ABUAD Journal of Engineering Research and Development (AJERD)
ISSN (online): 2645-2685; ISSN (print): 2756-6811

Volume 7, Issue 2, 216-228

Prediction of Urban House Rental Prices in Lagos - Nigeria: A Machine Learning Approach

Sunday OLUYELE, Juwon AKINGBADE, Victor AKINODE, Royal IDOGHOR

Department of Computer Engineering, Federal University Oye Ekiti, Oye, Ekiti State, Nigeria
Sunday.oluyele.2826@fuoye.edu.ng/juwon.akingbade.1011@fuoye.edu.ng/victorakinode@gmail.com /royalidoghor@gmail.com

Corresponding Author: juwon.akingbade.1011@fuoye.edu.ng, +2349032869229

Date Submitted: 21/06/2024
Date Accepted: 14/08/2024
Date Published: 24/08/2024

Abstract: *Often, prospective tenants need to know the rental price of an apartment, and homeowners need to know how best to price their apartments. This work aims to predict house rental prices in Lagos, Nigeria, using machine learning by examining the relationship between the rental price and features such as the number of bedrooms, bathrooms, toilets, location and house status(newly built, furnished, and/or serviced). Five machine learning models were trained and evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R-Square (R2); the random forest regression model outperformed the other four models with the lowest MAE, RMSE and the highest R2. This study also revealed that the number of bedrooms and the apartment's location are the most significant predictors, confirmed using the feature importance analysis. The developed model can be used to estimate the rental price of a property in Lagos, Nigeria.*

Keywords: *Machine Learning, Rental Price Prediction, Real Estate, Regression, Random Forest.*

1. INTRODUCTION

Finding inexpensive and acceptable accommodation is a perpetual issue in cities worldwide, and Lagos, one of Nigeria's economic powerhouse, is no exception. Lagos constantly expands, and the complex housing market hampers the hunt for renters [1]. Lagos' population is expected to reach 32.6 million by 2050, making it one of Africa's fastest-growing megacities, and its exponential growth outpaces infrastructural development, particularly in housing, resulting in informal settlements with even more complex rental dynamics [2]. This rapid growth has impacted many communities, from family-friendly Ajah to trendy Lekki and Victoria Island, making it more difficult to find affordable rentals. The traditional process of getting an apartment rental price in Lagos, Nigeria, based on property qualities, location and neighbourhood patterns, is very subjective, prone to errors and time-consuming. This always leads to wrong and inaccurate pricing, which might lead to potential losses for both landlords and tenants. Likewise, renting an economical and reliable apartment in Lagos keeps getting more problematic due to the fast and constant urbanisation and complex housing dynamics. Due to the issues pointed out here, an objective and efficient method is needed to predict the rent of residential apartments in Lagos. Therefore, this research investigates how machine learning algorithms can predict the rental price of residential apartments in Lagos. The objectives involve collecting and analysing accurate rental price data on residential properties in Lagos, pointing out the important features that affect the price of apartment rent, building and training machine learning models with this data for rent prediction, evaluating the effectiveness of the developed models by testing it on sample data and assess the impact of using machine learning in rent house prediction in Lagos, Nigeria.

Machine learning's capacity to effectively predict house values based on specific features makes it an increasingly valuable prediction tool, even when previous pricing data is unavailable [3]. Several studies have used machine learning approaches to estimate real estate prices, such as houses, apartments, condominiums, commercial real estate (office buildings, retail spaces), vehicles (cars, trucks), and other assets.

The next section of this research article discusses the theoretical background and reviews some selected and relevant related works. Section 3 discusses the methodology employed, while Section 4 discusses the model result. Section 5 concludes the work and highlights the major findings and recommendations for future work.

2. RELATED WORKS

A study by Nwanko et al. [4] investigates the application of machine learning algorithms to build models for predicting house prices in Lagos, Nigeria. It talks about how the kind of house, number of bedrooms, and parking space all impact the price of a house. Regression methods, variance inflation factor (VIF), and train-test split were among the machine learning approaches used in building the prediction models. Findings indicate a relationship between price and the number of bedrooms, parking spaces, and housing types. Furthermore, the study noted that house type plays a vital role in predicting the price of an apartment.

Support vector regression, linear regression, and artificial neural networks are practical modelling algorithms for predicting home values, according to Chowhaan et al. [5]. Real estate brokers, geography, location, amenities, and economic considerations significantly influence property values. SVM scored poorly (20.64%), but Random Forest (90.35%) and XGBoost (88.66%) performed better than the other models that were assessed. Moderate accuracy was shown by several models, including Lasso Regression (80.36%) and Linear Regression (79.04%). Citing daily fluctuations impacting homeowners and the real estate market, the study emphasised the need for data extraction and analysis to estimate housing expenses accurately.

The application of machine learning algorithms to predict home prices is examined by Rawool et al. [6]. The study emphasises how crucial it is to accurately estimate home prices in the erratic real estate market to protect oneself from losing money. Machine learning approaches like K-Means, Random Forest, Decision Trees, and Linear Regression are used to generate prediction models. The results demonstrate that Random Forest Regression, with a Root Mean Square Error (RMSE) of 2.9131988953, has the highest accuracy in price prediction. According to the study, the well-established model can reduce the likelihood of financial loss by assisting people in making well-informed decisions when buying a home within their means.

A comparative study on house price prediction was done by Joshi and Swarndeep [7] using machine learning algorithms, highlighting the importance of accurate prediction in real estate transactions. Different machine learning algorithms, including Random Forest, Extreme Gradient Boosting, and Hybrid Regression, were used to build models for accurate property price projections based on location, area, and amenities. It was suggested that ensemble learning be employed to combine multiple models for more accurate results, considering each model's strengths and limits, with the hope of improved accuracy in house price estimates.

A study by Aghav et al. [8] used machine learning methods to estimate housing prices. They tested three models, linear regression, Lasso, and Decision Tree, and found accuracies of 85%, 72%, and 71%, respectively. The linear regression model was chosen because it is simple to use and successful in predicting continuous values. The model was trained on some data, and its hyper-parameters were fine-tuned using Grid Search Cross-Validation.

Another work by Sisman and Sisman [9] demonstrated the presence of correlations between visual appearance and non-visual attributes such as crime statistics, housing prices, and population density within urban areas. For instance, their work "City Forensics: Using Visual Elements to Predict Non-Visual City Attributes" employed visual features to forecast property sale prices. Employing classification and regression algorithms, it was found that factors like living area square footage, roof content, and neighbourhoods significantly influenced estimating home selling prices. Furthermore, their analysis indicated that using Principal Component Analysis (PCA) techniques could enhance prediction accuracy.

Mora-Garcia et al. [10] focused on tackling the pressing challenges successfully predicted, marked by escalating house prices amidst solid demand and inadequate housing supply. The research aimed to automate property value estimation based on provided features, which is critical for real estate entities striving for accurate dwelling price predictions. Using machine learning regression algorithms and a deep learning model with the Keras framework, the study meticulously pre-processed data, selected influential variables, and trained models across various datasets. CatBoost emerged as the top-performing algorithm, surpassing expectations with notable accuracy rates and minimal error. In contrast, the neural network, despite expectations, yielded comparatively lower accuracy, underscoring the need for extensive datasets to optimise deep learning methods' performance.

A research Bali and Vashistha [11] aimed to estimate a residence's rent by examining the client's demands and financial situation. The study was to assist clients in learning the actual rent for the residence and assist owners in learning the rental rate that best suits the client's requirements. Among all the algorithms compared, Random Forest regression provides the best results. It has a lower mean absolute error and root mean square error than other algorithms and an R^2 score of 84%, making it the most effective algorithm.

In a systematic review, Abdulsalam et al. [12] identified 14 machine learning models used in real estate for price and rent predictions. Based on research from the past eleven years, models like Random Forest (RF), Decision Tree (DT), Linear Regression (LR), and Support Vector Machine (SVM) are frequently recommended. Less commonly used models include Elastic Net (EN), RIPPER (RP), and Adaptive Boosting (ADB). These findings help stakeholders, including investors and appraisers, to predict property prices and rentals more accurately.

A work by Renigier-Bilozor et al. [13] divided real estate market analysis models into conventional and advanced valuation systems. The conventional models include multiple regression and stepwise regression methods, while the advanced models consist of hedonic pricing, artificial neural networks (ANN), and spatial analysis methods. Choosing the right model is essential due to the diverse options available, with regression analyses such as multiple linear regression, support vector regression, and hedonic regression being commonly utilised. Moreover, machine learning models like Catboost, XGBoost, LightGBM, random forest, and ANN are frequently used. This study employs multiple regression analysis, ridge regression, LightGBM, and XGBoost to analyse and predict the relationship between house attributes and prices.

Oyedeji et al. [14] conducted a study on residential property rental value classification in Osogbo, Nigeria, employing multiple artificial intelligence models. The research utilized Artificial Neural Network (ANN), Support Vector Machine with two different configurations (SVM C1 and SVM C2), and Logistic Regression to classify rental values based on various property attributes.The study categorized rental values into four distinct groups: less than ₦50,000 per annum, between ₦51,000 and ₦100,000 per annum, between ₦101,000 and ₦200,000 per annum, and between ₦201,000 and ₦300,000 per annum. The classification model varied with different correct classification accuracy with ANN achieving 89%, SVM with configuration C1 with 87.27%, while SVM with configuration C2 achieving 89.70% and logistic regression with 86.06%. In their models, some property characteristics were considered, such as proximity to cultural sites, kind of home, age of structure, and neighborhood security. Notably, their sensitivity analysis showed that the most important factor influencing Osogbo rental values was the distance from cultural sites.

Mathotaarachchi et al. [15] investigated the use of advanced machine Learning Techniques for prediction of property prices to help make informed decision making in the real estate sector. He employed different machine learning regression models. The dataset used in this research was sourced from the UK government's Land Registry making it a reliable source. On performance testing of the regression models, Random Forest performs with an R-Square of 0.93, XGBoost with 0.77, LightGBM with 0.53, CatBoost with 0.49, Linear regression with 0.97 and Hybrid Regression with 0.67. It was noted that Random Forest performed best as it achieved a high R-Square value of 0.99 on the training set, however a slight decrease was discovered on the testing set which is 0.93 which suggests some over-fitting.

A comprehensive research by Kalidass et al. [16] explore the use of machine learning algorithm to predict house prices. The research point out use of Random Forest and Gradient Boosting algorithms were very effective due to their lower mean square error(MSE). It was noted that Random Forest would provide reliable predictions due to its ability to handle complex relationships while Gradient Boosting performs best on processing large datasets to improve accuracy. Also Ensemble methods were used to combine individual predictions, resulting in a more accurate final result.

In order to develop a more robust approach to prediction of real estate price, Dabreo et al. [17], explore the use of some machine learning algorithms using two Kaggle datasets: Boston and Melbourne. The results show that XGboost Regression performs best on both datasets, followed by Random Forest and Gradient Boosting. The Decision Tree algorithm has a low performance and linear regression was the least accurate. The work also noted that datasets quality has impacts on prediction accuracy and it uses cross-validation to check for over-fitting, stating the importance of these algorithms and high quality data in price prediction.

In a research conducted by Mouna et al. [18] on prediction of housing prices using three machine learning algorithms on the same Melbourne dataset used by Dabreo et al. [17] which has 34,857 property sales and 21 features. The result shows that Gradient Boosting outperformed both linear regression and Random forest in terms of accuracy. Gradient Boosting explains most of the Variability in housing prices and offers the best predictive performance with an R-Squared value of 0.828 and the lowest error metrics (RMSE and MAE). This result shows Gradient Boosting as the most efficient and effective model for predicting house prices over the other models tested.

Choy and Ho [19] used 24,936 housing transaction records to investigate the use of machine learning in real estate pricing. They assess how well the Random Forest, k-Nearest Neighbors, and Extra Trees algorithms perform in comparison to a conventional hedonic price model. With R-squared values rising between 6.62% and 12.9%, the results suggest that these machine learning algorithms perform noticeably better than more conventional techniques. They also exhibit gains in explanatory power and error minimization. The study also highlights how crucial it is to address moral concerns about machine learning, including those pertaining to fairness, privacy, and employment displacement. Though ethical issues need to be handled by suitable regulations and laws, these algorithms' accurate pricing forecasts can assist sustainable real estate practices and improve market efficiency.

One thing common to all the related works and existing solutions here is that all have explored different machine-learning algorithms to predict real estate or residential apartment prices for purchase. Nwanko et al. [4] and Chowhaan et al. [5] have shown us the effectiveness of support vector regression, linear regression, artificial neural networks, and other algorithms in predicting the price of an apartment. It further shows that property values can be influenced by factors like housing type, number of bedrooms, parking space, and location. While Oyedeji et al. [14] is more closed to this research using classification model on prediction of rental property using Osogbo as a case study, there are some things that needs to be addressed, while the paper does not explicitly states the number of dataset used, a limited sample size could affect the models generalizability. Also classifying rental values into four groups might make the models lose some granularity in prediction. This approach might not capture subtle variations in rental prices within each category. This research addresses some of these limitations by building and training regression models specially designed for the Lagos housing market apartment rental price. It aim to provide a more precise solution for the prediction of rental prices by collecting and analysing current rental price data, identifying important features that influence the price of rents and evaluating the effectiveness of these models. It will offer a feasible and impactful tool for anyone interested in understanding how predictive models can help predict rental prices, as predictive models for the Lagos housing market are built and evaluated in this research.

3. MATERIALS AND METHOD

The dataset used for this work was scraped from PropertyPro, a Nigerian real estate factor like housing type, number of king space, and location that can influence property values on the 2nd of May, 2024 and contains house rents of apartments within Lagos State, Nigeria. The data contains 12269 rows with nine columns. Table 1 below shows the first five rows in this dataset. This dataset contains the following features:

i. Title: Description of the property.
ii. Price: Rental price, originally in various formats (e.g., per year, per month, per square meter).
iii. PID: Property ID.
iv. Location: Detailed location of the property.
v. Bedroom: Number of bedrooms.
vi. Bathroom: Number of bathrooms.
vii. Toilet: Number of toilets.
viii. Status: Indicates whether the property is newly built, furnished, and/or serviced.

The features were selected based on their relevance to rental price and their availability in the data source which is PropertyPro. While these features offer the primary determinants for rental prices, the addition of more features can improve the effectiveness of this work (see Section 5 for more details).

Table 1: The first 5 apartments in the dataset

Title	Price	PID	Location	Bedroom	Bathroom	Toilet	Status
4 bedroom House for rent Ikota Lekki Lagos	5,500,000/year	7KZAE	Ikota Lekki Lagos	4	5	5	Furnished, Serviced, Newly Built
2 bedroom House for rent Ikate Lekki Lagos	6,000,000/year	1LCWR	Ikate Lekki Lagos	2	3	3	Furnished, Serviced, Newly Built
2 bedroom House for rent Ikate Lekki Lagos	6,000,000/year	4LBHX	Ikate Lekki Lagos	2	3	3	Furnished, Serviced, Newly Built
2 bedroom House for rent Ikate Lekki Lagos	6,000,000/year	4LENM	Ikate Lekki Lagos	2	2	2	Serviced, Newly Built
2 bedroom Flat / Apartment for rent Aguda Surulere Lagos	2,500,000/year	8LCCU	Aguda Surulere Lagos	2	0	0	NaN

3.1 Model

Certain libraries were used in this work to model and analyse the data. The tools used in this work include Pandas for data manipulation and pre-processing, NumPy for numerical computations, Scikit-learn for machine learning model training, evaluation and pre-processing, and Matplotlib and Seaborn for data visualisation. Machine learning models considered for this work include ridge regression, random forest regressor, support vector regressor, gradient boosting regressor and linear regression. The choice of these models was guided by the need to evaluate different machine learning models, and to measure their effectiveness in generalizing well on unseen data. Linear regression offers a baseline performance and insight into linear relationship between the features and the rental price, decision tree helps capture the non-linear relationships in the data, random forest – an ensemble model improves on the decision tree by aggregating multiple decision trees, gradient boosting builds model sequentially where new model correct errors from previous ones, ridge regression is a regularized version of the linear regression that prevents over-fitting. All these models were used so their strengths could be assessed, and the most effective one for predicting rental prices could be selected.

3.1.1 Data cleaning and feature engineering

Before training, some features needed to be dropped, while some needed to be encoded. The PID and title columns were dropped. While exploring the dataset, it was discovered that the Price column has additional information, such as prices per annum and others per month. The '/month' and '/year' was removed in the column and multiplied the rows with '/month' by 12 to have a uniform price that represents the rents of apartments yearly. The cities were extracted in this data from the location column. A Python function was written to perform this operation. The status column was split into 3 new columns to represent an apartment newly built, furnished and/or serviced apartment. Table 2 shows the dataset after splitting. Additionally, the newly created city column was encoded using the one-hot-encoding – this is a technique used in machine learning to convert categorical variables like 'city' into 42 cities binary columns; each column corresponds to one possible category, and only one of these columns can take the value 1 for a given row, with all others taking 0. One-hot-encoding was used because many machine learning models perform much better with binary features, this helps enhance model interpretability and helps to understand model's decision. For the dataset, there are 42 cities, meaning 42 new columns will be created. Each row in a column will have a value of 1 if the corresponding city is present, and 0 for all other cities. After this, presence of duplicate rows were checked for, and it was found there are 1336 rows duplicates; and this was removed. The dataset was checked for outliers; outliers are data points that significantly differ from the other data points in the dataset. This can arise due to variability in the data, and a boxplot was employed to visualise if the data had outliers. Figure 1 shows the boxplot for the data. This reveals that some data points are far from others; some apartments in Ajah, Ikoyi and Apapa that cost up to 2.8 billion, 1.5 billion and 1.5 billion naira to rent.

Table 2: The first five apartments after splitting the status column

Title	Location	City	Bedroom	Bathroom	Toilet	Newly Built	Furnished	Serviced	Price	Price_new
4 bedroom House for rent Ikota Lekki Lagos	Ikota Lekki	Lekki	4	5	5	1	1	1	5,500,000/year	5500000
2 bedroom House for rent Ikate Lekki Lagos	Ikate Lekki Lagos	Lekki	2	3	3	1	1	1	6,000,000/year	6000000
2 bedroom House for rent Ikate Lekki Lagos	Ikate Lekki Lagos	Lekki	2	2	2	1	0	1	6,000,000/year	6000000
2 bedroom Flat / Apartment for rent Aguda Surulere Lagos	Aguda Surulere Lagos	Surulere	2	0	0	0	0	0	2,500,000/year	2500000
4 bedroom House for rent Off Harris	Off Harris Drive VGC Lekki Lagos	Lekki	4	5	5	1	0	0	6,000,000/year	6000000

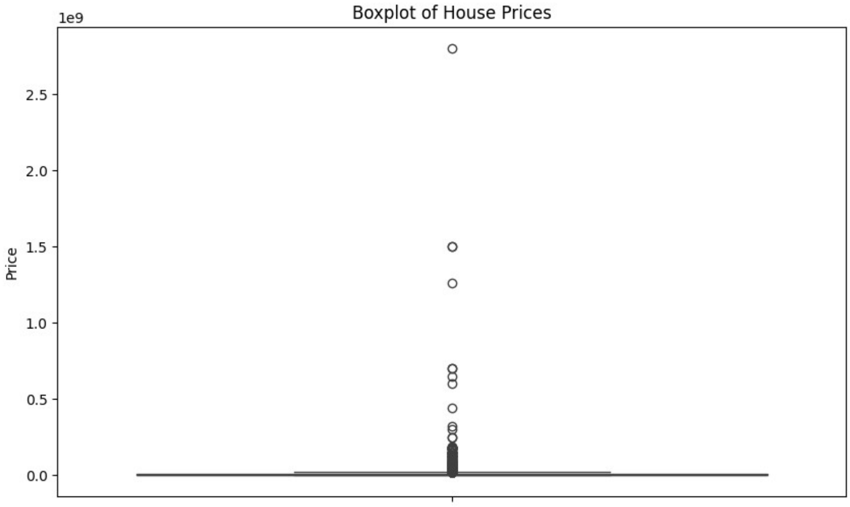


Figure 1: The boxplot of the data

These data points will significantly affect the machine-learning models by skewing the results, making them less representative of the actual data. The data points between the 10th and 90th percentile were removed to get rid of the outliers, reducing the total number of data points to 8295.

The value-count function in Python was used to count the number of data points in each city, as we have about 42 cities. Figure 2 shows a bar chart of the top 10 cities by number of houses; this shows that apartments in Lekki has a count of 3259, then Ajah with 1167 apartments, Yaba with 476, etc. Average prices per city was also computed; Figure 3 shows that Ikoyi takes the lead with approximately 12 million naira, followed by Victoria Island with approximately 10 million naira. Additionally, Orile appears to be a budget-friendly city with the lowest average price for apartment rentals.

Table 3 and Figure 4 also show the correlation matrix and heat map respectively. To investigate the correlation between different features of the dataset, it was seen that some variables correlate. For instance, the number of bathrooms is highly correlated with the number of bedrooms in the apartment; this means that as the number of bathrooms increases, the number of bedrooms also tends to increase. To investigate the presence of multi-colinearity in the data, the variance inflation factors (VIF) was used, which quantifies how much the variance of a regression coefficient is inflated due to multi-colinearity with other predictors in the model.

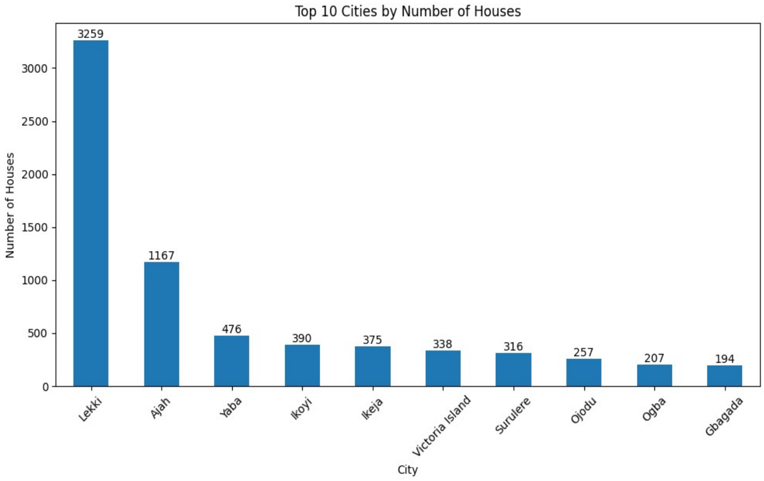


Figure 2: Bar chart of the top 10 cities by number of

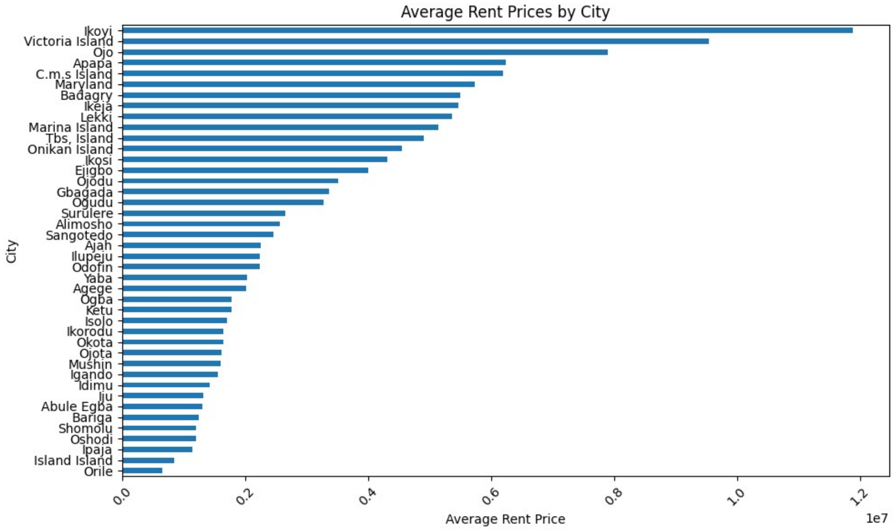


Figure 3: Bar chart of average rent prices by city

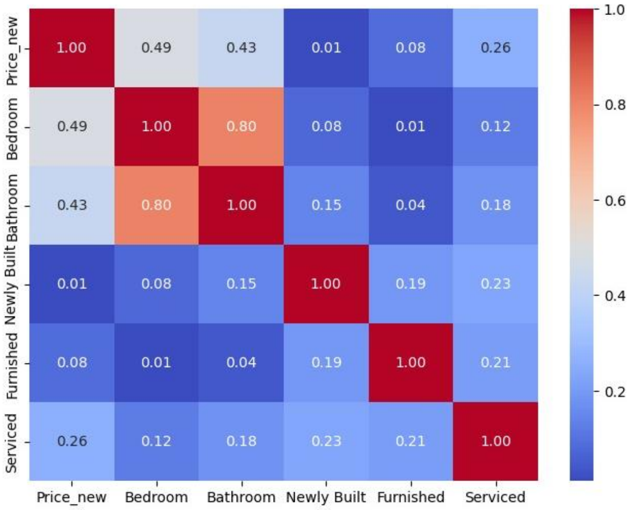


Figure 4: The correlation heat map

Table 3: The correlation matrix

	Price_new	Bedroom	Bathroom	Newly Built	Furnished	Serviced
Price_new	1.000000	0.489558	0.426901	0.014303	0.084425	0.257041
Bedroom	0.489558	1.000000	0.803193	0.076457	0.010937	0.120349
Bathroom	0.426901	0.803193	1.000000	0.145106	0.041540	0.178921
Newly Built	0.014303	0.076457	0.145106	1.000000	0.190575	0.230305
Furnished	0.084425	0.010937	0.041540	0.190575	1.000000	0.206841
Serviced	0.257041	0.120349	0.178921	0.230305	0.206841	1.000000

VIF is calculated as follows:

$$VIF(X_i) = \frac{1}{1 - R^2{}_i} \qquad (1)$$

Where $R^2{}_i$is the coefficient of determination obtained by regressing $X_i$on all the other predictor variables. The VIF values for the dataset are computed and shown in Table 4. The table does not show a VIF greater than ten, which indicates that there is no high multicollinearity [20].

Table 4: VIF values to check multi-colinearity

Features	VIF Value
Bedroom	6.547441
Bathroom	6.787154
Newly Built	1.469059
Furnished	1.190679
Serviced	1.463365

3.1.2 Model Training

The prediction models was built after completing all the preprocessing steps outlined above. This involves three stages which include:

1. **Data Sampling:** To evaluate the performance of the models effectively, a data sampling method was used that involves splitting the dataset into training and testing sets. The train_test_split function from the Scikit_learn tool provided by Python was used to divide the dataset into these sets. The training set comprises 80% of the original dataset, which is used to train the machine learning model, while the remaining 20% represents the testing set for the model. This set is reserved for testing purposes to evaluate the performance of the model. Table 5 showcases this distribution and reveals that 6636 apartments was used for the training and 1659 for testing.

Table 5: Train and test distribution

Set	Training
Training	6636
Testing	1659
Total	8295

2. **Model Fitting:** For this stage, five different models was explored to determine which performs best in generalizability. Five models were fitted: linear regression, decision tree regression, random forest regression, gradient boosting regression, and ridge regression.
3. **Model Evaluation:** To determine the efficacy of the trained models, their performance was evaluated using the following metrics:
   a. **Mean Absolute Error (MAE):** This measures the average magnitude of prediction errors without considering their direction. The equation is as follows:

   $$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \qquad (2)$$

   where $y_i$is the actual value, $\hat{y}$, is the predicted value, and n is the number of observations.

   b. **Root Mean Squared Error (RMSE):** This measures the square root of the average squared differences between predicted and actual values. The equation is as follows:

   $$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \qquad (3)$$

   where $y_i$is the actual value, $\hat{y}$, is the predicted value, and n is the number of observations.

   c. **R-Squared (R^2):** This measures the proportion of the variance in the dependent variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}\left(y_i - \underline{y}_i \right)^2} \qquad (4)$$

where y_i is the actual value, $\hat{y}$, is the predicted value, $\underline{y}_i$ is the mean of the actual values, and n is the number of observations.

A baseline Mean Absolute Error was computed for comparison to provide a reference point for the model evaluations. A baseline model was used for this operation, predicting the mean of the train data for all the test instances. The baseline MAE was calculated using the formula below:

$$Baseline\ MAE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \underline{y}_{train} \right) \qquad (5)$$

where y_i is the actual value, $\hat{y}$, is the predicted value, $\underline{y}_{train}$ is the mean of the training data, and n is the number of observations.

4. RESULTS AND DISCUSSION

Five machine learning models were trained for prediction, and the results gotten are shown in Table 6, which compares the models using the metrics mentioned in the previous section in terms of Mean Absolute Error, Root Mean Square Error and R-squared. Additionally, these metrics were visualize using bar charts, shown in Figure 5, 6 and 7. Table 6 shows that the Random Forest model performs better than other models for the dataset used. The lowest mean absolute error of 1,539,994.03, the lowest root mean square error of 2,395,925.34, and the highest R-squared of 0.63 were attained. This observation is further highlighted in Figure 5, 6 and 7 to show that the random forest is a better performer for the dataset.

Table 6: MAE, RMSE and R^2 for the trained models

Model	MAE	RMSE	R^2
Linear Regression	1855116.81	2724695.35	0.52
Decision Trees	1564603.27	2454366.19	0.61
Random Forest	1539994.03	2395925.34	0.63
Gradient Boosting	1643174.26	2476878.25	0.60
Ridge Regression	1859749.25	2724151.24	0.52

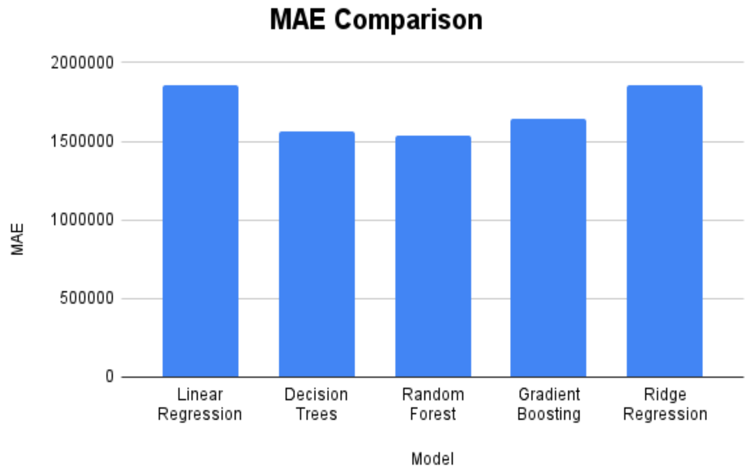


Figure 5: Bar charts showing the relationship between MAE among the five models

Comparing the baseline mean absolute error with the mean absolute error of the chosen model, insight can be gained into the effectiveness of the model compared to the simple baseline model. The baseline MAE was calculated using equation 5 to be 4438742.51; comparing this with the random forest MAE, which is 1539994.03, it can be seen that there is a substantial reduction in the MAE. This means that the developed model has captured significant patterns in the data and provides more accurate predictions than the baseline model. This further indicates that the model can predict more accurate house rental prices in Lagos, Nigeria.

20 instances from the test dataset were compared to observe the variation between the actual house rental prices and the predictions. The analysis showed that there are not many price variations, with predictions being close to the actual prices in most cases, as visualized in Figure 8.

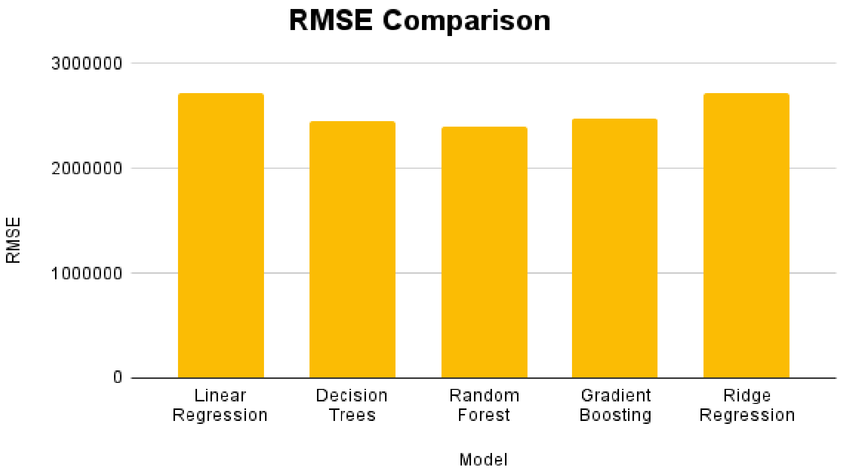


Figure 6: Bar charts showing the relationship between RMSE among the five models

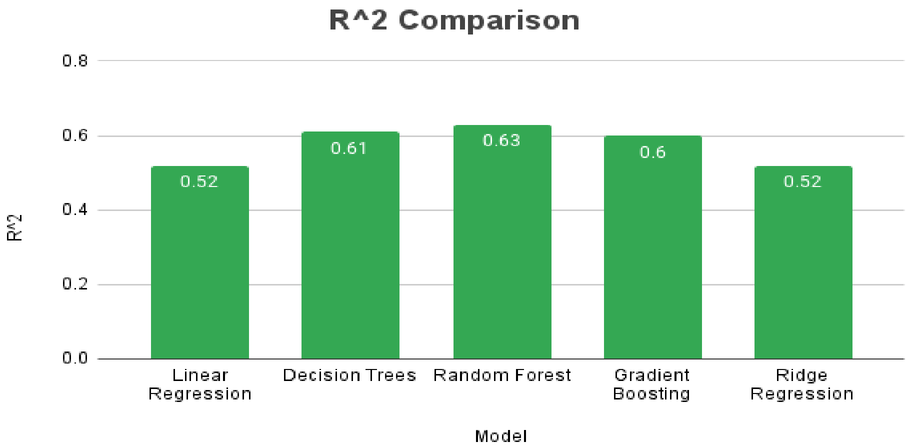


Figure 7: Bar charts showing the relationship between R2 among the five models

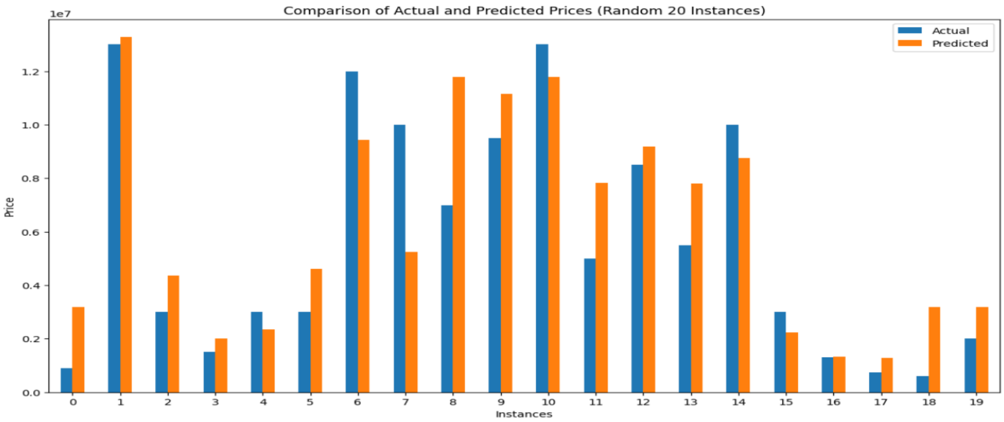


Figure 8: Bar chart comparing actual and predicted rental prices using 20 random instances

For the developed random forest model, importance of the feature was investigated. Feature importance is a concept in machine learning that helps us understand the impact of the features in predicting the target feature. This technique calculates a score for each feature used for prediction for a given model; these scores represent the 'importance' of that particular feature. A higher score means that the specific feature will have a more significant effect on the model in predicting a certain variable. In this case, the impact of the features such as bedroom, bathroom, toilet, newly built, services, furnished and city on the random forest model's prediction of house rental prices was examined. Figure 9 visualises this importance, and it was observed that the number of bedrooms and the city of location of the apartments have the most influence on the rental price of the apartments.

Real-life predictions were made to test the model and underscore the developed model's effectiveness. The model was tested using six arbitrary instances with features chosen using the interactive GUI called the ipywidgets available in Jupiter Notebook. The results of these predictions are found in Figure 10.

Figure 8 a, b, and c support the feature importance claims, indicating that location significantly impacts apartment prices in Lagos. A newly built apartment in Lekki can be rented for approximately 1.8 million naira, while the same apartment rents for approximately 827 thousand and 3 million naira in Shomolu and Victoria Island, respectively. In the same light, it can be seen that the impact of the number of bedrooms on the rental price of an apartment; taking Ajah as an example, in Figure 8 d, e and f, it can also be seen that for a bedroom apartment, the rent is predicted to be 733 thousand naira. In contrast, apartments in Ajah with 2 and 3 bedrooms go for approximately 1.6 and 3 million naira, respectively. This strengthens the initial claim that the major feature determining an apartment's rental price is the number of bedrooms and the location in the context of the dataset.

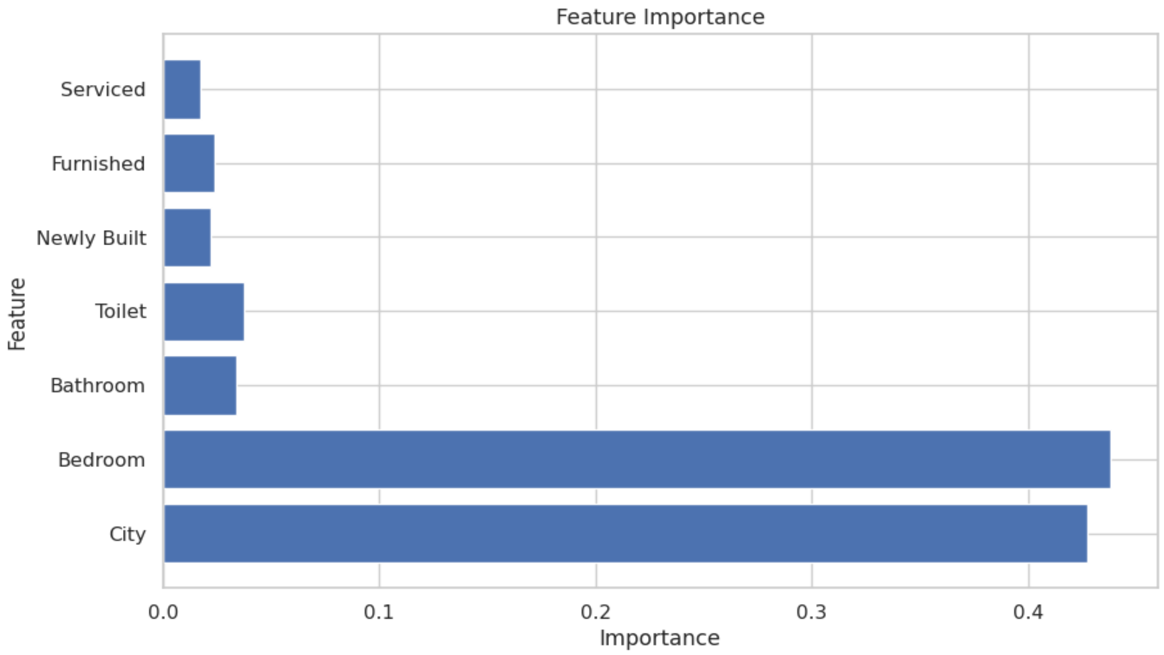
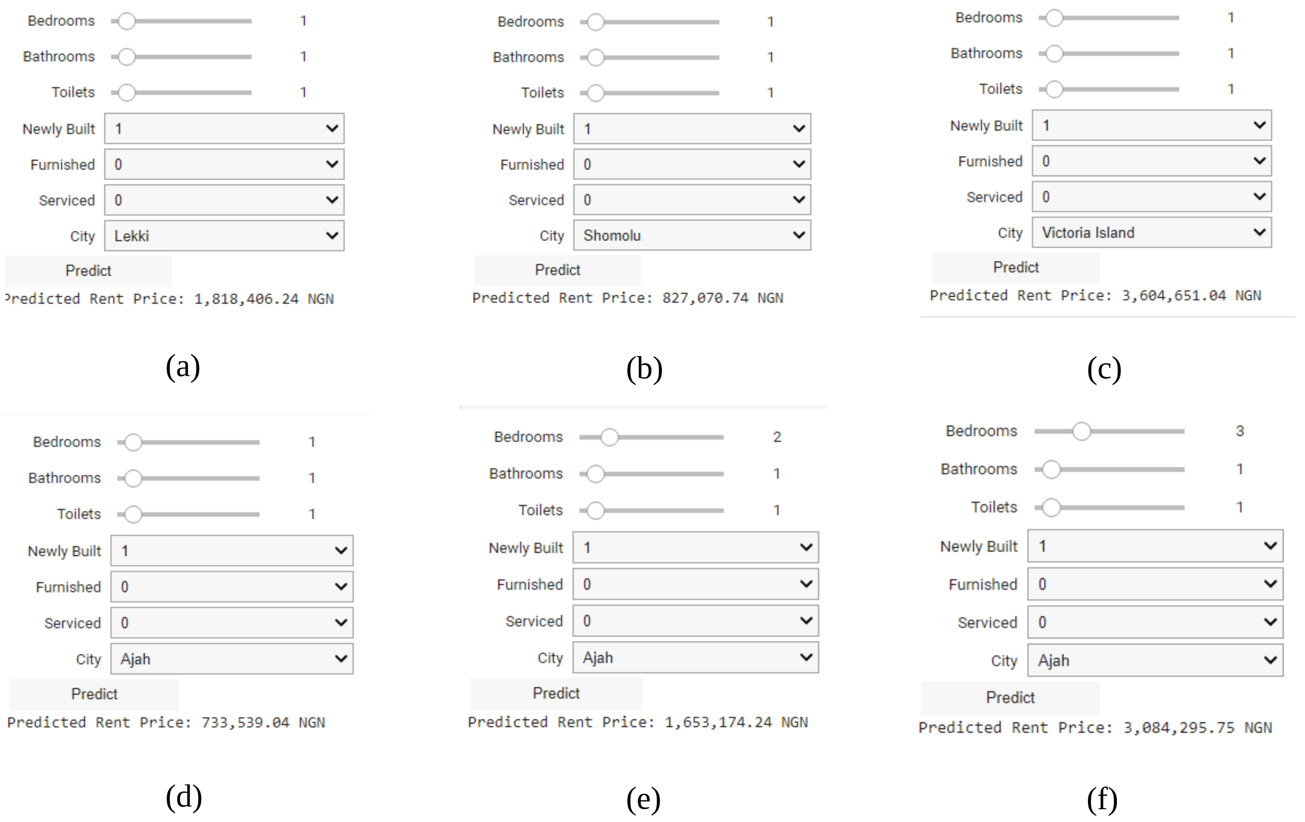


Figure 9: Bar chart showing feature importance

This work was compared with existing studies in this domain, and it was seen that all other works only focus on predicting the price of purchasing an apartment except Oyedeji et al. [14] which builds a classification model for property rental value in Osogbo, Osun State, Nigeria. Oyedeji et al. [14] uses a classification model and the classes were divided into 4 categories which are less than N50,000 p.a.; between N51,000 and N100,000 p.a.; between N101,000 and N200,000 p. a.; and between N201,000 and N300,000 p.a. Also, [14] uses support vector machine (SVM), artificial neural network (ANN), and logistic regression where SVM gave the highest accuracy of 89.7%. Oyedeji et al. [14] does not entirely align with this work because, it employs classification models and also dwells only in Osogbo, Osun State, Nigeria while this work uses regression models (better for price prediction) and dwells in Lagos State, Nigeria. Another work that addressed Nigeria's house price prediction is the work of Nwanko et al. [4]. The Mean Absolute Error was calculated to be 8525486.28, which is higher than the one gotten in this work - 1539994.03; this is because this work focuses primarily on rental price, which will be lower than the price of acquiring a new apartment.



Bedrooms 1
Bathrooms 1
Toilets 1
Newly Built 1
Furnished 0
Serviced 0
City Lekki
Predict
Predicted Rent Price: 1,818,406.24 NGN

(a)

Bedrooms 1
Bathrooms 1
Toilets 1
Newly Built 1
Furnished 0
Serviced 0
City Shomolu
Predict
Predicted Rent Price: 827,070.74 NGN

(b)

Bedrooms 1
Bathrooms 1
Toilets 1
Newly Built 1
Furnished 0
Serviced 0
City Victoria Island
Predict
Predicted Rent Price: 3,604,651.04 NGN

(c)

Bedrooms 1
Bathrooms 1
Toilets 1
Newly Built 1
Furnished 0
Serviced 0
City Ajah
Predict
Predicted Rent Price: 733,539.04 NGN

(d)

Bedrooms 2
Bathrooms 1
Toilets 1
Newly Built 1
Furnished 0
Serviced 0
City Ajah
Predict
Predicted Rent Price: 1,653,174.24 NGN

(e)

Bedrooms 3
Bathrooms 1
Toilets 1
Newly Built 1
Furnished 0
Serviced 0
City Ajah
Predict
Predicted Rent Price: 3,084,295.75 NGN

(f)

Figure 10: Real-time predictions

5. CONCLUSION

This work has successfully addressed the application of machine learning methods in predicting the price of renting an apartment in Lagos, Nigeria. Five models were trained successfully, and it was observed that the Random Forest model emerged as the most effective of the five. This effectiveness was demonstrated by the model exhibiting the lowest mean absolute error, root mean square error and the highest r-squared value. This model also shows a significant improvement when it was compared with the baseline model, which shows that the model will do well when used to predict the real-life rental price of an apartment in Lagos, Nigeria. Additionally, this model reveals that the two most important features that affect the price of an apartment in Lagos, Nigeria, are the location and the number of bedrooms in the context of the dataset.

The developed model serves as a valuable tool for real estate stakeholders, including property owners, prospective tenants, real estate developers, and policymakers who want an estimate of an apartment in a particular region in Lagos, Nigeria.

Future research could focus on refining the model by adding more features such as property age, property type, amenities and proximity to basic amenities to further improve the prediction accuracy. Additionally, more data can be acquired as there is a rapid change in the real estate field in Lagos, Nigeria; as new data are being discovered, they can be added to have a better generalisation capability of the machine learning model for better real-time predictive performance.

REFERENCES

[1] World Bank. *Lagos Diagnostic Study and Pathway for Transformation - A Rapid Multi-Sector Analytical Review of the Mega-City (English).* Washington, D.C.: World Bank Group. http://documents.worldbank.org/curated/en/099062123034023646/P1750310c8d0390000afa70e5c583aa3b87

[2] United Nations Human Settlements Programme (UN-Habitat). (2018). *The State of African Cities 2018: The Geography of African Investment.*

[3] Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science*, 174(1), 433-442.

[4] Nwankwo, M. P., Onyeizu, N. M., Asogwa, E. C., & Ejike, C. O. (2023). Prediction of House Prices in Lagos-Nigeria Using Machine Learning Models.*European Journal of Theoretical and Applied Sciences*, 1(5), 313-326. https://doi.org/10.59324/ejtas.2023.1(5).22 .

[5] Chowhaan, M. J., Nitish, D., Akash, G., Sreevidya, N., & Shaik, S. (2023). Machine Learning Approach for House Price Prediction. *Asian Journal of Research in Computer Science*, 16(2), 54-61.

[6] Rawool, A. G., Rogye, D. V., Rane, S. G., & Bharadi, V. A. (2021). House Price Prediction Using Machine Learning.*IRE Journals*, 4(11), 29-33.

[7] Joshi, H., & Swarndeep, S. (2022). A Comparative Study on House Price Prediction using Machine Learning.*International Research Journal of Engineering and Technology (IRJET)*, 9(11), 782-787.

[8] Aghav, V., Avhad, D., Nanaware, S., Gudekar, R., & Pawar, M. (2023). House Price Prediction Using Machine Learning. *International Research Journal of Modernization in Engineering Technology and Science*, 5(3), 3756-3760. https://doi.org/10.56726/IRJMETS35094 .

[9] Sisman, Y., & Sisman, A. (2016). Principal Component Analysis Approach in the Determination of House Value.*Ponte Multidisciplinary Journal of Sciences and Research*,72(3), 19-28. https://doi.org/10.21506/j.ponte.2016.3.23 .

[10] Mora-Garcia, R. T., Cespedes-Lopez, M.F., & Perez-Sanchez, V.R. (2022). Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times. *Land*, 11(2100), 22-32.

[11] Bali, K. K., & Vashistha, R. (2023, March). House Rent Prediction Using Machine Learning Algorithms – A Methodical Review. *Journal of Emerging Technologies and Innovative Research*, 10(3), 112-116.

[12] Abdulsalam, M. H., Thuraiya, M., Masrom, S., Johari, N., & Mohamad Saraf, M. H. (2022). Machine learning algorithms on price and rent predictions in real estate: A systematic literature review. *Virtual Go Green: Conference and Publication (v-Gogreen 2021) Rethinking Built Environment: Towards a Sustainable Future*, 29-30

[13] Renigier-Bilozor, M., Janowski, A., & Walacik, M. (2019). Geoscience Methods in Real Estate Market Analyses: Subjectivity Decrease. *Geosciences*, 9(3), 130. https://doi.org/10.3390/geosciences9030130 .

[14] Oyedeji, J. O., Oshodi, O. S., & Oloke, O. C. (2018). Property Rental Value Classification Model: A Case of Osogbo, Osun State, Nigeria. *Covenant Journal of Research in the Built Environment (CJRBE)*, 6(1), 52-64.

[15] Mathotaarachchi, K. V., Hasan, R., & Mahmood, S. (2024). Advanced Machine Learning Techniques for Predictive Modeling of Property Prices. *Information*, 15(6), 295. https://doi.org/10.3390/info15060295 .

[16] Kalidass, J., Dharshalini, T., Nivetha, R., & Subasri, A. (2024). House Price Prediction Using Machine Learning. *International Research Journal of Engineering and Technology (IRJET)*, 11(4), 1351-1358.

[17] Dabreo, S., Rodrigues, S., Rodrigues, V., & Shah, P. (2021). Real Estate Price Prediction. I*nternational Journal of Engineering Research & Technology (IJERT)*, 10(4), 644-649.

[18] Mouna, L. E., Silkan, H., Haynf, Y., Nann, M. F., & Tekouabou, S. C. K. (2023). A Comparative Study of Urban House Price Prediction using Machine Learning Algorithms. *E3S Web of Conferences*, 418, 03001. https://doi.org/10.1051/e3sconf/202341803001 .

[19] Choy, L. H. T., & Ho, W. K. O. (2023). The Use of Machine Learning in Real Estate Research. *Land*, 12(4), 740. https://doi.org/10.3390/land12040740 .

[20] Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling* (2nd ed.). Lawrence Erlbaum Associates, Inc.