from one address to another via a city street and is calculated using the following formula.

$$d(x,y) = \sum_{i=1}^{n}|y_i - x_i| \quad (10)$$

– *Minkowski distance*

The Minkowski distance measure is a generalization of the Euclidean and Manhattan distance metrics. The $p$ parameter in the formula below allows other distance measurements to be created. The Euclidean distance is represented by this formula when the p-value is equal to 2 and the Manhattan distance is shown as the $p$ value equal to 1.

$$d(x,y) = (\sum_{i=1}^{n}|y_i - x_i|^p)^{1/p} \quad (11)$$

### 2.5. Multiple Regression Analysis

Multiple regression analysis is a method used to model the relationship between two or more variables. Regression, which is also included in the statistical method group, is basically a classical machine learning model. Regression analysis can be divided into two types, simple and multiple regression. Multiple regression can be divided into three groups as linear, semi-logarithmic, and full logarithmic. Multiple regression is commonly preferred in real estate valuation studies. Studies have revealed that real estate value estimation is a non-linear problem in general (Čeh et al., 2018; Kontrimas & Verikas, 2011; Yu & Wu, 2016). In this study, 3 different regression estimations described below were tried. In the literature, generally linear and semi-logarithmic regression analysis were used for mass appraisal of real estate, and very limited studies were conducted with full logarithmic regression type. Therefore, another goal in this study is to determine which model

is more accurate for mass appraisal of real estate with regression models.

### 2.5.1. Linear Regression Analysis

Linear regression analysis is one of the basic regression analyses and commonly used in many research areas. The formula of the linear regression for mass appraisal valuation is shown below. In this formula, the value of n real estate (Value) is the dependent variable, using the weights of $\beta_0$ ,$\beta_1$,...,$\beta_n$ and X's show the independent variables, the parameters that affect the value of the real estate.

$$Value_i = \beta_0 + \beta_1 Area_1 + \beta_2 numberofrooms_2 + \cdots + b_n X_n + \varepsilon \quad (12)$$

### 2.5.2. Semi-logarithmic Regression Analysis

In nonlinear models, the models shown with the following formula, called semi-log, are used. In the equation below, a is the constant term, b is the slope coefficient and $\mathcal{E}$ is the error term. When estimating the real estate value, some researchers in the literature used linear models, while others tried non-linear models.

$$Log_y = a + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n + \varepsilon \quad (13)$$

### 2.5.3. Logarithmic Regression Analysis

Logarithmic regression is a type of regression used to model situations where growth or deterioration first accelerates and then slows down over time. It has been seen that it is used in the literature, although it is not very common in real estate mass appraisal; therefore, in this thesis, the data were tested and analyzed with full logarithmic regression.

$$Log_y = a + b_1 LogX_1 + b_2 LogX_2 + \cdots + b_n LogX_n + \varepsilon \quad (14)$$

## 2.6. Quality Analysis for Machine Learning Algorithms

Quality control analysis was performed to compare machine learning algorithms. For this comparison, the metrics shown with the formulas below are calculated for each of the ML algorithms. Thus, it was analyzed which of the methods was more successful.

- *Mean Square Error*

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (15)$$

In this equation, MSE is mean square error, N is the number of data tested, $y_i$ is the observed value, and $\hat{y}_i$ is the value estimated using the ML algorithms. In other words, MSE measures the mean square difference between known and predicted values. This formula measures the variance of the residuals.

- *Root Mean Square Error*

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (16)$$

In this equation, the RMSE is the Root Mean Squared Error, N is the number of data tested, $y_i$ is the observed value, and $\hat{y}_i$ is the value estimated using the ML algorithms. The root mean squared error is calculated as the square root of the mean squared error and serves to quantify the standard deviation of residuals.

- *Mean Absolute Error*

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (17)$$

In this equation, MSE is Mean Absolute Error, N is the number of data tested, $y_i$ is the observed value, and $\hat{y}_i$ is the value estimated using the ML algorithms, and measures the average of the residual values in the test data.

- *Coefficient of Determination ($R^2$):*

This is a widely known quality control method that measures the performance of regression models, also known as the coefficient of certainty. It is a value used to measure the performance of the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \quad (18)$$

In this formula, $y_i$ represents the observed values, $\hat{y}$ values calculated from the regression equation and, $\bar{y}$ shows the mean of the data.

- *Adjusted $R^2$*

$R^2$ is a statistical method that calculates the variance ratio for a dependent variable explained by the variables in a model. While the correlation checks the relationship between the independent and dependent variables, the number of variables is more prominent in the adjusted $R^2$ formula. Additionally, the quality of the model is tested with the corrected $R^2$.

$$R^2 = 1 - (1 - R^2) * \left[\frac{(n-1)}{n-k-1}\right] \quad (19)$$

In this formula, *n* represents the number of data and *k* represents the number of variables external to the dependent variable in the model. The coefficient of determination and the corrected $R^2$ are expected to be between 0 and 1. *An* $R^2$ value approaching 1 indicates the suitability and reliability of the model.

## 3. Study Area and Data

In this study, the study area was marked by the borders of Highway 20 (O-20) to cover the central districts of Ankara. The central districts of Ankara are: Altındağ, Çankaya, Etimesgut, Gölbaşı, Keçiören, Mamak, Sincan and Yenimahalle. The study area was created to cover the main centers of Gölbaşı, Sincan and Etimesgut districts and all the other districts. In Figure 2, the official administrative provincial border of Ankara Province and the boundaries of the study area are shown. As can be seen in the figure, the study area covers a large area, namely the center of Ankara. The study area is calculated at 699.03 km$^2$. Ankara's housing market is one of the rapidly rising provinces (Atasoy & Tanrıvermiş, 2024). In Ankara, 83,502 houses were sold in the year 2022, and many new houses were constructed (Tursun, 2023).

Real estate data of Endeksa for the last 5 years, which is a popular and confidential company of house sales in Turkey, were used. Residential type real estate values and their properties were obtained from the Endeksa company database for the province of Ankara. There is a total of 1,315,675 housing data. Data are shown in Figure 3 as point features. As can be seen in the figure, the data nearly the study area of Ankara province. Therefore, the data are very comprehensive, and their results will contribute to the literature.

There are 22 variables related to residential type real estate; these variables are shown in Table 1. Real estate variables can be numerical or categorical. Categorical data must have a numerical equivalent in order to be processed.

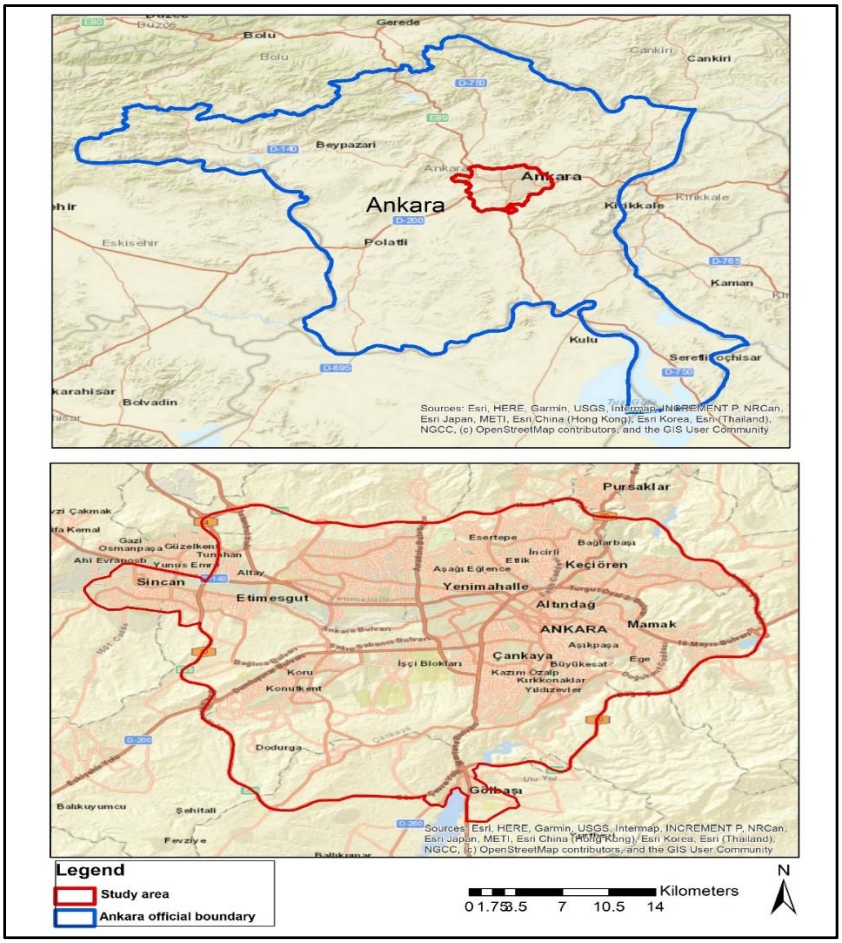


**Fig. 2.** Display of the study area with the official provincial border of Ankara. *Source*: own study.


eISSN: 2300-5289 | Received 2023-09-19 | Revised 2023-12-07 | Accepted 2024-02-16

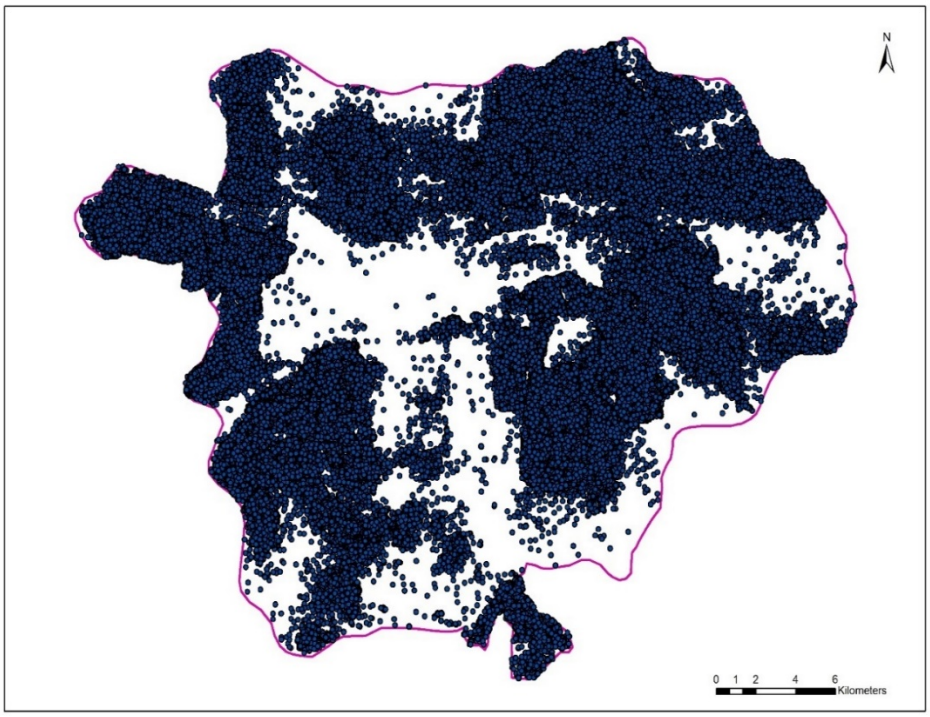


**Fig. 3.** Spatial representation of real estate data within the study area *Source*: own study.

**Table 1**

Variables of residential type real estate

| No | Variable | Unit | Variable Data Type |
|---|---|---|---|
| **1** | Residential area | $m^2$ | numeric |
| **2** | Number of living rooms | Piece | numeric |
| **3** | Number of bathrooms | Piece | numeric |
| **4** | Total number of rooms | Piece | numeric |
| **5** | Age | Number | numeric |
| **6** | Elevator | 0-1 (yes-no) | Categorical |
| **7** | Open parking lot | 0-1 (yes-no) | Categorical |
| **8** | Closed parking lot | 0-1 (yes-no) | Categorical |
| **9** | North facing status | 0-1 (yes-no) | Categorical |
| **10** | South facing status | 0-1 (yes-no) | Categorical |
| **11** | East facing status | 0-1 (yes-no) | Categorical |
| **12** | West facing status | 0-1 (yes-no) | Categorical |
| **13** | Floor number of the real estate | Number | numeric |
| **14** | Total number of floors | Number | numeric |
| **15** | The closest distance to parking areas | m. | numeric |
| **16** | The closest distance to university campuses | m. | numeric |
| **17** | The closest distance to places of worship | m. | numeric |
| **18** | The closest distance to schools | m. | numeric |
| **19** | The closest distance to shopping malls | m. | numeric |
| **20** | The closest distance to subway stops | m. | numeric |
| **21** | The closest distance to hospitals | m. | numeric |
| **22** | Heating Type | 10,20,30,40,50,60,70, 80, 90, 100, 110, 120, 130, 140 | Categorical |

*Source*: own study.

### 3.1. Parameter Selection in Machine Learning Algorithms

Two common parameter selection methods used in ML algorithms are the grid search and random search (Lerman, 1980). In this study, parameter estimation was carried out using the grid search method, which gives more optimum results (Chen et al., 2022). Random search is similar to grid search, but instead of using all points in the grid, it only tests a randomly

selected subset of those points. The smaller this subset, the faster the optimization, but also less accurate. A grid search is the process of scanning data to configure optimal parameters for a particular model. The grid search is applicable across machine learning to calculate the best parameters to use for any model, not just one type of model. The grid search generates a pattern on every possible combination of parameters. It iterates each parameter combination and creates a model for each combination, and determines the parameters of the most successful model.

### 3.2. Validation of the Data Sets

Cross-validation is a process that can be used to estimate the quality of machine learning (ML) algorithms. The results of cross-validation can be utilized to select the best parameter values set. K-fold cross-validation is a standard method for estimating the performance of ML algorithms on a dataset. The K-fold cross-validation procedure divides a limited dataset into K non-overlapping folds. While all other folds are collectively used as a training dataset, each of the K folds is given the opportunity to be utilized as a validation set. In this study, 5-fold cross-validation was employed, and a schematic of this method is illustrated in Figures 4 and 5.

In this study, data were tested and analyzed using 5-fold cross-validation. To explain more explicitly, 80% of the data was used for training, and 20% for testing. The 20% testing and 80% training portions were iteratively swapped five times to test the performance of Machine Learning (ML) algorithms. Error metrics were calculated and averaged in each iteration.

$$Error = \frac{1}{5}\sum_{i=1}^{5} Error_i \quad (20)$$

## 4. Results

Out of 1,315,675 housing data in total, 80% was used as training data and 20% as test data. In other words, for all ML algorithms, 1,052,540 of 1,315,675 residential properties were used to test the ML algorithms and create the model, while 263,135 were used for testing. As mentioned before, the grid search method was used for parameter estimation in ML algorithms, and the parameters obtained as a result of this method are shown in the table below. For residential real estate, the optimum parameters tried with the grid search and obtained as a result of the grid search are given for each ML algorithm (Table 2).

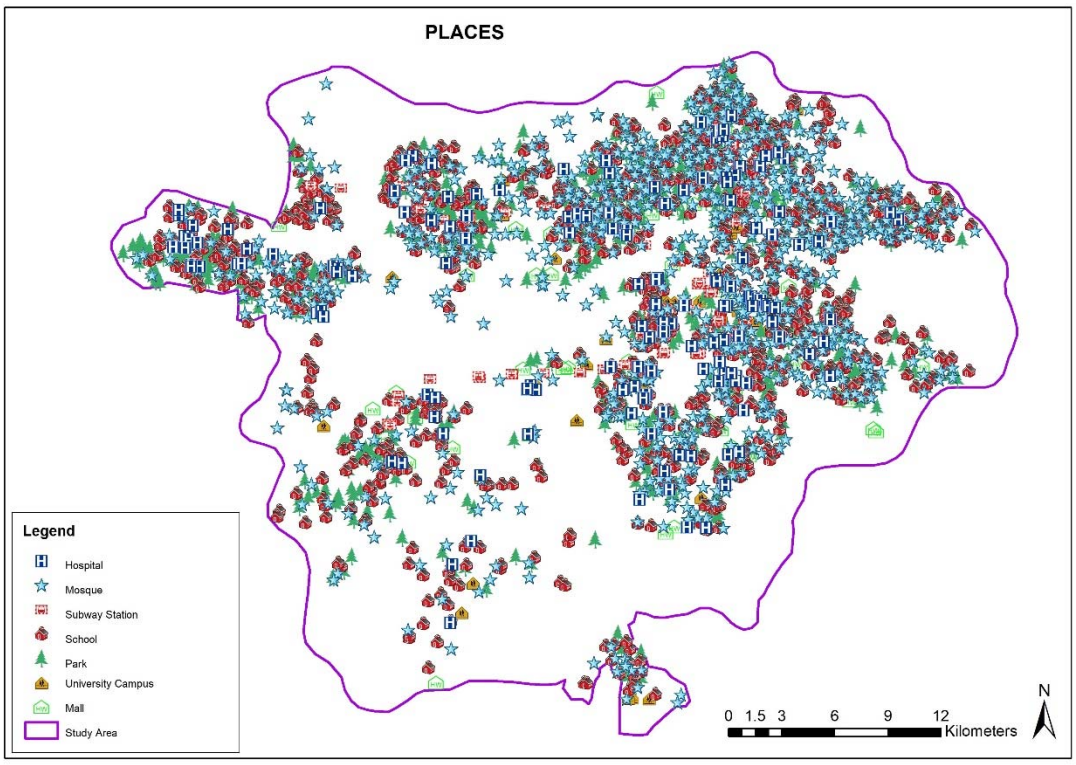


**Fig. 4.** Display of important places in the study area. *Source*: own study.


eISSN: 2300-5289 | Received 2023-09-19 | Revised 2023-12-07 | Accepted 2024-02-16

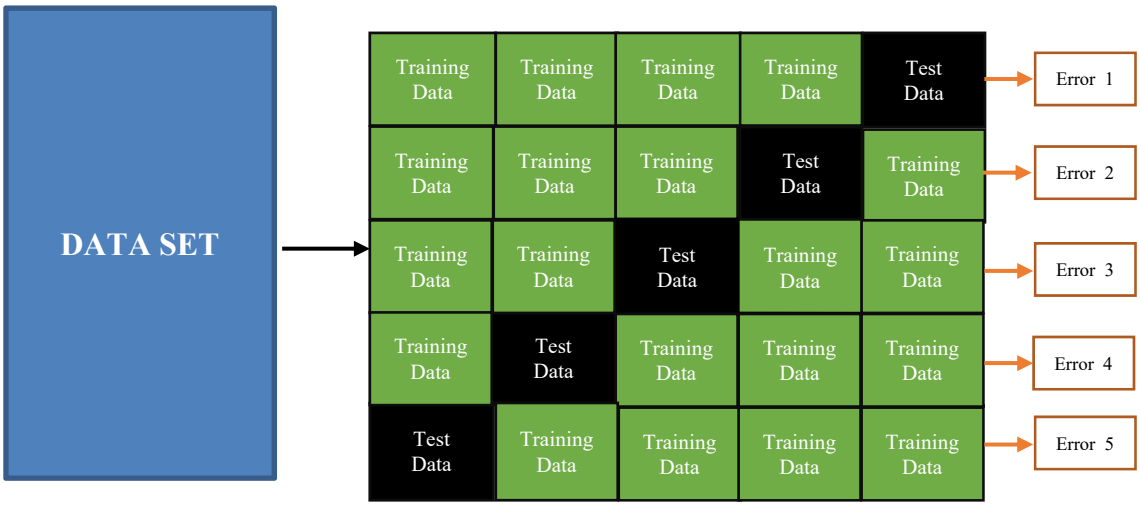


**Fig. 5.** The distribution of data as training and test sets in algorithms and cross-validation. *Source*: own study.

**Table 2**

ML algorithms hyper parameter selection for real estate data

| *Algorithm* | *Parameters for Grid Search* | *Parameter Used* |
|---|---|---|
| ***ANN*** | Number of neurons in hidden layers: 44,66,88 | Number of neurons in hidden layer: 66<br>Number of hidden layers: 2 |
| | Activation function: Identity, logistics, tanh, relu | Activation function: 66 |
| | Alpha: 0.001,0.0001,0.00001 | Alpha: 0.001 |
| | Learning rate: 0.01,0.001,0.0001 | Learning rate 0.01 |
| ***K-nn*** | N- neighborhood: 4,8,12,16,20,24,28,32 Distance: euclidean, manhattan | N-neighborhood: 8<br>Distance: euclidean |
| ***SVR*** | Kernel: linear, poly, rbf, sigmoid<br>Gamma: scale, auto | Kernel: linear<br>Gamma: scale |
| ***RF*** | N_estimators (number of trees): 50,100,150,200,250 | N_estimators: 200 |

*Source*: own study.

**Table 3**

Comparison of ML algorithms for mass appraisal of residential real estate

| | $R^2$ | Adjusted $R^2$ | MSE | RMSE | MAE |
|---|---|---|---|---|---|
| **RF** | 0.8126 | 0.8125 | 1,409,005 | 1187 | 763 |
| **ANN** | 0.7617 | 0.7615 | 1,791,277 | 1338 | 922 |
| **k-nn** | 0.6544 | 0.6543 | 2,598,203 | 1611 | 1122 |
| **Linear Regression** | 0.5256 | 0.5252 | 3,566,780 | 1888 | 1326 |
| **Semi-log Regression** | 0.4933 | 0.4932 | 3,785,642 | 1945 | 1321 |
| **SVR** | 0.2997 | 0.2997 | 5,231,685 | 2287 | 1532 |
| **Log-log Regression** | 0.2734 | 0.2733 | 5,863,194 | 2421 | 1620 |

*Source*: own study.

Residential real estate mass appraisal application was carried out using 1,315,675 houses and 22 variables belonging to these real estates data. For this purpose, 5 different ML algorithms were tested. Separate quality analyses were performed for these algorithms, and the comparison of quality metrics has been shown in Table 3. When the $R^2$ metric shown in this table and the other metrices is examined, it is seen that the most successful ML algorithm for residential type real estate is the RF algorithm, and the ANN algorithm is the second most successful algorithm, with the obtained model proving to be reliable. On the other hand, it has been determined that k-nn and SVR algorithms do not give reliable results for this data set and variables. The SVR algorithm, on the other hand, was the algorithm with the longest computation time in model estimation. Other metrics such as MSE, RMSE and MAE are very similar to each other, and these metrics also show the quality of the algorithms. MAE is an especially useful quality metric for commenting on the results. According to MAE, there is only a 763 Turkish liras average difference between the calculated value and the observed values.

### 4.1. Real Estate Value Map

At the last stage of this study, a value map with real estate values and housing type was created within the study area of the boundaries of the Ankara province. In raster-based maps, since all pixels will take a value, a value map can be obtained by using the nearby real estate values by using the interpolation method. Interpolation methods are very diverse and, in this study, unknown values were calculated using the known values with the Inverse Distance Weighting (IDW) method and a raster-based value map was created for the study area. In this way, a residential real estate value map was created for the province of Ankara, because there is sufficient and homogeneous data on the value of only the housing.

In Figure 6, the raster-formatted value map was created and shown as an example, using the IDW interpolation method for the data used in the study. On this map, the red values show the highest TL/m$^2$ value. It can be easily seen from the map that the housing values in Mustafa Kemal, Emek, and Beştepe locations are very high.

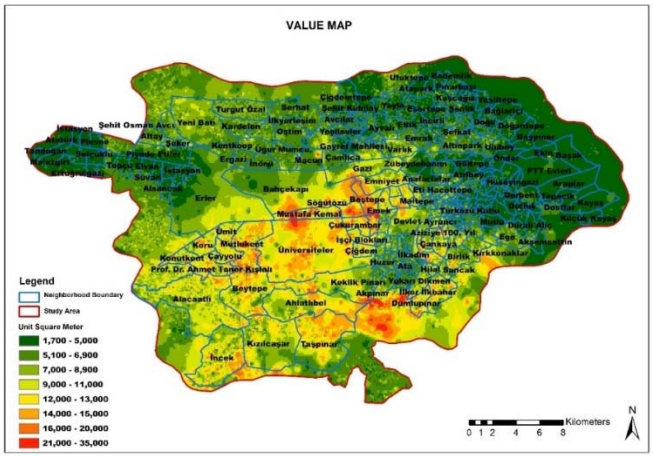


**Fig. 6.** Residential real estate value map for the study area. *Source*: own study.

### 5. Conclusions

Over recent years, traditional methods in various fields, including real estate, have given way to more advanced techniques, such as machine learning (ML). The advent of big data, facilitated by technological advancements and increased data storage capabilities, has highlighted the inefficiencies of processing data with traditional methods, leading to time and labor losses. In the real estate sector, the surge in property values and the proliferation of quantitative and qualitative variables influencing these values have generated significant amounts of big data. In the context of Turkey, there is a notable absence of a dedicated institution responsible for the mass appraisal of real estate. Recognizing this gap, the present study is of paramount importance to the country. The research focuses on leveraging ML algorithms to determine the value of real estate, specifically conducting an analysis on mass appraisal of real estate data in the capital of Turkey, employing a substantial dataset. Five ML algorithms were employed, and the findings revealed the RF algorithm's effectiveness, particularly in assessing residential real estate in Turkey ($R^2$=0.81). To achieve success in mass appraisal for Turkish real estate, the study identified the RF model with 200 trees as a successful tree-based model. The Artificial Neural Network (ANN) algorithm also emerged as an optimal approach, with the identified optimal parameters being 2 hidden layers comprising 66 neurons and a rectified linear unit (relu) activation function. Additionally, the study indicated that a linear regression model is more accurate for real estate appraisal in Turkey. Among the 22 real estate parameters analyzed, the study highlighted heating type and distances to important places as the most influential factors affecting real estate value. These parameters, as discerned by the RF algorithm, align with what prospective buyers prioritize when purchasing a house in Turkey. In summary, the study recommends the use of RF and ANN algorithms for the mass appraisal of Turkish real estate data across all cities. This approach aims to expedite the creation of a comprehensive value map for real estate, applicable in various areas within the real estate sector. Future studies could further refine the application of RF and ANN algorithms for Turkish real estate mass appraisal, exploring the potential integration of additional variables and fine-tuning model parameters to enhance accuracy. Additionally, an investigation into the scalability and adaptability of these algorithms across different regional nuances within Turkey's diverse real estate markets would contribute to a more detailed understanding of their effectiveness on a broader scale.

### References

Atasoy, T., & Tanrıvermiş, H. (2024). Gayrimenkul Türevleri, Gayrimenkul Türev Fiyatlandırma Modelleri Ve Türkiye'de Bir Uygulama (Real Estate Derivatives, Real Estate Derivative Pricing Models and an Application in Turkey). *The Journal of Academic Social Science Studies, 16* (98), 461-494.

Bilgilioğlu, S. S., & Yılmaz, H. M. (2023). Comparison of different machine learning models for mass appraisal of real estate.

*Survey Review, 55*(388), 32–43. https://doi.org/10.1080/00396265.2021.1996799

Borst, R. A. (1991). Artificial neural networks: The next modelling/calibration technology for the assessment community. *Property Tax Journal, 10*(1), 69–94.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., & Cutler, A. (2005). *Random Forests*. Berkeley. In.

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *ISPRS International Journal of Geo-Information, 7*(5), 168. https://doi.org/10.3390/ijgi7050168

Chen, H., Zhang, Z., Yin, W., Zhao, C., Wang, F., & Li, Y. (2022). A study on depth classification of defects by machine learning based on hyper-parameter search. *Measurement, 189*, 110660. https://doi.org/10.1016/j.measurement.2021.110660

Dambon, J. A., Fahrländer, S. S., Karlen, S., Lehner, M., Schlesinger, J., Sigrist, F., & Zimmermann, A. (2022). Examining the vintage effect in hedonic pricing using spatially varying coefficients models: A case study of single-family houses in the Canton of Zurich. *Swiss Journal of Economics and Statistics, 158*(1), 2. https://doi.org/10.1186/s41937-021-00080-2

Dellstad, M. (2018). Comparing three machine learning algorithms in the task of appraising commercial real estate. Degree project in computer science and engineering. Stockholm, Sweden.

Gnat, S. (2021). Property mass valuation on small markets. *Land (Basel), 10*(4), 388. https://doi.org/10.3390/land10040388

Gültekin, A., Dikmen, Ç., Erciyes, A., & Örgü, D. (2017). An examination on evolution of sustainability in the context of United Nations Sustainable Development Goals: Turkey case.

Hong, J., Choi, H., & Kim, W. (2020). 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. https://doi.org/10.3846/ijspm.2020.11544

Hong, J., & Kim, W. (2022). Combination of machine learning-based automatic valuation models for residential properties in South Korea. *International Journal of Strategic Property Management, 26*(5), 362–384.

Iban, M. C. (2022). An explainable model for the mass appraisal of residences: The application of tree-based Machine Learning algorithms and interpretation of value determinants. *Habitat International, 128*, 102660. https://doi.org/10.1016/j.habitatint.2022.102660

Kontrimas, V., & Verikas, A. (2007). Neural networks based screening of real estate transactions. *Neural Network World, 17*(1), 17.

Kontrimas, V., & Verikas, A. (2011). The mass appraisal of the real estate by computational intelligence. *Applied Soft Computing, 11*(1), 443–448. https://doi.org/10.1016/j.asoc.2009.12.003

Lam, K. C., Yu, C., & Lam, K. (2008). An artificial neural network and entropy model for residential property price forecasting in Hong Kong. *Journal of Property Research, 25*(4), 321–342. https://doi.org/10.1080/09599910902837051

Lenk, M. M., Worzala, E. M., & Silva, A. (1997). High-tech valuation: Should artificial neural networks bypass the human valuer? *Journal of Property Valuation and Investment, 15*, 8-26.

Lerman, P. (1980). Fitting segmented regression models by grid search. *Journal of the Royal Statistical Society. Series C, Applied Statistics, 29*(1), 77–84.

McCluskey, W. (1996). Predictive accuracy of machine learning models for the mass appraisal of residential property. *New Zealand Valuers Journal*, *16*(1), 41-47.

Morano, P., & Tajani, F. (2013). Bare ownership evaluation. Hedonic price model vs. artificial neural network. *International Journal of Business Intelligence and Data Mining, 8*(4), 340–362. https://doi.org/10.1504/IJBIDM.2013.059263

Musa, A. G., Daramola, O., Owoloko, A., & Olugbara, O. (2013). A neural-CBR system for real property valuation. *Journal of Emerging Trends in Computing and Information Sciences, 4*(8), 611–622.

Özkan, G., Yalpır, Ş., & Uygunol, O. (2007). *An investigation on the price estimation of residable real-estates by using artificial neural network and regression methods*. XIIth Applied Stochastic Models and Data Analysis International conference (ASMDA), Chania, Crete, Greece,

Ravikumar, A. S. (2017). *Real estate price prediction using machine learning*. Dublin, National College of Ireland.

Sampathkumar, V., Santhi, M. H., & Vanjinathan, J. (2015). Forecasting the land price using statistical and neural network software. *Procedia Computer Science, 57*, 112–121. https://doi.org/10.1016/j.procs.2015.07.377

Saraç, E. (2012). *Yapay sinir ağları metodu ile gayrimenkul değerleme (Real estate valuation with artificial neural networks method)* İstanbul Kültür Üniversitesi/Fen Bilimleri Enstitüsü/İnşaat Mühendisliği (Istanbul Kültür University/Institute of Natural and Applied Sciences/Civil Engineering)].

Sawant, R., Jangid, Y., Tiwari, T., Jain, S., & Gupta, A. (2018). *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),

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications, 36*(2), 2843–2852.

Sisman, S., Akar, A. U., & Yalpir, S. (2023). The novelty hybrid model development proposal for mass appraisal of real estates in sustainable land management. *Survey Review, 55*(388), 1–20. https://doi.org/10.1080/00396265.2021.1996797

Tabales, J. M. N., Caridad, J. M., & Carmona, F. J. R. (2013). Artificial neural networks for predicting real estate price. *Revista de Métodos Cuantitativos para la Economía y la Empresa, 15*, 29–44.

Tay, D.P.H. & Ho, D.K.H. (1992). Artificial intelligence and the mass appraisal of residential apartments. Journal of Property Valuation and Investment, 10(2), 525-540. https://doi.org/10.1108/14635789210031181

Torres-Pruñonosa, J., García-Estévez, P., & Prado-Román, C. (2021). Artificial neural network, quantile and semi-log regression modelling of mass appraisal in housing. *Mathematics, 9*(7), 783. https://doi.org/10.3390/math9070783

Tursun, A. (2023). Gayrimenkul Pazar Analizinde Sistem Dinamiği Yaklaşımı ve Uygulaması (System Dynamics Approach and Application in Real Estate Market Analysis). Nobel.

Unel, F. B., & Yalpir, S. (2023). Sustainable tax system design for use of mass real estate appraisal in land management. *Land Use Policy, 131*, 106734. https://doi.org/10.1016/j.landusepol.2023.106734

Valier, A. (2020). Who performs better? AVMs vs hedonic models. *Journal of Property Investment & Finance, 38*(3), 213–225. https://doi.org/10.1108/JPIF-12-2019-0157

Vapnik, V. (1998). *Statistical learning theory New York*. NY, Wiley.

Varma, A., Sarma, A., Doshi, S., & Nair, R. (2018). *House price prediction using machine learning and neural networks*. 2018

second international conference on inventive communication and computational technologies (ICICCT).

Wilson, I. D., Paris, S. D., Ware, J. A., & Jenkins, D. H. (2002). Residential property price time series forecasting with neural networks. In *Applications and Innovations in Intelligent Systems IX* (pp. 17-28). Springer.

Worzala, E., Lenk, M., & Silva, A. (1995). An exploration of neural networks and its application to real estate valuation. *Journal of Real Estate Research, 10*(2), 185–201. https://doi.org/10.1080/10835547.1995.12090782

Xin, J. G., & Runeson, G. (2004). Modeling property prices using neural network model for Hong Kong. *International Real Estate Review, 7*(1), 121–138.

Yılmaz, M., & Bostancı, B. (2023). Investigation of Real Estate Tax Leakage Loss Rates with ANNs. *Buildings, 13*(10), 2464. https://doi.org/10.3390/buildings13102464

Yu, H., & Wu, J. (2016). *Real estate price prediction with regression and classification.* CS229 (Machine Learning). Final Project Reports.