

# A Machine Learning-Based Bengaluru Real Estate Valuation and Booking System with Locality Analysis and Buyer CRM

Dr. R Rajeswari  
Professor, Dept. Of CSE  
Shridevi Institute of Engineering  
and Technology,  
Tumkur, India

Bhoomika H.V  
Dept. of CSE  
SIET, Tumkur

Bushra  
Dept. of CSE  
SIET, Tumkur

Dudekula Shabana  
Dept. of CSE  
SIET, Tumkur

Krishnaveni K.A  
Dept. of CSE  
SIET, Tumkur

**Abstract**—House-price estimation is difficult because property value is influenced by several interacting factors, including locality, floor area, bedroom configuration, bathrooms, availability and property type. This paper presents an implemented Bengaluru real-estate decision-support system that combines machine-learning-based valuation with locality exploration, property verification, booking management and a Buyer Customer Relationship Management (CRM) module. The supplied Bengaluru housing dataset contains 13,320 records and nine original attributes: area type, availability, location, size, society, total square feet, bathrooms, balconies and price. The raw file contains missing values and duplicate records, motivating a preprocessing stage before model development. Two regression approaches, Linear Regression and Random Forest Regressor, are evaluated. In the implemented benchmark, 10,234 cleaned records are evaluated. Random Forest obtains an  $R^2$  of 0.7706, MAE of INR 18.35 lakh, RMSE of INR 34.24 lakh and MAPE of 19.73%, compared with  $R^2$  0.3724, MAE of INR 26.35 lakh, RMSE of INR 56.64 lakh and MAPE of 26.61% for Linear Regression. The deployed interface demonstrates a Vijayanagar valuation of INR 89.71 lakh, a displayed 90% prediction interval of INR 85.4–94.0 lakh, a locality map, verification information and an active-booking dashboard. The study therefore extends house-price prediction from an isolated regression task to an integrated Bengaluru real-estate workflow.

**Index Terms**—Bengaluru real estate, house price prediction, Random Forest, Linear Regression, machine learning, property valuation, locality analysis, booking management, Buyer CRM.

## I. INTRODUCTION

Residential property valuation is an important practical problem for buyers, sellers, developers and property-service providers. A house cannot be priced reliably using floor area alone because two properties with similar sizes can have very different values when their localities, configurations and surrounding facilities differ. Traditional estimation frequently depends on broker experience or manual comparison, which can be time-consuming and difficult to reproduce consistently. Recent machine-learning studies have consequently examined regression and ensemble methods for learning relationships between property characteristics and observed prices [1]–[4].

The project from which this study is developed focuses on Bengaluru, a large and heterogeneous urban housing market. The initial project proposal identified location, size and room configuration as important variables and planned a comparison between Linear Regression and Random Forest.

It also identified a need to consider amenities, market trends and environmental context rather than relying only on basic property attributes. The implemented system goes beyond the original prediction-only workflow by adding locality maps, verification information, booking status and Buyer CRM.

The central objective of this paper is to analyze the implemented system using the actual dataset and displayed application outputs rather than describing a hypothetical model. The supplied CSV contains 13,320 records, while the implemented benchmark reports results on 10,234 cleaned records. The distinction is important because the raw dataset contains missing values, duplicate rows and inconsistent textual representations. The paper documents the data, preprocessing, model comparison, application outputs and limitations of the current implementation.

The major contributions are: (1) a Bengaluru-focused valuation workflow based on a relatively large housing dataset; (2) an empirical comparison of a linear baseline and a nonlinear ensemble; (3) an application that presents a predicted value together with locality and property context; and (4) integration of valuation with verified booking records and Buyer CRM functions.

## II. RELATED WORK

Machine-learning-based property valuation has progressed from simple regression models to ensemble, deep-learning and multimodal approaches. Linear Regression remains useful as a baseline because its coefficients are easy to interpret, but it assumes a comparatively simple relationship between predictors and price. Tree ensembles can model nonlinear interactions and threshold effects that are common in housing data [1].

Ningsih *et al.* studied house-price prediction in the Jabodetabek area using 999 records collected through web scraping. Their comparison between Multiple Linear Regression and Random Forest [1] reported a Random Forest  $R^2$  of 0.864 after hyperparameter tuning, and identified building size, land size, bedrooms, bathrooms, garage and location as important factors [3]. Their work demonstrates the value of nonlinear ensemble learning, but it addresses a different geographic market and does not provide the integrated booking and buyer-management workflow developed here.

The second supplied reference study analyzed the Melbourne Housing dataset containing 34,857 rows and 21 features. It compared Linear Regression, Decision Tree, Random Forest and XGBoost under different experimental scenarios and reported the highest  $R^2$  of 0.7678 for XGBoost in the 80% training-data scenario [4]. That study also emphasizes data cleaning, missing-value treatment and feature selection. The present work follows the same broad principle of careful preprocessing but focuses specifically on Bengaluru and evaluates the two models that were implemented in the project.

The project literature survey also considered LSTM, multi-modal deep learning, LLM-assisted prediction, web-search trends and SHAP-based explainability. These approaches indicate that future property systems can incorporate temporal, textual, image and explanatory information. However, increasing model complexity is not automatically beneficial if the available local dataset does not support the additional variables. The present implementation therefore prioritizes a reproducible baseline-versus-ensemble comparison and a usable application layer.

#### A. Research Gap

The literature does not indicate that machine learning[3] is absent from house-price prediction; rather, the gap is in how prediction is localized and operationalized. Existing studies generally evaluate algorithms and stop at predicted values or accuracy metrics. A practical Bengaluru system can provide greater decision value when the estimate is presented together with locality information and when the workflow supports subsequent actions such as booking and buyer management.

Accordingly, this work addresses the following gap: the development of a Bengaluru-specific property valuation application that combines machine-learning prediction with locality exploration, verification-oriented information, active booking management and Buyer CRM in one interface. This is an application-level gap supported by the contrast between the prediction-focused reference studies and the implemented output analyzed in this paper.

### III. DATASET AND DATA PREPARATION

#### A. Dataset Description

The supplied dataset contains 13,320 rows and nine columns. The target variable is *price*, expressed in lakh units in the file. The input attributes are *area \_ type*, *availability*, *location*, *size*, *society*, *total \_ sqft*, *bath*, and *balcony*. The raw file contains 1,295 distinct location strings before normalization. The application itself exposes a curated set of 241 Bengaluru localities for the user interface.

Table I summarizes the dataset structure. The raw data include examples from localities such as Whitefield, Electronic City, Vijayanagar, Rajaji Nagar, Yelahanka, Jayanagar, Marathahalli and many other Bengaluru areas. The dataset also contains both numeric and categorical information, making categorical encoding and cleaning necessary before regression.

TABLE I  
 STRUCTURE OF THE SUPPLIED BENGALURU HOUSING DATASET

| Attribute    | Description                                                               |
|--------------|---------------------------------------------------------------------------|
| area_type    | Area classification such as super built-up, built-up, plot or carpet area |
| availability | Availability or possession information                                    |
| location     | Bengaluru locality                                                        |
| size         | Bedroom/room configuration                                                |
| society      | Society or project name when available                                    |
| total_sqft   | Property area; some entries are ranges or mixed text                      |
| bath         | Number of bathrooms                                                       |
| balcony      | Number of balconies                                                       |
| price        | Target property price in lakh units                                       |

#### B. Data Quality

Inspection of the supplied CSV shows 529 duplicate rows. Missing values occur in location (1), size (16), society (5,502), bath (73) and balcony (609). The society field has the largest amount of missing information and therefore requires careful treatment if it is used as a predictor. The file also contains inconsistent representations such as “2 BHK” versus “2 Bedroom” and area entries containing ranges, for example “2100 - 2850”, as well as mixed units in a small number of records. These characteristics reinforce the need for preprocessing before model training.

The project output reports that the final benchmark was evaluated on 10,234 cleaned records. Because the exact code used to produce every cleaning rule is not included in the supplied materials, this paper does not claim a particular imputation formula beyond the documented objective of cleaning missing and inconsistent records. This distinction prevents the research report from overstating undocumented implementation details.

#### C. Feature Preparation

Categorical variables need numerical representation before conventional regression models can be trained. Location is particularly important because it carries substantial spatial information, but it also has high cardinality. The project methodology therefore treats location as a major feature and prepares categorical information for modeling. Numeric fields such as area, bathrooms and balconies are retained in numerical form after cleaning. The target variable is price.

The displayed application contains additional user-facing inputs such as facing direction, floor level, furnishing and age. These fields are part of the implemented interface and feature-importance display. They should be described as model inputs only to the extent that they were included in the trained feature matrix. The original nine-column CSV does not contain all of these fields; therefore, the paper separates dataset attributes from application-level attributes.

### IV. PROPOSED SYSTEM

The implemented system is organized as a sequence of data preparation, model training, evaluation and application modules. The original project architecture specifies user input, data

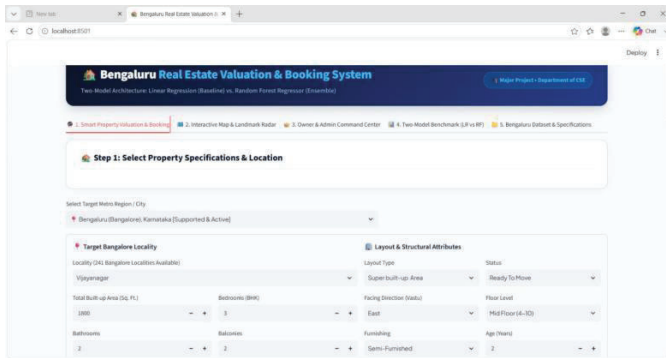


Fig. 1. Implemented property specification interface for Bengaluru valuation.

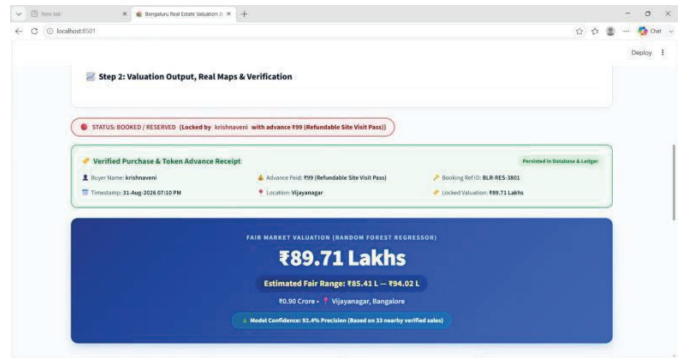


Fig. 2. Valuation output with booking status, buyer record reference and Random Forest estimate.

collection, preprocessing, feature engineering, model training, evaluation, best-model selection, prediction and user-interface presentation. The completed application adds locality maps, verification information, booking records and Buyer CRM.

#### A. Valuation Module

The valuation interface allows the user to select Bengaluru and a target locality and enter property characteristics. The supplied output demonstrates Vijayanagar as the target locality, with a 1,800 sq.ft. property, 3 BHK, 2 bathrooms, 2 balconies, semi-furnished status, East-facing orientation and a mid-floor selection. The application then produces a machine-learning valuation and related financial information.

#### B. Valuation and Locality Context

For the demonstrated Vijayanagar property, the application displays a Random Forest valuation of INR 89.71 lakh and a Linear Regression baseline of INR 91.98 lakh. It also displays a 90% prediction interval of INR 85.4–94.0 lakh, an estimated rate of INR 4,984 per sq.ft., and a monthly EMI estimate of INR 62,283 for 20 years at 8.5%. These values are application outputs and should be interpreted as estimates; the prediction interval should not be treated as statistically calibrated unless a separate calibration study is performed.

The locality module complements the numerical prediction with an interactive map and a micro-market radar. The supplied output shows Vijayanagar coordinates, a 1 km catchment zone and distances to major Bengaluru hubs. A second output demonstrates 1st Block Jayanagar with its own coordinates, metro information, technology-park proximity and distance estimates. This module addresses a practical limitation of prediction-only systems: the same predicted price is more useful when the user can understand the property's geographic context.

#### C. Property Verification

The valuation output includes a verification-oriented block containing a RERA identifier, utility information, land-title

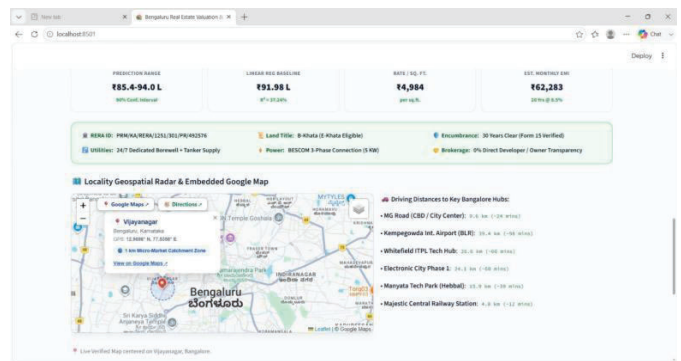


Fig. 3. Prediction interval, Linear Regression baseline, financial estimate and locality geospatial radar.

status, power connection, encumbrance status and brokerage information. These fields improve the information presented to a user, but they are not learned by the price model from the supplied CSV. In a production system, every verification field would need a trusted source, validation process and audit trail.

### V. MACHINE LEARNING MODELS AND EVALUATION

#### A. Linear Regression

Linear Regression is used as the parametric baseline. For an input vector  $x$ , the estimated price can be written as

$$\hat{y} = \theta_0 + \sum_{j=1} \theta_j x_j \quad (1)$$

where  $\theta_0$  is the intercept and  $\theta_j$  represents the contribution of the  $j$ th predictor under the linear assumption. The model

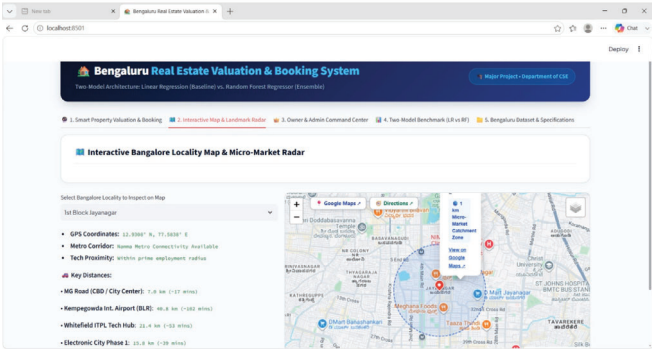


Fig. 4. Interactive Bengaluru locality map and micro-market information for 1st Block Jayanagar.

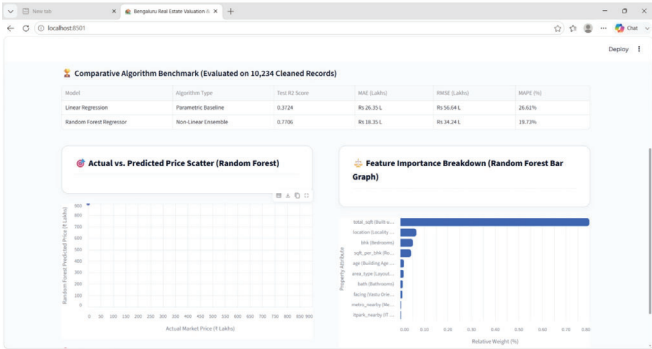


Fig. 5. Implemented two-model benchmark, actual-versus-predicted scatter and Random Forest feature-importance output.

TABLE II  
IMPLEMENTED MODEL BENCHMARK ON 10,234 CLEANED RECORDS

| Model             | $R^2$         | MAE                | RMSE               | MAPE          |
|-------------------|---------------|--------------------|--------------------|---------------|
| Linear Regression | 0.3724        | INR 26.35 L        | INR 56.64 L        | 26.61%        |
| Random Forest     | <b>0.7706</b> | <b>INR 18.35 L</b> | <b>INR 34.24 L</b> | <b>19.73%</b> |

is useful for establishing whether a more flexible algorithm provides a meaningful improvement.

B. Random Forest Regression

Random Forest is an ensemble of decision trees. Each tree produces a regression estimate, and the forest combines the individual outputs. The approach can represent nonlinear relationships and interactions without requiring a single global linear equation [1]. This characteristic is relevant to housing data, where the effect of area may differ by locality and room configuration.

C. Evaluation Metrics

The system uses four metrics. Mean Absolute Error (MAE) is

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|. \tag{2}$$

Root Mean Squared Error (RMSE) is

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}. \tag{3}$$

MAPE expresses average absolute error relative to the actual value, while  $R^2$  measures the proportion of target variance explained by the model. Using several metrics is preferable to relying on one score because RMSE, for example, is more sensitive to large errors than MAE.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

Table II shows a clear improvement from the linear baseline to Random Forest. The increase in  $R^2$  is 0.3982, while MAE decreases by INR 8.00 lakh, RMSE decreases by INR 22.40 lakh and MAPE decreases by 6.88 percentage points. This pattern indicates that the nonlinear ensemble captures relationships in the cleaned Bengaluru data that the linear model cannot represent adequately.

The actual-versus-predicted scatter shown in the project output provides a second visual check. Points closer to the ideal diagonal represent more accurate estimates, while larger deviations indicate cases where the model has difficulty. Such errors are expected in listing-style property data because unusual properties, inconsistent listings and high-value outliers can differ substantially from the majority of observations.

The feature-importance display in the output places total square feet and bedroom-related information among the prominent predictors, followed by variables such as price-per-square-foot, building age, area/layout, bathrooms, facing, metro proximity and technology-park proximity. This supports the project hypothesis that property value depends on more than a single size variable. However, feature importance from a tree model should be interpreted as model-specific importance, not as proof of causal influence.

A. Comparison With Previous Studies

The comparison in Table III is contextual rather than a claim that scores are directly interchangeable. The Jabodetabek study uses 999 records and reports Random Forest  $R^2 = 0.864$ ; the Melbourne study uses 34,857 records and reports XGBoost  $R^2 = 0.7678$  in its best scenario; the proposed system uses 13,320 raw Bengaluru records, with 10,234 cleaned records in the displayed benchmark, and reports Random Forest  $R^2 = 0.7706$ . Differences in geography, target scale, feature definitions, preprocessing and validation protocol make direct ranking inappropriate.

TABLE III  
CONTEXTUAL COMPARISON WITH SUPPLIED REFERENCE STUDIES

| Study                     | Data                                                     | Models                   | Reported best r e-<br>sult |
|---------------------------|----------------------------------------------------------|--------------------------|----------------------------|
| Ningsih <i>et al.</i> [4] | Jabodetabek; 999                                         | MLR, RF                  | RF, $R^2 = 0.864$          |
| Melbourne study [5]       | Melbourne;<br>34,857                                     | LR, DT, RF, XG-<br>Boost | XGBoost, $R^2 = 0.7678$    |
| Proposed system           | Bengaluru;<br>13,320 raw;<br>10,234 cleaned<br>benchmark | LR, RF                   | RF, $R^2 = 0.7706$         |

The proposed system is also different in scope. The reference papers concentrate primarily on prediction and algorithm evaluation. The present application connects the model output to locality exploration and transaction-oriented functions. Thus, its contribution is partly predictive and partly system-level.

VII. BOOKING MANAGEMENT AND BUYER CRM

The implemented application contains a dashboard for active verified bookings and a Buyer CRM. The output reports nine active units, a displayed portfolio value of INR 88.59 crore, an average valuation per unit and nine scheduled site visits. The CRM records include booking reference, buyer identifier, locality, configuration, advance amount, valuation, status and payment method.

For privacy in this paper, buyer names and phone numbers visible in the supplied demonstration are not reproduced. The output nevertheless demonstrates booking references, including a Vijayanagar booking with reference BLR-RES-3801, a 3 BHK 1,800 sq.ft. configuration, a displayed valuation of approximately INR 89.712 lakh and a confirmed status. Another part of the output shows a property transitioning from valuation to a booked/reserved state with a refundable site-visit pass.

This module is a meaningful system-level extension because a prediction becomes actionable when the user can proceed to a property-management step. The CRM can also support developers or administrators by keeping track of buyer interactions and booking status. However, it should be emphasized that the CRM is an application feature, not part of the machine-learning training target.

VIII. IMPLEMENTATION CONSIDERATIONS

The project identifies Python as the programming language and lists NumPy, Pandas, Matplotlib/Seaborn and Scikit-learn among the main libraries. The planned development environment includes Jupyter Notebook, VS Code or PyCharm. These tools are appropriate for a student-scale machine-learning application because they support data cleaning, model training, evaluation and visualization in a single workflow.

The implemented interface is web-based and is shown in the supplied output running locally. The system architecture separates model development from user-facing presentation. This separation is useful because a trained model can be loaded by the application without repeating the training process for every user request.

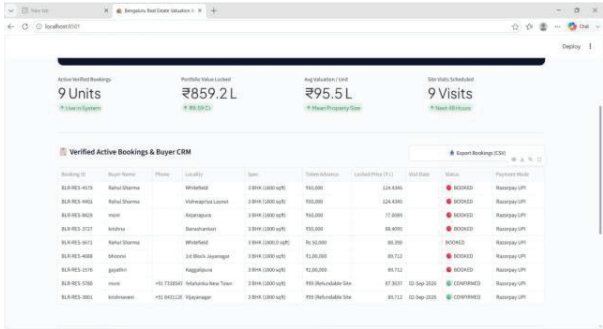


Fig. 6. Active verified bookings dashboard and Buyer CRM output from the implemented system.

For deployment beyond a prototype, model and application versions should be tracked together. Property predictions are time-sensitive, so a model trained on historical listings may degrade when the market changes. A production version would also require authentication, role-based access, audit logging, secure storage and controlled access to booking and buyer information.

IX. LIMITATIONS

Several limitations should be considered. First, the supplied dataset is listing-oriented and contains missing values, duplicates and inconsistent text. Listing prices are not necessarily identical to completed transaction prices. Second, the original dataset has only nine fields, so the application cannot automatically learn every external factor discussed in the project proposal unless those factors are added to the training data. Third, the displayed confidence percentage and prediction interval are useful application outputs but are not sufficient evidence of calibrated statistical uncertainty. Calibration should be tested on held-out data before describing such intervals as reliable probabilities. Fourth, the implemented benchmark compares only Linear Regression and Random Forest. The literature survey discusses XGBoost, LightGBM, LSTM and multimodal methods, but those algorithms were not part of the reported final benchmark. Finally, the system should not be interpreted as a replacement for professional valuation, legal verification or transaction due diligence. Its purpose is decision support based on available data.

X. FUTURE WORK

Future work can improve both the predictive and application layers. On the modeling side, hyperparameter tuning, XG-Boost and LightGBM comparisons, geographically separated validation sets and time-aware evaluation can be investigated. Adding verified amenity distances, public-transport accessibility,

environmental measures and market-trend variables would better match the research gap identified in the original project proposal.

Explainable AI is another useful extension. SHAP or related methods could be used to display how individual features change a particular prediction rather than only showing global feature importance. This would make the valuation easier for nontechnical users to understand.

For the application, booking and Buyer CRM can be strengthened with role-based permissions, secure document verification, audit logs, notification workflows and privacy-preserving storage. A periodic retraining pipeline could also detect model drift as Bengaluru's housing market changes.

#### XI. CONCLUSION

This paper analyzed an implemented Bengaluru real-estate valuation and booking system using the supplied 13,320-record dataset, project methodology and application outputs. The raw dataset contains nine attributes and several data-quality issues, while the displayed benchmark evaluates 10,234 cleaned records. Random Forest substantially outperforms the Linear Regression baseline, achieving  $R^2 = 0.7706$ , MAE of INR 18.35 lakh, RMSE of INR 34.24 lakh and MAPE of 19.73%.

The system output demonstrates that the project is broader than a conventional house-price predictor. A Vijayanagar property is valued at INR 89.71 lakh by Random Forest, the application provides a prediction range and locality context, and the completed system includes verification information, active bookings and Buyer CRM. The resulting contribution is an integrated Bengaluru real-estate decision-support prototype in

which machine-learning valuation is connected to practical property and transaction workflows. Future work should focus on richer local features, stronger validation, uncertainty calibration, explainability and secure production deployment.

#### ACKNOWLEDGMENT

The authors acknowledge the guidance of Dr. R. Rajeswari, Professor, Department of Computer Science and Engineering, SIET Tumkur and the academic support provided for the major project.

#### REFERENCES

- [1] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [2] X. Chen, Z. Zhang, and Y. Zhang, "House Price Prediction Using Machine Learning Algorithms," 2017.
- [3] I. R. Ningsih, A. Faqih, and A. R. Rinaldi, "House Price Prediction Analysis Using a Comparison of Machine Learning Algorithms in the Jabodetabek Area," *Journal of Artificial Intelligence and Engineering Applications*, vol. 4, no. 2, Feb. 2025.
- [4] N. I. Putra, N. A. Faiz, K. N. Salsabila, D. Pramudya, K. M. Runa, and M. Ridha, "Comparative Analysis of Machine Learning with Regression Algorithms in Housing Price Prediction," *Journal of AI-Driven Informatics and Management Information*, vol. 1, no. 1, pp. 34–39, Aug. 2026.
- [5] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [6] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [7] A. B. Adetunji *et al.*, "House Price Prediction using Random Forest Machine Learning Technique," *Procedia Computer Science*, pp. 806–813, 2022.