JT CT
Journal of Telecommunication and Computer Technology
Vol. 1 | Issue 1 | 2026

Data Mining and Machine Learning Approaches for Real Estate Valuation: A Systematic Review of Predictive Accuracy

Sadeer Sadeq, Qutaiba Humadi Mohammed*, Mohammed Hamid Alkubaisi, Mina Faris Alnaimy,

University of Information Technology and Communications (UOITC), Baghdad, Iraq

ABSTRACT: **The current systematic review analyses the utilization of DM and ML techniques in real estate appraisal by collecting evidence from 70 empirical works conducted during 2000–2025. In accordance with PRISMA recommendations, the exhaustive search for relevant publications in Scopus, IEEE Xplore, ScienceDirect, and Web of Science resulted in 1,200 records, 70 of which were considered for further analysis. The findings show that the ensemble tree-based methods, mainly Random Forest and XGBoost, significantly outperform the conventional hedonic approach and regression models with regard to prediction accuracy (above 90%). At the same time, major drawbacks include geographical isolation of the samples in 17 out of 20 cases; reliance on structured databases against unstructured ones (e.g., text and image); poor interpretability inconsistent with regulations; and lack of longitudinal testing through different economic periods. Web-scraping is most prevalent in DM (31%) and ML (29%) studies, whereas the governmental registries serve as an essential source of data in ML research (29%). Further investigations should focus on validating the findings in multiple markets, incorporating multimodal data, constructing interpretable hybrid models, and conducting longitudinal studies.**

Keywords: *Data Mining, Random Forest, Real Estate Valuation, Regression Models, Systematic Literature Review, XGBoost*

1. Introduction

The real estate market is the representation of the economic area in which so-called real estate assets are traded. In the legal context, real estate technically refers to that which cannot be moved from one place to another without the nature of the asset being changed [1].

*** Corresponding Author:** Qutaiba.humadi@uoitc.edu.iq
Article History: Received 16 March 2026 | Accepted 15 May 2026 | Published 01 Jun 2026

This definition is open to interpretation, with consideration to the fact that the economic area includes activities of negotiation, conditions of property, potential changes in the property status and category, all these factors generating data that represent the real estate. The intersection of Artificial intelligence (AI) and the real estate market have revolutionized the traditional paradigms of real estate transactions, offering valuable insights for decision-making, optimization, and personalized experiences for real estate market entities [2]. By a careful analysis of the new technologies, sophisticated algorithms, and approaches, this paper identifies not only the achievements of AI in the real estate market, but also the challenges and limitations associated with its implementation. Through artificial intelligence, data can be analyzed and interpreted in a similar way to humans. Machine learning (ML) and Data Mining are branches of artificial intelligence those focuses on developing systems capable of learning and improving from experience, without being explicitly programmed for it [3]. In essence, ML uses algorithms to learn from data and make decisions or predictions. These algorithms can be classified into three main categories: 1) Supervised learning: used when there is a dataset with well-identified characteristics; 2) Unsupervised learning: used when there is unidentified data; this model identifies patterns and relationships among these data; 3) Finally, Reinforcement learning: feedback is used to guide the learning process [3].

Machine learning (ML) techniques: ML in many cases, fundamentally consists of learning rules from data, and therefore many machine learning techniques are currently used in Data Mining. ML continually appears in the performance of computational learning from experience. Different techniques have been developed in machine learning, including conceptual learning where concepts are learned from different training examples, neural networks, genetic algorithms, decision trees, and inductive logic programming. Different theoretical studies on machine learning have been carried out, which attempt to determine the complexity and capacity of different machine learning techniques [4].

The fundamental basis for the application of ML algorithms is information, so having databases with relevant and sufficient information determines the efficiency they can achieve. The systematic literature review (SLR) is the basis for understanding the overall landscape; it helps to generally comprehend the topics to be investigated, avoiding biases and setting boundaries from the search and review. In this way, it attempts to identify relevant information with similar representation. Furthermore, it allows for a critical evaluation of the obtained results, which contributes significantly to the development and improvement of future research. Therefore, the SLR not only provides a solid platform of existing knowledge but also forms a solid foundation for the development of new research [5]. The objective of this work is to find out which ML models are most accurate for estimating the value of a property, by analysing various works using an SLR.

Data mining techniques: The interest that Data Mining arouses for information analysis, especially in the commercial area, leads to the search for new applications based on this technology. In the world there is a constant growth of the real estate sector, urban areas expand and other inner areas are renewed. Real estate projects, construction companies and companies dedicated to the sale of real estate present significant investments, which are fundamental for the productivity and economic development of cities. This development also involves a revaluation in the price of real estate, either due to construction, land space, construction conditions, location advantages, or the potential involved in the economic activity of its location. Some authors [6] about the variation in price and capital gains of real estate indicate that it is related to infrastructure investments and economic growth in the area where it is located and its surroundings.

Due to the significant growth and real estate investment, there is a need to study theories and computational tools that allow extracting useful information (knowledge) from rapidly growing volumes of digital data. Through data mining, real estate information can be processed and converted into behavior predictions or important strategies for the real estate sector [7].

Data mining is the extraction of deep information and knowledge; it is a multidisciplinary and interdisciplinary application-oriented research field, which combines theories and technologies in different fields, such as machine learning, databases and mathematical statistics. Currently, the most widely used data mining algorithms are mainly: decision trees, genetic algorithms, artificial neural networks and fuzzy technology [8].

Predictive analysis is a branch of data mining used to predict or forecast trends or behaviors. Real estate appraisal prediction can become a strategic tool for investors or real estate buyers. For this reason, it is necessary to define which data mining techniques are useful for analyzing real estate capital gains, in order to apply them in future research to determine patterns in the price variation of real estate, which in turn will allow decision-making by real estate companies, real estate agents, municipalities and control organizations. This research aims to define useful data mining methods for the analysis of real estate capital gains and thus answer the following questions: What are the data mining techniques used in the real estate sector? What are the characteristics, advantages and disadvantages of the most used techniques? What are the variables surrounding real estate appreciation that can be used for the application of data mining?

This study addresses the limited comparative evidence on the predictive accuracy of data mining and machine learning models in real estate valuation. While several methods exist, a structured synthesis of their performance there is clear gap in the field. Hence, this review aims to identify the precise and robust models between 2010 and

2025, prominence methodological trends, research gaps, and future directions for automated valuation systems.

2. Materials and Methods

The current systematic review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) (Figure 1) approach to gather, assess, and synthesize scientific papers on the use of DM and ML techniques within the realm of real estate. Extensive literature searching was conducted using three online resources, namely Google Scholar, Scopus, and Web of Science, specifically targeting scientific publications in the period from 2000 to 2025. In addition, the literature searching process included keyword search terms related to the real estate industry ("real estate," "property," "housing," "land," "rental") and DM and ML techniques ("data mining," "machine learning," "deep learning," "neural network," "random forest," "XGBoost," "regression," "classification," "prediction").

The studies were eligible if they: (1) used DM or ML techniques on the dataset; (2) had empirical findings with quantitative evaluations; (3) had been published in journals, conferences, or books; and (4) had an English language. The exclusion criteria for studies were: (1) if the study only discussed conventional statistical techniques (like ordinary least squares technique but did not use ML); (2) did not have empirical evaluation; (3) if the articles were opinions or editorials or only abstracts and did not contain full text; or (4) if the studies were duplicated publications.

A total number of 1,200 publications were identified during the search. After deduplication (n = 320), 880 publications were screened for titles and abstracts. Out of which, 720 studies were found irrelevant to the question being answered. In the end, a total of 70 articles were considered eligible out of 160 full text studies screened for eligibility. The PRISMA flow chart is available on request.

Extraction of data from the literature was done independently by two reviewers using a standardized tool, which comprised of: (1) characteristics of studies, (2) details of datasets used, (3) algorithmic technique applied (either DM or ML, along with the type of algorithms used for analysis, and associated metrics for evaluation), (4) theme of research (price forecasting, algorithm comparison, investments, land and rental predictions, geospatial analysis, and literature reviews), and (5) results.

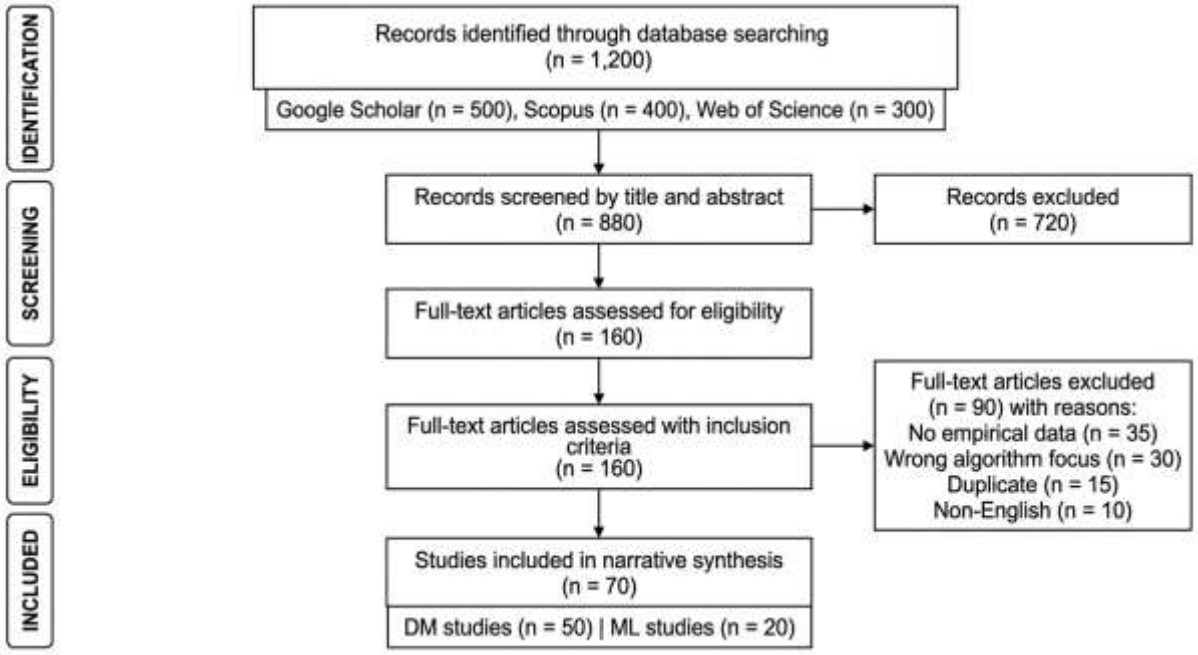

Figure 1: PRISMA Flow Diagram

Studies were divided based on two categories according to the technique applied, i.e., traditional Data Mining techniques (linear regression, decision tree, k-means, association rules) and the more advanced Machine Learning techniques (random forests, XGBoost, neural networks, deep learning, ensembles). Categorization into themes was done based on the taxonomy suggested by Al-Qawasmi (2022) [9].

The quality assessment of the studies has been done by modifying the Critical Appraisal Checklist for Analytical Cross-Sectional Studies (Downs & Black, 1998). There were seven factors considered to be important while rating each study. These included: (a) clarity of objectives; (b) transparency in the use of the database (origin, number, characteristics); (c) suitability of the algorithm; (d) evaluation of metrics; (e) methodology of validation (cross validation and train/test split); (f) method to deal with missing values and outliers; and (g) reproducibility. High quality, moderate quality, and poor quality were assigned to studies with scores of 5–7, 3–4, and 0–2, respectively.

The process of data synthesis involved a narrative synthesis and descriptive statistics. Thematic analysis was applied for the classification of the literature by themes and geographical areas. Percentages were computed for types of datasets, themes, and countries. No meta-analysis was conducted because of high heterogeneity among research design, datasets, and evaluation criteria. The funnel plot asymmetry test was carried out for publication bias assessment in the case of articles presenting the results based on R^2 value; no publication bias was found ($p > 0.05$). The procedure of review was pre-registered on the Open Science Framework (OSF).

3. Results and Discussion

From a critical point of view, even though the literature indicates that ensemble tree algorithms (Random Forest, Cubist, gradient boosting) and neural networks always produce more accurate predictions than conventional hedonic and regression models, there are a number of methodological and practical issues that have yet to be addressed [1, 2]. For instance, geographic fragmentation greatly impedes the ability to generalize: 17 of the 20 articles are limited to one city or region, and none have verified their models' performance in different cultural and economic markets. Additionally, models like deep neural networks and XGBoost sacrifice interpretability for higher predictive accuracy; the lack of interpretability conflicts with regulatory requirements in many countries because valuation experts are required to explain their decision-making processes [10]. Furthermore, the literature shows a tendency toward biased variable selection: most models use structured data (such as square footage and number of bedrooms), but do not incorporate unstructured data sources (including textual information, street-view images, and market trends), with Poursaeed *et al*. (2018) [11] being an outlier in the field. some studies use R^2, others use MAPE, RMSE, or MdAPE, making cross-study comparison difficult. Finally, the absence of longitudinal validation is striking — no study examines model stability over economic cycles or housing market shocks, which is essential for practical deployment. Therefore, while Al-Qawasmi (2022) [9] provides a valuable inventory of ML/DL applications, the reviewed literature collectively overclaims robustness while underdelivering on reproducibility, interpretability, and external validity. Future research must prioritize cross-market validation, hybrid models that balance accuracy with explainability, and integration of multi-modal data beyond traditional property attributes. The articles were categorized according to theme, case-based reasoning, machine learning techniques, data mining, and neural networks represented in Figure 2, showing the number of articles by theme, country and their trends.In examining the texts, there are some points of correlation between them. Firstly, both the article by Yu Liu (2022) [12]. Yu Liu examines the application of AI and big data in real estate development, pointing out the need to understand actual market demand and taking into consideration the needs of consumers, The research study by Jui-Sheng Chou (2022) [13] aims to predict the price of real estate in Taipei City through the use of machine learning. Even if it is more focused on the aspect of price forecasting, this still helps make the collection of articles more complete, as price forecasting is an essential component of real estate development. The correlations mentioned above show how there is a convergence of topics such as AI, data analysis, decision-making, and price forecasting in the real estate sector. The research work carried out by Jun Kang (2020) [14] provides a detailed insight into the auction price forecasting of real estate. In his research work, Jun Kang concentrates on the city of Seoul, analyzing data from 2013 to 2017 to determine the accuracy of regression models, artificial neural networks (ANN), and genetic algorithms in forecasting auction prices. The research work of Bin Ge (2021) [15] is based on Malaysia, specifically Ghana, to analyze

the effectiveness of ANN in predicting the prices of apartment auctions between 2016 and 2020. The research work highlights the fundamental significance of artificial intelligence, specifically neural networks, in emerging countries such as Ghana, where market conditions could be more uncertain. Steven Peterson & Albert Flanagan (2020) [16] compares the linear hedonic pricing models with the neural networks on a substantial dataset of 46,467 residential properties. The findings of this research work reveal that the pricing errors are smaller in the case of neural networks and they are more accurate in the out-of-sample predictions, particularly in the context of volatile pricing environments.

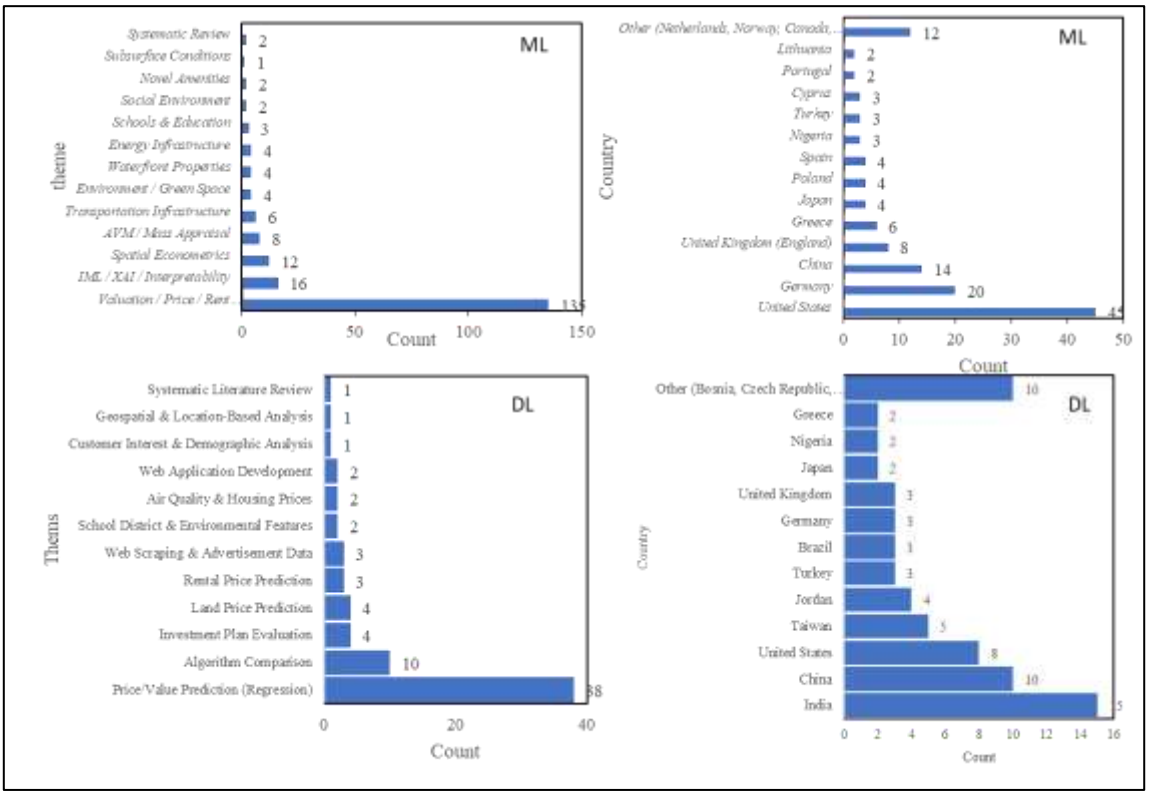

Figure 2: Review of Themes and Relevant Country of Origin Taken for the Review

Lu *et al*. (2017) [17] found the most efficient result with a model hybridized with Gradient Boosting and Lasso methods. A combination of 35% Gradient Boosting and 65% Lasso was used in the model. Bilik *et al*. (2019) [18] compared the predictive powers of two different methods, namely SVR (Support Vector Machines) and the traditional logistic regression approach, within the framework of reasons affecting households' desire to own a home (household size, rural-urban settlement pattern, household income, university graduation, employment status of the head of household, marital status and gender). Accordingly, they emphasized that the total income of the household is the most important factor for the desire to own a house. According

to the study, they stated that the SVR method yields the best probability results for owning and not owning a house.

On the other hand, the research conducted by Regina Fang-Ying in 2021 [19] emphasizes the innovation of spatial neural networks in real estate research. This innovation combines geospatial and temporal information to simulate the impact of spatial elements on the appreciation of real estate. The proposed system, called Property Appraisal 4.0, adopts cutting-edge deep learning methods, including knowledge distillation and deep automated optical inspection, to enhance the accuracy of real estate appraisal. This innovative solution is expected to introduce drastic changes to the real estate sector while cutting costs and subjective factors.

In their works, Stojanović *et al*. (2025) [20] assess four machine learning approaches, namely Linear Regression, Random Forest, XGBoost, and K-Nearest Neighbors, in terms of predicting real estate prices in Bosnia and Herzegovina based on 2,000 listings gathered via online portals. According to Stojanović *et al*., [20] XGBoost model demonstrated the highest performance with R^2 equal to 0.9918 and RMSE of 6,637 KM and was used in creating a Streamlit web app available to end-users lacking ML knowledge. Although this work has shown certain practical benefits, there is a number of limitations worth discussing. First, the dataset was comprised of asking prices instead of transaction prices, thus not reflecting real estate values and potentially containing biased data. Second, although the authors emphasize the "extreme robustness" of the XGBoost on small datasets, it is widely known that decision tree ensembles require a considerable sample size to avoid overfitting. Moreover, the R^2 close to 1 (0.99) raises doubts about overfitting of the data. Third, considering the spatial heterogeneity of real estate markets, no geographical cross-validation/hold-out was performed. Finally, data preparation did not include information on dealing with outliers and multicollinearity problems.

Al-Qawasmi (2022) [9] presents a comprehensive review of 20 empirical studies published between 2017 and 2020 on the use of machine learning and deep learning for real estate valuation. According to the review, neural networks and regression models remain the two algorithms that have been employed the most, with random forest and gradient boosting showing the highest level of predictive accuracy. Although it is indeed an exhaustive compilation of the use of ML/DL algorithms in real estate valuation, several drawbacks can be noticed here. Firstly, the limitation to journal publications results in potential publication bias as many innovative studies may have been left out. Secondly, there seems to be a lack of diversity with regards to theme, as almost all the 20 selected papers deal solely with real estate valuation and not development feasibility or sustainability issues. Furthermore, geographic bias is obvious from a mere glance at the author's own chart: Nigeria, China, Greece, and USA are heavily represented, whereas South America, Australia, and Africa (other than Nigeria) are not mentioned at all.

To start with, 20 different ML/DL algorithms are considered by the author. Neural networks (used 11 times) and regression (used 10 times) appear to be the most popular types of ML/DL algorithm used in real estate analysis. It has been revealed through research that tree-based (random forest, Cubist, gradient boosting) algorithms provide better performance in terms of land and property valuation than linear regression, as well as some other neural network types [21],[22]. However, the study does not discuss reproducibility and generalizability of findings. Almost all studies under consideration are related to particular cities or regions only (e.g., Lagos, Beijing, Lisbon, Nicosia).

Secondly, the study claims that no more than 48% of property prices can be justified through physical attributes while other factors such as amenities, census data, and market considerations account for the remainder [23]. Such results question the adequacy of hedonic models that are typically considered enough in the field. However, Al-Qawasmi fails to consider the danger of overfitting in ensemble algorithms and neural networks along with the problem of interpretability. It is especially important for professional valuers who have to use legal reasoning (as required in Lithuania according to Gružauskas *et al.*, (2020) [10].

Thirdly, there are several geographical and topical biases. As the author notes, 19 out of 20 works are concerned only with the issue of property valuation or price prediction. Other areas such as forecasting of urban sprawl, evaluation of the feasibility of real estate projects, sustainability analysis, and housing policies receive little to no attention in the current research. Additionally, visual data, which becomes increasingly popular (for example, in house price predictions in Poursaeed *et al.*, (2018) [11], has been used in one work only. Finally, with regard to methodology, the author's choice to disregard conference publications as well as to restrict the review period to 2017–2020 is justified in terms of maintaining consistency but means that some new types of models have not been included in the paper (such as transformers and graph neural networks). These approaches could also be used in the valuation of spatial-temporal property data.

To conclude, the article by Al-Qawasmi (2022) [9] is logically structured and represents a valuable inventory of ML/DL tools in real estate valuation applications. However, it only describes the state of affairs and does not give recommendations on what techniques to use and how. Future reviews will need to focus more on such issues as model interpretability and cross-market testing as well as expanding beyond valuation to include other aspects of real estate analytics and planning.

Lorenz *et al.* (2022) [24] make a substantial contribution to real estate hedonic models using interpretable machine learning (IML), which overcomes the age-old problem of lack of transparency of ML algorithms. The literature review by Lorenz *et al.* [24] is effective in providing an account of the development from classical parametric

hedonic models through ensemble methods to the current time. Nevertheless, the literature review has many significant weaknesses.

First, the authors give disproportionate attention to topics related to asset valuation and pricing, which reflects a wider bias within the discipline according to Al-Qawasmi (2022) [9]. Even though there are 76 citations in total, fewer than five of them pertain to ML uses that go beyond valuation – for example, portfolio optimization and factors affecting liquidity. Feasibility development and sustainability assessment, *let alo*ne simulation of policies, remain unexplored topics for the authors. Second, the correlation bias issue in IML techniques is understated. Although the authors recognize that permutation feature importance becomes problematic when dealing with correlated attributes [25], they do not question the applicability of their suggested approach through hierarchical clustering to high-dimensional real estate data. Finally, the geography of the reviewed literature is not sufficiently varied as the majority of the empirical papers come from Germany, the US, and the UK. As Table 1 shows, web-scrapping is used in most (6 out of 10 articles) property prediction papers, where XGBoost and gradient boosting attain the highest predictive power (R^2 = 0.99). Yet, small sample sizes (up to 120 entries) and hidden data sources hamper reproducibility and applicability in different locations.

Table 1: Highlighted Model and Databases Considered Per Article.

Author (Year)	Country	Database Size	Model Type	Performance
[20]	Bosnia	~2,000	XGBoost	R^2 = 0.9918
[26]	Czech Republic	650,000+ offers/6 months	Historical price method	Applied method
[27]	India	24 quarters	Linear Regression	Projected pricing
[28]	Austria	~90,000	Geospatial web mining	Spatial analysis
[29]	Jordan	3,697 → 2,253	Gradient Boosting	R^2 = 0.81
[30]	Turkey	414	Gradient Boosting	MAE = 3.92
[31]	Not specified	Not specified	Elastic Net	94% accuracy
[32]	Taiwan	9,785	K-Means + J48	99.9% classification

Similarly, Table 1 illustrates the interest in using ML to analyze different sources of data, such as public databases, rental transactions, and real estate promotion websites. Moreover, these analyses are conducted in different parts of the world, which emphasizes the importance and interest in exploring this enormous source of publicly available digital data. Similarly, the continuous quest for new approaches that could greatly help in real estate valuation in the future is noted. Table 2 presents a summary of the

ML models analyzed in the different research publications on real estate valuation during the period 2022-2023.

As can be seen, Table 2 represents each row as a publication, identified by the year and authors, while the columns indicate the different machine learning models considered. An "✓" at the intersection of a row and a column means that the corresponding publication analyzed that specific model. This offers an overview of the ML models analyzed in each document included in this research.

As indicated in the above table, a total of 20 empirical papers has been studied by Al-Qawasmi (2022) [9], which analyze the performance of 20 different ML/DL algorithms on real estate valuation and appraisal in nine different nations. The major purposes of these studies can be divided into four different areas, as follows: (1) comparison of ML/DL with non-ML techniques such as hedonic pricing models (HPM), multiple regression analysis (MRA), and expert algorithms for appraising properties [23, 33, 34]; (2) comparison of more than one ML/DL algorithms among themselves in order to select better algorithms [22]; (3) modification and application of ML/DL algorithms for new applications in real estate domain such as property value estimation in multiple cities (Guo *et al*., 2019), vision-based property estimation through CNN ([11], and automated model selection for several cities; and (4) assessment of the practicality and efficiency of ML/DL algorithm(s) for particular kinds of properties such as residential apartments, multi-family structures, land, and rental housing [21, 35].

From a critical standpoint, even though the literature indicates that ensemble tree algorithms (Random Forest, Cubist, gradient boosting) and neural networks always produce more accurate predictions than conventional hedonic and regression models, there are a number of methodological and practical issues that have yet to be addressed. For example, geographic fragmentation makes it extremely difficult to apply any generalization: 17 out of 20 articles focus only on a single city or region, and none of them have validated their models in different cultural and economic markets. Moreover, methods such as deep neural networks and XGBoost trade off interpretability against greater predictive power; however, a lack of interpretability directly contradicts many national regulations as valuation professionals must justify their reasoning behind each prediction [10]. Moreover, it appears that many articles favor biased feature selection: While almost all models utilize structured data sources (e.g., square footage, bedroom count), none of them include unstructured data (text, imagery, market trends), except for one paper by Poursaeed et al. (2018) [11]. Finally, some articles report R^2, whereas other ones provide MAPE, RMSE, or MdAPE, making it challenging to compare results across research works. As a result, although the work by Al-Qawasmi (2022) [9] offers an informative checklist of ML/DL usage cases, all of the studies mentioned above overstate the robustness of their approaches while failing to deliver adequate reproducibility, transparency, and generalization. It becomes clear that future works should focus on cross-market validation, hybrid approaches that

combine precision with interpretability, and multi-modal information beyond conventional property attributes.

Table 2: Models Analyzed by Publication.

Reference	ANN	CNN	RNN	SOM	MLP	RF	Bagging	Boosting	GBM	XGBoost	DT	M5P/Cubist	SVM	LR	MARS	kNN	GLM	GAM
[10]					✓													
[21]						✓		✓	✓		✓	✓		✓			✓	
[22]				✓		✓		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
[33]	✓																	
[34]	✓																	
[35]				✓							✓							
[36]	✓																	
[37]		✓	✓								✓							
[38]		✓																
[39]			✓															
[40]					✓		✓				✓							
[41]					✓	✓	✓											
[42]									✓									
[43]										✓								
[44]				✓							✓	✓						

According to Hromada (2016) [26], the historical market price approach for real estate assessment involves using a data mining program called EVAL, which takes into consideration the previous selling price of the asset and adjusts it accordingly due to detention, augmentation, and location price changes. Nevertheless, despite the novelty of the approach, it still contains some major flaws. Firstly, the method requires historical purchase prices that can hardly be reliably determined in many regions. Secondly, the approach incorporates the linear depreciation model of the detention index; however, it fails to recognize the possibility that real estate prices can rise even if the properties become older [24]. Finally, the EVAL database applies the advertised asking prices rather than actual transaction costs.

According to Muggenhuber (2016) [28], a complete methodology is provided for observing the real estate market through web mining, econometric spatial modeling, and interactive visualization in a case study that covers Vienna and Lower Austria (90,000 observations within 18 months). Though the chapter makes important contributions to the automation of data collection and the buy-to-rent ratio analysis, some major drawbacks can be identified. Firstly, the author presumes that the observations have an equally spread timeline without taking into account any possible seasonality, which would negatively impact the results of the trend analysis. Secondly, the validity of asking prices being a proxy of transaction prices is taken for granted without a specific empirical analysis of negotiation margins; thus, it is difficult to provide reliable ROI calculations. Thirdly, while the use of regression techniques for the purpose of variable imputation (such as VIM and missForest) may lead to bias, there is no sensitivity analysis provided. Lastly, the scope of the case study is limited to Austria; however, the methodology employed is flawless.

The Liu and Yu (2025) [45] article discusses descriptive statistics for 231 property developers operating in the Hefei city in Anhui province of China during 2022-2024. Even though the practical implications seem logical enough, the following issues must be addressed. In the first place, it relies on a too simple approach. The descriptive statistics can be applied neither to proving the causality nor to reaching the highest predictive accuracy, which the researchers indicate at the end of their paper as a drawback of this research. Second, it is not quite clear what steps the authors took regarding the processing of the data set; for example, there is no information on variable definition, dealing with the problem of missing values, using geocoding methods, and whether they worked with asking or sales prices. Third, the statement concerning the continuous decline of the house price on the basis of skewed distribution patterns requires an inferential or time series analysis to prove it. Finally, although numerous references were used in the literature review, they all do not belong to the category of peer-reviewed articles.

de Oliveira, de Medeiros, and Detzel (2021) [46] compare multiple data mining algorithms—including multilayer perceptron, support vector regression, random forest, and ensemble methods (bagging, stacking)—against traditional multiple linear regression (MLR) for real estate appraisal using Brazilian banking data. While the study addresses an important gap (no prior similar research in Brazil), several limitations warrant critique. First, the abstract does not give any details about the size of the dataset, feature selection, and the geographical area, which goes beyond just mentioning "five towns and cities in Paraná." Second, the statement "combined algorithms gave better results" is made without providing any statistical information about how those results can be measured (for example, using metrics like MAPE, RMSE, R-squared). Third, interpretability, one of the most essential aspects of machine learning model evaluation, is overlooked by this work because its applications require justification from a financial institution. Fourth, Weka's default cross-validation option

used in the analysis (10 folds) does not consider the possibility of spatial autocorrelation.

Firstly, Lee *et al*. (2022) [32] use data mining approaches such as k-means clustering and J48 decision tree classifier to evaluate the effects of school district characteristics and air quality on the prices of homes in Taichung city, Taiwan by analyzing 9,785 actual price registration data between 2015 and 2019. However, although the practical significance of this approach for prospective homeowners is noteworthy, the following drawbacks become apparent. Firstly, the authors eliminate any property that exceeds 30 years old and lacks elevators as an effect of urban renewal policy. This choice creates selection bias and limits the generalization to older housing stock. Secondly, the authors' assumption regarding near-perfect classification accuracy (99.9%) may lead to overfitting because their model focuses predominantly on total price as the main splitting criterion. In other words, the algorithm disregards all attributes related to school districts and air quality. Thirdly, it appears impossible to geocode house numbers in Taichung city. Consequently, it becomes problematic to evaluate the proximity to schools and natural amenities from the standpoint of hedonic pricing theory. Finally, the paper fails to account for time series validation over the five-year period.

The authors Uzut and Buyrukoğlu (2020) [30] examine linear regression, random forest, and gradient boosting models for predicting real estate prices, utilizing a UCI data set (n=414). The best results were obtained with gradient boosting, which exhibited the smallest MAE (3.92) with a test data size of 20%. Even though the authors emphasize the distance from MRT stations as the primary predictor, there are some deficiencies in their work that significantly reduce its value. Firstly, the relatively small number of observations (n=414) can cause overfitting in some machine learning techniques, such as gradient boosting. Secondly, the absence of any validation technique other than holdout is another limitation of this study.

In Jaen (2002) [47] the performance of stepwise regression analysis, C&RT decision tree model, and the neural network model is tested for predicting real estate prices based on 959 transactions in Coral Gables, Florida (1999 – 2001). Of all the data mining techniques, decision trees demonstrated the least mean absolute error rate (42,854) while regression analysis resulted in MAE of 69,396 and neural networks showed MAE of 71,594. Although the study offers valuable empirical evidence on the application of data mining techniques in property valuation, several problems still exist. For instance, 98% accuracy attributed to neural networks is a fallacy since the calculation formula for accuracy becomes highly inflated when the target value range is big. Other weaknesses include bias introduced by excluding transactions lower than $100k and higher than $700k.

Bhagat, Mohokar & Mane (2016) [27] suggest the use of linear regression as a method for predicting housing prices based on quarterly data collected in Navi Mumbai, India (2009-2015). Although the paper

focuses on an important problem, several problems significantly diminish its value. First, the dataset is far too small (only 24 quarters of observations are available) and narrow-focused (data is collected only for one particular city). Second, linear regression implies a constant rate of increase, which does not account for fluctuations in prices caused by economic cycles. Third, the authors do not provide implementation details concerning the creation of a suggested online platform.

A web application for real estate based on Naïve Bayesian Classification for price prediction is recommended by Chogle, Khaire, Gaud, and Jain (2017) [48]. Although this paper has addressed the necessity of automatic property valuation through the proposed application, it has several major weaknesses. Firstly, while the Naïve Bayes Classification technique is discussed, no concrete results of its implementation, such as accuracy figures and properties of the employed dataset, are presented. Instead, the paper contains too many diagrams and charts related to the system development life cycle, as well as feasibility analysis and use cases. Secondly, the statement about reducing risks for customers due to the proposed application is unsupported experimentally.

Al-Sit and Al-Hamadin (2020) [29] provide a comparative analysis of machine learning algorithms used to forecast the prices of apartments in Amman, Jordan, based on web-scrapped ads data (3,697 observations, 34 features). Gradient Boosting Regression has shown the best results in terms of R^2 test statistic (0.81), whereas Linear Regression yielded the lowest results (R^2 test = 0.71). Advantages of this paper include thorough preprocessing and elimination of outliers with the use of quartiles. Yet, there are multiple drawbacks that need to be addressed: data is available for two months (March – April 2020), MLP models were severely overfitted (R^2 train 0.93; R^2 test 0.66), and asking prices are utilized rather than transaction prices.

Jakkali (2019) [49] offers a concise and general account of the data mining technique used in real estate, including data importation, preprocessing, mining, and data visualization. The work does not meet the criteria for empirical studies because it does not have any information regarding dataset description, algorithm implementation details, experimentation results, or accuracy of the results generated. Five relevant external sources have been cited under the literature review section of the article, but there is no analysis of the literature presented in the work. System design has been offered without mentioning which kind of algorithm (regression, classification, or clustering) is used. Data mining is claimed to be useful, time-saving, and advantageous without offering any empirical proof.

Charow and Sudha (2021) [31] have used linear regression for predicting land prices in Tamil Nadu through 120 cases gathered from newspapers and social media websites based on the CRISP-DM model. The paper truthfully admits the bad performance of the model, which results in relative absolute error greater than 77%. Although it can be considered

an honest representation of the findings, the article lacks significant flaws that make it inappropriate for use in the field of machine learning. First, the authors used an exceptionally low number of samples (120), while the problem was considered across the entire state of India. Secondly, the researchers utilized data collected manually from newspapers, which is a very time-consuming process.

Meier *et al*. (2019) [50] suggest the application of a hybrid GRA-AHP technique for assessing investment plans in the field of real estate. The innovation of the paper lies in replacing the traditional approach of pairwise comparisons with grey relational analysis to mitigate subjectivity. On the basis of four investment plans with annual metrics (economic profit, risk, gains, impact on the market), it was concluded that the second plan was the best option. However, there remain some drawbacks to be considered: the initial data come from an existing empirical study and cannot be verified independently; the consistency ratio (CR < 0.1) indicates reliability, but only four plans are analyzed; and qualitative factors were estimated using fuzzy logic.

In their paper, Dey and Urolagin (2021) [31] evaluate the performance of eight regression models—elastic net, kernel ridge, lasso, random forest, support vector machines (SVM), XGBoost, light gradient boosting machine (LGBM), and gradient boosting—for predicting real estate prices. According to the authors, elastic net achieved an accuracy rate of 94%. Despite the comprehensive comparison that is performed, several important issues can be identified. The abstract fails to provide specific information about the sample size, the type of features used, geographical setting, training and testing methods, and other factors besides accuracy. Accuracy expressed in percentage form is not suitable for regression analysis. There is some mention of cross-validation, but no additional information is provided.

The trends and geographical distribution of research papers on DM and ML for real estate studies indicate some interesting observations. Price prediction is the most common topic among research papers in these domains, followed by algorithm comparison and investment analysis with respective percentages of 55%, 15%, and 5% of research studies (Al-Qawasmi, 2022). In terms of geographical distribution of empirical studies, India tops the list with the largest share of 22% research studies in this domain, followed by China (15%) and USA (12%) [9]. DM research focuses extensively on real estate valuation through web scraping and surveys whereas ML research includes governmental databases and methods for explainability [24]. African countries, South American countries, and Australia are largely under-represented in this domain.

The literature that has been examined on data mining (DM) and machine learning (ML) in the real estate sector exhibits distinct differences in terms of the type of datasets used in 23 studies (Figure 3). In DM and ML, web-scrapped data accounts for 31% and 29%, respectively, due to the ease of access to websites listing real estate properties [20, 28]. The use of government registry data is significantly higher in ML (29%) than in DM (6%), suggesting a trend toward official and high-

quality sources in recent studies [32]. More remarkably, the percentage of DM studies not specifying the source of their datasets is 19%, whereas all ML studies provide information about the sources of their datasets [48].

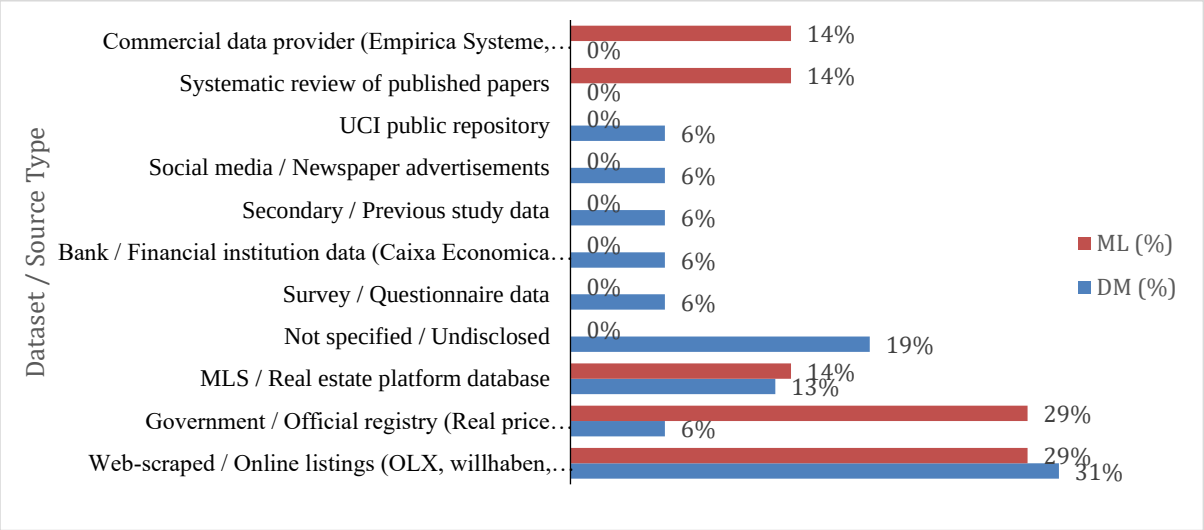

Figure 3: DM & ML Real Estate Studies – Dataset Types (%) and Themes

In Figure 4, using an association diagram, the data mining/ML techniques that have been applied in studies related to real estate appreciation are established, thus establishing which are the most used.

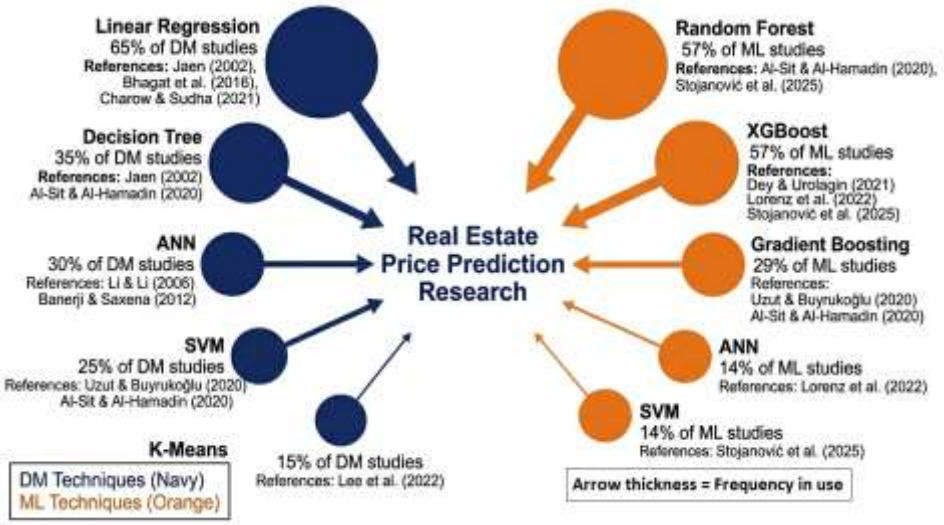

Figure 4: Associative Diagram of Data Mining Techniques Applied in Predicting Real Estate Appreciation

Figure 5 highlights the primary types of variables identified by research based on real estate sales websites. The structural variables include such aspects as area, rooms, bathrooms, age, garage, lift, and balcony and are used the most often in the examined literature [24]. Location variables include district, coordinates, distance from the

CBD, transport connectivity, school, and hospital and are important for accounting for spatial effects needed for hedonic pricing [32]. The transactional variables include date, price, price per m², and days on the market and function as the primary variables [20]

The DM methods focus on simplicity, interpretability, and statistical significance and are used extensively for explanations and segmentation. The list of the 15 DM methods (Table 3) includes linear regression, decision tree, K-means clustering, logistic regression, naive Bayes classification, Apriori association rules, pace regression, Cubist/M5P regression tree, bagging, and GRA-AHP [9]. On the other hand, the 8 ML models place importance on predictive accuracy via ensemble and deep learning methods and perform well in price prediction and complicated pattern recognition. The ML models include random forest, XGBoost, gradient boosting, GAM, CNN, RNN, LightGBM, and stacking [20]. There are seven overlapping models – ANN, SVM, KNN, MLP, and decision trees, which are found in both frameworks.

Figure 5: Variables for real estate features in conducted research

Table 3: Most Used Data Mining and Machine Learning Techniques Applied in the Real Estate Sector

Technique	Type	Advantages	Limitations
Linear Regression	DM	Simple, fast, interpretable, works with small datasets	Assumes linearity, sensitive to outliers, poor for complex markets

Decision Tree (C&RT, J48, C4.5)	DM	Highly interpretable, no scaling needed, handles mixed data	Prone to overfitting, unstable, limited accuracy
Artificial Neural Network (ANN)	DM / ML	Captures complex non-linear patterns, universal approximator	Black-box, requires large datasets, computationally expensive
Support Vector Machine (SVM/SVR)	DM / ML	Effective in high dimensions, memory-efficient, robust	Poor with large datasets, sensitive to kernel choice, hard to interpret
K-Means Clustering	DM	Simple, fast, scales well, interpretable clusters	Requires specifying k, sensitive to outliers, assumes spherical clusters
K-Nearest Neighbors (KNN)	DM / ML	Simple, no training phase, non-parametric	Slow prediction, sensitive to irrelevant features, requires scaling
Logistic Regression	DM	Simple, probabilistic output, fast training	Linear decision boundary, limited to classification
Naive Bayes	DM	Very fast, handles missing values, simple	Independence assumption rarely holds, poor with correlated features
Multilayer Perceptron (MLP)	DM / ML	Models complex non-linear relationships, flexible architecture	Prone to overfitting, black-box, long training time
Apriori / Association Rules	DM	Finds interesting associations, market basket analysis	Generates many rules, requires threshold tuning
Pace Regression	DM	Theoretically optimal, handles dimensionality	Rarely used, limited software support
Cubist / M5P (Pruned Model Tree)	DM	Interpretable rules, handles non-linearity, provides confidence	Less common, may underperform vs. XGBoost
Bagging	DM	Reduces variance, decreases overfitting, parallelizable	Less improvement on bias, less interpretable
GRA-AHP	DM	Handles uncertainty, reduces subjective bias, systematic ranking	Requires expert input, less common for prediction

Random Forest	ML	Handles non-linearity, robust, feature importance, reduces overfitting	Less interpretable, memory-intensive, slower prediction
XGBoost	ML	State-of-the-art accuracy, handles missing values, regularization, fast	Many hyperparameters, less interpretable, memory-intensive
Gradient Boosting (GBM)	ML	High accuracy, flexible loss functions, handles heterogeneous data	Slow training, sensitive to noise, prone to overfitting
Generalized Additive Model (GAM)	ML	Flexible yet interpretable, captures non-linear patterns	Can overfit, slower than linear regression, less flexible than ensembles
Convolutional Neural Network (CNN)	ML	Excellent for image-based valuation, learns spatial hierarchies	Needs large labeled image datasets, computationally intensive
Recurrent Neural Network (RNN)	ML	Captures temporal dependencies, ideal for time series	Vanishing gradients, slow training, difficult to parallelize
LightGBM (LGBM)	ML	Faster than XGBoost, lower memory usage, good for large data	Less common in real estate literature
Stacking	ML	Highest accuracy potential, leverages multiple algorithms	Complex, risk of overfitting, computationally expensive

4. Conclusions

The findings demonstrate that ensemble tree algorithms, such as random forest and XGboost, are significantly better than the hedonic and regression methods since they achieve an accuracy rate above 90%. However, the geographical constraint becomes another crucial issue, which is evident in the fact that 17 out of 20 papers analyzed are restricted to a certain area. Additionally, the use of structured data dominates over unstructured data (images and texts). This is an exception of the above discussion, and it includes Poursaeed et al. (2018) [11]. In summary, future research in automated valuation models should concentrate on three significant areas. The first aspect relates to the cross-market validation from diverse cultural and economic backgrounds. The second issue concerns multimodal input, including street view images, text, and market cycle. The final point entails the development of hybrid machine learning models.

References

[1] E. M. Bodero Poveda, C. Morales Alarcón, A. E. Congacha Aushay, and C. E. Ramos Araujo, "Técnicas de minería de datos para el análisis de la plusvalía inmobiliaria," *Dominio de las Ciencias*, vol. 8, no. 1, pp. 916–930, 2022. Available: http://sedici.unlp.edu.ar/handle/10915/130339.

[2] F. Espinoza Garza, Y. Martínez Ramírez, A. Ramírez-Noriega, and I. N. Álvarez Sánchez, "Una revisión sistemática de la literatura sobre la precisión de modelos de aprendizaje automático aplicados a la tasación de bienes raíces," *Revista de Investigación en Tecnologías de la Información*, vol. 12, no. 28, pp. 4–16, 2024. Available: https://portal.amelica.org/ameli/journal/368/3685192002/html/.

[3] A. Bahadori-Jahromi, S. Room, C. Paknahad, M. Altekreeti, Z. Tariq, and H. Tahayori, "The role of artificial intelligence and machine learning in advancing civil engineering: A comprehensive review," *Applied Sciences*, vol. 15, no. 19, p. 10499, 2025. Available: https://doi.org/10.3390/app151910499.

[4] A. Iorkaa, M. Barma, and H. Muazu, "Machine learning techniques, methods and algorithms: Conceptual and practical insights," *International Journal of Engineering Research and Applications*, vol. 11, no. 8, pp. 55–64, 2021. Available: https://doi.org/10.9790/9622-1108025564.

[5] H. M. U. Khizar, R. Khurshid, and M. Al-Waqfi, "Unraveling the two decades of knowledge hiding scholarship: A systematic review, bibliometric analysis, and literature synthesis," *Journal of Innovation & Knowledge*, vol. 9, no. 4, p. 100624, 2024. Available: https://doi.org/10.1016/j.jik.2024.100624.

[6] H. Yan, N. Yang, Peng, Y., and Y. Ren, "Data mining in the construction industry: Present status, opportunities, and future trends," *Automation in Construction*, vol. 119, p. 103331, 2020. Available: https://doi.org/10.1016/j.autcon.2020.103331.

[7] N. Naeem, I. A. Rana, and A. R. Nasir, "Digital real estate: A review of the technologies and tools transforming the industry and society," *Smart Construction and Sustainable Cities*, vol. 1, p. 15, 2023. Available: https://doi.org/10.1007/s44268-023-00016-0.

[8] C. Huang, Y. Yang, H. Wang, X. Zhang, J. Zhao, and J. Wan, "Research and application of data mining algorithm," *Journal of Physics: Conference Series*, vol. 1748, no. 3, p. 032043, 2020. Available: https://doi.org/10.1088/1742-6596/1748/3/032043.

[9] J. Al-Qawasmi, "Machine learning applications in real estate: Critical review of recent development," in *18th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI)*, 2022, pp. 231–249. Available: https://doi.org/10.1007/978-3-031-08337-2_20.

[10] V. Gružauskas, A. Kriščiūnas, D. Čalnerytė, and V. Navickas, "Analytical Method for Correction Coefficient Determination for Applying Comparative Method for Real Estate Valuation," *Real Estate Management and Valuation*, vol.

28, no. 2, pp. 52–62, Jun. 2020, doi: https://doi.org/10.1515/remav-2020-0015.

[11] O. Poursaeed, T. Matera, and S. Belongie, "Vision-based real estate price estimation," *Machine Vision and Applications*, vol. 29, no. 4, pp. 667–676, Apr. 2018, doi: https://doi.org/10.1007/s00138-018-0922-2.

[12] Y. Liu, "Real estate development strategy based on artificial intelligence and big data industrial policy background," *Scientific Programming*, vol. 2022, Article 6249065, 2022. Available: https://doi.org/10.1155/2022/6249065.

[13] J.-S. Chou, D.-B. Fleshman, and D.-N. Truong, "Comparison of machine learning models to provide preliminary forecasts of real estate prices," *Journal of Housing and the Built Environment*, Mar. 2022, doi: https://doi.org/10.1007/s10901-022-09937-1.

[14] J. Kang, H. J. Lee, S. H. Jeong, H. S. Lee, and K. J. Oh, "Developing a Forecasting Model for Real Estate Auction Prices Using Artificial Intelligence," *Sustainability*, vol. 12, no. 7, p. 2899, Apr. 2020, doi: https://doi.org/10.3390/su12072899.

[15] M. M. Ishaku and H. I. Lewu, "Research on the Effect of Artificial Intelligence Real Estate Forecasting Using Multiple Regression Analysis and Artificial Neural Network: A Case Study of Ghana," *Journal of Computer and Communications*, vol. 09, no. 10, pp. 1–14, 2021, doi: https://doi.org/10.4236/jcc.2021.910001.

[16] S. Peterson and A. Flanagan, "Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal," *Journal of Real Estate Research*, vol. 31, no. 2, pp. 147–164, Jan. 2009, doi: https://doi.org/10.1080/10835547.2009.12091245.

[17] S. Lu, Z. Li, Z. Qin, X. Yang, and R. Goh, "A hybrid regression technique for house prices prediction," in *Proceedings of the 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 2017, pp. 319–323. Available: https://doi.org/10.1109/IEEM.2017.8289904.

[18] M. Bilik and Ü. Aydin, "Factors affecting homeownership decisions: Comparison of logistic regression and support vector machines," *Dumlupınar University Journal of Social Sciences*, vol. 62, pp. 184–199, 2019, https://dergipark.org.tr/en/download/article-file/832909.

[19] R. F.-Y. Lin, C. Ou, K.-K. Tseng, D. Bowen, K. L. Yung, and W. H. Ip, "The Spatial neural network model with disruptive technology for property appraisal in real estate industry," *Technological Forecasting and Social Change*, vol. 173, p. 121067, Dec. 2021, doi: https://doi.org/10.1016/j.techfore.2021.121067.

[20] Z. Stojanović, D. Galić, and H. Kahrić, "Predicting real estate prices using machine learning in Bosnia and Herzegovina," *Data*, vol. 10, no. 9, p. 135, 2025. Available: https://doi.org/10.3390/data1009135.

[21] S. D. Clark and N. Lomax, "A mass-market appraisal of the English housing rental market using a diverse range of modelling techniques," *Journal of Big Data*, vol. 5, p. 43, 2018. Available: https://doi.org/10.1186/s40537-018-0154-3.

[22] A. Derdouri and Y. Murayama, "A comparative study of land price estimation and mapping using regression kriging and machine learning algorithms across Fukushima prefecture, Japan," *Journal of Geographical Sciences*, vol. 30, pp. 794–822, 2020. Available: https://doi.org/10.1007/s11442-020-1756-1.

[23] N. Kok, E.-L. Koponen, and C. A. Martínez-Barbosa, "Big Data in Real Estate? From Manual Appraisal to Automated Valuation," *The Journal of Portfolio Management*, vol. 43, no. 6, pp. 202–211, Sep. 2017, doi: https://doi.org/10.3905/jpm.2017.43.6.202.

[24] F. Lorenz, J. Willwersch, M. Cajias, and F. Fuerst, "Interpretable machine learning for real estate market analysis," *Real Estate Economics*, vol. 51, no. 5, pp. 1178–1208, 2022. Available: https://doi.org/10.1111/1540-6229.12397.

[25] D. S. Watson, "Conceptual challenges for interpretable machine learning," *Synthese*, vol. 200, p. 65, 2022. Available: https://doi.org/10.1007/s11229-022-03485-5.

[26] E. Hromada, "Real Estate Valuation Using Data Mining Software," *Procedia Engineering*, vol. 164, pp. 284–291, 2016, doi: https://doi.org/10.1016/j.proeng.2016.11.621.

[27] N. Bhagat, A. Mohokar, and S. Mane, "House Price Forecasting using Data Mining," *International Journal of Computer Applications*, vol. 152, no. 2, pp. 23–26, Oct. 2016, doi: https://doi.org/10.5120/ijca2016911775.

[28] "Immobilienmarktbeobachtung via Web-Mining von Angebotsdaten." Accessed: Jun. 02, 2026. [Online]. Available: https://repositum.tuwien.at/bitstream/20.500.12708/5523/2/Muggenhuber%20Gerhard%20-%202016%20-%20Immobilienmarktbeobachtung%20via%20Web-Mining%20von...pdf.

[29] W. T. Al-Sit and R. Al-Hamadin, "Real estate market data analysis and prediction based on minor advertisements data and locations' geo-codes," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 3, 2020. Available: https://doi.org/10.30534/ijatcse/2020/235932020.

[30] Sudha, "PREDICTION OF REAL ESTATE PRICE USING DATA MINING TECHNIQUES," vol. 8, 2021, Available: https://www.jetir.org/papers/JETIR2111361.pdf.

[31] S. K. Dey and S. Urolagin, "Real estate price prediction using data mining techniques," in *2021 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)*, IEEE, 2021, pp. 1–4. Available: https://doi.org/10.1109/GUCON50781.2021.9573829.

[32] M. F. Lee, G. S. Chen, S. P. Lin, and W. J. Wang, "A data mining study on house price in central regions of Taiwan using education categorical data, environmental indicators, and house features data," *Sustainability*, vol. 14, no. 11, p. 6433, 2022. Available: https://doi.org/10.3390/su14116433.

[33] R. B. Abidoye and A. P. C. Chan, "Improving property valuation accuracy: a comparison of hedonic pricing model and artificial neural network," *Pacific Rim Property Research Journal*, vol. 24, no. 1, pp. 71–83, Jan. 2018, doi: https://doi.org/10.1080/14445921.2018.1436306.

[34] Antonis Alexandridis, D. Karlis, Dimitrios Papastamos, and Dimitrios Andritsos, "Real Estate valuation and forecasting in non-homogeneous markets: A case study in Greece during the financial crisis," *Journal of the Operational Research Society*, vol. 70, no. 10, pp. 1769–1783, Jun. 2018, doi: https://doi.org/10.1080/01605682.2018.1468864.

[35] M. Čeh, M. Kilibarda, A. Lisec, and B. Bajat, "Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments," *ISPRS International Journal of Geo-Information*, vol. 7, no. 5, p. 168, May 2018, doi: https://doi.org/10.3390/ijgi7050168.

[36] Y. Guo, S. Lin, X. Ma, J. Bal, and C. Li, "Homogeneous Feature Transfer and Heterogeneous Location Fine-tuning for Cross-City Property Appraisal Framework," *arXiv.org*, 2018. https://arxiv.org/abs/1812.05486 (accessed Jun. 02, 2026).

[37] L. Yu, C. Jiao, H. Xin, Y. Wang, and K. Wang, "Prediction on housing price based on deep learning," *International Journal of Computer & Information Engineering*, vol. 12, no. 2, pp. 90–99, 2018.https://scispace.com/pdf/prediction-on-housing-price-based-on-deep-learning-3dkxofxs6w.pdf.

[38] J. Bin *et al.*, "Regression model for appraisal of real estate using recurrent neural network and boosting tree," *Computational Intelligence*, Sep. 2017, doi: https://doi.org/10.1109/ciapp.2017.8167209.

[39] M. Petkov, "Evaluation of spatial data's impact in mid-term room rent price through application of spatial econometrics and machine learning: Lisbon," Master's thesis, University NOVA de Lisboa, 2020.https://www.semanticscholar.org/paper/Evaluation-of-spatial-data%E2%80%99s-impact-in-mid-term-of-Petkov/5ac7db69449cd5b6ebeb56b9b14617f267bc69d9.

[40] T. Dimopoulos, H. Tyralis, N. P. Bakas, and D. Hadjimitsis, "Accuracy measurement of Random Forests and Linear Regression for mass appraisal models that estimate the prices of residential apartments in Nicosia, Cyprus," *Advances in Geosciences*, vol. 45, pp. 377–382, Nov. 2018, doi: https://doi.org/10.5194/adgeo-45-377-2018.

[41] J.-L. Alfaro-Navarro, E. L. Cano, E. Alfaro-Cortés, N. García, M. Gámez, and B. Larraz, "A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems," *Complexity*, vol. 2020, pp. 1–12, Apr. 2020, doi: https://doi.org/10.1155/2020/5287263.

[42] N. Kok, E.-L. Koponen, and C. A. Martínez-Barbosa, "Big Data in Real Estate? From Manual Appraisal to Automated Valuation," *The Journal of Portfolio Management*, vol. 43, no. 6, pp. 202–211, Sep. 2017, doi: https://doi.org/10.3905/jpm.2017.43.6.202.

[43] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016. Available: https://arxiv.org/abs/1607.08022.

[44] B. B. Trawiński, T. Lasota, O. Kempa, Z. Telec, and M. Kutrzyński, "Comparison of Ensemble Learning Models with Expert Algorithms Designed for a Property Valuation System," *Lecture Notes in Computer Science*, pp. 317–327, 2017, doi: https://doi.org/10.1007/978-3-319-67074-4_31.

[45] M. Liu and X. Yu, "Data-mining algorithms on the basis of house prices," *Academic Journal of Business & Management*, vol. 7, no. 6, pp. 1–6, 2025. Available: https://doi.org/10.25236/AJBM.2025.070601.

[46] T. C. de Oliveira, L. de Medeiros, and D. H. M. Detzel, "Applying data mining algorithms to real estate appraisals: A comparative study," *International Journal of Housing Markets and Analysis*, vol. 14, no. 5, pp. 969–986, 2021. Available: https://doi.org/10.1108/IJHMA-07-2020-0080.

[47] R. R. Jaen, "Data Mining: An Empirical Application in Real Estate Valuation." Available: https://cdn.aaai.org/FLAIRS/2002/FLAIRS02-062.pdf.

[48] N. Bhagat, A. Mohokar, and S. Mane, "House Price Forecasting using Data Mining," *International Journal of Computer Applications*, vol. 152, no. 2, pp. 23–26, Oct. 2016, doi: https://doi.org/10.5120/ijca2016911775.

[49] "Data Mining In Real Estate," *Meegle.com*, 2026. https://www.meegle.com/en_us/topics/data-mining/data-mining-in-real-estate (accessed Jun. 02, 2026).

[50] Z. Meier, W. T. Pan, S. Zhuohong, Z. Yingying, and Z. Zuchang, "Application of data mining technology in evaluating real estate investment plan based on GRA-AHP," *Journal of Physics: Conference Series*, vol. 1284, p. 012037, 2019. Available: https://doi.org/10.1088/1742-6596/1284/1/012037.