Prediction Value of a Real Estate in the City of Quito Post Pandemic

Wladimir Vilca(✉), Joe Carrion-Jumbo, Diego Riofrío-Luzcando, and César Guevara

Universidad Internacional SEK, Digital School Faculty, Calle Italia N31 - 125 y Av. Mariana de Jesús, Quito, Ecuador
wladimir.vilca@uisek.edu.ec
https://www.uisek.edu.ec/

Abstract. Many real estate projects were paralyzed due to a lack of funding, and sales dropped significantly due to COVID-19. This article provides a method to predict the value of real estate in the city of Quito post-pandemic, using a methodology that compares different data mining techniques to achieve the best accuracy. In the end, it has been possible to classify the properties in different sectors of the population under study with a good level of value prediction. It can be concluded that the study based on a review of the appropriate literature, the comparison of different techniques, and the segmentation of the population is a basis for other studies that apply other techniques to further improve the level of prediction.

Keywords: machine learning · data · real estate · pandemic · appraisal

1 Introduction

The real estate area has been affected by the pandemic, and unemployment has become part of the deterioration of people's economies. "The pandemic has been a catalyst for changes in the perception of housing and in buyers' habits... single-family home sales skyrocketed during the first 7 months of 2021 compared to previous years", [6] said a professor at EAE Business School. The formation of real estate prices responds to the same logic as that of financial assets, where speculation is common [14].

According to the Association of Real Estate Housing Developers of Ecuador (APIVE), the net reserves of the sector between January and June 2021 grew by 44% compared to 2020, while visits to projects grew by 11% during the same period in 2021 versus the year before [3].

The experience of qualified professionals in an appraisal is important, and the particularity of the place where the property is located must be considered

Supported by Organization X.

H. Florez and M. Leon (Eds.): ICAI 2023, CCIS 1874, pp. 197–212, 2024.
https://doi.org/10.1007/978-3-031-46813-1_14

since this can influence the moment of valuing a property and therefore cause its price to increase or decrease. Among other aspects of appraisal are mobility, measurement tools, etc. "Quito is located in the central region of the Ecuadorian highlands with a very irregular geographical context. (Metropolitan District of Quito, 2014).

This work proposes to apply machine learning techniques in order to forecast real estate values in the city of Quito. Applying algorithms is a simple way of facilitating the access and availability of information for the purpose of obtaining results in a more dynamic way.

The supervised learning process consists of training the neural network from a set of input data and their respective outputs. The learning algorithm adjusts the parameters of the network in such a way that the output generated by the ANN fits the data. Output is given some input [1].

In this type of learning, the figure "supervisor" is not conceived, and its main objective is to implement an output. It is based on the redundancy in the inputs, and the learning is extracted from the patterns that it obtains from the data input.

Below is a review of the literature related to the problem analyzed, the methodology applied, models, algorithms, and data used, then the results are presented for each model and geographical area under study, and the conclusions, limitations, and future work are presented.

1.1 Literature Review

There were three outstanding investigations that coincided with criteria and keywords such as "real estate + neural network" and applied variables such as: **area**, **price**, and others. The related works are the following:

1. Predictive analysis using Big Data for the real estate market during the COVID-19 pandemic. (Análisis predictivo utilizando Big Data para el mercado inmobiliario durante la pandemia del COVID-19) [16].
   It provides insight into the understanding of the applied machine learning models as well as which ones were the most accurate for real estate prediction. It collects data with web scraping, and the methodological steps were taken in building the machine learning models. Considering variables such as type of building, years old, size in square meters, room number, number of floors, garages, and other characteristics of the house, as well as proximity to universities, schools, shopping centers, and train stations as the characteristics most significant that affect the price.
2. Boost home price predictions by integrating geospatial networks [18]. The objective of this research was to predict the value of a house. The price is one of the most critical factors for buying a house, and highlighting the importance of the location and geospatial location, which is based on the installation of the neighborhood, also includes aspects such as room number, number of bathrooms, construction area, services, the proximity of the property to schools, train stations, and supermarkets, it is an area in a residential location whose environment is in harmony with the lifestyle of future buyers.

3. Real Estate Investing in Dubai: A Predictive and Time Series Analysis [19]. The objective of this research was to predict the house's value, considering that the price is a critical factor in the process of buying a house, as well as the geospatial location, which is based on the services of the neighborhood. Another aspect was considered, like the number of rooms and bathrooms, the construction area, etc. The study considers the proximity of the property to schools, train stations, and supermarkets as a residential location whose environment guarantees harmony for the future buyer's lifestyle.

2 Methodology

It is necessary to understand the process that an expert carries out and applies at the time of carrying out the appraisal. One of the methods they use is the so-called "comparison", which determines the value of the land in the value intervention area that does not have services and infrastructure "compared to another with homogeneous characteristics in the same area, for the same purpose and potentiality" [8].

The CRISP-DM (CRoss Industrial Standard Process for Data Mining) methodology will be used because it describes the key processes to carry out a data mining project

1. **Business understanding**: at the beginning, the objective is to obtain an overview of the real estate sector.
2. **Data preparation**: a phase that took a long time, since the selected data sets and their fields had to be cleaned before they could be used.
3. **Modelling**: in this phase, the analysis was carried out using the selected modeling techniques.
4. **Evaluation**: once the models were built and based on the metrics implemented, the evaluation was carried out.

2.1 Dataset

The web pages visited for the elaboration of this investigation were PROPERATI, PLUSVALIA, and REMAX, in which the web scraping tool was applied so that the information for the present project was collected.

2.2 Data Extraction

It was done by accessing the URL of the aforementioned real estate pages, where published information on houses is found and later saved in csv format. Among all of them, PROPERATI stands out because its portal allows downloading csv files, plus the data downloaded with the web scraping tool from the other web pages constitutes the data set for this project. It is necessary to highlight that the data obtained corresponds to the year 2020.

2.3 Data Transformation

In the transformation of the downloaded data set, the data of the collected properties was initially analyzed with the Excel tool; this was done with the purpose of visualizing the content of the information so that the starting point could be used to continue with the transformation process of the data.

2.4 Data Visualization

Once the data set was downloaded for the analysis of this work, it consisted of 288,984 observations and 24 columns or variables. The content of each column was analyzed, and its contribution to this research was verified. For a better understanding, all of this is described in detail in Table 1.

Table 1. Description of the variables.

Variable	Description
ad_type	Kind of property
Id	Automatically generated code numeric
start_date	Publication date
created_on	Post creation date
end_date	Finish date
Lat	Latitude
Lon	Length
l1	Country
l2	Province
l3	Canton
l4	Industry location name
l5	Industry name
l6	Neighborhood name
Rooms	Number of rooms
Bedrooms	Number of bathrooms
Area	Total area in of the property
Construction	Construction area
Price	Price in $
Surface	Area in
Currency	Currency USD
Description	Description of the user enters
Title	Post title
Price_period	Price periods

2.5 Column and Data Cleaning

For cleaning, we proceeded with the verification of the characteristics of the data. Columns were selected and eliminated due to their content, since many of these columns did not provide significant value for the present work. Table 2 shows the name of the variable and a brief description of its contents.

Table 2. Dropped Variables Table

Variable	Description
id	Generated Code Numeric and does not provide information.
created_on	Post creation date.
end_date	Finish date.
Surface	Area in m2
Currency	Currency USD
Description	Description that the user enters.
Title	post title
Price_period	Price periods

For the selection and cleaning of records within the data, the R-studio command was used, and the indications containing "9999-12-31" and empty were eliminated since they do not contribute to the data. According to the research "Model of Classification of Credit Risk Using Random Forest in a Finance Company of Ecuador", it should be checked if there are missing values that can generate errors when training the algorithm" [4].

2.6 Application of Automatic Methods

Some machine learning algorithms were applied, such as linear regression, decision trees, random forests, and neural networks, in order to respond to the proposed objective and later in the measurement of performance in models. In the implementation of the automatic methods, the open-source software R Studio was used, as well as Orange and Weka previously described in the theoretical framework section.

2.7 Algorithms and Key Parameters

A first aspect to consider is that there are several machine learning techniques that can classify, group, or predict data. The models used were:

– Linear Regression: The linear regression technique allows predicting the value of a variable from the values of another.

- Random Forest: The Random Forest algorithm works by aggregating the predictions made by multiple decision trees of variable depth. This method allows to correctly classify each variable based on a given objective.
- Decision Tree: Is a set of statistical algorithms that, through predictive methods, classifies data according to certain characteristics or properties, obtaining accurate and reliable models as a result. Initially.
- Neural network: Is a machine learning algorithm that can solve problems through machine learning, which trains neural networks with different structures for better results to be extracted with greater precision [11].

Also, the important parameters to verify for each method are:

- R2: A statistical measure that explains differences in one variable can be explained by the difference in a second variable.
- RMSE: Measure of the distance of the data points from the regression line
- CV: statistical measure of the dispersion of data points in a data series around the mean.

3 Results

The result of the transformation of the data through the visualization, analysis, and cleaning of the records allowed us to obtain a set of data ready to be used in the present project.

3.1 Data Analysis

Figure 1 shows the correlation of the variables that the data source has, with the largest circles being those with the best correlation and reaching the smallest circles with little correlation between the variables, obtaining the variables with the highest correlation of the area, construction, and price

Table 3 shows that for the large data set, we proceeded to segment some areas of the city of Quito into sectors; these sections are numbered from 1 to 7, since the downloaded data included said sectors. After that, their respective analyses were carried out.

Table 3. Sectors of Quito according to the downloaded data

Sector	Number
North of Quito	1
South of Quito	2
Valle de Tumbaco	3
Colonial Center	4
Valle de los Chillos	5
North Center	6
Others	7

3.2 Linear Regression

In this case, the price and area variables were used, and with both of them, it is possible to see the relationship between the **area** and the **price**.

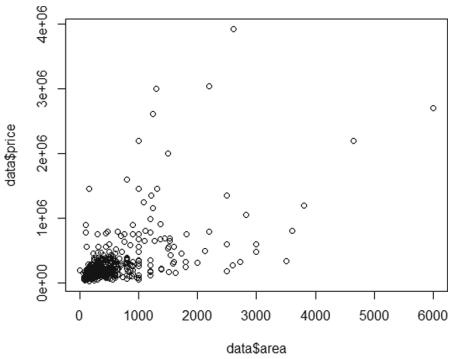

Fig. 1. Area and Price Scatter chart.

It shows, with small circles, the correlation dispersion of the data on the x axis. The area is in m2, and the y axis is the **price** in USD. We can observe the behavior of the data and its variations, which indicate that as the area of the property increases, so does the price.

Similarly, the correlation can be analyzed by means of Pearson's correlation, Table 4 shows the Pearson correlation by taking the **price** and **area** variables as parameters. When performing the correlation of the **price** and **area** variables, it is observed that there is a strong correlation between the two variables, and given that the p-value is less than 2.2e-16, it shows that there is a linear correlation as well.

Table 4. Pearson Coefficient Calculation

t	21.149
df	741
p Value	2.2e-16
Confidence Interval	95%
Cor	0.6135251

In this way, a linear regression model called "Modelo RL" can be proposed through the lm function of R, and later, the results can be displayed with summary (Modelo_RL), so that the results can be interpreted.

Table 5. Results of the estimation with the Linear Regression Model

Linear Regression Model	price vs Area
Estimated price	78644.45
Estimated Area	379.7
p Value	2e-16

Observing Table 5, once the Linear Regression model has been applied, the p-value is shown to be less than 0.05, considering that the model has a % confidence, as it has a significance value of less than 0.05. We can ensure that 0.05. We can ensure that the model is valid. To form the equation, we take the values of the coefficients. 78644.45 and 379.70 from Fig. 4.

$$x = 78644.45 + 379.70x \tag{1}$$

Equation 1 of the line allows us to predict the value of a piece of real estate with a certain area, according to the linear regression equation, just by moving x by the area in m2 of the real estate.

3.3 Random Forest

In the data set that is obtained when using the random forest, the possibilities of obtaining a more reasonable value for the properties by sectors are high since the decision is based on the results of multiple decision trees.

Table 6 shows the application of the Random Forest method. A partition of the data was created with 0.70 of the original data for building a model with Random Forest, where x represents the independent variables as well as the dependent variable. Ntree represents the number of trees that need to be generated. Later, we will make the prediction with the rest of the data. Resulting in the classification of real estate in each sector.

Table 6. Settings applied for Random Forest Model

Random Forest	Quito
Partition training data	0.7
X = Quito data	2:4
Y = Quito data	1
Number of trees	500

3.4 Decision Tree

We decided to use the area and construction variables of the real estate, but we had better influence using the **price**.

In the study of resolutions, it is possible to implement a decision tree to visually show the tree and thus, based on a set of variables, make the correct decision.

In Fig. 2, the algorithm shows the construction of the tree according to the data entered and its variables in the root node created by the tree. The **price** (real estate value)is taken as the main attribute, and at the first level we can observe 100% of the data, the first distinction made based on **price**. The properties with lower values than the one taken as a reference are 66% and the properties greater than said value correspond to 34%. The second distinction made corresponds to 34% and takes the **price** to be 25% lower than the referred **price** and 8% higher than this. For the third distinction, it takes construction, constructions smaller than 254 $m\hat{2}$ corresponds to 13%, and constructions greater than this value are 13%. The fourth distinction corresponds to 13% by taking real estate area, showing as a result that the area greater than 318 $m\hat{2}$ corresponds to 3% and less than this value is 9%.

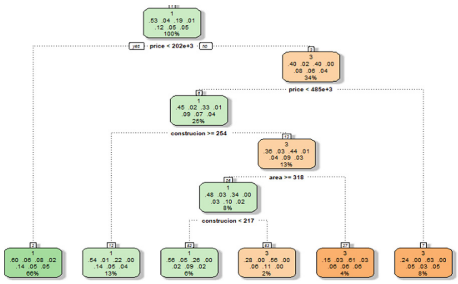

Fig. 2. Visualization tree using R

3.5 Neural Networks

Artificial networks are made up of nodes, or neurons," that receive, transmit, and send information; they are usually made up of layers. These networks are trained in a way that allows them to create, modify, or eliminate a connection between the nodes in order to provide a response.

The neural network method is applied. For the data presented in the present neural network, as input, there are all the independent variables that correlate with the **price** variable.

Figure 3 shows the input of the neural network and its output, which is the objective of the network. Its objective is the **price**, the presented neural network was carried out in 193 steps, which works with 9 neurons in the first layer and for the second layer 7 neurons.

4 Evaluation of Automatic Methods

This section shows the results of the performance measures after running the models on the seven data sets. The procedure for all models was the same in each case for both training and testing.

4.1 Results North Sector of Quito

Next, Table 7 shows the result of the performance measures generated by the models applied to the data in the northern sector of Quito.

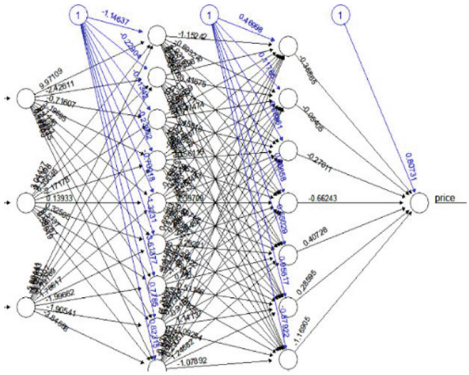

Fig. 3. Scheme neural network and its output

Table 7. Results obtained with the North Sector

	Linear Regression	Decision Tree	Random Forest	Neural network
R2	0.36	0.087	0.188	-0.57
RMSE	205148.504	245482.743	231397.99	322566.067
CV	1.0	1.25	1.18	1.65

Application result of the models concludes that the best determination coefficient is the Linear Regression with a value of 0.36, likewise, it can be seen that the smallest **RMSE** error is the Linear Regression. This allows verifying that the prediction value of the property is close to the real one in the sector, and finally it is seen how the coefficient of variation is also lower in the Linear Regression.

4.2 Results of South Sector of Quito

In Table 8 It is shown that the best determination coefficient is linear regression with a value of 0.24, followed by random forest. This makes it possible to verify

that the predicted value of the property is close to the actual value in the same sector. Also, it can be seen that the coefficient of variation is lower in random forests since the method coincides with the northern sector.

Table 8. Results obtained with the South Sector

	Linear Regression	Decision Tree	Random Forest	Neural network
R2	0.24	−0.40	0.125	−3.23
RMSE	44852.982	61047.422	48207.507	106028.419
CV	0.48	0.65	0.24	0.54

4.3 Results in the Valle de Tumbaco

Table 9 shows the performance of the Tumbaco Valley sector, in its results it indicates that the best coefficient of determination is Linear Regression with 0.62, followed by Random Forest with 0.54. For this sector, the neural network is the one with the lowest performance, it is shown that the value of the real estate with the southern sector of Quito is more economical than that of the Chillos Valley sector.

Table 9. Results in the Tumbaco Valley

	Linear Regression	Decision Tree	Random Forest	Neural network
R2	0.62	0.48	0.54	−0.78
RMSE	340741.160	396205.408	374709.079	490164.994
CV	0.69	0.80	0.21	0.99

4.4 Results Colonial Center Sector

As can be seen in Table 10, it is shown how the applied models do not contribute to the analysis; this is due to the fact that the data for this sector are small and consequently do not provide significant results, which is why they are excluded from this analysis.

Table 10. Results Colonial Center Sector

	Linear Regression	Decision Tree	Random Forest	Neural network
R2	−0.11	−0.19	−0.33	−12.97
RMSE	161446.168	49453.644	55661.92	163246.749
CV	0.98	0.30	0.34	0.33

4.5 Results of Valle de Los Chillos Sector

With the results of Table 11, it can be seen that the linear regression method has a higher coefficient of determination, followed by Random Forest; reviewing the previous results for the north sector 2.3, it can be deduced that the Linear Regression has a better impact on the sectors.

Table 11. Results of Valle de los Chillos Sector

	Linear Regression	Decision Tree	Random Forest	Neural network
R2	0.67	0.18	0.49	-1.2
RMSE	89926.373	143180.899	113115.389	240211.485
CV	0.49	0.79	0.62	1.3

4.6 Results Central North Sector

In Table 12 it is verified that the best determination coefficient is the linear regression method, and it can be seen that the low results are due to the scarce data sets of the north-central sector.

Table 12. Results Central North Sector

	Linear Regression	Decision Tree	Random Forest	Neural network
R2	0.07	−0.9	−0.4	−1.8
RMSE	147133.416	211231.11	183767.651	258027.961
CV	0,70	0.98	0.88	0.99

4.7 Other Sector Results

Table 13 As a result, for the other sector, the coefficient of determination with the greatest importance is Random Forest with 0.11 and its coefficient of variation is 0.86, which is lower compared to the other methods implemented.

Table 13. Other Sector Results.

	Linear Regression	Decision Tree	Random Forest	Neural network
R2	−0.13	−0.23	0.11	−1.17
RMSE	202503.134	211365.420	178830.719	280151.443
CV	0.98	0.99	0.86	1.36

4.8 Analysis of Results with ROC and AUC

From the analysis carried out and the results obtained with the applied methods, the so-called ROC analysis and AUC could be performed, which would allow verifying the classification performance of the applied methods.

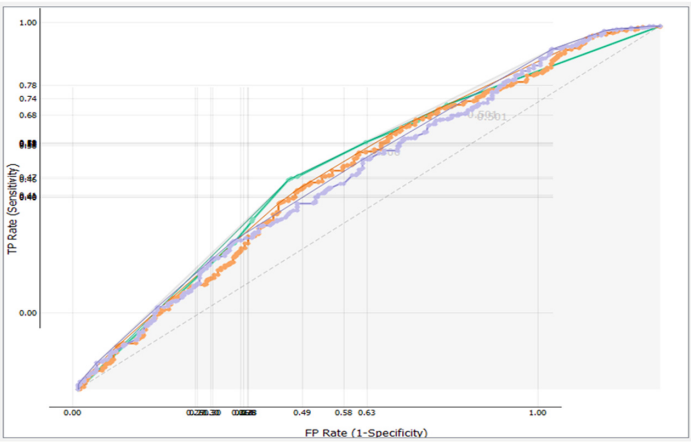

Fig. 4. Analysis of results with ROC and AUC

Figure 4 In the evaluation of the methods, it is observed that, for each method applied, the lower area of the **AUC** curve stands out with its respective value. In the ROC curve figure, the Y axis corresponds to the proportion of true positives over the total number of properties in the established sectors, and the X axis corresponds to the proportion of false positives over the total number of properties in the established sectors. Examining the ROC curve figure shows the "proportion of true positives" (Y axis) versus the "proportion of false positives" (X axis).

5 Discussion of Results

In this section, the techniques applied in the preparation of this work are detailed, as is their interpretation. In total, 1263 records were used, which, in the application of the chosen methods, had a partition of 70% for training and 30% for tests, since in this way it was possible to analyze the results obtained and their performance.

Table 14 was prepared as an explanation and summary of the techniques applied for each data set, showing the aforementioned percentages.

Table 14. Results obtained with the North Central sector.

Model	Regression, Decision Tree, Random Forest
Performance	Correlation of variables
Data	1263 individuals
Partition	Partition: 70%, Training 30%
Variables	l4 sectors de Quito, Price, Area m2, Construction m2

After obtaining the results of the applied methods, it can be observed that linear regression predominates over sectors 2, 3, 5, and 6, unlike the other sectors in which the random forest method was the best. In reference to the RMSE, it is verified that sectors 1, 3, and 4 obtained a high value in housing; the sectors with the highest value in terms of housing are those areas, with the model called linear regression predominating. With the results thrown by RMSE distributed for each sector, a value of the real estate by sector was obtained.

Table 15. Best adjustment to the price of homes according to RMSE

Sector	RMSE
Norte	$ 251148,83
Sur de Quito	$ 65034,082
Valle de Tumbaco	$ 400455,160
Valle los Chillos	$ 146608,53
Centro Norte	$ 200040,03
Otros	$ 218212,679

Table 15 shows the average value of different neighborhoods using the automatic learning methods. It is evident that the sector with high capital gains, according to the cost of housing, is Valle de Tumbaco, and the sector with the lowest capital gains is the south of Quito.

The ROC curve shows the performance of classification models Table 16.

Table 16. ROC and AUC analysis

Model	AUC	Accuracy
Decision Tree	0.59	0.45
Random Forest	0.64	0.40
Neural Network	0.65	0.39
Linear Regression	0.66	0.65

Through the ROC and AUC analysis produced by the models applied to the Quito data, the Linear Regression presents the highest percentage of precision when predicting the value of the property in the data set.

6 Conclusions and Future Work

In this study, a process of obtaining data through the use of web scraping was proposed. This tool provided data from the different web portals available that are involved in the real estate branch. These datasets were widely used at the time of training the implemented models. In order to obtain satisfactory results, we used different machine learning techniques like linear regression, random forest, neural networks, and decision trees with different libraries to analyze this data. As a result, linear regression was the most effective way to determine the value of the property in comparison with the performance measures of: $R\hat{2}$, RMSE and CV implanted in the data set. The cleaning of the information and data processing contemplated the reduction of repeated records and empty data and the selection of relevant characteristics of the data in order to obtain an optimal data set and apply the methods described in this work. It is recommended for future work to take into account more variables than those considered in this work, such as years old, geographical location, remodeling, internal finishes, and number of bathrooms. As well as the use of other machine learning models.

This work is a first stage of the real estate analysis, and the percentage of the precision indicators can be improved since currently the AUC and accuracy do not exceed 75%, so it is advisable to explore other models and algorithms such as principal component analysis or multivariate linear correlation.

References

1. Chiarazzoa, V., Caggiania, L., Marinellia, M.: Modelo basado en redes neuronales para la estimación de precios de bienes raíces. ScienceDirect 811 (2014)
2. Concepto Jurídico. Bienes Inmuebles. Obtenido de conceptojuridico.com (2021)
3. Ekos (2021). https://ekosnegocios.com/articulo/el-sector-inmobiliario-se-dinamiza
4. Freire, J., Guevara, B.: Modelo de Clasificación de Riesgo Crediticio Utilizando Random Forest en financiera del Ecuador (2021). https://repositorio.uisek.edu.ec/handle/123456789/4256
5. García, C.: indrscompany (2022). https://www.indracompany.com/
6. Higueras, J.C.: El sector inmobiliario crece luego del rezago ocasionado por la pandemia. El telegrafo (2021)
7. Koller, S.: ¿Qué es el web scraping? (2022). https://seranking.com/es/blog/web-scraping/
8. Norma técnica de Valoración de los Bienes Inmuebles en el MDMQ. Normas Técnica para la valoración de bienes inmuebles urbanoss y rurales del Distrito Metropolitano de Quito (2019)
9. Realia. ¿Qué es el mercado inmobiliario? (2022). https://www.realia.es/que-es-mercado-inmobiliario

10. Rodríguez, T.: Discutiendo entre el Machine Learning y el Deep Learning ¿A qué nos referimos con cada uno? (2017). https://www.xataka.com/robotica-e-ia/machine-learning-y-deep-learning-como-entender-las-claves-del-presente-y-futuro-de-la-inteligencia-artificial
11. Shia, Y., Zhangb, Y.: The neural network methods for solving Traveling Salesman. Elsevier, 2 (2022)
12. Spiegato (2022). https://spiegato.com/es/que-es-un-conjunto-de-datos
13. Vozmediano, J.R.: Concepto de Valoración Inmobiliaria y Normativa Aplicable (2018). https://negocioinmo.com/concepto-de-valoracion-inmobiliaria
14. Zakaria, F., Fatine, F.A.: Towards the hedonic modelling and determinants of real estate's price. Elserver, 1 (2021)
15. García, C.: indrscompany (2022). https://www.indracompany.com
16. Grybauskas, A., Pilinkienė, V., Stundžienė, A.: J. Big Data (2021)
17. Plan de Uso y Gestión de Suelo del Distrito Metropolitano de Quito (2021). https://quito.gob.ec
18. Das, S.S.S., Ali, M.E., Li, Y.-F., Kang, Y.-B., Sellis, T.: Boosting House Price Predictions using Geo-Spatial Network Embedding (2020)
19. Krishna, N.: Dubai Real Estate Investment: A Predictive and Time Series Analysis (2021)