NUANSA INFORMATIKA **p-ISSN :1858-3911 , e-ISSN : 2614-5405**
Volume 19 Number 2 July 2025 https://journal.fkom.uniku.ac.id/

A Data-Driven Approach to Comparative Evaluation of Regression Models for Accurate House Price Prediction

Tiara Permata Hati[1], Budiman*[2], Imannudin Akbar[3], Nur Alamsyah[4]
[1,2,3,4]Universitas Informatika dan Bisnis Indonesia, Indonesia
E-mail: [1]tiara.p21@student.unibi.ac.id, *[2]**budiman@unibi.ac.id**, [3]imannudin@unibi.ac.id
[4]nuralamsyah@unibi.ac.id

Abstract

This study aims to develop and evaluate a property price prediction model in Bandung by applying machine learning (ML) algorithms. The need for more accurate property price predictions is increasing due to fluctuations in the property market. This study analyzes property characteristics, including the number of bedrooms, bathrooms, land area, building area, and location, as well as their impact on house prices. The study evaluates four regression algorithms, including linear regression, K-Nearest Neighbors (KNN), Random Forest, and XGBoost. Finally, this study proposes price_per_m2 and building_land_ratio as new features recommended for improvement in accuracy. The bottleneck method is derived from the data collection area of the Rumah123.com website, encompassing data preprocessing and data exploration. The following metrics will be used to evaluate each model: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²). Based on our study, we conclude that both Random Forest Regression and XGBoost Regression achieve the highest accuracy, with R² values of 0.9941 and 0.9955, respectively, after adjustment. Conversely, Linear Regression and KNN Regression have the lowest accuracy, with KNN Regression being the least accurate. The primary contribution of this study is the development of a more accurate house price prediction model that can be applied in cities with similar market characteristics. These findings provide practical insights for property developers and buyers when making investment decisions.

Keywords— *house price prediction, machine learning, feature engineering, random forest* regression, *XGBoost* regression

Submission: June 05, 2025 *Accepted: June 20, 2025* *Published: July 10, 2025*

DOI Article: https://doi.org/10.25134/ilkom.v19i2.411

1. INTRODUCTION

In recent years, machine learning (ML) and artificial intelligence (AI) have emerged in the real estate sector, advancing predictive capabilities for property price estimations and valuations. Machine learning (ML) and artificial intelligence (AI) can enhance data-informed decision-making, which is crucial for stakeholders navigating shifting market conditions. Numerous studies have examined the success of various machine learning (ML) algorithms, including decision trees, random forests, and ensemble methods, in predicting housing prices more accurately and efficiently [1], [2], [3]. Random forests and gradient boosting models, which are more sophisticated than traditional methods, have been shown in numerous studies to outperform older methods in many situations due to the complex interactions of input features and their non-linear relationships in real estate data [4], [5].

Increasingly, machine learning models are being used globally, including India, the United States, and Saudi Arabia, to predict property prices and support real estate investment decisions [6], [7]. The use of techniques such as K-Nearest Neighbors (KNN), support vector regression (SVR), and linear regression all provide additional predictive value and accuracy dependent on various attributes of the dataset [8], [9], [10], [11]. Likewise, ensemble methods such as XGBoost and LightGBM outperform maximum simpler models like linear regression as a result of their ability to create complex feature interactions and great flexibility concerning non-linearity.

Multiple authors have highlighted the relevance of model selection and Feature Engineering use when the goal is to maximize prediction accuracy. Regression models and

p-ISSN :1858-3911 , e-ISSN : 2614-5405
https://journal.fkom.uniku.ac.id/

decision trees have been tested from a variety of perspectives to predict house prices in multiple cities, and it was found that model choice is significant in improving predictive outcomes [12], [13], [14]. This research aims to analyze and compare several machine learning (ML) models — specifically, decision trees, random forests, and linear regression — in predicting housing prices across multiple regions. The comparative outcomes could provide recommendations to a wide range of real estate stakeholders seeking to optimize their property price forecasting methodologies and continue advancing the field of real estate analytics [15], [16].

This study aims to design and evaluate a house price prediction model for Bandung based on property characteristics such as the number of bedrooms and bathrooms, land area, building area, carports, and location. The study aims to identify the primary predictors of house prices in Bandung and apply feature construction principles to enhance model performance. Additionally, the study compares the predictive performance of several regression models—linear regression, K-Nearest Neighbors, Random Forest, and XGBoost—to predict house prices. It optimizes model performance through tuning and cross-validation to evaluate the model's generalization performance on unseen data.

The impact of this research in the area of property pricing prediction through a data-driven approach is manifold. One of those contributions is the introduction of Feature Engineering techniques, which create new variables such as price_per_m2, building_land_ratio, and total_rooms. Including these variables can positively impact model performance on non-normally distributed data. Furthermore, this research compares four standard regression algorithms used for house price prediction and presents their advantages and disadvantages based on extensive evaluation metrics, e.g., MAE, MSE, RMSE, and R^2.

2. METHODS

Figure 1 illustrates the proposed method employed in this study, spanning from data collection to model evaluation.

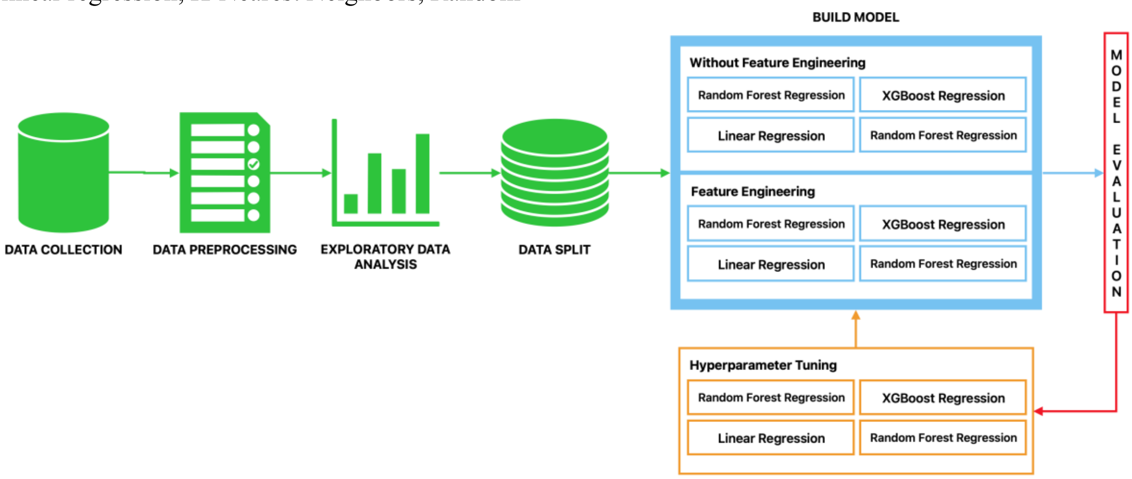

Figure 1. Proposed Method

2.1. Data Collection

This study utilized data sourced from the Rumah123.com website, which provides data related to house prices throughout the City of Bandung. This dataset alone consisted of 7,609 rows of data, including significant and relevant information to forecast house pricing based on property characteristics. The data from Rumah123.com provided a reasonably comprehensive picture of the factors that may affect house pricing, and the information will help to build a better prediction model to estimate house prices. The characteristics of the dataset used in the modeling are shown in Table 1.

Table 1 Feature Dataset

Feature	Description
Price	A target variable that represents the price of the house in Rupiah.
Bedroom Count	Indicates the number of bedrooms available in the house.
Bathroom Count	Indicates the number of bathrooms present on the property.
Carport Count	Describes the capacity of the available parking area or garage.

p-ISSN :1858-3911 , e-ISSN : 2614-5405
https://journal.fkom.uniku.ac.id/

Feature	Description
Land Area	The land area of the house is in square meters (m²).
Building Area	The area of the house is in square meters (m²).
Location	The house's geographical location in Bandung City is converted into numerical data through encoding techniques.

Price is the dependent or target variable in this research. It is the primary variable we want to study; we want to estimate the house price based on the other property characteristics. House prices will be expressed in Rupiah, which is influenced by many factors in the data, such as the number of bedrooms, the number of bathrooms, the size of the parking area or garage, the size of the land, and the size of the building, among many others. Thus, house price is a significant variable in this research and will be calculated or predicted using the various features from the dataset.

2.2. Data Preprocessing

Several key steps are involved in data preprocessing techniques to ensure high data quality. Those steps include missing value handling to avoid any missing values that would inaccurately reflect data and, in return, less accurate modeling; duplicate removal to avoid biased modeling due to duplicated data; and finally, for categorical data like location, encoding was performed using the Label Encoding technique so that categorical data could be encoded into numerical data to make it analyzable and easy to model; data normalization was used to normalize all of the different numerical features for the model to work reliably. Lastly, the data was split into training and testing datasets at an 80:20 ratio using the train_test_split function. This would give an equal number of test and training data to train and validate the model. The training dataset consists of 4740 samples, and the testing dataset comprises 1185 samples.

2.3. Exploratory Data Analysis

Figure 2 illustrates the distribution of the variables. The relationship of the variables can be studied using scatterplots and correlation heatmaps. It is evident that house prices positively correlate with several physical variables. There is a correlation between house prices and land_area, albeit not always proportionate, considering location and several other factors. The relationship between prices and building area appears stronger and more stable, involving homes with building areas greater than or equal to. On the other hand, bedroom_count shows a weaker relationship, as there seems to be high price variation even with the same number of rooms, which suggests that this metric is not a leading contributor. Bathroom_count positively correlates with prices - unlike bedroom_count, but does not appear to carry the same strength as building_area or land_area.

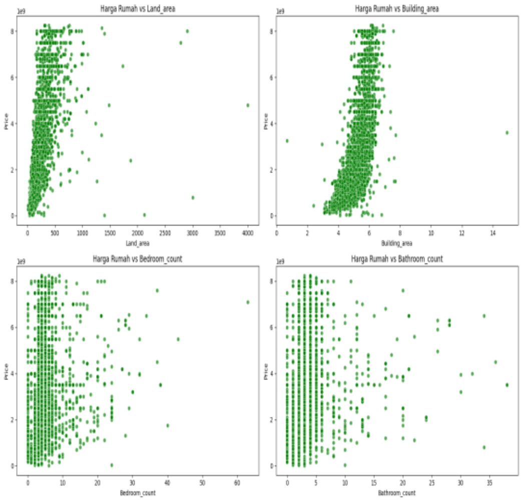

Figure 2. Distribution of Variables

Figure 3 presents a correlation matrix that highlights the strength of the relationship between variables and house prices (price). The correlation of 0.73 between price and building area and 0.63 with land area verifies that both variables are significant predictors of price in the dataset, with a slightly greater influence from building area than land area. There is a strong correlation of 0.79 with bedroom_count predicting bathroom_count. Thus, it would appear that homes with more bedrooms tend to have more bathrooms. The medium correlation of 0.59 between land area and building area appears to confirm that properties with larger land areas tend to have larger buildings as well. Carport_count has a weak correlation with other variables, with a price of 0.31, meaning carport_count appears to be the less significant variable in determining property value. Overall, the heatmap confirms what we already knew, with the two variables—building area and land area—appearing to have the most significant influence on house prices.

p-ISSN :1858-3911 , e-ISSN : 2614-5405
https://journal.fkom.uniku.ac.id/

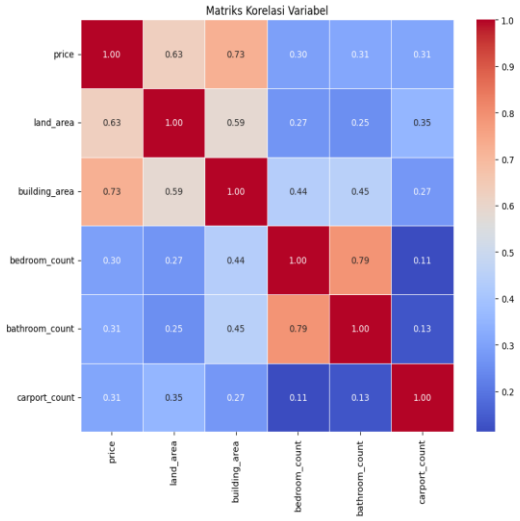

Figure 3. Heatmap Correlation

2.4. Feature Engineering (FE)

At this point, some supplementary features are created based on the relationships between the original variables, data transformation, and planned discretizing. This step enhanced the model's ability to identify patterns in the data and also addressed non-normal distributions. The newly created features are presented in Table 2.

No	Feature	Formula	Description
1	price_per_m2	price / land_area	Calculates the price per square meter (m²) to assess the relative value of a property based on land size.
2	building_land_ratio	building_area / land_area	The ratio between building and land area indicates how efficiently the land is utilized.
3	total_rooms	bedroom_count + bathroom_count	Total number of rooms (bedrooms and bathrooms), representing the comfort and capacity of the property.
4	total_facilities	bedroom_count + bathroom_count + carport_count	Total facilities in the property, including bedrooms, bathrooms, and carports.
5	land_size_category	Categorized using bins: [0, 50, 100, 200, 500, ∞]	Groups land size into defined categories (1: very small, 5: very large).
6	building_size_category	Categorized using bins: [0, 50, 100, 200, 500, ∞]	Groups building size into defined categories (1: very small, 5: very large).
7	land_bedroom_interaction	land_area * bedroom_count	Interaction between land size and number of bedrooms, illustrating whether large properties are proportionally equipped.
8	building_bathroom_interaction	building_area * bathroom_count	The interaction between building size and the number of bathrooms indicates adequate sanitary facilities.
9	price_scaled	np.log1p(price)	Scaled transformation of price to manage skewed price distribution.
10	land_area_scaled	np.log1p(land_area)	Scaled transformation of land area to handle non-normal data distribution.
11	building_area_scaled	np.log1p(building_area)	Scaled transformation of building area to improve interpretability and model stability for skewed data.

Table 2. Feature Engineering

2.5. Modelling

The modeling of this study makes predictions for house prices in Bandung using the number of bedrooms, bathrooms, land area, building area, number of carports, and location as variables. The model primarily aimed to achieve a high level of predictive accuracy while also providing prospective property buyers or developers with an indication of the key factors that drive house prices. Four regression algorithms were used in this regard: XGBoost Regression, K-Nearest Neighbors Regression, Linear Regression, and Random Forest Regression.

In-house price forecasting, as shown in equation (1), utilizes linear regression to acknowledge a linear relationship between house prices and the features of the land area, building area, and number of bedrooms and bathrooms. This model estimates the βi coefficient as the average change in price resulting from a one-unit change in each feature while holding the other features constant. For instance, if β2 refers to the building area, β2 will indicate how much the price is expected to increase for each additional square meter of building area.

$$\hat{y} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \quad (1)$$

Equation (2) is a KNN Regression used for predicting the price of a house based on similar houses in its feature space - houses with identical values for relevant characteristics or variables, such as land area, number of bedrooms, or position. For example, to predict the price of a house with three bedrooms and a land area of 150 m², KNN Regression simply identifies K houses that are nearby with similar feature values for all measures of interest, then calculates the mean average of the K relevant nearest houses' prices, producing an expected cost.

$$\hat{y} = \frac{1}{K} \sum_{i \in \mathcal{N}_{\mathcal{K}}(x)} y_i \quad (2)$$

The Random Forest in equation (3) involves several decision trees that make divisions on the data based on thresholds of any features, e.g., whether the building area is greater than 100 m² or whether the house has more than three bedrooms. Each tree provides a price prediction, and the final price estimation is the average of all trees.

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x) \quad (3)$$

XGBoost Regression in equation (4) constructs an ensemble of trees sequentially, with each new tree attempting to rectify the errors made by the previous tree. For house prices, the model improves its predictions by first making predictions based on the essential features and gradually predicting more accurately by focussing on properties that the previous trees did not perform as expected.

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \quad (4)$$

2.6. Model Evaluation

The analysis of the model is performed with pre-separated test data to identify if the model can be generalized to new data. In the evaluation, three principal metric assessments are used. The first is the Mean Absolute Error (MAE), which measures the absolute average error, or the average distance between the prediction value and the actual value, providing an overall sense of how far the model's prediction is from the actual price. The second is the Mean Squared Error (MSE), which calculates the average square of the prediction error; thus, the MSE penalizes the extremes in error and is more sensitive to outliers, e.g., home prices that have deviated significantly from the expected value. Finally, the R-squared (R^2) Score indicates how much variation in house prices could be explained by the model. When the R^2 score is close to 1, it means reasonable predictive ability. Thus, these three evaluation metrics determine the accuracy, stability, and predictive ability of house prices.

3. RESULTS

To evaluate how effective the regression models are at predicting house prices, they will be assessed using a few performance metrics: the R^2 statistic, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), all based on the transformed price values that were used to achieve more normally distributed price data. This was done to reduce the influence of extreme values and facilitate more accurate model comparisons. The evaluated models included Linear Regression, K-Nearest Neighbors Regression, Random Forest Regression, and XGBoost Regression—with and without Feature Engineering and hyperparameter tuning. The results obtained by the four models are presented in Table 3.

Model	Type	MAE	MSE	RMSE	R2
Linear Regression	No FE	830.4	0.6616	0.8134	0.5766
	Tuning No FE	830.4	0.6616	0.8134	0.5766
	FE	4763	4.3347	2.082	0.8676
	Tuning FE	495.5	0.3013	0.5489	0.8564
	Cross Validation No FE	825.8	0.6372	0.7983	0.5863
	Cross Validation FE	475.2	0.2722	0.5217	0.8701

Model	Type	MAE	MSE	RMSE	R2
KNN Regression	No FE	687.5	0.5514	0.7426	0.6743
	Tuning No FE	580.9	0.5052	0.7107	0.712
	FE	9797	7.5489	2.7475	0.4272
	Tuning FE	675.7	0.6657	0.8159	0.5728
	Cross Validation No FE	559.2	0.4343	0.659	0.7593
	Cross Validation FE	741.7	0.7124	0.8441	0.5145
Random Forest Regression	No FE	533.1	0.4156	0.6447	0.7799
	Tuning No FE	533	0.4156	0.6447	0.7799
	FE	**468.5**	**0.8148**	**1.3725**	**0.9942**
	Tuning FE	47	0.0191	0.1387	0.9941
	Cross Validation No FE	521.1	0.381	0.6172	0.7979
	Cross Validation FE	48.4	0.0205	0.1433	0.9932
XGBoost Regression	No FE	598.6	0.6528	0.8083	0.7722
	Tuning No FE	539.3	0.6333	0.8015	0.7902
	FE	1053.2	1.1716	1.3725	0.9867
	Tuning FE	**45.4**	**0.0166**	**0.129**	**0.9955**
	Cross Validation No FE	530.9	0.6196	0.7872	0.796
	Cross Validation FE	48.2	0.1611	0.0259	0.9911

Table 3 Model Evaluation

The basic Linear Regression model without any Feature Engineering (No FE) yields an MAE of 830.4 and an R^2 of 0.5766, indicating that the model can explain approximately 57.66% of the variance in house prices. The alternative model with Feature Engineering increases the MAE to an enormous 4763, while the R^2 improves to 0.8676. The difference in R^2 indicates that the model may be more suitable for the data, but it also suggests the possibility of overfitting. When this model is tuned, the MAE is reduced to 495.5, and the R^2 value is 0.8564, indicating a firm reliance on the training data results. Cross-validation provided a more stable result with $R^2 = 0.8701$, suggesting that with some improvement, using Feature Engineering with linear regression can be challenging, and documenting performance decreases in-house price is essential.

For KNN Regression, with no Feature Engineering, the model yields an MAE of 687.5 and an R^2 of 0.6743, which is only marginally better than linear regression. With Feature Engineering, the model's performance suffers significantly, with the MAE increasing to 9797 and the R^2 dropping to 0.4272. This suggests that KNN Regression struggles with the complexity of Feature Engineering, and it appears that the model is overfitting. The cross-validation results demonstrate that KNN Regression with Feature Engineering is particularly bad (R^2 of 0.5145), and KNN Regression without Feature Engineering is better (R^2 of 0.7593). This suggests that Feature Engineering is not always beneficial for KNN regression, and a more thorough evaluation is necessary to understand its implications properly.

Random Forest Regression has outperformed the prior models with and without Feature Engineering. When using the Random Forest model without Feature Engineering, the MAE was 533.1, and the R^2 was 0.7799, which indicates that the model is quite effective in managing the variability of housing prices. After including Feature Engineering, the MAE improved to 468.5, and R^2 increased almost to the perfect margin of 0.9942, indicating this model managed to take additional features in stride. However, it also begs the question of overfitting due to the vast difference in training and test data results. Random Forest tuning with Feature Engineering only showed improvement with MAE at an astonishingly low 47 and R^2 close to 1 at 0.9941. This again reinforces that Random Forest Regression performs incredibly well when provided with the appropriate features.

XGBoost Regression with feature engineering yields strong results but shares similar shortcomings with the Random Forest Regression model. Without Feature Engineering, the XGBoost Regression models have an MAE of 598.6 and an R^2 of 0.7722, indicating that they perform well but not as well as the Random Forest Regression model. With Feature Engineering, the XGBoost Regression shows better performance R^2 of 0.9867. However, like the Random Forest Regression model, there is a risk of overfitting the training data, as indicated by the significant difference between the training and test results. Tuning with Feature Engineering produces the best results, with a very low MAE of 45.4 and an R^2 of 0.9955, which demonstrates the agnostic truthfulness of XGBoost Regression in predicting a very accurate outcome when correctly semantically predicted. However, any degree of overfitting would risk incorrect findings without proper validation.

One of the most important takeaways from these evaluation results is that Feature Engineering enhances model performance. In some cases, such as with Random Forest Regression and XGBoost Regression, feature engineering can add features and significantly improve model accuracy. In other instances, KNN Regression can be hindered by the addition of features, indicating that all models are limited by the number of features that can be added while still yielding reasonable performance. In other words, the outcome of Feature Engineering will not always be strictly positive, as it depends on the model.

Overall, cross-validation is a crucial component in mitigating the risk of overfitting, which is often evidenced in Feature Engineering models. The cross-validation process provides more stable and reliable results, indicating that it is vital to use Cross-validation to illustrate that a model will perform on "real" unseen data and not just training data. The difference in performance following the application of Cross-validation is most evident in Random Forest Regression and XGBoost Regression, as the performance was far more stable after Cross-validation was applied.

In conclusion, XGBoost Regression and Random Forest Regression are the top-performing models for predicting house prices. XGBoost Regression shows a slight edge after tuning and applying Feature Engineering. However, it is essential to note that while Feature Engineering can improve model accuracy, the risk of overfitting remains, and techniques such as cross-validation and regularization are necessary to ensure that the model performs well on new data and is not overly reliant on the training dataset.

4. DISCUSSION

Table 4 compares studies on house price prediction using machine learning (ML) models. It provides essential information about the algorithms used, evaluation metrics, outcomes, and R^2 accuracy values of R^2 for each study. The results of the comparison highlight performance differences among the algorithms and the factors considered in each study. The results from the studies presented in the table highlight the strengths and limitations of the findings, indicating which models offer some accuracy.

Study	Models Used	Evaluation Metrics	Accuracy Results (R^2)
This Study	Linear Regression, KNN Regression, Random Forest Regression, XGBoost Regression	MAE, MSE, RMSE, R^2, Cross Validation	Random Forest Regression: 0.9941 XGBoost Regression: 0.9955
[8]	Multiple Linear Regression, Random Forest	R^2	Multiple Linear Regression: 0.73 Random Forest: 0.69
[17]	Multiple Linear Regression, Ridge, Lasso, Random Forest, Polynomial Regression	R-squared, RMSE, Cross-Validation	Multiple Linear Regression & Lasso Regression: 0.6947 Random Forest: 0.69 (Cross-Validation)
[18]	Decision Tree Regression, Linear Regression, Multiple Linear Regression, Random Forest Regression	R-Squared	Random Forest Regression: 82.09% Decision Tree Regression: 73.32%

Table 4 Comparison of House Price Prediction

From the comparison, it is evident that the results from the current study (2025) demonstrate greater accuracy. The Random Forest Regression R^2 value and the XGBoost Regression R^2 value are 0.9941 and 0.9955, respectively, and provide significantly more

p-ISSN :1858-3911 , e-ISSN : 2614-5405
https://journal.fkom.uniku.ac.id/

accuracy than the previous studies, which only reported lower R^2 values (i.e., 0.69 - 0.73) [8], [17], [18]. While the Xiaoyan Ouyang (2024) study bounded the potential of basic machine learning (ML) models, such as Linear Regression and Random Forest, it also yielded moderate results [8]. Similarly, both Mindi Richia Putri (2023) and Naeem Akhtar et al. (2023) identified Random Forest as the best performer; however, the accuracy was low in comparison to our study [18], [17]. Therefore, with sound engineering features, we have increased the value of using advanced models, such as XGBoost Regression and Random Forest Regression, to improve the accuracy of our in-house price forecasting predictions.

Cross-validation is a highly effective method for addressing the problem of overfitting. By dividing the dataset into several subsets, the model is trained on some of the data and then tested on the remaining data. This process is repeated multiple times to ensure that all data is utilized for both training and testing. Through this method, cross-validation provides a more accurate estimate of the model's performance on unseen data, helping to detect overfitting earlier. Additionally, cross-validation enables the selection of more optimal hyperparameters, resulting in a more robust model that can adapt to various data conditions.

5. CONCLUSION

This research study aims to develop a house price prediction model in Bandung using multiple regression techniques, including linear regression, K-nearest neighbors (KNN) Regression, Random Forest Regression, and XGBoost Regression. In terms of the results, Random Forest Regression and XGBoost Regression were the best-performing models after Feature Engineering and tuning, with the highest R^2 values of 0.9941 and 0.9955, respectively. The models were found to sufficiently model the non-linear relationships of variables, including the interactions between building area, land area, and location. Linear regression performed better after Feature Engineering but performed worse than decision tree-based models, mainly due to its inability to capture non-linear interactions well. KNN Regression performed poorly, with a very low R^2, and is therefore unsuitable for this research.

This research will add value by highlighting additional factors influencing house prices in Bandung and proposing predictive models to develop a more accurate house price estimation model. This research sheds light on the Feature Engineering process, utilizing price_per_m2 and building_land_ratio, which have been valuable in furthering model prediction accuracy, particularly in correcting for the non-normal data distribution of the current relevant data. The findings in the present study can also serve as helpful reflections for property market stakeholders, such as satisfied buyers, frustrated developers, and investors, enabling them to make more informed decisions when estimating house prices in local markets or cities with similar market features.

6. SUGGESTS

In light of the findings of this research, potential avenues for further investigation could include developing a more detailed prediction model for house prices, particularly by incorporating more complex local factors. For instance, a research element surrounding the relationship between house prices and macroeconomic factors such as interest rates, inflation, or governmental policies affecting the property market in Bandung might be interesting. In addition, researchers could consider using deep learning models, such as neural networks or more advanced techniques, to leverage the ability to utilize large datasets and the complexity of the relationships between features that traditional models, such as Random Forest Regression or XGBoost Regression, may not have captured. More broadly, researchers could expand on the dataset by considering parts of the world beyond Bandung to examine whether house price prediction patterns are more global or tied to local contexts.

Furthermore, future research may want to investigate how ensemble methods can potentially improve multiple prediction models, thereby making them less prone to overfitting and more robust. For example, we may be able to test hybrid models that combine Random Forest Regression and XGBoost Regression features or utilize stacking models to incorporate different predictive models, thereby adding possibilities for enhancing house price predictions. Researchers could also consider developing real-time predictive systems that enable monitoring of house price changes from period to period in the property market and include AI-based recommendations to better assist users in making informed decisions about their property investment opportunities.

REFERENCE

[1] O. Yalgashev, A. K. Natarajan, and M. G. Galety, 'Foundations of AI and Machine Learning in Real Estate Valuation: An Analysis Using the California Housing Prices Dataset With Python Implementations', in *Advances in Computational Intelligence and Robotics*, M. G. Galety, J. H. Claver, A. V. Sriharsha, N. R. Vajjhala, and A. K. Natarajan, Eds., IGI Global, 2024, pp. 1–36. doi: 10.4018/979-8-3693-6215-0.ch001.

[2] N. T. M. Sagala and L. H. Cendriawan, 'House Price Prediction Using Linier Regression', in *2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED)*, Sukabumi, Indonesia: IEEE, Jul. 2022, pp. 1–5. doi: 10.1109/ICCED56140.2022.10010684.

[3] A. Banerjee, A. Pandey, B. Prakash, S. Banerjee, A. Bandyopadhyay, and P. Chakraborty, 'Enhancing Zillow Zestimates: Leveraging Machine-Learning for Precise Property Valuation Predictions', in *2024 IEEE Students Conference on Engineering and Systems (SCES)*, Prayagraj, India: IEEE, Jun. 2024, pp. 1–6. doi: 10.1109/SCES61914.2024.10652318.

[4] I. Zitoune and M. K. Arabov, 'Comparative Analysis of Ensemble and Linear Machine Learning Models in the Task of House Price Prediction', in *2024 International Russian Automation Conference (RusAutoCon)*, Sochi, Russian Federation: IEEE, Sep. 2024, pp. 50–55. doi: 10.1109/RusAutoCon61949.2024.10694522.

[5] M. Sharma, D. Sharma, R. Burle, P. Patil, I. Joge, and C. Puri, 'Predicting House Price Model : A Comprehensive Analysis with Random Forest and Decision Tree Method', in *2024 3rd International Conference for Innovation in Technology (INOCON)*, Bangalore, India: IEEE, Mar. 2024, pp. 1–6. doi: 10.1109/INOCON60754.2024.10511732.

[6] A. Sul, V. Jagtap, P. Jesalpura, A. Nema, and R. R, 'Optimizing House Price Prediction: Comparative Analysis of Machine Learning Techniques', in *2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, Trichirappalli, India: IEEE, Jul. 2024, pp. 1–7. doi: 10.1109/ICEEICT61591.2024.10718610.

[7] T. Alshammari, 'Evaluating machine learning algorithms for predicting house prices in Saudi Arabia', in *2023 International Conference on Smart Computing and Application (ICSCA)*, Hail, Saudi Arabia: IEEE, Feb. 2023, pp. 1–5. doi: 10.1109/ICSCA57840.2023.10087486.

[8] X. Ouyang, 'House Price Prediction Based on Machine Learning Models', *HSET*, vol. 85, pp. 870–878, Mar. 2024, doi: 10.54097/ftyf9665.

[9] N. Saeed, A. Shahzad, M. Bashir, and L. A. Caceres-Najarro, 'Predictive Modeling of House Prices; A Regression Approach', in *2023 18th International Conference on Emerging Technologies (ICET)*, Peshawar, Pakistan: IEEE, Nov. 2023, pp. 159–164. doi: 10.1109/ICET59753.2023.10374881.

[10] A. Elnaeem Balila and A. B. Shabri, 'Comparative analysis of machine learning algorithms for predicting Dubai property prices', *Front. Appl. Math. Stat.*, vol. 10, p. 1327376, Feb. 2024, doi: 10.3389/fams.2024.1327376.

[11] N. Chuhan, 'House price prediction based on different models of machine learning', *ACE*, vol. 49, no. 1, pp. 47–57, Mar. 2024, doi: 10.54254/2755-2721/49/20241058.

[12] N. Alamsyah, Budiman, V. R. Danestiara, I. Akbar, A. R. Sinaga, and R. Nursyanti, 'Driven Multivariate Regression - Feature Engineering with Random Forest and XGBoost for Accurate Weather Prediction', in *2024 6th International Conference on Cybernetics and Intelligent System (ICORIS)*, Nov. 2024, pp. 1–6. doi: 10.1109/ICORIS63540.2024.10903855.

[13] V. Hoxha, 'Comparative analysis of machine learning models in predicting housing prices: a case study of Prishtina's real estate market', *IJHMA*, vol. 18, no. 3, pp. 694–711, Mar. 2025, doi: 10.1108/IJHMA-09-2023-0120.

[14] R.-T. Mora-Garcia, M.-F. Cespedes-Lopez, and V. R. Perez-Sanchez, 'Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times', *Land*, vol. 11, no. 11, p. 2100, Nov. 2022, doi: 10.3390/land11112100.

[15] S. Zhong, 'Factors influencing housing prices: A comparative study using multiple linear regression and random forest', *TNS*,

vol. 51, no. 1, pp. 73–79, Nov. 2024, doi: 10.54254/2753-8818/51/2024CH0174.

[16] M. Čeh, M. Kilibarda, A. Lisec, and B. Bajat, 'Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments', *IJGI*, vol. 7, no. 5, p. 168, May 2018, doi: 10.3390/ijgi7050168.

[17] D. A. Putri, D. A. Kristiyanti, E. Indrayuni, A. Nurhadi, and D. R. Hadinata, 'Comparison of Naive Bayes Algorithm and Support Vector Machine using PSO Feature Selection for Sentiment Analysis on E-Wallet Review', *J. Phys.: Conf. Ser.*, vol. 1641, no. 1, p. 012085, Nov. 2020, doi: 10.1088/1742-6596/1641/1/012085.

[18] N. Th. Yousir, S. M. Abdulameer, and S. A. Mostafa, 'Data Mining Approach in Predicting House Price for Automated Property Appraiser Systems', in *International Conference on Innovative Computing and Communications*, vol. 537, A. E. Hassanien, O. Castillo, S. Anand, and A. Jaiswal, Eds., in Lecture Notes in Networks and Systems, vol. 537. , Singapore: Springer Nature Singapore, 2023, pp. 543–554. doi: 10.1007/978-981-99-3010-4_45.