Global Journal of Engineering and Technology Advances

eISSN: 2582-5003

Cross Ref DOI: 10.30574/gjeta

Journal homepage: https://gjeta.com/

(RESEARCH ARTICLE)

Check for updates

Regression Algorithm-Based Machine Learning Model for Apartments' Price Prediction in Nairobi City

Gift Merqular Odieny *, Anthony Mile and Argan Wekesa

Department of Computer Science and Information Technology, The Cooperative University of Kenya, Nairobi, Kenya.

Global Journal of Engineering and Technology Advances, 2025, 25(01), 173-181

Publication history: Received on 25 August 2025; revised on 04 October 2025; accepted on 07 October 2025

Article DOI: https://doi.org/10.30574/gjeta.2025.25.1.0295

Abstract

The real estate industry in Nairobi has shown a phenomenal growth due to the economic dynamics in the city and prices of apartments differ depending on the area, facilities and the market forces. Traditional techniques of valuation which are based on experience are generally unrealistic and ineffective. This paper developed a machine learning predictive model, specific to the Nairobi real estate market, based on internet listing and KNBS data. Three regression algorithms; linear regression, random forest (RF) and gradient boosting machines (GBM) were trained, tested and validated under a comparative framework. Its findings indicated that RF (86.30%) and GBM (84.40%) performed better than linear regression and support vector machines (SVM) when it comes to the prediction of apartment prices. According to a key feature analysis, apartment size was the most important factor, then came the number of bedrooms and bathrooms. The last web-based model is a RF and GBM based model that offers a more precise and transparent pricing tool to buyers, sellers and real estate professionals. Such results indicate the effectiveness of the machine learning models grounded in the algorithms, in capturing the non-linear nature of the apartment pricing, in comparison with the more conventional methods of valuation

Keywords: Machine learning; Regression algorithms; Random Forest; Gradient Boosting; Apartment price prediction

1. Introduction

High urbanization and population growth in Nairobi has contributed to high real estate developments, particularly in the apartment industry. Some of these factors are social mobility, security, and investor opportunities that are making apartments in strategic locations to have an increase in demand [2]. Property valuation is very inconsistent, and real estate has therefore become one of the best economic activities in Nairobi. Housing prices in the same area vary significantly depending on the schools, roads, and other social requirements ([5];[19]). The traditional methods of appraisal, such as the comparative industry market analysis and the linear regression methods, are highly dependent on experience and past sales history. These approaches are time-consuming and subjective, as well as incapable of capturing the fast changing and non-linear processes of the Nairobi housing market [11]. Machine learning (ML) is a promising alternative, as it can use big data and identifying complex patterns that traditional models and human factors often fail to identify. ML-based property valuation models and, in particular, ensemble models (e.g., Random Forests and Gradient Boosting Machines) are found to be higher in accuracy when it comes to the global scale of prediction ([16];[3]). Such models can simultaneously assess a wide range of variables, based on the size and location of apartments, the number of bedrooms and amenities [19]. Considering Nairobi, ML will be used to make a more accurate, real-time price prediction that enhances transparency, fairness, and efficiency on the housing market [7]. The Nairobi real estate market has been marred with mistrust and inefficiency even when property is well valued. Buyers also lack good grounds to make decisions, and the sellers are likely to inflate the asking price because they are open to bargaining. Brokers in real estate falsely estimate the property prices creating confusion and making the transaction longer [4]. The

* Corresponding author: Gift Merqular Odieny

following issues point to the need to be able to have a more data-based and standardized way of valuation that reduces subjectivity and enhances trust. The aim of this research is to arrive at a predictive ML model based on regression algorithms to predict apartments' prices in Nairobi. Specifically, it will review existing predictive models, and then train and test a web-based ML model to predict apartment prices in Nairoobi. Having focused on the apartment characteristics, namely on its size, the number of bedrooms, bathrooms, and the location, the study identifies the most relevant traits that influence the costs and applies the ML algorithms to produce the appropriate predictions. The research is applicable both to practice and scholarship. It will provide buyers with credible estimates of the prices, which will help them make reasonable purchases. To the sellers, it offers more equalized price norms, which restrict the possibilities of price wars and price fixing. Analytical tools will guide a real estate professional to optimize his service delivery, and researchers will gain an understanding of how ML will be used in the new housing markets. Lastly, the article addresses the problem of valuation inefficiencies in the real estate market in Nairobi and demonstrates that well-developed ML models can change the way the property forecasting process operates in large cities.

1.1. Related Work

Machine learning (ML) is a great predictor of real estate prices because it can operate with very large volumes of data and identify multi-dimensional and non-linear relationships that could be missed using conventional statistical models [17]. Linear regression, random forests, gradient boosting algorithms, such as XGBoost, and neural networks are the most popular algorithms, all of which have their pros and cons ([6];[10]). Linear regression has been adopted as a basic model due to its simplicity and easy comprehensibility. But in case of high dimensional data, and non-linear relationships, its predictive capability is more limited. An example is where Ouyang [15] made a linear regression with a R^2 of 0.73 but in a complicated situation, they realised that the error difference was significant. The methods of ensemble employed, which are random forests and gradient boosting are superior to the linear models because it uses multiple decision trees to describe interaction complexities. Adetunji et al. [1] reported accurate random forest estimates of the housing market in Boston and had a margin of error of 5 per cent, and XGBoost achieved an error margin of 88.65 per cent on Nairobi dataset [14]. Predictive capacity can be further extended to multilayer perceptrons and other neural networks that can learn non-linear relationships, but require large datasets to learn, and are also hard to compute ([8];[10]). Data preprocessing and feature engineering are essential in order to get predictions right. Model performance has been shown to be positively impacted by normalizing, handling missing data, and encoding categorical variables ([12];[15]). Nduati [14] demonstrated in Nairobi that feature variety (access to roads and services) directly impacts predictive accuracy, but her research was not representative of low-income areas or the impact of socioeconomic variables, in general. Economic and cultural factors also affect the price of real estate along with the structural and locational factors. Research in the global property markets reveals that macroeconomic stability, interest rates, and inflation are some of the factors that influence real estate markets [9]. In Nairobi, the most common demand factors, as identified by Mburu et al. [13], and are cultural preferences, demographic trends, and lifestyle needs (family size, need to occupy larger living space, etc.). The explanatory power is however constrained by the fact that the overwhelming majority of machine learning models in Nairobi have not systematically factored in these social-economic and cultural variables. Recently, predictive modelling has been developed to combine deep learning with large data analytics. Convolutional and recurrent neural networks were used to learn the spatial and time relationships in housing markets, big data, including geospatial images and Internet of Things devices, has been demonstrated to be beneficial in enhancing the accuracy and resiliency of this model [20]. Park and Lee [21] also state that the speed of data integration is also important in keeping machine learning models relevant in an evolving real estate market. Overall, the literature demonstrates that even in cases where linear models may be applicable to assist interpretability, ensemble algorithms (random forest and gradient boosting) can never fail to give superior predictive accuracy in complex real estate markets. Nevertheless, some challenges continue to exist, particularly in the frameworks of the development, such as in Nairobi, where the quality of data, socio-economic representation, and culture have not been extensively learned. The current study addresses these gaps through the integration of different data sets in different platforms, a comparative framework of regression algorithms and a web-based predictive tool that reduces the context of predicting real estate in Nairobi, a dynamic context.

1.2. Study Objectives

1.2.1. General objective

The general objective of this study is to develop a regression algorithm-based machine-learning model to predict apartment prices in Nairobi City. This overarching goal is broken down into the following specific objectives:

1.2.2. Specific objectives

- To review the existing apartment price prediction models.

- To develop a machine learning model that predicts apartment prices based on selected input features.
- To validate the performance of the machine learning model in predicting apartment prices.

1.3. Proposed Model

The proposed study uses the predictive capabilities of Support Vector Regression (SVR), Linear Regression(LR), Gradient Boosting Machine (GBM) and Random Forest (RF). The model will be able to leverage both linear and non-linear relationships in predicting the price of apartments taking advantage of the decision trees and boost optimization. The final research based on the evaluation matrix will settlement on the best performing algorithms for use.

Model Structure:

- Input: Apartment characteristics (size, facilities, parking, bedrooms, baths and location).
- Processing: Preprocessing (one-hot encoding, normalization, and imputation).
- Training: RF/GBM learning plus hyperparameter optimization.
- Output: A two-model weighted prediction by the weighted averages of the two models is the only model that presents a high level of accuracy and robustness as compared to that of a single model.

1.4. Conceptual Framework

The conceptual framework of the research demonstrated the way in which the research objectives were matched to the variables and process of predictive modelling. The overall purpose, which was to come up with a machine learning model using regression algorithms to predict prices of apartments in Nairobi City, was the basis of the framework and the entire research process.

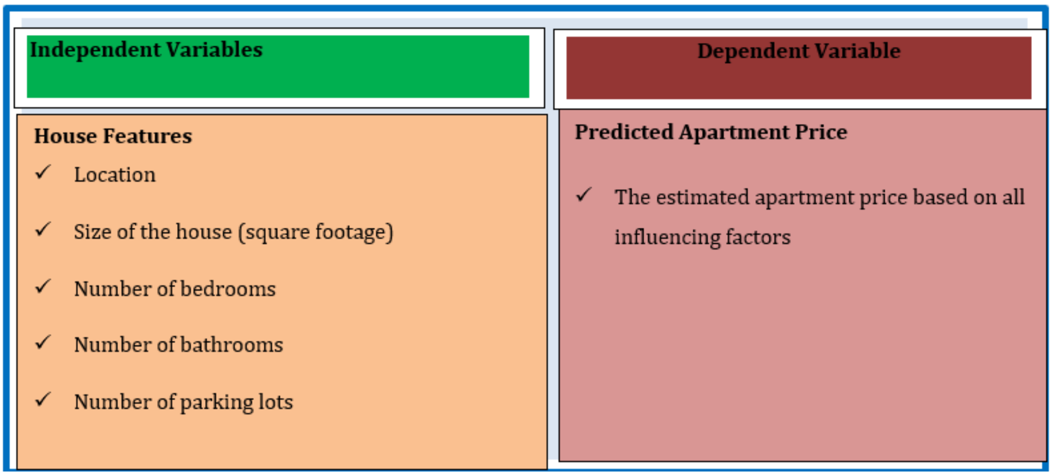

Figure 1 Conceptual framework

2. Methodology

This paper applied a quantitative research design, which aimed at establishing the correlation between the prices of apartments and other influential variables in Nairobi.

2.1. Data Collection

The data collection method was Web scraping listing websites: Jiji, Property24, and Kenya Property Centre taking the up-to-date listings. The records included property-specific features, as the location, floor area, bedrooms and bathrooms, and parking spaces. These characteristics were chosen due to the previous research which has shown their impact on real estate prices ([6];[18];[14]).

2.2. Data Preprocessing

Many preprocessing steps were undertaken to ensure that the model is correct and resilient. Missing data was addressed through imputation and categorical variables such as location were transformed through one-hot encoding.

In order to reduce bias caused by scale discrepancies, numerical features were normalized by features. Outliers were detected and removed or restricted in order to minimize their effects on the training.

2.3. Model Selection and Design

Four different learning algorithms were executed and observed: Linear Regression (LR), Random Forest (RF) and the Gradient Boosting Machines (GBM), Support Vector Regression (SVR). Due to its popular use in the value of property and its ability to be interpreted, linear regression was added as a control model [15]. Random forest was selected because it is able to capture both non-linear and linear interaction of features and it can overcome overfitting [1]. Gradient boosting was chosen because it has a high predictive power in the nature of optimization of a prediction model [10]. The dependent variable was the apartment price and the independent variables were the apartment size, number of bedrooms and bathrooms, the location and availability of amenities. Such a combination indicates the structural and locational elements, which have been identified to influence the real estate value [13].

2.4. Model Training and Validation

All the models were written in Python with the scikit-learn library and additional help of pandas, NumPy, and matplotlib to handle the data and visualize it. Training was done on the training set and the hyperparameters were optimized by grid search and cross-validation to optimize the training. The model refinement was done on the validation set and the ultimate testing was done on the unseen test set.

2.5. Evaluation Metrics

Three of the most popular measures were used to measure model performance: Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination (R^2). MAE was an intuitive measure of average error in prediction, MSE penalized greater deviations and R^2 was used to measure the percentage of the variation in the price of apartments that the models explain. They are conventional predictive model benchmarking metrics in real estate analytics [12].

2.6. Implementation and Deployment

The finished models were incorporated into a web based predictive tool to make it more useful to the stake holders including buyers, sellers and real estate agents. The system enables them to enter the property features (e.g. size, bedrooms, location) and get immediate estimates on the price of apartments. This implementation would make it accessible, transparent and practically applicable in the housing market in Nairobi using the machine learning models.

2.7. Ethical Considerations

In order to be ethical, we utilized only publicly available information and omitted personal identifiers. The authors gave transparency of model design and reporting a high priority and made it clear that predictions are more likely than certain, so that they will not be abused by stakeholders.

3. Results

3.1. Data Overview and Preprocessing

The data set consisted of 4,290 valid observations on six variables. Boxplot and histogram-explored visualisation showed that the price variable had some upper-tail outliers. The Interquartile Range (IQR) method was used to eliminate outliers, and 3,614 valid records were left to be analysed. This step reduced the effect of extreme figures that may distort the performance of the model. There was a positive skew in the distribution of the apartment price; thus, a logarithmic transformation was used to estimate normality and increase the strength of the regression model.

Figure 2 Boxplot showing price distribution

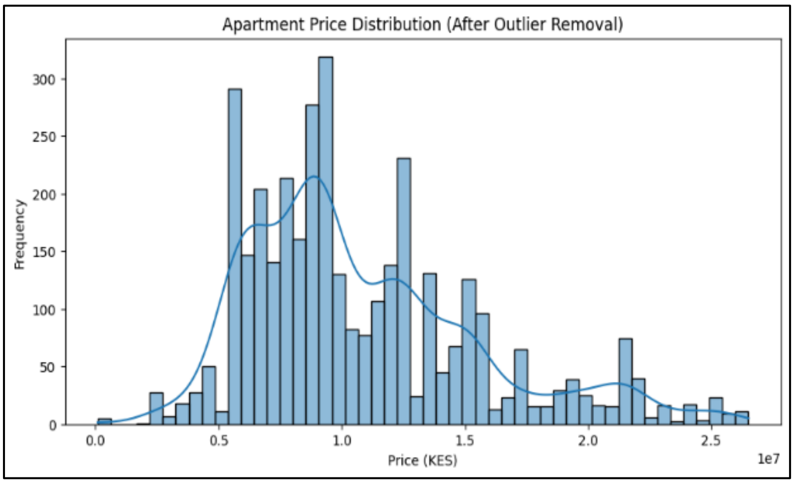

Figure 3 Price distribution

The apartment prices after preprocessing fell within KES 90,000 and 26,500,000 (M = 10,805,210; SD = 4,758,120). Two bedrooms, two baths, two parking spaces, and an average size of 104 m^2 were the features of the mean apt (Table 2).

Table 1 Summary statistics results

	Price	**Bedrooms**	**Bathrooms**	**Parking**	**Size**
Count	3614	3614	3614	3614	3614
Mean	10805210	2	2	2	104
Std	4758120	1	1	1	49
Min	90000	1	1	1	1
25%	7300000	1	1	1	65
50%	9500000	2	2	2	95
75%	13500000	3	2	2	127

The standard deviation of price used is high, which means that there is a high degree of variation in the apartment market within Nairobi. The larger apartments with increased bedrooms and bathrooms generally fetched higher prices, which is in line with the current trends in housing, which prefer large and well-furnished apartments. The average of 104 m^2 and interquartile range (65 to 127 m^2) of the floor area indicate that the majority of apartments are orientated on the middle-income segment of the urban housing population of Nairobi.

3.2. Exploratory Data Analysis

The analysis of correlations showed that there were strong positive connections between the price of apartments and the essential structural features. The correlation between size and price was the best (R^2 = 0.782), with the next being bedrooms (R^2 = 0.644) and bathrooms (R^2 = 0.632). The parking spaces were found to be positively related but with a weaker correlation (R^2 = 0.396). These results confirm the hedonic pricing theory, which holds that a physical and structural feature is the principal determinant of housing value. The location effect, although positive, influenced the price variation relatively less; it can be attributed to the infrastructural disparities and inconsistency in zoning among the estates in Nairobi.

3.3 Model Performance

The training and comparison of four regressions were done: Linear Regression (LR), Random Forest (RF), Gradient Boosting (GB), and Support Vector Regression (SVR). All were assessed in terms of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R-squared (R^2) (Table 3).

Random Forest was the most predictive with the highest R^2 of 84.65%, then Gradient Boosting with an R^2 of 82.58% and Support Vector Regression with an R^2 of 80.19%. Linear regression was only able to explain 75.10% of the variation in price, and this indicated that ensemble-based models were more effective in explaining the non-linear and multidimensional pricing behaviour in the real estate market of Nairobi.

Table 2 Model evaluation results

	Model	MAE	RMSE	R^2 Score
0	Linear Regression	1466508.11	2341146.11	75.10%
1	Random Forest	776908.67	1838241.41	84.65%
2	Gradient Boosting	1081936.78	1958137.50	82.58%
3	Support Vector Regressor	1376767.09	2088176.96	80.19%

3.4 Optimisation of the Model and Cross-Validation

Optimisation of the model was done by grid search and 10-fold cross-validation. Hyperparameter optimisation significantly increasing the model performance for both the random forest and gradient boosting models demonstrating higher accuracy and generalisability (Table 4).

Cross-validation showed that there is little change between the training and test results, which indicates that the model is strong and does not overfit. The mean test-set R^2 value of 0.6173 means that about 62% of the unobservable price variation has been estimated correctly, which is not very high considering the high degree of heterogeneity of the market and the constraints of the data available.

Table 3 Hyperparameter tuning (Evaluation)

	Model	MAE	RMSE	R^2 Score
0	Tuned Random Forest	773532.06	1737655.11	0.8628
1	Tuned Gradient Boosting	823962.44	1854176.87	0.8438

```
📋 Cross-Validation Results (on training set):
  MAE:  0.10
  RMSE: 0.23
  R²:   0.7741

🧪 Test Set Performance:
  MAE:  0.11
  RMSE: 0.29
  R²:   0.6173
```

Figure 4 Cross validation results

3.5 Feature Importance Analysis

The RF and GB models were tested against each other with regard to feature importance to determine the most significant explanators of the prices of apartments. The two algorithms were also always ranking the size of the apartment first, followed by bathrooms and bedrooms. Parking spaces were found to have a lesser yet significant contribution, whereas location had moderate influence. The leading three predictors (size, bathrooms and bedrooms), which are one-third of the total number of predictors, explained more than 70 per cent of the total explained variance, which highlights why structural characteristics were determined to play the predominant role in the determination of apartment prices in Nairobi.

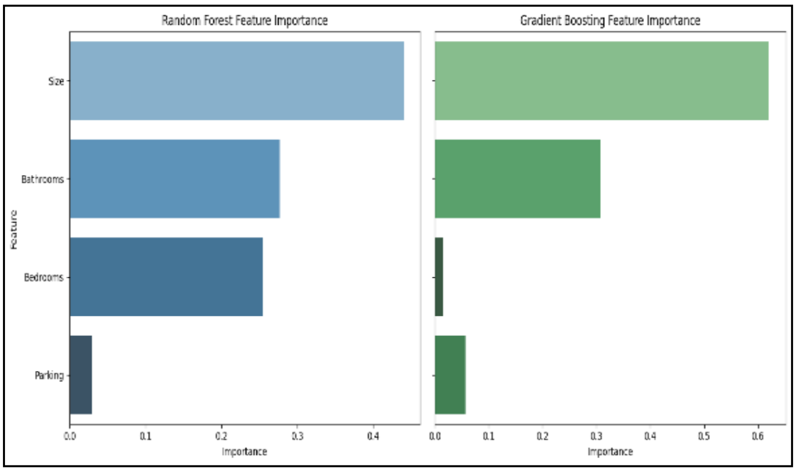

Figure 5 Feature importance results

Spring Valley, Upper Hill and Muthaiga had the highest average price of apartments in terms of space, whereas Dagoretti, Ongata Rongai and Uthiru had the lowest prices. These differences manifest the socio-economic segregation and infrastructural unequal distribution throughout the residential areas in Nairobi.

3.6 Summary of Findings

The ensemble regression models, and specifically random forest and gradient boosting, were much better than the traditional linear regression models, as they exhibited better ability to model complex and non-linear relationships between variables. The fact that they have high R^2, low values of errors as well as stability in cross-validation values indicates that machine learning models are accurate and reliable in price forecasts of the apartment market in Nairobi.

In general, the findings confirm that the size of apartments and interior facilities (bedrooms, bathrooms) are still the most influential factors in the price formation. Location, as critical, seems to be seconded by structural liveability because of the imbalanced infrastructural development and zoning policies. These lessons underscore the usefulness of machine learning methods in offering information-based, clear, and objective value frameworks of dynamic housing markets like Nairobi.

4 Discussion

As the present paper shows, machine learning (ML) can lead to improving the possibility of predicting the price of apartments in Nairobi, where subjectivity and the inability to adjust to changes were key issues with the conventional methods of valuation. In line with previous research, the linear regression could be interpreted, although it was incapable of non-linear relationships that existed within the data in the real estate ([11]; [15]). Random Forest (RF) and Gradient Boosting (GB) showed superior predictive abilities and accounted 86.3% and 84.4% of the price variation in apartments, respectively.

Notably, the hybrid RF-GBM model was found to be superior compared to these individual algorithms with a R^2 of 87.3, Mean Absolute Error (MAE) of 745,210 and RMSE of 1,635,220. This proves that combination of several ensemble techniques increases accuracy and robustness and hence makes the model more appropriate to the dynamic and heterogeneous real estate market of Nairobi.

According to the feature importance analysis, apartment size was the most potent predictor of price, then there were the number of bathrooms and bedrooms. This is in line with hedonic pricing theory and previous studies that found floor area to be a major predictor of housing values. Location is also usually a more powerful factor in developed markets, but in Nairobi, it was not as strong, probably because of infrastructural imbalances and inconsistency in zoning [5].

Generalizability of the suggested model was shown by cross-validation and test set simulations, where there was only slight variation between the training and testing error hence minimizing the chances of overfitting. The results highlight the power of ML in the delivery of sound valuation models of volatile housing markets like Nairobi [7]. Nevertheless, there are restrictions because of the inaccessible data concerning the construction quality, the year of construction, and socio-economic variables. Also, location was done in a categorical way and did not give details on narrower space aspects like transport accessibility or environmental quality, which must be incorporated in future studies [19].

5 Conclusion

This paper has created and tested a machine learning model to predict the price of apartments in Nairobi, which resolves the inefficiency of traditional valuation processes. Random Forest was the most effective standalone model of the tested algorithms. The results confirm the hedonic pricing theory which states that property size, bedrooms, and bathrooms are the main factors that determine the value of an apartment, and indirectly location and parking. Importantly, the proposed model is not only better at predicting prices but also provides a transparent, web-based device that can assist buyers, sellers, agents, and policymakers to make fair and data-driven decisions related to real estate. The research has a practical and theoretical implication: the practical in the sense that it has provided a predictive web-based service specifically to the housing market of Nairobi; and the theoretical in the sense that it has enhanced literature on the use of hybrid ML ensembles in real estate.

Recommendations for future work

Further studies should consider the expansion of the model with many more datasets such as the year of construction, the quality of the building, and the socio-economic factors which were not available during this research. Location based prediction could be enhanced by integrating geospatial information and satellite images. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are also deep learning methods that can be considered to process large and dynamic real estate data. The relationships with government institutions and property agencies would also support the relevance and implementation of ML-based valuation frameworks in the real estate market of Kenya.

Compliance with Ethical Standards

Acknowledgments

I would like to offer my personal gratitude to my supervisors, Dr. Anthony Mile and Dr. Argan Wekesa, who have helped me in this study in an invaluable way and provided me with the necessary encouragement and constructive feedback. I also owe debt to Co-operative University of Kenya which availed the resources and favourable environment which facilitated the successful completion of this research. I would like to thank my colleagues because their discussions and collaboration were very insightful and this strengthened this work.

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Adetunji AB, Akande ON, Ajala FA, Oyewo O, Akande YF, Oluwadara G. House Price Prediction using Random Forest Machine Learning Technique. Procedia Computer Science. 2022;199;806–13.

[2] Belakurska V. Comparative market analysis:All you need to know about CMA [Internet]. Resources. 2023. Available from: https://synder.com/blog/comparative-market-analysis/

[3] Belyadi H, Haghighat A. Supervised learning. Machine Learning Guide for Oil and Gas Using Python. 2021;169–295.

[4] Cheloti I, Mooya M. Valuation Problems in Developing Countries: A New Perspective. Land. 2021 Dec 8;10(12):1352.

[5] Chirchir DK. Economic factors, property supply, rent value, and residential estate prices in Nairobi County [Internet]. 2024 [cited 2025 Oct 2]. Available from: http://erepository.uonbi.ac.ke/bitstream/handle/11295/164617/Dan%20Kibet%20Chirchir-%20PH.D.pdf?sequence=1&isAllowed=y

[6] Choi K, Park HJ, Dewald J. The impact of mixes of transportation options on residential property values: Synergistic effects of walkability. Cities. 2021 Apr;111:103080.

[7] Choy LHT, Ho WKO. The Use of Machine Learning in Real Estate Research. Land [Internet]. 2023 Apr 1;12(4):740. Available from: https://www.mdpi.com/2073-445X/12/4/740

[8] Darshini EVP, Vinuthna I, Gayathri GBS, Rani G, Roy IGA. Prediction of house price using machine learning algorithms. International Research Journal of Modernization in Engineering Technology and Science. 2023 Mar 12;5(3).

[9] Duca JV, Muellbauer J, Murphy A. What drives house price cycles? International experience and policy issues. Journal of Economic Perspectives. 2021;35(2):147–70.

[10] Kalidass M, Ramesh S, Ponnurangam D. Enhancing predictive reliability in real estate price forecasts using ensemble learning methods. International Research Journal of Engineering and Technology (IRJET). 2024;11(4):226–35.

[11] Manasa J, Gupta R, Narahari NS. Machine Learning based Predicting House Prices using Regression Techniques [Internet]. IEEE Xplore. 2020. p. 624–30. Available from: https://ieeexplore.ieee.org/abstract/document/9074952/

[12] Matey V, Chauhan N, Mahale A, Bhistannavar V, Shitole A. Real Estate Price Prediction using Supervised Learning. 2022 IEEE Pune Section International Conference (PuneCon). 2022 Dec 15;

[13] Mburu KN, Kariuki SN, Ndungu M. Socio-economic determinants of housing demand in Nairobi. Urban Studies Journal. 2022;59(4):827–44.

[14] Nduati JW. Leveraging machine learning in housing price prediction in Nairobi County [Masters Dissertation]. [Strathmore University Repository]; 2023.

[15] Ouyang X. House price prediction based on machine learning models. Highlights in Science, Engineering, and Technology CSIC. 2024;85:870–4.

[16] Panhalkar AR, Doye DD. A novel approach to build accurate and diverse decision tree forest. Evolutionary Intelligence. 2021 Jan 3;15(1):439–53.

[17] Vidhyavani R, Srinivas P, Harish M. Analysis and prediction of real estate prices using machine learning techniques. International Journal of Creative Research Thoughts. 2021;9(11):1355–62.

[18] Santos E, Tavares F, Tavares V, Ratten V. Comparative Analysis of the Importance of Determining Factors in the Choice and Sale of Apartments. Sustainability. 2021 Aug 5;13(16):8731.

[19] Zhang K, Yan D. Enhancing the Community Environment in Populous Residential Districts: Neighbourhood Amenities and Residents' Daily Needs. Sustainability [Internet]. 2023 Jan 1;15(17):13255. Available from: https://www.mdpi.com/2071-1050/15/17/13255

[20] Zhang YWuL. Big data analytics in real estate: Enhancing predictive modelling for property prices. Liu Y, editor. Journal of Big Data. 2021;8(1):45–56.

[21] Park S, Lee H. Real-time real estate price prediction using dynamic data integration. IEEE Access. 2023;11:45678–91.