AESTIMUM
CENTRO STUDI DI ESTIMO E DI ECONOMIA TERRITORIALE - Ce.S.E.T.

Firenze University Press
www.fupress.com/ceset

OPEN ACCESS

Citation: Yağmur A., Kayakuş M., Terzioğlu M. (2022) House price prediction modeling using machine learning techniques: a comparative study. *Aestimum* 81: 39-51. doi: 10.36253/aestim-13703

Received: September 3, 2022

Accepted: December 4, 2022

Published: March 10, 2023

Data Availability Statement: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

ORCID:
AY: 0000-0003-2138-240X
MK: 0000-0003-0394-5862
MT: 0000-0002-4614-7185

House price prediction modeling using machine learning techniques: a comparative study

Ayten Yağmur[1], Mehmet Kayakuş[2,*], Mustafa Terzioğlu[3]

[1] *Department of Labour Economics and Industrial Relations, Akdeniz University, Turkey*
[2] *Department of Management Information Systems, Akdeniz University, Turkey*
[3] *Accounting and Tax Department, Akdeniz University, Turkey*
E-mail: mehmetkayakus@akdeniz.edu.tr, aytenyagmur@akdeniz.edu.tr, mterzioglu@akdeniz.edu.tr
*Corresponding author

Abstract. In the literature, there are two basic approaches regarding the determination of house prices. One of them is the prediction of house price using macroeconomic variables in the country where the house is produced, and another one is the price prediction models, which we can express as micro-variables, by considering the features of the house. In this study, the price of the house was attempted to be predicted using machine learning methods by establishing a model with micro variables that reveal the features of the house. The study was conducted in Turkey' Antalya province, where household housing demand of foreigners is also high. The house advertisements in locations belonging to the lower, middle- and upper-income groups were selected as the sample. In the results, it was observed that the artificial neural network (ANN) method made predictions with more meaningful results compared to support vector regression (SVR) and multiple linear regression (MLR). These results appear to be a viable model for institutions that supply housing, mediate housing sales, and provide housing financing and valuation. It is considered that this model, which can be used to predict fluctuating house prices, especially in developing countries, will regulate the housing market.

Keywords: Home price, Prediction, Support Vector Regression, Artificial Neural Networks, Multiple Linear Regression.
JEL codes: O18, R33.

1. INTRODUCTION

Human needs are endless, however, some basic needs such as nutrition, shelter and protection should be first met for the continuation of their lives. Housing need is a multidimensional problem that is necessary for people's shelter, health, security and various socio-cultural needs. People want to buy a house in order to have their own house only when their welfare reaches a certain level. At this stage, the important thing is to choose a house that will meet the budget they have and the needs of their family members. In this respect, affordable housing prices are very important for households. Hous-

ing suppliers prioritize the needs of these households while designing the house they will produce architecturally. It is very important that the house to be produced meets these needs appropriately and that it is built in the right location in terms of the costs to be incurred. It is of vital importance for these institutions to determine the housing price correctly, to meet these costs, to sell the produced houses easily and to achieve a desired amount of profit margin. Because the institutions that supply housing make huge capital investments and the wrong construction projects that cannot be sold cause these institutions to go bankrupt very quickly. Banks, mortgage and real estate companies that provide housing financing allocate loans to households that demand housing based on the housing price and the appraisal valuation they will make. Therefore, the creation of an effective and effective credit policy by these financial institutions directly depends on the accurate price prediction of the house. Since the maturities of these loans will be medium and long term, incorrect loan allocation will reduce their assets and reduce their direct return on assets (ROA), because of these companies making an inefficient use of assets on their balance sheets in the long run. Thus, the main deciding factor for the three important actors in the housing market is the sale price of the house.

The factors that determine the sale price of the house are primarily the basic features of the house. The first of them is the location of the house. In general, it is seen that houses are built according to the lower, middle- and upper-income groups and their needs depending on the features of the location (by the sea, in the forest, distance from the city center, school, hospital, religious places, and proximity to organized industrial zones, which are production zones, etc.). Another factor is the volume and situation of the house. The usable size and the number of rooms of the house are a direct factor for the selling price due to both the demographic characteristics of the demanding households and the cost of the housing to be built. Furthermore, the fact that the house sold is a new or secondary house directly affects the firm sale price of the house in high-type houses. Moreover, whether the house is designed as a complex of buildings (security, pool, Spa, gym, etc.) is a determining factor on the sales price of the house.

These variables, which we describe as micro-variables, were made into a model in the study. With this model, it was attempted to predict the house prices using machine learning methods, which are among the advanced prediction techniques. It is considered that the obtained results will contribute to correct pricing in terms of housing suppliers, mediators in house sales and institutions that provide financing. The results of the model created in the study are also important in terms of an effective and active housing market. Especially in housing markets where price fluctuations are high and there are housing supply and demand imbalances, the use of advance price prediction mechanisms will ensure the proper operating of the markets.

The use of three different machine techniques in the study and especially the testing of the support vector regression technique in this regard differs from similar studies in the literature. The aim of this study is to create a model that can accurately predict the housing prices in the locations in the portfolios of the institutions that offer housing and mediate its sale. Testing the success of the designed model using machine learning methods is the second main objective of the study. At the same time, it is aimed to be an exemplary reference study for more appropriate housing production planning by considering the preferences of those who supply housing and those who demand it. To achieve these goals, the main hypothesis of the study is that the variables that reveal the characteristics of the house in the estimation of housing prices will be successfully predicted using machine learning methods.

In the second section of the study, reference was made to the studies on the basic dynamics affecting housing prices. In addition, studies using machine learning and other methods for housing prices are included. Section 3 describes the model of the study and the machine learning methods used by focusing on the data set of this model. In section 4, the results obtained by machine learning are included in the study and these findings are discussed. Section 5 presents the conclusions drawn from the study and the policies and recommendations drawn from these conclusions.

2. LITERATURE

In this section, first of all, the basic economic structure affecting housing prices is emphasized.

Houses meet the shelter needs of people and are also an investment tool. The housing market differs from other markets in that housing is both a consumption and an investment good. Housing markets differ from other markets in that the housing supply is very costly, the housing is permanent and continuous, heterogeneous, fixed, causes growth in the secondary markets, and is used as a guarantee (Iacoviello, 2000).

The housing market is formed through a mechanism of housing supply and demand. In the housing market, unlike the goods and services market, the housing sup-

ply is inelastic. Supply and demand for housing change and develop over time depending on the economic, social, cultural, geographical, and demographic realities of the countries. Meeting the housing demand is associated with housing policies and economic conditions. Housing demand arises for different purposes such as consumption, investment, and wealth accumulation. The supply and demand factors change according to the type of housing demand. In addition to the input costs of the house as a product, the determination of the price of the house is affected by many variables such as people's income level, marital status, industrialization of the society and agricultural employment rate, interest rates, population growth and migration, and all variables also affect the price. Since changes in housing prices affect both socio-economic conditions and national economic conditions, it is an important issue that concerns governments and individuals (Kim and Park, 2005). Housing demand arises for different purposes such as consumption, investment, and wealth accumulation.

In this part of the literature, some studies that estimate housing prices are cited.

The prediction of houses with real factors is important for the studies. With the developments in artificial intelligence methods, it now allows the solution of many problems in daily life such as purchasing a house. The competitive nature of the housing sector helps the data mining process in this industry, processing this data and predicting its future trends. Regression is a machine learning tool that encourages to build expectations from available measurable information by taking the links between the target parameter and many different independent parameters. The cost of a house is based on several parameters. Machine learning is one of the most important areas to apply ideas on how to increase costs and predict with high accuracy.

Machine learning method is one of the recent methods used for prediction. It is used to interpret and analyze highly complex data structures and patterns (Ngiam and Khor, 2019). Machine learning predicts that computers learn and behave like humans (Feggella, 2019). Machine learning means providing valid dataset, and moreover predictions are based on it, machine learns how important a particular event might be on the whole system based on pre-loaded data and predicts the outcome accordingly. Various modern applications of this technique include predicting stock prices, predicting the probability of an earthquake, and predicting company sales, and the list has infinite possibilities (Shiller, 2007).

Unlike traditional econometrics models, machine learning algorithms do not require the training data to be normally distributed. Many statistical tests rely on the assumption of normality. If the data are not normally distributed, these statistical tests will fail and become invalid. These processes used to take a long time, however, today they can be completed quickly with the high-speed computing power of modern computers and therefore this technique is less costly and less timely to use.

Rafiei and Adeli (2016) used SVR to determine whether a property developer should build a new development or stop the construction at the beginning of a project based on the prediction of future house prices. The study, in which data from 350 apartment houses built in Tehran (Iran) between 1993 and 2008 were used, had 26 features such as zip code, gross floor area, land area, estimated cost of construction, construction time, and property prices. Its results revealed that SVR was a suitable method for making home price predictions since the loss of prediction (error) was as low as 3.6% of the test data. Therefore, the prediction results provide valuable input to the property developer's decision-making process.

Cechin et al. (2000) analyzed the data of buildings for sale and rental in Porto Alegre, Brazil, using linear regression and artificial neural network methods. They used parameters such as the size of the house, district, geographical location, environmental arrangement, number of rooms, building construction date and total area of use. According to the study, they reported that the artificial neural network method was more useful compared to linear regression.

Yu and Wu (2016) used the classification and regression algorithms. According to the analysis, living area square meter, roof content and neighborhood have the greatest statistical significance in predicting the selling price of a house, and the prediction analysis can be improved by the Principal Component Analysis (PCA) technique. Because the value of a particular property is closely associated with the infrastructure facilities surrounding the property.

Koktashev et al. (2019) attempted to predict the house values in the city of Krasnoyarsk by using 1,970 housing transaction records. The number of rooms, total area, floor, parking lot, type of repair, number of balconies, type of bathroom, number of elevators, garbage disposal, year of construction and accident rate of the house were discussed as the features in that study. They applied random forest, ridge regression, and linear regression to predict the property prices. Their study concluded that the random forest outperformed the other two algorithms, as evaluated by the Mean Absolute Error (MAE).

Park and Bae (2015) developed a house price prediction model with machine learning algorithms in

real estate research and compared their performance in terms of classification accuracy. Their study aimed at helping real estate sellers or real estate agents to make rational decisions in real estate transactions. The tests showed that the accuracy-based Repeated Incremental Pruning to Produce Error Reduction (RIPPER) consistently outperformed other models in house price prediction performance.

Bhagat et al. (2016) studied on linear regression algorithms for house prediction. The aim of the study was to predict the effective price of the real estate for clients based on their budget and priorities. They indicated that the linear regression technique of the analysis of past market trends and price ranges could be used to determine future house prices.

In their study, Mora-Esperanza and Gallego (2004) analyzed house prices in Madrid using 12 parameters. The parameters they used were the distance to the city center, road, size of the district, construction class, age of the building, renovation status, housing area, terrace area, location within the district, housing design, the floor and the presence of outbuildings. The dataset was created assuming that the sales values of 100 houses for sale in the region were the real values. Researchers, who used the ANN and linear regression analysis technically, reported that the ANN technique was more successful and achieved an average agreement of 95% and an accuracy of 86%.

Wang and Wu (2018) used 27,649 data on home appraisal price from Airlington County, Virginia, USA in 2015 and suggested that Random Forest outperformed the linear regression in terms of accuracy.

In their study in the case of Mumbai, India, Varma et al. (2018) attempted to predict the price of the house by using various regression techniques (Linear Regression, Forest regression, boosted regression) and artificial neural network technique based on the features of the house (usage area, number of rooms, number of bathrooms, parking lot, elevator, furniture). In conclusion, they determined that the efficiency of the algorithm with the use of artificial neural networks was higher compared to other regression techniques. They also revealed that the system prevented the risk of investing in the wrong house by providing the right output.

Thamarai and Malarvizhi (2020) attempted to predict the prices of houses from real-time data after the large fluctuation in house price increases in 2018 at the Tadepalligudem location of West Godavari District in Andhrapradesh, India using the features of the number of bedrooms, age of the house, transportation facilities, nearby schools, and shopping opportunities. They applied these models in decision tree regression and multiple linear regression techniques, which are among the machine learning techniques. They suggested that the performance of multiple linear regression was better than decision tree regression in predicting the house prices.

As examined in the literature, the general characteristics of the housing are often used as a model in the estimation of housing prices. Therefore, in the study, a model was created over the variables that contain the characteristics of the housing rather than the general economic conditions. However, unlike other studies, three different machine learning methods were used to compare the success of these methods against each other. While creating the model, different economic, location and social cultural neighbourhoods are selected and the work from other studies is made original.

3. MATERIAL AND METHOD

3.1 Case study data set

In the study, the data of the three biggest and most advertised neighborhoods of three largest districts in Antalya province of Turkey were selected. These districts are the locations of houses that appeal to different income groups and have different features. In particular, these neighborhoods with a heterogeneous demographic and economic structure were selected to test the machine learning techniques to be used in the model created. Kepez, Erenköy and Ahatlı districts of Kepez district, Çağlayan, Fener and Meydankavağı districts of Muratpaşa district, and Gürsu, Hurma and Uncalı districts of Konyaaltı district were selected as locations. According to economic characteristics, Kepez district is low income, while Muratpaşa and Konyaaltı districts are preferred by middle- and high-income groups. Since Antalya is a tourism city, culturally these districts have a heterogeneous structure and receive migration from both different provinces and countries. Especially the citizens of Ukraine, Russia, Arab Countries, and Iran prefer Antalya, which is a tourism city, due to the economic and political reasons experienced in the world and in Turkey. This creates a very cultural diversity in all three districts. Konyaaltı and Muratpaşa districts have a seacoast. Kepez district has no coast to the sea. In addition, Konyaaltı and Muratpaşa districts have more alternatives than Kepez district in terms of art, sports, and recreation areas. The locations of the districts and their neighbourhoods subject to the study are presented in Figure 1.

The numbered districts and neighbourhoods' information in Figure 1 is shown in Table 2.

The data of a total of 900 house for sale advertisements in these locations were obtained from the website

Table 1. Literature synthesis table.

References	Summary of Findings
Cechin, A., Antonio, S. & Gonzales, M. A. (2000)	They used parameters such as the size of the house, district, geographical location, environmental arrangement, number of rooms, building construction date and total area of use. According to the study, they reported that the artificial neural network method was more useful compared to linear regression.
Yu, H., & Wu., J. (2016)	They used classification and regression algorithms. According to the analysis, in the estimation of the sale price of a house, the living space square meter, the roof content and the neighbourhood are of the greatest statistical importance.
Koktashev, V., Makee, V., Shchepin, E., Peresunko, P., & Tynchenko, V. V. (2019).	They tried to estimate the housing values in his city. In this study, the number of rooms, total area, floor, parking, repair type, number of balconies, bathroom type, number of elevators, garbage disposal, construction year and accident rate of the house were discussed. They applied random forest, ridge regression, and linear regression to predict property prices. Their study concluded that the random forest performed better than the other two algorithms, as evaluated by mean absolute error (MAE).
Park, B. H., & Bae, J. K. (2015).	In his real estate research, he developed a residential price prediction model with machine learning algorithms and compared their performance in terms of classification accuracy. His work aims to help real estate sellers or real estate agents make rational decisions in real estate transactions. Experiments show that the RIPPER algorithm based on accuracy consistently outperforms other models in housing price prediction performance
Bhagat, N., Mohokar, A., & Mane, S. (2016)	They worked on linear regression algorithms for the prediction of homes. The purpose of the article is to estimate the effective price of real estate for clients according to their budgets and priorities. Analysis of past market trends and price ranges, predicted future home prices.
Mora-Esperanza, J. G., & Gallego, J. (2004)	In their study, they analysed housing prices in Madrid using 12 parameters. The data set is the actual value of the sale values of 100 houses for sale in the region. The results were more successful than the regression analysis, with an average compliance rate of 95% and an accuracy rate of 86%.
Wang, C. C., & Wu, H. (2018).	In 2015, they used 27,649 home valuation price data from Airlington County, Virginia, and suggest that Random Forest outperforms linear regression in terms of accuracy.
Rafiei, M. H., & Adeli, H. (2016).	He used SVR to determine whether a property developer should build a new development or stop construction at the start of a project based on a forecast of future home prices. Using data from 350 apartments built in Tehran (Iran) between 1993 and 2008, the research trained a model with 26 characteristics such as zip code, gross floor area, land area, estimated construction cost, construction time, real estate prices, etc. Nearby housing developments, exchange rate and demographic factors. Their results showed that SVR is a viable method for making house price predictions, as the loss of prediction (error) is as low as 3.6% of the test data. Forecast results, therefore, provide valuable input into the property developer's decision-making process.

Figure 1. Locations of districts and neighbourhoods.

sahibinden.com, Turkey's largest advertisement site, and analyzed. As data, housing sales announcements in the period of August 2022 were examined. The variables of location, usable area, number of rooms, age of residence, floor and social facilities, and the presence in complex buildings, which are the main features of the house, were used as the input unit in the model. The advertised sales price of the house was used as the output unit. A total of 5,400 data entries were made for these input and output units. The variables used in the study and the characteristics of the houses in the input variables obtained from the housing sales announcements and the number of data are shown in Table 3.

Table 2. Numbered districts and neighbourhood information in Figure 4.

1- Kepez- Kepez	6- Konyaaltı- Gürsu
2- Kepez- Erenköy	7- Muratpaşa- Çağlayan
3- Kepez- Ahatlı	8- Muratpaşa- Fener
4- Konyaaltı- Uncalı	9-Muratpaşa- Meydankavağı
5- Konyaaltı- Hurma	

The data set of the studies was manually entered into the Excel program from the advertisements on the

Table 3. Features of housing.

Output Variable	Data Entry Range
House Price	0-2,000,000 Turkish Liras: 297 data 2,000,001-4,000,000 Turkish Liras: 348 data +4,000,000 Turkish Liras: 255 data
Input Variables	**Data Entry Range**
Location	Kepez-Kepez neighbourhood 100 data Kepez-Erenköy neighbourhood 100 data Kepez-Ahatlı neighbourhood 100 data Konyaaltı-Uncalı neighbourhood 100 data Konyaaltı-Hurma neighbourhood 100 data Konyaaltı-Gürsu neighbourhood 100 data Muratpaşa-Çağlayan neighbourhood 100 data Muratpaşa-Fener neighbourhood 100 data Muratpaşa-Meydankavağı neighbourhood 100 data
Usable Area	40-100 m2: 270 data 101-150 m2: 282 data +150 m2: 348 data
Number of Rooms	2 rooms: 78 data 3 rooms: 320 data 4 rooms: 364 data +5 rooms: 138 data
Age of residence	0-4 years: 290 data 5-10 years: 274 data 11-15 years: 170 data +16 years: 166 data
Floor	0-3rd floor: 624 data 4th-7th floor: 198 data +8th floor: 78 data
Social facilities, and the presence in complex buildings	Yes: 357 data No: 543 data

sahibinden.com site. The relevant machine learning methods were run in the open-source program Knime.

3.2 Method

Three different machine learning techniques were used to perform the prediction of house prices. Artificial neural networks, multiple linear regression and support vector regression were used in the study and the most successful model with the least error was determined.

3.2.1. Artificial neural networks

Cybernetics refers to analyzing the behavior of living things, modeling them mathematically, and producing similar artificial models. Artificial neural networks (ANNs) have emerged because of mathematical modelling of the learning process by taking the human brain as an example. The machines are intended to be trained, learn, and make decisions through artificial neural networks, just like humans (Jain, 1996; Kayakuş and Terzioğlu, 2021). It mimics the abilities to learn, remember and generalize the structure of biological neural networks in the brain. Artificial Neural Network applications are mostly used for prediction, classification, data association, data interpretation and data filtering.

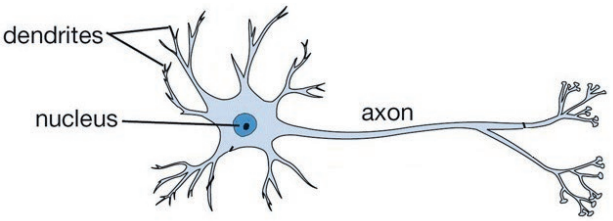

Figure 2. Biological Representation of a Nerve Cell (Abraham, 2005).

The structure of a human nerve cell (neuron) is presented in Figure 2.

Dendrites are the system inputs that collect signals from other cells. Nucleus provides periodic reproduction of marks along the axon. Synapse provides the connection of the axons of the cells with other dendrites. Axon is the system output from which the output pulses are generated.

Artificial nerve cells form the structure of ANN with the connections they have established. Artificial nerve cells have five basic elements. Each artificial nerve cell has inputs that receive external information, weights that process incoming information and create connections, summation function, activation function, and outputs or output elements that present the processed information to the outside world (Krogh, 2008).

The inputs represent information from other cells or the outside world. The summation function is the function that calculates the net input into the cell. Various functions can be used according to the ANN model to be applied. Generally, the summation function is the sum of the information coming into the cell by multiplying the weights of that information. Equation 1 shows the calculation of the net input value in the kernel.

$$NET = \sum_{i}^{N} X_i W_i \tag{1}$$

Here, X represents the entries, W represents the weight value, and N represents the total number of entries in a cell.

The activation function establishes the connection between input and output. It generates output information by processing the information from the summation function. This function, like the summation function, has various functions according to the ANN model to be applied. The "Sigmoid function" is generally used as the activation function in the "Multilayer perceptron" model, which is the most widely used today. Sigmoid function is presented in Equation 2.

$$f(NET) = \frac{1}{1 + e^{-NET}} \quad (2)$$

Output is the values generated by the activation function. The working principle of ANN is presented in Figure 3.

Artificial neural network models can be examined in four groups as single-layer perceptrons, multi-layer perceptrons, feed-forward neural networks and feedback artificial neural networks. Single-layer networks consist of input and output. They may have more than one input value. In single layer perceptron, the output function is linear and takes a value of 1 or -1. Multilayer neural networks consist of input layer, hidden layers, and output layers. Multilayer artificial neural networks are used to solve complex problems. Therefore, they are preferred for nonlinear problems. They may have multiple inputs and hidden layers. The number of hidden layers can be increased or decreased according to the flow of the problem. The hidden layer enables the problem to be processed with different functions and transferred to the output layer according to its structure (Kayakuş et al., 2022).

In feedforward neural networks, neurons are in the form of regular layers from input to output. There is only a connection from one layer to the next layers. The information coming to the input of the artificial neural network is transmitted to the middle point, in other words to the cells in the hidden layer, without any change. It is then processed through the output layer, respectively, and transferred to the external environment. In feedback artificial neural networks, unlike feedforward networks, the output of a neuron is not only given as an input to the next neuron layer. It can be connected to any neuron in the previous layer or its own layer as an input. With this structure, feedback artificial neural networks display a nonlinear dynamic behavior. According to the connection type of the connections that give the feedback feature, feedback artificial neural networks with different behavior and structure can be obtained with the same artificial neural network (Hassoun, 1995).

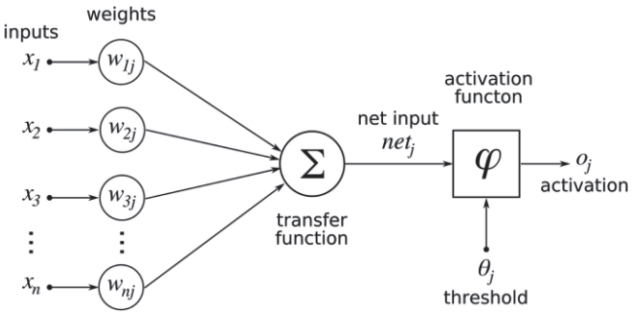

Figure 3. ANN working principle.

3.2.2. Multiple linear regression

Linear regression analysis is one of the statistical methods that are commonly used in the analysis of normally distributed dependent variables.

In simple linear regression, a bivariate model is established to predict an independent variable (x) and a dependent variable (y). If the model contains more than one independent variable to predict the dependent variable (y), then multiple linear regression techniques can be used (Eberly, 2007; Kayakuş and Terzioğlu, 2021).

Multiple regression analysis is a type of analysis for predicting the dependent variable based on two or more independent variables associated with the dependent variable. It enables to interpret the total variance explained by the independent variables in the dependent variable and to comment on the direction of the relationship between the independent variables and the dependent variable. In the regression analysis, it is aimed to establish the best model that can predict the dependent variable from the independent variables or to determine which independent variables are more affected by the dependent variable (Kayakuş, 2022; Tranmer and Elliot, 2008).

The mathematical model showing the true linear relationship can be written for n independent variables as follows:

$$y=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\beta_3X_{i3}+\cdots+\beta_kX_{ik}+\varepsilon_{ij} \quad (3)$$

defined as i=1,2,3,…,n and j=1,2,3,…,n. X_{ij}, j. represents the value of the independent variable at the i. level, B_j, j. represents the regression coefficient, e_{ij}, represents the error term, k, represents the number of independent variables. The coefficient β refers to the amount of change that will occur in Y in terms of its unit, as opposed to 1 unit change in X in its unit.

Multicollinearity may lead to incorrect estimation of the regression coefficients, exaggeration of the stand-

ard errors of the regression coefficients, resulting in an increase in the confidence intervals and a decrease in the t-test value. The increase in the standard error may cause statistically significant regression coefficients to be insignificant, thus resulting in incorrect results. The correlation matrix between the independent variables is used to detect whether there is a multicollinearity (Olive, 2017).

The basic assumptions of the multiple linear regression model are that the error term is normally distributed with zero mean and constant variance, there is no autocorrelation between the error terms, there is no relationship between the error term and the independent variables, there is no multicollinearity between the independent variables, in other words, the absence of a linear relationship between the independent variables (Bahçecitapar and Aktaş, 2017; Yamane, 1969)

3.2.3. Support vector regression

Support vector machine (SVM) is a supervised machine learning algorithm that can be used for classification or regression problems. It can be used for linear or non-linear classification and regression problems. SVM is mainly used to separate data belonging to two classes in the most appropriate way. To this end, decision boundaries, or in other words, hyper planes are determined. In other words, it can be defined as a vector space-based machine learning method that finds a decision boundary between the two classes that are furthest from any point in the training data. Support vector machines were first proposed by Vapnik (1995). They are based on statistical learning theory. This method was originally designed to solve classification and regression problems, and then, Support Vector Regression (SVR), which is used for prediction, was developed (Drucker et al., 1997). SVR ensures that the range we will draw includes the maximum point.

SVR is the regression model that allows to define how much error can be accepted in the model created. According to the errors entered, it finds a suitable line or creates a hyperplane. Therefore, the SVR method attempts to minimize the error of estimation and thus aims to find a function that approximates the training data set. In this process, the flatness of the function is maximized and the risk of getting stuck in local values is reduced (Çoban and Demir, 2021; Demir and Akkaş, 2018).

Consider a $\{(x_1, y_1), \ldots, (x_l, y_l)\}$ dataset of training points with $x_i \in R^n$ vector and $y_i \in R$ target output. The nonlinear relationship between input and output data is formulated with a linear function. A nonlinear SVR is presented in Figure 4.

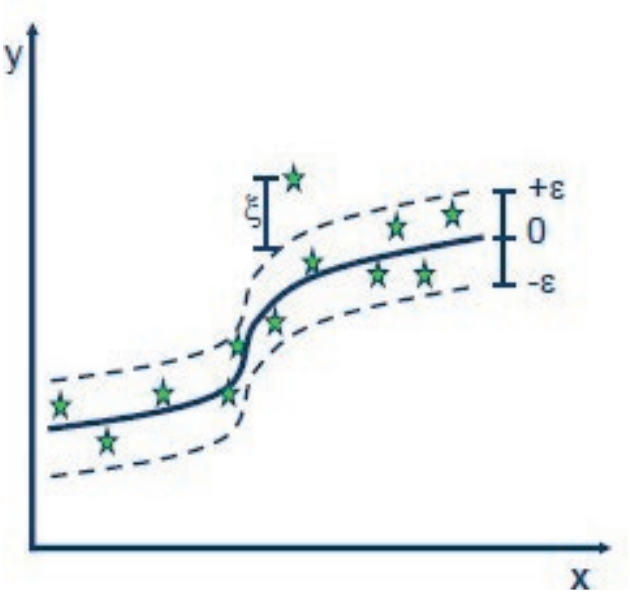

Figure 4. Non-linear SVR.

The function showing this relationship is presented in Equation 4.

$$f(x)=w^T \phi(x)+b \tag{4}$$

where, f(x) is the predicted values. Φ; is the nonlinear mapping function and $w(w \in R^n)$ and $b(b \in R)$ are adjustable coefficients. The standard form of the SVR is defined as below, with C>0 and ε>0:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i + \xi_i^* \tag{5}$$

Constraints,

$$\begin{cases} w^T \phi(x_i) + b - y_i \leq \epsilon + \xi_i \\ y_i - w^T \phi(x_i) - b \leq \epsilon + \xi^* \\ \xi_i, \xi_i^* \geq 0; i = 1,2,3, \ldots, l \end{cases} \tag{6}$$

ξ_i^* represents the training errors on ε, ξ_i represents the training errors under ε.

After solving the above quadratic optimization problem with inequality constraints, the parameter vector *w* in Equation 4 is found by Equation 7.

$$w = \sum_{i=1}^{l} (\lambda_i^* - \lambda_i)\, \phi(x_i) \tag{7}$$

where, λ_i^* and λ_i are Lagrange multipliers. Thus, the SVR formula is obtained as in Equation 8.

$$f(x) = \sum_{i=1}^{l} (\lambda_i^* - \lambda_i)\, K(x_i, x_j) + b \quad (8)$$

In Equation 8, $K(x_i, x_j)=\exp(-y\|x_i-x_j\|^2)$ function refers to the radial basis kernel function (RBF). In this method, classes that are normally not linearly separable are made linearly separable by applying the kernel function, and more successful results are obtained. The four basic kernel functions used in SVR are linear, polynomial, radial basis function (RBF) and sigmoid. In the literature, it is seen that the RBF kernel function is frequently used because it produces more satisfactory and successful results compared to other kernel functions (Abut et al., 2016).

4. RESULTS AND DISCUSSION

In the study, the data on a total of 900 house advertisements for Antalya province of Turkey were used. In the model established in the study, the property price was used as the dependent variable, and seven independent variables were used to predict this dependent variable. Three different machine learning techniques including artificial neural networks, multiple linear regression and support vector regression were used in the study. R^2, MSE and MAPE statistical methods were used to analyze the success and error of the models. The main working structure of the study is presented in Figure 5.

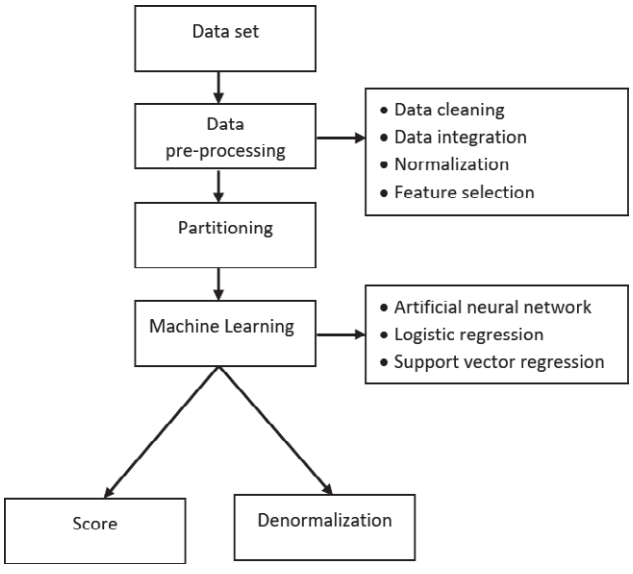

Figure 5. Main working structure of the study.

As can be seen in Figure 5, first process to be done after the dataset is created is the preprocessing stage on the dataset. The first step of this stage is data removal. Inconsistent and erroneous data in the dataset are called noise. For removing the noise in the data, the records with missing values may be excluded, missing values can be replaced with a constant value, this value can be written instead of the missing data by calculating the average of the other data, and it can be used instead of missing data by making an appropriate estimation of the data.

The second stage is the data integration stage. It is the process of converting different types of data into a single type so that the data obtained from different datasets or data sources can be evaluated together. While the price information in our dataset contains numerical information, whether it is included on the site contains yes/no information, that is, textual information. At this stage, all information in the dataset has been converted to numerical format.

The third stage of data pre-processing consists of the normalization stage. The size and value ranges of the data in the dataset may vary. Thus, numerical features of different scales may reduce the performance of the model by affecting the model applied in the learning process in an unbalanced way. The distribution of numerical features can be standardized by observing certain limit values according to the characteristics of the problem to be solved. In this study, the data were linearly normalized using the Min-Max method. It is the normalization of all values in a group of data according to the largest and smallest value in this group. While the minimum is the lowest value that a data can take, the maximum refers to the highest value that the data can take. The values to be generated here will be between 0-1. Min-Max normalization formula:

$$x' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (9)$$

where, x_i represents the normalized data, x_i represents the input value, x_{min} represents the smallest number in the input set, x_{max} represents the largest number in the input set.

Various methods have been developed for variable selection. These methods are examined in two groups according to calculation techniques: classical methods and stepwise methods. If stepwise methods are (Alpar, 2003):

1) Forward selection method
2) Backward selection method
3) Stepwise selection method

In the forward selection method, it is desired to find the most appropriate regression model by adding one independent variable each time. In the backward selection method, the opposite is the case with the forward selection method. The model starts with all independent variables. In the stepwise selection method, both the forward selection method and the backward selection method are used simultaneously (Çakır Zeytinoğlu, 2007; Kayaalp et al., 2015). The forward selection method was used in this study.

Another stage of the study was dividing the data into two as training and testing. It is the stage where the ideal parameters for the machine learning chosen during the training phase are determined and the error is reduced to the minimum level. In the test phase, the parameters determined during the training phase are tested on the data that have not been used before in the dataset and are statistically evaluated. In the literature, it is accepted to divide the dataset into 70% Training, 30% Test or 80% Training, and 20% Test data. While doing this application, when the dataset was divided into 70% Training and 30% Test data, our current dataset included 630 Training and 270 Test data. Different methods are used to divide the data into two as training and testing. Take from top, linear sampling and draw randomly are some of the data selection methods that can be used. In the study, the linear sampling method was preferred in data selection to compare the results of the two models.

R^2 (Coefficient of determination), MSE (Mean Squared Error) and MAPE (Root Mean Square Error) techniques were used to analyze and interpret the results of the study.

R^2 is the coefficient of determination, which is the measure of how much the independent variable x explains the dependent variable y with the regression model. R^2 takes values between 0 and 1 ($0<R^2<1$). The fact that R^2 value approaches 1 when there is a linear relationship between the variables indicates that most of the variation in the dependent variable is explained by the independent variables. R^2 describes the extent to which the variance of one variable explains the variance of the second variable. R^2 formula is presented in Equation 10:

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}} \tag{10}$$

MSE refers to how close a regression curve is to a set of points. MSE measures the performance of a machine learning model, estimator, and is always positive. It can be said that estimators with an MSE value close to zero perform better. MSE gives an absolute number of how much your predicted results differ from the actual number. MSE formula is presented in Equation 11:

$$\text{MSE} = \frac{1}{n}\sum_{j=1}^{n} e_j^2 \tag{11}$$

Here, n is the number of data and e is the error value.

The MAPE statistic eliminates the disadvantages that may arise in the comparison of models with different unit values. Among the listed criteria, the fact that MAPE has a meaning on its own as it expresses the prediction errors as a percentage is considered as its superiority over other criteria. While the models with MAPE <10% are classified as "very good", the models with 10% < MAPE < 20% are classified as "good", the models with 20% < MAPE < 50% are classified as "acceptable", and the models above 50%<MAPE are classified as "false and faulty". MAPE formula is presented in Equation 12:

$$\text{MAPE} = \frac{100}{n}\sum_{j}^{n} \frac{|e_j|}{|A_j|} \tag{12}$$

Three different machine learning techniques were used in the study. They were neural networks, support vector regression and multiple linear regression.

A feedback model consisting of seven inputs and one output neuron was developed for the artificial neural network method. The number of hidden layers in the model and the number of neurons in the hidden layer were determined by trial-and-error method. As a result of the study, it was seen that the structure with two hidden layers and two neurons in each layer produced more successful results since it gave the most successful result. The structure of the developed model is presented in Figure 6.

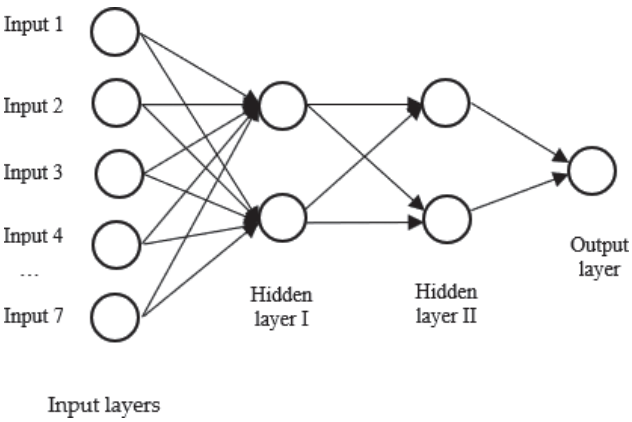

Figure 6. ANN model designed.

The sigmoid function was used after testing various functions for the Activation Function. Back-propagation algorithm was preferred as the learning algorithm, and its parameters were determined automatically through the software. 1.000 iterations were performed to get the best results in the model.

Nonlinear SVR was used for support vector regression, which was another machine learning method used in the study. Polynomial, hyper tangent, radial basis function (RBF) was tested for the kernel function and RBF was preferred in the study since it was determined as the most successful model. Other parameters of the model were chosen as overlapping penalty value of 100 and RBF sigma value of 0.1.

A significance value was first determined to measure the effects of variables on the system in multiple linear regression. The variable with the current highest p-value (probability value) now was determined and if P>SL, the variable was removed from the system. The model was established again and then this step was repeated. The elimination was terminated when it was P<SL for all variables. Since there were no independent values below 0.05 for p values in the designed model, the model was found to be significant.

The success and error values of the models according to the ANN, SVR and MLR methods are presented in Table 4.

An R^2 value of 1, which indicates how well the data fit a linear curve, indicates that the test data have provided a linear curve. As a result of the study, the R^2 value was 82.7% for ANN, 72.1% for MLR, 74.9% for SVR and it was seen to be very close to the ideal value. MSE measures the performance of a machine learning model, estimator, it is always positive, and it can be said that estimators with MSE value close to zero perform better. Therefore, the MSE value is desired to be close to zero. In the study, it was observed that the MSE value was 0.006 for ANN, 0.008 for MLR and 0.027 for SVR, which was close to the ideal value. The models with a MAPE of less than 10 percent are considered very well. In the study, it was observed to be 4.86 for ANN, 6.69 for MLR and 6.69 for SVR. It was considered that the MAPE value was very good in all three models. Considering the error and success values, it was observed that the most successful and least error models were ANN, SVR and MLR, respectively.

Table 4. Comparison of success and error of the models.

	Artificial neural network	Multiple linear regression	Support vector regression
R^2	0.827	0.721	0.749
MSE	0.006	0.027	0.008
MAPE	0.486	0.669	0.567

Since the results of the study were normalized, they are shown between 0 and 1. Denormalization is performed to adapt the results to real values and make sense of them. Thus, users reach the real values.

5. CONCLUSION

Housing is the basic need of households. The basic features of the house in the living areas are decisive in meeting these needs. These features are the value and other physical characteristics of the house. Therefore, households prefer and seek houses that are suitable for their income and meet their needs. The location, size, number of rooms, age, floor, independent property or whether it is in a site, which we call as the micro-variables of the house, were used to predict the price of the house in this study. Artificial neural networks, multiple linear regression, and support vector regression techniques, which are among the machine learning methods were used for this prediction.

The success and error analyses of machine learning methods were statistically performed based on the R^2, MSE and RMSE criteria. The most successful and least erroneous models used in the study were neural networks, support vector regression and multiple linear regression, respectively. The models are considered acceptable value according to the R^2 value. It was observed that the change in the independent variables used in the model affected the dependent variable in all three methods. According to the MSE and RMSE criteria, they are considered as the methods that predict with low error coefficient, in other words, with less error. These results of the study are like the studies of Cechin et al. (2000), Mora-Esperanza and Gallego (2004), and Varma et al (2018) that revealed the features the house as variables and uses ANN and multiple linear techniques in the literature.

With this study, institutions that supply housing and mediate their sales will make the price fluctuations as stable as possible and prevent speculative movements in the market by accurately estimating the prices of subjects with similar features. Housing suppliers will supply houses according to the preferences and price expectations of the households. Institutions that act as an intermediary for those seeking housing will work more effectively in finding houses according to the budget of households. They will be able to help their customers in making the right decision by using the artificial neural

networks machine learning technique. Financial institutions involved in housing finance will have more accurate appraisal results by using this model. Thus, it will reduce capital costs by providing the right number of financial resources to house demanders.

REFERENCES

Abut, F., Akay, M. F., & George, J. (2016). Developing new VO_2max prediction models from maximal, submaximal and questionnaire variables using support vector machines combined with feature selection. *Computers in Biology and Medicine*, 79, 182–192.

Abraham, A. (2005). Artificial neural networks. In Sydenham, P. H.,& Thorn, R. (Eds.). *Handbook of measuring system design*. London, John Wiley and Sons, 901–908.

Alpar, R. (2003). Introduction to Applied Multivariate Statistical Methods-1. Ankara, Turkey, Nobel Publication.

Bahçecitapar, M., & Aktaş, S. (2017). Use of linear mixed model in multicollinearity and an application. *Sakarya University Journal of Science*, 21(6), 1349–1359.

Bhagat, N., Mohokar, A., & Mane, S. (2016). House price forecasting using data mining. *International Journal of Computer Applications*, 152(2), 23–26.

Bircan, H. (2004). Studies of the Logistic Regression Analysis and its application on the medical data. *Kocaeli University Journal of Social Sciences*, 8, 185–208.

Cechin, A., Antonio, S., & Gonzales, M. A. (2000). Real estate value at Porto Alegre City using Artificial Neural Networks. In *Proceedings, Vol. 1, Sixth Brazilian Symposium on IEEE on Neural Networks, November 2000*. 237–242.

Çakır Zeytinoğlu, F. (2007). The effects of current assets of companies on sales: selection of the best regression equation and sectoral comparison. *Marmara University Journal of Economic and Administrative Sciences*, 23(2), 331–349.

Çoban, F., & Demir, L. (2021). Demand forecasting with Artificial Neural Networks and Support Vector Regression: an application in a food company. *Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering*, 23(67), 327–338.

Demir L., Akkaş S. (2018). A comparison of sales forecasting methods for a feed company: a case study. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 24(4), 705–712.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. In Moser, M., Jordan, J., & Petsche, T. (Eds.). *Neural information processing systems*, Vol. 9. Cambridge, MA, USA, MIT Press, 155–161.

Eberly, L. E. (2007). Multiple linear regression. In Ambrosius, W. T. (Ed.). *Methods in Molecular Biology, vol. 404: Topics in Biostatistics*. Totowa, NJ, USA, Humana Press, 165–187.

Mora-Esperanza, J. G., & Gallego, J. (2004). Artificial intelligence applied to real estate valuation an example fort he appraisal of Madrid. *Catastro*, 1, 255–265.

Feggella, D. (2019). What is machine learning? Available at: https://emerj.com/ai-glossaryterms/what-is-machine-learning/ (Accessed 11 April 2022)

Hassoun, M. H. (1995). *Fundamentals of artificial neural networks*. Cambridge, MA, USA, MIT press.

Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: a tutorial. *Computer*, 29(3), 31–44.

Iacoviello, M. (2000). House prices and the macroeconomy in Europe: results from a Structural VAR Analysis. ECB Working Paper No. 18.

Kayaalp, G. T., Güney, M. Ç., & Cebeci, Z. (2015). Variable selection in Multiple Regression Model application of animal science. *Çukurova University Journal of the Faculty of Agriculture*, 30(1), 1–8.

Kayakuş, M., & Terzioğlu, M. (2021). The prediction of pension fund net asset values using Artificial Neural Networks and Multiple Linear Regression methods. *Journal of Information Technologies*, 14(1), 95–103.

Kayakuş, M. (2022). Estimating the changes in the number of visitors on the websites of the tourism agencies in the COVID-19 process by Machine Learning Methods. *Sosyoekonomi*, 30(53), 11–26.

Kayakuş, M., Terzioğlu, M., & Yetiz, F. (2022). Forecasting housing prices in Turkey by machine learning methods. *Aestimum*, 80, 33-44.

Kim, K., & Park, J. (2005). Segmentation of the housing market and its determinants: Seoul and its neighbouring new towns in Korea. *Australian Geographer*, 36(2), 221–232.

Koktashev, V., Makee, V., Shchepin, E., Peresunko, P., & Tynchenko, V. V. (2019). Pricing modeling in the housing market with urban infrastructure effect. *Journal of Physics*. Conference Series, 1353(12139), 1–6.

Krogh, A. (2008). What are artificial neural networks? *Nature Biotechnology*, 26(2), 195–197.

Ngiam, K. Y., & Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, 20(5), 262–273.

Olive, D. J. (2017). Multiple linear regression. In Olive, D. J. (Ed.). *Linear regression*. Cham, Switzerland, Springer, 17–83.

Park, B. H., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: the case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, 42(6), 2928–2934.

Rafiei, M. H., & Adeli, H. (2016). A novel machine learning model for estimation of sale prices of real estate units. *Journal of Construction Engineering and Management*, 142(2), 04015066.

Shiller, R. J. (2007). Understanding recent trends in house prices and home ownership. National Bureau of Economic Research, Working Paper 13553.

Thamarai, M., & Malarvizhi, S. P. (2020). House price prediction modeling using machine learning. *International Journal of Information Engineering & Electronic Business*, 12(2), 15–20.

Tranmer, M., & Elliot, M. (2008). Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR), 5(5), 1–5.

Wang, C. C., & Wu, H. (2018). A new machine learning approach to house price estimation. *New Trends in Mathematical Sciences*, 6(4), 165–171.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York, USA, Springer.

Varma, A., Sarma, A., Doshi, S., & Nair, R. (2018). House price prediction using machine learning and neural networks. 2. In *Proceedings of the 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), April 20–21, 2018, Coimbatore, India.* 1936–1939. IEEE.

Yamane, T. (1969). *Statistics: an introductory analysis*. New York, USA, Harper & Row.

Yu, H., & Wu., J. (2016). Real estate price prediction with regression and classification CS 229 Autumn 2016 Project Final Report 1–5. Available at: http://cs229.stanford.edu/proj2016/report/WuYu_HousingPrice_report.pdf (Accessed 6 July 2022)