CIBES 2025 / 5th Current Issues in Business and Economic Studies Conference

Development of a Predictive Tool for Real Estate Analysis Using Machine Learning Techniques

Ricardo Reier[a*] Gregorio Acedo[b] Luis Miguel Garay[c] Sergio Náñez[d]

[a] PhD, Catholic University of Ávila, Spain, ricardo.reier@ucavila.es, Orcid: 0000-0002-4790-0351
[*] Corresponding Author
[b] Professor, Catholic University of Ávila, Spain
[c] PhD, International University of La Rioja (UNIR), Spain, Orcid: 0000-0002-2119-4532
[d] PhD, Catholic University of Ávila, Spain, Orcid: 0000-0001-5353-2017

Abstract

Housing markets in many countries are currently facing severe affordability challenges, particularly in large urban areas where prices have risen faster than wages. In Spain, and especially in cities such as Madrid, housing prices have reached levels comparable to the 2007 bubble, intensifying concerns about access and urban inequality. This study develops a predictive tool for real estate valuation based on Big Data and Machine Learning techniques. Using automated data collection, spatial analysis, and Gradient Boosting algorithms, the proposed system estimates property market values in real time by integrating structural and georeferenced variables. Madrid is selected as a case study due to its size, economic relevance, and highly dynamic housing market. The results demonstrate that machine learning models effectively capture intra-urban price heterogeneity and outperform traditional valuation approaches in predictive accuracy. The study also presents an interactive application that translates academic research into a practical decision-support tool for buyers, sellers, investors, and policymakers. By combining methodological rigor with applied relevance, this research contributes to the literature on PropTech and real estate analytics, while highlighting both the potential and the limitations of AI-based tools in addressing structural housing market challenges.

Keywords: machine learning, sustainability, innovation, real estate market

Cited: Reier, R., Acedo, G., Garay, L. M., & Náñez, S. (2026). Development of a predictive tool for real estate analysis using machine learning techniques. *Sustainability, Organization, Business and Economic Research (SOBER), 3*, 119-128. https://doi.org/10.66414/sober.291165

Selection and peer-review under responsibility of the 5th Current Issues in Business and Economic Studies Conference.

1. INTRODUCTION

The housing market is currently facing severe affordability challenges in many countries, characterized by rising prices and reduced access to housing. In Spain, and particularly in major cities such as Madrid, property prices have surpassed levels observed during the 2007 housing bubble, while access to housing has become increasingly difficult (Byrne & Norris, 2022). In 2024 alone, housing costs rose by over 9%, significantly outpacing wage growth and excluding a growing share of the population from the market. By the end of that year, housing affordability had become the primary socio-economic concern for more than one-fifth of Spaniards (Capellán et al., 2021).

This situation is not unique to Spain but reflects a broader, international pattern. Research across advanced economies highlights a persistent deterioration in housing affordability, driven by structural factors such as financialization, demographic pressures, migration flows, and policy decisions. Studies document diverse manifestations of this phenomenon, including ethnic discrimination in rental markets, price shocks caused by refugee inflows, health crises affecting property values, and the stabilizing—though limited—role of social housing systems (Rampini & Re Cecconi, 2022). Furthermore, evidence from countries such as the Netherlands and the United Kingdom suggests that housing price dynamics increasingly depend on institutional and policy frameworks rather than traditional demand-side or monetary explanations.

Taken together, these findings indicate that housing affordability has become a systemic and multifaceted challenge with global dimensions. As a result, there is a growing need for analytical tools capable of capturing complex price dynamics and supporting informed decision-making in the real estate sector. Access to real-time, granular, and reliable housing market data is now essential for buyers, sellers, investors, real estate professionals, and public administrations alike.

This study contributes to the existing literature on real estate analytics by combining automated data collection, fine-grained geospatial analysis, and advanced Gradient Boosting machine learning models within a single, operational decision-support application. Unlike many previous studies that focus either on methodological accuracy or on descriptive market analysis, this research integrates predictive modeling with an interactive tool designed for real-time use by market participants. The originality of the approach lies not only in the empirical application to Madrid, but also in the translation of machine learning outputs into a transparent, user-oriented system capable of supporting informed decision-making in complex urban housing markets.

Madrid is selected as the reference case due to its size, economic relevance, and highly dynamic real estate market, making it an ideal environment to test and validate the proposed system. The overarching objective of the study is to create a decision-support tool that enhances transparency and efficiency in the housing market (Reisenbichler, 2021). The research

hypothesizes that advanced machine learning methods can successfully deliver reliable real-time property valuations, offering meaningful value to market participants and policymakers. The paper is structured to present the relevant literature, methodology, empirical results, and conclusions, emphasizing the study's contribution to both academic research and practical applications in the real estate sector.

2. LITERATURE REVIEW

The housing market plays a central role in modern economies due to its strong connection with GDP, employment, household wealth, and financial stability (Rey-Blanco et al., 2024). In developed countries, residential investment and real estate-related activities account for a substantial share of economic output, while housing price fluctuations influence consumption and investment through wealth effects (Mach, 2019). As demonstrated by the 2008 financial crisis, housing booms can stimulate growth, whereas abrupt downturns may trigger deep recessions and prolonged economic instability. Beyond its macroeconomic relevance, housing constitutes the main form of wealth accumulation for most households, making price dynamics particularly consequential for social welfare and inequality (Kettunen & Ruonavaara, 2021).

In Europe, the economic and social importance of housing is reinforced by high homeownership rates and the sector's contribution to employment across the value chain (James et al., 2024). However, strong regional differences exist, shaped by institutional frameworks, rental market structures, and public housing provision. Over recent decades, many European cities have experienced sustained price increases driven by low interest rates, international capital flows, urban land constraints, and demographic pressures. These trends have intensified affordability problems, particularly for young and low-income households, placing housing access at the center of public policy debates. Additional pressures, such as tourism expansion, migration flows, and the growing financialization of housing, have further reduced affordability and weakened the role of social housing in several countries (Acharya et al., 2024).

Spain exemplifies both the economic relevance and the vulnerabilities of the housing sector. With one of the highest homeownership rates in Europe, real estate represents the dominant component of household wealth. The housing boom of the early 2000s fueled consumption and employment but also created structural imbalances that culminated in a severe crisis after 2008. Although the market recovered after 2014 under more conservative credit conditions, recent price increases in major cities and tourist areas have reignited affordability concerns. Policy responses, including rent controls and tenant protections, reflect growing awareness of these challenges, while new analytical tools are increasingly used to monitor market risks (Bogatyreva et al., 2021).

Housing markets are inherently cyclical, alternating between phases of expansion and contraction. These cycles are often amplified by speculative behavior, abundant credit, and optimistic expectations that detach prices from economic fundamentals. When housing bubbles

burst, the consequences extend beyond the sector, eroding household wealth, constraining credit, and depressing public revenues. The Spanish experience highlights how recovery phases can coexist with renewed risks when demand rebounds faster than supply, underscoring the need for effective monitoring and policy intervention (Gil García & Martínez López, 2021).

Technological innovation, particularly through PropTech and Artificial Intelligence (AI), is reshaping the real estate sector. Digital platforms, big data analytics, and machine learning models have improved efficiency, transparency, and price estimation accuracy, often outperforming traditional valuation methods. AI applications now support real-time valuations, market monitoring, customer interaction, and asset management, while emerging explainable and fairness-oriented approaches seek to mitigate bias and enhance trust (Zhang & Buyuklieva, 2025). Nevertheless, these advances also raise concerns related to data concentration, regulatory compliance, and social outcomes, especially in highly regulated and privacy-sensitive European contexts.

Overall, the housing market emerges as a complex system where economic cycles, social dynamics, and technological innovation interact. While AI-driven tools offer significant potential to improve market transparency and decision-making, their deployment must be accompanied by robust regulatory frameworks to ensure fairness, data protection, and alignment with broader social objectives.

3. MATERIALS AND METHODOLOGY

The main objective of this study is the development of an advanced application capable of accurately estimating the market value of residential properties by leveraging data analytics and machine learning techniques. The proposed tool is designed to process large volumes of heterogeneous data and generate precise, real-time property valuations that support informed decision-making and anticipate market trends.

To achieve this objective, the study combines traditional real estate valuation principles with modern computational methods. First, a comparative property valuation approach is employed, whereby the value of a given property is estimated by comparing it with recently sold properties with similar characteristics in the same geographic area, adjusting for differences in size, location, and features. This method is enhanced through spatial analysis using the Haversine formula, which calculates the geodesic distance between properties based on latitude and longitude. This allows the model to account for the critical role of location in determining property value, including proximity to services, amenities, and neighborhood conditions.

Geographic accuracy is further improved through the integration of geocoding tools, specifically the Mapbox API, which converts addresses into precise geographic coordinates. This step enables spatial comparisons between properties and improves the predictive performance of machine learning models.

The methodological process follows several key stages. First, data collection is conducted using open-source online real estate platforms, primarily specialized property listing websites focused on the Madrid market. Next, data preprocessing ensures data quality through cleaning, transformation, encoding of categorical variables, and duplicate removal. Proper data governance practices are applied to guarantee consistency and reliability.

Subsequently, predictive models are developed and validated using advanced machine learning techniques, particularly Gradient Boosting algorithms such as the HistGradientBoostingRegressor. Model performance is assessed through cross-validation to prevent overfitting and ensure accurate price predictions. Geocoding is then applied to enrich the dataset with spatial variables, enabling more refined analysis of urban context and accessibility.

The final stage involves application development, implemented using Streamlit to create an interactive interface. The application allows users to modify property characteristics, visualize locations on an interactive map of Madrid, and obtain real-time valuation estimates.

To automate data extraction efficiently, the study employs Robotic Process Automation (RPA) using UiPath. This approach enables large-scale web scraping with reduced error rates and greater efficiency compared to manual methods, resulting in a robust dataset for model training. Overall, the methodology integrates automated data collection, precise spatial analysis, and advanced machine learning, aligning with established academic practices and demonstrating clear advantages over traditional valuation approaches.

Image 1. Code used in UiPath

Source: Own elaboration

The study employs advanced Gradient Boosting–based machine learning algorithms to analyze and predict real estate prices in the city of Madrid. Specifically, three models are used: XGBRegressor, LGBMRegressor, and HistGradientBoostingRegressor. These algorithms are recognized for their high predictive accuracy, efficiency in handling large datasets, and capacity to capture complex, non-linear relationships between property characteristics and prices. Their selection is supported by previous academic research demonstrating that Gradient Boosting

methods outperform traditional statistical approaches, particularly when combined with georeferenced variables.

Prior studies show that incorporating spatial information—such as geographic coordinates—into Gradient Boosting models significantly improves prediction accuracy, in some cases by up to 50%, as evidenced in applications to French and Spanish real estate markets. The integration of machine learning with geospatial analysis enhances the explanatory power of valuation models by accounting for spatial heterogeneity that conventional methods fail to capture, while maintaining a strong balance between accuracy and computational efficiency.

To ensure robustness and generalizability, the model training process relies on K-fold cross-validation, a widely accepted evaluation technique in real estate price modeling. The dataset is divided into five folds, with iterative training and testing across different partitions to prevent overfitting and obtain reliable performance estimates. This methodological choice aligns with best practices in the literature and strengthens the validity of the predictive results prior to final deployment of the model.

```
# Realizar validación cruzada
"""Se utiliza la validación cruzada k-fold con 5 pliegues (folds) utilizando la clase KFold de scikit-learn. Esto divide los
 La opción shuffle=True indica que los datos se barajan aleatoriamente antes de dividirlos en pliegues."""
cv = KFold(n_splits=5, shuffle=True, random_state=42)  # Dividir los datos en 5 pliegues (folds) para validación cruzada

 # Calcular MSE, RMSE, MAE y MAPE utilizando validación cruzada
mse_scores = cross_val_score(hist_gradient_boosting, X, y, scoring='neg_mean_squared_error', cv=cv)
rmse_scores = np.sqrt(-mse_scores)
mae_scores = cross_val_score(hist_gradient_boosting, X, y, scoring='neg_mean_absolute_error', cv=cv)
mape_scores = cross_val_score(hist_gradient_boosting, X, y, scoring='neg_mean_absolute_percentage_error', cv=cv)
r2_scores = cross_val_score(hist_gradient_boosting, X, y, scoring='r2', cv=cv)

# Calcular la media de los puntajes de validación cruzada
mse_mean = -mse_scores.mean()
rmse_mean = rmse_scores.mean()
mae_mean = -mae_scores.mean()
mape_mean = -mape_scores.mean()
r2_mean = r2_scores.mean()

print("Error cuadrático medio promedio (MSE):", mse_mean)
print("Root Mean Square Error promedio (RMSE):", rmse_mean)
print("Mean Absolute Error promedio (MAE):", mae_mean)
print("Mean Absolute Percentage Error promedio (MAPE):", mape_mean)
print("Coeficiente de determinación promedio (R²):", r2_mean)
```

Image 2. Development of K-Fold code with n=5

Source: Own elaboration

4. RESULTS

The configuration with a learning rate of 0.05 and a maximum depth of 7 was identified as the optimal setup for the HistGradientBoostingRegressor model. This configuration achieved the best overall performance across key evaluation metrics, including MSE, RMSE, MAE, MAPE, and R^2, outperforming alternative parameter combinations by minimizing prediction errors while maximizing explanatory power.

From a methodological perspective, this configuration offers an effective balance between bias and variance. The relatively low learning rate reduces the risk of overfitting, while a tree depth of 7 enables the model to capture the complex, non-linear relationships characteristic of the real

estate market in Madrid. Such complexity arises from the interaction between structural property attributes, location, and neighborhood effects, making deeper trees necessary for accurate modeling.

Building on this optimized model, the study successfully developed a web-based application designed to estimate property market values in Madrid using advanced data analysis and machine learning techniques. The application, publicly accessible online, features an intuitive and user-friendly interface, allowing users with no technical background to input property characteristics through sliders and checkboxes. Users can easily customize variables such as property size, number of rooms, and available amenities.

A key functionality of the application is the integration of interactive geospatial visualization. Using Folium and the Leaflet.js library, the tool displays the property's exact location on an interactive map of Madrid and highlights the five geographically closest comparable properties. This feature enhances transparency and provides users with valuable spatial context when interpreting price estimates.

In addition, the application includes dynamic visualizations that present real estate market data clearly and intuitively, such as price distributions by location and relationships between property features and values. Together, these elements demonstrate the practical applicability of the proposed system and confirm its potential as a decision-support tool for buyers, sellers, investors, and other stakeholders in the housing market.

5. DISCUSSION

The housing market currently faces severe challenges worldwide, marked by rising prices and declining affordability. In Spain, particularly in major cities such as Madrid, housing prices have reached levels comparable to the 2007 bubble, while access to housing has become increasingly difficult. In 2024 alone, prices rose by around 9.3%, far exceeding wage growth and excluding large segments of the population from the market, making housing the main socioeconomic concern for a significant share of Spaniards (Fernandez-Perez et al., 2025).

This situation reflects a broader, international phenomenon. Research shows that many advanced economies experience similar affordability crises, driven by financial, demographic, and policy-related factors. Cases from Ireland, Poland, the United States, Austria, the Netherlands, and the United Kingdom illustrate how external shocks, migration flows, health crises, financialization, and regulatory frameworks interact with local housing dynamics, producing heterogeneous but persistent pressures on access and prices (Vergara-Perucich, 2023). Consequently, housing affordability has become a systemic and increasingly global challenge.

In this context, the application of Big Data and Artificial Intelligence (AI) emerges as a key response to the complexity of modern housing markets (Hromada et al., 2023). These

technologies enable the processing of large volumes of heterogeneous data in real time, significantly improving price estimation accuracy, trend detection, and risk assessment compared to traditional valuation methods. The PropTech revolution exemplifies how digitalization and data-driven approaches are transforming real estate transactions, asset management, and market transparency.

The study demonstrates that Big Data and machine learning techniques can overcome the limitations of conventional models by capturing the structural complexity and spatial heterogeneity of urban housing markets, as illustrated by the case of Madrid. The results confirm strong intra-urban price segmentation, reinforcing evidence of growing inequalities between central, well-connected areas and peripheral zones with lower activity. This highlights the need for valuation models that integrate socio-spatial variables alongside economic factors. This study contributes to the existing literature by combining automated data extraction, fine-grained geospatial analysis, and Gradient Boosting–based machine learning models into an integrated and operational decision-support system for real estate valuation. Unlike many previous studies that focus either on methodological performance or on descriptive market analysis, this work bridges both dimensions by delivering a fully functional, user-oriented application capable of producing real-time property valuations. The originality of the research lies not only in the empirical application to a highly dynamic urban market such as Madrid, but also in its emphasis on practical usability, spatial transparency, and policy relevance, positioning the tool as a scalable framework adaptable to other metropolitan contexts.

AI-based models prove particularly effective in identifying complex, non-linear relationships and anticipating price dynamics in highly volatile urban environments. However, challenges related to interpretability, robustness, and fairness remain, underscoring the importance of explainable and transparent algorithms. Overall, the findings align with existing literature on housing financialization, speculative pressures, and supply constraints, suggesting that technological tools can support more informed decision-making but cannot fully offset deeper structural imbalances in housing markets.

6. CONCLUSION

This study contributes to the existing literature on real estate analytics and PropTech in three main ways. First, it advances machine learning–based valuation research by integrating automated data extraction, geospatial analysis, and Gradient Boosting models within a single, end-to-end framework, moving beyond purely methodological comparisons toward an applied decision-support system. Second, by focusing on Madrid as a highly dynamic and spatially heterogeneous urban market, the paper provides empirical evidence on intra-urban price segmentation using explainable, data-driven techniques, reinforcing the relevance of socio-spatial variables in valuation accuracy. Third, the development of an interactive application demonstrates the practical transferability of academic models to real-world contexts, offering a scalable tool for market participants and policymakers. In this sense, the study not only confirms

the predictive capacity of machine learning in housing markets but also bridges the gap between technical innovation and applied urban and housing policy analysis.

The results of this study confirm that machine learning models can capture the structural complexity and spatial heterogeneity of urban housing markets, as shown in the case of Madrid. Strong intra-urban price segmentation highlights growing inequalities between central and peripheral areas, underscoring the need for valuation models that integrate socio-spatial factors alongside economic variables. While AI-based tools offer substantial improvements in predictive performance and decision support, challenges related to interpretability, robustness, and fairness remain. Overall, technological innovation can support more informed decision-making but cannot, by itself, resolve the deeper structural imbalances affecting housing affordability.

REFERENCES

Acharya, D. B., Divya, B., & Kuppan, K. (2024). Explainable and fair AI: Balancing performance in financial and real estate machine learning models. *IEEE Access, 12*, 154022-154034. https://doi.org/10.1109/access.2024.3484409

Bogatyreva, M. V., Leskinen, M. I., & Kolmakov, M. A. (2021). The domestic real estate market during financial crises. *IOP Conference Series: Earth and Environmental Science, 751*(1), 012134. https://doi.org/10.1088/1755-1315/751/1/012134

Byrne, M., & Norris, M. (2022). Housing market financialization, neoliberalism and everyday retrenchment of social housing. *Environment and Planning A: Economy and Space, 54*(1), 182-198. https://doi.org/10.1177/0308518x19832614

Capellán, R. U., Luis Sánchez Ollero, J., & Pozo, A. G. (2021). The influence of the real estate investment trust in the real estate sector on the Costa del Sol. *European Research on Management and Business Economics, 27*(1), 100133. https://doi.org/10.1016/j.iedeen.2020.10.003

Fernandez-Perez, A., Gómez-Puig, M., & Sosvilla-Rivero, S. (2025). *El Clasico of housing: Bubbles in Madrid and Barcelona's real estate markets* (Preprint). SSRN. https://doi.org/10.2139/ssrn.5239332

Gil García, J., & Martínez López, M. A. (2021). State-Led actions reigniting the financialization of housing in Spain. *Housing, Theory and Society, 40*(1), 1-21. https://doi.org/10.1080/14036096.2021.2013316

Hromada, E., Heralová, R. S., Čermáková, K., Piecha, M., & Kadeřábková, B. (2023). Impacts of crisis on the real estate market depending on the development of the region. *Buildings, 13*(4), 896. https://doi.org/10.3390/buildings13040896

James, B. V., Joseph, D., & Daniel, N. (2024). Young adults' experience of housing and real estate chatbots in India: Effort expectancy moderated model. *International Journal of Housing Markets and Analysis, 17*(4), 1050-1066. https://doi.org/10.1108/ijhma-01-2023-0004

Kettunen, H., & Ruonavaara, H. (2021). Rent regulation in 21stcentury Europe. Comparative perspectives. *Housing Studies, 36*(9), 1446–1468. https://doi.org/10.1080/02673037.2020.1769564

Mach, Ł. (2019). Measuring and assessing the impact of the global economic crisis on European real property market. *Journal of Business Economics and Management, 20*(6), 1189-1209. https://doi.org/10.3846/jbem.2019.11234

Rampini, L., & Re Cecconi, F. (2022). Artificial intelligence algorithms to predict Italian real estate market prices. *Journal of Property Investment & Finance, 40*(6), 588-611. https://doi.org/10.1108/jpif-08-2021-0073

Reisenbichler, A. (2021). The politics of quantitative easing and housing stimulus by the Federal Reserve and European Central Bank, 2008–2018. In A. Johnston, & P. Kurzer (Eds.), *Bricks in the wall* (pp. 190-210). Routledge. https://doi.org/10.4324/9781003157182-8

Rey-Blanco, D., Arbués, P., López, F. A., & Páez, A. (2024). Using machine learning to identify spatial market segments. A reproducible study of major Spanish markets. *Environment and Planning B: Urban Analytics and City Science, 51*(1), 89-108. https://doi.org/10.1177/23998083231166952

Vergara-Perucich, J. F. (2023). A systematic bibliometric analysis of the real estate bubble phenomenon: A comprehensive review of the literature from 2007 to 2022. *International Journal of Financial Studies, 11*(3), 106. https://doi.org/10.3390/ijfs11030106

Zhang, Y., & Buyuklieva, B. (2025). Spatial cluster pattern and influencing factors of the housing market: An empirical study from the Chinese city of Shanghai. *Buildings, 15*(5), 708. https://doi.org/10.3390/buildings15050708