Konya Mühendislik Bilimleri Dergisi, c. 10, s. 1, 147-161, 2022
Konya Journal of Engineering Sciences, v. 10, n. 1, 147-161, 2022
ISSN: 2667-8055 (Electronic)
DOI: 10.36306/konjes.1013833

A C4.5 – CART DECISION TREE MODEL FOR REAL ESTATE PRICE PREDICTION AND THE ANALYSIS OF THE UNDERLYING FEATURES

[1]Sait Can YÜCEBAŞ , [2]Melike DOĞAN , [3]Levent GENÇ

[1] *Canakkale Onsekiz Mart University, Faculty of Engineering, Computer Engineering, Canakkale, TÜRKİYE*
[2] *Laren Engineering Map Design, Mugla, TÜRKİYE*
[3] *Canakkale Onsekiz Mart University, Faculty of Architecture and Design, City and Region Planning, Canakkale, TÜRKİYE*
[1]can@comu.edu.tr, [2]larenharita@gmail.com, [3]leventgc@comu.edu.tr

(Geliş/Received: 23.10.2021; Kabul/Accepted in Revised Form: 23.01.2022)

ABSTRACT: The machine learning approaches are used in different domains for price prediction. Real estate price prediction comes to fore in recent years. However, most of the studies focus on the prediction performance and the factors affecting the price are often ignored. In this study, a C4.5 – CART model to predict the residential real estate prices is developed. This model is capable of predicting both numeric and categorical price for real estate properties. In addition, the factors affecting the price are reveled and analyzed in detail. The performance of the developed model is compared to Direct Capitalization model, which is used as a gold standard in the domain. Both models are tested on a dataset that includes updated real time data that is gathered by a web scraper. For numeric prediction, RMSE of the developed model is 13.169 and 358.69 for the Direct Capitalization model. KAPPA and accuracy is used for the categorical prediction. The model has 81% KAPPA and 88% accuracy.

Keywords: *Machine learning, Decision tree, C4.5, CART, Direct Capitalization*

Gayrimenkul Fiyat Tahmini ve Alttaki Özelliklerin Analizi İçin C4.5 – CART Karar Ağacı Modeli

ÖZ: Fiyat tahmini için makina öğrenmesi uygulamaları farklı alanlarda kullanılmaktadır. Gayrimenkul alanında fiyat tahmini son yıllarda ön plana çıkmaktadır. Ancak, çalışmaların büyük bölümü tahmin performansına odaklanmış olup fiyata etki eden faktörlerin incelenmesi göz ardı edilmiştir. Bu çalışmada gayrimenkul fiyat tahmin için bir C4.5 – CART ağacı modeli geliştirilmiştir. Bu model hem nümerik hem de kategorik fiyat tahmini yapabilmektedir. Ek olarak fiyata etki eden faktörler detaylıca analiz edilerek ortaya çıkarılmıştır. İlgili modelin performansı bu alanda bir altın standart olan Direkt Kapitalazyon modeli ile karşılaştırılmıştır. Her iki model web kazıyıcı tarafından elde edilen güncel gerçek zamanlı veri kümeleri üzerinde test edilmiştir. Nümerik tahmin için geliştirilen modelin kök ortalama kare hatası 13.169 iken Direkt Kapitalizasyon için 359,69 bulunmuştur. Kategorik tahmin için kesinlik ve KAPPA metrikleri kullanılmıştır. Modelin KAPPA sayısı %81 ve kesinlik değeri %88'dir.

Anahtar Kelimeler: *Makine öğrenmesi, Karar ağacı, C4.5, CART, Direkt kapitalizasyon*

1. INTRODUCTION

In today's volatile economy, determining the prices of properties by certain standards is very important for economic stability. The seller, the buyer and the intermediary stakeholders determine property prices for sales. Stakeholders offer prices within a certain range by evaluating the pros and cons of the real estate by comparing them to other similar sales. However, the absence of a standard for the pricing makes price determination difficult and leads to extremely low or extremely high-end values in the market.

It is quite clear that a model should be created to establish a standard for the pricing. This model should determine the most effective parameters on the price for different conditions. When the studies in the literature are examined, it is seen that different methods are used to determine the real estate prices.

In the hedonic model (HM), price estimation is based on regression calculations. The parameters that are thought to affect the price is evaluated extensively (Ward and Gleditsch, 2019). In addition, a specific weight constant is determined for each of these parameters (Ward and Gleditsch, 2019). An expert opinion is required for this weighting scheme (Mayer et al., 2019). The disadvantages of these methods are the need for vast amount of parameters to determine the correct price and the need for expert opinion to calculate the weight of parameters.

Apart from HM there are studies that use machine learning (ML) for the price estimation. When the studies in this domain are examined, we see that mostly Artificial Neural Networks (ANN) (Varma et al., 2018; Wang et al., 2019), Deep Learning (DL) (Piao et al., 2019) and Regression (Rg) (Madhuri et al., 2019) methods are used. There are also studies that compare the ML methods with each other (Truong, 2020; Phan, 2018; Park and Bae, 2015).

Most of the studies, whether ML or HM, focus on the performance of the model in price prediction. Some studies (He et al., 2021; Sawant el al., 2018) list the factors that affect the price; however, the relation between these factors are not analyzed as a whole in most of the studies. Another deficit is that these studies focus on numerical price and are far from making a categorical evaluation.

Based on these shortcomings, we aim to create a model with high prediction performance and can learn by itself without the need for expert opinion and shows the effect of the parameters on this prediction. In order to achieve these goals, a model that uses CART (Breiman et al., 1984) for numerical price prediction and C4.5 (Salzberg, 1994) for categorical price classification is developed. To reveal the performance of the model, it is compared to the Direct Capitalization (DC), which is a sub branch of HM. Specific web scrapper is designed to gather the real time data from the web. Both methods are run on the residential ads for sale in Cumhuriyet district in Canakkale.

2. RELATED WORK

Machine learning methods used for property price prediction can be grouped into regression, black box, and decision tree approaches. The paragraphs below, present the current machine learning studies for real estate price prediction in terms of the methods used, their advantages and drawbacks.

Most of the regression-based studies consider factors independently. They examine the relation between the price and a specific factor once at a time. (Rave et al., 2019) applied a linear regression-based approach to predict real estate prices. Although the focus of the study was price prediction, their main contribution was on big data regression.

(Wu et al., 2019) applied a similar model on spatiotemporal determinants. They used weights for geographical and temporal variables. With the use of weights, they stated that the regression model becomes more flexible in determining spatial and temporal variables.

(Manasa et al., 2020) compared the prediction performance of regression based models. Losso, ridge and linear regression models were compared in terms of error metrics and no significant difference was found. In general regression based methods can be used for numeric price prediction, however, they fail to perform well in categorical price prediction.

The black box approaches such as artificial neural networks (ANN) and its derivatives are also used in the related domain. (Li and Chu, 2017) used ANN for price prediction. Financial variables such as income, loan and growth rates were used to predict the house price indexes. They compared back propagation with radial basis networks and no significant differences were found.

In another ANN study (Zhang et al., 2012) factors other than the structural characteristics of the property were used for price prediction. In this study, parameters such as income, population, and gross domestic product were emphasized. Since studies using black box approaches focus on prediction performance, there are many studies comparing ANN with other machine learning methods. (Peter et al., 2020, Abidoye and Chan, 2017) presented a detail analysis of these studies.

In (Mukhlishin, 2017) study, ANN was compared to fuzzy and nearest neighbor models. Fuzzy model outperformed the ANN and, the effect of the variables used on the prediction result were not given.

In another study (Khalafallah, 2008) the effect of ANN architecture and hyper parameter tuning over the result was discussed. This study revealed that the architectural design and the hyper parameters could dramatically affect the prediction performance.

Among these studies, only (Abidoye and Chan, 2017) focused on the variable importance. These were given as relative importance measures. However, it did not show the relationship of the factors with each other or their cumulative effect on the result. Related studies show that black box approaches produce successful results in price prediction. However, the biggest disadvantage of these methods is the difficulty to show the effect of the variables on the outcome. The relative importance of the variables can be calculated. However, this imposes a large computational cost. As the hidden layers in the architecture increase, it becomes very difficult to make the relevant calculations.

Decision tree (DT) based methods are preferred in real estate price prediction as they can show the effects of variables on each other and on the result as rules. In addition, they both numerical and categorical predictions can be made. When the studies in the literature are examined, it is seen that CART and Random Forest (RF) methods are frequently used in real estate price prediction.

(Afonso et al., 2019) compared the prediction performances of RF and Recurrent Neural Network (RNN) based on root mean square log error. RF outperformed RNN. The effect of variables on the outcome was not examined for either method.

Hog et al. (2020) compared the performances of RF and Hedonic methods. With a 6%, deviation in the hit rate RF showed a better prediction. In the related study, the relative importance of the variables was also calculated. In this way, the effect of the variables on the result was determined. However, the interactions of the variables with each other were not analyzed.

(Levantesi and Piscopo, 2020) examined the effect of socioeconomic variables on price using RF. In this study, variables such as demand, population growth, and migration were used. (Breiman, 2001) variable importance was run on the RF model and the relative effect of the variables on the result was shown.

In another study, (Sawant et al., 2018) the prediction performance of the RF model was compared with DT. R-Squared and mean absolute error metrics were used, and in both metrics, RF gave slightly better results. Variable importance was also calculated. However, it is used for feature reduction rather than showing the effect of variables on the prediction.

When the studies in the literature are examined, it is seen that regression, black box and DT-based methods are frequently used in real estate price estimation. The regression-based methods suffer from poor performance of categorical prediction. Black box approaches are quite successful in learning nonlinear relationships in high-dimensional data. However, it is very difficult to show the effect of the variables on the result due to the black box characteristics. Although relative importance can be calculated for the variables, this calculation becomes very costly as the number of hidden layers and the number of nodes in each layer increase. DT is one of the methods that can show the effect of the variables on the result according to their interactions with each other. The RF, in the form of DT ensembles, further improves the prediction performance. However, since the ensemble is a kind of black-box approach, it becomes difficult to explain the effect of the variables on the result in the form of rules as in the DT. In the light of this information, DT was preferred in our study to make both categorical and numeric predictions and to reveal the importance of the variables on the result, as inter related rules.

3. MATERIAL and METHOD

3.1. Material

In order to create a price prediction model, residential for sale category in Canakkale Cumhuriyet district is gathered by a web scraper based on python scrapy library. There are 61 residential ads for sale in the dataset and 11 attributes for each. The attributes and data types are presented in Table 1.

Table 1. Attributes and their data types

Attribute	Data Type
Unit Price (TL)	Numeric
Residential Type	Categorical
Number of Rooms	Categorical
Area (gross) (m^2)	Numeric
Current Floor	Categorical
Building Age	Numeric
Heating	Categorical
Number of Floors	Numeric
Deed	Categorical
Facade	Categorical
Fuel Type	Categorical

Two different predictions are made, numerical and categorical, in the related study. For categorical prediction, unit price is divided into three classes as "High", "Medium" and "Low". In order to determine the relevant class labels, standard deviation (σ) and mean ($\bar{x}$) is used for the value range of the unit price. The relevant calculation is done as follows:

Low = $[\mathrm{Min}_{\mathrm{Unit\_Price}}, \mathrm{Min}_{\mathrm{Unit\_Price}}+ \sigma]$

Medium = $[\mathrm{Min}_{\mathrm{Unit\_Price}}+ \sigma+1, \bar{x}+ \sigma]$

High = $[\bar{x}+ \sigma+1, \mathrm{Max}_{\mathrm{Unit\_Price}}]$

When the data set with categorical class labels is constructed according to the calculation above, there are 29 "Low", 20 "Medium" and 12 "High" class residential.

3.2. Method

In this study, decision tree (DT) is used to predict the sale prices both numerically and categorically and to examine the factors affecting the prediction. In addition, the performance of the model is compared with the DC, which is frequently used in the price prediction for real estate. In this section, the details of both models are given.

3.2.1 Decision Tree Models

Two different models based on DT are developed in order to make both numeric and categorical prediction. DTs are preferred because of their visual interpretation. By this way, the parameters that affect the pricing are determined.

The CART (Breiman et al., 1984) is used because the "Unit Price" attribute is numerical. The branching criterion in the related tree is determined as least squares. Suppose that n is the number of data points, x_i denotes each single data point, C is the class label, V is the attribute vector and f is the prediction function. Then the error rate between the actual unit price and prediction is calculated as in Equation 1.

$$\sum_{i=1}^{n}\left(C_i - f(v,x_i)\right)^2 \tag{1}$$

The maximum depth for the regression tree is set as 10. The pruning algorithm (Salzberg, 1994) is used to minimize the repetitive paths. The minimum size required for a node to be divided into sub-branches is 4, and the minimum data number for a leaf is 2.

Since there are both numeric and categorical attributes, C4.5 (Salzberg, 1994) is used for classification. For the relevant model, the information gain ratio is used as the branch criterion. By this way, bias towards attributes with larger value range is prevented. Suppose that C is the class label and D is the data belonging to a certain class. The probability of the data i to belong class C_i, is p_i. Then relevant ratio is calculated as in Equation 2.

$$-\left(\sum_{i=1}^{m} p_i logp_i\right) - \left[\left(\sum_{i=1}^{m} p_i logp_i\right)\left(\sum_{i=1}^{k}\frac{D_{1i}}{D}\right)\right] \tag{2}$$

Similar to the regression tree in the first model, pruning is also used in the categorical model, the smallest node size is determined as 4 and the smallest leaf size is 2.

3.2.2 Direct Capitalization Model

DC is a sub branch of Hedonic calculation and widely preferred because it better states the financial and monetary condition of the property (Mayer et al., 2019). In this method, the value of a real estate is based on the annual rental income (Pınar and Demir, 2014). While applying the method, the expenses of the real estate and the rental losses due to its vacancy can be deducted (Onurlu, 2006; Yalçın et al., 2018). However, for the calculations made in the market, the expenses, loss caused by risk factors cannot always be estimated. For this reason, rental income can be used directly in the DC method (Michaletz and Artemenkov, 2018). Expenses and vacancy-rent loss are not taken into account in this study.

Capitalization rate (CR) is the rate calculated by dividing the annual rental yield of the real estate by the value. The most accurate approach to determine CR is to collect information from for sale and rental peers in the region. In the study, we also collected the rental information of the properties in the same region. The average unit price (AUP) is then calculated by the Equation 3. Here, p_r indicates the sale price and m gives the gross flat area.

$$AUP = \frac{\sum_{i=0}^{j=n} p_r}{\sum_{i=0}^{j=n} m} \tag{3}$$

AUP metric is used to calculate DC rate, which is the ratio between annual income and the AUP.

4. RESULTS

In this section, the results obtained from the DT model for categorical classification and unit price prediction are given and compared with the DC.

4.1. Unit Price Prediction

Since the unit price is numeric, CART with least squares is used. The maximum depth, the smallest leaf size and min. number of examples for branching are determined as 10, 2 and 4 respectively.

In order to create the model and determine the performance criteria, training and test datasets should be produced from the original dataset. It is important to preserve the class distribution of the original dataset in training and test sets to avoid any bias. Since this established model makes unit price estimation, the training and test data are created according to the distribution of this attribute in the original data. For this, the values given in Table 2 are calculated in the first step.

Table 2. The distribution of the unit price

Min Price (TL)	Max Price (TL)	Mean	Standard Deviation
2350	6847,826	3495,591	810,990

After the CART is built, we see that the most discriminative feature in the unit price prediction is the "Number of Rooms" attribute. According to the established tree, if the “Number of Room” attribute is “1+0” then the unit price is 3088 TL (Turkish Liras). For the values “3+2” and “5+1” the unit prices are predicted respectively as 6847 TL and 3111 TL. The tree structure is represented as sub-paths since it occupies a large space according to the rules specified in the page borders. The first two levels of the tree is given in Figure 1.

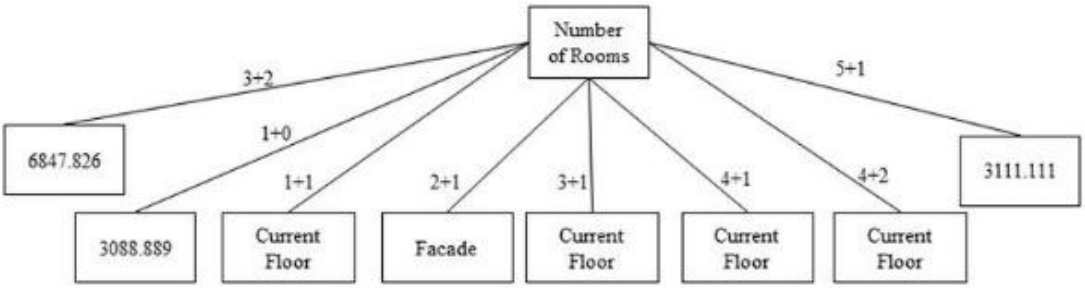

Figure 1. The first two levels of the regression tree

Given in Figure 1, apart from the branches for which direct price estimation is made, for some room numbers, branches are formed according to the "Floor" and "Facade" features. The subtrees formed for each room number value are presented in the following figures.

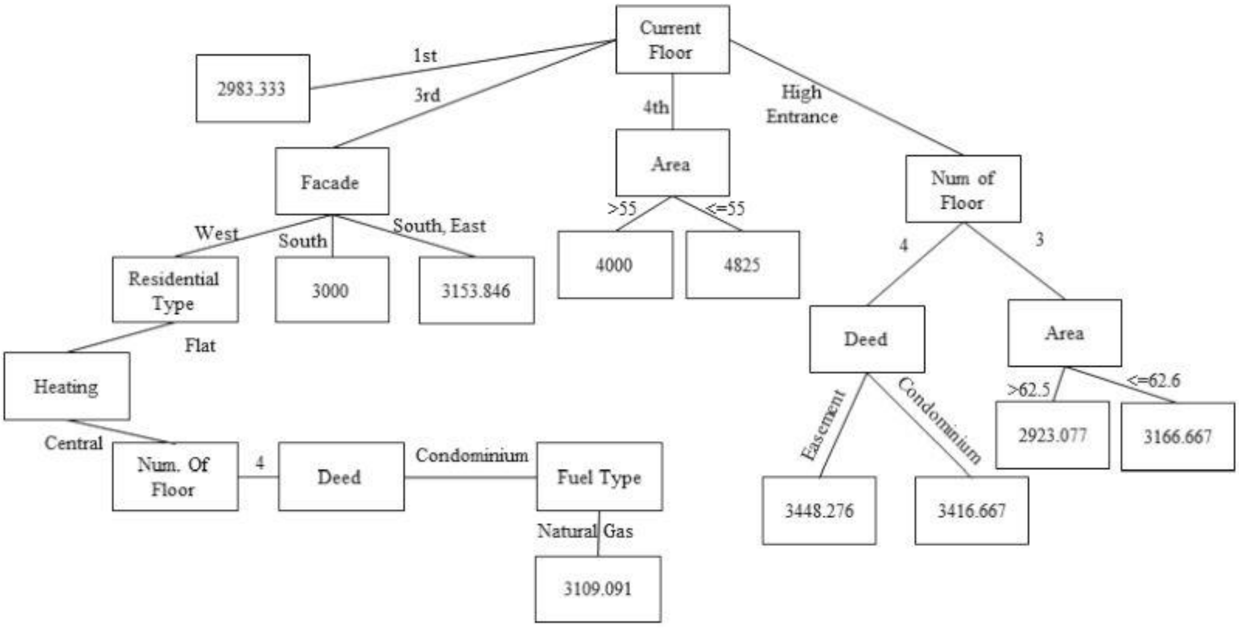

Figure 2. Sub regression tree for residences with 1+1 rooms

According to Figure 2, the most important attribute in the price prediction for the houses with 1+1 rooms is the "Floor". Accordingly, if the related house is on the first floor, the unit price prediction is given as 2983 TL. If the house is on the fourth floor, this time the area of the house is looked at, if the area is larger than 55 m2, the unit price is 4000 TL, if it is smaller, the unit price is 4825 TL. If an apartment with 1+1 rooms is on the third floor, the "Facade" becomes important. If it is at a high entrance, the number of floors of the building gains importance. The subtree formed for an apartment with 2+1 rooms is given in Figure 3.

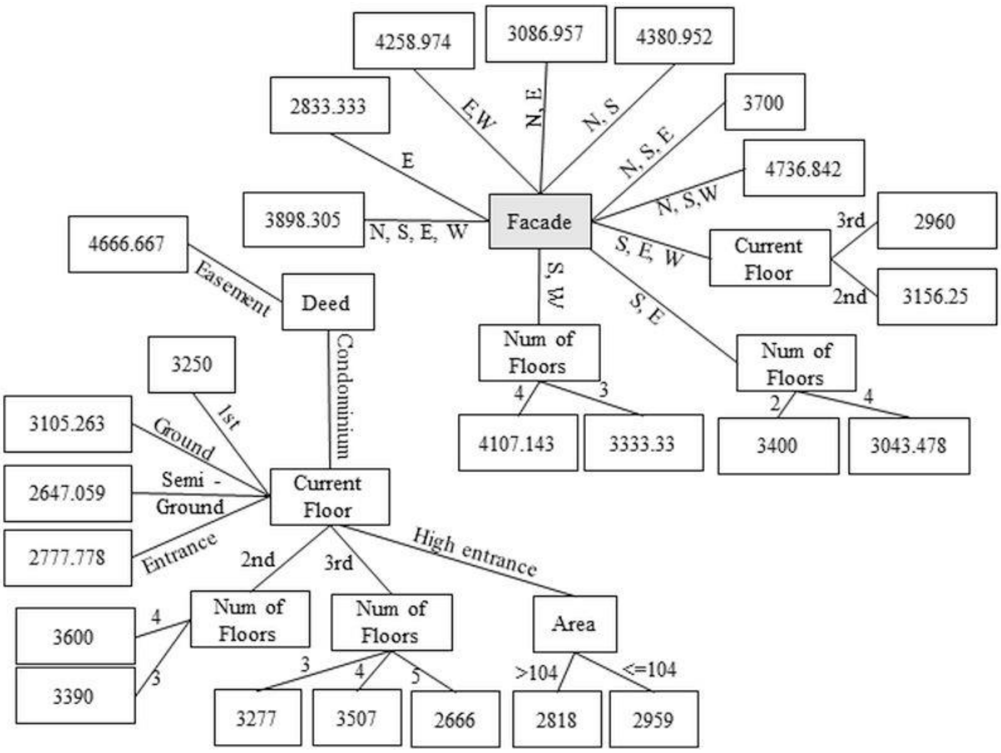

Figure 3. Sub-regression tree for residences with 2+1 rooms

According to Figure 3, "Facade" is the most important feature in price prediction for 2+1 flats. When the figure is examined, direct unit price prediction can be made for facades "East", "East-West", "North-East", "North-South-East", "North-South-West". For facades other than these, price predictions can be made by looking at the branches of the "Number of Floors", "Deeds" and "Floor" attributes.

The resulting subtree for the houses with 3+1 room number in Cumhuriyet district is given in Figure 4.

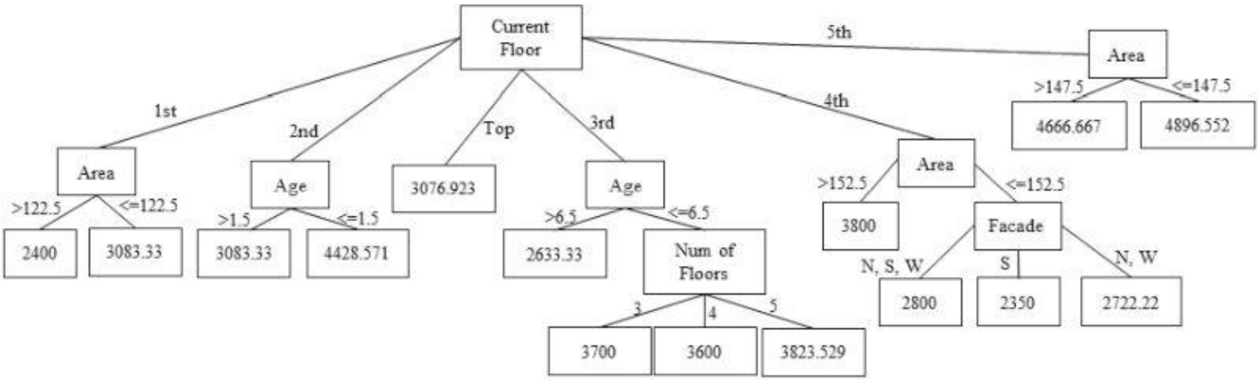

Figure 4. Sub-regression tree for residences with 3+1 rooms

For the flats with 3+1 rooms, the most important attribute in the price prediction is the "Current Floor" of the flat. While the unit price is predicted as 3076 TL for the flats on the top floor for the other floors, the prediction is made depending on the branching of the attributes of the flat area, “Building Age”, and “Number of Floors”.

For the flats with 4+1 and 4+2 rooms, the more compact subtrees are formed. That is because these flat types are very low in number Figure 5 represents the tree structures for these floor types.

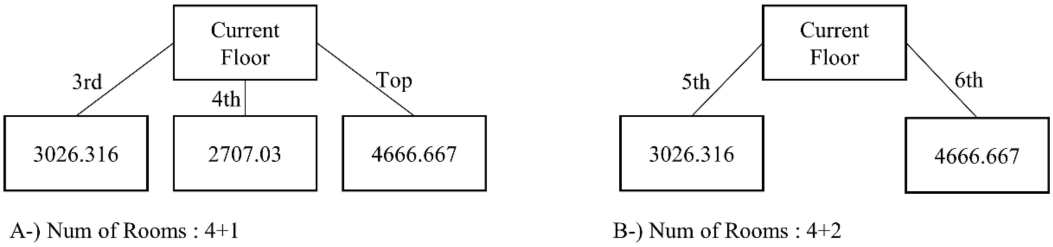

Figure 5. Sub-Regression tree for houses with 4+1 and 4+2 Rooms. A-) 4+1 flats B-) 4+2 flats

The price prediction is based on the "Current Floor" attribute for both flat types. When estimating the price for both types of flats, a price estimation is made according to the "Floor" attribute. If the 4+1 flat is on the third floor, the predicted unit price is 3026 TL. The prediction for the fourth and top floors are 2702 TL and 4666 TL, respectively. For 4+2 flats, the predicted unit price for those located on the fifth floor is 5121 TL. Those located on the sixth floor are 4444 TL.

4.2. Categorical Price Prediction

In order to group the unit prices under three classes (as explained in section 2.1), we calculated the distribution of the prices. This distribution is given in the Figure 6.

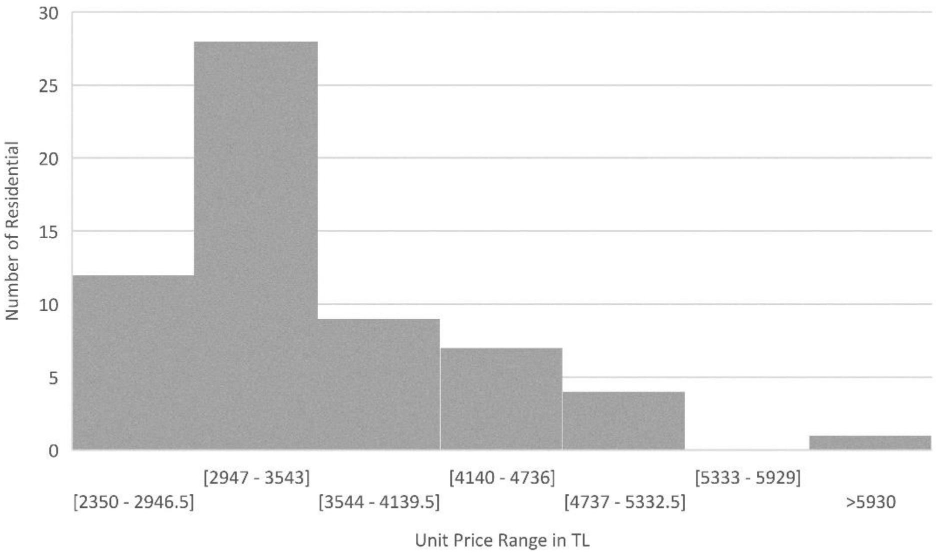

Figure 6. Distribution of unit prices of houses for sale in Cumhuriyet district

Figure 6 shows that most of the data is grouped close to the average unit price. Data as far as the standard deviation in the +/- direction from the mean is grouped as normal, and unit price values outside these limits are grouped as low and high. The classes are formed accordingly and the histogram graph is given in the Figure 7.

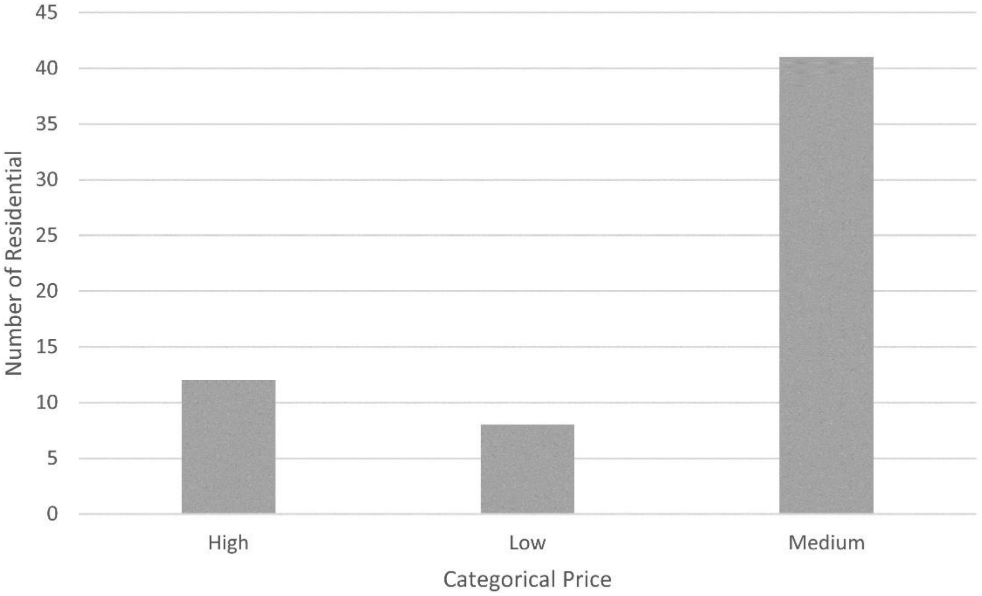

Figure 7. Distribution of unit prices of houses for sale in Cumhuriyet district by categorical classes

As given in the Figure 7, 20% of the data classified as high and 13% of them are classified as low. In the training and test datasets, these ratios are maintained to avoid bias and data are randomly selected.

C4.5 (Salzberg, 1994) is used as the decision tree since there are both categorical and numerical features in the related data set. In order to be compatible with the regression-based decision tree, the highest depth is determined as 10, the smallest leaf size is 2, and the lowest number required for branching is 4.

After the decision tree is built, we see that the most important feature in categorical unit price estimation is “Residential Type”. The first five levels of the tree is given in the Figure 8.

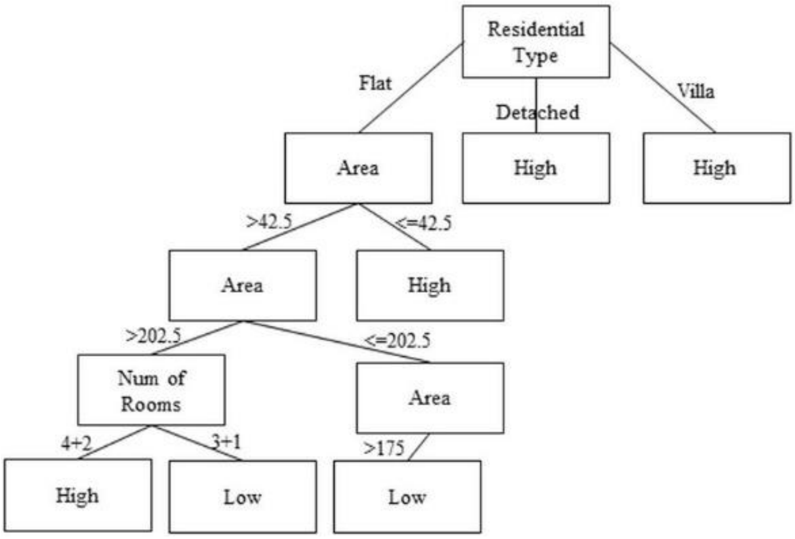

Figure 8. The C4.5 tree for the categorical unit price

Figure 8 shows that the prices of detached houses or villas are in the "High" class. Other attributes that are effective for unit price prediction are “Area” and “Number of Rooms”. The "Area" attribute at the fifth level of the tree is branched according to whether it is greater or less than 175 m^2. The two subtrees after this branching are given in Figure 9 and Figure 10.

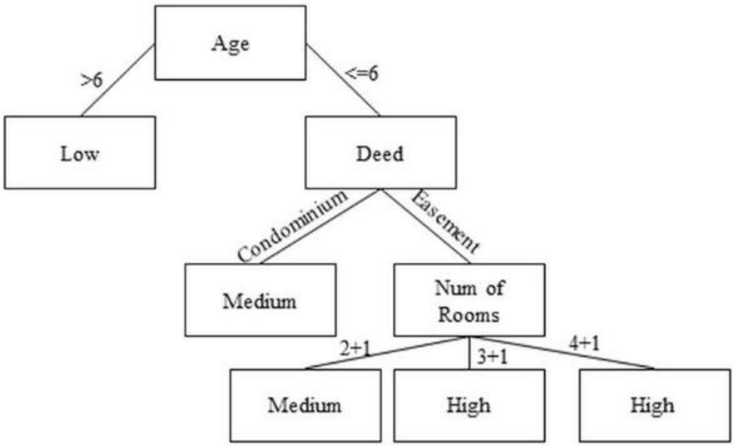

Figure 9. Sub tree of figure 8 when the area is larger than 175 m^2

According to Figure 9, if the age of the residential is older than six, then these residential grouped under "Low" class. Otherwise, the "Deed" attribute is checked. If there is a condominium, the price is in "Middle" class. If the deed is condominium and the number of rooms 3+1 and 4+1, these residential belong to class "High". While those with 2+1 are in the "Middle" class.

When we look to the bottom right level of the C4.5 given in the Figure 8, another subtree is formed for cases where the "Area" attribute is less than 175 m^2. This sub tree is given in Figure 10.

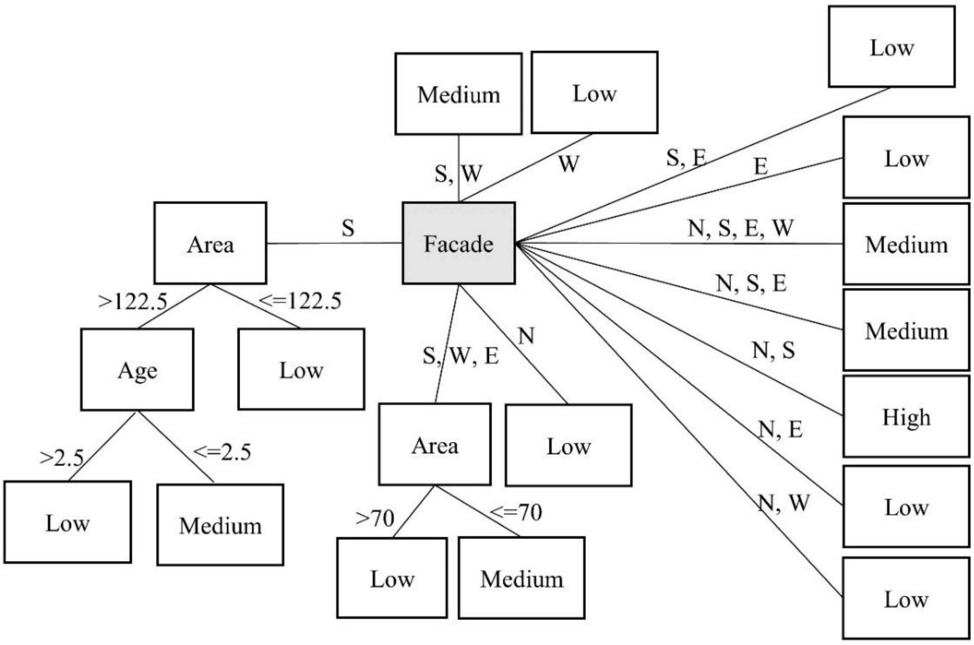

Figure 10. Sub tree of figure 8 when the area is smaller than 175 m^2

According to Figure 10, the most important attribute in price classification is "Facade". For residential on the west front, there is no probability of entering the "High" class, while the probability of entering the "Low" and "Middle "class is equal. The classification for the south facade is made according to the "Area" and "Building Age" attributes. Accordingly, flats on the south facade and larger than 122.5 m^2 are assigned to the "Middle" class.

4.3. Direct Capitalization Price Prediction

For DC calculation, the web scraper automatically gathers the information of the houses for sale and rental in the Cumhuriyet district. The unit prices for sale and rental residential in terms of TL/m^2 are given in the Table 3.

Table 3. The min, max and average unit prices for rental and sale residential

	Average	Min Average	Max Average
Rental	3495,591	10,625	16,667
Sale	13,239	2350	6847,826

The capitalization rate of the Cumhuriyet district, based on the numbers in Table 3, is calculated as 0.045447096. This rate is used in Equation 3 and the prediction of the DC is calculated as 3488,235 TL/ m^2 for residential in sale.

5. PERFORMANCE COMPARISONS

The regression tree is used for the unit price prediction. Considering the related problem within the framework of multi-classification, there are 61 classes, which is the total number of advertisements in the dataset. In this case, traditional performance metrics might not yield healthy results. Therefore, we preferred Root Mean Square Error (RMSE) (Armstrong and Collopy, 1992), as the performance metric.

RMSE for CART model is 13.169, and 358.69 for the DC. The error rate of the regression tree is quite low compared to the DC for the numerical unit price prediction.

The categorical price prediction is a multi-classification problem with three classes. Kappa (Vanbelle, 2017) and accuracy metrics are used to compare DC and C4.5. The results of C4.5 are given in Table 4.

Table 4. The performance metrics for C4.5

Accuracy	0,8852			
Kappa	0.811			
		Actual		
		Low	**Medium**	**High**
Model Prediction	**Low**	29	6	1
	Medium	0	14	0
	High	0	0	11

DC can predict lower and higher price for a given residential. Thus, DC gives two prediction models for the classification problem at the hand. The performance result of the first DC model, based on lower price prediction, is given in the Table 5.

Table 5. Performance criteria of the DC model established according to the lowest price prediction

Accuracy	0,754			
Kappa	0.633			
		Actual		
		Low	**Medium**	**High**
Model Prediction	**Low**	29	10	0
	Medium	0	10	5
	High	0	0	7

According to Table 5, DC based on the lower price successfully classifies the prices of residential in the low class. However, the classification performance decreases for the middle and high classes.

The results of the DC model for the highest price are presented in Table 6.

Table 6. Performance criteria of the DC model established according to the highest price prediction

Accuracy	0,639			
Kappa	0.470			
		Actual		
		Low	**Medium**	**High**
Model Prediction	**Low**	11	10	0
	Medium	18	16	0
	High	0	4	12

Table 6 shows that the DC model based on highest price is only successful for classifying the residential in high class. However, overall performance of the model is very low.

When the relevant tables are examined, the C4.5 model has a much better classification performance with 88% accuracy and 81% Kappa. On the other hand, DC has an average accuracy of 69.7% and average Kappa of 55%. These results indicate that the C4.5 classification model outperforms the domain standard, DC model.

6. CONCLUSION

There is no gold standard used by all stakeholders in estimating property prices for sale. Therefore, the general approach is to determine a price based on similar sale ads. The features of the similar properties are cross-compared and the price range is determined accordingly. However, the lack of a generally accepted standard may lead to incorrect pricing.

According to the literature search, it is seen that DC, is used to predict real estate prices (Adetiloye and Eke, 2014; Arslan, 2016). However, the success of this method depends on knowing the rental price of the property, as well as many other direct and indirect parameters (Wang et al., 2019; Yılmaz, 2019). If there is not enough expert opinion, the parameters are not examined, or the rental income is incorrect, the price is calculated incorrectly. In addition, the relevancy of the parameters are strongly dependent expert opinion which is subjective. Thus, DC itself cannot determine how the parameters affect the price. Apart from the traditional methods such as DC and Hedonic models, machine learning has just started to be used in the related field. However, studies focused on the prediction performance of the models rather than the factors affecting pricing.

In this study, our aim is to construct a model for the real estate price prediction and to find the parameters that affect pricing. In this way, a model is designed that can make price prediction without the need for rental information and an expert opinion.

The developed model is based on decision trees. Apart from the other studies, this model is able to make both numeric price prediction and categorical price classification. In order to make both estimations, CART (Breiman et al., 1984) is used for unit price prediction and C4.5 (Salzberg, 1994) is used for categorical price classification.

The relevant models are tested on the real estates for sale in the Cumhuriyet Neighborhood of Canakkale. In the numerical price prediction, the most important parameters are the "Number of Rooms", "Current Floor" and "Facade" attributes. According to the categorical classification, the parameters of “Residential Type”, “Area” and “Number of Rooms” come to the fore.

The decision tree model is compared with the DC. The RMSE metric is used for this comparison. The RMSE for decision tree model is 13.169, while it is 358.69 for DC. In categorical price classification, KAPPA and accuracy is used as performance criteria. Accordingly, the decision tree model has 81% Kappa and 88% accuracy, while these metrics for DC is only 55% and 69.7%, respectively.

Considering the results obtained, the developed model shows a superior performance in both numerical and categorical price prediction. In addition, the parameters that affect the pricing are reveled and analyzed in detail.

We plan to expand the relevant study further and turn it into a model that is applied to the whole province and then to the whole of Turkey. After this stage, the related model is planned as an application that can produce reports on different roles and detail levels by making predictions for sale and rental of different types of real estate for users.

REFERENCES

Abidoye, R.B., Chan, A.P.C., 2017, "Modelling property values in Nigeria using artificial neural network", *Journal of Property Research*, vol. 34, no. 1, pp. 36-53. doi: 10.1080/09599916.2017.1286366

Adetiloye, K.A., Eke, P.D., 2014, "A Review of Real Estate Valuation And Optimal Pricing Techniques", *Asian Economic and Financial Review*, vol. 4, no. 12, pp. 1878-1893. doi: https://doi.org/10.1108/JERER-08-2018-0035

Afonso, B.K.A., Melo, L.C., Oliveira1, W.D.G., Sousa, S.B.S., Berton, L., 2019, "Housing Prices Prediction with a Deep Learning and Random Forest Ensemble", web adresi: https://www.researchgate.net/publication/335527230_Housing_Prices_Prediction_with_a_Deep_Learning_and_Random_Forest_Ensemble, Ziyaret Tarihi: 20.12.20201

Armstrong, S., Collopy, F., 1992, "Error Measures For Generalizing About Forecasting Methods: Empirical Comparisons", *International Journal of Forecasting*, vol.8, no.1, pp. 69-80, 1992. https://doi.org/10.1016/0169-2070 (92)90008-W

Arslan, A., 2016, "*Kentsel Alanlarda Taşınmaz Değerlemesi*", Yüksek Lisans Tezi, Balıkesir Üniversitesi, Fen Bilimleri Enstitüsü, Balıkesir

Breiman, L., 2001, "Random Forests", *Machine Learning*, vol. 45, pp. 5–32

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984, "Classification And Regression Trees", 1st ed., Brooks/Cole Publishing, Monterey, CA, USA.

He, H.M., Chen, Y., Xiao, J.Y., Chen, X.Q. Lee, Z.J., 2021, "Data Analysis on the Influencing Factors of the Real Estate Price", *Artificial Intelligence Evolution* [Internet]. 2021Sep.10 [cited 2021Dec.23]; 2(2):52-66. Available from: https://ojs.wiserpub.com/index.php/AIE/article/view/966

Hong, J., Choi, H., Kim, W., 2020, "A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea", *International Journal of Strategic Property Management*, vol. 24, no. 3, pp 140-152. https://doi.org/10.3846/ijspm.2020.11544

Khalafallah,A., 2008, "Neural network based model for predicting housing market performance", *Tsinghua Science and Technology*, vol. 13, no. S1, pp. 325-328. doi: 10.1016/S1007-0214(08)70169-X

Levantesi, S., Piscopo, G., 2020, "The Importance of Economic Variables on London Real Estate Market: A Random Forest Approach", *Risks*, vol. 8, pp. 112. https://doi.org/10.3390/risks8040112

Li, L., Chu, K., "Prediction of Real Estate Price Variation Based on Economic Parameters", *2017 International Conference on Applied System Innovation (ICASI)*, Sapporo, Japan, 87-90, 2020. doi: 10.1109/ICASI.2017.7988353

Madhuri, C.R., Anuradha, G., Pujitha, M.V., 2019, "House Price Prediction Using Regression Techniques: A Comparative Study", International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 1-5, 14-15 March 2019. doi: 10.1109/ICSSS.2019.8882834

Manasa, J., Gupta, R., Narahari, N.S., "Machine Learning Based Predicting House Prices Using Regression Techniques", *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, Bangalore, India, 624-630, 2020. doi: 10.1109/ICIMIA48430 .2020.9074952

Mayer M., Bourassa, M., Hoesli, D., Scognamiglio, D., 2019, "Estimation and Updating

Methods for Hedonic Valuation", Journal of European Real Estate Research, vol. 12, no. 1, pp. 134-150. https://doi.org/10.1108/JERER-08-2018-0035.

Michaletz, V.B., Artemenkov, A., 2018, "The Transactional Assets Pricing Approach and Income Capitalization Models In Professional Valuation: Towards A Quick Income Capitalization Format", De Gruyter, vol. 26, no. 1, pp. 89-107. doi: 10.2478/remav-2018-0008.

Mukhlishin, M.F., Saputra, R., Wibowo, A., "Predicting House Sale Price Using Fuzzy Logic, Artificial Neural Network and K-Nearest Neighbor", 2017 *1st International Conference on Informatics and Computational Sciences (ICICoS)*, Semarang, Indonesia, pp. 171-176, 2017. doi: 10.1109/ICICOS.2017.8276357

Onurlu, Ö., 2006, Uluslararası Değerleme Standartlarının Türkiye'de Uygulanması Sürecinde Gelir Kapitalizasyonu Yaklaşımının İrdelenmesi, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Park, B., Bae, J.K., 2015, "Using Machine Learning Algorithms for Housing Price Prediction: The Case of Fairfax County, Virginia Housing Data", *Expert Systems with Applications*, vol. 42, no. 6, pp. 2928-2934. https://doi.org/10.1016/j.eswa.2014.11.040

Peter, N.J., Okagbue, H.I., Obasi, E. C.M., Akinola, A.O., 2020, "Review on the Application of Artificial Neural Networks in Real Estate Valuation", *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 3, pp. 2918–2925. https://doi.org/10.30534/IJATCSE/2020/66932020)

Pınar, A., Demir, M., 2014, "Konut Sektöründe Kapitalizasyon Oranlarını Belirleyen Faktörler: Türkiye için Bir Mikro-Veri Analizi," *Sosyoekonomi*, vol. 22, no. 22, pp. 386-398.

Piao, Y., Chen, A., Shang, Z., "Housing Price Prediction Based on CNN", *9th International Conference on Information Science and Technology (ICIST)*, Hulunbuir, China, 491-495, 2-5 Aug. 2019. doi: 10.1109/ICIST.2019.8836731

Phan, T.D., "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia", *International Conference on Machine Learning and Data Engineering (iCMLDE)*, Sydney, NSW, Australia, 35-42, 3-7 Dec. 2018. doi: 10.1109/iCMLDE.2018.00017

Rave, J.I.P., Morales, J.C.C., Echavarría, F.G., 2019, "A Machine Learning Approach to Big Data Regression Analysis of Real Estate Prices for Inferential and Predictive Purposes, *Journal of Property Research*, vol. 36, no. 1, pp. 59- 96, DOI: 10.1080/09599916.2019.1587489

Salzberg, S.L, 1994, "C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993", *Machine Learning*, vol. 16, pp. 235 – 240. https://doi.org/10.1007/BF00993309

Sawant, R. Jangid,Y., Tiwari, T., Jain, S., Gupta A., "Comprehensive Analysis of Housing Price Prediction in Pune Using Multi-Featured Random Forest Approach," *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Pune, India, 1-5, 2018. doi: 10.1109/ICCUBEA.2018.8697402.

Truong, Q., Nguyen, M., Dang, H., Mei, B., 2020, "Housing Price Prediction via Improved Machine Learning Techniques", *Procedia Computer Science*, vol. 174, pp. 433-442. https://doi.org/10.1016/j.procs.2020.06.111

Vanbelle, S., 2017, "Comparing Dependent Kappa Coefficients Obtained On Multilevel Data" Biom J., vol. 59, no. 5, pp. 1016- 1034. https://doi.org/10.1002/bimj.201600093

Wang, F., Zou, Y., Zhang, H., Shi, H., "House Price Prediction Approach Based on Deep Learning And ARIMA Model", IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), Dalian, China, 303-307, 19-20 Oct. 2019. doi: 10.1109/ICCSNT47585.2019.8962443

Ward, M.D., Gleditsch, K.S., 2019, *Spatial Regression Models*, 2nd ed., Sage Publications, Thousand Oaks, CA, USA.

Varma, A., Sarma, A., Doshi, S., Nair, R., "House Price Prediction Using Machine Learning and Neural Networks", 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 1936-1939, 20-21 April 2018. doi: 10.1109/ICICCT.2018.8473231.

Wu, C., Ren, F., Hu, W., Du, Q., 2019, "Multiscale Geographically and Temporally Weighted Regression: Exploring the Spatiotemporal Determinants of Housing Prices", *International Journal of Geographical Information Science*, vol. 33, no. 3, pp. 489-511, DOI: 10.1080/13658816.2018 .1545158

Yalçın, G., Selçuk, O., Şentürk, E., 2018, "Bursa İli Mustafakemalpaşa İlçesi Tarım Arazilerinde Kapitalizasyon Oranının Tespiti," *Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi*, vol. 18, no. 2, pp. 548-560. doi: 10.5578/fmbd.67386

Yılmaz, M., 2019, "*Gayrimenkul Değerleme Yöntemleri Ve Bir Uygulama*", Yüksek Lisans Tezi, Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul

Zhang P., Ma, W., Zhang, T., 2012, "Application of Artificial Neural Network to Predict Real Estate Investment in Qingdao", *Future Communication, Computing, Control and Management. Lecture Notes in Electrical Engineering*, 141, Editör: Zhang, Y., Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-27311-7_28