Enhancing Housing Price Prediction Accuracy Using Decision Tree Regression with Multivariate Real Estate Attributes

Ahmar Dwi Utomo[1,*], B Herawan Hayadi[2], Eko Priyanto[3]

[1]*Informatics Department, University of AMIKOM Purwokerto, Jl. Let. Jend. Pol. Soemarto No.126, Purwokerto 53127, Indonesia*

[2]*Primary School Teacher Education,Universitas Bina Bangsa, Serang, Indonesia,*

[3]*Ma'arif University of Nahdlatul Ulama, Kebumen, Indonesia*

(Received April 15, 2024; Revised July 20, 2024; Accepted October 25, 2024; Available online December 1, 2024)

Abstract

The real estate sector functions as a critical barometer of a nation's economic performance; however, its inherent volatility and intricate pricing mechanisms often hinder precise valuation—particularly in developing urban markets. In the context of Indonesia, where the property industry contributes substantially to national GDP, deriving fair and data-driven housing price estimates remains a persistent challenge. Traditional appraisal methods, which rely predominantly on subjective human judgment, frequently fall short in reflecting market dynamics accurately. This research seeks to construct an interpretable machine learning framework for predicting residential housing prices by employing a Decision Tree Regression (DTR) model. The DTR method was chosen for its transparent and hierarchical structure, allowing for a clear understanding of how individual property characteristics affect price outcomes. The study utilizes a public dataset from Kaggle containing key housing attributes, including land area, building size, number of rooms, and location variables. The methodological steps encompass data preprocessing (cleaning and encoding using One-Hot Encoding), data partitioning into training and testing sets with an 80:20 ratio, and model performance evaluation using standard regression metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and the Coefficient of Determination (R^2). The model attained an R^2 value of 0.385, suggesting that the selected features explain approximately 38.5% of the variance in housing prices. While this indicates moderate predictive capability, the DTR model offers valuable interpretive insights—particularly in identifying land area as the most influential predictor of price. The findings highlight that interpretable machine learning approaches can serve as effective analytical tools for property valuation in emerging markets, balancing predictive accuracy with transparency. Moreover, this study lays the groundwork for the future development of ensemble and hybrid predictive models, as well as the integration of AI-based analytics into decision-support systems for property valuation, investment forecasting, and urban development planning in Indonesia's evolving real estate landscape.

Keywords: House Price Prediction, Machine Learning, Decision Tree Regression, One-Hot Encoding.

1. Introduction

Real-estate markets serve as a fundamental barometer of a nation's economic performance and financial resilience. Fluctuations in residential property prices not only mirror macroeconomic trends but also influence employment levels, household consumption, and investment confidence. A healthy housing sector typically signals rising disposable income and overall economic stability, while volatility can amplify financial uncertainty and social inequality [1], [2]. In recent years, the increasing complexity of global property markets has prompted the adoption of data analytics and artificial intelligence (AI) to reduce uncertainty and enhance decision-making. AI-driven predictive models can uncover non-linear patterns within massive datasets, offering stakeholders—developers, investors, and policymakers—data-supported insights that improve market efficiency and foresight [3], [4].

Within the Indonesian context, the property sector constitutes a key component of the national economy, contributing roughly 15 percent to GDP [5]. Indonesia's rapid urbanization, expanding at an estimated 4 percent annually, has intensified housing demand in major cities [6]. Bandung, as one of the country's fastest-growing urban centers, illustrates this trend vividly. The city's housing market is characterized by accelerating demographic shifts, infrastructural expansion, and heightened socio-economic diversity. Consequently, property valuation in Bandung has

***Corresponding author: Ahmar Dwi Utomo (dwiutomoahmar@gmail.com)**
DOI: https://doi.org/10.47738/ijiis.v7i4.226
This is an open access article under the CC-BY license (https://creativecommons.org/licenses/by/4.0/).
© Authors retain all copyrights

become increasingly multifaceted, shaped by economic cycles, regulatory adjustments, and varying neighborhood characteristics that collectively influence price dynamics [7].

Despite its economic importance, the Indonesian housing market continues to rely heavily on traditional appraisal techniques that are often inadequate in capturing such complexity. Conventional valuation methods—typically based on comparative market analysis—depend on limited comparable data and the subjective judgment of appraisers [8]. These methods are time-consuming and prone to inconsistency, leading to valuation disparities and reduced stakeholder trust [9]. In volatile and data-rich environments like Bandung, such subjectivity can distort price benchmarks and limit the effectiveness of property investment decisions. Therefore, a methodological shift toward more objective, data-driven approaches is increasingly warranted [10].

Machine learning has emerged as a promising alternative for property valuation, offering automated analysis capable of recognizing intricate patterns across diverse variables. Through algorithms that learn from historical data, these systems can identify non-obvious relationships between property attributes and prices [11]. Machine learning models are highly scalable and adaptive, enabling rapid recalibration when new data become available, thus reflecting current market realities [12]. Prominent global applications include Zillow's "Zestimate" system in the U.S. and similar initiatives by the U.K. Land Registry, where algorithmic valuation has enhanced pricing transparency and predictive reliability [13], [14]. Such developments highlight the transformative potential of AI in real-estate analytics.

However, the integration of machine learning into Indonesia's property sector remains limited, particularly concerning urban areas like Bandung where publicly available, structured datasets are scarce. Most existing studies emphasize predictive accuracy but overlook interpretability—an essential aspect for end-users such as appraisers, agents, and investors who require transparency in decision processes [15]. Consequently, there is a pressing need to develop models that not only deliver accurate price forecasts but also explain how specific features influence valuation outcomes. Addressing this gap can strengthen trust and facilitate broader acceptance of AI-based appraisal systems within local contexts.

This study aims to implement a Decision Tree Regression model to predict residential property prices in Bandung. The Decision Tree algorithm was chosen for its intuitive, rule-based structure that mirrors human reasoning, its compatibility with both numerical and categorical features, and its inherent interpretability [16]. By applying a systematic data-processing pipeline that includes feature transformation, training, and validation, the research seeks to produce a transparent model that balances analytical precision with user comprehension. The emphasis on interpretability differentiates this study from more opaque "black-box" techniques such as deep learning.

2. Literature Review

2.1 Overview of Machine Learning in Property Valuation

The evolution of property valuation methods has shifted markedly from traditional hedonic pricing models to more sophisticated data-driven techniques. The hedonic model, which estimates property value based on structural and locational characteristics, laid the theoretical foundation for quantitative valuation but was constrained by its assumption of linearity and inability to capture complex, non-linear relationships [1]. These limitations became increasingly evident as modern housing markets grew more dynamic and data-rich. Machine learning (ML) approaches have emerged as a powerful alternative, capable of detecting hidden patterns in large, multidimensional datasets and modeling complex interactions that traditional econometric techniques overlook [2].

Among the most widely used algorithms in property valuation are Linear Regression, Random Forest, Gradient Boosting, and Deep Learning models. Linear Regression often serves as a baseline due to its simplicity and interpretability, whereas ensemble algorithms such as Random Forest and Gradient Boosting enhance predictive accuracy by combining multiple weak learners to reduce variance and bias [3], [4]. Deep Learning extends these capabilities further by incorporating unstructured data—such as aerial imagery, textual property descriptions, and social media indicators—allowing for more holistic valuation frameworks [5].

Empirical studies have consistently reported that ML-based models outperform traditional methods in predictive accuracy and adaptability. Navarro et al. [6] demonstrated that ensemble regression methods significantly reduced error

rates compared with classical hedonic models, while Mao et al. [7] found that hybrid architectures combining regression and boosting techniques provided superior generalization on heterogeneous property datasets. These advancements underscore a global shift toward automated, data-driven real estate valuation as a cornerstone of modern housing analytics.

2.2 Decision Tree Models for House Price Prediction

Decision Tree Regression (DTR) represents one of the most intuitive and interpretable machine learning methods, relying on a hierarchical structure that partitions data based on feature-driven decision rules. Each internal node corresponds to a condition on an attribute (e.g., land area or number of bedrooms), and each leaf node represents a predicted output value [8]. This structure enables the model to handle both categorical and numerical variables while providing transparent, rule-based explanations for each prediction—an essential attribute for real estate stakeholders who require clarity in valuation outcomes [9].

Despite its simplicity, DTR forms the basis for more complex ensemble methods. Random Forest, XGBoost, and LightGBM improve upon the Decision Tree by aggregating multiple trees to enhance predictive performance and mitigate overfitting [10]. Random Forest constructs numerous uncorrelated trees and averages their predictions, leading to stable results even in noisy datasets, while XGBoost introduces gradient boosting and regularization mechanisms to optimize accuracy and prevent model drift [11]. LightGBM, developed for high-speed performance on large datasets, further extends scalability through leaf-wise tree growth strategies [12].

Empirical evaluations in various housing markets highlight these models' predictive capabilities. Studies have shown Random Forest achieving R^2 scores exceeding 0.90 in structured datasets [13], while XGBoost and LightGBM have been reported to outperform baseline models across multiple metrics, including Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) [14]. However, these ensemble methods often compromise interpretability—an increasingly important concern for explainable AI applications. To address this issue, interpretive frameworks such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) are now frequently used to quantify feature importance and visualize how input variables influence predicted outcomes [15].

2.3 Research Gap and Positioning of the Current Study

Although machine learning has demonstrated considerable promise in real estate analytics, research remains unevenly distributed across global contexts. The majority of prior studies have focused on mature markets in North America, Europe, and East Asia, where high-quality datasets and standardized property information are available [16]. In contrast, developing regions—such as Indonesia—have received comparatively limited scholarly attention, despite exhibiting unique market characteristics, irregular data structures, and varying valuation standards. Moreover, many existing works emphasize predictive performance using complex "black-box" models like deep neural networks, often at the expense of interpretability and usability for practitioners [17].

This lack of transparency poses a practical challenge in the Indonesian housing market, where buyers, appraisers, and local governments rely on clear, explainable justifications for valuation outcomes. Few empirical studies have examined interpretable machine learning models in Indonesian urban settings, particularly those undergoing rapid demographic and infrastructural change such as Bandung. As a result, there is a critical need for approaches that balance predictive precision with transparency and local adaptability [18].

The present study addresses this gap by developing and evaluating a Decision Tree Regression model tailored to the Bandung housing market. The model emphasizes interpretability and accessibility, allowing users to trace how each input variable contributes to the final predicted price. By applying structured preprocessing, One-Hot Encoding for categorical variables, and quantitative evaluation using MAE, MSE, and R^2 metrics, this study offers a replicable framework for property valuation within data-limited environments. The findings aim to support real estate professionals, investors, and policy institutions in making evidence-based decisions while establishing a methodological baseline for future research on ensemble and hybrid modeling techniques in Indonesia's property sector.

 ISSN 2579-7069

3. Method

The research methodology was systematically structured to ensure a reproducible and scientifically valid process. Each step was implemented using the Python programming language, supported by key libraries such as pandas for data manipulation and scikit-learn for machine learning modeling

3.1 Research Design

The machine learning workflow for house price prediction in this study follows a systematic and reproducible sequence of steps designed to ensure that data is properly cleaned, transformed, and utilized for building a robust predictive model. As illustrated in Figure 1, the process begins with the import of essential libraries and the clear definition of the research objective—predicting residential property prices based on multiple features, including land area, building area, number of rooms, and location. The data collection stage involves obtaining raw housing data from a publicly available Kaggle dataset containing various attributes of residential properties in Bandung. This dataset forms the foundation for subsequent processing and model training. Before standard preprocessing, the dataset undergoes a pre-preprocessing phase to ensure data quality and consistency. This stage includes removing irrelevant columns such as property names or identification numbers that do not contribute to price prediction, detecting and eliminating duplicate records to prevent data leakage or bias, and standardizing measurement units (e.g., converting square feet to square meters) to maintain uniformity across all records.

Following this initial cleaning, the dataset is divided into training and testing subsets using an 80:20 ratio, with a fixed random seed to ensure reproducibility. The training set is used to fit the model, while the testing set is reserved for evaluating performance on unseen data. Once the split is completed, the preprocessing phase begins, where categorical attributes such as location are transformed using the One-Hot Encoding technique, and numerical attributes are standardized where appropriate. This transformation is implemented using a ColumnTransformer within a scikit-learn pipeline, which ensures that all preprocessing steps are applied systematically and prevents data leakage between the training and testing stages. In the model training phase, a Decision Tree Regressor is fitted using the processed training data. The algorithm learns patterns and relationships within the dataset by iteratively splitting the data into smaller subsets based on the most informative features, generating a tree-like structure that represents the decision-making process behind the price predictions. Once trained, the model proceeds to the evaluation stage, where its performance is assessed using three standard regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and the Coefficient of Determination (R^2 Score). These metrics collectively measure the model's predictive accuracy, error magnitude, and explanatory power.

The results are then visualized to enhance interpretability. Visualization outputs include the structural representation of the decision tree, feature importance plots that identify the most influential variables in price determination, and graphical summaries illustrating predicted price distributions. These visual insights not only validate the model's internal logic but also make the findings more accessible to non-technical stakeholders such as real estate professionals and policy planners. Finally, the workflow concludes with the generation of a fully trained and evaluated model, which serves as a foundation for future development and deployment. This structured research design provides a transparent, modular, and replicable framework that aligns with best practices in data science, allowing for further optimization through ensemble methods or additional feature engineering in subsequent studies.

 ISSN 2579-7069

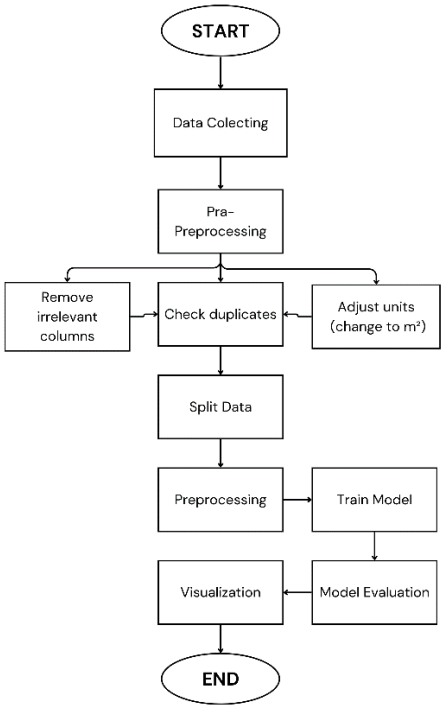

Figure 1. Research Method Flowchart

This modular workflow design ensures clarity, traceability, and the ability to test or improve individual components independently. It reflects best practices in data science pipelines and supports the reproducibility of results.

3.2 Data Collection and Preparation

The dataset used in this study was obtained from an open-source repository on Kaggle, containing detailed information on residential properties located in Bandung, Indonesia. This dataset was imported into a Python environment using the pandas library, which facilitated efficient data handling, exploration, and manipulation. Upon initial inspection, the dataset was organized into two primary variable groups—features (X) and target (y). The feature variables represent the independent attributes that influence property value, including bedroom_count, bathroom_count, carport_count, land_area, building_area (m^2), and location. Each of these variables captures a distinct aspect of the property's physical or locational characteristics. The house_name feature, which served merely as a unique identifier for each record, was deliberately excluded from the analysis because it does not provide any predictive value and could potentially introduce unnecessary noise into the model. The target variable (y) is the price of each property, expressed in Indonesian Rupiah (IDR), which the model aims to predict based on the given features. Before modeling, the dataset was carefully examined to ensure completeness, consistency, and validity. Records containing missing or duplicate values were identified and addressed to prevent data leakage or model bias. Outliers, such as extremely high or low property prices, were analyzed and retained only if they represented realistic market conditions. Additionally, continuous variables like land and building area were standardized to ensure that all numerical features shared a consistent scale, while categorical variables were reviewed for appropriate representation of location data. This step was crucial to maintain the reliability of the model's predictive performance and to prevent distortions arising from inconsistent input formats.

3.3 Preprocessing Data

Preprocessing represents a critical stage in the data pipeline, ensuring that the raw input conforms to the format expected by the machine learning algorithm. In this study, the dataset's attributes were first classified into two types: numeric features—including bedroom_count, bathroom_count, carport_count, land_area, and building_area—and categorical features, represented by location. Since most machine learning models, including Decision Tree Regression, require numeric input, categorical data must be encoded into numerical form. The One-Hot Encoding (OHE) method

was applied to transform the location feature into a series of binary variables, each representing a specific neighborhood or area within Bandung. For each property, a value of 1 in a binary column indicates the presence of that location, while 0 indicates absence. This approach prevents the algorithm from inferring any ordinal or hierarchical relationships between locations, which could mislead the model's interpretation of feature importance. To ensure robustness, the handle_unknown="ignore" parameter was enabled within the encoder, allowing the model to handle any previously unseen location categories that might appear in the testing or deployment phases. Through this encoding process, the categorical feature was converted into a machine-readable format without losing interpretability, ensuring that the model could leverage locational context effectively.

3.4 Pipeline Creation and Data Sharing

To streamline the entire modeling workflow, a scikit-learn Pipeline was constructed to integrate preprocessing and regression into a single unified framework. This pipeline automatically executes each stage sequentially, from data transformation to model training, thus ensuring consistency, reducing manual errors, and preventing data leakage between the training and testing phases. The ColumnTransformer component was embedded in the pipeline to specifically apply One-Hot Encoding to categorical features while allowing numerical features to pass through unaltered. After preprocessing, the dataset was divided into training and testing subsets using the train_test_split() function provided by scikit-learn. An 80:20 split ratio was applied, meaning that 80% of the data was used to train the model while the remaining 20% was reserved for evaluating its performance on unseen samples. This division strikes a balance between providing sufficient data for model learning and ensuring a reliable evaluation phase. The random_state=42 parameter was set to maintain reproducibility, guaranteeing that identical results can be obtained across repeated experiments. By structuring the pipeline in this way, the study ensures a clean, modular workflow that aligns with best practices in data science experimentation and model validation.

3.5 Modeling and Training

The predictive model employed in this research is the Decision Tree Regressor (DTR), a widely recognized algorithm known for its interpretability and straightforward logical structure. The DTR works by recursively partitioning the dataset into smaller subsets based on feature values, ultimately forming a tree-like structure where each node represents a decision rule and each leaf node corresponds to a predicted price. To control model complexity and prevent overfitting, the maximum tree depth parameter was limited to three levels (max_depth = 3). This constraint ensures that the resulting tree remains interpretable and avoids excessive fragmentation of data, which could lead to overly specific predictions that do not generalize well. In addition, a random_state = 42 was specified to maintain experimental consistency and reproducibility throughout the training process. The training phase involved invoking the .fit() function on the training subset (X_train, y_train), allowing the model to learn from the historical relationships between property features and their corresponding prices. During this phase, the Decision Tree algorithm evaluated all possible splits in the feature space to identify those that minimized prediction error, effectively building a decision hierarchy that reflects the underlying structure of the housing market. Once the model completed its training, it was ready for performance evaluation and visualization, providing both quantitative metrics and graphical insights into the rules governing price prediction.

4. Results And Discussion

This section presents quantitative results of the model performance and provides an in-depth interpretation of these results as well as a visualization of the model.

4.1 Quantitative Performance Analysis

This section presents a detailed evaluation of the Decision Tree Regression model's performance in predicting house prices based on key statistical metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and the Coefficient of Determination (R^2). These indicators collectively measure the accuracy, consistency, and explanatory power of the model.

The Mean Absolute Error (MAE) of the model is recorded at approximately IDR 2,454,094,695.01, indicating that, on average, the predicted house prices deviate from the actual values by around IDR 2.45 billion. Although this represents

 ISSN 2579-7069

a relatively large error margin for precise property valuation at the individual level, it is still informative for broader market trend analysis and general estimation. Such a magnitude of error suggests that while the model is capable of identifying overall pricing tendencies, it may not yet be suitable for applications requiring highly granular precision, such as mortgage lending or property taxation. Future improvements in feature representation and data diversity could help reduce this margin and enhance predictive accuracy.

The Mean Squared Error (MSE) is reported at 6.51×10^{19}, a considerably large value attributed to the squaring of residuals in the computation process. Because MSE disproportionately penalizes larger errors, this figure implies that a few extreme outliers—properties whose characteristics differ substantially from the majority of the dataset—may have significantly influenced the model's performance. These high-error predictions highlight the need for better feature normalization, potential removal of anomalies, or model refinement using ensemble techniques such as Random Forest or Gradient Boosting to achieve greater stability and robustness.

The Coefficient of Determination (R^2) stands at 0.385, meaning that the model explains approximately 38.5% of the variance in housing prices across the dataset. This moderate level of explanatory power suggests that the Decision Tree captures essential relationships between property features (such as land area, building size, and location) and their corresponding market prices, but that other latent variables not included in the model—such as property age, neighborhood socioeconomic status, accessibility, and infrastructure development—likely contribute to the remaining unexplained variance. While this value may seem modest compared to more complex ensemble models, it demonstrates the Decision Tree's ability to produce interpretable, transparent results while maintaining reasonable predictive power.

Overall, the quantitative analysis underscores a trade-off between model simplicity and predictive accuracy. The Decision Tree Regression model, though limited in precision, successfully identifies dominant predictors and general pricing structures within the Bandung housing market. This outcome aligns with the study's objective of developing an interpretable baseline model that prioritizes transparency over complexity, laying a solid foundation for future model optimization and integration into decision-support systems.

4.2 Interpretation of Decision Tree Models

The visualization of the Decision Tree model (Figure 2) provides a clear and interpretable representation of how the algorithm arrives at its predictions for house prices. Each node within the tree is color-coded to reflect the magnitude of the predicted price at that point. Nodes displayed in lighter shades, close to white or pale tones, indicate groups of properties with relatively low predicted average prices—typically around IDR 2.5 billion in this dataset. In contrast, nodes shaded in darker orange or deeper hues correspond to higher predicted price levels, such as IDR 16.5 billion. This color gradient enables users to visually identify which sections of the tree represent higher or lower value properties, enhancing the interpretability of the model's decision patterns.

The decision structure of the tree further illustrates how the model classifies and differentiates properties based on their features. At the root node, the first and most influential decision criterion is the variable land_area (m^2) $\leq$ 220.0, confirming that land area is the dominant factor affecting property value in this dataset. Properties with land areas less than or equal to 220 m^2 are separated from those with larger plots, forming the initial split that drives subsequent decisions. As the data flow continues through the branches and child nodes, the model applies additional criteria—such as building area, number of rooms, and location—to further refine its predictions. Each complete path from the root to a leaf node represents a distinct rule set that leads to a specific predicted price. For example, a house with a land area of 150 m^2 would follow the "True" branch under this initial rule and, depending on its other attributes, terminate at a leaf node corresponding to a precise price estimation.

Each node in the visualization also contains detailed quantitative information that deepens the understanding of the model's decision logic. The "samples" value indicates the number of data points meeting the conditions up to that node—for instance, the root node encompasses 3,829 training data samples. The "squared_error" metric reflects the variability of predicted prices within the node, where a smaller value suggests that the properties grouped there have similar price levels. Finally, the "Price" value represents the model's average predicted price for properties in that

 ISSN 2579-7069

subset. For example, properties with land areas exceeding 220 m² have an average predicted value of approximately IDR 8.3 billion. Overall, this interpretive analysis transforms the Decision Tree model from a mere computational "black box" into a transparent, explainable decision-making framework. Stakeholders such as appraisers, investors, and real estate agents can clearly trace how each attribute contributes to the final price estimation, thereby increasing trust in the model's results and supporting more informed, data-driven decision-making processes.

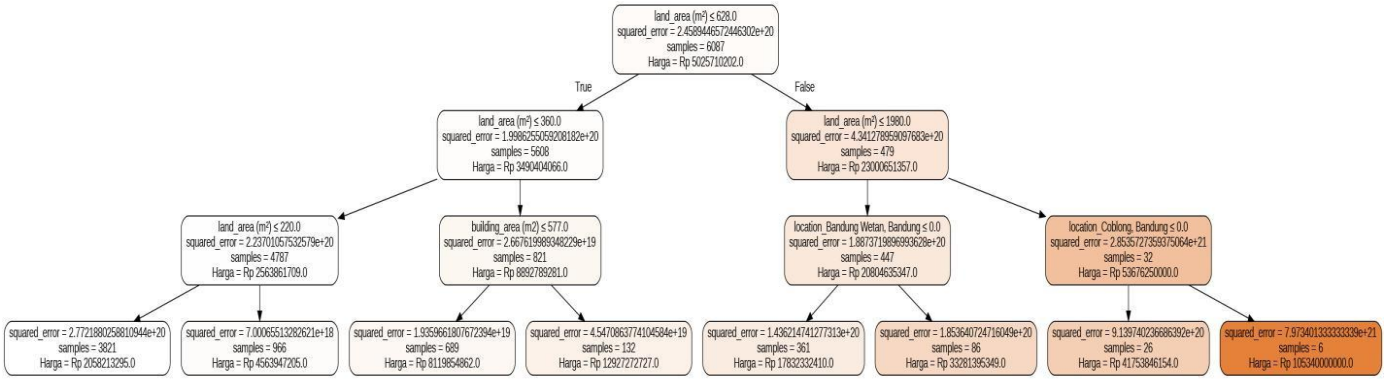

Figure 2. Visualization of Decision Tree Model

4.3 Future Research Suggestion

Based on the findings and limitations of this study, several directions for future development are proposed to enhance the performance, generalizability, and applicability of the house price prediction model. First, future research should consider utilizing more complex machine learning algorithms to improve predictive accuracy. Ensemble techniques such as Random Forest and Gradient Boosting offer promising alternatives. The Random Forest algorithm builds multiple decision trees and averages their predictions, effectively reducing overfitting and increasing model stability. Meanwhile, Gradient Boosting operates sequentially, where each new model iteratively corrects the errors of its predecessor, often leading to superior accuracy compared to single-tree approaches.

Second, implementing cross-validation techniques such as k-fold cross-validation can yield more reliable and robust performance assessments than a single train-test split. This approach ensures that the model's predictive ability is not overly dependent on how data are divided, thus improving its consistency across different samples. In parallel, hyperparameter tuning—such as adjusting parameters like maximum tree depth or minimum samples per split—can further refine model performance and optimize generalization on unseen data.

Third, enhancing the model through feature engineering is a crucial step toward greater predictive power. Introducing more context-aware variables can capture additional dimensions of property value that are not represented by the existing attributes. Examples of potential new features include the age of the building, proximity to key amenities such as schools, shopping centers, and public transportation, legal ownership status, historical regional price trends, and floor level for multi-story residences. These supplementary features would provide a more holistic representation of market dynamics and better account for value variations across neighborhoods.

A fourth avenue for development involves the real-world deployment of a Decision Support System (DSS). A web-based or mobile-based DSS could allow users—such as real estate agents, property investors, or potential buyers—to input property details and instantly obtain price estimates along with interactive visualizations of the decision tree pathway that led to the prediction. The system could also provide textual explanations (e.g., "land area > 220 m² → higher price tier") to enhance transparency and user trust. Such an application would transform the model from an analytical tool into a practical instrument for day-to-day property assessment and investment planning.

Fifth, future studies could explore advanced ensemble algorithms to further improve predictive accuracy. State-of-the-art methods such as Random Forest Regressor, Gradient Boosting, XGBoost, and LightGBM have demonstrated strong performance in various regression tasks. XGBoost and LightGBM, in particular, offer efficient computation and enhanced scalability, enabling faster training on large datasets without sacrificing accuracy.

Another promising direction is the integration of temporal and spatial data. Incorporating time-series information can enable the model to capture housing price trends over different periods, while spatial data—such as geographical information system (GIS) layers—can account for locational influences and zoning variations. Together, these additions would enable the development of a more dynamic and geographically sensitive property valuation model.

Finally, expanding the dataset scope and diversity remains a critical step for future work. The current dataset, although useful for preliminary modeling, is limited in size and variability. Collaborations with local property agencies, government institutions, or online listing platforms could help gather more comprehensive, up-to-date, and verified datasets. A richer data foundation would enhance the model's generalizability, reliability, and real-world applicability, ensuring that it can serve as a robust reference for both academic research and practical implementation in Indonesia's evolving housing market.

5. Conclusion

This study successfully implemented a Decision Tree Regressor model to predict house prices in Bandung. The model demonstrated a measurable statistical relationship between property attributes such as land area, building area, and location and the corresponding selling price. With an R^2 score of 0.385, the model has moderate predictive capability and successfully explains some of the price variation in the dataset. Its main strengths lie in its transparency and interpretability, which allow for a deeper understanding of the price drivers based on available data. This makes it particularly valuable in real estate contexts, where explainability is essential for user trust. Moreover, the interpretability of the Decision Tree allows stakeholders such as home buyers, real estate agents, or developers to understand the reasoning behind a specific price estimation. Instead of relying on opaque predictions from black-box models, users can trace each decision node and evaluate whether the estimated value aligns with their market understanding. This transparency is especially important in real estate, where decisions involve high financial stakes

6. Declarations

6.1. Author Contributions

Author Contributions: Conceptualization, A.D.U., B.H.H., and E.P.; Methodology, A.D.U. and B.H.H.; Software, E.P. and B.H.H.; Validation, B.H.H. and E.P.; Formal Analysis, A.D.U.; Investigation, E.P. and B.H.H.; Resources, B.H.H. and E.P.; Data Curation, E.P.; Writing—Original Draft Preparation, A.D.U.; Writing—Review and Editing, B.H.H. and E.P.; Visualization, B.H.H. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. L. Alfaro Navarro, E. L. Cano, E. Alfaro, N. García, M. Gámez, and B. Larraz, "A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems," Complexity, 2020, doi: 10.1155/2020/5287263.

[2] N. Arcuri, M. D. Ruggiero, F. Salvo, and R. Zinno, "Automated Valuation Methods Through the Cost Approach in a BIM and GIS Integration Framework for Smart City Appraisals," Sustainability, 2020, doi: 10.3390/su12187546.

[3] T. Bogdanova, A. Kamalova, T. Kravchenko, and A. Poltorak, "Problems of Modeling the Valuation of Residential Properties," Business Informatics, 2020, doi: 10.17323/2587-814x.2020.3.7.23.

[4] I. A. Boitan, "Residential Property Prices' Modeling: Evidence From Selected European Countries," Journal of European Real Estate Research, 2016, doi: 10.1108/jerer-01-2016-0001.

[5] J. Ćetković et al., "Assessment of the Real Estate Market Value in the European Market by Artificial Neural Networks Application," Complexity, 2018, doi: 10.1155/2018/1472957.

[6] C. Chen et al., "Experimental Research on the Impact of Interest Rate on Real Estate Market Transactions," Discrete Dynamics in Nature and Society, 2022, doi: 10.1155/2022/9946703.

[7] C. ÇILGIN and H. Gökċen, "Machine Learning Methods for Prediction Real Estate Sales Prices in Turkey," RDLC, 2023, doi: 10.7764/rdlc.22.1.163.

[8] J. Cong, "Comparative Analysis of the Real Estate Market in Different Countries," Highlights in Business Economics and Management, 2023, doi: 10.54097/hbem.v19i.11881.

[9] D. Demetriou, "A Spatially Based Artificial Neural Network Mass Valuation Model for Land Consolidation," Environment and Planning B Urban Analytics and City Science, 2016, doi: 10.1177/0265813516652115.

[10] T. Dimopoulos and N. Bakas, "Sensitivity Analysis of Machine Learning Models for the Mass Appraisal of Real Estate. Case Study of Residential Units in Nicosia, Cyprus," Remote Sensing, 2019, doi: 10.3390/rs11243047.

[11] L. Fang, "Machine Learning Models for House Price Prediction," Applied and Computational Engineering, 2023, doi: 10.54254/2755-2721/4/20230505.

[12] V. D. Giudice, P. D. Paola, F. Torrieri, P. Nijkamp, and A. Shapira, "Real Estate Investment Choices and Decision Support Systems," Sustainability, 2019, doi: 10.3390/su11113110.

[13] Y. He, "A Polynomial Linear Prediction Model for Housing Price in the USA," Applied and Computational Engineering, 2023, doi: 10.54254/2755-2721/2/20220593.

[14] H. Li, "House Price Prediction and Analysis Based on Random Forest and XGBoost Models," Highlights in Business Economics and Management, 2023, doi: 10.54097/hbem.v21i.14837.

[15] Y. Mao, Y. Duan, Y. Guo, X. Wang, and S. Gao, "A Study on the Prediction of House Price Index in First-Tier Cities in China Based on Heterogeneous Integrated Learning Model," Journal of Mathematics, 2022, doi: 10.1155/2022/2068353.

[16] N. Razali, S. Ismail, and A. Mustapha, "Machine Learning Approach for Flood Risks Prediction," Iaes International Journal of Artificial Intelligence (Ij-Ai), 2020, doi: 10.11591/ijai.v9.i1.pp73-80.

[17] T. Sadayuki, K. Harano, and F. Yamazaki, "Market Transparency and International Real Estate Investment," Journal of Property Investment & Finance, 2019, doi: 10.1108/jpif-04-2019-0043.

[18] M. Tekin and İ. U. Sarı, "Real Estate Market Price Prediction Model of Istanbul," Real Estate Management and Valuation, 2022, doi: 10.2478/remav-2022-0025.

[19] A. C. ÜZÜMCÜ and N. Eligüzel, "Predictive Analysis Using Web Scraping for the Real Estate Market in Gaziantep," Bitlis Eren Üniversitesi Fen Bilimleri Dergisi, 2023, doi: 10.17798/bitlisfen.1155725.

[20] L. Zhang, "Housing Price Prediction Using Machine Learning Algorithm," Journal of World Economy, 2023, doi: 10.56397/jwe.2023.09.03.

[21] Y. Zhang, J. Huang, J. Zhang, S. Liu, and S. Shorman, "Analysis and Prediction of Second-Hand House Price Based on Random Forest," Applied Mathematics and Nonlinear Sciences, 2022, doi: 10.2478/amns.2022.1.00052.

[22] Y. Zong, "House Prices Prediction Advanced Regression Techniques," Advances in Economics Management and Political Sciences, 2023, doi: 10.54254/2754-1169/50/20230580.

[23] C. Zou, "The House Price Prediction Using Machine Learning Algorithm: The Case of Jinan, China," Highlights in Science Engineering and Technology, 2023, doi: 10.54097/hset.v39i.6549.