Megaron, Vol. 19, No. 2, pp. 259–274, June 2024

Megaron
https://megaron.yildiz.edu.tr - https://megaronjournal.com
DOI: https://doi.org/10.14744/megaron.2024.22316

Article

Determination of the relationship between housing characteristics and housing prices before and after the Kahramanmaraş earthquake using machine learning: A case study of Adana, Türkiye

Simge DOĞAN[1], Levent GENÇ[2,*], Sait Can YÜCEBAŞ[3], Metin UŞAKLI[1], Cengizhan DUMLU[4]

[1]1Instute of Sciences, Department of Real Estate Development, Canakkale Onsekiz Mart University, Computer- Agriculture- Environment- Plan Research Group (ComAgEnPlan) Çanakkale, Türkiye
[2]Faculty of Architecture and Design, Department of Urban and Regional Planning, Computer- Agriculture- Environment- Plan Research Group (ComAgEnPlan), Canakkale Onsekiz Mart University, Çanakkale, Türkiye
[3]Faculty of Engineering, Computer Engineering Department, Canakkale Onsekiz Mart University, Computer- Agriculture- Environment- Plan Research Group (ComAgEnPlan), Çanakkale, Türkiye
[4]Institute of Sciences, Depertment of Computer Engineering, Yıldız Technical University, Istanbul, Türkiye

ARTICLE INFO

Article history
Received: 05 September 2023
Revised: 20 March 2024
Accepted: 13 May 2024

Key words:
Adana, earthquake, housing preference criteria, machine learning, real estate.

ABSTRACT

Earthquakes have a significant impact on the real estate sector. Damage caused by earthquakes leads to an imbalance in the supply and demand for housing, thus temporarily causing stagnation in the real estate sector. Two earthquakes occurred in the Pazarcık and Elbistan districts of Kahramanmaraş on February 6, 2023, at 04:17 am with a magnitude of 7.7 and at 13:24 pm with a magnitude of 7.6. A machine learning-based model was created to analyze the change in house prices and the variables affecting the price during the earthquake, which is called "the Disaster of the Century." After the earthquake, the prices of houses for sale in the central districts of Adana province (Seyhan, Yüreğir, Sarıçam, and Çukurova), where there was the least damage, were collected from the relevant website with a web scraper. These data were classified as categorical and numerical datasets, and the necessary pre-processing stage for machine learning algorithms was performed. The characteristics that change and are effective in housing preferences before the earthquake (February 2022) and after the earthquake (February 2023) were determined by the decision tree method, which is one of the machine learning algorithms. In this context, it is aimed to determine the housing variables that are effective in before- and after-earthquake pricing in the central districts of Adana province. In the study, while 'Building Age and Number of Rooms' are effective in determining the price in 2022, 'Housing Shape and Facade' features come to the fore in 2023. The housing characteristics that affect the price change in two years. The change in housing preference criteria after the earthquake shows that the lifestyle in cities has also changed. According to this change, it requires the development of new approaches in urban design and planning approaches and is expected to be a reference for future studies.

Cite this article as: Doğan, S., Genç, L., Yücebaş, S. C., Uşaklı, M., Dumlu, C. (2024). Determination of the relationship between housing characteristics and housing prices before and after the Kahramanmaraş earthquake using machine learning: A case study of Adana, Türkiye. Megaron, 19(2), 259–274.

***Corresponding author**
*E-mail adres: leventgc@comu.edu.tr

INTRODUCTION

Türkiye is geographically located on the North Anatolian, East Anatolian, and West Anatolian fault lines. Among these, the North Anatolian and East Anatolian Faults are the active fault lines (Caglar et al., 2023). There were 269 earthquakes that caused damage in Türkiye between 1900 and 2023. According to the Republic of Türkiye Presidency Strategy and Budget Directorate (TCCSBB) Kahramanmaraş and Hatay Earthquakes Report (TCCSBB, 2023), the largest earthquakes in terms of damage were, respectively, the Kahramanmaraş, 2023, Gölcük 1999, and Erzincan-centred 1939 earthquakes. On February 6, 2023, earthquakes occurred in Türkiye at 04:17 and 13:24 Turkish time, demonstrating this damage once again. According to the TCCSBB, (2023), two earthquakes of magnitude 7.7 and 7.6 occurred in districts of Kahramanmaraş (Pazarcık and Elbistan). This severe earthquake was felt over a wide area in southeastern Türkiye (Adana, Adıyaman, Diyarbakır, Elazığ, Gaziantep, Hatay, Malatya, Kahramanmaraş, Kilis, Osmaniye, Şanlıurfa provinces) and northwestern Syria. According to Turkish Statistical Institute (TÜİK) data, 14,013,196 people were affected (KRDAE, 2023; TÜİK, 2023a). The official report on the Kahramanmaraş earthquake stated that many people were trapped under the buildings due to the severity of the earthquake.

This caused extensive damage to the region, with over 45,000 people reported dead and many injured (Dal Zilio & Ampuero, 2023). According to the TCCSBB Kahramanmaraş and Hatay Earthquake Report, the cities of Hatay, Kahramanmaraş, Adıyaman, and Malatya were identified as the cities with the highest destruction. In total, damage assessment studies were conducted on 1,712,182 buildings in 11 provinces. Accordingly, it was determined that 35,355 buildings were destroyed, 31,421 buildings were slightly damaged, 40,228 buildings were moderately damaged, and 179,786 buildings were severely damaged. 46,640 independent units in 6849 buildings were moderately damaged and 439,647 independent units in 59,995 buildings were slightly damaged (TCCSBB, 2023).

Earthquakes, which cause large-scale damages, negatively affect the economy at regional and national levels and many sectors. One of these is the real estate sector. After the earthquake, there are changes in the prices of housing and land in the disaster area, causing a decrease in sales rates as people will not prefer to buy real estate in the earthquake zone (Beron et al., 1997; Nakagawa et al., 2009; Naoi et al., 2009). In addition, the purchase and sale of real estate for investment purposes will slow down in areas with high earthquake risk.

Related Work

There are many studies examining the impact of earthquakes on house prices, but most of these do not use machine learning methods.

The study by Brookshire et al. (1985) calculated the change in housing prices in earthquake-prone areas in California. They concluded that earthquakes directly affect prices. In another study conducted in California, the hedonic model was used to analyze housing prices before and after the earthquake. This study especially focused on the post-earthquake market values of single-family houses (Beron et al., 1997). There are many similar studies conducted in California. These studies emphasize that earthquake risk reduces housing prices (Beron et al., 1997; Brookshire et al., 1985; Fekrazad, 2019; Jung & Smith, 2022; Murdoch et al., 1993; Singh, 2019).

Similar results were obtained in Japan, a country where earthquakes are quite frequent. After severe earthquakes in Japan, earthquake risk maps were created using geographic information systems (GIS) and associated with house prices using a hedonic model. Accordingly, it was revealed that housing prices were lower in areas with high earthquake risk (Nakagawa et al., 2007). In another study, the effect of earthquake risk on land prices in the Tokyo metropolitan area was examined (Nakagawa et al., 2009). This study revealed that land with high earthquake risk is priced approximately 8% less than land with lower risk.

However, these studies were not very successful in establishing a relationship between earthquake risk and variables affecting house prices. The effect of such variables (housing variables, environmental variables, etc.) on house prices and their relation with earthquake risk maps in Japan was analyzed using a hedonic price model (Naoi et al., 2009). This study showed that housing prices decreased after the earthquake risk maps were created and stated that houses in earthquake risk areas were more discounted.

Hidano et al. (2015) investigated the impact of earthquake risk on real estate prices in Tokyo. They compared the Two-Dimensional Regression Discontinuity (2DRD) model, hedonic model, and the traditional one-dimensional regression (RD) approach. They found that housing prices fall in earthquake-prone areas and that newly constructed buildings are more resistant to earthquakes. The study emphasized that 2DRD was better than traditional approaches.

In another study examining the effects of earthquakes on the economy and housing, three different results were obtained with the hedonic model (Koster & Ommeren, 2015). According to this: (a) houses are sold at lower prices due to the damage caused by earthquakes, (b) people prefer to sell their houses as damaged buildings are repaired after earthquakes, leading to an oversupply, (c) demand for houses in the region decreases due to the fear that future earthquakes will cause more damage, leading to lower prices.

Lara-Pulido et al. (2022) analyzed the short- and medium-term effects of the earthquake in Mexico on the housing market using a hedonic model. As a result, they stated that people's lack of sufficient information about the earthquake risk or the lack of proper understanding of the earthquake risk had an impact on real estate prices.

Several studies have reported that the hedonic model (HM) is often used to identify post-earthquake price changes (Beron et al., 1997; Hidano et al., 2015; Lara-Pulido et al., 2022; Murdoch et al., 1993; Nakagawa et al., 2007; Nakagawa et al., 2009; Naoi et al., 2009). However, hedonic models such as regression-based computation focus on numerical prices and their categorical price classification performance is generally poor (Yücebaş et al., 2022). Therefore, regression-based price estimation is not preferred in this study. According to the methods and analyses used in current studies, it is determined that real estate prices depreciated due to the earthquake (Beron et al., 1997; Murdoch et al., 1993; Nakagawa et al., 2007, 2009; Naoi et al., 2009).

However, the existing literature has not analyzed in detail the features affecting house prices before and after the earthquake using machine learning methods.

Although Adana has the largest population in the earthquake zone, it was the least affected by earthquake damage. In this study, a machine learning-based model is developed to analyze the change in house prices and the variables affecting the price before and after the earthquake. For the related study, pre-earthquake (February 2022) and post-earthquake (February 2023) house prices in the central districts of Adana province were collected via a web scraper. A machine learning model based on decision trees is built over this data to reveal the variables that affect house prices before and after the earthquake and to reveal any changes.

This study is designed to fill the gaps explained above and to reveal the potential of the decision tree model to explain the housing variables that are effective in pre- and post-earthquake pricing.

MATERIALS AND METHODS

Study Area

Adana province is located at coordinates 37.5005°N 35.715°D in the Çukurova, the Mediterranean Region in southern Türkiye (Figure 1). It is bordered by Kayseri to the north, Osmaniye to the east, Kahramanmaraş to the northeast, Hatay to the southeast, Niğde to the northwest, and Mersin to the west. According to the Turkish Statistical Institute (TÜİK) data for the year 2022, Adana is the 7th largest city in Türkiye with a population of 2,274,106. The area of the province is 13,844 km^2. It has 15 districts, five of which are central (Seyhan, Yüreğir, Sarıçam, Çukurova) and has 831 neighborhoods (Wikipedia, 2023; TÜİK, 2023b). In 2022, the population of Çukurova, Sarıçam, Seyhan, and Yüreğir districts was

Figure 1. Location of the study area.

389,195, 221,733, 795,012, and 404,726 respectively. Sarıçam district has the highest annual population growth rate of 62.8% in Adana province (TÜİK, 2023a).

Material

In order to build a decision tree model, two datasets covering house sales in the four main districts of Adana (Seyhan, Yüreğir, Çukurova, and Sarıçam) before the earthquake (February 2022) and after the earthquake (February 2023) were created through a web scraper. Two datasets, pre- and post-earthquake, were created by removing extreme data and uncommon variables (pre-processing) from the house sale prices collected from the open access internet sales site (HepsiEmlak, 2023) using the web scraper. In the pre-earthquake dataset, 3017 sales data and in the post-earthquake dataset, 3391 sales data were made ready for analysis with 11 variables (Table 1).

Table 1. Variables and data types

Variables	Data Type
Current Floor	Numeric
Number of Floors	Numeric
Unit Price (TL)	Categorical
Area Attribute (m2)	Categorical
Residential Type	Categorical
Building Age	Categorical
Heating	Categorical
Number of Rooms	Categorical
Building Type	Categorical
Facade	Categorical
District	Categorical

As Table 1 indicates, data types are numeric and categorical. Furthermore, the range of values of numeric variables is also inherently wide. When we tested the available data, we found that this caused the decision tree to branch too much and, in some cases, the data was memorized by the software. To avoid this, some of the numeric data sets have been converted into categorical data types. The price and area variables are categorized into three classes as 'High,' 'Medium,' and 'Low.' Standard deviation (σ) and mean ($\bar{x}$) are used to determine the range of values in these classes. The formulas (1,2,3) used are given below (Yücebaş et al., 2022).

Low = $[MinUnit\_Price, MinUnit\_Price + \sigma]$ (1)

Medium = $[MinUnit\_Price + \sigma + 1, \bar{x} + \sigma]$ (2)

High = $[\bar{x} + \sigma + 1, MaxUnit\_Price]$ (3)

After the transformation, there are 1537 low-priced, 1167 medium-priced, and 315 high-priced houses in the pre-earthquake dataset. In the post-earthquake dataset, there are 1786 low-priced, 1345 mid-priced, and 260 high-priced houses. Using the same formulas for the area attribute, there are 76 low, 2723 medium, and 217 high area attribute houses in the pre-earthquake dataset. In the post-earthquake dataset, there are 439 low, 2687 medium, and 265 high area attribute houses. In the case of building age, the range of values in the attribute is very sparse. The average building age in the pre-earthquake (2022) dataset is 10, and the average building age in the post-earthquake (2023) dataset is similarly 10. Therefore, the boundary value 10 was accepted as the average value and divided into 2 classes. Apart from these variables, for the 'Building Age' variable, buildings newer than 10 years are categorized as 'New,' and buildings older than 10 years are categorized as 'Old.' With this transformation, there are 1540 new and 1476 old buildings in the pre-earthquake dataset; and there are 1924 new and 1467 old buildings in the post-earthquake dataset. Variables, their value ranges, and frequencies of each value range are given in Table 2.

According to Table 2, the most preferred residential type in the central districts of Adana is 'Flat,' the most common heating type is 'Combi boiler,' and the most preferred building type is 'Reinforced Concrete' structures. In addition, the most common number of rooms is '3+1,' and these dwellings are mostly '3 Facade.'

In this dataset, the number of houses with a 1+1 number of rooms between 2022 and 2023 is higher. This may be an indication that affordable houses are being built in Adana. In addition, the increase in the number of houses with 4 facades in 2023 compared to 2022 is also noteworthy in the real estate sector. The tables in which the variables are analyzed in detail in terms of frequency, mean, standard deviation, minimum, and maximum values of the datasets obtained with the web scraper before and after the earthquake are given below. The purpose of analyzing these tables is to provide a better understanding of the datasets used for decision trees (Table 3 and Table 4).

Since the mean, standard deviation, minimum, and maximum values with less than 2 frequency cannot be calculated, they are not shown in Table 3 and Table 4. In Pre-Earthquake, the maximum price was 5,000,000 and the minimum price was 400,000, while in Post-Earthquake the maximum and minimum prices were 15,000,000 and 600,000 respectively. In Pre-Earthquake, the minimum value for the area attribute was 45 m^2 and the maximum was 950 m^2. In Post-Earthquake, the minimum value for the area attribute was 35 m^2 and the maximum area attribute was 980 m^2. For the number of rooms variable, it is observed that the average price and area attributes generally increase as the number of rooms increases. However, the opposite was observed for the current floor variable. In Pre-Earthquake and Post-Earthquake, the minimum current floor of the houses was current floor (0), while the maximum current floor of the houses was on the 21st floor. As the number of rooms increases, the current floor number and the number of floors generally decrease. In Pre-Earthquake, the houses with the highest number of rooms were detached houses. The building age variable varies according to the differences in the number of rooms. The minimum building age was 0, while the maximum building age was 45. In the number of floors variable, it is seen that in Pre-Earthquake and Post-Earthquake, there were buildings with a minimum number of 1 floor and a maximum number of 27 floors (Table 3 and Table 4).

The change in the average price of houses for sale in Adana province before and after the earthquake is analyzed and presented in Figure 2.

In 2022, the average prices in Çukurova, Sarıçam, Seyhan, and Yüreğir districts are similar. However, in 2023, there is a significant increase in the prices of houses for sale, especially in Çukurova and Seyhan districts.

Method

Decision trees are among the most widely used algorithms in machine learning (Salzberg L., 1994). It is preferred because the result of the decision tree is visually easier to interpret. In this study, decision trees were used to analyze the change in house prices and to determine the affecting variables before and after the earthquake. The decision tree determines the most discriminative variable in the training set (T) and assigns it to the root node of the tree. While there are several metrics to calculate the discriminative power of the variables, entropy-based Information Gain is widely used (Salzberg L., 1994). If we assume the target variable as X, the number of values that the target variable can take as n, and the number of values that the predictor variable can take as v (Paul &

Table 2. Species and frequency values in the dataset

Variable	Type	Frequency Pre Earthquake (2022)	Frequency Post Earthquake (2023)
Residential Type	Flat	2773	3132
	Detached House	128	175
	Residance	10	10
	Villa	107	75
Heating	Not Specified	63	50
	Natural Gaz Stove	21	31
	Solar Energy	33	38
	No Heating	105	110
	Floor Heating	10	19
	Air Conditioning	490	629
	Boiler	2177	2362
	Central	116	91
	Stove	39	83
	Underfloor Heating	1	10
Building Type	Reinforced Concrete	2986	3381
	Steel	4	3
	Brick	8	3
	Stone	19	5
Facade	Single Facades	103	179
	2 Facades	523	525
	3 Facades	1932	2073
	4 Facades	458	615
Number of Room	1+0	0	1
	1+1	76	214
	2+1	420	643
	3+1	1691	1565
	3+2	2	2
	4+1	628	848
	4+2	14	6
	5+1	99	56
	5+2	13	11
	5+3	0	2
	6+1	15	13
	6+2	7	6
	6+3	8	5
	7+1	6	2
	7+2	3	4
	7+3	2	2
	8+1	23	0
	8+2	1	4
	8+3	3	4
	9+3	6	2

Table 3. Standard (std.) deviation, mean, min-max values by number of rooms (Pre-earthquake: February 2022)

Number of Room	Frequency	Parameter	Unit Price	Area Attribute (m^2)	Current Floor	Building Age	Number of Floor
1+1	76	Mean	574.753	63	5	2	10
		Std. Dev.	90.972	13	4	2	4
		Min.	400.000	45	0	0	3
		Max.	720.000	120	12	6	18
2+1	420	Mean	732.550	123	5	7	10
		Std. Dev.	264.307	34	4	9	4
		Min.	405.000	65	0	0	1
		Max.	3,250,00	366	18	40	20
3+1	1691	Mean	1.009.438	170	6	12	10
		Std. Dev.	323.896	32	4	10	4
		Min.	410.000	100	0	0	1
		Max.	3.750.000	510	20	45	24
4+1	628	Mean	1.620.302	225	7	12	11
		Std. Dev.	562.625	173	4	8	4
		Min.	440.000	130	0	0	1
		Max.	4.250.000	950	18	40	26
4+2	14	Mean	1.184.540	368	3	11	5
		Std. Dev.	632.372	226	4	9	5
		Min.	435.000	130	1	1	2
		Max.	4.873.000	726	18	30	19
5+1	98	Mean	2.331.727	299	6	12	9
		Std. Dev.	1.132.498	102	4	9	5
		Min.	620.000	130	1	0	1
		Max.	4.950.000	720	21	35	25
5+2	13	Mean	1.194.878	305	3	11	6
		Std. Dev.	623.119	109	3	10	5
		Min.	465.000	100	2	0	2
		Max.	4.750.000	450	11	30	16
6+1	15	Mean	1.241.750	357	9	13	8
		Std. Dev.	648.766	111	5	9	6
		Min.	550.000	110	2	0	2
		Max.	5.000.000	550	21	35	22
6+2	7	Mean	1.446.645	383	5	14	3
		Std. Dev.	415.929	167	2	7	1
		Min.	980.000	220	1	6	2
		Max.	3.000.000	700	6	25	4
6+3	8	Mean	954.741	222	6	21	3
		Std. Dev.	253	123	2	9	0
		Min.	490.000	100	2	7	3
		Max.	1.425.000	420	6	35	4
7+1	6	Mean	2.066.609	378	9	8	8

Table 3. Standard (std.) deviation, mean, min-max values by number of rooms (Pre-earthquake: February 2022) (Cont.)

Number of Room	Frequency	Parameter	Unit Price	Area Attribute (m²)	Current Floor	Building Age	Number of Floor
		Std. Dev.	596.628	65	5	5	6
		Min.	1.450.000	300	6	0	4
		Max.	3.800.000	450	19	13	20
7+2	3	Mean	2.076.818	450	2	8	3
		Std. Dev.	467.310	87	0	6	0
		Min.	1.569.000	400	2	2	3
		Max.	3.400.000	500	2	13	4
8+1	23	Mean	1.159.346	397	6	11	4
		Std. Dev.	550.631	67	1	7	0
		Min.	450.000	210	1	5	2
		Max.	3.500.000	500	6	40	4
8+3	3	Mean	1.372.111	245	2	25	3
		Std. Dev.	146.004	153	0	9	1
		Min.	1.150.000	140	2	15	2
		Max.	1.700.000	420	2	30	4
9+3	6	Mean	1.157.057	323	2	19	3
		Std. Dev.	329.561	123	3	8	0
		Min.	685.000	193	2	5	3
		Max.	2.150.000	50	3	25	3

Thomas, 2016), then information gain can be calculated as follows:

$$\Sigma_{i=1}^{m} (pi \log_2 pi) \quad (1)$$

$$\Sigma_{j=1}^{v} \frac{|xj|}{|x|} \; x \; Entropy\ (Xj) \quad (2)$$

$$Information\ Gain\ (T,X)=Entropy\ (T)\text{-}Entropy\ (T,X) \quad (3)$$

The tree starts branching according to the predictor variable that provides the highest information gain. The process tests all predicted variables to form sub-branches. In this study, the maximum depth for the decision tree was set to 8, and pre-pruning was applied to prevent overtraining. The minimum gain variable for pre-pruning was set as 0.01, and the minimum number of data for a leaf (min samples leaf) was set as 2.

The dataset contains both categorical and numeric data. Due to the high number of numeric data, it is difficult to read the tree as it increases the branching in the decision tree. For this reason, some numeric variables were converted to categorical data types. The remaining numeric variables (current floor and number of rooms) were left as numeric since they are the main factors affecting housing preferences. Since both numeric and categorical data are used together, the C4.5 algorithm (Salzberg L., 1994) with the information gain ratio was used for classification.

Decision Tree Modeling

In this section, decision tree models established with pre-earthquake (2022) and post-earthquake (2023) data are compared (Figure 3-Figure 10). Since the size of the decision tree models is very large, they do not fit in a single figure. For this reason, the sub-branches of the trees are given separately. In all figures, % indicate the decision certainty of the price prediction. In some branches of the decision tree, the classification rate may be 100% when going to the leaves. This indicates that all examples belonging to that leaf belong to the same class and does not mean overfitting. It indicates that the samples in that leaf have a homogeneous distribution under the decision path. Overfitting and underfitting are terms for an entire learning model (Montesinos López et al., 2022). They are measured through the overall performance metrics of the model (Aliferis & Simon, 2024). If these metrics, for example, accuracy and/or precision, are very high (close to 100%), and/or the test performance of the model is higher than the training performance, overfitting may be suspected. However, when we look at the performance metrics given in Table 5 and Table 6, there is no concern of overfitting. Overfitting can be suspected if all samples belonging to any class are 100% distributed across all leaves in the tree.

Table 4. Standard (std.) deviation, mean, min-max values according to the number of rooms + halls (Post-erathquake: February 2023)

Number of Room	Frequency	Parameter	Unit Price	Area Attribute (m2)	Current Floor	Building Age	Number of Floor
1+1	214	Mean	965.981	66	4	2	11
		Std. Dev.	302.022	17	3	4	3
		Min.	600.000	35	0	0	1
		Max.	2.000.000	203	13	26	22
2+1	643	Mean	1.441.740	118	5	7	9
		Std. Dev.	505.666	28	4	9	4
		Min.	610.000	50	0	0	1
		Max.	4.500.000	300	21	40	22
3+1	1566	Mean	2.178.079	170	5	12	10
		Std. Dev.	910.008	43	4	10	4
		Min.	620.000	70	0	0	1
		Max.	10.900.000	750	21	45	27
4+1	848	Mean	3.538.704	221	6	12	11
		Std. Dev.	1.494.036	48	4	8	4
		Min.	930.000	125	0	0	1
		Max.	15.000.000	800	21	42	27
4+2	6	Mean	2.011.667	249	1	23	3
		Std. Dev.	1.371.210	111	1	8	4
		Min.	750.000	150	0	10	2
		Max.	4.120.000	418	2	31	11
5+1	56	Mean	5.945.482	301	6	13	10
		Std. Dev.	3.383.456	92	4	11	5
		Min.	1.200.000	150	0	0	2
		Max.	15.000.000	750	16	45	21
5+2	11	Mean	5.405.909	368	3	13	5
		Std. Dev.	2.806.926	190	4	10	4
		Min.	1.400.000	165	0	0	2
		Max.	9.500.000	800	15	26	15
6+1	13	Mean	8.567.308	420	6	7	8
		Std. Dev.	4.158.392	150	6	6	7
		Min.	3.100.000	240	0	0	2
		Max.	15.000.000	750	21	20	25
6+2	6	Mean	4.208.333	329	4	4	5
		Std. Dev.	2.164.351	139	4	10	3
		Min.	2.000.000	150	2	6	2
		Max.	7.400.000	550	11	31	11
6+3	5	Mean	6.126.000	395	2	11	3
		Std. Dev.	5.725.022	202	1	8	1
		Min.	730.000	140	0	0	3
		Max.	14.500.000	600	2	20	4

Table 4. Standard (std.) deviation, mean, min-max values according to the number of rooms + halls (Post-erathquake: February 2023) (Cont.)

Number of Room	Frequency	Parameter	Unit Price	Area Attribute (m2)	Current Floor	Building Age	Number of Floor
7+2	4	Mean	8.900.000	367	2	6	3
		Std. Dev.	2.802.380	217	0	5	1
		Min.	5.400.000	120	2	0	2
		Max.	12.000.000	650	2	10	4
8+2	4	Mean	10.198.750	702	3	15	9
		Std. Dev.	3.203.853	210	3	12	6
		Min.	8.500.000	540	0	0	4
		Max.	15.000.000	980	6	25	14
8+3	4	Mean	10.198.750	702	3	15	9
		Std. Dev.	3.203.853	210	3	12	6
		Min.	8.500.000	540	0	0	4
		Max.	15.000.000	980	6	25	14

The most distinctive variable for both years is the 'Area' variable (Figure 3). The pre-earthquake decision tree is divided into three branches (Number of Rooms, District, and Heating) according to whether the area variable is low, medium, or high. The post-earthquake decision tree is branched to Facade, Heating, and Residential Type. When the variables that form the branches in the first levels of these trees are examined, it is seen that variables affecting the house price changed after the earthquake, and the importance of residential type and facade variables increased (Figure 3).

In 2022 (pre-earthquake), for houses with a low area variable, the number of rooms is the most important variable as it is included in the tree root. 1+1 and 2+1 houses are projected to be low-priced. The price prediction of houses with a 3+1

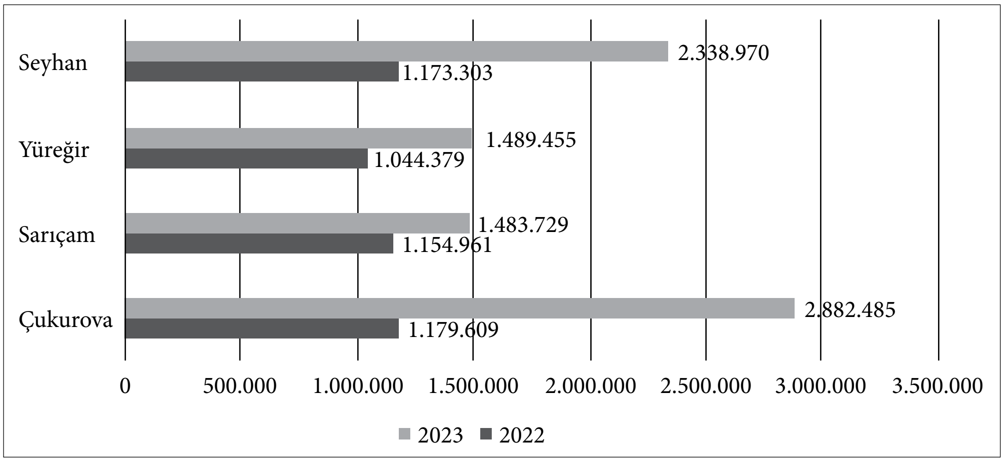

Figure 2. Change in the average price of houses for sale before and after the earthquake.

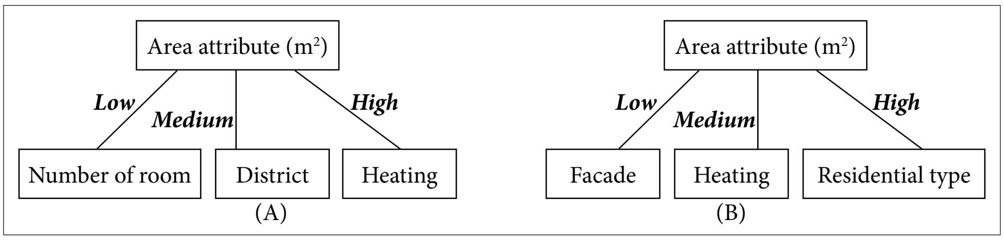

Figure 3. First two branches of the decision tree A: Pre-earthquake and B: Post-earthquake.

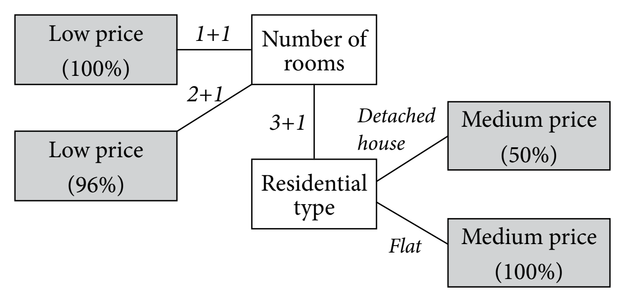

Figure 4. Pre-earthquake, sub tree for low area attribute.

number of rooms is based on the residential type variable. For these, detached houses and flats are considered to be medium-priced (Figure 4).

In the decision tree model in Figure 5, the most important variable for houses with a medium area variable in 2022 is the district. At the 2nd level, the facade in Sarıçam district, the number of floors in Yüreğir and Çukurova, and the number of floors in Seyhan were found to be important variables. Since the branching in the tree is very high and the district with the highest price change between 2022 and 2023 is Çukurova district (Figure 2), the decision tree model for this district is given.

In 2022, the most important variable affecting the price of flats for sale in Çukurova, which has a medium area variable, is the floor variable. Facade and building age variables are at the 2nd level of the tree. Number of floors, number of rooms, and residential type are assigned to lower levels. Accordingly, detached houses with an older building age are moderately priced; villas with more than 2.5 floors are low priced, while houses with less than 2.5 floors are high priced. In flats, it is seen that houses with a 2+1 number of rooms are low priced, while the branching continues in 3+1 and 4+1 according to different variables. The decision tree model is given in Figure 6.

It is found that the heating type variable affects price for both pre- and post-earthquake. This variable is related to the high area variable for the pre-earthquake dataset while it is related to the medium area variable for the post-earthquake dataset. The decision tree for dwellings with a high area variable in 2022 is given in Figure 7.

The floor and residential type variables are found at levels 2 and 3 of the tree. Accordingly, houses with 2 facades were determined as medium priced, houses with 4 facades and a floor greater than 2.5 were determined as high priced, and houses with a floor less than 2.5 were determined as low priced. For 3-fronted dwellings, if the floor is less than or equal to 3, it is moderately priced, and if it is greater than 3, the residential type is considered.

The 2023 sub-tree for houses with a low area variable is presented in Figure 8. According to the figure, the most

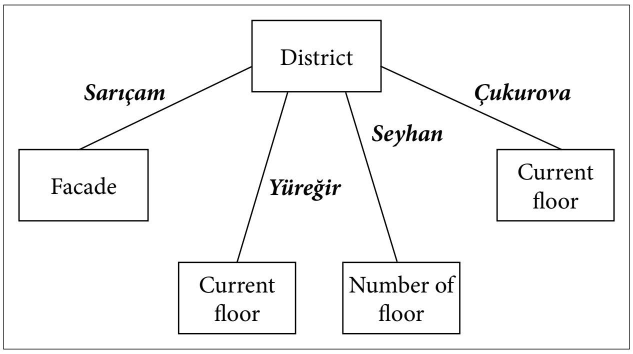

Figure 5. Pre earthquake, decision tree for dwellings with medium area attribute.

important variable affecting the price for houses with a low area attribute is the facade. However, the most important variable for houses with a low area attribute in 2022 is the number of rooms (Figure 4). In the tree for 2023, when the sub-branches of the area variable are analyzed, houses with one facade, 2 and 4 facades are classified as low priced. For houses with 3 facades, the district variable comes forward.

When the District sub-branch is followed, the price prediction of Çukurova and Yüreğir districts is determined as low priced. In Sarıçam district, residential type stands out, with flat and detached houses being low priced. In Seyhan district, houses with less than 3.5 floors are considered to be low priced.

For the 2023 dataset, the most important variable affecting the price for houses with a medium area variable is the heating type (Figure 9). For the 2022 dataset, the most important variable affecting the price of houses with a medium area variable is the district (Figure 5). It is seen that houses with 4 facades and a 4+1 number of rooms with air conditioning heating type are moderately priced. Houses with a 2+1 number of rooms are low priced. When the number of rooms variable takes the value of 3+1, the houses in Seyhan, Yüreğir, and Çukurova districts are determined as low priced. In Seyhan district, the floor is checked as the next variable. Houses with 2.5 floors or less are determined as low priced, and the certainty rate is 100%. Buildings above 2.5 floors are also considered low priced, but the certainty rate is 50%.

In 2023, the most important variable affecting the price of houses with a medium area variable is the heating type. While in 2022, the most important variable affecting the price of houses with a medium area variable is the district (Figure 5).

When Figure 9 is analyzed, it is seen that the most important variables in pricing are residential type, number of floors, number of rooms, and the floor on which it is located. Other variables are the age of the building and the district. It is seen that houses with 4 facades and a 4+1 number of rooms with

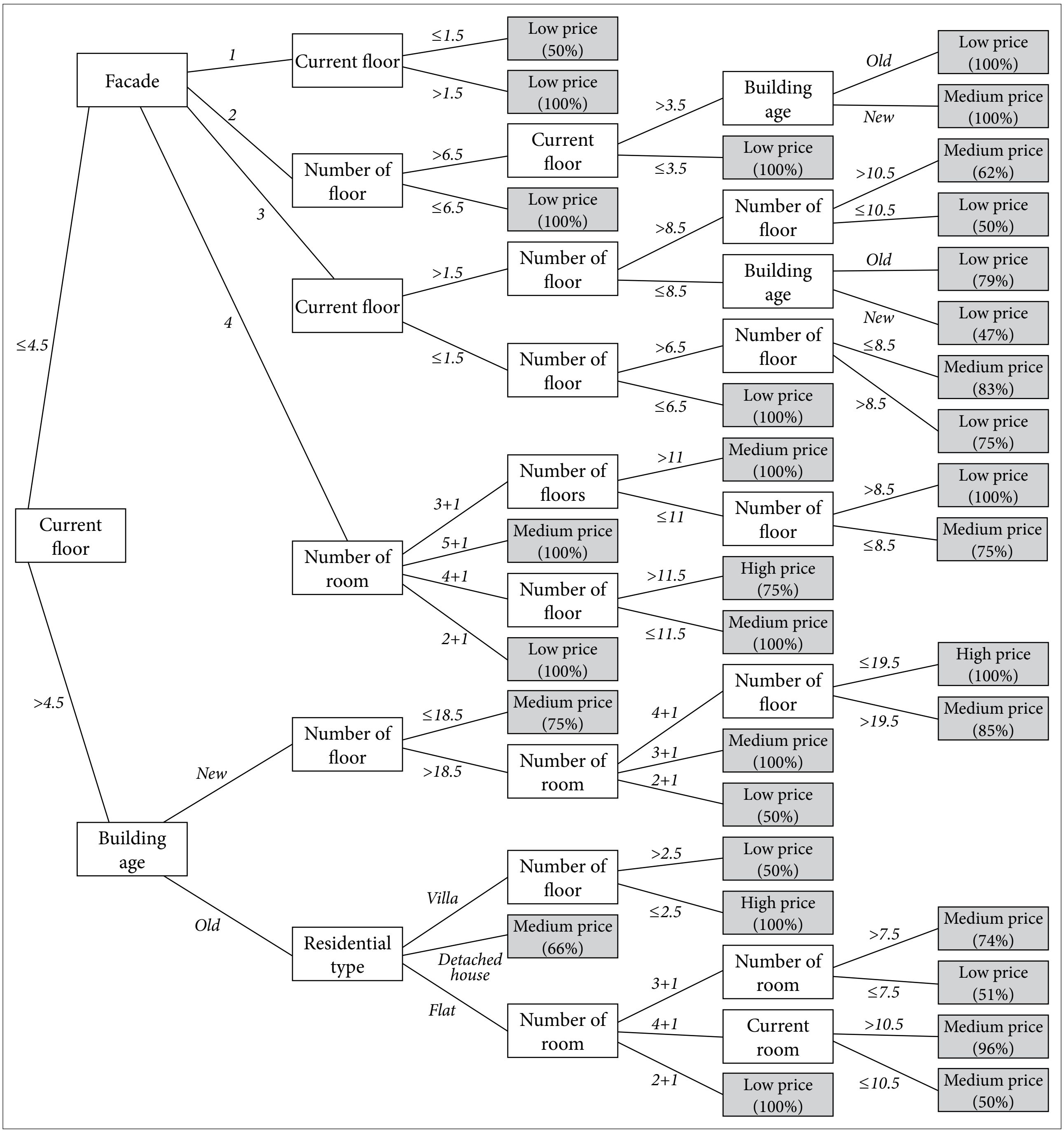

Figure 6. Pre-earthquake; Decision tree for Çukurova province with medium area attribute.

air conditioning are moderately priced. Houses with a 2+1 number of rooms are low priced. When the number of rooms variable takes the value of 3+1, the houses in Seyhan, Yüreğir, and Çukurova districts are determined as low priced. In Seyhan district, the floor is checked as the next variable. Houses with 2.5 floors or less are determined as low priced. Buildings above 2.5 floors are also considered low priced.

For houses with 3 facades, the district variable is checked first. In Yüreğir district, houses with 3 facades with air conditioning are predicted to be low priced (Figure 9). In Sarıçam district, the next variable, the type of housing, is checked, and it is seen that detached and flat-type houses are low priced. In Seyhan district, the next variable, which is the floor, is checked. Houses with 4.5 floors and below are low priced. For houses with a floor above 9.5, it is low priced, and if it is below 9.5, the number of floors variable is important. Houses with more than 6.5 floors are considered moderately priced, while houses with less than 6.5 floors are considered low priced. In Çukurova

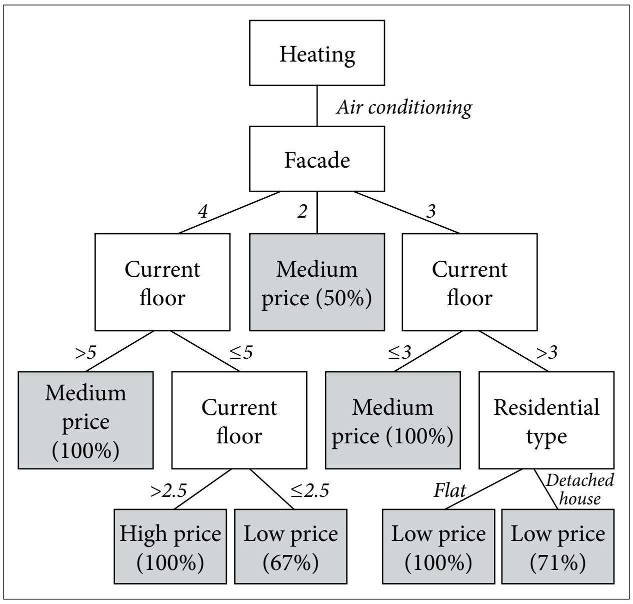

Figure 7. Decision tree for dwellings with high area attribute in 2022.

district, the floor variable is checked. If the current floor is above 9.5, the medium price is determined, while if the current floor is less than 9.5, the current floor becomes more important. Houses with less than 3 floors are predicted to be low priced, while for houses with more than 3 floors, the number of rooms variable is important. In Çukurova, houses with a 3+1 and 2+1 number of rooms are considered low priced, while 4+1 houses are predicted as medium priced (Figure 9). Similarly, the other branches of the tree can be read by following the lines to the right.

In 2023, the most important variable for dwellings with a high area variable is the residential type. For detached houses, flats, and villas, the most important variable is the number of floors. The details of the corresponding subtree are given in Figure 10. Residence-type houses are considered to be high priced. District, building age, heating, facade, and floor variables are assigned to lower levels.

Evaluation of Decision Tree Models

The variables affecting the price in the decision tree models for February 2022 before the earthquake and February 2023 after the earthquake are shown in Figure 11. The importance level of these variables decreases from the root to the lower levels, and the levels in the model are shown between Level 1 and Level 5.

Pre- and post-earthquake trees are branched according to the 'Area' variable. In the pre-earthquake period, the variables 'Number of Rooms,' 'District,' and 'Heating Type' are at the first level, while in the post-earthquake period, 'Facade,' 'Heating Type,' and 'Residential Type' variables are important at the first level. For both datasets, the variables affecting the price vary, but it is noteworthy that the heating type variable is at the first level for both datasets. It can be said that the heating type is important in affecting the price due to the weather conditions of the region.

On February 6, 2023, the earthquake and the demolition of flimsy high-rise buildings increased the tendency of people to prefer detached houses. In this case, residential type stands out among the variables affecting the price in 2023. The fact that detached houses generally have multiple facades shows that the 'Facade' variable is also a significant determinant of the house price in 2023.

After the earthquake, the effect of building age on the price changed due to the demolition of non-durable buildings, regardless of whether they were new or old. In the 2022 decision tree model, the age of the building is at a higher level, while in 2023 its importance in determining the price decreases.

In 2022, the number of rooms is important in the factors affecting house prices. However, in 2023, this factor is

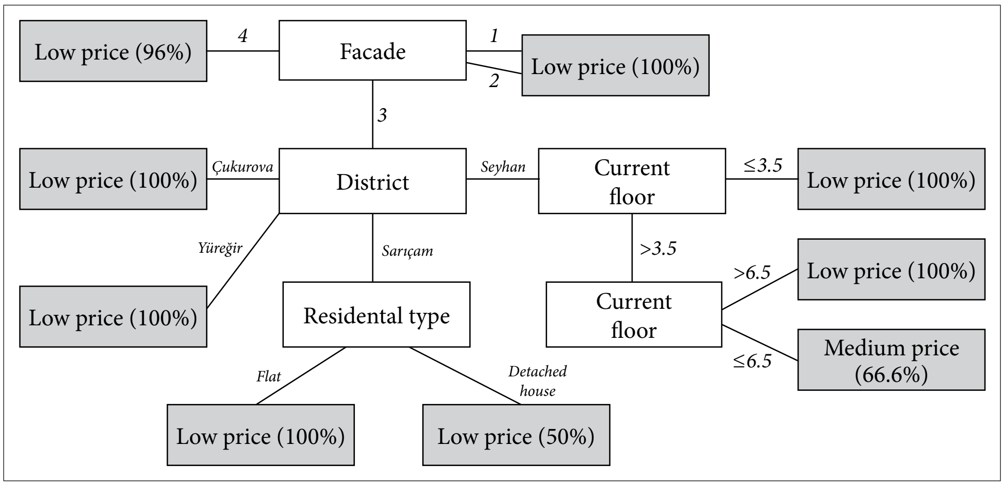

Figure 8. Decision tree for houses with low area attribute in 2023.

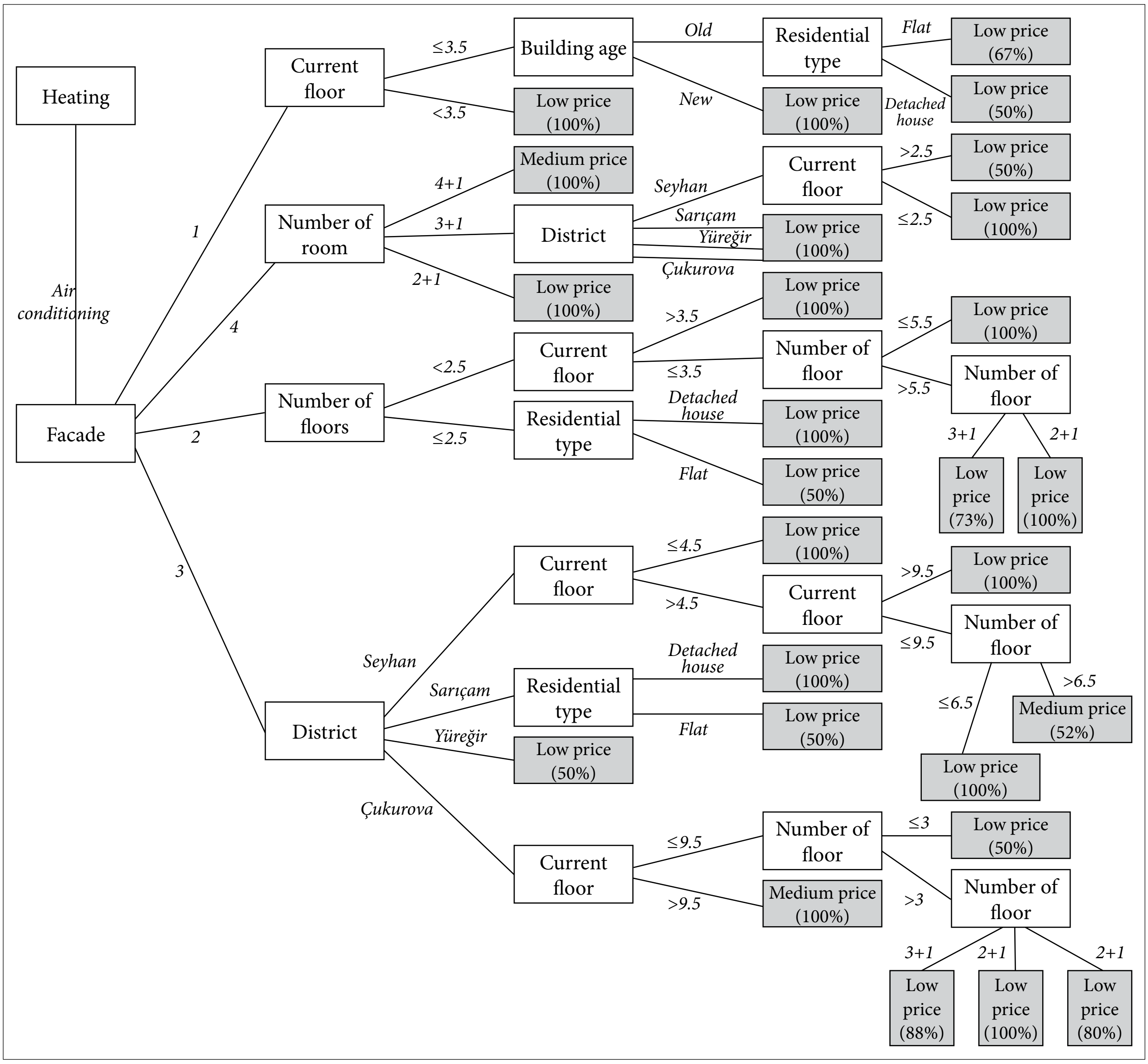

Figure 9. Decision tree for houses with medium area attribute in 2023.

found at lower levels of the decision tree. This shows that the number of rooms variable was more effective in the pre-earthquake period. The fact that this variable is less effective in housing preference criteria after the earthquake shows that the lifestyle in cities has changed.

Performance Evaluation of the Models

Accuracy was used to measure the prediction performance of decision tree models. Classification accuracy shows the rate of correct classification of the data in the sample. True Positive (TP) represents the result that the model correctly predicts the positive class, True Negative (TN) represents the result that correctly predicts the negative class, False Positive (FP) represents the result that incorrectly predicts the positive class, and False Negative (FN) represents the result that the model incorrectly predicts the negative class, and the way it is calculated is shown in formula (4) (Solanki et al., 2021).

$$\frac{\textit{True Positive(TP)+True Negative (TN)}}{\textit{True Positive+False Positive (FP)+ True Negative+ False Negative (FN)}} \quad (4)$$

The accuracy values used to measure the success criteria of the decision tree model in this study are given in Table 5 and Table 6.

There are a total of 3 classes in the pre-earthquake cluster for the year 2022. There are 615 data in low class, 466 data in medium class and 126 data in high class. The accuracy of the pre-earthquake decision tree model is 71.33%.

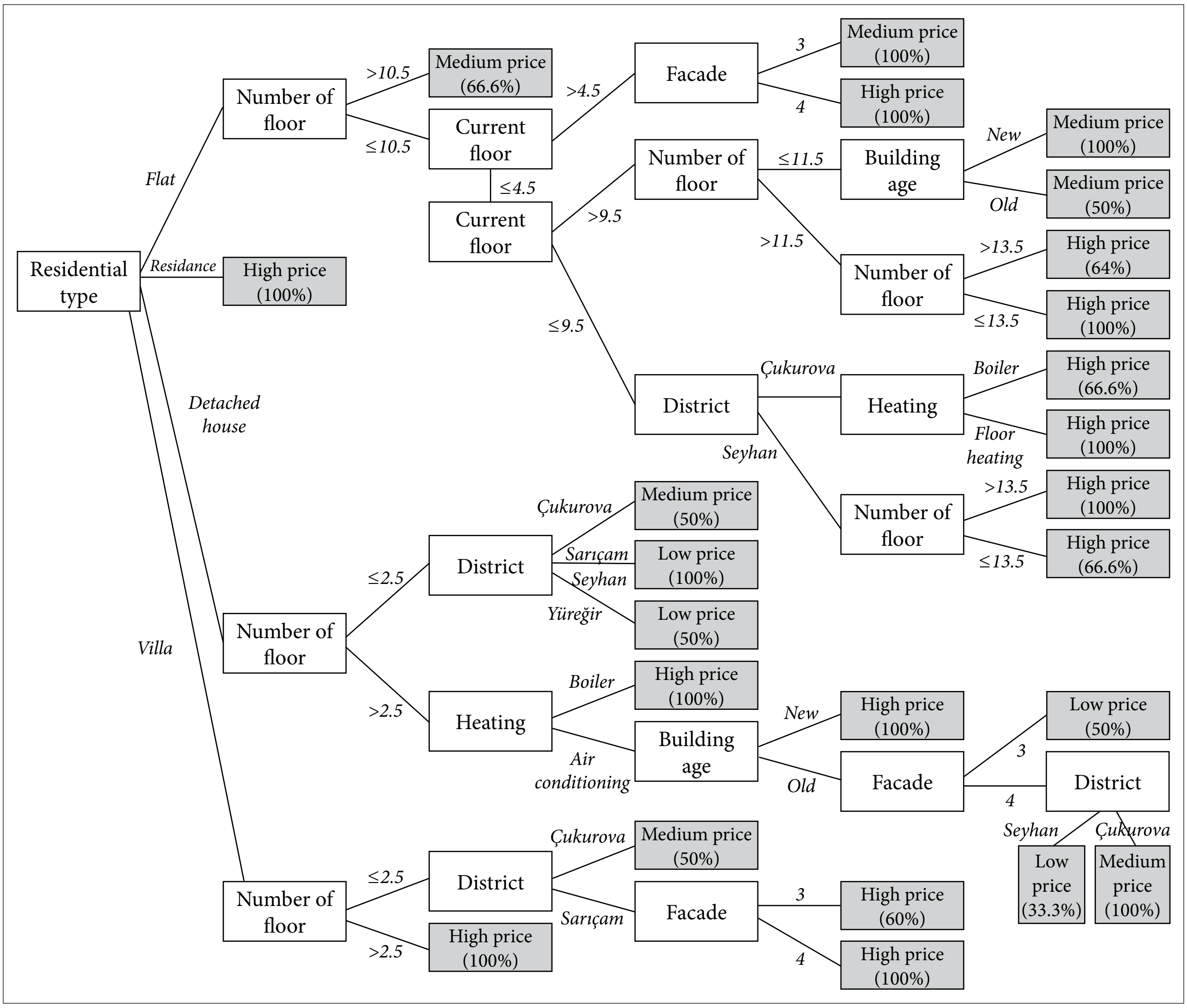

Figure 10. Decision tree for houses with high area attribute in 2023.

Table 5. Decision tree model performance values for 2022

Accuracy: %71.33	**Value**	**Actual**			
		Low	**Medium**	**High**	**Class Precision**
Model Prediction	Low	515	162	25	%73.36
	Medium	100	290	45	%66.67
	High	0	14	56	%80.00

Table 6. Decision tree model performance values for 2023

Accuracy: %74.63	**Value**	**Actual**			
		Low	**Medium**	**High**	**Class Precision**
Model Prediction	Low	584	144	18	78.28%
	Medium	128	383	41	69.38%
	High	2	11	45	77.59%

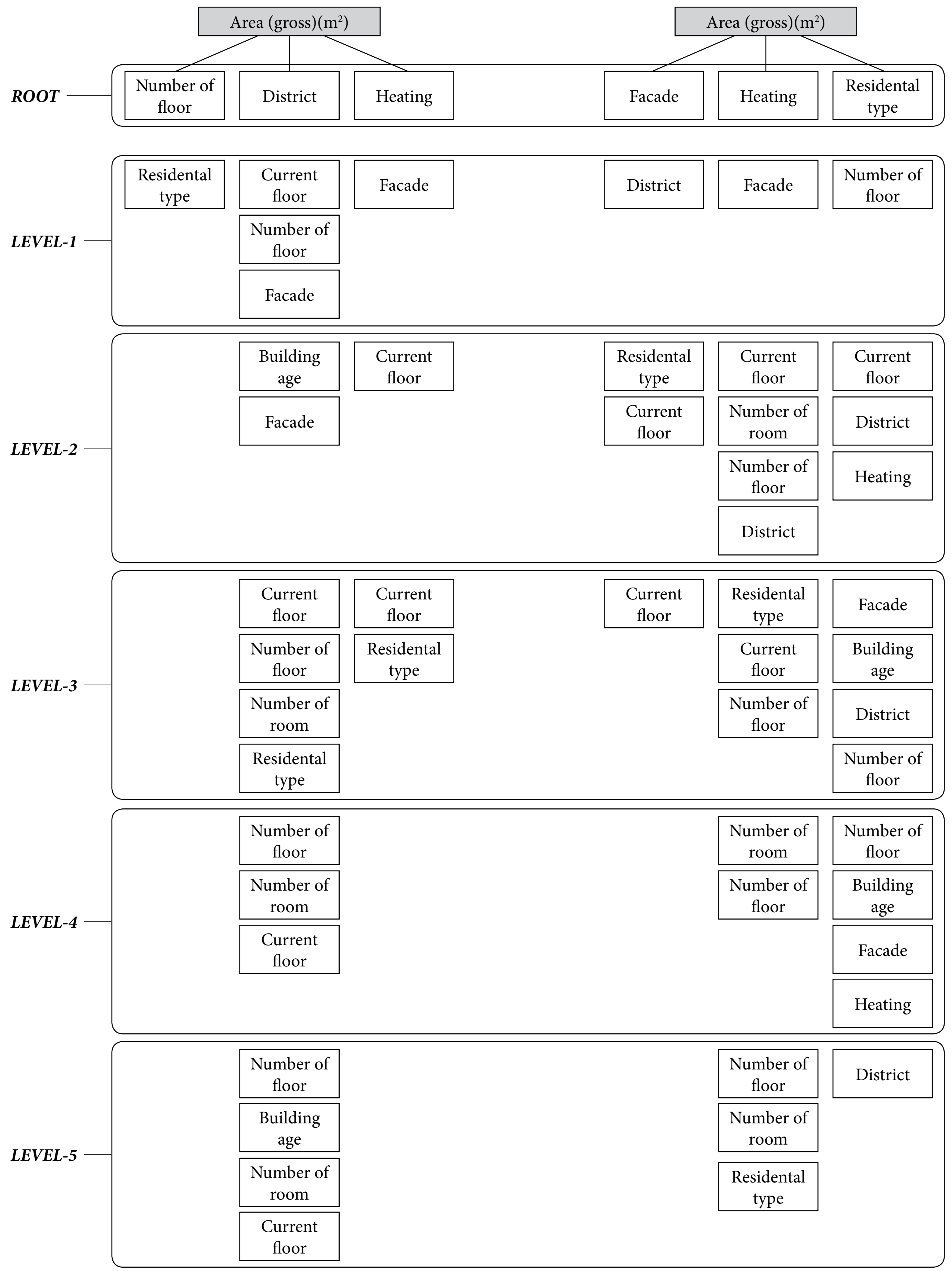

Figure 11. General representation of decision trees.

In the February 2023 dataset, there are 714 data in the low class, 538 data in the medium class, and 104 data in the high class. The accuracy of the decision tree was calculated as 74.63%. When the relevant tables are analyzed, it is seen that the model for 2023 has a better classification performance.

DISCUSSION AND CONCLUSION'

As houses in the region were directly or indirectly damaged after the 7.7 and 7.6 earthquakes in Kahramanmaraş and Gaziantep, Türkiye on February 6, 2023, the real estate sector in the region was also affected. The earthquake caused changes in house prices and variables affecting these prices. Price changes and the variables affecting the price were analyzed using machine learning.

The decision tree approach is used to compare the price change of house sales and variables affecting the price for both pre- and post-earthquake datasets. The central districts of Adana province (Seyhan, Yüreğir, Sarıçam, Çukurova) were chosen as the study area because it is the largest province in the earthquake zone, yet it suffered relatively less damage.

The prediction performance of the pre-earthquake model is 71.33%, while it is 74.63% in post-earthquake. For both models, the first branching occurred according to the area variable. The importance of this feature can be explained by the high population in the region. The socio-economic structure in the region may cause large families to live in a single house, and therefore structures with a high number of rooms may be preferred.

Features such as the number of rooms, district, and heating type were the determinants of housing preferences before the earthquake. Facade, heating type, and housing type were the reasons for preference in post-earthquake housing purchases. Due to the fact that Adana province is in a temperate region, housing purchase preferences, especially with air conditioning heating, have come to the fore. After the earthquake, it has been important in the sense that it strengthens the idea that there will be preferences for low-rise undamaged buildings, especially for detached houses. The preference for detached houses after the earthquake has automatically brought the facade feature to the forefront because detached houses have at least three facades. In this way, in addition to earthquake resistance standards, the variables revealed by the model can also be given importance in the new construction to be built in the region.

This study analyzes not only the earthquake effect but also the changes in house sales preferences before and after the earthquake and their impact on pricing. It cannot be said that these concepts are directly caused by the earthquake effect, but all the prominent features in house sales preferences are evaluated according to the results of decision trees.

Decision tree models were constructed to analyze the changes in house prices before and after the earthquake and to find the variables affecting the prices. Covering 11 provinces, Adana was chosen to be a quick representation and preliminary study of the disaster area. The study is a pioneering study that shows that machine learning models can be used successfully to reveal the price differences before and after the earthquake and to identify the variables affecting this change.

In order to reflect the overall situation in the earthquake region, more comprehensive datasets should be used, including the variability in the destruction rate and socioeconomic structure of all provinces in the region. The housing characteristics obtained for the creation of the database used in the study are limited to the characteristics provided by the open-source website. Another limiting factor is that the datasets were not collected in more than one month. The web scraper was able to access February 2022 and February 2023 housing data for a month before and after the earthquake. It is possible to obtain different results according to the damage in different areas of the earthquake zone and the data on the real estate website.

In parallel with increasing data size and detail, different machine learning models may also need to be used. Efforts have been initiated to create relevant datasets and model them with different machine learning methods.

ETHICS: There are no ethical issues with the publication of this manuscript.

PEER-REVIEW: Externally peer-reviewed.

CONFLICT OF INTEREST: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FINANCIAL DISCLOSURE: The authors declared that this study has received no financial support.

REFERENCES

Aliferis, C., & Simon, G. (2024). Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI. In G. Y. Simon, & C. Aliferis (Eds.). Artificial intelligence and machine learning in health care and medical sciences: Best practices and pitfalls (pp. 477–524). Springer.

Beron, K. J., Murdoch, J. C., Thayer, & M. A., Vijverberg, W. P. (1997). An analysis of the housing market before and after the 1989 Loma Prieta Earthquake. Land Econ, 73(1), 101–113.

Brookshire, D. S., Thayer, M. A., Tschirhart, J., & Schulze, W. D. (1985). A test of the expected utility model:

Evidence from earhquake risks. J Pol Econ, 93(2), 369–389.

Caglar, N., Vural, I., Kirtel, O., Saribiyik, A., & Sumer, Y. (2023). Structural damages observed in buildings after the January 24, 2020 Elazığ-Sivrice Earthquake in Türkiye. Case Stud Constr Mater, 18, e01886.

Dal Zilio, L., & Ampuero, J. P. (2023). Earthquake doublet in Turkey and Syria. Commun Earth Environ, 4, 71.

Fekrazad, A. (2019). Earthquake-risk salience and housing prices: Evidence from California. J Behav Exp Econ, 78, 104–113.

HepsiEmlak. (2023). Satılık Konut. https://www.hepsiemlak.com/satilik

Hidano, N., Hoshino, T., & Sugiura, A. (2015). The effect of seismic hazard risk information on property prices: Evidence from a spatial regression discontinuity design. Reg Sci Urban Econ, 53, 113–122.

Jung, J. H., & Smith, G. (2022). Earthquakes and home prices: The Napa and Ridgecrest Quakes. J Hous Res, 31(1), 74–87.

Koster, H. R., & Van Ommeren, J. (2015). A shaky business: Natural gas extraction, earthquakes and house prices. Eur Econ Rev, 80, 120–139.

KRDAE. (2023). Boğaziçi Üniversitesi Kandilli Rasathanesi Ve Deprem Araştirma Enstitüsü 6 Şubat 2023 Mw=7.6 Ekinözü Kahramanmaraş Depremi. Bölgesel Deprem-Tsunami İzleme ve Değerlendirme Merkezi. http://www.koeri.boun.edu.tr/sismo/2/06-subat-2023-ml7-5-ekinozu-kahramanmaras-depremi/

Lara-Pulido, J. A., Arias-Martelo, C., Jiménez-Quiroga, C. I., Soloaga, I., Uribe-Sánchez, M., & Guevara-Sanguinés, A. (2022). Short-and mid-term effects of the September 19 earthquake in Mexico City on the housing market. Int J Disaster Risk Reduct, 79, 103178.

Montesinos López, O. A., Montesinos, A., & Crossa, J. (2022). Overfitting, model tuning, and evaluation of prediction performance. In O. A. M. Lopez, A. M. Lopez, & J. Crossa, (Eds.). Multivariate statistical machine learning methods for genomic prediction (pp. 109–139). Springer.

Murdoch, J. C., Singh, H., & Thayer, M. (1993). The impact of natural hazards on housing values: The Loma Prieta Earthquake. Real Estate Econ, 21(2), 167–184.

Nakagawa, M., Saito, M., & Yamaga, H. (2007). Earthquake risk and housing rents: Evidence from the Tokyo metropolitan area. Reg Sci Urban Econ, 37(1), 87–99.

Nakagawa, M., Saito, & M., Yamaga, H. (2009). Earthquake risks and land prices: Evidence from the Tokyo metropolitan area. Jpn Econ Rev, 60, 208–222.

Naoi, M., Seko, M., & Sumita, K. (2009). Earthquake risk and housing prices in Japan: Evidence before and after massive earthquakes. Reg Sci Urban Econ, 39(6), 658–669.

Paul, J., & Thomas, P. Y. (2016). Sharma-mittal entropy properties on record values. Statistica, 76(3), 273–287.

Salzberg, S. L. (1994). C4.5: Programs for machine learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. Mach Learn, 16, 235–240.

Singh, R. (2019). Seismic risk and house prices: Evidence from earthquake fault zoning. Reg Sci Urban Econ, 75, 187–209.

Solanki, Y. S., Chakrabarti, P., Jasinski, M., Leonowicz, Z., Bolshev, V., Vinogradov, A., & Nami, M. (2021). A hybrid supervised machine learning classifier system for breast cancer prognosis using feature selection and data imbalance handling approaches. Electronics, 10(6), 699.

TCCSBB. (2023). Türkiye Cumhuriyeti Cumhurbaşkanlığı Strateji ve Bütçe Başkanlığı: Kahramanmaraş ve Hatay Depremleri Raporu. https://www.sbb.gov.tr/wp-content/uploads/2023/03/2023-Kahramanmaras-ve-Hatay-Depremleri-Raporu.pdf

TÜİK. (2023a). Adrese dayalı nüfus kayıt sistemi sonuçları, 2021. https://data.tuik.gov.tr/Bulten/Index?p=Adrese-Dayali-Nufus-Kayit-Sistemi-Sonuclari-2021-45500

TÜİK. (2023b). İstatistiksel Tablolar, Adrese Dayalı Nüfus Kayıt Sistemi, İl yaş ve cinsiyete göre nüfus. https://data.tuik.gov.tr/Bulten/Index?p=Adrese-Dayali-Nufus-Kayit-Sistemi-Sonuclari-2021-45500

Yücebaş, S., Doğan, M., & Levent, G. (2022). A C4. 5–Cart decision tree model for real estate price prediction and the analysis of the underlying features. Konya J Eng Sci, 10(1), 147–161.

Wikipedia. (2023). Adana. https://tr.wikipedia.org/wiki/Adana