EPH–International Journal of Science & Engineering
Published by: E-Pub Solutions Private Limited
ISSN (Online): **2454-2016**
Volume 12 Issue 1 April 2026

PREDICTION OF RESIDENTIAL HOUSE PRICES USING MACHINE LEARNING REGRESSION TECHNIQUES

N. B. Karthik Babu*

*Department of Mechanical Engineering, Rajiv Gandhi Institute of Petroleum Technology, Sivasagar, Assam, 785697, India

Article History:

Article Type: **Research**

Received Date: **03/03/2026**

Revised Date: **15/04/2026**

Accepted Date: **22/04/2026**

Published Date: **29/04/2026**

Keywords: House Price Prediction, Machine Learning, Linear Regression, Real Estate Analytics, Regression Models, Data Analysis.

ABSTRACT

The precise forecasting of residential house prices is a significant issue in the real estate industry as it helps in informed decision making by buyers, sellers, and investors. The purpose of the study is to create a machine learning-based model to predict house prices on the basis of a structured housing data. The dataset in this study comprises 545 records with an array of properties, such as area, number of bedrooms, bathrooms, stories, and other housing characteristics. The dataset was prepared by data preprocessing methods like encoding categorical variables and feature selection, to use in modeling. Three machine learning regression algorithms - Linear Regression, Random Forest Regressor and Decision Tree Regressor - were deployed and compared. Evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2) were used to evaluate the performance of these models. Results show that the Linear Regression model performed better than the rest with R^2 of around 0.65 which shows moderate predictive power. The results indicate that machine learning methods can be effective in predicting the prices of residential properties, and it can be used as a valuable tool in the analysis and decision-making of real estate.

1. Introduction

Proper pricing of residential properties is a very important component of the real estate industry since it affects the actions of buyers, sellers, investors, financial institutions and policy-makers. Conventional property valuation techniques can rely on subjective, historical market comparison, and manual evaluation, which can be subjective, time intensive and ineffective when dealing with large datasets. Over the past few years, machine learning started to be an effective method of predicting the price of a house since it could process numerous features related to a house and pinpoint trends based on the previous history (Truong et al., 2020).

The prices of houses depend on various factors such as the area, number of bedrooms, number of bathrooms, number of stories, accessibility of parking space, accessibility of roads, furnishing condition, and other residential amenities. These variables can influence the end price in various manners, so the task of prediction is complicated. This kind of problem can be addressed using machine learning regression models since they can predict continuous numerical values like housing prices, given input features (Adetunji et al., 2022).

Some of the studies have indicated that machine learning methods can be effectively used in real estate price prediction. Such models as Linear Regression, Decision Tree, and Random Forest, and other sophisticated algorithms have been employed to enhance the accuracy of predictions in the housing datasets (Thamarai & Malarvizhi, 2020). On the same note, research conducted by survey has indicated that machine learning models are being increasingly applied in automated house price estimation since it has the capability to deal with structured real estate data (Zulkifley et al., 2020).

The machine learning models have also been previously utilized on city-specific housing data and have shown that data-driven techniques can be useful in helping to value real estate. To illustrate the usefulness of predictive models in urban property markets, (Phan, 2018)used machine learning algorithms to predict housing price using the data of Melbourne housing, demonstrating the usefulness of prediction algorithms in housing markets. (Ho et al., 2021)also highlighted that machine learning algorithms have the potential to enhance price prediction of the property by extracting connections between housing features and the market prices.

Alongside machine learning-based solutions, comparative analyses have been conducted on the traditional ways of valuation of property and artificial intelligence models to enhance the accuracy of the property valuation process. Abidoye and Chan compared hedonic pricing models to artificial neural networks and demonstrated that data-driven models can be used to help in the more accurate valuation of a property. These studies show that machine learning may prove to be a valuable tool when it comes to estimating the price of a residence provided that the right features and assessment techniques are used(Abidoye & Chan, 2018).

The aim of the current research is to create and compare machine learning regression models to predict the price of residential houses with a housing dataset of 545 records. The research uses Linear Regression, Random Forest Regressor, and Decision Tree Regressor in the prediction of the prices of houses using the characteristics of the residential property. Mean Absolute Error, Mean Squared Error, Root Mean Squared Error and R2 score is used to evaluate the performance of the models. This research will attempt to find the most appropriate model to use on the chosen data and illustrate how machine learning methods can be useful in estimating house prices. The purposes of our Study are–

- To develop a machine learning-based model for predicting residential house prices using structured housing data.
- To implement and compare multiple regression techniques, including Linear Regression, Random Forest Regressor, and Decision Tree Regressor.
- To evaluate the performance of these models using statistical metrics (MAE, MSE, RMSE, and R^2) and identify the most effective model for accurate price prediction.

2. Literature Review

Recent research has demonstrated that machine learning tools are becoming more and more valuable in terms of their use in residential property price prediction. (Mora-Garcia et al., 2022) studied housing prices prediction with the help of machine learning algorithms in the COVID-19 era and emphasized the importance of data-driven approaches in the volatile market. Likewise, Foryś contrasted regression models with neural networks to analyze the price of houses and demonstrated that various machine learning models could have varying performance based on the properties of the data(Foryś, 2022).

A number of researchers have used machine learning models on real estate and housing data to enhance prediction accuracy. Singh et al. have also talked about the importance of big data analytics in real estate price forecasting and highlighted the importance of large-scale data processing in estimating property values(Singh et al., 2020). Alfiyatin et al. performed regression analysis and optimization to predict the house prices, showing that regression-based approaches are still applicable in estimating house prices(Alfiyatin et al., 2017). The article by Kutasi and Badics also referred to valuation techniques in the housing market and demonstrated the need to choose the valuation techniques which are applicable to the real estate markets(Kutasi & Badics, 2016).

Various model types and datasets have also been investigated to predict real estate by using machine learning. (Pai & Wang, 2020) utilized machine learning models based on real transaction data in predicting real estate prices, which shows that real market data can enhance realistic predictability. (Jafary et al., 2024)juxtaposed machine and deep learning approaches to automated land valuation and discovered that various factors that influence valuation performance may be used. The predictive models can help assess properties automatically, further supported the use of machine learning in real estate valuation. The recent review and sophisticated modeling research have also suggested the increasing significance of artificial intelligence in real estate forecasting. The article by Tekouabou et al. is a systematic survey of AI-based machine learning approaches to predicting real estate in cities and highlighted the growing importance of intelligent models in property markets(Tekouabou et al., 2024). Rodriguez-Serrano proposed a prototype-based learning model of real estate valuation, with an interest in explainability of price prediction(Rodriguez-Serrano, 2025).

Nevertheless, there are limitations that are also suggested by existing studies. Most of the more sophisticated models are either large data, complicated preprocessing, or computationally intensive, and not always feasible with small structured datasets. (Fotheringham et al., 2015) also observed that the scale, spatial context, and heterogeneity might influence the house price modeling, implying that the performance of the prediction can differ depending on the datasets and locations. Using these past studies, the current study makes a contribution by using simple and understandable machine learning regression models on a structured housing dataset of 545 records. This study, in contrast to other studies where complex models are predominantly studied, compares Linear Regression, Random Forest Regressor, and Decision Tree Regressor based on common evaluation metrics. The aim is to establish a realistic and simple to apply model to the prediction of the residential house prices without ambiguity of interpretation of the findings.

3. Methodology

3.1 Dataset Description

The data employed in this work is a structured residential housing data with 545 records and various attributes of property characteristics (Harish Kumar DataLab, 2023). The dataset has variables like the area, number of bedrooms, number of bathrooms, number of stories, availability of parking and other housing related variables. Also, some categorical characteristics like access to main road, availability of guest rooms, basement, air conditioning and furnishing condition are added.The dependent variable that should be predicted is the house price, which is the target variable of the dataset. The dataset offers an appropriate foundation to use machine learning regression methods because it is a combination of numeric and categorical variables that can be applied in the valuation of real estate.

3.2 Data Preprocessing

Preprocessing of data is a crucial procedure to prepare the dataset to undergo machine learning modeling. In this analysis, missing values were initially checked on the dataset, and none of the missing values were found thus completeness of the data.

Binary encoding was used to encode categorical variables like yes and no, where yes was coded as 1 and no as 0. The furnishing status feature, which has several categories, was also converted with the help of one-hot encoding, which created new binary columns with the various furnishing conditions.

Following preprocessing, everything was transformed into numerical data, which enabled the dataset to be used in regression-based machine learning models.

3.3 Model Development

Three machine learning regression models were used to forecast residential house prices in this research, namely: Linear Regression, Random Forest Regressor and Decision Tree Regressor. Linear Regression was taken as a baseline model because it is simple and has the capability to determine a linear relationship between the independent variables and the target variable. It offers results that can be interpreted and used to comprehend the impact of various characteristics on house prices. Besides this, ensemble and tree-based models were also used to model non-linear relationships in the data. The ensemble technique, a random forest regressor, builds a series of decision trees and combines their results to enhance the prediction and minimize overfitting. The Decision Tree Regressor, however, divides the data into hierarchical decision rules according to feature values, and it is applicable in learning about the patterns of decisions in the data. It was split into training and testing data in the ratio of 80:20 with the models being trained on the training and tested on the testing data.

3.4 Evaluation Metrics

The developed models were evaluated using various statistical evaluation measures to guarantee a thorough evaluation. The average magnitude of the prediction errors without taking into account the direction of the errors was used to evaluate the model accuracy using the Mean Absolute Error (MAE) as a measure of the average. The error was penalized by the Mean Squared Error (MSE) to give more weight to larger errors, and this is sensitive to outliers. Prediction errors were expressed using the square root of MSE (Root Mean Squared Error (RMSE)) to facilitate easier interpretation of the error within the context of house prices. Further, the coefficient of determinants (R2 score) was also taken as a key performance indicator to measure the degree to which the models explain the variation in house prices. The greater the value of R2, the better the performance of the model and the better the prediction of the target variable. All these metrics give a valid foundation to the comparison of the performance of various regression models applied in this study.

4. Results

The results of the applied machine learning models in predicting the price of residential houses are presented in this section. The outputs comprise quantitative analysis based on statistical measures and graphical analysis to gain a better insight into the model performance and the impact of features. The models were evaluated based mainly on the coefficient of determination (R2 score) among other error measures as noted in the methodology. A comparative analysis was conducted to determine the most efficient model to use with the specified data.

The data were split into training and testing data and the three models, Linear Regression, Random Forest Regressor and Decision Tree Regressor, were trained and tested on the same conditions. The comparison of performance shows the predictive power of each model and gives an understanding of whether they are applicable to structured housing data or not.

4.1 Model Performance Comparison

The R^2 scores obtained from the three regression models are presented in Table 1. This comparison provides a clear understanding of how well each model explains the variance in house prices.

Table 1. Performance Comparison of Regression Models Based on R^2 Score

Model	R^2 Score
Linear Regression	0.6529
Random Forest Regressor	0.6114
Decision Tree Regressor	0.4771

Table 1 presents the comparison of machine learning regression models based on their R^2 scores. It shows that Linear Regression achieved the highest performance, followed by Random Forest Regressor and Decision Tree Regressor.

The results indicate that Linear Regression performed the best, achieving an R^2 score of approximately 0.65, which suggests a moderate level of prediction accuracy. Random Forest Regressor also performed reasonably well but did not surpass Linear Regression. The Decision Tree Regressor showed the lowest performance, indicating limited generalization capability on the testing data.

4.2 Graphical Analysis of Model Performance

Graphical representations are used to further analyze model performance and feature relationships within the dataset.

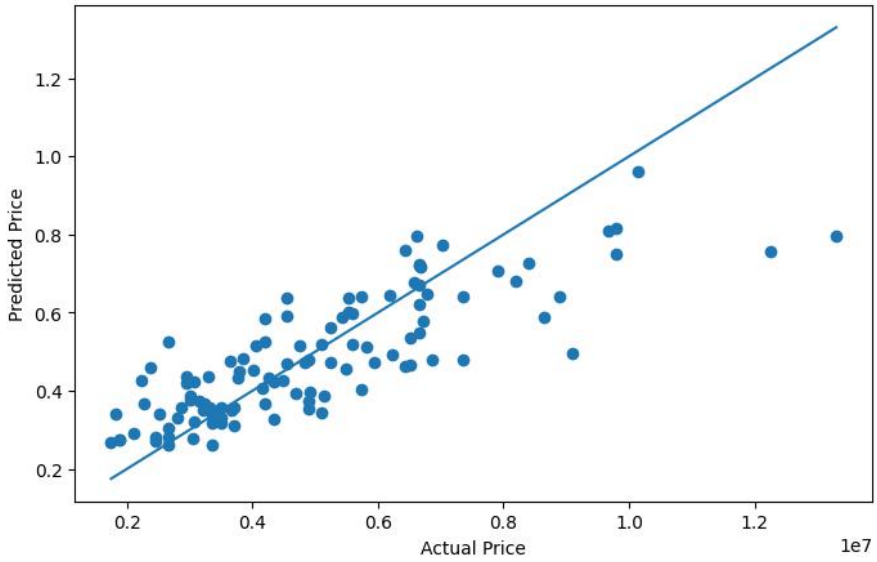

Figure 1. Actual vs Predicted House Prices Using Linear Regression

Figure 1 shows the relationship between actual house prices and predicted values obtained from the Linear Regression model. Data points closer to the diagonal line indicate better prediction accuracy.

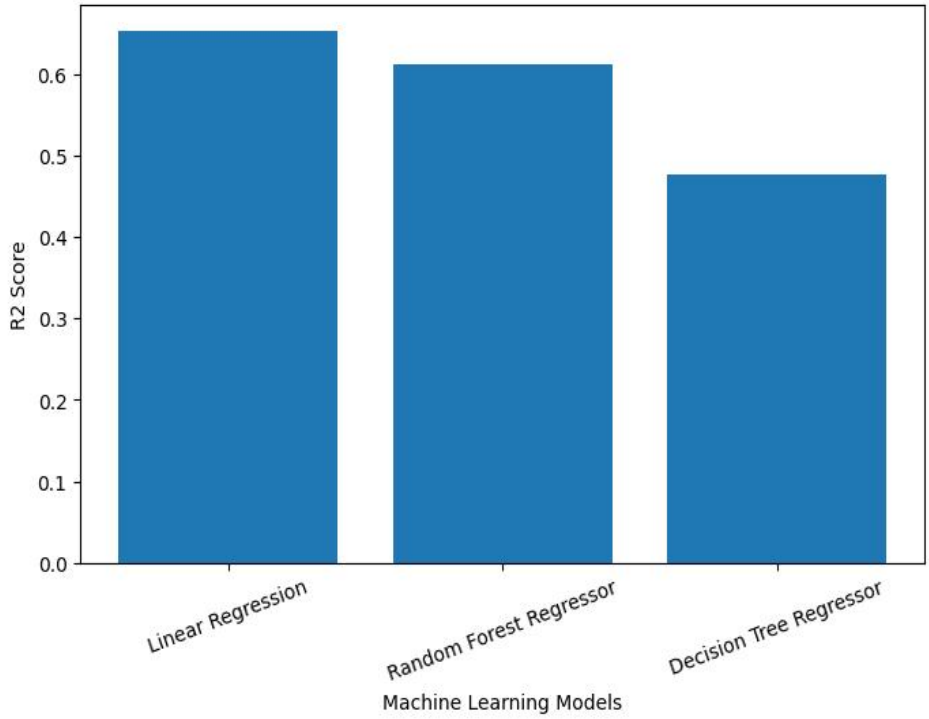

Figure 2: Comparison of Regression Models Based on R² Score

Figure 2 of this bar chart compares the R^2 scores of Linear Regression, Random Forest Regressor, and Decision Tree Regressor. It visually demonstrates that Linear Regression outperforms the other models.

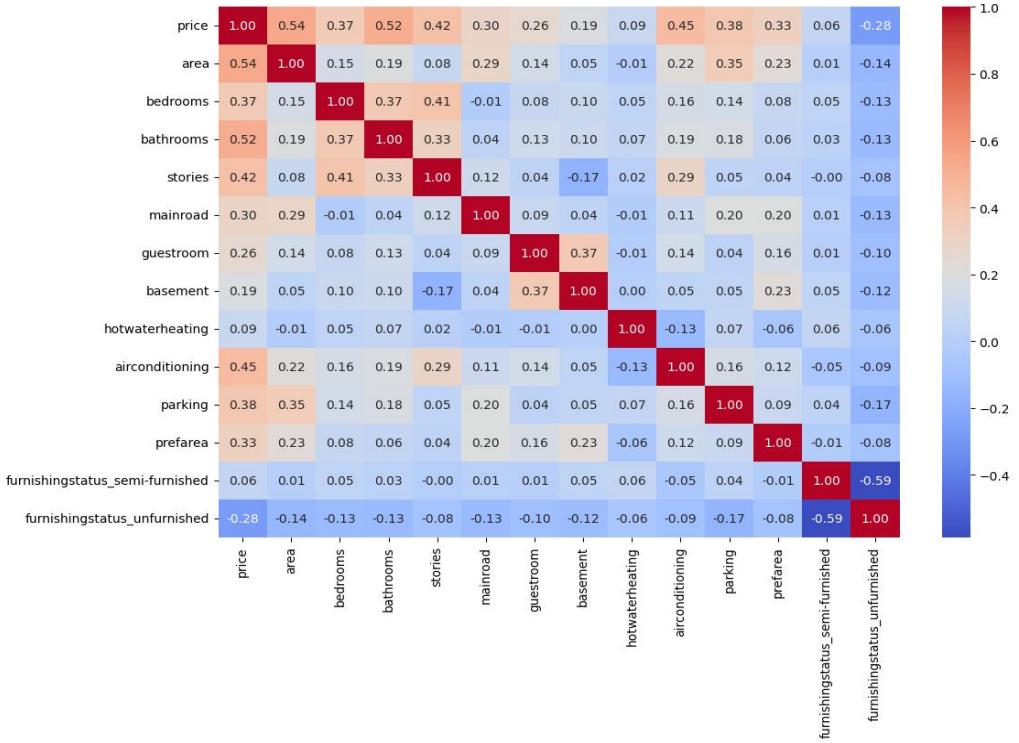

Figure 3: Correlation Heatmap of Housing Dataset

Figure 3 shows heatmap which illustrates the correlation between different features in the dataset. It helps identify variables that have strong positive or negative relationships with house price.

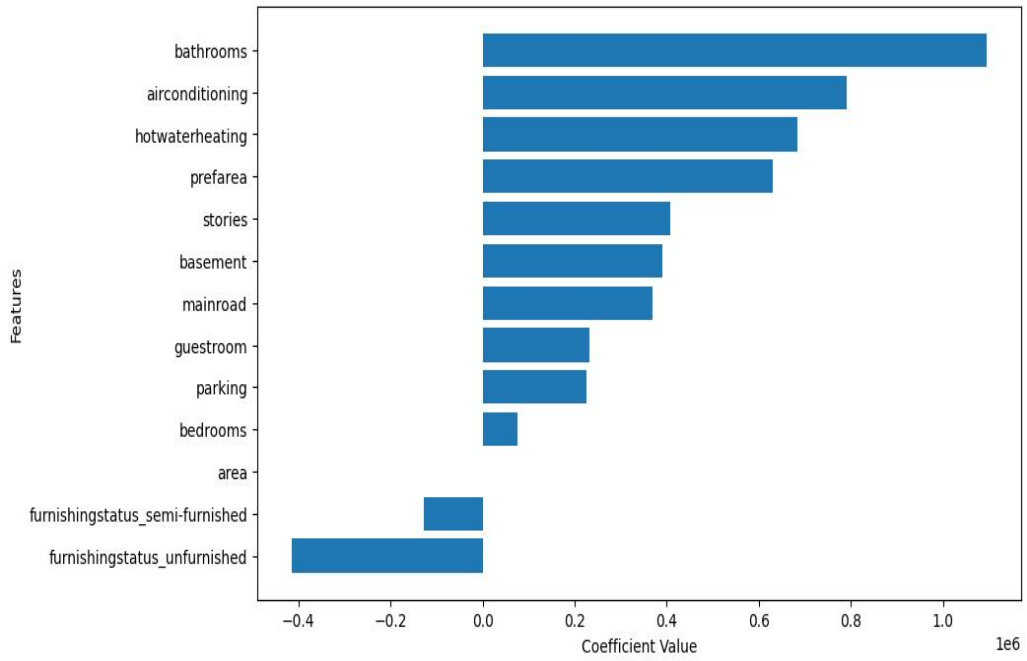

Figure 4: Feature Importance Based on Linear Regression Coefficients

This figure 4 represents the influence of different features on house price prediction based on Linear Regression coefficients. Positive coefficients indicate features that increase house prices, while negative coefficients indicate features that decrease them.

4.3 Summary of Results

The overall results demonstrate that Linear Regression provides the most reliable predictions for the given dataset. With an R^2 score of approximately 0.65, it outperforms the other models while maintaining simplicity and interpretability. The graphical analysis further supports these findings by showing a reasonable alignment between actual and predicted values and highlighting the influence of key features on house prices.

5. Discussion

The findings of this paper suggest that Linear Regression was the best among the models of Random Forest and Decision Tree with the highest R2 of about 0.65. The nature of the dataset could be one of the reasons behind this performance. The housing data involved in this research is relatively small and organized, where most of the relationships among the input variables and the target variable are linear. Linear Regression works well with this type of data as it effectively models any linear relationship without over-fitting. More complex models like Random Forest and Decision Tree, on the other hand, are more apt to work well with large datasets with non-linear relationships, and they do not necessarily generalize well when the dataset is small or the relationships are not that complex (Foryś, 2022; Pai & Wang, 2020).

The feature analysis also offers valuable information on the factors that influence the prices of houses. The findings indicate that the number of bathrooms, air conditioning, and preferred area are the variables that have strong positive effects on house prices. This implies that houses that have superior facilities and good location are likely to be priced higher. Conversely, the unfurnished status is negatively co-efficient meaning that houses that are not furnished have lower prices. These results are aligned with the past research, which has emphasized the role of structural characteristics and amenities in the determination of property value (Ho et al., 2021; Mora-Garcia et al., 2022).

Although the models employed in this study have performed reasonably, the models have some limitations. A significant weakness is the moderate value of R 2 that shows that some part of the variance of house prices cannot be explained by the chosen features. This could be because of the lack of critical variables like precise location coordinates, neighborhood aspects, economic, and market trends. Also, the size of the dataset is

quite limited, which can limit the capabilities of more sophisticated machine learning models to learn more intricate patterns.

In future work, there are a number of improvements that can be made to enhance the predictive performance. To begin with, bigger and more varied datasets that include other features like geographical location, distance to facilities and socio-economic data can enhance the accuracy of models. Second, more sophisticated models like Gradient Boosting, XGBoost, or deep learning approaches can be investigated to learn non-linear relationships better. Lastly, feature engineering and hyperparameter optimization can also be used to further optimize model performance. Such enhancements may be used to create more robust and precise house price prediction models to be applied in real life.

6. Conclusion

This paper was aimed at estimating the prices of residential houses based on machine learning regression models applied to a well-structured housing dataset. Three models were used, including Linear Regression, Random Forest Regressor, and Decision Tree Regressor, which were tested and measured by the standard performance metrics. These findings revealed that Linear Regression performed the best with a score of about 0.65 in the R2, which means that the model is effective in encompassing the relationship between housing features and price in the dataset used. It was also found that the number of bathrooms, availability of air conditioning and preferred area are some of the features that have a strong positive effect on the prices of houses whereas unfurnished properties have a negative effect. These results indicate that residential qualities are important factors in estimating prices. The model that has been developed could be applicable in real-life situations like property evaluation, real estate analysis, and buyer and seller decisions. To make future work, it is possible to use larger datasets, add more features, and apply more advanced machine learning models to enhance the accuracy of predictions and make the models more robust.

References

1. Abidoye, R. B., & Chan, A. P. C. (2018). Improving property valuation accuracy: A comparison of hedonic pricing model and artificial neural network. *Pacific Rim Property Research Journal*, *24*(1), 71–83. https://doi.org/10.1080/14445921.2018.1436306
2. Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House Price Prediction using Random Forest Machine Learning Technique. *Procedia Computer Science, The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021): Developing Global Digital Economy after COVID-19*, *199*, 806–813. https://doi.org/10.1016/j.procs.2022.01.100
3. Alfiyatin, A. N., Febrita, R. E., Taufiq, H., & Mahmudy, W. F. (2017). Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study: Malang, East Java, Indonesia. *International Journal of Advanced Computer Science and Applications (Ijacsa)*, *8*(10). https://doi.org/10.14569/IJACSA.2017.081042
4. Foryś, I. (2022). Machine learning in house price analysis: Regression models versus neural networks. *Procedia Computer Science, Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 26th International Conference KES2022*, *207*, 435–445. https://doi.org/10.1016/j.procs.2022.09.078

5. Fotheringham, A. S., Crespo, R., & Yao, J. (2015). Exploring, modelling and predicting spatiotemporal variations in house prices. *The Annals of Regional Science*, *54*(2), 417–436. https://doi.org/10.1007/s00168-015-0660-6
6. Harish Kumar DataLab. (2023). *Housing price prediction*. https://www.kaggle.com/datasets/harishkumardatalab/housing-price-prediction
7. Ho, W. K. O., Tang, B.-S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, *38*(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558
8. Jafary, P., Shojaei, D., Rajabifard, A., & Ngo, T. (2024). Automated land valuation models: A comparative study of four machine learning and deep learning methods based on a comprehensive range of influential factors. *Cities*, *151*, 105115. https://doi.org/10.1016/j.cities.2024.105115
9. Kutasi, D., & Badics, M. C. (2016). *Valuation methods for the housing market: Evidence from Budapest*. https://doi.org/10.1556/032.2016.66.3.8
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*(11), 2100. https://doi.org/10.3390/land11112100
11. Pai, P.-F., & Wang, W.-C. (2020). Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices. *Applied Sciences*, *10*(17), 5832. https://doi.org/10.3390/app10175832
12. Phan, T. D. (2018). Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*, 35–42. https://doi.org/10.1109/iCMLDE.2018.00017
13. Rodriguez-Serrano, J. A. (2025). Prototype-based learning for real estate valuation: A machine learning model that explains prices. *Annals of Operations Research*, *344*(1), 287–311. https://doi.org/10.1007/s10479-024-06273-1
14. Singh, A., Sharma, A., & Dubey, G. (2020). Big data analytics predicting real estate prices. *International Journal of System Assurance Engineering and Management*, *11*(2), 208–219. https://doi.org/10.1007/s13198-020-00946-3
15. Tekouabou, S. C. K., Gherghina, Ş. C., Kameni, E. D., Filali, Y., & Idrissi Gartoumi, K. (2024). AI-Based on Machine Learning Methods for Urban Real Estate Prediction: A Systematic Survey. *Archives of Computational Methods in Engineering*, *31*(2), 1079–1095. https://doi.org/10.1007/s11831-023-10010-5

16. Thamarai, M., & Malarvizhi, S. P. (2020). House Price Prediction Modeling Using Machine Learning. *International Journal of Information Engineering and Electronic Business*, *12*(2), 15–20. https://doi.org/10.5815/ijieeb.2020.02.03
17. Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science, 2019 International Conference on Identification, Information and Knowledge in the Internet of Things*, *174*, 433–442. https://doi.org/10.1016/j.procs.2020.06.111
18. Zulkifley, N. H., Rahman, S. A., Ubaidullah, N. H., & Ibrahim, I. (2020). House Price Prediction using a Machine Learning Model: A Survey of Literature. *International Journal of Modern Education and Computer Science*, *12*(6), 46.