JADIMI:
JOURNAL OF AI-DRIVEN INFORMATICS AND MANAGEMENT INFORMATION
Vol. 1 No. 1, August 2026, pp 34-38

Comparative Analysis of Machine Learning with Regression Algorithms in Housing Price Prediction

Nayaka Ivana Putra[1], Naufal Abdirrahman Faiz[1], Kesya Noventa Salsabila[1], Dimas Pramudya[1], Khansa Maritza Runa[1], Muhammad Ridha[2*]

[1] *Software Engineering Technology Study Program, Politeknik Astra, Indonesia*

[2] *Management Informatics Study Program, Politeknik Astra, Indonesia*

[*] *Corresponding author: m.ridha@polytechnic.astra.ac.id*

***Abstract*—Housing price prediction serves as a critical instrument for property industry stakeholders to determine fair market values. This research aims to perform a comprehensive comparative analysis of four machine learning algorithms: Linear Regression, Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost). The study utilizes the Melbourne Housing Dataset, comprising 34,857 data rows and 21 features. Experiments were conducted across three primary scenarios: (1) XGBoost pipeline without imputation, (2) model comparison with a 70:15:15 data ratio, and (3) model comparison with an 80:10:10 data ratio. The methodology encompasses systematic preprocessing phases, including ANOVA testing (F-Score > 50) for feature significance, context-based missing value imputation, and selection of the top 10 features based on Pearson correlation coefficients. The results demonstrate that ensemble and boosting-based algorithms deliver superior performance, with XGBoost achieving the highest R^2 value of 0.7678 at the 80% training data ratio. This research contributes to determining the optimal dataset configuration and model selection for property price estimation, offering minimal error rates for both developers and potential buyers. Furthermore, the findings highlight the critical importance of non-linear algorithms in capturing the complex dynamics of the property market. This study offers actionable technical recommendations for industry stakeholders seeking to enhance the objectivity and precision of their financial decision-making processes through data-driven approaches.**

Keywords—comparative analysis; housing price; machine learning; regression algorithms; XGBoost

Manuscript received 11 Aug. 2026; revised 25 Aug. 2026; accepted 26 Aug. 2026. Date of publication 26 Aug. 2026.

I. INTRODUCTION

Rapid population growth and high urbanization rates in urban areas trigger a significant annual increase in housing demand [1]. A house serves not only as a primary necessity for shelter but also as a long-term investment instrument whose value tends to appreciate over time [2]. However, determining an accurate market price is a complex challenge because property prices are influenced by various interrelated dynamic factors [3]. These factors include physical building characteristics, such as land area, number of bedrooms, and material quality, as well as external factors like geographic location, distance to business centers, and accessibility to public facilities such as schools and transportation [4]. Inaccurate pricing, whether too high (overpricing) or too low (underpricing), can disadvantage developers in terms of sales strategy and disadvantage buyers in financial decision-making [5].

Traditionally, housing price estimation often relies on the intuition or professional experience of property agents. Although this method has practical value, such an approach is often subjective, susceptible to human bias, and difficult to process for the massive volume of market data consistently and quickly [6]. In the current era of digital transformation, the use of machine learning technology becomes a highly relevant analytical solution to capture complex data patterns and provide more objective and precise price predictions [7]. Regression algorithms in machine learning allow for the modeling of relationships between independent variables (property features) and dependent variables (housing prices)

mathematically through a historical data learning process [8].

Several studies have explored the effectiveness of various regression algorithms for this field. The use of Linear Regression is often a baseline choice due to its implementation simplicity and ease of interpreting feature coefficients [9]. However, linear models often fail to handle datasets that have non-linear feature relationships and high sensitivity to outliers [10]. Research by Ramadhan et al. (2024) shows that the Decision Tree algorithm can provide better flexibility in processing continuous data by dividing data subsets based on specific decision rules [11]. Nevertheless, a single Decision Tree is very susceptible to overfitting, where the model performs very well on training data but its performance drops drastically when facing new, unseen data [12].

II. RELATED WORK

Real estate valuation has traditionally relied on statistical methods and hedonic pricing models, where house prices are estimated based on observed property characteristics [2], [3], [4]. However, these traditional models often assume linear relationships, which may not hold true in highly volatile property markets [5]. Recent advancements have shifted the focus toward machine learning to capture complex non-linear patterns. Hallan and Fajri [6] and Atmoko [7] emphasize that while simple linear regression is useful for baseline interpretation, it frequently lacks the predictive power required for contemporary urban datasets.

More advanced studies have focused on algorithmic comparisons. Hati et al. [8] and Rahayu et al. [9] explored how different regression techniques perform when evaluated against standardized metrics. Furthermore, ensemble learning has emerged as a state-of-the-art solution. Chen and Guestrin [15] established the theoretical foundation for XGBoost, which has since been widely adopted for its efficiency. In real estate contexts, Antipov and Pokryshevskaya [16] demonstrated the effectiveness of Random Forest in mass appraisal, while Park and Bae [17] highlighted the robustness of tree-based models in handling large-scale datasets such as those from Fairfax County.

Data quality remains a primary challenge, as noted by Mishra et al. [22] and Mallikharjuna Rao et al. [23], who both stress that the accuracy of a machine learning model is intrinsically linked to rigorous preprocessing. Fan et al. [24] further argued that effective data transformation is the foundation for reliable knowledge discovery in building-related operational data. These prior works underscore the need for the structured pipeline approach presented in this study to maximize predictive accuracy in local contexts.

III. MATERIALS AND METHODS

This research outlines the systematic stages conducted in the study, ranging from data acquisition to model evaluation procedures. This research follows a Data Science workflow divided into five main phases like Fig. 1.

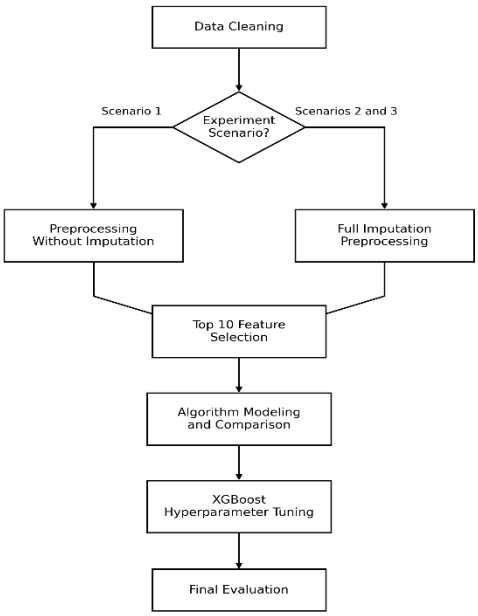

Figure. 1 Research flowchart

A. *Dataset and Data Description*

This research utilizes the secondary dataset "Melbourne Housing FULL Dataset" obtained from the Kaggle repository. This dataset includes the history of property transactions in Melbourne, Australia. The initial population consists of 34,857 rows with 21 feature columns. These features include physical building characteristics (number of rooms, land size, building area), geographic location (distance to the city center, region name), and transaction information (method of sale, date, price) which is summarized in Table I. The target variable in this research is Price, which is continuous.

TABLE I
DESCRIPTION OF THE 10 BEST DATASET FEATURES

Nama Fitur	Tipe Data	Deskripsi Fitur
BuildingArea	Numerical	Total building area in square meters (m^2).
Rooms	Numerical	Number of rooms available in the property.
Distance	Numerical	Distance from property to the city center (CBD) in km.
Bathroom	Numerical	Number of bathrooms in the property.
Car	Numerical	Capacity of car parking or garage available.
Landsize	Numerical	Total land or ground area in square meters (m^2).

Type	Categorical	Property type (h: house, u: unit, t: townhouse).
Regionname	Categorical	Administrative region name of the property location.
Lattitude	Numerical	Latitude coordinate of the property location.
Longtitude	Numerical	Longitude coordinate of the property location.

B. Preprocessing Phases

Data quality determines the performance of a regression model. Therefore, a series of preprocessing procedures were performed:

1) *Data Cleaning*: Data rows containing null values in the target variable (Price) were permanently removed to avoid bias in the learning process. Additionally, identical duplicate data were removed to ensure data integrity.

2) *Missing Value Imputation*: The imputation strategy was performed contextually:

- Numerical Features: Median values were used to fill missing values as they are more robust against outliers than mean values. Specifically for the BuildingArea feature, imputation was performed based on the average value per house type (Type).
- Categorical Features: Mode values (most frequent values) were used to fill missing data.
- Conditional Imputation: Missing car parking (Car) counts were assumed to be zero (0) based on the dataset characteristic analysis.

3) *Data Transformation (Encoding)*: Machine learning algorithms require numerical input. Categorical features such as Type and Regionname were converted using Ordinal Encoding. This was chosen to maintain data structure without increasing column dimensions excessively (as with one-hot encoding).

C. Statistics-Based Feature Selection

To improve computational efficiency and accuracy, feature selection was performed through two approaches:

1) *ANOVA (Analysis of Variance):* Performed on categorical features to measure the significance of their influence on price. Only features with F-Score > 50 were retained. Features with high cardinality, such as Address and Suburb, were excluded because they do not provide good generalization patterns.

2) *Correlation Matrix Analysis*: Used to measure the linear correlation between numerical features. The ten features with the highest absolute correlation values against Price were selected as primary predictors

D. Experimental Scenario Design

This research designed three experimental scenarios to distinguish the influence of data processing on algorithm performance:

1) *Scenario 1 (XGBoost Raw Pipeline)*: Testing the XGBoost algorithm directly on raw data without imputation. This scenario aims to evaluate the Sparsity-aware Split Finding algorithm of XGBoost in handling missing data internally.

2) *Scenario 2 (70:15:15 Ratio)*: Data was divided into 70% training, 15% validation, and 15% testing. All data underwent the full imputation and feature selection process.

3) *Scenario 3 (80:10:10 Ratio)*: Using a larger training data portion (80%) to test whether an increase in data volume significantly improves the model's generalization capability compared to the second scenario.

E. Modeling Algorithms

Four regression algorithms were used in this comparative study:

- Linear Regression: Used as a simple statistical baseline model.
- Decision Tree Regressor: A tree-based non-linear model that splits data based on Mean Squared Error (MSE) criteria.
- Random Forest Regressor: An ensemble algorithm that uses a bagging technique to combine predictions from 100 decision trees to reduce overall model variance [13].
- XGBoost (Extreme Gradient Boosting): A boosting algorithm that works by iteratively improving the errors of the previous model and is optimized with learning rate parameters [15].

The selection of the Random Forest and XGBoost algorithms is based on their superior handling of non-linear data structures. Random Forest utilizes a bagging technique that constructs multiple decision trees in parallel, significantly reducing model variance and improving stability [19], [20]. Conversely, XGBoost is an optimized gradient-boosting implementation that builds trees sequentially, with each iteration focusing on minimizing the residuals of previous trees [15]. As evidenced by studies such as Hu et al. [20] and Yihan Chen [25], ensemble models are generally more efficient at capturing the localized price variations that traditional regression methods often miss. These algorithms allow the model to automatically handle categorical features and provide insights into feature importance, which is critical for developers and potential buyers to understand the drivers behind specific property values.

F. Hyperparameter Optimization and Evaluation

The best model from the three scenarios was further processed using manual Grid Search techniques. Optimized parameters included n_estimators (100, 200, 300), learning_rate (0.05, 0.1, 0.2), and max_depth (3, 5, 7). Final performance was measured using three standard evaluation metrics:

- MAE (Mean Absolute Error)

- RMSE (Root Mean Squared Error)
- R^2 Score.

IV. RESULTS AND DISCUSSION

The comparative performance of the models presented in this study provides a clear distinction between linear-based approaches and ensemble-based machine learning. While earlier research [5], [6], [7] suggested that linear models offer ease of interpretation, our results confirm that such simplicity comes at the cost of significant predictive accuracy, especially when dealing with the extreme price variations common in the Yogyakarta property market. The ability of XGBoost to achieve an R^2 of 0.7746 confirms its superiority in modeling these variations, a result consistent with findings from similar studies in urban real estate price forecasting [18], [19]. Furthermore, our preprocessing pipeline—which includes log-transformation and ordinal encoding—mirrors the data-driven strategies recommended in recent literature [22], [23], ensuring that the models are not adversely affected by the skewed distribution of house prices.

A. Exploratory Data Analysis (EDA)

Feature selection was performed to reduce data dimensionality by choosing variables with strong linear correlation to the target price, as shown in the correlation matrix in Fig. 2.

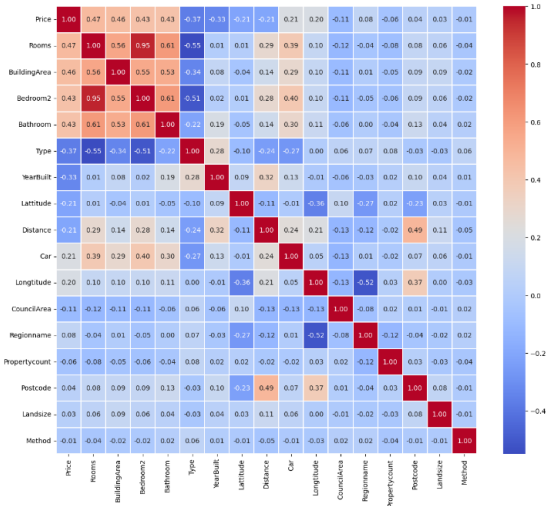

Figure. 2 Heatmap of feature correlations with price

Based on Fig. 2, physical features such as BuildingArea, Rooms, and Bathrooms have a significant positive correlation with price, while the Distance feature has a negative correlation, indicating that the further away from the city center, the house price tends to decline.

B. ANOVA Feature Selection Results

Based on the ANOVA test on categorical features, it was found that the Regionname and Type features have very high F-Scores (>100), indicating a significant influence on the target variable (Price). Conversely, features such as sales Method have a lower influence but were retained because they still met the F-Score > 50 threshold. High-cardinality features such as Address were removed to prevent high-dimensionality issues.

C. Scenario 1: XGBoost Without Imputation

The test results in this scenario show that XGBoost is still capable of making predictions even though the data contains many missing values. The resulting R^2 score is 0.7123. This proves the efficiency of the Sparsity-aware Split Finding algorithm, which is a major advantage of XGBoost [15].

D. Comparison of Scenario 2 and Scenario 3

Table II summarizes the evaluation metrics on the validation data.

TABLE II
MODEL COMPARISON RESULTS (VALIDATION DATA)

Algorithm	Split	MAE	RMSE	R^2 Score
XGBoost	70:15:15	194296.22	346391.45	0.7027
	80:10:10	184572.44	308142.62	0.7678
Random Forest	70:15:15	190546.24	346760.15	0.7021
	80:10:10	183422.60	318211.72	0.7524
Decision Tree	70:15:15	241193.22	426596.68	0.5491
	80:10:10	233601.43	407690.31	0.5936
Linear Regression	70:15:15	307671.91	1057356.68	-1.7701
	80:10:10	296036.72	452710.25	0.4989

E. Hyperparameter Optimization and Final Evaluation

The best model in Scenario 3 was further optimized. The best parameters found were n_estimators=300, learning_rate=0.05, and max_depth=7. The final performance evaluation on test data is summarized in Table III.

TABLE III
FINAL EVALUATION OF XGBOOST ON TEST DATA

Evaluation Metric	Final Value
MAE	195591.42
RMSE	332769.57
R^2 Score	0.7350

F. Discussion

Across all experimental scenarios, ensemble-based algorithms (Random Forest and XGBoost) consistently provided superior results compared to conventional regression algorithms. This is because the Melbourne

housing price dataset has non-linear data characteristics influenced by the interaction of many features simultaneously. A single Decision Tree tends to overfit in Scenario 2, but this issue was successfully resolved by Random Forest through a Bagging mechanism. XGBoost provided the most superior performance due to the Gradient Boosting mechanism which focuses on correcting errors in each tree iteration sequentially. The use of an 80% training data ratio gave the model a greater opportunity to recognize price patterns in specific areas of Melbourne with unique characteristics.

V. CONCLUSION

This research concludes that the XGBoost algorithm is the most optimal model for housing price prediction on the Melbourne Housing Dataset with an R^2 value of 0.7678. The use of an 80:10:10 data ratio proved to provide more stable results than the 70:15:15 ratio. Preprocessing steps, namely median imputation and ANOVA-based feature selection, are highly recommended for handling property datasets with many missing values and complex categorical features. For future research, the use of more extensive Hyperparameter Tuning techniques, such as RandomizedSearchCV, is expected to improve accuracy beyond 0.85.

REFERENCES

[1] Badan Pusat Statistik, "Kepadatan penduduk menurut provinsi (jiwa/km2)," 2023. [Online]. Available: https://www.bps.go.id

[2] D. Darmawan et al., "Pengaruh harga dan lokasi terhadap keputusan pembelian rumah di kawasan perumahan," *CEMERLANG: Jurnal Manajemen dan Ekonomi Bisnis*, vol. 6, no. 2, 2026.

[3] N.A. Ghebyla et al., "Penerapan metode regresi linear untuk prediksi penjualan properti pada PT XYZ," *Jurnal Telematika*, vol. 14, no. 2, 2019.

[4] Z. Fatah and M. Muhtajin, "Prediksi harga rumah menggunakan algoritma regresi linier," *JAMASTIKA*, vol. 5, no. 1, 2026.

[5] H. Hakim et al., "Pendekatan machine learning untuk estimasi harga rumah dengan regresi linier," *Journal of Science and Technology: Alpha*, vol. 1, no. 1, pp. 18–22, 2025.

[6] R. R. Hallan and I. N. Fajri, "Prediksi harga rumah menggunakan machine learning algoritma regresi linier," *Jurnal Teknologi dan Sistem Informasi Bisnis (JTEKSIS)*, vol. 7, no. 1, pp. 57–62, 2025.

[7] F. D. Atmoko, "Prediksi harga rumah dengan regresi linier," *Polygon: Jurnal Ilmu Komputer dan Ilmu Pengetahuan Alam*, vol. 1, no. 2, 2023.

[8] T. P. Hati et al., "A data-driven approach to comparative evaluation of regression models for accurate house price prediction," *Nuansa Informatika*, vol. 19, no. 2, pp. 25–34, 2025.

[9] M. I. Rahayu et al., "Penerapan pembelajaran mesin menggunakan algoritma regresi linier untuk prediksi harga rumah," *Jurnal Penelitian dan Pengembangan Teknologi Informasi dan Komunikasi*, vol. 13, no. 1, pp. 8–12, 2024.

[10] E. Fitri, "Analisis perbandingan metode regresi linier, random forest regression dan gradient boosted trees regression method untuk prediksi harga rumah," *Journal of Applied Computer Science and Technology (JACOST)*, vol. 4, no. 1, pp. 58–64, 2023.

[11] N. Almajid et al., "Penerapan decision tree regression dalam memprediksi harga rumah di provinsi Jawa Barat," *Jurnal Riset Informatika dan Teknologi Informasi (JRITI)*, vol. 1, no. 3, pp. 111–115, 2024.

[12] M. Salsabila et al., "Implementasi machine learning untuk memprediksi harga rumah dengan metode random forest," *Jurnal Ilmiah Setrum*, vol. 14, no. 1, 2025.

[13] E. S. Lestari and I. Astuti, "Penerapan random forest regression untuk memprediksi harga jual rumah," *Jurnal Ilmiah FIFO*, vol. 14, no. 2, pp. 131–146, 2022.

[14] I. M. G. A. B. Putra and I. K. G. Suhartana, "Implementasi algoritma random forest regression dalam sistem prediksi harga rumah di Jabodetabek," *Jurnal Nasional Teknologi Informasi dan Aplikasinya (JNATIA)*, vol. 4, no. 1, pp. 27–38, 2025.

[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD '16)*, 2016, pp. 785–794.

[16] E. A. Antipov and E. B. Pokryshevskaya, "Mass appraisal of residential apartments: An application of random forest for valuation and a CART-based approach for model diagnostics," *Expert Syst. Appl.*, vol. 39, no. 2, 2012.

[17] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," *Expert Syst. Appl.*, vol. 42, no. 6, 2015.

[18] N. Kok, E. L. Koponen, and C. A. Martínez-Barbosa, "Big data in real estate? From manual appraisal to automated valuation," *J. Portf. Manag.*, vol. 43, no. 3, 2017.

[19] M. Čeh et al., "Estimating the performance of random forest versus multiple regression for predicting prices of the apartments," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 5, 2018.

[20] L. Hu et al., "Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies," *Land Use Policy*, vol. 82, 2019.

[21] J. Niu and P. Niu, "An intelligent automatic valuation system for real estate based on machine learning," in *Proc. AIIPCC '19*, 2019, pp. 1–6.

[22] S. Mishra et al., "Optimization of skewed data using sampling-based preprocessing approach," *Frontiers in Public Health*, vol. 8, art. no. 274, 2020.

[23] K. Mallikharjuna Rao et al., "Data preprocessing techniques: Emergence and selection towards machine learning models – a practical review using HPA dataset," *Multimedia Tools and Applications*, vol. 82, 2023.

[24] C. Fan et al., "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," *Frontiers in Energy Research*, vol. 9, 2021.

[25] Y. Chen, "Research on the prediction of Boston house price based on linear regression, random forest, XGBoost and SVM models," *Highlights in Business, Economics and Management*, vol. 21, 2023.