International Journal of Management, Entrepreneurship, Social Science and Humanities 9 No. 1 (2025) https://doi.org/10.31098/ ijmesh..v9i1.3567

Check for updates

Research Paper

Improving Housing Price Prediction with Machine Learning: Evidence from Yogyakarta and Implications for Emerging Urban Markets

Galuh Sudarawerti[1], Fahmi Arif[2]
[1]Telkom University, Indonesia
[2]Institut Teknologi Nasional, Indonesia

Received: July 1, 2025	Revised: August 9, 2025	Accepted: September 30, 2025	Online: October 11, 2025

Abstract

Predicting housing prices accurately remains a major challenge in the real estate industry, especially in fast-urbanizing areas where both structural and locational factors are at play. This study focuses on the Special Region of Yogyakarta, Indonesia—a city with varied land use and dynamic housing market conditions—to explore how machine learning (ML) can support better price forecasting. Using the CRISP-DM framework, we analyzed data from 2,020 residential listings, incorporating variables such as building area, land size, number of rooms, and district. Among the several classification models tested, Random Forest achieved the highest accuracy and F1-score. According to the feature importance analysis, building area, land area, and district emerged as the strongest predictors, while vertical features suchs as the number of floors, had relatively little effect. These findings suggest that in Yogyakarta's market, observable physical features may play a bigger role in price determination than location-specific factors. While the study offers a practical framework that real estate professionals can apply, its reliance on structural data and a single-region focus limits how broadly the findings can be applied. Future research could expand on this by including socioeconomic or environmental variables to enhance model performance and relevance across various markets.

Keywords: *house price prediction; machine learning; real estate analytics; housing market*

INTRODUCTION

House price variability poses a significant challenge for real estate professionals worldwide. Various factors, Various factors, including accessibility, environmental quality, infrastructure, location, and socio-economic dynamics, influence the variation in house prices (Liu et al., 2025; Yu et al., 2024a, 2024b). These variations pose challenges for the housing industry, particularly for real estate businesses, in determining the most appropriate pricing strategies. A price set too high may deter buyers, while underpricing may result in lost revenue and inefficient inventory turnover.

This study selected the Special Region of Yogyakarta as the case study due to its rapidly developing urban structure, diverse land-use patterns, and increasing land-use diversity, resulting in variations in housing prices (Chandraderia et al., 2022; Suparmono et al., 2017). Although it is a medium-sized city, its spatial and economic dynamics represent broader phenomena observed in many other urban areas across Indonesia and Southeast Asia (Chandraderia et al., 2022; Guarini et al., 2025; Hidayati & Bagaskara, 2024). Moreover, Yogyakarta exhibits polycentric development, characterized by increasing traffic congestion and a shift in accessibility patterns (Suparmono et al., 2017). The overall character of the city provides a relevant setting for testing new pricing strategies grounded in real market behavior.

Due to this phenomenon, real estate businesses, including agents, brokers, and developers, frequently struggle to set accurate house prices that actually represent the value of the properties. The asymmetric information regarding aspects that determine the asset's value, combined with rapid market fluctuations, makes it challenging to determine the price. This situation leads to inconsistent valuations and prolonged sales cycles, since house pricing is determined based on informal benchmarks and subjective experience.

Numerous studies have been conducted to explore the determinants of housing prices. These studies are predominantly utilizing conventional econometric and statistical methods, such as regression analysis and hedonic pricing models (Balqis & Purwono, 2021; Capozza et al., 2002; Cohen & Karpavičiūtė, 2017). These approaches rely heavily on linear assumptions and may overlook complex and nonlinear patterns in real estate markets. To overcome this limitation, this study adopted a machine learning approach that leverages data-driven techniques to uncover hidden patterns in housing turnover, thus supporting a more objective data-driven decision-making process.

Data analytics, including machine learning (ML), is gaining momentum with the growing availability of online real estate data. This approach offers more accurate, scalable, replicable, and adaptive house price modeling. This approach might overcome limitations embedded in a more traditional static model, such as regression-based modeling, including the Hedonic Pricing Method (HPM), which often struggles with capturing nonlinear relationships or interactions between predictors. ML-based algorithms such as Random Forest, Decision Trees, and Gradient Boosting have the ability to overcome these limitations and uncover complex patterns in large-scale datasets (Cui & Wang, 2025; Kinasih et al., 2024).

This study selected the Special Region of Yogyakarta as the case study due to its rapidly developing urban structure, diverse land use patterns, and increasing land-use diversity, resulting in the variations of housing prices (Chandraderia et al., 2022; Suparmono et al., 2017). Although it is a medium-sized city, its spatial and economic dynamics represent broader phenomena observed in many other urban areas across Indonesia and Southeast Asia (Chandraderia et al., 2022; Guarini et al., 2025; Hidayati & Bagaskara, 2024). Moreover, Yogyakarta has polycentric development with increasing traffic congestion and has experienced a shift in accessibility patterns (Suparmono et al., 2017). Yogyakarta's polycentric growth, suburban expansion (especially in Sleman and Bantul), and increasing land-use diversity—where residential, commercial, and mixed-use developments coexist within the same district—contribute to wide variations in housing prices. These conditions make Yogyakarta a representative case for other Indonesian cities undergoing similar transitions, where rapid urbanization leads to heterogeneous property markets and pricing uncertainty. The overall character of the city provides a relevant setting for testing new pricing strategies grounded in real market behavior.

While ML has been widely applied in housing price prediction globally, its application in Indonesia remains limited, where most prior studies continue to rely on regression-based or hedonic pricing models. These conventional approaches often assume linear relationships and overlook the nonlinear interactions between attributes that are common in emerging markets. In the case of Yogyakarta, property valuation practices are further complicated by reliance on informal benchmarks, such as neighbourhood gossip, agents' personal judgment, or developer mark-ups, which often lead to inconsistent pricing. Addressing this gap, the present study systematically applies machine learning models to the Yogyakarta housing market, demonstrating how predictive modelling can reduce information asymmetry and provide more reliable benchmarks for practitioners. Against this background, this study aims to address the following research questions:

1. Can structural and locational housing attributes be effectively used to predict housing prices in Yogyakarta using machine learning?
2. Which attributes exert the strongest influence on house price prediction?
3. How does the predictive performance of machine learning models?

By answering these questions, this study is expected to make both theoretical and practical contributions. Theoretically, it enriches the house price literature by introducing machine learning as an alternate approach and technique in determining house price behavior. In practice, this study will support the real estate businesses in designing more accurate pricing strategies by identifying key attributes that determine the housing value. Overall, this study supports the data-informed decision-making process in the housing industry, particularly in cities like Yogyakarta that present spatial and market variability.

LITERATURE REVIEW

House Price Determinants and Variability

Understanding the factors that influence house prices is crucial in supporting real estate businesses. Therefore, developing a rigorous methodology to predict and support decision-making in this area is crucial. House price formation is inherently complex, influenced by various factors ranging from structural characteristics (e.g., building size, number of rooms and surface area), to local attributes, market sentiments, and broader socio-economic aspects (Liu et al., 2025; Phipps, 2020). Recent literature has begun to unravel these dynamics by employing several approaches and methodologies.

A spatiotemporal dynamics analysis across the market was employed by Liu et al. (2025) to identify pivotal factors influencing housing prices. Their finding shows the dominance of agglomeration effects, transportation accessibility, and regional economic indicators as the prominent factors. Meanwhile, Mattera and Franses (2025) highlighted the importance of regional clustering in housing markets by proposing a novel forecasting model that incorporates global and cluster-specific latent factors using spatiotemporal clustering techniques. Based on this work, the predictability of house prices can be increased when both macroeconomic trends and region-specific dynamics are considered. Work by Huang et al. (2024a) revealed that several factors, including spaciousness, convenience, and diversity, known as livability factors, significantly influence housing prices across core and non-core urban areas. Another approach incorporating media sentiment has emerged as a significant predictor of housing price fluctuations. A study by Biktimirov et al. (2024) that utilized topic modeling in the case of real estate-related news articles in Canada and Australia showed that the sentiment implied in media narratives exhibits a predictive relationship with house price movements. This effect varies across different countries. Moreover, another aspect, such as the government's policy, also potentially affects the house price dynamics through land-use policy, infrastructure investment, and spatial planning (Yu et al., 2024). Confirming this finding, another study conducted by Yu et al. (2024) proposed the concept of "urban critical zone," suggesting that house prices are influenced not only by market mechanisms but also by systemic environmental and governance factors. These works confirm hedonic pricing theory, which conceptualizes property prices as the accumulation of valued attributes (Capozza et al., 2002).

Taken together, prior studies underline that house prices are determined by various context-specific aspects that are multi-layered and nonlinear. The determinants span from measurable structural aspects to intangible socio-cultural value. Capturing this multi-layered complexity of house price strategy is important. Focusing on more consistent and observable structural aspects potentially offers a practical entry point for predictive modeling. In this context, applying a machine learning analytical approach is valuable in identifying patterns within structured datasets and improving consistency and decision-making related to price categorization and determining aspects.

Machine Learning Approaches In Housing Prediction

Research on house price prediction and its determinants has predominantly employed traditional statistical models, such as linear regression, hedonic pricing, or time-series forecasting. Despite their value in offering interpretability and handling small datasets, these models rely on assumptions of linearity, stationarity, and other assumptions in determining prior variable selection. Recent work even questions the ability of conventional econometric models to capture nonlinear interactions and spatial heterogeneity. Wu and Deng (2024), for instance, show the effect of urban renewal policies on price trajectories, in which such variation is difficult to model with linear regression. These characteristics limit their ability to handle complex, nonlinear, and high-dimensional data (Mattera & Franses, 2025). As urban housing markets became more heterogeneous and increased in volume as well as complexity, these limitations constrained the predictive performance of the model utilized.

Machine learning (ML) models, on the other hand, offer an alternative approach that allows for data-driven pattern recognition without limitation on predetermined parametric assumptions. The strict linearity assumption used in the regression model often fails to capture the complex dependency in property data, particularly when using structural variables for general valuation (Forys, 2022). Compared to the statistical approach. The ML approach offers various advantages, such as the ability to handle large amounts of data with high complexity and non-linearity, incorporate a large number of predictors, and handle various data structures. Furthermore, ML models consistently show better accuracy and sensitivity (Lee et al., 2020; Ong et al., 2024). The comparison between ML and conventional statistical analytical methods is presented in Table 1.

Table 1. Comparison of conventional statistical models with machine learning models

Aspect	Conventional statistical models	Machine learning models
Underlying assumptions	Parametric: require linearity, normality, and error independence	Non-parametric: do not need strict assumptions
Variable selection	Predefined and conceptually determined	Automatic feature selection and ranking
Complexity handling	Limited in capturing non-linear and high-dimensional interactions	Capable of modeling complex, non-linear, and high-dimensional patterns
Data requirements	Suitable for small to medium-sized datasets	Suitable for larger datasets
Interpretability	Highly interpretable coefficients	Less interpretable: feature importance analysis is needed
Predictive accuracy	Moderate and sensitive to multicollinearity	Typically higher predictive accuracy

Source: author's syntheses

In terms of house price categorization, ML classification models such as Decision Tree (DT) and Random Forest (RF) are amongst the relevant models. Using if-then rules based on feature thresholds, DTs partition data, allowing for intuitive interpretation of the classification process. However, this model exhibits limitations, as single-tree models are prone to overfitting and instability. RFs overcome these limitations through ensemble learning by averaging predictions from multiple constructed trees. This ensemble learning improves generalization and reduces variance, thus providing a more robust result (Ong et al., 2024). In the context of house prediction, RF models showed strong performance in classifying the data based on structural features such as building size, number of rooms, and property type. Studies in housing price prediction using ML, utilizing RF and gradient boosting models, resulted in high accuracy classification compared to logistic regression (Lee et al., 2020; Ong et al., 2024).

In the field of house pricing determination, there is a growing numbers of studies applying ML models not only for prediction analysis but also for interoperability and scenario-based analysis. Huang et al. (2024b), for example, are developing machine learning-enabled hedonic pricing models by incorporating spatial features. The study shows the positive impact of high road density and functional diversity on housing prices in core areas, with different effects across the peripheral zone. This result demonstrates the utility of ML in improving the house pricing model.

While studies using statistical methods offer valuable insights and complexity, several gaps remain in the existing literature. Prior studies on price determination mostly rely on conventional statistical models, such as regression analysis and hedonic pricing (Balqis & Purwono, 2021; Capozza et al., 2002; Cohen & Karpavičiūtė, 2017). These approaches are limited by a predefined assumption that potentially oversimplifies the complexity of property value determinants. A more context-specific approach has been employed in more recent studies by incorporating macroeconomic indicators, regional clustering, spatial accessibility, and sentiment analysis to improve prediction accuracy (Biktimirov et al., 2024; Mattera & Franses, 2025). Despite its advanced improvements, maintaining the consistency of data availability that meets the requirements of this approach may be challenging. Therefore, this study addresses the gap by incorporating a machine learning-based classification model. Using only structural property attributes such as land area, building size, and number of rooms, this study provides a practical and replicable framework to support real estate businesses in enhancing their decision-making process regarding house price determination. This study also contributes to the literature by demonstrating the use of machine learning as an alternative approach in enhancing decision support, even in the absence of more complex variables. This study provides evidence that even when limited to structural features, machine learning methods can produce practical and scalable results for real estate data analytics (Forys, 2022).

Theoretical Benchmark

The conceptualization of this study is grounded in the hedonic pricing model (HPM) proposed by Capozza et al. (2002), which posits that the value of property depends on a bundle of attributes, ranging from structural and locational attributes. Structural attributes include land area, building size, and number of rooms. The locational attributes comprise environmental characteristics such as accessibility, neighborhood quality, and proximity to services. This theory is widely adopted in various studies related to housing prices due to its comprehensive coverage and sound conceptualization in structuring property prices decomposition into the marginal contributions of individual attributes (Cohen & Karpavičiūtė, 2017). As a result, implementing HPM without further development may underestimate or misrepresent the dynamic complexity of price formation, involving the interaction between structural and locational factors.

In response to these limitations, the HPM approach should be further extended. Machine learning (ML) approaches offer tools that have potential in complementing hedonic theory by capturing and addressing attribute interactions that traditional statistical modeling often overlooks (Forys, 2022; Ong et al., 2024). ML approaches retain the hedonic assumptions that property values derive from their attributes, but further extend this notion by allowing data-driven discovery of complex, non-linear relationships. Standing on this argument, ML can be seen as the computational extension of the hedonic pricing model by enhancing its explanatory and predictive capacity in heterogeneous and dynamic urban markets. Considering the strengths of ML-based approaches in addressing the complexities and dynamics of the housing market, this study adopted ML techniques as a complementary extension to hedonic pricing. This adoption enables the study to situate its analysis within both economic theory and computational modeling.

Research Propositions

Building on the theoretical foundation, this study adopts ML approaches in conducting its analysis based on the hedonic pricing methods in Yogyakarta's housing market. To operationalize this research, two propositions are advanced:

1. Proposition 1: Structural attributes are expected to exert stronger predictive power in housing price categorization in Yogyakarta than vertical attributes.
2. Proposition 2: Among machine learning approaches, ensemble-based models such as Random Forest are expected to outperform single-tree or linear classifiers in terms of their predictive accuracy and robustness

RESEARCH METHOD

This study employs a quantitative methodology, utilizing a predictive exploratory design, to evaluate how structural and locational attributes influence housing price categorization in Yogyakarta. This study focused on predictive modeling rather than hypothesis testing, aiming to compare multiple machine learning algorithms and identify attributes that exert the strongest predictive power. Accordingly, any causal relationships or theory-driven hypotheses are not being carried out in this research. It explores how the accuracy and consistency of housing price estimation can be improved using computational approaches in a real-world dataset.

CRISP-DM (Cross Industry Standard Process for Data Mining) protocol are utilized in this study. CRISP-DM provides a structured, domain-independent framework for developing data mining models (Schröer et al., 2021). In general, CRISP-DM methodology is constructed based on six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The stages of data analysis based on CRISP-DM can be seen in Figure 1. Following this protocol, this section elaborates on how each phase was applied in this study to support housing price categorization using machine learning.

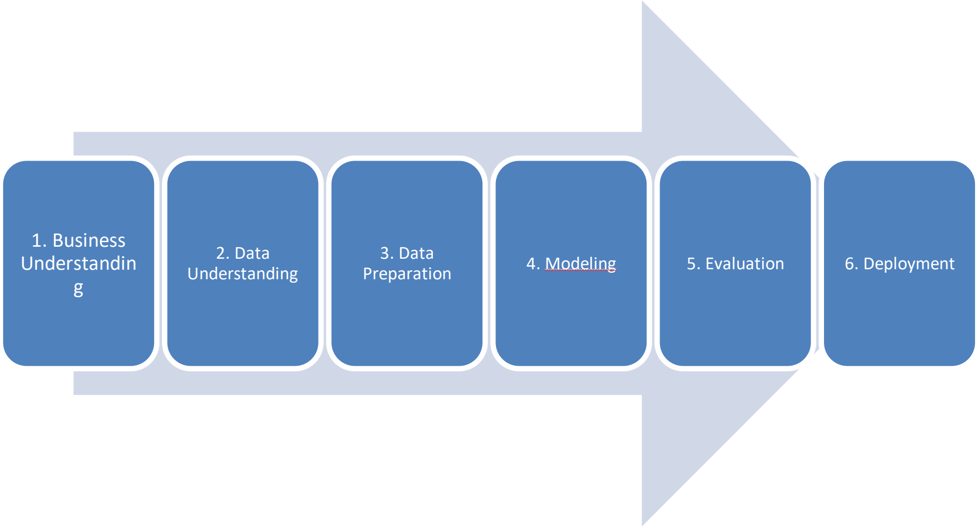


Figure 1. Data analysis protocol

Business Understanding

Addressing the challenge faced by real estate professionals in determining rational house prices, this study aims to build a classification model that enables categorization of house prices using a machine learning approach. In many cases, especially in urban areas like Yogyakarta, price determination is often unstructured, inconsistent, and not supported by accountable data and analytical processes. According to Schröer et al. (2021), this phase focuses on translating business objectives into data mining goals by defining success criteria. In this study, the classification accuracy and feature relevance are defined as the success criteria.

Data Understanding

This study uses data collected from Kaggle by Dika (2024). Rather than a purposive subsample, the dataset collected represents a full census of the publicly available housing listings for Yogyakarta on Kaggle at the time of collection. The data listed 2,020 properties in Yogyakarta. Each of the listings includes price, district, number of bedrooms, bathrooms, garage capacity, land area, and building area. The data size is adequate for the scope of the study with seven main predictors for 2,020 valid records. The minimum recommended observations are 10-30 per variable to ensure the sufficient power and stability of classification models (Silvey & Liu, 2024).

To gain valuable insight from the collected data, initial exploration involved descriptive statistics and visualization to identify data distribution, missing values, and outliers. This is a crucial step since it determines the decision made in the data preparation steps as suggested by Schröer et al. (2021).

Data Preparation

This stage includes data cleansing, in which anomalies and data with missing values were removed. After cleansing, 1,711 valid records were retained. After data cleansing, data tuning was carried out. 'Price' was converted into millions of IDR, 'Kabupaten' was removed due to redundancy, and a new variable, namely 'number of house levels,' was calculated by dividing building area with land area. The price variable was transformed into 15 categories using Jenks natural breaks method to address wide price dispersion. These preprocessing steps are actualizing the third stage in the CRISP-DM protocol, emphasizing data transformation and feature engineering to optimize modeling outcomes (Casonatto et al., 2024; Solano et al., 2021).

Modeling

In developing the classification model. Several machine learning classifiers were tested, including Naïve Bayes, logistic regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Decision Tree, Random Forest, Gradient Boosting, and CatBoost. The choice of these algorithms was based on their suitability for the dataset and research objectives. Some algorithms, including Naïve Bayes, Logistic Regression, LDA, and QDA, are represented as baseline statistical classifiers that rely on linear or probabilistic assumptions (Muhajir et al., 2021). The Decision Tree was included due to its interpretability and ability to capture non-linear rules (Dumitrescu et al., 2022; Yang et al., 2021). Random Forest and Gradient Boosting were selected as ensemble-based models with their ability to handle high-dimensional interactions, reduce variance, and are known to have a better predictive performance, especially for tabular data (Ong et al., 2024; Shrivastav & Kumar, 2021). CatBoost was included due to its specific design that is suitable in handling categorical features efficiently and reducing overfitting in smaller datasets (Hancock & Khoshgoftaar, 2020; Farahani et al., 2025). This combination of different algorithms is needed to be involved in balancing the comparison between simpler and more interpretable models and more complex and high-performing methods.

Each model was trained and evaluated to ensure the validity and reliability of the models. Several established practice was carried out. First, 10-fold cross-validation was applied, focusing on accuracy and F1-score. This procedure is undertaken to minimize overfitting and to provide a more robust estimate of model generalization. This procedure is needed to prepare the data into the training and testing process. Second, multiple confusion matrices were applied in each models to validate the classification consistency across price categories, as well as to provide interpretability of misclassification. While the accuracy provides an overall measure of classification performance, the F1 score balances models' precision and recall. The precision and recall test need to be addressed to show how well minority categories are predicted. The combination of these procedures not only to the accuracy of the models, but also to measure the consistency of the model. Therefore, the combined procedures will be able to enhance the statistical soundness and the practical reliability of the modeling outcomes (Ong et al., 2024; Silvey & Liu, 2024).

Random Forest and CatBoost were further tuned due to their superior baseline performance. After the model performance was being measured, then the results are visualized in a comparative bar chart. The process of selecting the most appropriate model reflects CRISP-DM's iterative experimentation principle as the best practice in aligning technique with data characteristics (Abdelhadi & Almomani, 2023; Solano et al., 2021).

Evaluation

At this stage, the models were evaluated based on classification accuracy, F1-score, and confusion metrics. Random forest consistently shows the highest performance. Combining the performance metrics with visual tools like confusion matrices is essential to assess generalizability and error distribution (Solano et al., 2021). In this phase, the model was being verified to see whether it matches the business objective.

Deployment

Deploying the model into a production environment is beyond this study's limitations. However, through the model development and evaluation, this study provides actionable insights into the most influential predictors of house prices, including building area, land area, house level, bedrooms, garage, and district. Schröer et al. (2021) emphasized the importance of knowledge delivery and documentation at this stage, rather than merely focusing on technical implementation. Supporting this notion, the results of this study aim to support property practitioners in making informed, data-driven decisions.

FINDINGS AND DISCUSSION

This study aims to propose the most suitable method for determining house prices. The housing prices in the Yogyakarta region are highly varied due to asymmetric information, resulting in a high degree of uncertainty. This phenomenon is illustrated in Figure 2, which presents the scattered results of house prices in Yogyakarta based on their price and location. Figure 2 illustrates the scattered distribution of house price levels, presented on a nominal scale. The scatter plot analysis suggested that location (district) is not a reliable determinant of price, as nearly all price levels are present in each district. This result highlights the complexities of determining house prices in Yogyakarta.

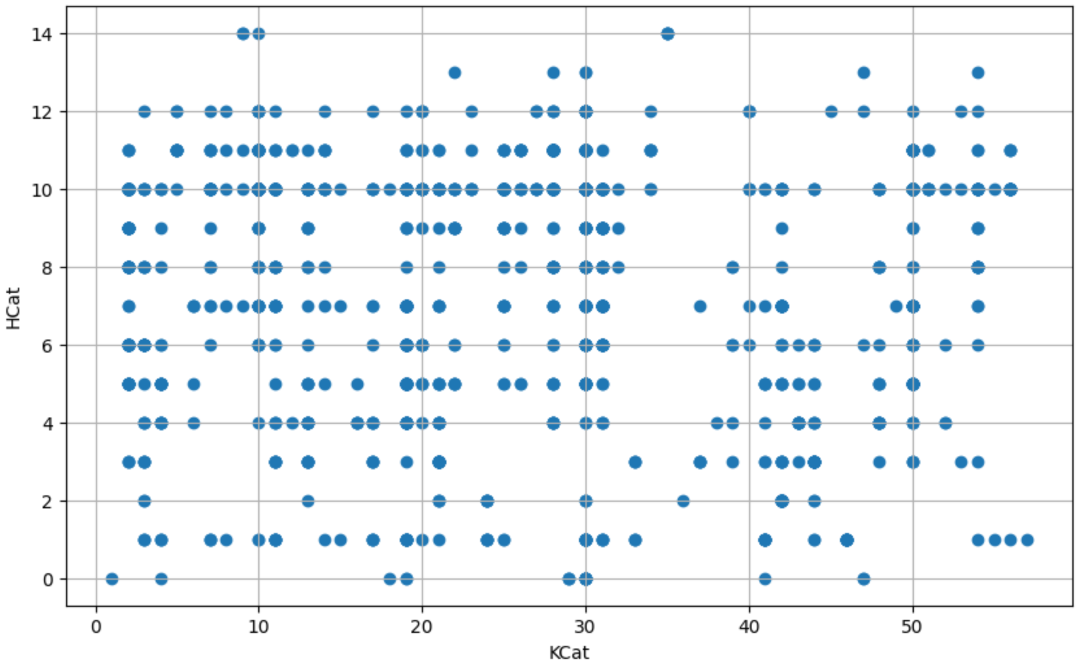


Figure 2. Scatter Diagram of House Prices and Locations

Dataset Characteristics

The 1,710 cleaned dataset comprises valid housing records across Yogyakarta. Each entry consists of structural attributes (land area, building area, number of bedrooms, bathrooms, and carports), as well as locational information. The dataset was then categorized using the Jenks natural breaks method to facilitate the examination of price prediction as both a continuous variable and meaningful market segments. The distribution of records by district, along with a summary of property attributes, will be provided in the following tables.

Table 2. Distribution of Records by District

District	Approx. Share (%)
Sleman	largest share (~40–45%)
Bantul	substantial (~30%)
Yogyakarta City	moderate (~18%)
Kulon Progo	small (<5%)
Gunungkidul	small (<5%)
Total	100%

Table 2 shows that the dataset is not evenly distributed across the district. The majority of housing records are coming from Sleman and Bantul resident, reflecting their key roles in the suburban expansion in Yogyakarta. Kulonprogo and Gunung Kidul show fewer listings, reflecting their slower housing development in the district. The imbalance within the distribution reflects the actual housing market structure in Yogyakarta, which is dominated by suburban areas such as Sleman and Bantul. To mitigate potential bias, the model utilizes cross-validation to ensure that uneven distributions do not compromise its predictive reliability.

Table 3. Summary statistics of property attributes

Attribute	Average (Approx.)	Typical Range
Land Area (m^2)	~100–150	50 – 300+

Attribute	Average (Approx.)	Typical Range
Building Area (m^2)	~80–120	40 – 250+
Bedrooms	2–3	1 – 6
Bathrooms	1–2	1 – 4
Carports	1	0 – 2

Table 3 provides the descriptive characteristics of the dataset. The most common housing configuration in the district consists of two or three bedrooms with one or two bathrooms, showing the modest land and building size. Consistent with the prevalence of compact and low-rise housing preferences, most properties have a single carport. This descriptive figure shows the characteristics of the representative Yogyakarta residential market, which is dominated by middle-sized and low-density housing.

Comparative Model Performance

To ensure the effectiveness of the model and to select the most appropriate model, this study compares eight classification models to be tested and evaluated using accuracy and macro-averaged F1-score. The comparative performance is presented in Figure 3. Random Forest shows the best performance parameter with the highest values of both measurements, followed by Gradient Boosting, which also shows good results. This result provides valuable insight into the next step in the CRISP-DM stages to guide model selection. These results are shown in Table 4 and Figure 3.

Table 4. Model Performance Comparison

Algorithm	Accuracy	F1-score	Notes
Random Forest (Optimized)	0.948	0.922	Strongest performance
Random Forest	0.751	0.634	High baseline ensemble
Gradient Boosting	0.742	0.603	Good boosting model
Decision Tree	0.706	0.563	Interpretable but less robust
CatBoost	0.624	0.523	Handles categories, moderate
k-NN	0.642	0.492	Distance-based, moderate
Naive Bayes	0.389	0.229	Weak performance
Logistic Regression	0.479	0.188	Limited by linearity
QDA	0.293	0.181	Very weak
LDA	0.407	0.136	Very weak

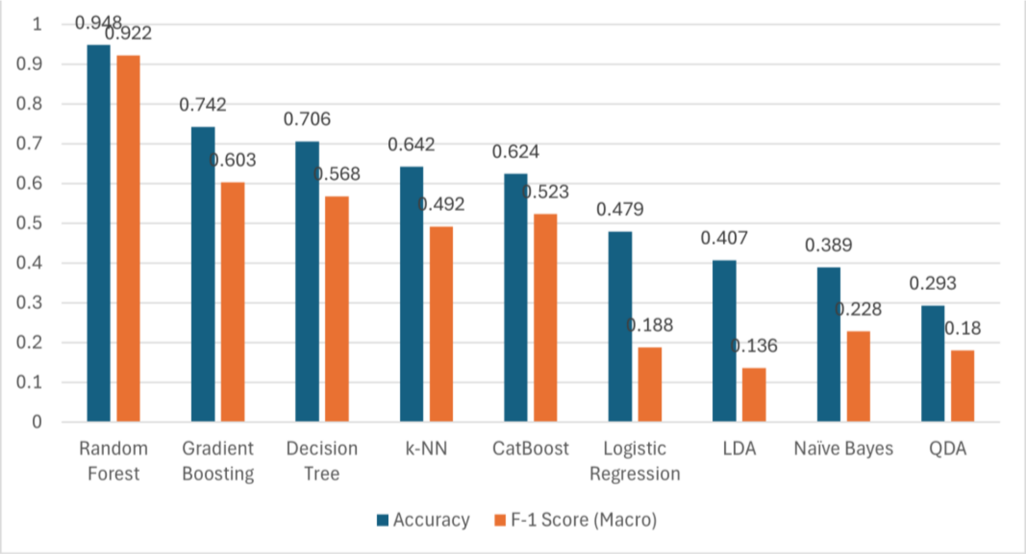


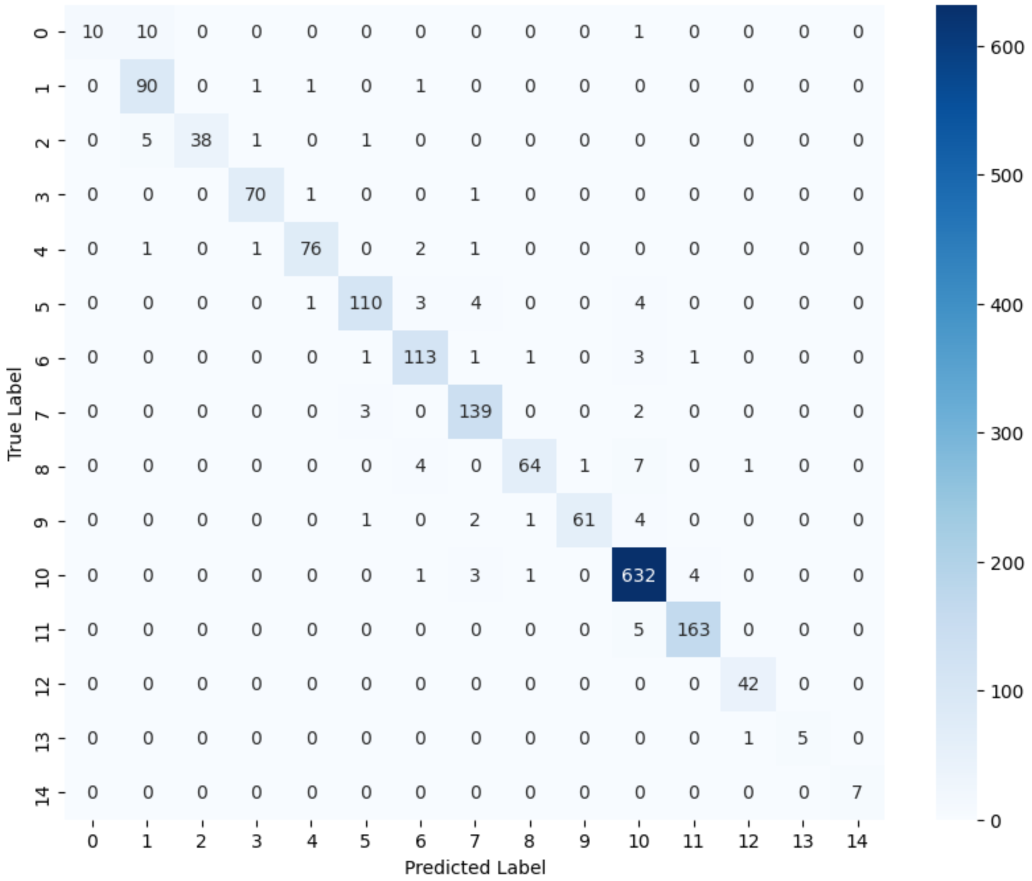
Figure 3. Algorithm Models Comparison Result

Error Distribution through Confusion Matrix

After conducting a comparison of various algorithms, this study found that Random Forest has the best performance; hence, the next stage in CRISP-DM will involve Random Forest. In using random forest, the confusion matrix is then calculated. This matrix is commonly used to assess the accuracy, precision, and recall of a random forest classification model; thus, the classifier's performance can be evaluated (Gope, 2025; Mahapatra et al., 2025; Yu et al., 2025). The confusion matrix assesses classification efficacy by tracking true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) (Gope, 2025). Using this matrix, in a multi-class setting, the diagonal elements represent correctly predicted instances, while off-diagonal elements indicate misclassifications. The visualization of confusion metrics from this study is presented in Figure 4, showing the accuracy of class prediction across 15 housing price categories.

The confusion matrix indicates strong model performance, characterized by diagonal dominance. In this study, price categories 10 and 11 exhibit notably high accurate favorable rates, as indicated by 632 and 163 correct predictors, respectively. These categories represent the mid-to-upper-class housing market, particularly in Sleman and Bantul, two districts with the highest suburban growth. Properties in these brackets are typically large (above 150 m^2 for land and 120 m^2 for building area) and often complemented by additional amenities such as carports. The distinct physical features, along with their prevalence in the dataset, make them easier to classify.

In contrast, occasional misclassifications are observed between neighboring categories, namely categories 5, 6, 7, and 8, where structural differences between units are less pronounced. In these brackets, houses with land areas between 90-110 m^2 and building areas of around 80–100 m^2 may have overlapping pricing brackets, leading to confusion with neighboring categories. These errors, however, are concentrated in the adjacent classes, suggesting that the model's misclassification remains within plausible pricing bands. This condition is acceptable in practical real estate valuation.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	10	10	0	0	0	0	0	0	0	0	1	0	0	0	0
1	0	90	0	1	1	0	1	0	0	0	0	0	0	0	0
2	0	5	38	1	0	1	0	0	0	0	0	0	0	0	0
3	0	0	0	70	1	0	0	1	0	0	0	0	0	0	0
4	0	1	0	1	76	0	2	1	0	0	0	0	0	0	0
5	0	0	0	0	1	110	3	4	0	0	4	0	0	0	0
6	0	0	0	0	0	1	113	1	1	0	3	1	0	0	0
7	0	0	0	0	0	3	0	139	0	0	2	0	0	0	0
8	0	0	0	0	0	0	4	0	64	1	7	0	1	0	0
9	0	0	0	0	0	1	0	2	1	61	4	0	0	0	0
10	0	0	0	0	0	0	1	3	1	0	632	4	0	0	0
11	0	0	0	0	0	0	0	0	0	0	5	163	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	42	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	1	5	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7

Figure 4. Confusion Matrix for Random Forest Model

Feature Importance and Dominant Price Drivers

The result of Random Forest classification provides high prediction performance with interpretable results. The result ranks the relative importance of predictors in determining the house price category. As seen in Figure 5, the most influential variable includes

1. Building area (LB)
2. Land area (LT)
3. District name transformed into a categorical variable (KCat)
4. Number of bathrooms (KM)
5. Number of bedrooms (KT)
6. The availability of carports (Cp)
7. Number of floors in the building, derived by dividing the building area with the land area (Lan).

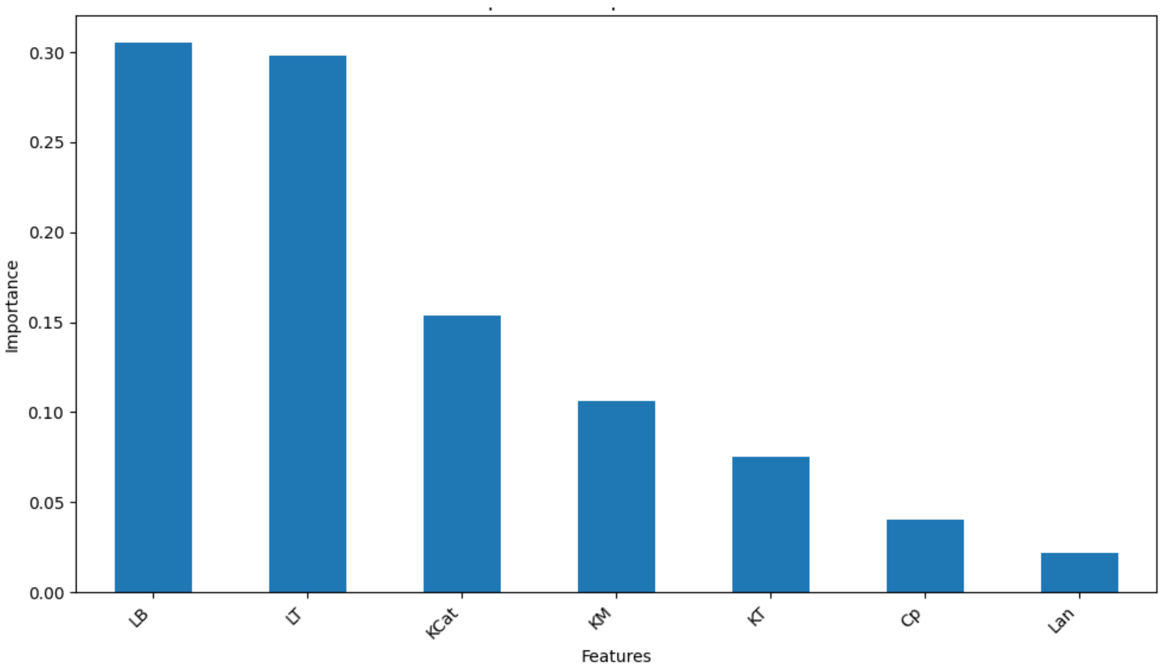


Figure 5. Feature Importance Using the Random Forest Algorithm

The result shows that LB, LT, and KCat are the most influential variables. This finding suggests that physical housing features have a significant influence on price perception, particularly in the Yogyakarta Region. This finding is similar to Yu et al. (2025), whose research found that land area, location, and capacity-related factors were highly influential in assessing contamination risk. Thus, this study reinforces the notion that spatial characteristics have the potential to be reliable predictors in multi-class classification. Furthermore, this finding is also consistent with He (2025), who employs Random Forest to identify dominant features in assessing financial risk. Both works highlighted the ability of Random Forest modeling in filtering key structural elements while assigning minimal weight to redundant or weak contributors.

Another interesting finding is related to the role of KCat as a derived feature indicating the house level. The results suggest that interaction terms may potentially improve classification performance. The least significant variables, such as KM, KT, and CP, also make meaningful contributions as supplementary indicators that refine the model's performance, as suggested by Gope (2025) and He (2025). These variables are still considered important differentiators in determining prices within the interior structure of Yogyakarta, with a tight segmentation of housing markets.

On the other hand, an intriguing finding emerges where Lan, which represents the number of floors in the house, does not have a strong role in determining house prices. This finding suggests that in the context of Yogyakarta's housing market, vertical dimensions contribute less to price categorization than horizontal factors such as land area (LT). This finding indicates that architectural attributes exhibit lower variance and weaker correlation, with limited predictive value, a finding consistent with He (2025). This result also suggests that in the context of Yogyakarta, floor number may not be a strong differentiator, due to two reasons:

1. In the context of Yogyakarta, most residential units are low-rise due to the preference of most residents, who like to live in low-rise buildings or landed houses that are connected to the ground or garden (Swasto, 2017).
2. Citizens have greater emphasis on horizontal attributes than vertical attributes.

Wang et al. (2025), in their study related to urban terrain and housing prices, postulated that the vertical dimensions matter more in mountainous terrain. Thus, a city like Yogyakarta, dominated by horizontal expansion, tends to show less price sensitivity to the building's height, thus resulting in the low level of its relevance in determining the house price.

3. Markets focus more on observable features
   Features like land area and building area are more observable than fine-grained attributes like floor number. According to Broxterman & Zhou (2023), in a market with high levels of asymmetric information (just like the Yogyakarta housing market), price formation tends to be strongly determined by broad and easily observed characteristics rather than micro-level attributes. This explains the less concerned perspective on the floor number in the Yogyakarta housing market.

The scatter plot analysis indicates that location (represented by district) is not a reliable primary determinant for house prices due to the existence of all price levels in each district. However, the Random Forest model identified the district category (KCat) as one of the top predictors. Instead of being contradictory, this finding reflects the different roles' locations play in the market. At the descriptive level, prices are highly heterogeneous across different areas. Thus, the district alone does not cleanly separate housing prices. However, in the predictive analysis setting, the district interacts with structural features such as land area and building area. This setting enhanced the model's ability to discriminate between categories. For example, a 150 m^2 house in Sleman typically falls into a higher price bracket than a house of the same size in Gunungkidul, even if both belong to the same structural category. Thus, location is an influential predictor when combined with structural features, confirming that price formation in Yogyakarta depends on both physical and locational attributes.

CONCLUSIONS

Predicting the determinants of house prices is crucial for the real estate industry. Without a proper and well-structured methodology, the decision regarding price determination will be ill-informed, resulting in either overpricing or underpricing, which can be detrimental for both real estate agents and customers. This study examined how machine learning can improve housing price prediction in Yogyakarta's dynamic housing market. Using the CRISP-DM framework and multiple classification models, this study found that Random Forest consistently outperformed other ML algorithms and conventional statistical approaches such as logistic regression and discriminant analysis. He result shows that the most influential predictors in this context comprises of a combination between structural and locational attributes, particularly land area, building area, and district. On the other hand, vertical attributes such as the number of floors play a limited role.

The result shows that relying on the locations alone is not adequate to determine the house price, since the price level is scattered across different districts. Instead, in the context of Yogyakarta, the building area and land area are the most significant predictors. This result is confirmed by other works that observe the Yogyakarta citizens' behavior, who prefer to live in a low-rise-level building. This finding provides valuable insight for real estate agents in modeling the price determination of housing, especially in urban areas like Yogyakarta.

Conceptually, the study demonstrates how ML extends the hedonic pricing model, whereas the traditional HPM decomposes value into additive linear contributions. ML approaches relax linearity assumptions, can capture complex attribute interactions, and handle uneven data distributions. This ability thereby provides a computational extension of hedonic theory that is more suitable for heterogeneous urban markets.

The findings also provide practical insights, offering actionable guidance for real estate practitioners. In districts with the most development, such as Sleman and Bantul, building size and land size emerge as the most reliable determinants of value. Thus, the agent should use structural benchmarks rather than informal heuristics. Meanwhile, in Yogyakarta, locational effects remain more pronounced, suggesting that agents should adjust their pricing strategy according to the district-specific dynamics. Applying data-driven modeling will contribute to in-house decision-making regarding pricing by reducing information asymmetry, shortening the sales cycle, and achieving more consistent valuations.

In conclusion, the study emphasizes the importance of integrating foundational economic theory with practical and replicable machine learning techniques. The hedonic pricing model provides a conceptual foundation for understanding how different kinds of attributes influence value. This conceptual foundation is supported by machine learning, which offers methodological capabilities for improving prediction outcomes with greater accuracy. Thus, combining the sound conceptual model with machine learning will provide a more robust framework for both researchers and practitioners in housing market analysis.

LIMITATION & FURTHER RESEARCH

Despite the valuable insights and contributions offered, this study should acknowledge several limitations. The first limitation is regarding the geographic scope of its dataset, which focuses on a single city, Yogyakarta. Thus, the generalizability of its findings is limited. Second, this study focuses merely on the structural aspects. Incorporating non-structural aspects, such as economic conditions, socio-cultural factors, and environmental considerations, may enrich the developed model.

Future research could address these limitations in several ways. Expanding the dataset by including multiple cities or regions with different market dynamics would provide a wide range of bases for model validation and better generalization. Providing additional variables, whether structured or unstructured, such as spatial features and neighborhood characteristics, can further enhance the model's accuracy and explanatory power. Furthermore, comparing the ML models and the hybrid approach that integrates ML and statistical methods may provide new insights and fresh perspectives on the strengths and weaknesses of each approach. By addressing these limitations, future research will be able to enhance the predictive power of machine learning models and broaden their applicability to various markets. Building upon this study, future research has the potential to establish more comprehensive and replicable data-driven tools that improve data-driven decision-making in the real estate industry.

REFERENCES

Abdelhadi, A., & Almomani, M. (2023). Selection of suppliers using crisp gradual means integral in conjunction with clustering algorithms. *MethodsX, 11*(October), 102442. https://doi.org/10.1016/j.mex.2023.102442

Balqis, S. F., & Purwono, R. (2021). Determinant of residential property price index in five Asian emerging market countries: A demand and supply approach. *International Journal of Social Science and Economics Invention, 7*(8), 169–177. https://doi.org/10.23958/ijssei/vol07-i08/313

Biktimirov, E. N., Sokolyk, T., & Ayanso, A. (2024). Unpacking the relation between media sentiment and house prices: A topic modeling approach. *Journal of Housing Economics, 66*(September), 102025. https://doi.org/10.1016/j.jhe.2024.102025

Broxterman, D., & Zhou, T. (2023). Information frictions in real estate markets: Recent evidence and issues. *Journal of Real Estate Finance and Economics, 66*(2), 223–243. https://doi.org/10.1007/s11146-022-09918-9

Capozza, D. R., Hendershott, P. H., Mack, C., & Mayer, C. J. (2002). Determinants of real house price dynamics. *NBER Working Paper Series,* October, 1–35. http://www.nber.org/papers/w9262.pdf

Casonatto, R. A., Souza, T. D. P. G., & Mariano, A. M. (2024). Quality and risk management in data mining: A CRISP-DM perspective. *Procedia Computer Science, 242,* 161–168. https://doi.org/10.1016/j.procs.2024.08.257

Chandraderia, D., Siwi, V. N., & Fevriera, S. (2022). Analisis faktor-faktor yang mempengaruhi harga rumah di area aglomerasi Yogyakarta. *Jurnal Pembangunan Wilayah dan Kota, 18*(2), 128–139. https://doi.org/10.14710/pwk.v18i2.37603

Cohen, V., & Karpavičiūtė, L. (2017). The analysis of the determinants of housing prices. *Independent Journal of Management & Production, 8*(1), 49–63. https://doi.org/10.14807/ijmp.v8i1.521

Cui, G., & Wang, C. (2025). The machine learning algorithm based on decision tree optimization for pattern recognition in track and field sports. *PLOS ONE, 20*(2), e0317414. https://doi.org/10.1371/journal.pone.0317414

Dika, P. (2024). *Yogyakarta housing price (Indonesia)* [Data set]. Kaggle. https://www.kaggle.com/datasets/pramudyadika/yogyakarta-housing-price-ndonesia

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research, 297*(3), 1178–1192. https://doi.org/10.1016/j.ejor.2021.06.053

Forys, I. (2022). Machine learning in house price analysis: Regression models versus neural networks. *Procedia Computer Science, 207,* 435–445. https://doi.org/10.1016/j.procs.2022.09.078

Gope, A. (2025). Image-based texture analysis of vowel spectrograms in Sylheti using random forest classifier. *Procedia Computer Science, 260,* 399–405. https://doi.org/10.1016/j.procs.2025.03.216

Guarini, M. R., Roma, A., Sabatelli, E., & Segura-de-la-Cal, A. (2025). Intrinsic and extrinsic attributes in real estate pricing: Insights for sustainable urban planning strategies. *Land Use Policy, 153*(April), 107543. https://doi.org/10.1016/j.landusepol.2025.107543

Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: An interdisciplinary review. *Journal of Big Data, 7*(1), 94. https://doi.org/10.1186/s40537-020-00369-8

He, C. (2025). Enterprise financial risk warning based on random forest algorithm. *Procedia Computer Science, 261,*1229–1237. https://doi.org/10.1016/j.procs.2025.04.709

Hidayati, W., & Bagaskara, B. (2024). Consumer preference for mid-cost housing based on their social stratification in Special Region of Yogyakarta. *EKO-REGIONAL: Jurnal Pembangunan Ekonomi Wilayah, 19*(1), 1–10. https://doi.org/10.32424/1.erjpe.2024.19.1.3368

Huang, J., Wang, Y., Wu, K., Yue, X., & Zhang, H. (2024a). Livability-oriented urban built environment: What kind of built environment can increase the housing prices? *Journal of Urban Management, 13*(3), 357–371. https://doi.org/10.1016/j.jum.2024.04.001

Huang, X., Geng, C., Li, X., & Dong, J. (2024b). Multiple factor flows and heterogeneous premium of urban house prices: Empirical evidence from Chinese cities. *Finance Research Letters, 69*(PB), 106287. https://doi.org/10.1016/j.frl.2024.106287

Kinasih, A. N. S., Handayani, A. N., Ardiansah, J. T., & Damanhuri, N. S. (2024). Comparative analysis of decision tree and random forest classifiers for structured data classification in machine learning. *Science in Information Technology Letters, 5*(2), 13–24. https://doi.org/10.31763/sitech.v5i2.1746

Lee, D., Cho, S. M., Austin, P., Abdel-Qadir, H., Taheri, C., Freitas, C., Tomlinson, G., Chicco, D., Wang, B., Epelman, S., Lawler, P. R., Billia, F., Gramolini, A., & Ross, H. J. (2020). Comparison of machine learning (ML) methods with conventional statistical models (CSM) for prediction of mortality in myocardial infarction (MI) patients. *Journal of the American College of Cardiology, 75*(11), 3538. https://doi.org/10.1016/s0735-1097(20)34165-6

Liu, T., Wang, J., Liu, L., Peng, Z., & Wu, H. (2025). What are the pivotal factors influencing housing prices? A spatiotemporal dynamic analysis across market cycles from upturn to downturn in Wuhan. *Land, 14*(2), 356. https://doi.org/10.3390/land14020356

Mattera, R., & Franses, P. H. (2025). Forecasting house price growth rates with factor models and spatio-temporal clustering. *International Journal of Forecasting*, *41*(1), 398–417. https://doi.org/10.1016/j.ijforecast.2024.09.003

Mahapatra, S., Majhi, B. K., Sarkar, M. S., Datta, D., Mishra, A. P., & Rathnayake, U. (2025). Understanding forest fragmentation dynamics and identifying drivers for forest cover loss using random forest models to develop effective forest management strategies in North-East India. *Results in Engineering*, *26*, 104640. https://doi.org/10.1016/j.rineng.2025.104640

Muhajir, D., Akbar, M., Bagaskara, A., & Vinarti, R. (2021). Improving classification algorithm on education dataset using hyperparameter tuning. *Procedia Computer Science*, *197*, 538–544. https://doi.org/10.1016/j.procs.2021.12.171

Ong, W. J. D., How, C. H., Chong, W. H. K., Khan, F. A., Ngiam, K. Y., & Kansal, A. (2024). Outcome prediction for adult mechanically ventilated patients using machine learning models and comparison with conventional statistical methods: A single-centre retrospective study. *Intelligence-Based Medicine, 10,* 100165. https://doi.org/10.1016/j.ibmed.2024.100165

Phipps, A. G. (2020). Inner-city neighbourhood changes predicted from house prices in Windsor, Ontario, since the early- or mid-1980s. *Journal of Building Construction and Planning Research, 8*(2), 138–160. https://doi.org/10.4236/jbcpr.2020.82009

Rashin Gholijani Farahani, R., Ghazanfari, P., & Zarrabi, A. (2025). A report on CatBoost: Unbiased boosting with categorical features. *ResearchGate,* February. https://doi.org/10.13140/RG.2.2.30029.96485

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science, 181,* 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Shrivastav, L. K., & Kumar, R. (2021). An ensemble of random forest, gradient boosting machine and deep learning methods for stock price prediction. *Journal of Information Technology Research, 15*(1), 1–19. https://doi.org/10.4018/jitr.2022010102

Silvey, S., & Liu, J. (2024). Sample size requirements for popular classification algorithms in tabular clinical data: Empirical study. *Journal of Medical Internet Research, 26,* e60231. https://doi.org/10.2196/60231

Solano, J. A., Lancheros Cuesta, D. J., Umaña Ibáñez, S. F., & Coronado-Hernández, J. R. (2021). Predictive models assessment based on CRISP-DM methodology for students' performance in Colombia – Saber 11 Test. *Procedia Computer Science, 198,* 512–517. https://doi.org/10.1016/j.procs.2021.12.278

Suparmono, Darsono, Sarungu, J., & Riyanto, G. (2017). Land accessibility: The effects of distance versus traveling time to CBD and facility attributes on residential land price. *International Journal of Applied Business and Economic Research, 15*(15), 27–39.

Swasto, D. F. (2017). Friendly vertical housing: Case of walk-up flat housing development in Yogyakarta. *IOP Conference Series: Earth and Environmental Science, 8,* 012074. https://doi.org/10.1088/1755-1315/8/1/012074

Wang, X., Wen, H., Gui, B., Liu, Z., & Yang, L. (2025). Urban terrain, mountain landscape, and housing price: A heterogeneous investigation of the amenity effects in a mountainous city (Guiyang) from the vertical dimension. *Applied Geography, 174,* 103479. https://doi.org/10.1016/j.apgeog.2024.103479

Wu, S. M., & Deng, Y. (2024). Typological differentiation and time-series effects of urban renewal on housing prices. *Cities, 145,* 104668. https://doi.org/10.1016/j.cities.2023.104668

Yang, L., Liang, Y., Zhu, Q., & Chu, X. (2021). Machine learning for inference: Using gradient boosting decision tree to assess non-linear effects of bus rapid transit on house prices. *Annals of GIS, 27*(3), 273–284. https://doi.org/10.1080/19475683.2021.1906746

Yu, H., Zhu, S., Li, J. V., & Wang, L. (2024). Dynamics of urban sprawl: Deciphering the role of land prices and transportation costs in government-led urbanization. *Journal of Urban Management, 13*(4), 736–754. https://doi.org/10.1016/j.jum.2024.08.002

Yu, P., Wei, Y., Ma, L., Wang, B., Yung, E. H. K., & Chen, Y. (2024). Urbanization and the urban critical zone. *Earth Critical Zone, 1*(1), 100011. https://doi.org/10.1016/j.ecz.2024.100011

Yu, T. K., Chang, I. C., Chen, S. D., Chen, H. L., & Yu, T. Y. (2025). Predicting potential soil and groundwater contamination risks from gas stations using three machine learning models (XGBoost, LightGBM, and Random Forest). *Process Safety and Environmental Protection, 199,* 107249. https://doi.org/10.1016/j.psep.2025.107249