Article

A Data Mining Study on House Price in Central Regions of Taiwan Using Education Categorical Data, Environmental Indicators, and House Features Data

Min-feng Lee [1,*], Guey-shya Chen [2], Shao-pin Lin [3] and Wei-jie Wang [3]

1   The National Museum of Natural Science, Taichung City 404023, Taiwan
2   The Institute of Educational Information and Statistics, National Taichung University of Education, Taichung City 40306, Taiwan; grace@mail.ntcu.edu.tw
3   Graduate Institute of Educational Information and Measurement, National Taichung University of Education, Taichung City 40306, Taiwan; bms110108@mail.ntcu.edu.tw (S.-p.L.); bms110100@mail.ntcu.edu.tw (W.-j.W.)
*   Correspondence: antonio@nmns.edu.tw

Abstract: This study takes the city of Taichung, Taiwan, as the research area, combines the survey results about the demand for residential houses for the next year, and uses relevant parameters and data of real price registration as the prediction results. In this study, eight types of school district features (such as teachers and students of secondary and elementary schools) and five types of air pollution features are selected and processed with a data mining method to discover the total transactions of real estate properties in various districts of Taichung. The results of K-means clustering and decision tree classification reveal that the four districts of the old Taichung City, namely, Beitun District, North District, Xitun District, and Nantun District, have houses meeting the conditions of egg yolk districts; houses in the old Taichung County have attributes of egg white districts. The results of decision tree classification show that the total price is the most important attribute influencing egg yolk and egg white districts.

Keywords: data mining; real price registration; school districts features; air pollution features

1. Research Background and Objectives

House prices have long been an issue of importance to Taiwanese society. The "justice of living" has been frequently brought up for discussion during each election. However, under the free market mechanism, buyers' and sellers' subjective will is not the same and sometimes hugely differ from each other, leading to the trend of continuously rising house prices in Taiwan. Under these circumstances, where a consensus on price is hard to achieve, the government has introduced the *Act of Real Price Registration* to improve information transparency and transaction equality of house transactions in hopes of alleviating the situation of transaction opacity.

In 2011, the Legislative Yuan passed the revised provisions for implementing the "three regulations of land administration on real price registration". The implementation concerns three regulations, the *Real Estate Broker Management Act*, the *Land Act*, and the *Equalization of Land Rights Act.* Buyers and sellers of real estate, relevant land administration agents, and real estate brokers must record the actual transaction price in the registration system; the above behaviors are called "three acts of land administration on real price registration". Within 30 days of the house transaction and completion of all procedures of ownership transfer, the proprietor must take the initiative to declare to authorities the relevant information, including the actual transaction price. The transfer of "land ownership" and "creation of pawning rights" can be excluded from the declaration. The proprietor and obligor must declare the current value of the land transfer within 30 days of the date of the deed. In the case of presale homes, the actual transaction price shall be submitted to authorities for auditing within 30 days of expiration and termination of the "commission contract".

According to the survey on the residential house demand trend in 2019 conducted by the Ministry of the Interior among people who planned to rent houses in the following year, over 40% are residents of Taichung. Moreover, the survey found that the total population of Taichung reached 2.815 million with a population growth of 11,000 new residents, and the general public has no professional knowledge of the real estate market. According to a report issued by Citibank in 2016, "in the real estate market in 2016, up to 60% of the buyers hope that the ideal price of the house they purchase is lower by 10% than the transaction price recorded in the real price registration system, while 60% of the sellers still believe that the market price is 10% higher or lower than the price recorded in the real price registration system. This shows that buyers have a high degree of expectation for a price cut, and there is a gap in terms of price awareness" [1]. There are also studies on other important factors affecting the price of rental houses, such as the financial crisis in 2008 and the COVID-19 emergency in 2020 [2].

According to the statistics of the World Health Organization (WHO), in 2019, about 7 million people worldwide died due to air pollution, higher than the combined population of five cities and counties in the central regions of Taiwan (Miaoli County, Taichung City, Changhua County, Nantou County, and Yunlin County). In 2016, the International Agency for Research on Cancer of the WHO classified fine suspended particulates (PM2.5) as a class-1 carcinogen, indicating that it is one of the main environmental factors contributing to cancer deaths. The higher the level of PM2.5 concentration in the air, the higher the relevant risks of lung cancer, stroke, ischemic heart disease, and chronic lung disease. The air pollution issue in Taichung City has been repeatedly reported by the media. Some architecture firms in the Taichung region also noticed the air pollution problem and started to provide air pollution protection equipment, including total heat exchangers with filters, external air filtration systems, and nanometer-level window screens, and used this as a new selling point of their projects. Certainly, consumers in Taichung City have included the air pollution issue as one of their considerations while purchasing houses.

Education has always been the biggest concern for Taiwanese people when it comes to the next generation. The education expenditure provided by parents starts to increase from the time that their children reach school age. It is highly possible that parents worldwide hope to live near schools so that they need not take their children to and pick them up after school. As a result, houses close to schools have good opportunities to sell at a better price. In Taiwan, junior high schools aim for the goals of "equal opportunities for education", "realization of national education", "popularization of education", and so forth.

This study uses the data of five monitoring stations in Taichung City set up by the Environmental Protection Agency of the Executive Yuan. The five stations are located in Xitun, Chongming, Fengyuan, Shalu, and Dali. The data are air quality index (AQI) data of ozone (O_3), fine suspended particulate matter (PM2.5), suspended particulate matter (PM10), carbon monoxide (CO), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2) during the years 2015–2018. According to the degree of impact of these pollutants on human health, this study calculates their sub-index respectively and then uses the maximum level of various sub-indexes of the day as the station's AQI on that day [3].

In line with the "Open Government" and "Open Data" principles, the Taichung Municipal government actively releases public data from various municipal authorities. This study uses the number of national primary and secondary schools, the number of male and female teachers, and the total number of students in each district of Taichung City from 2015 to 2019 provided by the Taichung City Department of Education [4].

The study processes the real price registration data with the data mining method, combines the features influencing house prices as concluded in the literature review, and discusses the features affecting house prices in the Taichung region. It is hoped that the study can provide the findings to house buyers for reference and also verify whether the features of real price registration can effectively serve as the basis for house valuation.

2. Research Literature: Real Price Registration

As the name suggests, "real price registration" means that the buyer and seller of real estate property and land administration agents must declare the actual transaction price into the registration system; this is called "three acts of land administration on real price registration". The proprietor shall take the initiative to declare to authorities the relevant information within 30 days of the house transaction and the completion of all procedures of ownership transfer, and the information shall include the actual transaction price while the transfer of "land ownership" and "creation of pawning rights" can be excluded from the declaration. The proprietor and obligor must declare the current value of the land transfer within 30 days of the date of the deed. In the case of presale homes, the actual transaction price shall be submitted to authorities for auditing within 30 days of expiration and termination of the "commission contract".

The relevant real estate information that buyers and sellers need to register is listed in Table 1.

Table 1. Data attributes of real price registration.

Target of Transaction	Target Information	Price Information
House number at a land section	Total area of land transfer	Total price of house transaction
House number at a building section	Total area of building transfer	Total price of real estate transaction
Immovable property mark	Total area of parking space transfer	Total price of building transaction
Number of buildings per transaction	Division of use area	Total price of parking space transaction
	Current layout of the building	Unit price per square meter
	Type of parking space	Year and month of the transaction
	Type of community management	

Zhu-hua, who is the author of Taiwan's real estate policy of "200,000 residential houses in the society in 8 years", has the following opinion on real price registration: "The implementation of real price registration can indeed effectively release concerns from the society about the data source of house price, and it also provides confidence to the government in releasing relevant statistical information. Although before the implementation of real price registration, in practice, some alternative data could be used to carry out the same analysis, and the results obtained did not necessarily deviate from the actual cases by a lot. Nonetheless, the information released by the government assuredly needs to conform to stricter 'public credibility' standards. More importantly, under the premise of 'public credibility,' the legitimacy, integrity, and universality of utilizing relevant information are also improved. This is the most precious and important significance of the real price registration system. This is a start for people to pay attention to the 'public credibility' of market information, and the core foundation for the integrity and universality of market information" [5].

The theory of "hedonic pricing" was put forth by Rosen in 1974 [6], and the hedonic demand function was developed with respect to buyers and sellers. Buyers pursue high product performance, and sellers pursue high prices, and the two parties will decide the product features and price. However, the market transaction price represents a balanced hedonic price; accordingly, the following equation was put forth [7]:

$$P_i = (X_{(1,i)}, X_{(2,i)}, \ldots, X_{(m,i)} + \varepsilon) \tag{1}$$

Since then, a number of foreign studies have used the hedonic pricing method, as summarized in Table 2.

Based on the above, commodity prices are influenced by various kinds of features. Moreover, when one feature changes, the commodity price changes as well. A study by Wu in 2020 on high-rise and low-rise collective residential buildings from 2012 to 2019 found that apartments on the 10th floor sell at a higher price in a high-rise building than in a low-rise one.

Table 2. Foreign research literature on the hedonic pricing method.

Researcher	Region	Research Time	Research Subject	Research Results
Estes and Smith (1996) [8]	Arizona, United States	1994	Fruits and vegetables	The price of fruits and vegetables will be affected by "packaging, size, and organic product label".
Combris, Lecocg And Visser (1997) [9]	Bordeaux, France	1992	Bordeaux wine	Wine price will be affected by the objective quality indicated on the bottle.
Gibbs, Halstead, Boyle And Hung (2002) [10]	New Hampshire, United States	1990–1995	Cleanness of lake water	Cleanness of lake water will affect the price of houses nearby.
Freccia, Jacobsen and Kilby (2003) [11]	Cigar production places	1992–1999	Cigars	The effect of cigars made in Cuba has the largest effect among all features.
T. Connell-Variy, B. Berggren, and T. McGough (2021) [12]	Queensland, Australia	2000–2018	Local mineral products	By comparing the resource reliance on the community in various countries regarding two independent resource areas, house price area is studied through resource relation.

2.1. House Features

Based on the hedonic pricing theory, the features of real estate property determine its price. The features can be divided into two major categories, internal and external features. Internal features include the bedroom, living room, shower, building area, floor, the land area of the entire building, house age, division of use area, building height, parking space, and building materials. There are five types of external features, "yes in my backyard" (YIMBY) facilities, "not in my backyard" (NIMBY) facilities, environment quality, population, and overall environment. Consumers prefer YIMBY facilities when choosing real estate property; such facilities include schools, parks, and other public facilities, as well as transportation facilities. More factors include road width, green space, and the distance between downtown and the workplace, which affect living convenience around the house. Some examples of NIMBY facilities are funeral parlors, crematoriums, waste yards, sewage treatment factories, and power substations. In terms of environmental quality, it includes the level of ambient noise, air quality, chances of flooding, and demographic structure of the community (e.g., education level, race, disposable income, and type of occupation). The overall factors are, for example, tax revenue, foreign exchange rate, stock index, consumer price index, and interest rate [13], and some studies also identified a significantly negative correlation between house price and marriage rate [14].

2.2. School District Features

In 1956, Tiebout suggested that in terms of the consideration of moving, the selection of school districts is included to achieve the result of "voting by feet" [15]. In the action of moving, the house transaction is the largest cost. In 1969, a study by Oates (1969) indicated that the number of people and the cost of going to school in the school district affect house prices [16]. Since then, much literature has proved that school districts affects house prices [17,18].

In terms of domestic research, the study by Ku found that in New Taipei city, elementary school district features have a significant impact on house prices in those areas with high real estate property prices. Several domestic and foreign sources mention that factors that measure the quality of a school district include examination scores, education expenditure per student, teaching experience of the faculty, ethnicity, and others. In Taiwan, the subsidies schools receive and data on students enrolling in a higher institution are not made public, so society widely perceives a sought-after school district as equivalent to a

good school district. Most people believe that compared with the regular school district, going to schools in a sought-after school district means better performance in the entrance exam to a higher school.

2.3. Air Quality Features

In 1990, the U.S. Congress revised and passed the *Clean Air Act*, which stipulates that environmental protection agencies must guarantee people's right to know about air quality. Therefore, the U.S. Environmental Protection Agency formulated the National Ambient Air Quality Standards. Moreover, to facilitate the understanding of the general public, the Pollutant Standards Index was developed [19].

The domestic studies on air quality and real estate property price features are summarized in Table 3.

Table 3. Relevant domestic studies on real estate property price features.

Author	Real Estate Property Price Features	Analysis Factor
Yeh (1993) [20]	Residential house transaction price in 1991	PM10
Lin (1992) [21]	Survey data of the Directorate General of Budget, Accounting, and Statistics in 1989	Air pollution and odor
Wu (1995) [22]	Adjusted residential house price in 1994	TSP
Gieng, Wang, and Lin (2000) [23]	Investigation of residential house status in Kaohsiung region in 1994	CO, PM10
Lin (2008) [24]	Town House Real Estates in the Old CBD Area of Taichung City	Housing prices and other housing features
Chen (2012) [25]	Central Taiwan Science Park on Local Housing Price from 2003 to 2012	Impact of Central Taiwan Science Park on Home Prices
Tasi (2015) [26]	The value assessment of climatic conditions and air quality in the Taiwan metropolitan area from 2003 to 2012	Temperature, Rainfall, Air Quality
Wu (2020) [27]	The effect of air pollution on housing prices in Taichung City from 2016 to 2018	Rainfall, Season, and Air Pollution Factors

2.4. Literature on Data Mining

"Data mining" can find information that has not been discovered before or that has potential value. In recent years, the trend of using big data technology to conduct mining on decipherable data has been growing. Table 4 summarizes the definitions by foreign scholars on data mining.

Table 4. Studies by scholars on data mining.

Scholar	Time	Definition
W. Frawley, et al. [28]	1992	Extract potentially useful and non-general information from the past unknown information implied by data.
D. Hand, et al. [29]	2001	Data mining is a science that searches for useful information from the big data database.
R. Grossman [30]	2001	Data mining uses a semi-automated extraction model on data to discover correlated and statistically meaningful datasets.
F. Guevara-Viejó, J. D. Valenzuela-Cobos, A. Grijalva-Endara, P. Vicente-Galindo, and P. Galindo-Villardón [31]	2022	The K-means clustering algorithm and PCA Biplot discover the result value stably produced through observation value of different parameters.
Y.-S. Chen, C.-K. Lin, Y.-S. Lin, S.-F. Chen, and H.-H. Tsao [32]	2022	This study consolidates the calculation of 7 kinds of data mining technologies, such as decision tree, Bayes, Function, Lazy, Meta, Mise, and Rule, and 23 kinds of important clustering algorithms (or classifier), and finds out the best classifier among them.

Source of data: Summarized by this study.

When applying statistical analysis, establishing a hypothetical model is often needed before conducting the research. However, this is not necessary for the field of data mining, so there is no predetermined standpoint and no need to establish a hypothesis. It only requires researchers to select the analysis and calculation method. Another feature of data mining is the unpredictability of the calculation result. The processes of data mining can be summarized into six steps, comprising data cleaning, data consolidation, data selection, data conversion, data mining, and explanation and validation [33].

3. Data Mining and Feature Engineering

The data mining methods used in the research are multiple linear regression, k-means, and decision tree. Initially, multiple linear regression was used to calculate the features correlation coefficients in the real price registration, and the features were clustered by the k-means method. Then the decision tree was used for clusters' condition classification. This chapter introduces in detail the data extraction, merging, and sampling methods of the data source, quantity, and method of data usage, and further explains the data mining tools in this research and the method of manipulating the data.

3.1. Data Extraction, Consolidation, and Sampling

The research period of this study was 2015–2019. We used the data made public by the government, employing the six steps of data mining to evaluate the impact of the school district and air quality on the transaction price of houses. Data extraction was conducted through the websites of real price registration, Taichung Municipal Education Bureau, and the public data website of the Environmental Protection Agency. The study consolidated the data by year, and the attribute data are given numerical values. K-means and decision tree methods were utilized to perform data mining, and finally, the explanation of the results.

There are three types of data in the study, education category data, environmental indicators, and housing characteristics. The data obtained from the real price registration are the housing characteristics, the data from the Taichung City Department of Education are the education category data, and the Environmental Protection Agency of the Executive Yuan's public website data are the environmental indicators. All indicators are shown in Figure 1.

housing characteristics	
• Administrative area	• Total square meters of building transfer
• Year of the transfer	• Number of living rooms
• Quarter of the transfer	• Number of bathrooms
• Parking space	• Month of the transaction
• Total price in NT$	• Number of bedrooms
• Total square meters of land transfer	• Unit price per square meter
• Floor of transfer	• Whether it has partition
• Main use	• House age

education category data	environmental indicators
• number of elementary schools	• O^3
• number of elementary school teachers	• PM 2.5
• number of elementary school students	• PM 10
• number of junior high schools	• CO
• number of junior high school teachers	• SO^2
• number of junior high school students	• NO^2

Figure 1. The three types of data in this research (education category data, environmental indicators, and housing characteristics).

In 2017, the Ministry of the Interior introduced the *Implementation Rules of Regulations on Accelerating the Reconstruction of Dilapidated and Old Buildings in Urban Area*, enabling buildings without elevators and older than 30 years in the planned urban area to apply for

reconstruction. Moreover, the buildings approved for reconstruction can have incentives for floor area ratio, better building coverage ratio, and tax reduction (exemption of the land-value tax during the construction period and 50% reduction of land-value tax and house tax for two years). Given the above reasons, the price of an eligible building in very old condition without an elevator can have similar transaction prices to houses with better conditions. Considering this, data mining excludes transactions of houses older than 30 years and without elevators. A total of 9785 sample cases obtained through data extraction in this research. Furthermore, in the education database, there is contact information and statistics of non-teacher staff of various schools; this study only extracts the number of schools, number of male and female teachers, and students, removing other kinds of data. The AQI of the Environmental Protection Agency is calculated based on the observed values of O_3, PM 2.5, PM 10, CO, SO_2, and NO_2. Therefore, this study also obtains various observation values of AQI in the air quality database, filtering out other items. Table 5 shows the extraction of real price registration samples.

Table 5. Extraction of real price registration samples.

Real Price Registration Item	Item Description
Administrative area	The administrative area where the building being transacted is located
Year of the transfer	The year when the transaction takes place
Quarter of the transfer	The quarter when the transaction takes place
Parking space	Form of the parking space
Total price in NT	Total transaction price
Total square meters of land transfer	Total floor area of the house
Floor of transfer	The floor where the house being transacted is located
Main use	Division of land-use area
Whether it has community management	Whether or not it has community management
Total square meters of building transfer	Indoor area of the house
Number of living rooms	Number of living rooms
Number of bathrooms	Number of bathrooms and toilets
Month of the transaction	The month when the transaction takes place
Number of bedrooms	Number of bedrooms
Unit price per square meter	Selling price per square meter of the architecture interior
Whether it has partition	Whether it has partition
House age	The gap between the year/month of the transaction and the year/month of completion

3.2. Data Mining Tools

Waikato Environment for Knowledge Analysis (WEKA), developed by the University of Waikato in New Zealand, is an open-source that uses JAVA language and can be applied in the fields of machine learning and data mining. This experiment uses the WEKA toolkit to conduct the data exploration process.

There are 17 item rows in the real price registration database, excluding the "total price in NT$", and there are 16 internal features that may affect the total transaction price. In the section on price features, it is mentioned that commodity price is composed of multiple attributes that have different impacts. To examine the degree of influence of various attributes, this study takes the 16 internal features as independent variables and the "total price in NT$" as the dependent variable. Regression analysis was used to calculate the relevant coefficients of various variables, as shown in Table 6.

The *p*-value for "internal characteristics" is less than the significance level of 0.025, which means that the "total value" of this attribute is significant. Then, comparing the correlation coefficient of the significant attributes, it can be known that the "total square meter of building transfer" has the highest value (0.445), indicating that the "total square meter of building transfer" has the biggest influence on the total transaction price.

The clustering of data requires researchers to determine the number of clusters. In the process of buying real estate property, brokers or commission agents will use the terms "egg

yolk district" and "egg white district" when introducing them to the intended consumers. The two terms classify the district where the house is located, that is, high-end district or affordable district. Some studies classify the residential house districts in the administrative area of Taipei city into two kinds, luxury mansion district and regular house district, and then further divide the luxury mansion district into the egg yolk district and egg white district [34].

Table 6. Coefficients of internal features.

Statistical Parameter\Descriptive Statistical Coefficient, R-Value at 0.913	Correlation Coefficients	Standard Deviation (SD)	t Value	*p*-Value	Significance Level α/2 = 0.025	Confidence Level (0.975)
Administrative area	−0.0131	0.001	−9.504	0	−0.016	−0.01
Total square meters of land transfer	0.0691	0.006	12.076	0	0.058	0.08
Year of the transfer	0.0078	0.002	3.351	0.001	0.003	0.012
Quarter of the transfer	−0.003	0.002	−1.775	0.076	−0.006	0
Floor of transfer	0.0168	0.004	3.999	0	0.009	0.025
Total floors of the building	0.0035	0.001	6.68	0	0.002	0.005
Main use	−0.0147	0.059	−0.247	0.804	−0.131	0.102
House age	−0.003	0	−10.302	0	−0.004	−0.002
Total square meters of building transfer	0.445	0.024	18.583	0	0.398	0.492
Number of bedrooms	0.0162	0	113.126	0	0.016	0.016
Number of living rooms	0.0081	0	26.198	0	0.008	0.009
Number of bathrooms	0.0025	0.0000385	65.42	0	0.002	0.003
Whether it has partition	0.0004	0	1.103	0.27	0	0.001
Whether it has community management	0.0061	0.01	0.618	0.537	−0.013	0.026
Unit price per square meter	0.0966	0.003	30.977	0	0.09	0.103
Parking space	0.0219	0.002	−12.153	0	−0.025	−0.018

The number of clustering in the "K-means method" is set based on the classification of "egg yolk district" and "egg white district"; the smaller the "within-cluster sum of squared errors (WSS)" is, the closer the distance from the falling point of each cluster to K point, and the better the clustering effect. We use the 17 "internal features" attributes for clustering, the 2 attributes with the highest WSS, and the correlation coefficient to obtain the values of WSS. After which, the study used the two attributes with the highest correlation coefficient, that is, "total square meter of building transfer" and "total price in NT$", to perform clustering. The result is that the WSS of using "total square meter of building transfer" and "total price in NT$" is the smallest. Because the smaller the WSS, the better the effect, the study adopted the clustering result of using "total square meter of building transfer" and "total price in NT$", as shown in Table 7. The output of WEKA is in AIFF format, which is a pure text file format used by WEKA; hence, it was converted into a CSV file to facilitate subsequent processing.

After performing k-means clustering on the data of 9785 transaction cases from 2015 to 2019, each transaction case has an additional cluster attribute, that is, Cluster0 for egg yolk district and Cluster1 for egg white district. Then, we performed the decision tree classification on the dataset with cluster attributes, and the input data are the following: administrative area, total square meters of land transfer, year of the transfer, quarter of the transfer, floor of transfer, total floors of the building, main use, house age, total square meters of building transfer, number of bedrooms, number of living rooms, number of

bathrooms, whether it has a partition, whether it has community management, the unit price per square meter of parking space, total price in NT$, number of elementary schools, number of elementary school teachers, number of elementary school students, number of junior high schools, number of junior high school students, O_3, PM 2.5, PM 10, CO, SO_2, and NO_2.

Table 7. WSS of clustering.

| The Attributes for Clustering and Their Quantity	Number of Clusters	Intra-Group Square and WSS
17 internal features	2	9203
17 internal features	3	8430
17 internal features	4	8220
17 internal features	5	7950
2 features, "total square meter of building transfer" and "total price in NT$"	2	22

We chose the decision tree algorithm J48 to classify data in WEKA, as shown in Figure 2. It derived the statistical data of attributes of the dataset, such as maximum value, minimum value, mean value, and SD. Take the example of "total price in NT$", its maximum value, minimum value, mean value, and SD are 113,680,000, 28,800, 11,327,445.555, and 8,285,525.513, respectively. After the J48 algorithm classified that present the results of accuracy, correctly classified instances, and incorrectly classified instances. Of them, 9775 cases are correctly classified instances, accounting for 99.8978% of the total cases, and 10 are incorrectly classified instances, accounting for 0.1022% of the total cases. The tree has 7 leaves, and its size is 13, with a calculation time of 0.06 s.
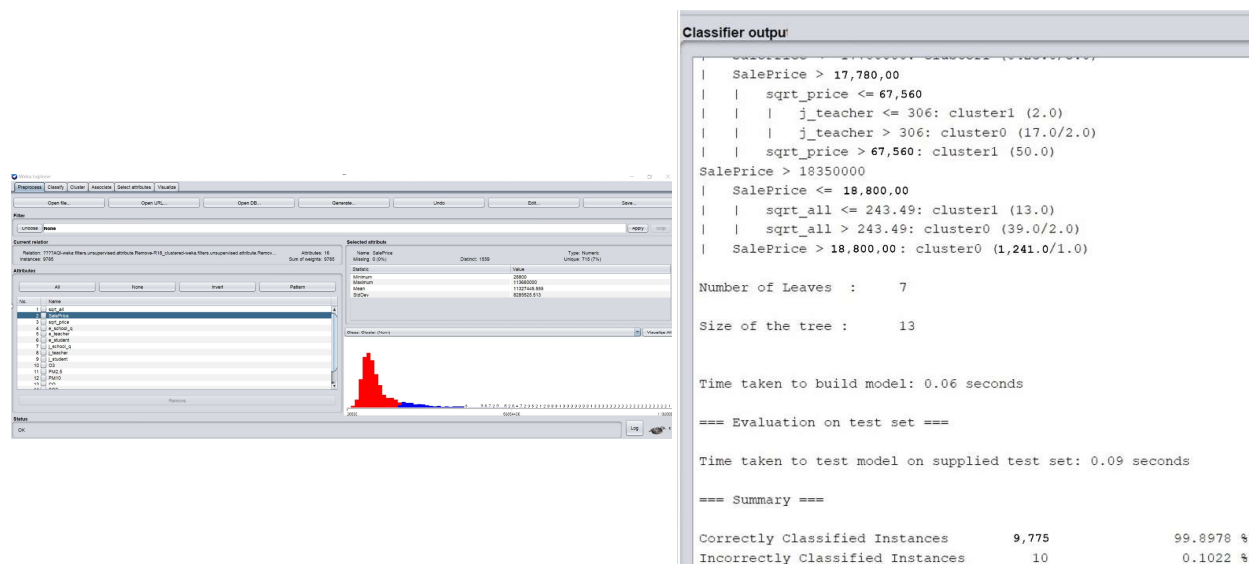
Figure 2. Decision Tree function in WEKA.

4. Results

The data period of "real price registration" was 2015–2019; in the database of real price registration, buildings older than 30 years with less than 11 floors are excluded. In total, there are 9785 transaction cases, and their distribution by administrative area is illustrated by the bar chart in Figure 3.
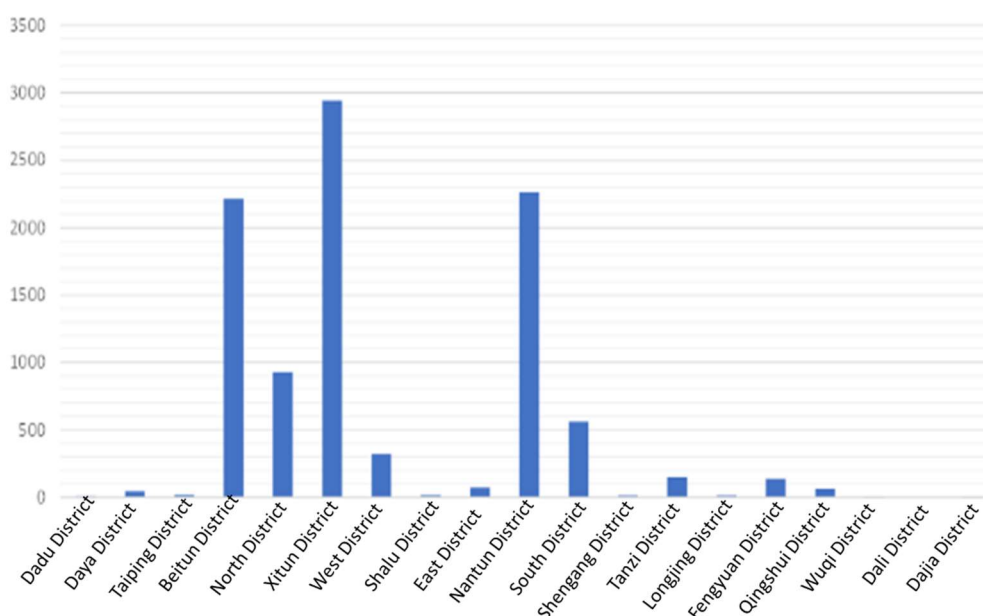
Figure 3. Bar chart of the number of transactions by administrative area.

　　　From 2015 to 2019, there were 9785 transaction cases that met the requirements; that is, the building has at least 11 floors and a house age of fewer than 30 years. The samples were distributed in 19 administrative areas of Taichung City, of which Xitun District, Nantun District, and Beitun District had the largest number of transaction cases. In terms of the total price, the highest was found in one case (NT\$ 113,680,000) in Xitun District, while the lowest was found in two cases in North District (NT\$ 28,800), and the mean of the total price was NT\$ 11,327,445.

4.1. Results of Clustering by k-Means

　　　The k-means method classifies the clusters of "real price registration" into egg yolk districts and egg white districts. Egg yolk districts have features such as large building areas, high unit prices, and high total prices, while egg white districts are relatively low in building area, unit price, and total price. Regarding the difference in the mean values of "total square meter of building transfer", "unit price per square meter", and "total price in NT\$" between egg yolk districts and egg white districts, egg yolk districts have much larger building area and much higher unit price per square meter as well as higher total price than egg white districts.

　　　Of the six observation items of "air quality features", only the O_3 level is slightly lower in the yolk regions than in the egg white regions, and the rest of the indicators in egg yolk districts are all slightly higher than those in egg white districts. Overall, the difference in air quality between the two kinds of districts is not significant; in other words, there is no difference.

　　　The study uses the algorithm of k-means to sort out two clusters, that is, egg white districts and egg yolk districts. Then, the clustering result of the data is summarized by administrative area, as shown in Table 8. From 2015 to 2019, egg yolk districts had a total of 1297 transaction cases distributed in 6 administrative areas, while egg yolk districts had a total of 8488 transaction cases distributed in 19 administrative areas. Xitun District, Nantun District, and Beitun District had the most egg yolk districts, and Xitun District, Beitun District, and Nantun District had the most egg white districts.

Table 8. Distribution of egg yolk districts and egg white districts by administrative area.

Administrative Area	Egg Yolk District	Egg White District	Total
Dadu District	0	6	6
Daya District	0	46	46
Taiping District	0	18	18
Beitun District	85	2131	2216
North District	92	838	930
Xitun District	615	2328	2943
West District	35	289	324
Shalu District	0	17	17
East District	0	72	72
Nantun District	469	1795	2264
South District	1	561	562
Shengang District	0	14	14
Tanzi District	0	150	150
Longjing District	0	14	14
Fengyuan District	0	136	136
Qingshui District	0	62	62
Wuqi District	0	5	5
Dali District	0	2	2
Dajia District	0	4	4
Total	1297	8488	9785

In December 2010, the old Taichung City and old Taichung County were merged into the Taichung City of today. From the distribution of egg yolk districts and egg white districts by administrative area, it can be found that in the administrative areas that once belonged to the old Taichung County, no house fulfills the attributes of the egg yolk district, and all houses with the attributes of egg yolk district are located in administrative areas that were part of the old Taichung City. The output of the decision tree has 13 tree nodes and 7 leaf nodes. In terms of the number of "correctly classified instances" and "incorrectly classified instances", there are 9775 and 10 cases, respectively.

4.2. Results of Decision Tree Rules

The classification rules of the decision tree are the following:

Rule 1: The total price is below NT$ 17,780,000. In total, 8418 cases eligible for this condition belong to egg white districts, accounting for 99% of the total transaction cases of egg white districts. The total transaction price is not affected by school district features or air quality features. The other five cases fulfilling this condition belong to egg yolk districts.

Rule 2: The total price is between NT$ 17,780,000 and NT$ 18,350,000, the unit price per square meter of the real estate property is lower than NT$ 67,560, and there are less than 306 junior high school teachers in the administrative area where the real estate property is located. There are 2 cases fulfilling such conditions in egg white districts, accounting for less than 1% of the total transaction cases in egg white districts.

Rule 3: The total price is between NT$ 17,780,000 and NT$ 18,350,000, the unit price per square meter of the real estate property is lower than NT$ 67,560, and there are more than 306 junior high school teachers in the administrative area where the real estate properties are located. Moreover, 15 cases fulfilling such conditions are in egg yolk districts, accounting for 1.1% of the total transaction cases in egg yolk districts. There are 2 cases fulfilling such conditions in egg white districts, accounting for less than 1% of the total transaction cases in egg white districts.

Rule 4: The total price is between NT$ 17,780,000 and NT$ 18,350,000, and the unit price per square meter of the real estate property is higher than NT$ 67,560. There are 50 cases fulfilling such conditions in egg white districts, accounting for 0.5% of the total transactions in egg white districts.

Rule 5: The total price is between NT$ 18,350,000 and NT$ 18,800,000, and the total building area of the real estate property is less than 243.49 square meters; 13 cases fulfilling

such conditions are in egg white districts, accounting for 0.1% of the total transactions in egg white districts.

Rule 6: The total price is between NT$ 18,350,000 and NT$ 18,800,000, and the total building area of the real estate property is more than 243.49 square meters; 37 such cases are in egg yolk districts. There are 2 such cases in egg white districts, accounting for 0.02% of the total transactions in egg white districts.

Rule 7: The total price is higher than NT$ 18,800,000; 1240 cases fulfilling such conditions are in egg yolk districts, accounting for 95% of the total transactions in egg yolk districts. One case is in the egg white district.

5. Conclusions and Discussion

This study adopts the data mining method to interpret the phenomena that can be demonstrated by the transaction data of real price registration, categorical education data, and environmental indicators, aiming to provide consumers a judgment basis in addition to speculating price features of houses in egg yolk districts and egg white districts. Moreover, it can provide cross-references with studies on hedonic pricing, such as the research on the impact of airplane noise on the quality of life of residents and the structure of houses close to the airport [35].

The limitation of this research is mainly due to the fact that although the real price registration database has the registration section house number, the Taiwan house number code is messy, and there is no conversion system that is accurate and can handle large amounts of data. The house number is converted into latitude and longitude, so it is impossible to judge the influence of the total transaction price caused by the external characteristics of real estate distance and price characteristics.

In addition, the data of this study can be used as the basis for future research, and other price features that affect the transaction price can be added so that both real estate buyers and sellers can more comprehensively understand that real estate prices in Taichung City will be affected by those characteristics, the degree of influence, and the indirect contribution. For example, the distance characteristics of buildings to schools, the characteristics of roads adjacent to buildings, and the distinction between construction before and after the 1999 the 921 Taiwan earthquake can be increased in terms of time conditions. Based on this study, areas and houses with affordable and good housing can be classified.

5.1. Features of House Price District

Egg yolk districts have 1297 transaction cases in total; of them, 615 cases are in Xitun District, 469 in Nantun District, and 92 in North District, ranking top 3 in the number of cases. Xitun District has the largest number of buyers. Although North District ranks third in the number of buyers, compared with Beitun District, the difference in the number of transaction cases is merely 7 according to the 5-year statistics. Before clustering, the number of transaction cases in Xitun District, Nantun District, North District, and Beitun District is 2943, 2264, 930, and 2216, respectively. Although the difference in the number of buyers between Beitun District and North District is merely seven, the total number of house buyers in North District is far less than that of Beitun District, which shows that the number of houses in egg yolk districts in North District is similar to that in Beitun District, but the number of house sellers of North District is much less than that of Beitun District. In recent years, Beitun District has developed many rezoning areas. Because the North District was developed quite early, it contains few large construction sites. Xitun District has the seventh stage of rezoning area, and Nantun District has the eighth stage of rezoning area, which explains why they have most of the transaction cases fulfilling conditions of the egg yolk district.

5.2. Features of Education Category Data

On average, the building area of houses in egg yolk districts is larger than that of the egg white districts, and the unit price per square meter of the former is also much

higher than that of the latter. However, the number of schools nearby houses is more or less the same in the two kinds of districts. In terms of teachers and students of secondary and elementary schools, more of them are located in egg yolk districts. However, the total transaction cases of egg yolk districts are less than that of egg white districts by 7191. The interpretation from the results of data mining indicates that education practitioners and families with secondary/elementary school students are willing to spend more resources in the selection of house areas, thereby choosing to stay in egg yolk districts.

5.3. Air Quality

In terms of air quality features, the influence on egg yolk districts and egg white districts is similar, and the influence is not significant compared with that of school district features. As a result, air quality does not significantly impact the number and price of house transactions in Taichung City.

5.4. Attribute Features

Regarding the results of clustering and the results of the decision, only 10 transaction cases are classified differently. Based on the results of the decision tree, in terms of attributes influencing the classification of egg yolk districts and egg white districts, the most influential attribute is the total price in NT$ of the real estate property. If NT$ 18,350,000 is used as the division criteria, then 9658 cases can be filtered out, accounting for 98% of the total transaction cases. Furthermore, only 127 cases are affected by the attributes of "unit price per square meter of the real estate property", "total square meters of the building", and "number of secondary school teachers".

Author Contributions: Conceptualization, M.-f.L. and G.-s.C.; methodology, M.-f.L. and G.-s.C.; software, S.-p.L. and W.-j.W.; validation, M.-f.L.; formal analysis, M.-f.L. and G.-s.C.; investigation, M.-f.L. and G.-s.C.; resources, M.-f.L.; data curation, S.-p.L. and W.-j.W.; writing—original draft preparation, S.-p.L. and W.-j.W.; writing—review and editing, M.-f.L. and G.-s.C.; visualization, S.-p.L. and W.-j.W.; supervision, M.-f.L. and G.-s.C.; project administration, M.-f.L. and G.-s.C.; All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Citibanker. 2016 Global Market Outlook Adapting to Local Conditions and Flexible Layout. Available online: https://www.citibank.com.tw/sim/citigold/pdf/citibanker-2016-spring.pdf (accessed on 17 March 2022).
2. Tajani, F.; Di Liddo, F.; Ranieri, R.; Anelli, D. An automatic tool for the determination of housing rental prices: An analysis of the Italian context. *Sustainability* **2021**, *14*, 309. [CrossRef]
3. R.O.C. Environmental Protection Administration Executive Yuan and Taiwan. Environmental Protection Administration Environmental Information Open Platform. Available online: https://data.epa.gov.tw/ (accessed on 17 March 2022).
4. Education Bureau of Taichung City Government. Available online: https://english.taichung.gov.tw/education (accessed on 17 March 2022).
5. Hua, C.-C. The Importance of Real Price Registration Data to the Compilation and Release of House Price Affordability Indicators. Available online: https://blog.xuite.net/fullland/twblog/173123430-%E5%AF%A6%E5%83%B9%E7%99%BB%E9%8C%84%E8%B3%87%E6%96%99%E5%B0%8D%E6%88%BF%E5%83%B9%E8%B2%A0%E6%93%94%E8%83%BD%E5%8A%9B%E6%8C%87%E6%A8%99%E7%B7%A8%E8%A3%BD%E8%88%87%E7%99%BC%E5%B8%83%E7%9A%84%E9%87%8D%E8%A6%81%E6%80%A7 (accessed on 17 March 2022).
6. Rosen, S. Hedonic prices and implicit markets: Product differentiation in pure competition. *J. Pol. Econ.* **1974**, *82*, 34–55. [CrossRef]

7. Gu, M.-F. The Effect of Characteristics of Elementary Schools on House Price—The Case of High-Price Districts of New Taipei City. Available online: https://ndltd.ncl.edu.tw/cgi-bin/gs32/gsweb.cgi/login?o=dnclcdr&s=id=%22107SHU00389008%22.& searchmode=basic (accessed on 17 March 2022).
8. Estes, E.A.; Smith, V.K. Price, quality, and pesticide related health risk considerations in fruit and vegetable purchases: An hedonic analysis of Tucson, Arizona supermarkets. *J. Food Distrib. Res.* **1996**, *27*, 59–76.
9. Combris, P.; Lecocq, S.; Visser, M. Estimation of a hedonic price equation for Bordeaux wine: Does quality matter. In *World Scientific Reference on Handbook of the Economics of Wine: Volume 1: Prices, Finance, and Expert Opinion*; World Scientific: Singapore, 1997; pp. 167–183. [CrossRef]
10. Gibbs, J.P.; Halstead, J.M.; Boyle, K.J.; Huang, J.-C. An hedonic analysis of the effects of lake water clarity on New Hampshire lakefront properties. *Agric. Resour. Econ. Rev.* **2002**, *31*, 39–46. [CrossRef]
11. Freccia, D.M.; Jacobsen, J.P.; Kilby, P. Exploring the relationship between price and quality for the case of hand-rolled cigars. *Q. Rev. Econ. Financ.* **2003**, *43*, 169–189. [CrossRef]
12. Connell-Variy, T.; Berggren, B.; McGough, T. Housing markets and resource sector fluctuations: A cross-border comparative analysis. *Sustainability* **2021**, *13*, 8918. [CrossRef]
13. Lin, J.-J.; Chang, Y.-C. The Shop Rents Analysis of Underground Arcades in Taipei Metro System: Application of Hedonic Price Approach. Available online: https://www.airitilibrary.com/Publication/alDetailedMesh?docid=16068238-200606-7-1-47-69-a (accessed on 17 March 2022).
14. González-Val, R. House prices and marriage in Spain. *Sustainability* **2022**, *14*, 2848. [CrossRef]
15. Tiebout, C.M. A pure theory of local expenditures. *J. Pol. Econ.* **1956**, *64*, 416–424. [CrossRef]
16. Oates, W.E. The effects of property taxes and local public spending on property values: An empirical study of tax capitalization and the Tiebout hypothesis. *J. Pol. Econ.* **1969**, *77*, 957–971. [CrossRef]
17. Reback, R. House prices and the provision of local public services: Capitalization under school choice programs. *J. Urban Econ.* **2005**, *57*, 275–301. [CrossRef]
18. Gravel, N.; Michelangeli, A.; Trannoy, A. Measuring the social value of local public goods: An empirical analysis within Paris metropolitan area. *Appl. Econ.* **2006**, *38*, 1945–1961. [CrossRef]
19. Lin, L.-W. Applying the Hedonic Price Method to Assess the Benefits of Air Quality Improvement in Taiwan's Metropolitan Area. Available online: https://ndltd.ncl.edu.tw/cgi-bin/gs32/gsweb.cgi/login?o=dnclcdr&s=id=%22091NTPU0399001%22. &searchmode=basic (accessed on 17 March 2022).
20. Yeh, H.S. Estimating the Impact of Air Pollution on Housing Price—An Application of Hedonic Price Method. Available online: https://ndltd.ncl.edu.tw/cgi-bin/gs32/gsweb.cgi/login?o=dnclcdr&s=id=%22081NCCU0303007%22.&searchmode= basic (accessed on 17 March 2022).
21. Qiu, Z.H. A Study of Housing Imputed Rent in Taipei City and Taiwan. Available online: http://nccur.lib.nccu.edu.tw/handle/ 140.119/64366 (accessed on 17 March 2022).
22. Sent-ian, W. Price Estimation of Air Pollution in Taipei Metropolitan Area—Application of Hedonic Price Method. Available online: https://ndltd.ncl.edu.tw/cgi-bin/gs32/gsweb.cgi/ccd=IP8Kq7/record?r1=1&h1=0 (accessed on 17 March 2022).
23. Chiang, Y.-S.; Wang, S.-E.; Lin, Y.-L. The Direct Effect of the Air Pollution Control Fees on Air Quality Improvement. Available online: https://tpl.ncl.edu.tw/NclService/JournalContentDetail?SysId=A00018855&ji%5B0%5D=%E9%81%8B%E8%BC%B8 %E8%A8%88%E5%8A%83&cn%5B0%5D=567&q%5B0%5D.f=KW&q%5B0%5D.i=%E7%A9%BA%E6%B0%A3%E6%B1%A1 %E6%9F%93&page=1&pageSize=1&orderField=score&orderType=desc (accessed on 17 March 2022).
24. Lin, C.-W. A Spatial Analysis of Land Price Based on the Hedonic Price Theory: With the Case Study of the Town House Real Estates in the Old CBD Area of Taichung City in 2008. Master's Thesis, Graduate Institute of Earth Science, Chinese Culture University, Taipei, Taiwan, 2010. Available online: https://hdl.handle.net/11296/r343he (accessed on 19 May 2022).
25. Chen, S.-M. The Effect of Central Taiwan Science Park on Local Housing Price Using Hedonic Price Method. Master's Thesis, Tunghai University, Taichung, Taiwan, 2012. Available online: https://hdl.handle.net/11296/wjdrab (accessed on 20 April 2022).
26. Tsai, M.-C. The Valuation of Climate and Air Quality in Taiwan—An Application of the Hedonic Price Method. Master's Thesis, Insitiue of Natural Resources Management, National Taipei University, Taipei, Taiwan, 2015. Available online: https: //hdl.handle.net/11296/r6fn54 (accessed on 20 April 2022).
27. Wu, Y.-P. The Impact of Air Pollution on Housing Price—A Case Study of Taichung City. Master's Thesis, Business Administration, National Chung Hsing University, Taichung, Taiwan, 2020. Available online: https://hdl.handle.net/11296/3z9jeb (accessed on 20 April 2022).
28. Frawley, W.J.; Piatetsky-Shapiro, G.; Matheus, C.J. Knowledge discovery in databases: An overview. *AI Mag.* **1992**, *13*, 57. [CrossRef]
29. Hand, D.; Mannila, H.; Smyth, P. *Principles of Data Mining*; MIT Press: Cambridge, MA, USA, 2001.
30. Yehuda, R.; Halligan, S.L.; Grossman, R. Childhood trauma and risk for PTSD: Relationship to intergenerational effects of trauma, parental PTSD, and cortisol excretion. *Dev. Psychopathol.* **2001**, *13*, 733–753. [CrossRef] [PubMed]
31. Guevara-Viejó, F.; Valenzuela-Cobos, J.D.; Grijalva-Endara, A.; Vicente-Galindo, P.; Galindo-Villardón, P. Data mining techniques: New method to identify the effects of aquaculture binder with sardine on diets of juvenile litopenaeus vannamei. *Sustainability* **2022**, *14*, 4203. [CrossRef]

32. Chen, Y.-S.; Lin, C.-K.; Lin, Y.-S.; Chen, S.-F.; Tsao, H.-H. Identification of potential valid clients for a sustainable insurance policy using an advanced mixed classification model. *Sustainability* **2022**, *14*, 3964. [CrossRef]

33. Li, M.-F. Analyzing the Learner's Emotions and Color Relation Framework Uses Data Mining Models. Available online: https://ndltd.ncl.edu.tw/cgi-bin/gs32/gsweb.cgi?o=dnclcdr&s=id=%22104NTCT0629001%22.&searchmode=basic (accessed on 17 March 2022).

34. Chiang, M.-C. Can Luxury Tax Effectively Suppress Rising Housing Prices?—A Case Study in Taipei Residence. Available online: https://ndltd.ncl.edu.tw/cgi-bin/gs32/gsweb.cgi/login?o=dnclcdr&s=id=%22103YUNT0304007%22.&searchmode=basic (accessed on 17 March 2022).

35. Tsao, H.-C.; Lu, C.-J. Assessing the impact of aviation noise on housing prices using new estimated noise value: The case of Taiwan Taoyuan International Airport. *Sustainability* **2022**, *14*, 1713. [CrossRef]