Journal of Universal Computer Science, vol. 30, no. 4 (2024), 531-560
submitted: 12/12/2022, accepted: 15/11/2023, appeared: 28/4/2024 CC BY-ND 4.0

Price Prediction and Determination of the Affecting Variables of the Real Estate by Using X-Means Clustering and CART Decision Trees

Sait Can Yucebas
(Canakkale Onsekiz Mart University, Canakkale, Turkey
https://orcid.org/0000-0002-1030-3545, can@comu.edu.tr)

Sukran Yalpir
Konya Technical Univeristy, Konya, Turkey
https://orcid.org/0000-0003-2998-3197, syalpir@ktun.edu.tr)

Levent Genc
(Canakkale Onsekiz Mart University, Canakkale, Turkey
https://orcid.org/0000-0002-0074-0987, leventgc@comu.edu.tr)

Melike Dogan
(Laren Engineering Map Design, Mugla, Tukey
https://orcid.org/0000-0002-1945-6927, melikedidiemdogan@gmail.com)

Abstract: The use of machine learning in real estate is quite new. When the working area is large, the factors affecting the price may vary according to the geographical regions and socioeconomic factors. It is thought that the price prediction performance of a model that will reflect these differences will be more successful than a general model. Unsupervised learning methods can be used both to increase performance and to show the variation of different factors affecting the price according to regions. With this aim, a hybrid model of X-Means clustering and CART decision trees was established in this study. This model successfully learned the geographical and physical variables that affect the price. The prediction performance of the model was compared with the direct capitalization method, which is the gold standard in the domain. The hybrid model has a superior performance over direct capitalization in terms of mean square error, root mean square error and adjusted R-Squared metrics. The scores were 72.86, 0.0057 and 0.978, respectively. The effect of clustering was also examined. Clustering increased the prediction performance by 36%.

Keywords: Machine learning, Classification and regression tree, X-Means clustering, prediction methods, Real estate
Categories: I.2, I.2.1, I.2.6
DOI: 10.3897/jucs.98733

1 Introduction

Today machine learning (ML) is used to analyze vigorous amount of data in different domains such as medicine [Garg and Mago, 2018] bioinformatics [Lan et al., 2018], finance [Ozbayoglu et al., 2020], security [Liu et al., 2018], e-commerce [Leung et al., 2020], sports [Beal et al., 2019] etc. Real estate is a new domain where the advantages of ML are explored recently. In this domain, ML is used mainly for price prediction instead of traditional methods like direct capitalization (DC). DC can reveal accurate predictions. However, it is often not applicable because it requires vast amount of variables [Abidoye and Chan, 2017]. In addition, expert opinion is needed for the initial evaluation of the variables [Abidoye and Chan, 2017; Abidoye and Chan, 2018]. It is inevitable that the DC makes false predictions when one or more of these requirements are missing.

The use of ML in real estate price prediction gained momentum because it eliminates manual analysis of vast amount of variables and the need for expert opinion. When the studies are examined, it is seen that regression (LR) and black box approaches such as Support Vector Machine (SVM) [Phan, 2018], Artificial Neural Network (ANN) [Li and Chu, 2017] and Deep Learning (DL) [Zhao et al., 2019] [Manasa et al., 2020] are frequently preferred. This indicates that supervised learning methods are generally preferred, and unsupervised learning is often omitted. However, ML studies in different domains reveal that unsupervised learning could contribute positively to performance [Erhan et al. 2009]. Variables affecting the price were determined by decision tree method in the study by Yucebas et al. [2022]. However, instead of numerical price prediction, the price scale was divided into three classes (low, normal, high) and these classes were estimated. The study has two shortcomings. The first is that there is no unsupervised support in the learning model, and the second is that the dataset used covers a very narrow region.

When current studies were examined three main gaps were identified. These studies focused mainly on the prediction performance and the variables affecting the result often overlooked. In addition, although the use of unsupervised learning could improve performance, blending supervised and unsupervised methods was not done in most of the studies. More importantly, studies covering large areas were limited to a general model for the entire region. However, the prediction performance of a model that will reflect the geographical, socio-economic etc. differences of certain regions will be higher.

To address these issues in the related domain, a hybrid model of supervised and unsupervised learning was developed. Main motivation of the study is to reveal the variables that affect the price while establishing a solid prediction performance. The prediction performance of the hybrid model was compared with the domain gold standard DC. The effect of unsupervised learning on the performance was also evaluated.

There are two main contributions of the study. First, instead of establishing a general suit to all model, clusters that represents different socio-economic and geographical regions were formed. X-Means clustering with an adoption of DILCA method was used to form these clusters. By this way, problems such as optimum k value and distance between discrete variables were addressed. Second, a specific CART model was established for each cluster. Then, the variables affecting the price were analyzed in detail for each model.

The organization of this paper is follows: Section 2 gives a background of related studies. The details of the data and the methods are given in the Section 3. Section 4 and Section 5 gives results and the performance comparison respectively. Conclusion is given in Section 6 and the paper ends with future work given in Section 7.

2 Related Work

[Li and Chu, 2017] used ANN for price prediction. Apart from other studies, economic variables such as income rate, economic growth and loan rates were included. Instead of predicting the actual price, price indexes were estimated. ANN model was used in the study. The focus of the study was on the predictive performance, however the authors indicated that the accuracy of the models were questionable. This study fails to provide the variables that affect pricing.

The study by [Phan, 2018] compared the performance of different regression based models, SVM and ANN. The discrete attributes are critical for regression based models because they require data transformation. However, in the study, it was not specified how discrete data were handled. The variable importance were calculated, but it was used for feature reduction.

[Manasa et al., 2020] also compared the regression-based methods. The biggest drawback of these methods is the use of discrete data. The discrete to numeric conversion was strongly needed because it affects the performance of the model. Methods like one hot encoding [Hancock and Khoshgoftaar, 2020] or Jaccard transformation [Ahmad and Khan, 2019] could fail to represent the differences between discrete attributes [Cerda and Varoquaux, 2022]. Another drawback, also indicated by the authors, was the lack of the residential type in the dataset. This information is important in price prediction because prices change according to residential type.

Another performance comparison was conducted between Random Forest (RF) and linear regression [Wang and Wu, 2018]. The number of predictive variables was very few. The performance of the models was compared and RF gave better performance in terms on R^2 and root mean square (RMSE) metrics. Model parameters were set to default and no parameter optimization was conducted. Most importantly, the study did not reveal any clues on variables that affect the price.

A study [Varma et al., 2018] compared the performance of RF, LR and ANN. They suggested the use of hybrid models, which could increase the prediction performance. One drawback of the study was the dataset used. Study area had a limited border and real estates were not diverse. Therefore, the proposed model will not be able to make high performance predictions for larger areas such as neighborhood or entire city.

The performance of SVM, RF and Gradient Boosted Machine (GBM) was compared in the study by [Ho et al., 2021]. The samples were taken from the Hong Kong city. The algorithms were compared by different error metrics. In terms of prediction performance, RF and GBM performed better. SVM stood out with time efficiency. Authors stated that the use of machine learning in real estate price prediction is still in its infancy state. In addition, they emphasized that each of the methods has different disadvantages. This idea constitutes the proof of the necessity of a hybrid system.

The study by [Mohd et al., 2020] is among the most comprehensive studies that compared several ML methods in real estate price prediction. Fourteen models were compared in detail. Their usage, advantages and disadvantages were discussed in terms of prediction performance. However, detecting the variables that affect the prediction was overlooked.

A study [Yucebas et al., 2022] compared the prediction performance of decision tree and hedonic model. The study proposed a method that converts continuous values of unit price to discrete. Models were compared based on the discrete unit price classification. Kappa and accuracy metrics were used. Decision tree outperformed hedonic model for all given classes. Although it is one of the pioneering studies in terms of examining the variables that affect the price, the study has some shortcomings. The study area was limited to a single neighborhood. The parameters of the models were not optimized and the effect of unsupervised learning was ignored.

The hybrid model developed by [Lee, 2021] is the closest work to our study. The authors stated that unsupervised learning could improve model's performance. To prove the concept, a hybrid model of ANN and principal component analysis (PCA) was used. This study has three main differences from ours. Unsupervised learning was used for feature reduction. However, in our study, unsupervised learning was used to reveal the differences on the dataset. In Lee's study, a single supervised model was established on the entire data set. In our study, a specific model was established for each cluster. While the land prices in a single neighborhood were selected as the study area, our study predicts the price of residential scattered to the entire city.

Related studies given above can be grouped under two categories. Studies that use single model for price prediction and the studies that compare the prediction performance of several models. These studies are summarized in Table 1.

Reference	Methods	Criticism	Hybrid	Unsupervised Learning	Parameter Optimization	Determination of the Variables that Affect Price
[Li and Chu 2017]	ANN	Low Accuracy	X	X	X	X
[Phan 2018]	LR, SVM, ANN	No data transformation, Handling of discrete variables	X	X	X	Variable importance was calculated but used for feature reduction
[Manasa et. al. 2020]	LR	Handling of discrete variables, Residential type was missing	X	X	X	X
[Wang and Wu, 2018]	LR, RF	Low number of variables	X	X	X	X
[Varma et al., 2018]	LR, RF, ANN	Low diversity of the dataset	X	Suggested to use hybrid models	X	X
[Ho et al., 2021]	RF, SVM, GBM	X	X	X	X	X
[Mohd et al., 2020]	14 different models	Default parameters of the models were not given	X	X	X	X
[Yucebas et al., 2022]	DT, HM	Limited study area,	X	X	X	V

		Continuous values were not used, No unsupervised learning				
[Lee, 2021]	ANN, PCA	Unsupervised phase for feature reduction, Limited study area, Single model for entire dataset	V	V	X	X
Proposed Model	X-Means, CART		V	V	V	V

Table 1: The comparison of the proposed study to the current studies V: refers "yes" and X: refers "no"

In terms of methods, regression-based and black box approaches were frequently preferred. In the regression-based methods, there was a lack of discrete to numerical conversion. Hyper-parameter optimization was skipped in black box methods. In addition, majority of the studies used supervised learning and contribution of unsupervised learning was ignored. More importantly, no matter which method was used, these studies focused on prediction performance and the examination of the variables that affect the price was missing.

Based on these gaps, a hybrid model was developed to analyze the variables that affect the price. Supervised learning was used to perform the related analysis, while unsupervised learning was used to increase prediction performance and to reveal the differences between neighborhoods and to reflect the geographical and socioeconomic situation.

The prediction performance of the established model was also considered. To ensure the best performance, hyper-parameter optimization was used in supervised learning phase. For the same purpose, DILCA method was used in unsupervised learning phase.

3 Method

The main aim of the study is to establish a hybrid model that predicts the unit price and to reveal the variables that affect the price tag. The infrastructure of the hybrid model is given in Figure 1.

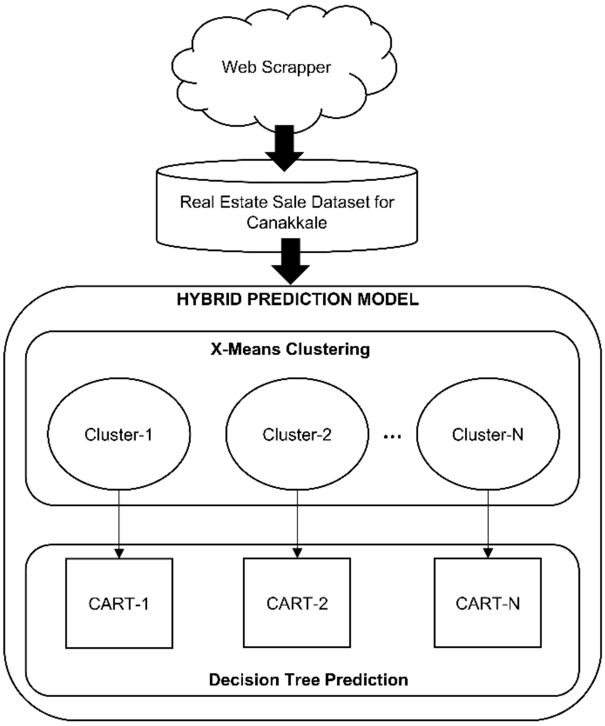

Figure 1: The infrastructure of the hybrid model

A web scrapper, based on Python scrapy library, was designed to retrieve the real estate ads. Clustering was used for grouping similar sale ads and to reflect different socio-economic and geographical regions. For each cluster, a specific decision tree model was established to predict the prices and to reveal the variables that affect the price. In order to meet these aims a hybrid model of X-Means clustering and classification regression tree (CART) was established. The CART approach was used for price prediction and for the analysis of the variables that affect the price. Clustering was used to reveal the differences among the regions and to increase the prediction performance.

3.1 Clustering Approach

In terms of the prediction performance, when the data set is very extensive, using a single model for prediction could lead to poor predictive performance [Fong and Hong, 2021]. Therefore clustering was used to divide the dataset into more homogenous subsets and a CART model was built for each cluster.

In the clustering phase, an adoption of X-Means algorithm [Pelleg and Moore, 2000] was used. One of the biggest problems of centroid methods is finding the optimum number of clusters. X-Means solves this problem by finding the optimum k value. In this method, an upper and lower limit is determined for the k. Then, starting from the lower bound, centroid-based clustering logic is applied for each k. The first centroids are determined according to k. Then, each centroid creates partitions within itself and clusters are formed. The Bayesian Information Criteria (BIC) determines the quality of the newly formed clusters. For each model, the number of data points is denoted with N, the number of parameters is given by par and log likelihood is L. Then the BIC is calculated as follows:

$$BIC = L - \frac{1}{2} par \log N \tag{1}$$

It is assumed that the data points in the clusters fit the Gaussian distribution in order to find the optimum k number and to avoid the problems in centroid calculation due to the mean [Raykov et al., 2016]. Based on this assumption, L value was calculated by Equation 2. In this equation p(x_{ik}) is the probability of the data point x_i belongs to cluster k.

$$L = \log \prod_k \prod_i p(x_{ik}) \tag{2}$$

The k value, which creates the best quality clusters according to the BIC criterion, is determined as the optimum value [Patibandla and Veeranjaneyulu, 2018].

The second problem in clustering is the distance calculation between discrete variables. For example, in a color variable, blue is closer to the dark blue and is distant to the white. However, in traditional approaches the distance between all different values are calculated as one. This shows the difficulty in reflecting the content of the relevant category. Methods such as DILCA [Battaglia e. al., 2021] and DVD [Xavier et al., 2013] were proposed to overcome this challenge. The algorithmic complexity of DVD is high; thus, DILCA was used in this study. DILCA calculates entropy and content value for each discrete variable. By calculating the Euclidean distance between these content values, the distance between the discrete variables was found. If we assume that C denotes a class set, D is the data belonging to class C, P_i is the probability of data belong to class C_i, E_A is the entropy of attribute A, and E_{AB} is the entropy of attribute A with respect to the attribute B. Then information gain is calculated as follows:

$$E_A = -\sum_{i=1}^{k} p_i \log_2 p_i \tag{3}$$

$$E_{AB} = (-\sum_{i=1}^{k} p_i \log p_i)(\sum_{i=1}^{k} \frac{D_{1i} + ... + D_{mi}}{D}) \tag{4}$$

$$InformationGain = E_A - E_{AB} \tag{5}$$

In order to prevent the bias through attributes with more values, the ratio given in the equation 6 is used:

$$R(A,B) = 2 * \frac{E_{AB}}{E_A + E_B} \tag{6}$$

This ratio is known as the uncertainty value and its mean was used to measure the context of a given attribute. After contexts were determined, the distance was measured by Euclidian distance as given in the Equation 7:

$$Dist(x_i, x_j) = \sqrt{\sum_{B \in Contxt(a)} \sum_{B_k \in B} (p(x_i \mid B_k) - p(x_j \mid B_k))^2} \tag{7}$$

The dataset consists of both numeric and discrete values. Traditional clustering methods could fail to assign discrete data to clusters mostly because they fail to represent the context of these attributes. To solve this problem DILCA method was used. By this way, discrete attributes were represented by entropy values. Thus, all data was converted to numeric values in the clustering phase. Then the optimum K value was calculated by the equations 1and 2, and distance between data points was calculated by Euclidean Distance.

In the second phase of the hybrid model, CART was established for each cluster. By this way, the examination of the variables that affect the sale price was carried out in more detail.

3.2 Decision Tree Approach

For nonlinear problems, decision trees can provide promising results when compared to much complex algorithms. Their advantage over other learning algorithms is they can visually present the variables that affect the classification and/or prediction. Decision logic can be presented as rule sets by following the branches of the tree.

The basic idea behind the decision tree is finding the best split that will divide the dataset into homogenous subsets. Different splitting criteria such as Gini Index [Jain et al., 2018], Information Gain [Jain et al., 2018] and Information Gain Ratio [Mienye et al., 2019] can be used. In the related study, to prevent a bias in favor of attributes with large value range, the information gain ratio (IGR) was used.

As given in the materials section, the data set consists of continuous and discrete attributes. In order to handle the attributes with continuous values and to prevent a bias towards them, a penalty term (PT) given in Equation 8 [Quinlan, 1996] was used.

$$PT = -\log_2 \frac{n-1}{|D|} \tag{8}$$

To predict the continuous values, branches was formed according to the actual value convergence. The level of convergence was calculated by the mean square error (MSE) metric. Suppose that the number of data is denoted by n, x denotes each data tuple in the set, C is the class tag, V is the collection of attributes that represents a tuple and F is the prediction function. Then the error is calculated as follows:

$$\sum\nolimits_{i=1}^{n} (C_i - F(U, x_i))^2 \tag{9}$$

To achieve the maximum performance from the CART, an evolutionary hyper-parameter optimization (HPO) was conducted. In evolutionary HPO, the number of solution alternatives is reduced by random search and the best solution is granted by mutation and crossover [Yang and Shami, 2020]. In the HPO algorithm, Gaussian mutation with tournament selection with a fraction of 0.25 and cross over probability of 0.9 was used. As a result, the depth of the CART model was determined as ten, minimum leaf size was two and the minimum size for split was four. To test the performance, tenfold cross validation was used. Each fold was constructed by stratified sampling. The results of the CART model are given in the following section.

3.3 Direct Capitalization Approach

The prediction performance of the proposed hybrid model was compared with DC model which is known to be a gold standard for real estate price prediction [Lennhoff, (2011)]. DC predicts the price according to the rental income [Pınar and Demir, 2014]. The risk of the flat being empty can also be calculated. However, this has not been taken into account as it may reduce the prediction performance of the DC method [Yalcin et al., 2018; Michaletz and Artemenkov, 2018]. In order to calculate DC, average unit price (AUP) must be calculated first as in the Equation 10.

$$AUP = \frac{\sum_{i=0}^{n} sp}{\sum_{j=0}^{n} fa} \tag{10}$$

In above equation sp is the sale price, fa is flat area and n is the number of properties. After AUP is found, DC is calculated as given in the Equation 11.

$$DC = \frac{AI}{V} \tag{11}$$

AI indicates annual income, V indicates overall capitalization rate which is based on AUP. DC values of the dataset were calculated by a domain expert.

3.4 Performance Metrics

Different metrics can be preferred to compare the performance of the models. However, price tag is continuous that makes prediction task a multiple regression. In this case, traditional performance metrics such as accuracy could fail to compare the models. The majority of studies recommend root mean squared error (RMSE), mean absolute percentage error (MAPE) and adjusted R2 metrics to compare regression models.

.Particularly for MAPE and RMSE, there are different opinions as to which of these metrics is more valid. Some studies advocate the use of MAPE because RMSE values become too large when error magnitudes increase [Willmott et al., 2009], there are also studies showing that RMSE is a better metric when the error distribution is normal [Hodson, 2022]. Both metrics were used because there was no consensus on the relevant metrics.

RMSE calculates the root of squared difference between actual measurement (a_{ct}) and the predication (p_{re}) for n values, as given in equation 12.

$$RMSE = \sqrt{\frac{(a_{ct} - p_{re})^2}{n}} \tag{12}$$

MAPE is used to show how much the prediction (p_{re}) deviates from the actual measurement (a_{ct}) as a percentage. MAPE calculation for n distinct values is given in equation 13.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left| \frac{a_{ct_i} - p_{re_i}}{a_{ct_i}} \right| \tag{13}$$

R^2 metric is used to identify the variance of the prediction result based on the given attributes [Chicco et al., 2021]. However, the number of attributes added to the model can relatively increase the R2 value [Akossou et al., 2013]. Although this situation looks good in terms of performance, it may introduce overfitting [Miles, 2005]. When the number of attributes is increased, adjusted R^2 is used to see if the performance increase is due to chance or by overfitting [Miles, 2005]. Adjusted R^2 can be calculated as in equation 14.

$$Adjusted\ R^2 = 1 - (1 - R^2)\frac{n-1}{n - attr_{\#} - 1} \quad (14)$$

In equation 14 number of attributes was denoted as $attr_{\#}$. As this equation indicates R^2 must also be calculated. The formula to calculate R^2 is given in equation 15. In the equation actual measurement is denoted by a_{ct}, mean of actual measurement is denoted by μa_{ct} and prediction is denoted by p_{re}.

$$R^2 = \frac{\Sigma(a_{ct_i} - \mu a_{ct})^2 - \Sigma(a_{ct_i} - p_{re})^2}{\Sigma(a_{ct_i} - \mu a_{ct})^2} \quad (15)$$

4 Material

In the study, the residential ads for sale were retrieved from the Canakkale province between 01 Spt. 2021 and 30. Nov. 2022. Canakkale is one of the provinces in the Marmara region on the Northwest Anatolian side of Turkey. Figure 2 shows the geographical location of Canakkale on the map.

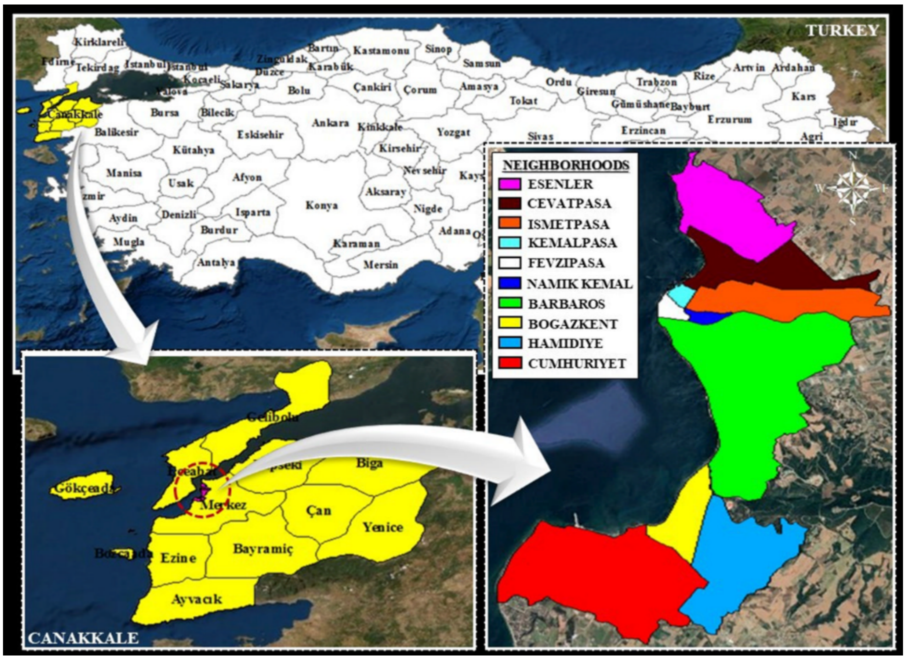

Figure 2: Geographical location and neighborhoods of Canakkale

As shown in Figure 2, there are ten neighborhoods in the study area. The neighborhoods Cumhuriyet, Hamidiye and Bogazkent constitute the region known as Kepez. The neighborhoods in the city center are Barbaros, Ismetpasa, Cevatpasa, Fevzipasa, Kemalpasa, Namık Kemal and Esenler. The area of Kemalpasa, Fevzipasa and Namık Kemal neighborhoods is quite small. Thus, the number of sale ads in the relevant neighborhoods was very low. In order to obtain sufficient number of ads, these neighborhoods were combined and named as ”UnitedNH”.

The residential for sale in Canakkale was retrieved from a website that covers all rental and sale real estates of Tukey. A web scrapper based on the python scrapy library was developed for the given website. The scrapper was used to retrieve the tabular information for all rental or sale residential in Canakkale. The information in free text area was omitted because it was left blank for most of the ads.

Web scrapping gathered a dataset of 500 residential sales ads. For each residential, twelve variables were used. Table 2 gives details of the variables.

Variable	***DataType***	***Min - Max***	***Mean***	***Std. Dev.***
Unit Price	Numeric	2,166.67-6,847.83	3,651.33	833.63
Area (m^2)	Numeric	35.00 - 800.00	123.18	66.54
Current Floor	Numeric	-2.00 - 10.00	2.72	1.85
Number of Floors	Numeric	1.00 - 10.00	4.79	1.44
Age of Building	Integer	0.00 - 50.00	7.26	8.69
Variable	***Data Type***	***Number of Values***	**Min Freq.**	**Max Freq.**
Neighborhood	Nominal	10	Namık Kemal (3)	Barbaros(116)
Residential Type	Nominal	7	Summer House (1)	Flat (485)
Number ofRooms	Nominal	12	8+3 (1), 9+3 (1)	2+1 (160)
Heating	Nominal	6	Room Heater (1)	Boiler (379)
Deed Type	Nominal	3	Land (3)	Condominium (216)
Facade	Nominal	13	North-East-South (5)	South (128)
Fuel Type	Nominal	2	Coal - Wood (8)	Natural Gas (426)

Table 2: Numeric and discrete variables of the dataset. Each numeric variable is summarized by max, min, mean and standard deviation. For discrete variables, number of values, min and max frequencies are given

Some of the variables in Table 2 have the same value for most of the samples. This may create the impression that these variables are not self-explanatory and only increase the complexity of the model. However, the model’s ability to distinguish these rare values will show its discrimination power.

Economic variables such as gross national product, gross domestic product, exchange rates and stock market indices were also retrieved. However, the analyzes showed that the three-month period was not sufficient enough to show the effect of economic variables on the price. Therefore, economic variables were excluded. A future study is planned, in which data are retrieved for a longer period. Thus, the effect of economic variables can be examined.

According to the experts and literature review, one of the variables that affect price is the neighborhood [Macpherson and Sirmans, 2001; Fernandez et al., 2013]. Thus, statistical distribution of the unit price among neighborhoods was calculated and is given in Figure 3.

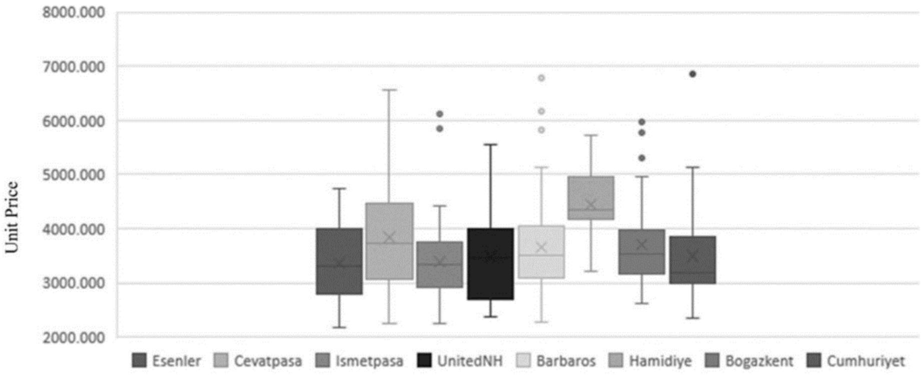

Figure 3: The box plot of the unit price among neighborhoods

There are both old and new constructions in Cumhuriyet, Barbaros and Cevatpasa neighborhoods. Thus, they were preferred by wide variety of socio-economic segments. This situation explains the wide price range of the relevant neighborhoods. On the other hand, Hamidiye neighborhood consists of only new constructions and is preferred by high socio-economic segments. Therefore, the price range in the relevant region was narrower.

5 Results

In clustering phase, an adoption of X-Means Clustering was used to find the optimum number of clusters. To avoid content loss of the discrete variables, DILCA distance was applied. All variables except the unit price was included in this phase. Since the hybrid model makes price predictions, this variable was not included in the clustering step to avoid any bias. As a result, three clusters were formed.

The distribution of the neighborhoods among clusters reflected the exact geographical boundaries. This distribution is given in Figure 4.

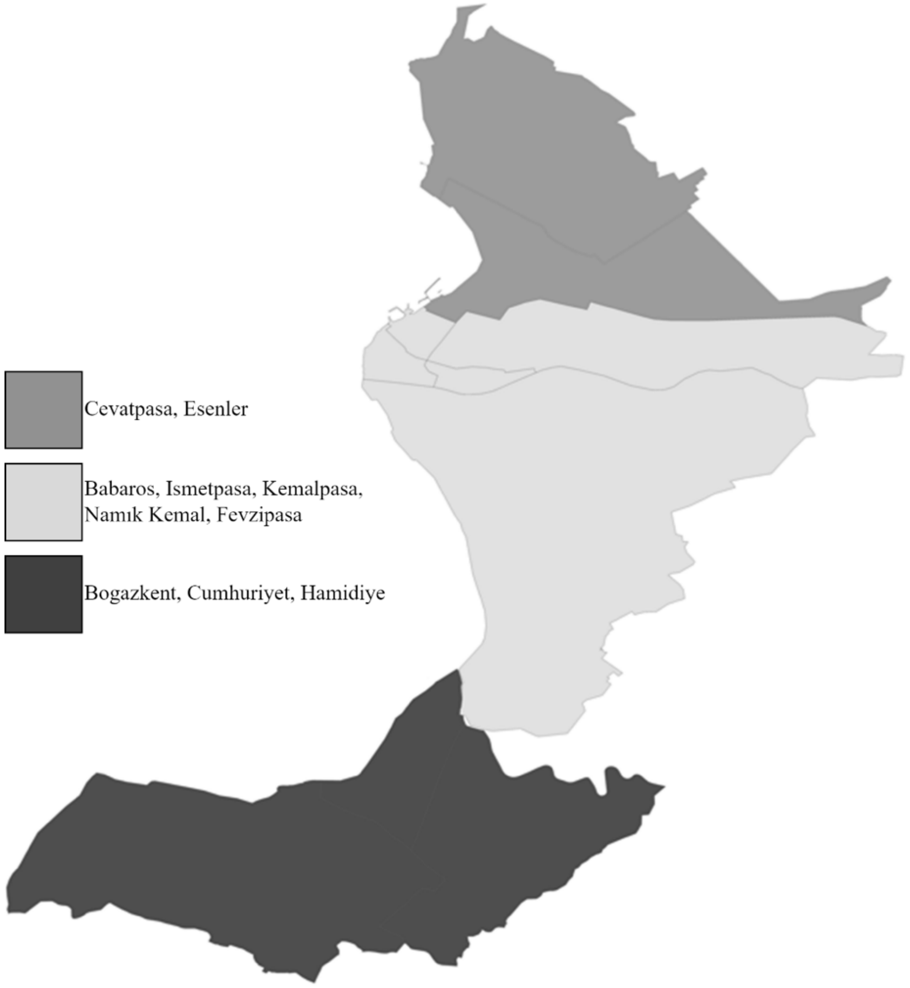
Cevatpasa, Esenler
Babaros, Ismetpasa, Kemalpasa, Namık Kemal, Fevzipasa
Bogazkent, Cumhuriyet, Hamidiye

Figure 4: The distribution of the neighborhoods among clusters

The first cluster covers Hamidiye, Cumhuriyet and Bogazkent neighborhoods. Cluster- 2 covers Cevatpasa and Esenler, Cluter-3 covers Barbaros, Ismetpasa and the united districts (Fevzipasa, Namık Kemal, Kemalpasa).

In Section 3.1, it was assumed that the data in the clusters would be normally distributed. To test this assumption, Quantile-Quantile (Q-Q) plot was used. Q-Q plots are widely used to prove the distribution of data [Yuan et al., 2021]. In this method, the theoretical and sample quantiles are compared. If they match to form a straight line, the data is normally distributed. In the case of convex curve, data is right skewed and data is left skewed if the curve is concave [Marimuthu et al., 2022]. A comparative Q-QP plot for the eleven variables in each cluster are given in Figure 5 to Figure 8.

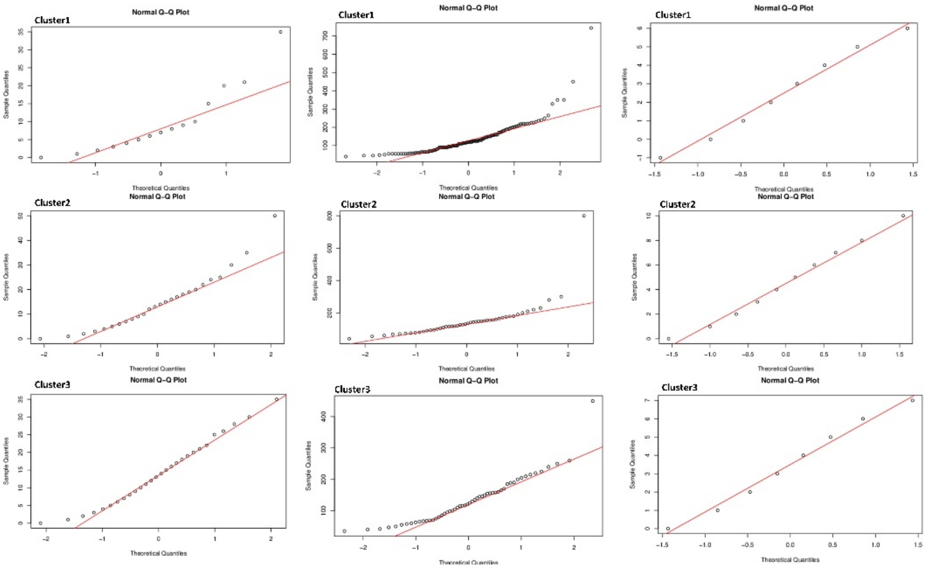

Figure 5: Cluster-based q-q plot for Age - Area - Current Floor variables. Rows indicate clusters and columns indicate related variables respectively.

In general, it can be said that all variables are close to normal distribution (Figure 5). Although the Area variable follows a normal distribution for Cluster-2, there is a slight convex slope in Cluster-1. Age and Current-Floor variables showed a diagonal distribution.

Deed, Facade and Floor-Number variables were also tested and given in Figure 6.

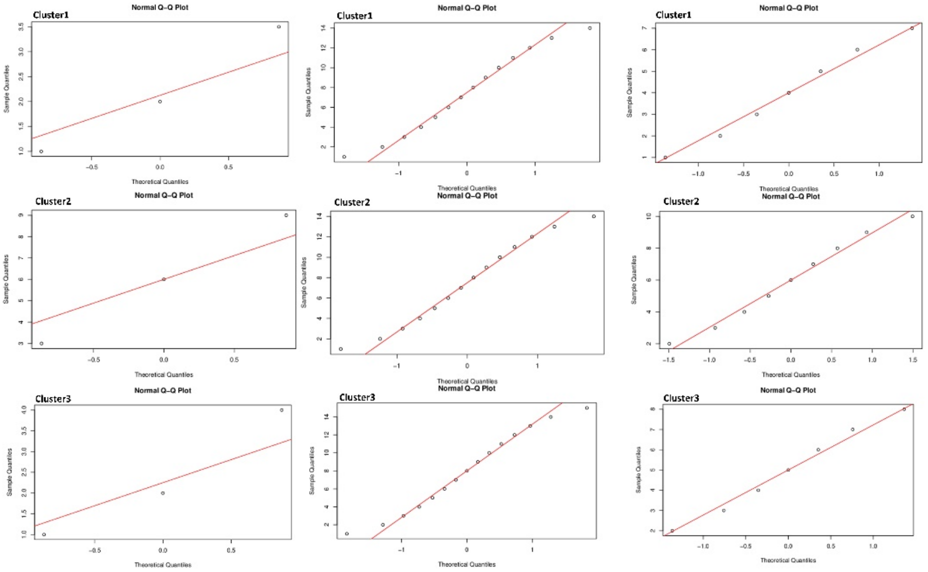

Figure 6: Cluster-based q-q plot for Deed-Façade-FloorNumber variables. Rows indicate clusters and columns indicate related variables respectively

Figure 6 indicates all three attributes (Deed-Façade-FloorNumber) are normally distributed. However, for Deed attribute some values diverges from theoretical quantiles, which can be interpreted as the outliers. For performance comparisons, these outliers were removed from the data set. However, no significant gain was achieved. The same can be concluded from Figure 7 for the Fuel-Type attribute for all three clusters.

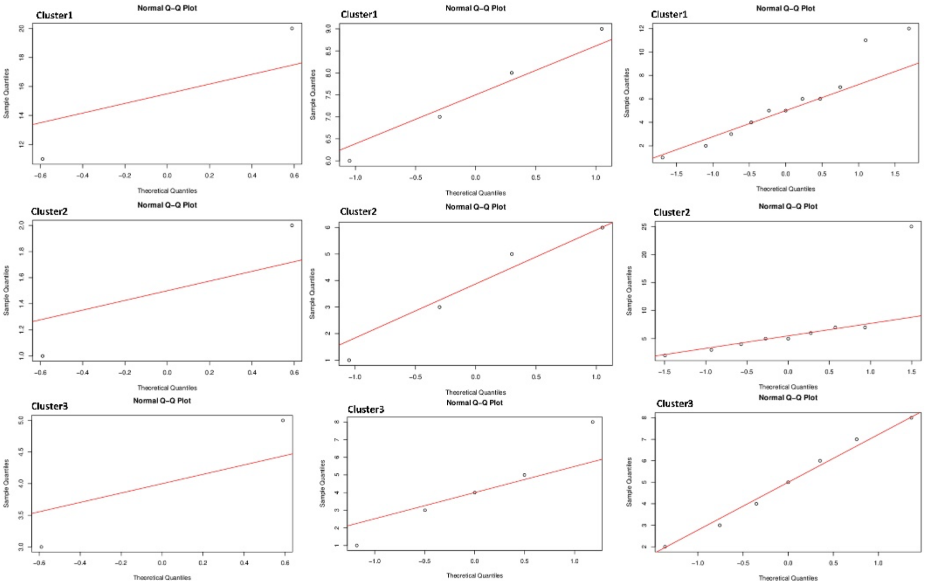

Figure 7: Cluster-based q-q for FuelType – HetaingType - NumberOfRooms variables. Rows indicate clusters and columns indicate related variables respectively

Almost identical distribution was observed for Heating-Type variable in Cluster-1 and Cluster-2. The distribution of the NumberOfRooms variable in Cluster-2 and Cluster-3 is very close to normal. However, there are some outliers in Cluster-1 for the given variable.

The last variables tested are the Unit Price and Residential Type. Their plot is given in Figure 8.

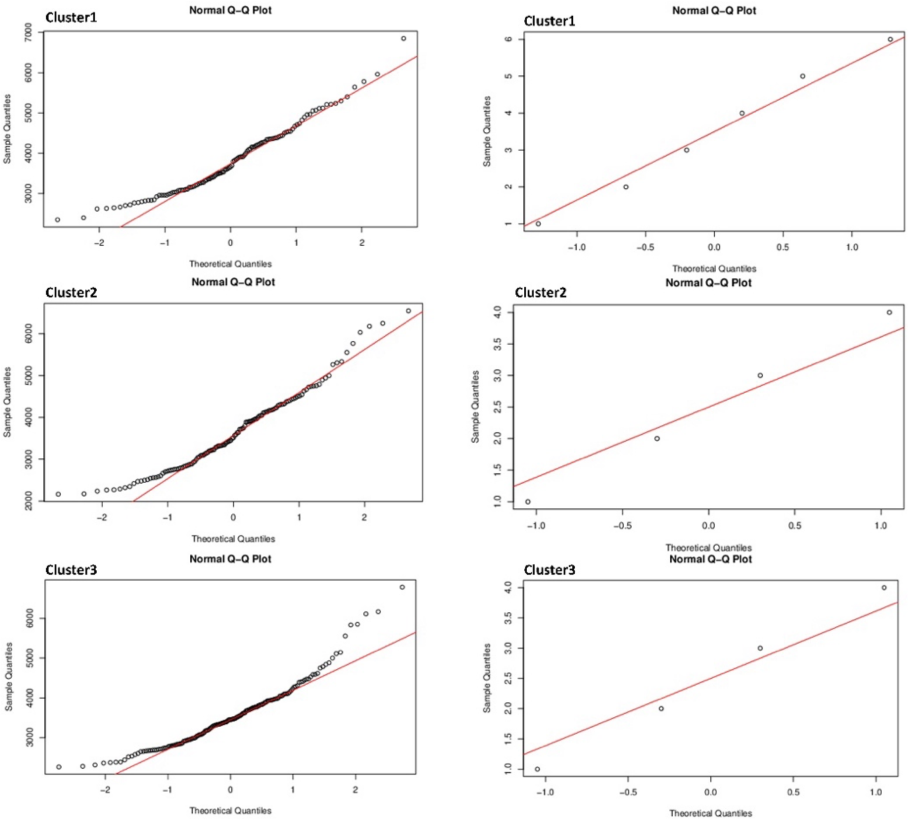

Figure 8: Cluster-based q-q plot for UnitPrice and ResidentialType variables. Rows indicate clusters and columns indicate related variables respectively.

As Figure 8 shows, Unit Price converges to a normal distribution. When outliers are removed, it is seen that the relevant variable is normally distributed for all three clusters. For the Residential Type variable, it is seen that there is a normal distribution for all three clusters.

The distribution graphs of the variables in the clusters (Figure 5 to figure 8) show that almost all variables are normally distributed or close to normally distributed.

In prediction phase, a CART model was established for each cluster. All CART models had the neighborhood attribute at the root of the tree. However, the variables that affect the price varied for each neighborhood. The CART models formed for each cluster are given in the following subsections.

Since the resulting CART models were too large to fit into a single figure, they were partitioned and given as sub-figures. For each cluster, the first two levels containing the root node was given in a single figure. Sub-branches were given as separate figures. The organizations of the figures are as follows:

CART of Cluster-1: Root and second level was given in Figure 9, sub-trees were given form Figure 10 to Figure 12.

CART of Cluster-2: Root and second level was given in Figure-13, sub-trees were given in Figure 14 and Figure 15.

CART of Cluster-3: Root and second level was given in Figure 16, sub-trees were given from Figure 17 to Figure 19.

Tree structures of the clusters were outlined by the figures from 10 to 18. Since the full description of each figure would be too long, only the important findings were given under the relevant paragraphs of the figures. However, one branch of Figure 10 was described from root to leaf as an example of how the entire figure can be interpreted. Similar inferences for other figures can be made by tracing the branches from root to leaves.

5.1 CART Model for Cluster 1

The first cluster consisted of neighborhoods in Kepez district, which is far from city center. When the CART model was established for this cluster, the "neighborhood" attribute was located at the root of the tree. The top branching is given in the Figure 9.

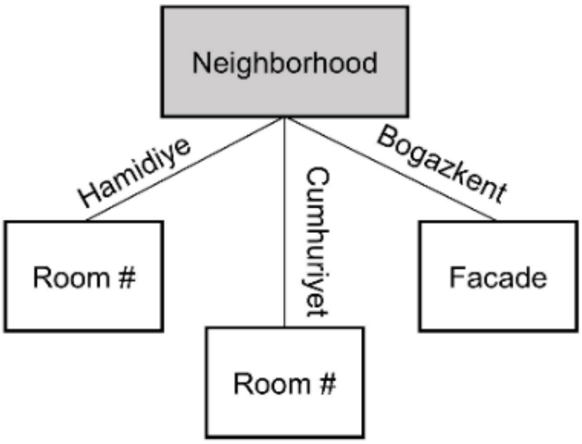

Figure 9: The top branching of CART for Cluster 1

The actual size of the regression tree does not fit into a single figure. Thus, sub-trees of each neighborhood are given separately.

The residential in Hamidiye neighborhood are newer when compared to other districts. Thus, the variation among attributes is much higher. Regression tree, given in Figure 10, reflected this situation.

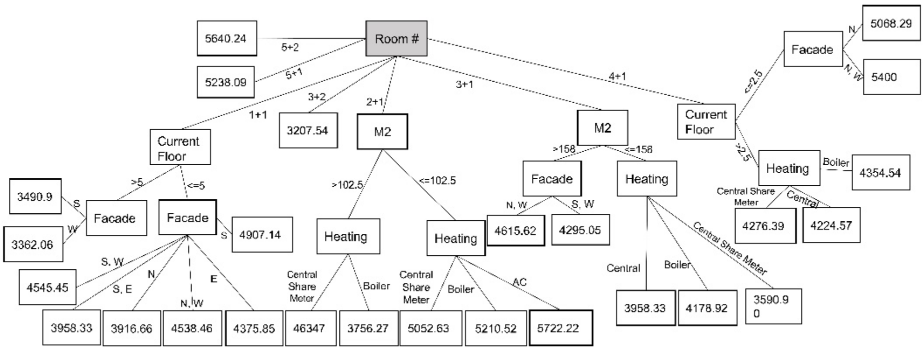

Figure 10: CART for Hamidiye neighborhood

Figure 10 indicates that the most important variables that affected the price were number of rooms, area of the residential in m^2 and current floor. Other variables were heating and facade. In Figure 6, when the "Number of Rooms" attribute takes the value "1+1", "Current Floor" is checked as the next attribute. When this attribute is greater than 5, the decision is made according to "Facade" attribute. In this case, if the “Facade” value is “South”, the unit price is predicted as 3,490 TL, and if the “Facade” value is “West”, the unit price is predicted as 3,362.06 TL. When the "Number of Rooms" attribute takes the value "2+1", the "Area (m^2)" attribute is checked. If this attribute is larger than 102.5 m^2, prediction is made based on the values of the “Heating” attribute. If “Heating” is “central share meter”, the price is predicted as 4,6347 TL, if it is “boiler” the prediction is 376 TL. Similarly, by following the other branches of the tree from top to bottom, the predictions of the whole structure can be deduced in the form of rules.

The CART model for Cumhuriyet neighborhood is given in Figure 11.

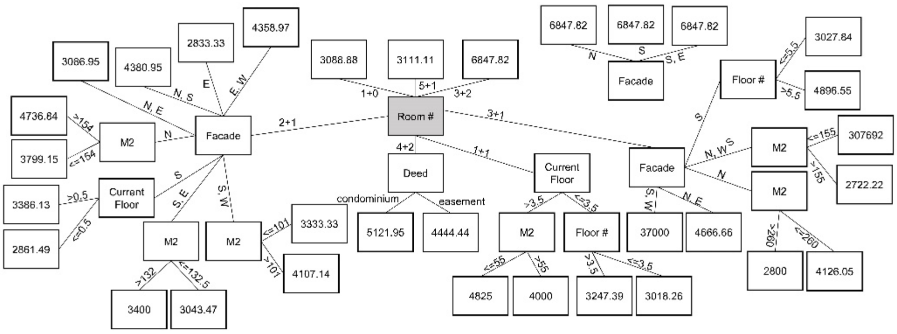

Figure 11: CART for Cumhuriyet neighborhood

The most important attribute for the unit price prediction in Cumhuriyet neighborhood was the number of rooms. This attribute was followed by deed, facade and current floor.

Since the attribute "facade" was located at the root of the tree, it was the most important attribute for price prediction in Bogazkent neighborhood. The CART model for this neighborhood is given in Figure 12.

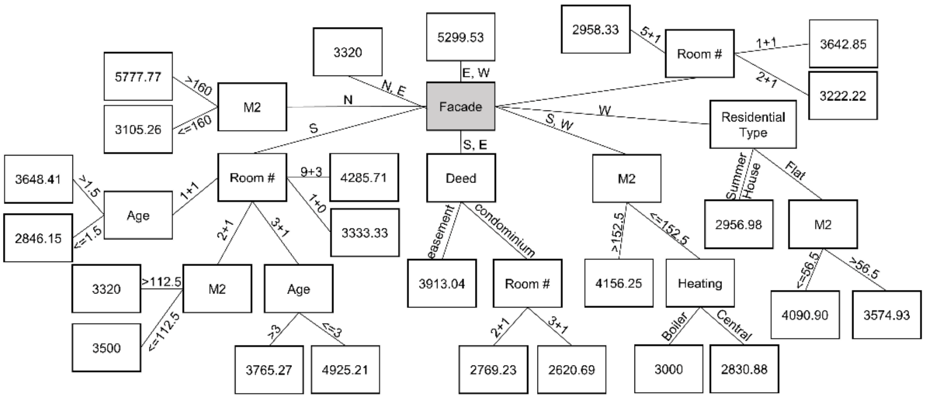

Figure 12: CART for Bogazkent neighborhood

The attributes in the second level of the CART were number of rooms, area, deed and residential type. Heating and building age were assigned to lower levels of the CART.

5.2 CART Model for Cluster 2

Cluster-2 covered Cevatpasa and Esenler neighborhoods. The most discriminative attribute for this cluster was also neighborhood as given in the Figure 13.

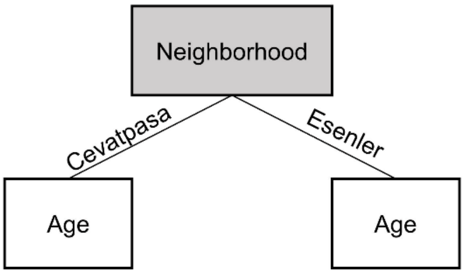

Figure 13: CART for Cluster 2

For both neighborhoods, age of the building was the second important attribute for unit price prediction.

Cevatpasa covers a large area with both new and old settlements. Therefore, age of the building was found as the most important attribute affecting the price. The sub-tree for Cevatpasa is given in Figure 14.

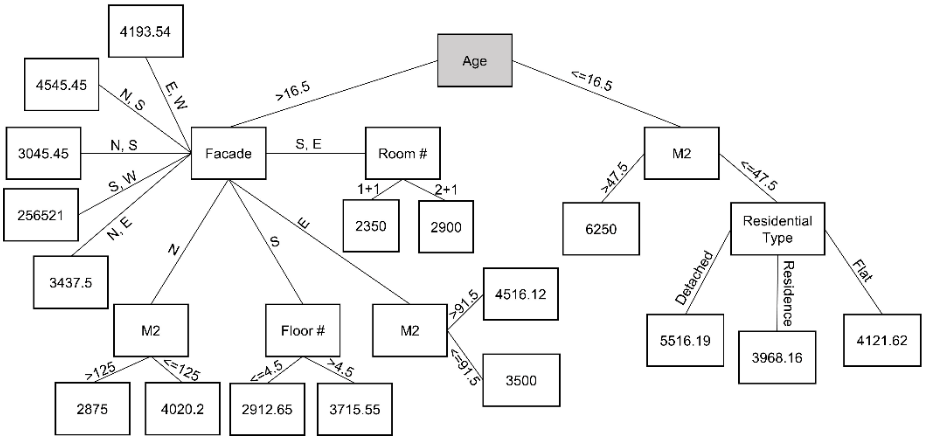

Figure 14: CART for Cevatpasa neighborhood

For old settlements, price prediction was based on the attributes facade, area and number of floors. For new settlements, area and residential type were used.

Esenler neighborhood, like Cevatpasa, covers both old and new settlements. Thus, age attribute was assigned to the root of the CART. The sub-tree for Esenler is given Figure 15.

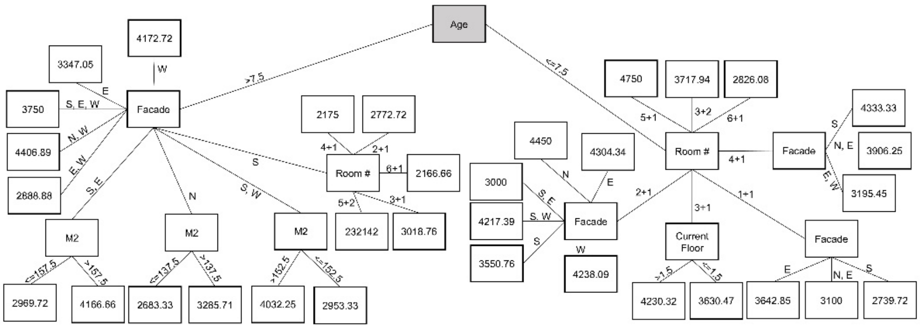

Figure 15: CART for Esenler neighborhood

Facade and number of rooms formed the branches at the second level of the tree. Branching of the third level was based on area and current floor attributes.

5.3 CART Model for Cluster 3

Cluster 3 consisted of Babaros, Ismetpasa and united districts. Neighborhood attribute was assigned to the root of the tree. Unlike the CART models of other clusters, each branching after the root had a different attribute. The top level branching for the Cluster 3 is given in Figure 16.

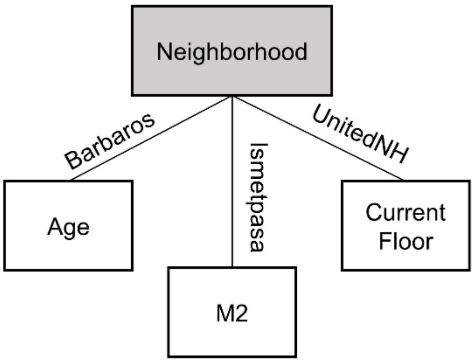

Figure 16: CART for Cluster 3

The sub-tree for Barbaros starts with the age attribute. CART model of this neighborhood is given in the Figure 17.

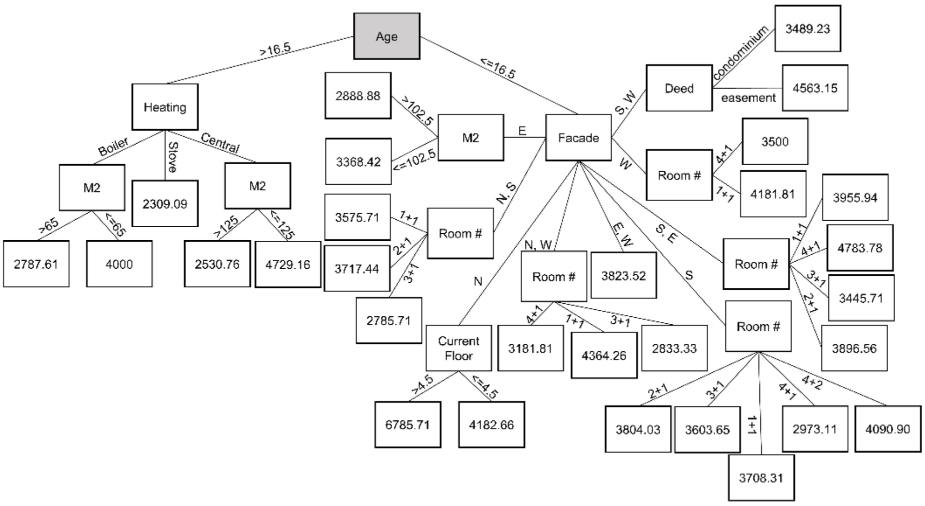

Figure 17: CART for Barbaros neighborhood

The nodes in the second level of the tree were composed of heating and facade attributes.

The unit price prediction in Ismetpasa neighborhood was strongly related to the area of the residential as given in Figure 18.

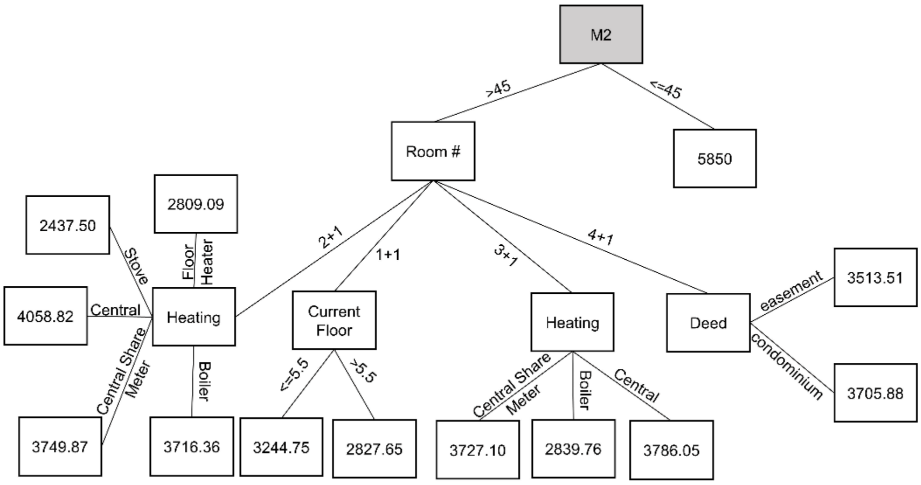

Figure 18: CART for Ismetpasa neighborhood

If the area of a given property is equal to or less than 45m^2, the unit price was predicted as 5,850 TL without further branching. In other cases the prediction was based on the number of rooms, heating, current floor and deed attributes.

The districts of Fevzipasa, Namık Kemal and Kemalpasa are close to each other. The number of residential sales were not enough for each of these neighborhoods. Thus, they were combined to form a single district. The CART model of the united neighborhoods is given in the Figure 19.

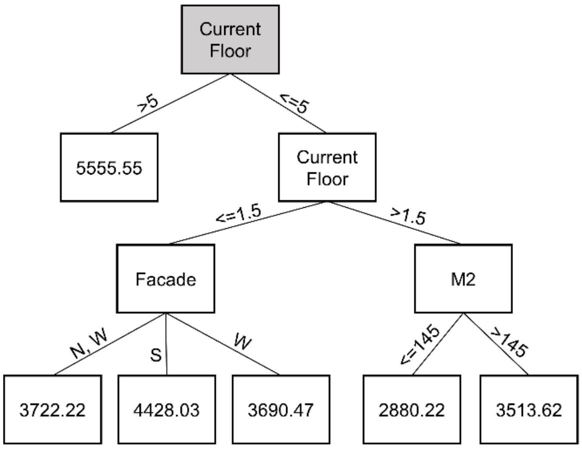

Figure 19: CART for united neighborhoods

For the united neighborhoods, current floor, facade and area attributes were used for unit price prediction.

5.4 Summary of CART Models

The hybrid model divided the dataset into three clusters. Neighborhoods within each cluster are composed of neighborhoods that are geographically close to each other (Figure 4). CART models were established for each cluster. All CARTs started with the neighborhood attribute. However, other attributes varied for each CART. The important attributes in each CART model are given in Figure 20.

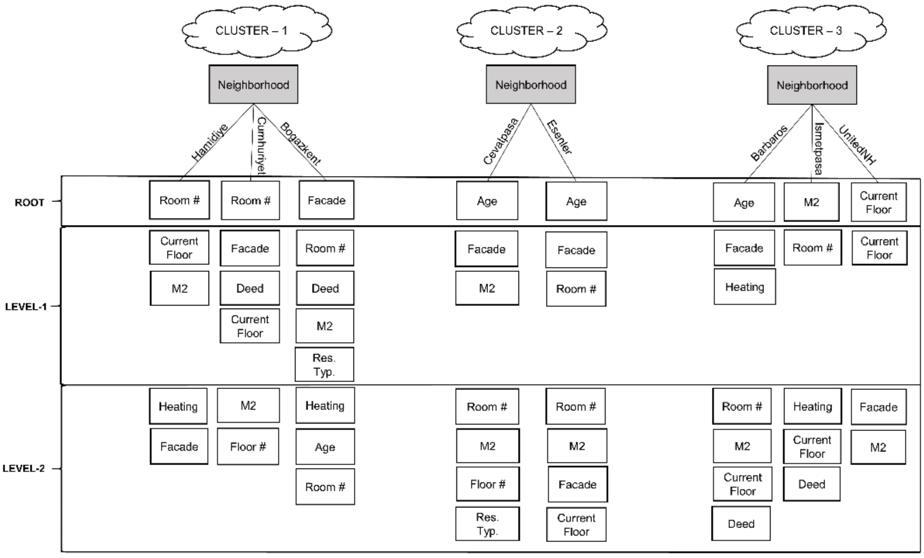

Figure 20: The general representation of CARTs for each cluster. The attributes affecting the price are listed in order of importance. The importance decreases from root to lower levles

The CARTs were formed to reflect the characteristics of each neighborhood. The Hamidiye neighborhood in Kepez is a region with new constructions. Citizens of high socio-economic groups generally prefer this neighborhood. People in this group prefer larger houses. In accordance with this situation CART model for this neighborhood found the number of rooms attribute as the most important variable. Cevatpasa and Esenler covers both new and old settlements. Thus, CART model of these neighborhoods, were based on building age attribute. Both models revealed that, the price of for older settlements were affected by facade. Number of rooms and area of the residential were the attributes that affect price of the new settlements. The income level in the combined neighborhoods (Fevzipasa, Namık Kemal and Kemalpasa) is lower than in other regions. Therefore, the residential in this region have more standardized structure when compared to other regions. CART models also reflected this situation. The models for other neighborhoods were more complex in structure; however, the model for united neighborhood was much simpler. The variables that affect the price for this neighborhood were current floor, facade and area of the flat.

The importance of the attributes within each cluster was also calculated and given in Figure 21. Clusters were not directly used for price prediction. Instead, a CART model was established for each cluster. The importance of the attributes in the CART of each cluster was calculated relatively to branching logic of the CART.

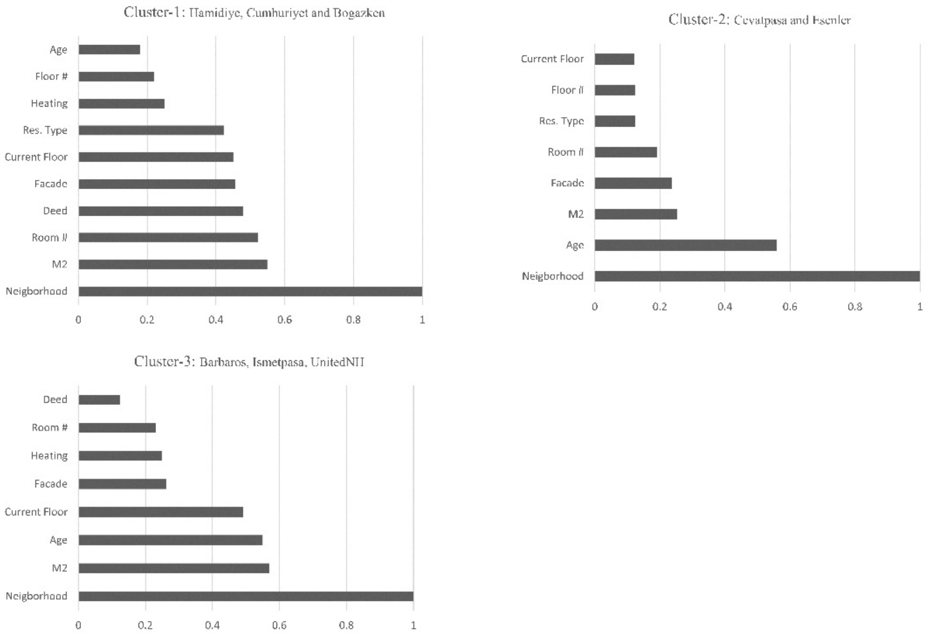

Figure 21: Variable importance of attributes in CART models for each cluster

For all clusters the most important attribute was “Neighborhood”. It was fallowed by “Area (m^2), “Number of Rooms (Rom #)” and “Deed” in Cluster – 1. The most important attributes of the second cluster were “Age”, Area (m^2) and “Facade”. In the third cluster, the most important attributes were ranked as “Area (m^2)”, “Age” and “Current Floor”.

When Figure 21 was examined, it was seen that the importance of attributes was different in each cluster. For example, "Age" was the least important attribute for the first cluster, while it is the most important attribute after "neighborhood" in the second cluster. If a single decision model were used, the prediction performance would also decrease, as it would be difficult to model extreme cases like the above. However, with the hybrid approach, the differences between regions was revealed and it was shown that high prediction performance could be achieved.

6 Performance Comparison

The prediction performance of the hybrid model was compared with DC, which is the gold standard in real estate price prediction. For this comparison, a separate dataset (not previously used in training and testing) was used. This dataset had identical features as the original.

The prediction of the unit price, TL per square meter (TL/m2), is a multiclassification task. However, price tag is continuous that makes prediction task a multiple regression. For this reason, models were compared in terms of RMSE, MAPE and adjusted R^2 metrics. Related metrics were calculated separately for specific CARTs established for each cluster. Afterwards, they were averaged to show the overall performance of the model. Comparison result is given in Table 3.

	RMSE (TL/m2)		MAPE (TL/m2)		Adjusted R^2	
	Train	Validation	Train	Validation	Train	Validation
Hybrid Model	71.249	72.867	0.0053	0.0057	0.982	0.978
Direct Capitalization	382.12	374.5	0.087	0.0909	0.837	0.797

Table 3: The performance comparison of the hybrid model and direct capitalization

Diebold-Mariano Test [Diebold and Mariano, 1995] was used to check if the prediction performances of DC and Hybrid model were significantly different. For this purpose the order of h taken as four. The result of the test $p < 0.05$. Thus, the prediction performance of the models was significantly different. The prediction performance of the Hybrid model was superior to the DC in terms of RMSE, MAPE and R^2 metrics. Considering that DC is used as a main standard in price prediction, the potential of the hybrid model in the real estate domain is quite strong.

One of the important contributions of this work is the unsupervised learning used in the hybrid model. Clustering was used to reveal the differences of properties. In this way, different variables affecting the price for each cluster were revealed. In addition, it was predicted that the clustering approach should increase the prediction performance. A CART model without clustering was established and its prediction performance was compared with the hybrid model. To avoid any bias, hyper-parameters of the single CART Model was optimized by evolutionary HPO. As a result, the depth, minimum leaf size and minimum size for split hyper-parameters for the single CART Model were found to be 13, 2 and 2 respectively. The comparison is given in Table 4.

Prediction Models	RMSE (Train – Validation)	
Hybrid Model	71.249	72.867
Single CART Model	105.720	112.983

Table 4: The performance comparison of the hybrid model and single CART to reveal the effect of clustering on prediction performance

The RMSE metric for the single CART is 112.983. This metric proves two points. First, even a single CART can perform better than the DC. Thus, use of machine learning approaches in real estate price prediction is promising. Second and more importantly, the use of unsupervised learning increased the prediction performance by 36%.

7 Conclusion

The use of ML in real estate price prediction gained speed in recent years, as it eliminates the disadvantages of DC such as expert opinion and manual analysis of data. However, majority of ML studies use only supervised learning and focus on prediction performance, ignoring the contribution of unsupervised learning on performance and detailed analysis of the variables that affect price.

In this study, a hybrid model of CART and X-Means was established to combine supervised and unsupervised learning. Clustering was used to determine the geographic, socio economic and other differences in the dataset. A specific CART model was constructed for each cluster. In this way, the prediction performance was increased and the variables that affect the price were examined in more detail. The hybrid model outperforms DC in terms of RMSE, MAPE and adjusted R^2. The RMSE of the hybrid model was 72.867 while for DC it was 374.5. The contribution of clustering on performance was also analyzed. X-Means clustering increased the performance of the model by 36%.

In addition to prediction performance, the hybrid model was successful in revealing the variables that affect the price. The branching of each CART model started with the neighborhood variable. Then variables specific to each neighborhood came to the fore. The most dominant variables over the price were the number of rooms, facade, age, current floor and area.

This study was carried out to predict the price based on the physical properties of the property. In order to increase the prediction performance, a hybrid model consisting of un-supervised and supervised learning was established. In addition, the features that affect the price were retrieved from the hybrid model. The prediction performance of the hybrid model was quite high. The variables affecting the price successfully reflected socio-economic conditions and geographical regions. These results show the importance of using ML models in real estate.

8 Future Work

This model is designed as a pioneer of future studies. As a future study, we aim to expand the study area to include different provinces and develop more comprehensive model by including economic, non-numerical and spatial variables. In addition, we plan to change the methods to be used in the model. The dataset covering many provinces in Turkey is quite large, so it is planned to use deep learning methods. Since the examples of some provinces are very few, it is also planned to provide synthetic data to establish more balanced datasets. For this purpose, it is planned to use Generative Adversarial Network (GAN) based methods.

When the studies were evaluated, it was seen that variables such as economic, spatial and social variables (non-numeric), had an effect on the price. Being aware of this, we tried to include economic variables in our study. However, the data set was established for three-month period, thus, it was not sufficient to reveal the effect of economic variables.

For this reason, our first aim as a future study is to retrieve the dataset for a long period and to show the effect of economic variables. Another aim is to establish the dataset to include different provinces representing different geographical regions of Turkey. In this way, the effect of socio-economic variables as well as the physical properties of the property will be examined and the results will be generalized for all provinces in Turkey. We are working to achieve our future goals. The web scraper has been retrieving the real estate ads for the selected provinces and combining them with daily economic data.

References

[Abidoye and Chan, 2017] Abidoye, R. B., Chan, A. P. C.: “Critical Review of Hedonic Pricing Model Application in Property Price Appraisal: A Case of Nigeria”; International Journal of Sustainable Built Environment, 6, 1, (2017), 250–259. https://doi.org/10.1016/j.ijsbe.2017.02.007.

[Abidoye and Chan, 2018], Abidoye, R. B., Chan, A. P. C.: “Improving Property Valuation Accuracy: A Comparison of Hedonic Pricing Model and Artificial Neural Network”; Pacific Rim Property Research Journal, 24, 1, (2018), 71–83. https://doi.org/10.1080/14445921.2018.1436306.

[Ahmad and Khan, 2019] Ahmad, A., Khan, S. S.: “Survey of State-Of-The-Art Mixed Data Clustering Algorithms”; IEEE Access, 7, (2019), 31883–31902. Doi: 10.1109/ACCESS.2019.2903568.

[Akossou et al., 2013] Akossou, A. Y. J., Palm, R.: “Impact Of Data Structure on the Estimators R-Square and Adjusted R-Square in Linear Regression”; Int. J. Math. Comput, 20, 3, (2013), 84-93.

[Battaglia et al., 2021] Battaglia, E., Celano, S., Pensa, R.: “Differentially Private Distance Learning in Categorical Data”; Data Mining and Knowledge Discovery, 35, 5, (2021), 2050–2088. https://doi.org/10.1007/s10618-021-00778-0

[Beal et al., 2019] Beal, L. R., Norman, T., Ramchurn, S.: “Artificial Intelligence for Team Sports: A survey”; The Knowledge Engineering Review, 34, 2019, E28. Doi:10.1017/S0269888919000225.

[Cerda and Varoquaux, 2022] Cerda, P., Varoquaux, G.: “Encoding High-Cardinality String Categorical Variables”, IEEE Transactions on Knowledge and Data Engineering, 34, 3, (2022), 1164 – 1176. Doi: 10.1109/TKDE.2020.2992529

[Chicco et al., 2021] Chicco, D., Warrens, M. J., Jurman, G.: “The Coefficient of Determination R-Squared is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation”; PeerJ Computer Science, 7, e623, (2021). https://doi.org/10.7717/peerj-cs.623

[Diebold and Mariano, 1995] Diebold, F. X., Mariano, R. S.: "Comparing Predictive Accuracy"; Journal of Business and Economic Statistics, 13, 3, (1995), 253-263. doi: 10.1080/07350015.1995.10524599

[Erhan et al., 2009] Erhan, D., Manzagol, P., Bengio, Y., Bengio, S., Vincent, P.: "The Difficulty of Training Deep Architectures and The Effect of Unsupervised Pretraining"; Proc. of the 12th International Conference on Artificial Intelligence and Statistics, PMLR Press, Florida (2009), 153–160.

[Fernandez et al., 2013] Fernandez, L., Cutter, B., Sharma, R., Scott, T.: "Land Preservation Policy Effect or Neighborhood Dynamics: A Repeat Sales Hedonic Matching Approach"; Journal of Environmental Economics and Management, 88, (2013), 311–326. https://doi.org/10.1016/j.jeem.2018.01.001.

[Fong and Hong, 2021] Fong, A. C. M., Hong, G.: "Boosted Supervised Intentional Learning Sup- Ported by Unsupervised Learning"; International Journal of Machine Learning and Computing, 11, 2, (2021), 98–102. Doi: 10.18178/ijmlc.2021.11.2.1020.

[Garg and Mago, 2018] Garg, A., Mago, V.: "Role of Machine Learning in Medical Research: A Survey"; Computer Science Review, 40, 7, (2021), 100370. https://doi.org/10.1016/j.cosrev.2021.100370.

[Hancock and Khoshgoftaar, 2020] Hancock, J. T., Khoshgoftaar, T. M.: "Survey on Categorical Data for Neural Networks"; Journal of Big Data, 7, 1, (2020), 1-41. https://doi.org/10.1186/s40537-020-00305-w

[Ho et al., 2021] Ho, W. K. O., Tang, B. S., Wong, S. W.: "Predicting Property Prices with Machine Learning Algorithms", Journal of Property Research, 38, 1, (2021), 48–70. https://doi.org/10.1080/09599916.2020.1832558

[Hodson, 2022] Hodson, T. O.: "Root-Mean-Square Error (RMSE) or Mean Absolute Error (MAE): When to Use Them or Not"; Geoscientific Model Development, 15, 14, (2022), 5481-5487. https://doi.org/10.5194/gmd-15-5481-2022.

[Jain et al., 2018] Jain, V., Phophalia, A., Bhatt, J. S.: "Investigation of a Joint Split- Ting Criteria for Decision Tree Classifier Use of Information Gain and Gini Index"; Proc. TENCON 2018 - 2018 IEEE Region 10 Conference, IEEE, Jeju Island (2018), 2187–2192. Doi: 10.1109/TENCON.2018.8650485.

[Lan et al., 2018] Lan, K., Wang, D. T., Fong, S., Liu, L. S. Wong, K., Dey, N.: "A Survey of Data Mining and Deep Learning in Bioinformatics"; Journal of Medical Systems, 42, 8, (2018), 139. Doi: 10.1007/s10916-018-1003-9.

[Lee, 2021] Lee, C.: "Predicting Land Prices and Measuring Uncertainty by Combining Supervised and Unsupervised Learning"; International Journal of Strategic Property Management, 25, 2, (2021), 169–178. Doi: 10.3846/ijspm.2021.14293.

[Lennhoff, 2011] Lennhoff, D. C.:" Direct Capitalization: It Might Be Simple But It Isn't That Easy"; Appraisal Journal, 79, 1, (2011).

[Leung et al., 2020] Leung, K. H., Mo, D. Y., Ho, G. T. S., Wu, C. H., Huang, G. Q.: "Modelling Near-Real-Time Order Arrival Demand in E-Commerce Context: A Machine Learning Predictive Methodology"; Industrial Management Data Systems, 120, 6, (2020) 1149–1174. Doi: 10.1108/IMDS-12-2019-0646

[Li and Chu, 2017] Li, L., Chu, K.: "Prediction of Real Estate Price Variation Based On Economic Parameters", Proc. 2017 International Conference on Applied System Innovation (ICASI)"; IEEE, Sapporo (2017), 87–90. Doi: 10.1109/ICASI.2017.7988353

[Liu et al., 2018] Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., Leung, V. C. M.: "A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View"; IEEE Access, 6, (2018), 12103–12117. Doi: 10.1109/ACCESS.2018.2805680.

[Macpherson and Sirmans, 2001] Macpherson, D. A., Sirmans, G. S.: "Neighborhood Diversity and House Price Appreciation"; The Journal of Real Estate Finance and Economics, 22, 1, (2001), 81–97. https://doi.org/10.1023/A:1007831410843.

[Manasa et al., 2020] Manasa, J., Gupta, J. R., Narahari, N. S.: "Machine Learning Based Predicting House Prices Using Regression Techniques", Proc. 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)"; IEEE Xplore, Bangalore (2020), 624-630. Doi: 10.1109/ICIMIA48430.2020.9074952.

[Marimuthu et al., 2022] Marimuthu, S., Mani, T., Sudarsanam, T.D., George, S., Jeyaseelan, L.: "Preferring Box-Cox Transformation, Instead Of Log Transformation to Convert Skewed Distribution of Outcomes to Normal in Medical Research"; Clinical Epidemiology and Global Health, 15, (2022). https://doi.org/10.1016/j.cegh.2022.101043.

[Michaletz and Artemenkov, 2018] Michaletz, V. B., Artemenkov, A.: "The Transactional Assets Pricing Approach and Income Capitalization Models in Professional Valuation: Towards a Quick Income Capitalization Format", De Gruyter , 26, 1, (2018), 89–107. Doi: 10.2478/remav-2018-0008

[Mienye et al., 2019] Mienye, I. D., Yanxia, S., Wang, Z.: "Prediction Performance of Improved Decision Tree Based Algorithms: A Review"; Procedia Manufacturing, 35, (2019), 698– 703. https://doi.org/10.1016/j.promfg.2019.06.011.

[Miles, 2005] Miles, J.: "R-squared, adjusted R-squared"; Encyclopedia of statistics in behavioral science". https://doi.org/10.1002/0470013192.bsa526.

[Mohd et al., 2020] Mohd, T., Jamil, N.S., Johari, N., Abdullah, L.: "An Overview of Real Estate Modelling Techniques for House Price Prediction"; Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-3859-9_28.

[Ozbayoglu et al., 2020] Ozbayoglu, A. M., Gudelek, M. U., Sezer, O. B.: "Deep Learning for Financial Applications : A Survey"; Applied Soft Computing, 93, (2020), 106384. https://doi.org/10.1016/j.asoc.2020.106384.

[Patibandla and Veeranjaneyulu, 2008] Patibandla, R. S. M. L., Veeranjaneyulu, N.: "Survey on Clustering Algorithms for Unstructured Data"; Springer, Singapore (2008). https://doi.org/10.1007/978-981-10-7566-7_41

[Pelleg and Moore 2000] Pelleg, D., Moore, A. W.: "X-means: Extending k-means with efficient estimation of the number of clusters"; Icml, 1, (2000), 727-734

[Phan, 2018] Phan, T. D.: "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia"; Proc. International Conference on Machine Learning and Data Engineering (iCMLDE), IEEE, Sydney (2018), 35–42. Doi: 10.1109/iCMLDE.2018.00017.

[Pınar and Demir, 2014] Pınar, A., Demir, M.: "Konut Sektorunde Kapitalizasyon Oranlarını Belirleyen Faktorler: Turkiye Icin Bir Mikroveri Analizi", Sosyoekonomi , 22, 22, (2014), 386–398. Doi: 10.17233/se.93073.

[Quinlan, 1996] Quinlan, J. R.: "Improved Use of Continuous Attributes in C4.5", Journal of Artificial Intelligence Research, 4, (1996), 77–90. doi/10.5555/1622737.1622742.

[Raykov et al., 2016] Raykov, Y. P., Boukouvalas, A., Baig, F., Little, M. A.: "What To Do When K-Means Clustering Fails: A Simple Yet Principled Alternative Algorithm"; PLOS ONE, 11, 9 (2016), e0162259. Doi: 10.1371/journal.pone.0162259.

[Varma et al., 2018] Varma, A., Sarma, A., Doshi, S., Nair, R.: "House Price Prediction Using Machine Learning and Neural Networks", Proc. Second International Conference on Inventive Communication and Computational Technologies (ICICCT), IEEE, Coimbatore (2018), 1936–1939. Doi: 10.1109/ICICCT.2018.8473231.

[Wang and Wu, 2018] Wang, C., Wu, H.: "A New Machine Learning Approach to House Price Estimation", New Trends in Mathematical Science, 6, 4, (2018), 165–171. http://dx.doi.org/10.20852/ntmsci.2018.327.

[Willmott et al., 2009] Willmott, C. J., Matsuura, K., Robeson, S. M.: "Ambiguities Inherent in Sums-Of-Squares-Based Error Statistics";. Atmospheric Environment, 43, 3, (2009), 749-752. https://doi.org/10.1016/j.atmosenv.2008.10.005.

[Xavier et al., 2013] Xavier, J. C. , Canuto, A. M. P., Almeida, N. D., Goncalves, L. M. G.; "A Comparative Analysis of Dissimilarity Measures for Clustering Categorical Data"; Proc. The 2013 International Joint Conference on Neural Networks (IJCNN), IEEE, Dallas (2013), 1–8. Doi: 10.1109/IJCNN.2013.6707039

[Yalcin et al., 2018] Yalcin, G., Selcuk, O., Senturk, E.: "Bursa Ili Mustafakemalpasa Ilcesi Tarım Arazilerinde Kapitalizasyon Oranının Tespiti"; Afyon Kocatepe Universitesi Fen ve Muhendislik Bilimleri Dergisi, 18, 2, (2018), 548–560. Doi: 10.5578/fmbd.67386.

[Yang and Shami, 2020] Yang, L., Shami, A.: "On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice"; Neurocomputing, 415, (2020), 295-316. https://doi.org/10.1016/j.neucom.2020.07.061.

[Yuan et al., 2021] Yuan, K. H., Liu, H., Han, Y.:" Differential Item Functioning Analysis Without A Priori Information on Anchor Items: QQ Plots and Graphical Test"; Psychometrika, 86, (2021). https://doi.org/10.1007/s11336-021-09746-5

[Yucebas et al., 2022] Yucebas, S., Dogan, M., Genc, L.: "A C4.5 – Cart Decision Tree Model for Real Estate Price Prediction and The Analysis of The Underlying Features"; Konya Journal of Engineering Sciences, 10, 1, (2022), 147-161. https://doi.org/10.36306/konjes.1013833.

[Zhao et al. 2019] Zhao, Y., Chetty, G., Tran, D.: "Deep Learning With Xgboost for Real Estate Appraisal"; Proc. 2019 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, Xiamen (2019), 1396–1401. Doi: 10.1109/SSCI44817.2019.9002790.