ISSN 2007-9737

Supervised Learning Applied to Real Estate Price Classification in Bogotá, Colombia

Gabriel Elías Chanchí Golondrino[1,*], Manuel Alejandro Ospina Alarcón[1], Rafael Gaitán Ospina[2]

[1] Universidad de Cartagena,
Facultad de Ingeniería,
Colombia

[2] Escuela de Ciencias Básicas,
Tecnología e Ingeniería – ECTBI,
Colombia

{gchanchig, mospinaa}@unicartagena.edu.co, rafael.gaitan@unad.edu.co

Abstract. Currently, machine learning models have gained relevance in various fields of application, with real estate market price prediction being a key application for both property sellers and potential buyers. This article implements a machine learning workflow on a Bogotá real estate dataset, evaluating three classic models (kNN, Decision Trees, and Logistic Regression) and two ensemble methods (Random Forest and AdaBoost). The CRISP-DM data mining methodology was adapted into four phases: business and data understanding; data preparation; modeling, and evaluation and deployment. Using the Orange tool, the two ensemble models achieved the best performance, with AdaBoost obtaining the highest accuracy, precision, and recall scores reaching a value of 0.720. Property type and number of rooms were identified as the most relevant attributes. This study serves as a reference for the real estate sector, providing a decision-making tool based on current market evidence and machine learning techniques.

Keywords. Real estate market, visual orange, price estimation, ensemble methods, machine learning.

1 Introduction

The real estate market faces significant challenges in accurately estimating property prices due to the variability of factors influencing a property's value, such as location, physical characteristics, market conditions, and economic trends (Adetunji et al., 2022). In this context, machine learning models provide notable advantages, enabling buyers and sellers to obtain more accurate and well-informed estimates (Wang et al., 2021). Thus, one of the most widely used approaches in machine learning is supervised learning, a technique that utilizes labeled data to train models capable of predicting or classifying new data (Aguilar et al., 2018; Laura Ochoa et al., 2017; Suthaharan, 2016).

These models can analyze vast amounts of data, identifying complex patterns that traditional methods cannot capture, resulting in more precise and reliable price predictions (Rizun & Baj-Rogowska, 2021).

In this regard, manual real estate valuations often lack a solid data basis and are not updated as quickly, whereas machine learning models provide more precise estimations based on current market evidence (Guijarro Martínez, 2023).

In addition to the above, these models are capable of processing a wide range of variables and factors that can affect property prices, allowing for more complete and holistic models compared to traditional methods. (Solano Sanchez et al., 2021).

The predictive capabilities of these models are crucial for informed decision-making among real estate investors, buyers, and sellers (Kang et al., 2021). Thus, the implementation of these models enables a faster and more cost-effective valuation of real estate properties, proving highly useful for

stakeholders (Choy & Ho, 2023). For buyers, this means being able to make more informed purchasing decisions that align with their needs and budgets. For sellers, it involves setting competitive prices in line with market conditions, optimizing both sales time and return on investment. The importance of developing machine learning studies in this field is particularly significant, given the widespread availability of open datasets and tools that facilitate the implementation of these models (Zhan et al., 2023). This promotes innovation and continuous improvement in real estate valuation methodologies, contributing to a more transparent and efficient market.

Recent advances in the application of machine learning to the real estate market have led to significant improvements in price prediction models. (Adetunji et al., 2022) employed Random Forest algorithms to predict house prices, achieving notable accuracy improvements by selecting relevant property attributes from structured datasets. Similarly, (Wang et al., 2021) proposed a deep learning approach based on heterogeneous data analysis and joint self-attention mechanisms, enhancing predictive performance when combining various property and market features.

In another study, (Suthaharan, 2016) analyzed the predictive power of web search queries for forecasting price trends in the real estate market, demonstrating the effectiveness of integrating external unstructured data sources. (Zhan et al., 2023) presented a hybrid machine learning framework that combined multiple algorithms to forecast house prices more reliably, indicating that ensemble and hybrid models generally outperform individual classical models.

These studies underline the importance of selecting appropriate variables, adopting ensemble learning techniques, and integrating external data sources for improving the predictive accuracy of real estate valuation models. In contrast to previous works, the present study focuses on the use of classical supervised learning algorithms and ensemble methods (Random Forest and AdaBoost) applied specifically to structured data from the Bogotá real estate market, leveraging the Orange visual programming tool for workflow deployment and evaluation.

This study explores and characterizes the dataset corresponding to the Bogotá real estate market obtained from the Kaggle portal, aiming to identify the factors that directly influence property valuation. Building on these factors and utilizing the Orange visual programming tool for machine learning, various classic and ensemble supervised learning models (kNN, Logistic Regression, Decision Trees, Random Forest, and AdaBoost) were evaluated to assess their fitting and classification capability for real estate prices (expensive, cheap, and moderate).

The inclusion of ensemble models is justified by the fact that these methods represent an evolution of classical machine learning approaches, as they address the limitations of individual models by combining them to improve performance and robustness (Pintelas & Livieris, 2020; Sagi & Rokach, 2018). The Orange tool was selected considering its various advantages in the visual creation of machine learning workflows, allowing for the straightforward construction of complex data processing pipelines through an intuitive graphical interface designed for both novice and expert users (Demšar et al., 2004; Dobesova, 2024; Reena Thakur, 2023).

The developed work aims to support decision-making for both buyers and sellers regarding the appropriate price at which they can purchase or sell a property in Bogotá. The approach used in this article, based on ensemble methods, can be extrapolated to datasets from other cities, considering similar property-related variables. Additionally, incorporating new economic variables could further enhance the model's accuracy, providing more comprehensive insights into real estate valuation.

The remainder of the article is organized as follows: Section 2 presents the methodological phases considered for the development of this research. Section 3 presents the results obtained in this study, including the identification of the factors that have the greatest influence on real estate price classification. This section also includes the deployment of the machine learning workflow and the evaluation of the fitting capability of the deployed models. Finally, Section 4 presents the conclusions and future work derived from this research.

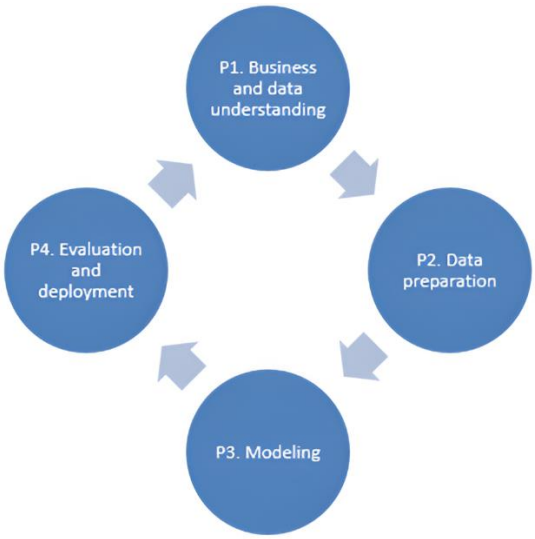

Fig. 1. Considered methodology

Table 1. Description of dataset attributes

Attribute	Description
Type	The type of property, which can be apartment, house, office, local, lot, warehouse, etc.
Description	A brief description of the property.
Rooms	The number of rooms in the property.
Bathrooms	The number of bathrooms in the property.
Area	The size of the area in square meters of the property.
Neighborhood	The name of the neighborhood where the property is located.
UPZ	The Zoning Planning Unit where the property is located.
Value	The value of the property in Colombian pesos.

```
0 s   df_inm.isna().sum()

Tipo            0
Habitaciones    0
Baños           0
Área            0
Barrio          0
Valor           0
dtype: int64
```

Fig. 2. Checking missing values of the dataset

2 Methodology

For the development of this study, an adaptation of the CRISP-DM data science project methodology (Calvillo et al., 2016; Herawati et al., 2024; Lundén et al., 2023; Martinez-Plumed et al., 2021) was applied, defining four phases: P1. Business and data understanding, P2. Data preparation, P3. Modeling, and P4. Evaluation and deployment (see Fig. 1). Although the CRISP-DM methodology originally includes six phases, in this study an adaptation was applied considering only four. The phases related to Deployment and Maintenance were excluded, as the scope of the project focused on exploratory data analysis and model evaluation within an academic research context, without the need for operational deployment in a production environment.

In phase 1 of the methodology, the real estate dataset for the city of Bogotá was obtained from the Kaggle website using the developer key provided by the portal, ensuring that the dataset can be automatically updated if any changes are made by the dataset creator. Additionally, the dataset was characterized and described. This dataset, by default, contains a total of 9,520 instances or records, each with 8 attributes or columns, which are detailed in Table 1.

Following the dataset loading, the data preparation phase involved data cleaning. This included removing the “Description” and “UPZ” attributes, as they contain information that is already included in the “Barrio” (Neighborhood) and “Tipo” (Type) attributes, making them redundant. The elimination of these columns is critical to ensure the dataset's efficiency by reducing redundancy, which can lead to unnecessary computational overhead and storage inefficiency. Additionally, removing textual attributes like “Description” streamlines data analysis, as these columns often introduce noise and complexity, especially when they do not contribute new, unique insights. This process was effectively executed using Python's Pandas library, which offers powerful, efficient methods for data manipulation and transformation, enabling a cleaner and more concise dataset for subsequent analysis.

After converting the “Valor” attribute to numeric format, an inspection was performed to check if

Table 2. Description of dataset attributes

Id	Attribute	Data type
1	Type	Categorical
2	Rooms	Numerical
3	Bathrooms	Numerical
4	Area	Numerical
5	Price_Category	Categorical

	Type	Rooms	Bathrooms	Area	Price_category
0	Apartamento	0.018349	0.222222	0.000409	Moderate
1	Casa	0.027523	0.333333	0.001011	Moderate
2	Apartamento	0.018349	0.333333	0.000854	Expensive
3	Apartamento	0.018349	0.222222	0.000914	Expensive
4	Apartamento	0.009174	0.333333	0.000758	Expensive

Fig. 3. Dataset obtained after the coding and scaling process

any of the attributes contained NA (Not Available) values. The results showed that none of the attributes had missing values, as illustrated in Figure 2. Therefore, it was not necessary to apply any imputation methods, as the dataset was complete.

After determining that imputation processes were not necessary for the dataset, the next step was the generation of the "Value_Dollars" and "Price_per_m2" columns, which were essential for deriving the classification predictor attribute, given that the default dataset was originally configured for regression tasks. In this process, the "Value_Dollars" column represents the property price in dollars and was obtained by dividing the "Value" column by the current exchange rate of the dollar. Similarly, the "Price_per_m2" column was calculated by dividing the "Value_Dollars" column by the "Area" column, ensuring a standardized measure of price per square meter.

For the creation of the predictor attribute "Price_category", the "Price_per_m2" attribute was used as a reference, categorizing its values into three levels: "Cheap", "Moderate", and "Expensive" based on the 33rd and 66th percentiles, leveraging the capabilities provided by the pandas library in Python. It is important to note that since the "Value", "Value_Dollars", and "Price_per_m2" columns were used to derive the predictor attribute "Price_category", maintaining a directly proportional relationship, they were not included in the final dataset for model fitting. Additionally, the "Rooms", "Bathrooms", and "Area" variables were normalized to ensure they shared the same numerical scale and did not affect the predictive capacity of the models. Thus, Table 2 presents the attributes considered for dataset construction along with their data types. It is also worth mentioning that the "Neighborhood" attribute was excluded, as it contains a large number of categories, and the tool used for deploying the machine learning workflow only supports attributes with a maximum of 16 categories.

In accordance with the above, the resulting dataset with the attributes listed in Table 2 is presented in Fig. 3.

The final dataset used for model training included the following attributes: Type, Rooms, Bathrooms, and Area. The Price_Category attribute was used as the target variable for classification. These selected variables, after data preparation and feature selection, were utilized for fitting all supervised learning models evaluated in this study.

After categorizing the predictor attribute and normalizing the numerical attributes, the next step was to identify the variables with the greatest impact on classification using violin plots, box plots, and information gain methods. This analysis leveraged the "Violin Plot", "Box Plot", and "Rank" components provided by the Orange visual programming tool.

In Phase 3 of the methodology, once the most relevant attributes were identified, the first step was the complete deployment of the machine learning workflow in the Orange tool. This process included dataset loading, variable selection, sampling through cross-validation, model training using machine learning algorithms, and evaluation based on the metrics derived from the confusion matrix.

Finally, in Phase 4 of the methodology, the metrics obtained for each model were compared using the confusion matrix (accuracy, precision, recall) to determine the model with the best fit for the given dataset. It is important to note that for the evaluation of these methods, cross-validation was configured with 10 folds, meaning 10 iterations in which the training and test sets were varied (1 fold for testing and 9 folds for training). The cross-validation approach was chosen considering that,

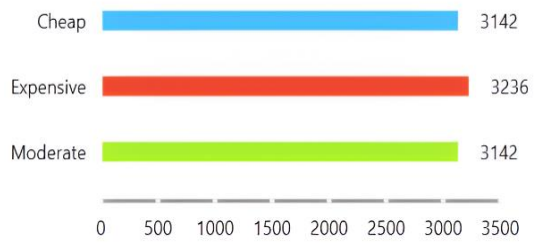

Fig. 4. Distribution of instances across the categories of the "Price_category" attribute

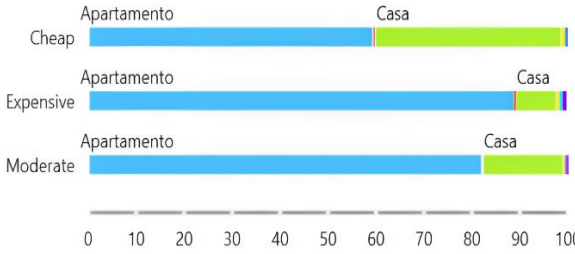

Fig. 5. Distribution of property types across the categories of the "Price_category" attribute

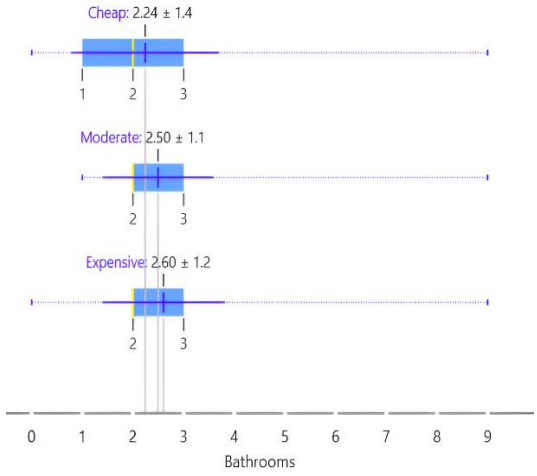

Fig. 6. Box plot diagram of the "Bathrooms" attribute

although it is less computationally efficient, it provides a more realistic and robust assessment of model performance on new validation datasets (Korjus et al., 2016; Mohr & van Rijn, 2023; Qiu, 2024).

3 Results

Regarding the results obtained in this research and based on the constructed dataset, the first step was to analyze the distribution of the variables, as well as the relationship between the dataset's numerical attributes ("Rooms", "Bathrooms", and "Area") and the categorical variables ("Type" and "Price_category"). Accordingly, Fig. 4 presents the distribution of instances across each category of the "Price_category" attribute.

From Fig. 4, it can be observed that the dataset exhibits a comparable percentage of instances across each category of the "Price_category" attribute ("Cheap", "Expensive", and "Moderate"), with the number of instances ranging between 3,142 and 3,236. This indicates that the dataset is generally balanced, ensuring that the supervised learning models are not biased toward selecting one category over the others due to an imbalance in the number of instances. In this regard, traditional methods tend to classify the class with the most instances more accurately, while classes with fewer instances exhibit lower performance (Beyan & Fisher, 2015).

Fig. 5 presents a diagram illustrating the distribution of property types across each category of the "Price_category" attribute. It can be observed that, although apartments are the predominant property type in all three categories, their number is significantly higher compared to houses and other property types.

The difference between the number of apartments and houses is most pronounced in the "Expensive" and "Moderate" categories. It is worth mentioning that both Fig. 4 and Fig. 5 were generated using the Box Plot component for categorical variables.

Now, when relating the "Bathrooms" attribute to the predictor attribute "Price_category" using a box plot, the resulting diagram is presented in Fig. 6.

In the diagram presented in Fig. 6, it can be observed that as the property becomes more expensive, the median number of bathrooms increases. This is evident as the median for properties in the "Cheap" category is 2.4, whereas for those in the "Expensive" category, it rises to 2.6. The difference in medians is also explained by the fact that the number of bathrooms in "Cheap" properties ranges between 1 and 3,

while in the other categories, it varies between 2 and 3.

Now, when relating the "Rooms" attribute to the different categories of the predictor attribute "Price_category", the resulting violin plot is presented in Fig. 7.

The violin plot presented in Fig. 7 shows that more expensive properties in the "Expensive" category tend to have fewer rooms, whereas properties in the "Moderate" category not only have the highest median but also the largest number of extreme values in room count. Meanwhile, the "Cheap" category exhibits a more concentrated distribution with some variability. This suggests that "Expensive" properties may be more influenced by factors such as location and luxury, while "Moderate" properties encompass a wider range of sizes, including those with a higher number of rooms.

Now, when analyzing the relationship between the "Area" attribute and each category of the "Price_category" attribute using the diagram generated by the “Mosaic Display” component, the resulting graph is presented in Fig. 8.

According to the results presented in Fig. 8, it can be observed that properties with an area smaller than 57.5 square meters have a higher proportion of listings in the "Cheap" category and a lower proportion in the "Moderate" category. On the other hand, for properties with an area between 80.5 and 136.5 square meters, the proportion of "Cheap" properties decreases, while the presence of "Moderate" and "Expensive" categories increases. Additionally, for properties with an area of 136.5 square meters or more, the proportion of "Expensive" properties decreases compared to the previous range, while the "Cheap" and "Moderate" categories show a relative increase. This suggests that, although property prices tend to rise with size, larger properties do not exclusively belong to the most expensive category, indicating the possible influence of other factors, such as location and construction quality, in price classification.

Now, to analyze the relationship between the categories of the "Type" variable and the categories of the "Price_category" variable, the violin plot presented in Fig. 9 was generated using the "Violin Plot" component in Orange.

From Fig. 9, it can be observed that apartments exhibit the greatest dispersion in values, with some

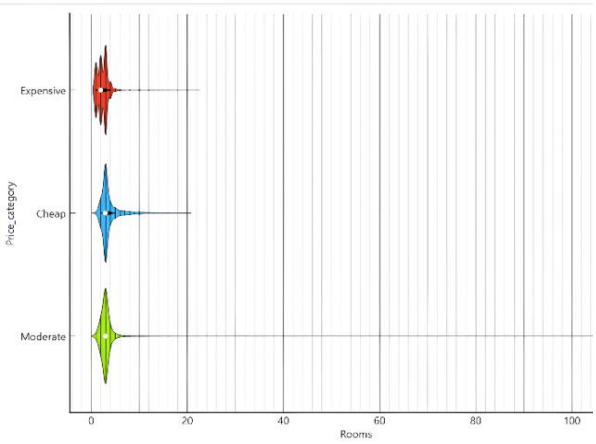

Fig. 7. Violin plot of the number of rooms in properties

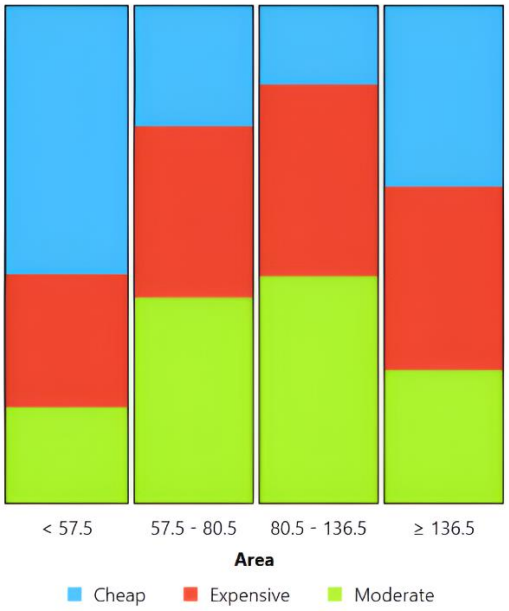

Fig. 8. Distribution of Area attribute values across the categories of the "Price_category" attribute

extreme cases of significantly high prices compared to other property types. In contrast, warehouses, lots, and farms show lower variability and generally lower prices.

Additionally, in the case of buildings and office/consulting spaces, prices per square meter tend to be higher and display a more concentrated distribution, indicating greater stability in these segments. Meanwhile, commercial premises

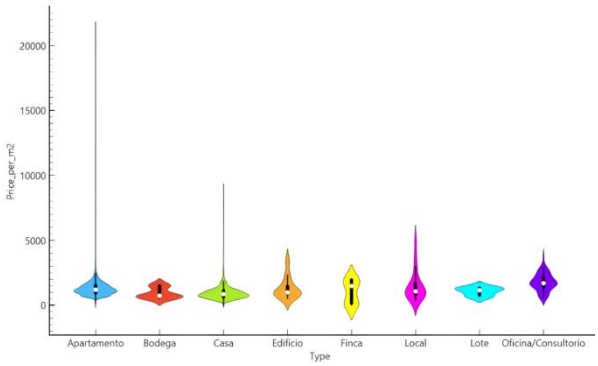

Fig. 9. Violin plot of the Type attribute categories in relation to the Price_per_m2 attribute

		#	Gain ratio	χ^2	ReliefF
1	C Type	8	0.083	1051.323	0.002
2	N Rooms		0.067	906.573	0.002
3	N Bathrooms		0.038	219.146	-0.000
4	N Area		0.038	178.963	0.000

Fig. 10. Application of information gain methods to the dataset attributes

exhibit a more uniform distribution with moderate price variability.

Overall, the graph suggests that property type significantly influences price dispersion, with apartments displaying the highest variability, while lots and warehouses show more homogeneous and lower values in terms of price per square meter.

Now, when applying information gain methods: (Gain Ratio, Chi-Square, and ReliefF) to the dataset attributes, leveraging the advantages provided by the Rank module in the Orange tool, the results obtained are presented in Fig. 10. Information gain methods are commonly used techniques for feature selection in classification models, helping to improve model accuracy and reduce execution times (Kurniabudi et al., 2020; Putra & Kadnyanana, 2021).

According to Fig. 10, it can be observed that among the dataset attributes, “Type" and “Rooms” have the greatest influence on property valuation prediction. This is evident as these two attributes consistently appear as the most relevant across all three information gain methods applied. This finding suggests that for property valuation classification, the type of property and the number of rooms are key attributes.

Now, to evaluate the fitting capability of the four considered models (kNN, Logistic Regression, Decision Trees, Random Forest, and AdaBoost) the machine learning workflow was deployed using the Orange visual programming tool, as presented in Fig. 11.

In Fig. 11, the machine learning workflow includes various processes, such as dataset loading using the "File" component, data visualization through the "Data Table" component, and attribute selection, including the predictor variable, using the "Select Columns" component. Additionally, information gain methods and violin plots were applied using the "Violin Plot" and "Rank" components. The dataset was then sampled using cross-validation through the "Data Sampler" component, followed by model training using the respective components for each algorithm. The decision tree model output was visualized using the "Tree Viewer" component, while model validation was performed through the "Test and Score" component. Finally, the confusion matrix was generated using the "Confusion Matrix" component to assess classification performance.

In the same context, Fig. 12 presents the results obtained for the different models, which were generated using the "Test and Score" component.

Based on the results presented in Fig. 12, it is observed that in terms of the Accuracy (CA) metric, the AdaBoost and Random Forest models achieve the best performance, with values of 0.720 and 0.712, respectively. In contrast, Logistic Regression exhibits the worst performance, with a value of 0.514. Similarly, for the Precision and Recall metrics, the best-performing models remain AdaBoost and Random Forest, both with values of 0.720 and 0.712, respectively.

This indicates that these models not only achieve a higher overall accuracy rate but also maintain high precision and sensitivity in classification. In comparison, the Decision Tree model shows slightly lower values in these metrics (0.691), while kNN and Logistic Regression display the weakest performance, with Precision values of 0.656 and 0.498, and Recall values of 0.657 and

ISSN 2007-9737

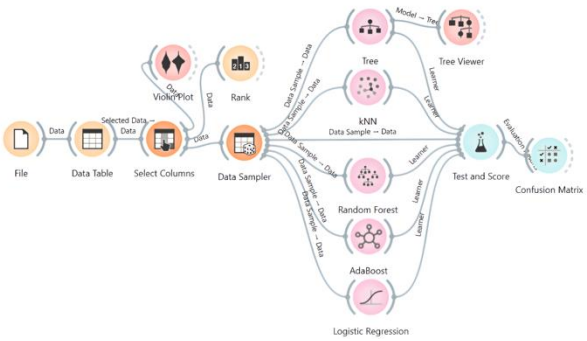

Fig. 11. Machine learning workflow deployed in the Orange tool

Model	AUC	CA	F1	Prec	Recall
AdaBoost	0.826	0.720	0.720	0.720	0.720
Random Forest	0.860	0.712	0.712	0.712	0.712
Tree	0.842	0.693	0.691	0.691	0.693
kNN	0.804	0.657	0.656	0.656	0.657
Logistic Regression	0.726	0.514	0.471	0.498	0.514

Fig. 12. Results obtained from the evaluation of the different supervised learning models

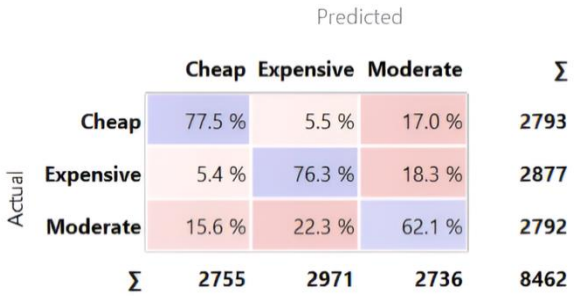

Fig. 13. Confusion matrix obtained for the AdaBoost model

0.514, respectively. These results confirm that, in terms of overall performance, AdaBoost and Random Forest are the most effective and well-balanced models for real estate valuation classification.

Finally, Fig. 13 presents the confusion matrix obtained from the evaluation of the AdaBoost model.

According to the results presented in Fig. 13, the AdaBoost model demonstrates better performance in classifying properties within the "Cheap" and "Expensive" categories, achieving accuracy rates of approximately 77%. In contrast, for the "Moderate" category, the accuracy rate is 62%. Additionally, it can be observed that misclassification errors in the "Cheap" and "Expensive" categories are primarily due to instances being misclassified as "Moderate". Similarly, for misclassifications in the "Moderate" category, the model tends to confuse these instances with the "Expensive" category. Despite these misclassifications, it is important to highlight that the model achieves a classification accuracy above 62% across all three categories.

Compared to recent studies, such as (Adetunji et al., 2022), who achieved an accuracy of 0.75 using Random Forest algorithms for house price prediction, the ensemble models evaluated in this study (AdaBoost and Random Forest) achieved accuracy scores of 0.720 and 0.712, respectively. Similarly, (Zhan et al., 2023) proposed a hybrid machine learning framework that combined multiple algorithms to improve house price forecasting, reporting accuracy rates exceeding 0.76. Although our models exhibit slightly lower accuracy, it is important to note that this study focused solely on structured data from the Bogotá real estate market, without integrating heterogeneous data sources or hybrid modeling techniques. Therefore, the results demonstrate the strength and practical applicability of the models within a more constrained and realistic context.

4 Conclusions

Due to the challenges faced by the real estate market in effectively estimating property prices, given the variety of factors that influence a property's value, such as location, physical characteristics, and market trends, machine learning models have become a key tool for buyers and sellers to estimate property purchase and sale prices. This work proposes a study based on machine learning for the characterization and prediction of property prices in the city of Bogotá. The relevant variables affecting price were identified, and classical models were compared with ensemble methods. This study aims to serve as a reference for promoting the use and application of machine learning in the real estate market.

One of the main contributions of this work was identifying the key attributes affecting property prices in Bogotá. Thus, by applying information gain methods (Gain Ratio, Chi-Square, and ReliefF), it was determined that the most significant attributes influencing property valuation in Bogotá were property type and number of rooms. These methods were implemented leveraging the advantages provided by the Rank component of the Orange visual programming tool.

Among the classic machine learning models, the Decision Tree model achieved the best performance, with an Accuracy score of 0.693 and Precision and Recall values of 0.691. In contrast, Logistic Regression had the lowest performance, with an Accuracy score of 0.514 and respective Precision and Recall values of 0.498 and 0.514. Additionally, in the final results, the Decision Tree model ranked third, surpassed by Random Forest and AdaBoost, which demonstrated superior performance.

Regarding the ensemble methods, it is important to highlight that the AdaBoost and Random Forest models achieved respective Accuracy scores of 0.720 and 0.712. Similarly, for the Precision and Recall metrics, both models obtained values identical to their Accuracy scores, indicating a strong ability to differentiate between the dataset categories. Additionally, based on the confusion matrix, it was determined that the AdaBoost model achieved accuracy rates close to 77% in the "Cheap" and "Expensive" categories, while in the "Moderate" category, it exhibited an accuracy rate exceeding 62%. It is also noteworthy that the two evaluated ensemble models consistently outperformed the classic machine learning models, demonstrating superior performance across all evaluation metrics.

This study demonstrated the effectiveness of the Orange open-source visual programming tool for conducting exploratory data analysis and deploying machine learning workflows. In this regard, the workflow enabled the rapid adjustment and evaluation of three classic machine learning models and two ensemble methods. Additionally, Orange facilitates the generation of various statistical visualizations, including bar charts, box plots, and violin plots, among others. Thus, the use of Orange aims to be promoted and extended for machine learning experimentation in universities and research centers.

This study contributes to the field by demonstrating the applicability of classical and ensemble machine learning models to structured real estate datasets, using a visual programming approach. By focusing on the Bogotá real estate market and leveraging the Orange tool, it provides a replicable and practical framework for similar urban contexts. It is important to acknowledge that the models developed exhibited an average classification error of approximately 28–30%, which is consistent with the complexity and variability typically found in real-world real estate datasets. Future studies incorporating additional socioeconomic variables and external data sources are expected to further reduce this error margin.

It is important to note that the present study focused exclusively on a single dataset corresponding to the Bogotá real estate market. As future work derived from this research, the following objectives are proposed: a) Enriching the dataset by incorporating additional socioeconomic attributes or variables that may influence property prices, b) Include experiments using additional datasets from other cities or sources to validate and generalize the findings obtained and c) Evaluating the predictive capacity of other models based on neural networks, leveraging Orange's capabilities to support this type of model.

Acknowledgments

The authors of this article express their gratitude to the University of Cartagena and the National Open and Distance University for their support in the development of this research.

References

1. **Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., Oluwadara, G. (2022).** House Price Prediction using Random Forest Machine Learning Technique. Procedia Computer Science, Vol. 199, pp. 806–813. Doi: 10.1016/j.procs.2022.01.100.

2. **Aguilar, R. M., Torres, J. M., Martín, C. A. (2018).** Aprendizaje Automático en la Identificación de Sistemas. Un Caso de Estudio en la Predicción de la Generación Eléctrica de un Parque Eólico. Revista Iberoamericana de Automática e Informática Industrial, Vol. 16, No. 1, pp. 114. Doi: 10.4995/riai.2018.9421.
3. **Beyan, C., Fisher, R. (2015).** Classifying imbalanced data sets using similarity based hierarchical decomposition. Pattern Recognition, Vol. 48, No. 5, pp. 1653–1672. Doi: 10.1016/j.patcog.2014.10.032.
4. **Calvillo, E. A., Mendoza, R., Munoz, J., Martinez, J. C., Vargas, M., Rodriguez, L. C. (2016).** Automatic Algorithm to Classify and Locate Research Papers Using Natural Language. IEEE Latin America Transactions, Vol. 14, No. 3, pp. 1367–1371. Doi: 10.1109/TLA.2016.7459622.
5. **Choy, L. H. T., Ho, W. K. O. (2023).** The Use of Machine Learning in Real Estate Research. Land, Vol. 12, No. 4, pp. 740. Doi: 10.3390/land12040740.
6. **Demšar, J., Zupan, B., Leban, G., Curk, T. (2004).** Orange: From Experimental Machine Learning to Interactive. Data Mining, pp. 537–539. Doi: 10.1007/978-3-540-30116-5_58.
7. **Dobesova, Z. (2024).** Evaluation of Orange Data Mining Software and Examples for Lecturing Machine Learning Tasks in Geoinformatics. Computer Applications in Engineering Education, Vol. 32, No. 4. Doi: 10.1002/cae.22735.
8. **Guijarro Martínez, F. (2023).** Valoración automática de inmuebles residenciales mediante modelos de Machine Learning. Revista de Estudios Empresariales, pp. 27–40. Doi: 10.17561/ree.n2. 2023.7823.
9. **Herawati, N. A., Gary, A. A. P., Hikmawati, E., Surendro, K. (2024).** A Hybrid Predictive Model as an Emission Reduction Strategy Based on Power Plants' Fuel Consumption Activity. IEEE Access, Vol. 12, pp. 47119–47133. Doi: 10.1109/ACCESS. 2024.3380809.
10. **Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., Ratti, C. (2021).** Understanding House Price Appreciation Using Multi-Source Big Geo-Data and Machine Learning. Land Use Policy, Vol. 111, pp. 104919. Doi: 10.1016/j.landusepol.2020. 104919.
11. **Korjus, K., Hebart, M. N., Vicente, R. (2016).** An Efficient Data Partitioning to Improve Classification Performance While Keeping Parameters Interpretable. PLOS ONE, Vol. 11, No. 8, e0161788. Doi: 10.1371/journal.pone.0161788.
12. **Kurniabudi Stiawan, D., Darmawijoyo Bin Idris, M. Y., Bamhdi, A. M., Budiarto, R. (2020).** CICIDS-2017 Dataset Feature Analysis With Information Gain for Anomaly Detection. IEEE Access, Vol. 8, pp. 132911–132921. Doi: 10.1109/ACCESS.2020. 3009843.
13. **Laura Ochoa, L., Rosas Paredes, K., Baluarte Araya, C. (2017).** Evaluación de Técnicas de Minería de Datos para la Predicción del Rendimiento Académico. Proceedings of the 15th LACCEI International Multi-Conference for Engineering, Education, and Technology: Global Partnership for Development and Engineering Education. Doi: 10.18687/LACCEI2017.1.1.368.
14. **Lundén, N., Bekar, E. T., Skoogh, A., Bokrantz, J. (2023).** Domain Knowledge in CRISP-DM: An Application Case in Manufacturing. IFAC-PapersOnLine, Vol. 56, No. 2, pp. 7603–7608. Doi: 10.1016/j.ifacol.2023.10.1156.
15. **Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., Flach, P. (2021).** CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering, Vol. 33, No. 8, pp. 3048–3061. Doi: 10.1109/TKDE.2019.2962680.
16. **Mohr, F., van Rijn, J. N. (2023).** Fast and Informative Model Selection Using Learning Curve Cross-Validation. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 45, No. 8, pp. 9669–9680. Doi: 10.1109/TPAMI.2023. 3251957.
17. **Pintelas, P., Livieris, I. E. (2020).** Special Issue on Ensemble Learning and Applications.

Algorithms, Vol. 13, No. 6, pp. 140. Doi: 10.3390/a13060140.

18. **Putra, D., Kadnyanana, I. G. A. G. A. (2021).** Implementation of Feature Selection using Information Gain Algorithm and Discretization with NSL-KDD Intrusion Detection System. JELIKU (Jurnal Elektronik Ilmu Komputer Udayana), Vol. 9, No. 3, pp. 359. Doi: 10.24843/JLK.2021.v09. i03.p06.
19. **Qiu, J. (2024).** An Analysis of Model Evaluation with Cross-Validation: Techniques, Applications, and Recent Advances. Advances in Economics, Management and Political Sciences, Vol. 99, No. 1, pp. 69–72. Doi: 10.54254/2754-1169/99/2024 OX0213.
20. **Reena Thakur, E. al. (2023).** A Comprehensive Analysis to Image Classification: Understanding Techniques and Explore Data Preprocessing a Non-linear Approach. Advances in Nonlinear Variational Inequalities, Vol. 26, No. 2, pp. 110–122. Doi: 10.52783/anvi.v26.i2.287.
21. **Rizun, N., Baj-Rogowska, A. (2021).** Can Web Search Queries Predict Prices Change on the Real Estate Market? IEEE Access, Vol. 9, pp. 70095–70117. Doi: 10.1109/ACCESS.2021.3077860.
22. **Sagi, O., Rokach, L. (2018).** Ensemble learning: A survey. WIREs Data Mining and Knowledge Discovery, Vol. 8, No. 4. Doi: 10.1002/widm.1249.
23. **Solano Sanchez, M. Á., Núñez Tabales, J. M., Caridad y Ocerin, J. M. (2021).** Un modelo hedónico para los alquileres turísticos en la ciudad de Sevilla. Revista de Métodos Cuantitativos Para La Economía y La Empresa, Vol. 31, pp. 144–160. Doi: 10.46661/revmetodoscuanteconempresa. 4043.
24. **Suthaharan, S. (2016).** Supervised Learning Models. pp. 145–181. Doi: 10.1007/978-1-4899-7641-3_7.
25. **Wang, P.-Y., Chen, C.-T., Su, J.-W., Wang, T.-Y., Huang, S.-H. (2021).** Deep Learning Model for House Price Prediction Using Heterogeneous Data Analysis Along With Joint Self-Attention Mechanism. IEEE Access, Vol. 9, pp. 55244–55259. Doi: 10.1109/ACCESS.2021.3071306.
26. **Zhan, C., Liu, Y., Wu, Z., Zhao, M., Chow, T. W. S. (2023).** A hybrid machine learning framework for forecasting house price. Expert Systems with Applications, Vol. 233, pp. 120981. Doi: 10.1016/j.eswa.2023.120981.

Article received on 10/06/2025; accepted on 06/10/2025.
**Corresponding author is Gabriel Elías Chanchí Golondrino.*