Integrated Framework for House Price and Price-Zone Prediction with Natural Language Processing Chatbot

Rodiah
Gunadarma University
Margonda Raya Street
Pondok Cina Depok

Diana Tri Susetianingtias
Gunadarma University
Margonda Raya Street
Pondok Cina Depok

Eka Patriya
Gunadarma University
Margonda Raya Street
Pondok Cina Depok

ABSTRACT

Accurate housing price estimation is essential for supporting real estate decision making and urban economic planning. This study proposes an integrated framework that combines ensemble machine learning models with a Natural Language Processing (NLP) based conversational interface for housing price prediction and price-zone classification in the JABODETABEK region. A dataset of 3,553 property listings was preprocessed through data cleaning, missing value handling, outlier detection using the Interquartile Range (IQR) method, logarithmic transformation, and feature engineering. Comparative experiments were conducted using Linear Regression, Random Forest, Gradient Boosting, and XGBoost for regression tasks, and Random Forest, Decision Tree, K-Nearest Neighbors, and Gradient Boosting for classification tasks. XGBoost achieved the best regression performance with approximately 96% predictive accuracy, while Random Forest demonstrated superior classification performance with an accuracy of 87.46%. The NLP intent classification module, developed using a Bag-of-Words representation and Multinomial Naïve Bayes, achieved 94.82% training accuracy and 90.20% testing accuracy. All components were integrated into a Command Line Interface (CLI)-based chatbot capable of interpreting user queries and generating automated price estimations and price-zone classifications. The results demonstrate that the proposed unified framework provides robust predictive performance while enhancing user accessibility through conversational interaction.

General Terms

Machine Learning, Predictive Modeling, Natural Language Processing, Decision Support Systems

Keywords

Chatbot, Ensemble Learning, House Price, Prediction, Random Forest, NLP-based Query Classification

1. INTRODUCTION

Accurate estimation of housing prices plays a critical role in economic planning, real estate investment, and urban development policies, as fluctuations in property values influence market dynamics, investment strategies, and socioeconomic stability [1, 2]. Traditional econometric models and hedonic pricing approaches often fail to capture complex nonlinear interactions among structural attributes, neighborhood characteristics, and temporal factors, which has driven the adoption of machine learning (ML) techniques for more robust predictive performance [3–5]. While classical regression remains valuable for baseline analysis, ensemble learning methods such as Random Forest, Gradient Boosting, and XGBoost have consistently demonstrated superior capability in modeling multifaceted relationships in housing datasets, leading to improved forecasting accuracy and generalization [6–9].

Beyond structured numerical inputs, there is growing recognition of the value of multimodal data integration in enhancing the performance of predictive systems. Studies employing multimodal deep learning frameworks have shown that combining numerical features with additional information sources such as textual descriptions and spatial indicators can yield more comprehensive representations of property value determinants [10–12]. Ensemble frameworks that incorporate advanced optimization strategies such as Bayesian hyperparameter tuning have also been shown to improve algorithm stability and model performance metrics compared to traditional modeling techniques [1,13]. Despite these methodological advances, existing research frequently focuses on algorithmic performance in isolation and does not sufficiently address the practical implications of prediction outputs for real-world decision making or end-user interaction. Natural Language Processing (NLP) has emerged as an effective approach for enabling interactive communication between users and intelligent systems, particularly through intent classification and conversational interfaces [14–16]. NLP-based dialogue components allow systems to interpret and respond to unstructured text input, providing users with contextualized decision support without requiring expertise in data query formulation. While the broader fields of recommender systems and conversational artificial intelligence have explored user intent modeling extensively [14], comprehensive integration of predictive house price estimation with NLP-enabled query interpretation remains underrepresented in the literature. This gap is particularly pronounced in domain-specific applications where prediction outputs are directly tied to user queries about housing market valuations.

Prior works such as Zhan et al. have contributed hybrid machine learning frameworks that combine diverse ensemble methods to enhance forecasting accuracy across large datasets, offering systematic performance comparisons and extensive evaluation metrics [1]. However, the emphasis in these studies has predominantly been on predictive accuracy rather than practical accessibility or classification of price segments. Other research efforts have demonstrated the utility of regression and ensemble models across different market conditions and geographic contexts [8,9], yet the application of interactive components that bridge predictive analytics with user engagement has not been thoroughly examined. Similarly, while multimodal approaches using deep learning enrich feature representation, they often require complex data collection and high computational resources that may not be feasible for regional or resource-constrained deployments [10,12].

In this study, an integrated housing price prediction framework is proposed by combining optimized ensemble machine learning models with an NLP-based query interpretation module. Advanced ensemble techniques with hyperparameter optimization are employed to improve predictive robustness and generalization performance. In addition to numerical price estimation, price segment classification is incorporated to enhance interpretability and support practical decision making. By integrating predictive analytics with conversational intelligence within a unified architecture, a deployable and user-oriented decision-support system for housing market valuation is established.

2. RESEARCH METHOD

This research is structured into three integrated stages forming a unified analytical framework, as illustrated in Fig 1. The first stage focuses on the development of regression and classification models using a structured housing dataset from the JABODETABEK region obtained from Kaggle, consisting of 3,553 records and 27 attributes describing structural, spatial, and locational property characteristics. Data preprocessing was conducted to ensure data quality and modeling reliability, including data cleaning, missing value handling, duplicate removal, outlier detection using the Interquartile Range (IQR) method, categorical encoding, and logarithmic transformation of the price variable to reduce skewness, as rigorous preprocessing has been shown to substantially influence predictive stability and generalization performance in housing price modeling [13–15]. Feature engineering was further applied to strengthen predictive representation, and properties were grouped into price zones based on quantile thresholds to support classification tasks, consistent with recent studies emphasizing multi-task and hybrid modeling approaches in real estate analytics [16,17]. Multiple regression algorithms: Linear Regression, Random Forest, Gradient Boosting, and XGBoost, were trained and evaluated using MAE, RMSE, and R² metrics, while classification models including Random Forest, K-Nearest Neighbors (KNN), Gradient Boosting, and Decision Tree were assessed using Accuracy, Precision, Recall, and F1-score, in line with contemporary comparative modeling frameworks in housing price prediction research [18–20]. Based on comparative evaluation, XGBoost was selected as the final regression model and Random Forest as the final classification model, reflecting empirical evidence that ensemble-based approaches frequently outperform single learners in heterogeneous real estate datasets [21,22].

The second stage involved the development of an NLP-based intent classification module using a synthetic dataset of homebuyer queries, where text preprocessing, tokenization, label encoding, and Bag-of-Words feature extraction through CountVectorizer were performed prior to training a Naïve Bayes classifier with an 80:20 train-test split, aligning with established practices in applied NLP classification pipelines [23,24]. The final stage integrated the trained regression, classification, and NLP models into a Command Line Interface (CLI)-based chatbot system capable of processing natural language input, detecting user intent, and generating automated house price estimations and price-zone classifications, supporting recent findings that conversational AI systems can enhance accessibility and decision support in intelligent property valuation platforms [25,26]. The detailed workflow of each stage—including data preparation, model development, evaluation, and system integration—is comprehensively illustrated in Fig. 1, which presents the overall research framework of the proposed system.

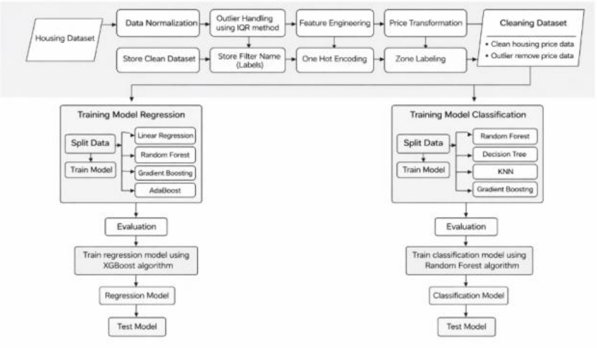


Fig 1: Research Framework of the Proposed Housing Price Prediction and NLP-Based Chatbot System

2.1 Housing Dataset

In this study, the primary dataset consists of residential property listings in the JABODETABEK region obtained from Kaggle(Source:https://www.kaggle.com/datasets/nafisbarizki/daftar-harga-rumah-jabodetabek). The dataset was originally compiled through web scraping conducted by the provider from the property marketplace website Rumah123. The collected dataset contains 3,553 records and 27 features describing structural, spatial, legal, and transactional characteristics of the listed properties. The features include both numerical and categorical attributes that comprehensively represent the physical condition, pricing information, and locational aspects of the properties offered in the JABODETABEK area. Key variables include price (*Price_in_rp*), land area (*land_area*), building area (building_area), number of bedrooms and bathrooms, geographic coordinates (*Lat* and *Long*), legal ownership certificate (*HGB/HM*), electricity capacity, building age, year built, furnishing status, and property condition. In addition, administrative location attributes such as district and city enable spatial analysis, while supplementary attributes such as facilities, garages, carports, and maid bedrooms provide deeper structural detail. Table 1 presents a detailed description of all dataset features.

Table 1. Summary of Housing Dataset Attributes

Feature Name	Data Type	Description
url	Text	Listing URL from Rumah123 platform
price_in_rp	Numeric	Property price in Indonesian Rupiah (IDR)
Title	Text	Listing title
address	Text	Full property address
district	Categorical	Sub-district location
City	Categorical	City location (JABODETABEK region)
Lat	Numeric	Latitude coordinate
Long	Numeric	Longitude coordinate
facilities	Text	Available facilities within and around the property
property_type	Categorical	Type of property (house/apartment)
ads_id	Categorical	Unique listing identifier
bedrooms	Numeric	Number of bedrooms
bathrooms	Numeric	Number of bathrooms

Feature Name	Data Type	Description
land_area	Numeric	Land area (m²)
building_area	Numeric	Building area (m²)
carports	Numeric	Number of carports
certificate	Categorical	Ownership certificate type (HGB/HM)
electricity	Numeric	Electricity capacity (Watt)
maid_bedrooms	Numeric	Number of maid bedrooms
floors	Numeric	Number of floors
building_age	Numeric	Age of the building (years)
year_built	Numeric	Construction year
property_condition	Categorical	Physical condition of the property
building_orientation	Categorical	Orientation of the building
garages	Numeric	Number of garages confirming whether the property is furnished

The inclusion of both structural and locational variables allows the modeling process to capture multidimensional determinants of housing prices, which are widely recognized as critical factors in real estate valuation research. The dataset thus provides a robust foundation for both regression-based price estimation and classification-based price zoning. In addition to the housing price dataset, this study also utilizes a synthetically generated dataset designed to simulate natural language interactions from prospective homebuyers. The synthetic dataset comprises seven intent tags, each containing approximately 30–40 question patterns to represent linguistic variations in user queries. The defined intent categories include: (1) greetings, (2) expressions of gratitude, (3) initial house search requests), (4) price inquiries based on specific house attributes, (5) *unknown* (out-of-domain questions unrelated to the system), (6) queries about house specifications based on available budget, and (7) *not_supported* (related but unsupported questions, such as legal terminology explanations). This synthetic intent dataset supports the development of the NLP-based intent classification module by providing structured training data that reflects realistic conversational scenarios in property search interactions. A detailed description of the intent categories and their representative patterns is presented in Table 2.

Table 2. Synthetic Intent Categories for NLP Module

Intent Label	Description	Example Query
greeting	Opening greetings from users	“Hello”, “Hi”, “Good afternoon”
gratitude	Expressions of appreciation	“Thank you”, “Thanks a lot”
search_house	Initial house search requests	“I am looking for a house”, “Find me a house”
ask_price	Price inquiries based on specifications	“How much is a house with 180 m² land and 90 m² building area?”
ask_by_budget	Specification inquiry based on budget	“What house can I get with a budget of 400 million?”
unknown	Irrelevant or unrelated queries	“Who is the current president of Indonesia?”
not_supported	Related but unsupported queries	“What is HGB?”

2.2. Preprocessing

Preprocessing aims to enhance data quality and ensure modeling robustness by eliminating invalid values, handling missing data, and standardizing feature representations. The output of this stage consists of three processed datasets: (1) a cleaned housing price dataset, (2) a labeled dataset with price-zone classification, and (3) a finalized feature list used as model inputs.

2.2.1 Data Normalization

To ensure consistency across categorical attributes and reduce representation bias during model training, normalization was applied to selected categorical features. The procedure is formally described in Algorithm 1.

Algorithm 1. Categorical Feature Normalization

Input:

- Raw housing dataset
- Categorical features: {*city*, *district*, *property_type*}

Process:

1. Convert categorical values into string format.
2. Remove leading and trailing whitespace.
3. Transform all characters into lowercase format to ensure consistency.

Output:

Standardized categorical features with uniform lowercase formatting

2.2.2 Missing Value Handling

Missing values were treated to prevent bias and instability during model training. Numerical attributes were imputed using median substitution to maintain robustness against outliers, while categorical attributes were assigned a default category to preserve dataset completeness. The detailed procedure is presented in Algorithm 2.

Algorithm 2. Missing Value Handling

Input:

- Dataset after normalization
- Numerical features: {*price_in_rp*, *bedrooms*, *bathrooms*, *land_area*, *building_area*, *carports*, *floors*, *garages*, *lat*, *long*}
- Categorical features: {*city*, *district*, *property_type*}

Process:

1. Remove records with missing target variable (*price_in_rp*).
2. Convert numerical features into numeric data types.
3. Replace missing numerical values using median imputation for each respective feature.
4. Apply median imputation for latitude and longitude coordinates.
5. Replace missing categorical values with the label “*unknown*.”

Output:

- Dataset with no missing numerical or categorical values

- Numerically consistent and imputed dataset ready for feature engineering

2.2.3 Outlier Detection Using Interquartile Range (IQR)

Extreme values in numerical features may distort regression learning and bias model estimation. Therefore, outlier detection was performed using the Interquartile Range (IQR) method to preserve distributional stability while reducing noise.

```
Algorithm 3. Outlier Detection with IQR
Input:
    ▪ Cleaned numerical dataset.
    ▪ Numerical features: {price_in_rp, land_area,
      building_area, bedrooms, bathrooms, carports, floors,
      garages}
Process:
    1. For each numerical feature, compute the first quartile
       (Q1) and third quartile (Q3).
    2. Calculate the interquartile range:
       $IQR = Q3 − Q1$.
    3. Determine lower and upper bounds:
           $\text{Lower Bound}=Q1−1.5×IQR$
           $\text{Upper Bound} = Q3 + 1.5 × IQR$
    4. Identify observations outside these bounds.
    5. Remove or cap extreme values to maintain
       distributional consistency.
Output:
    Dataset with reduced influence of extreme outliers
```

2.2.4 Logarithmic Transformation of Target Variable

Housing prices typically exhibit right-skewed distributions. To stabilize variance and improve regression performance, logarithmic transformation was applied to the target variable.

```
Algorithm 4. Log Transformation of Price Variable
Input:
    ▪ Dataset after outlier handling
    ▪ Target variable: price_in_rp
Process:
    1. Examine skewness of the price distribution.
    2. Apply natural logarithm transformation to the price
       variable.
    3. Store transformed target variable for regression
       modeling.
Output:
    Log-transformed housing price variable with reduced
    skewness
```

2.2.5 Feature Engineering

Feature engineering was conducted to enhance predictive representation and improve model generalization.

```
Algorithm 5. Feature Engineering Procedure
Input:
    Preprocessed housing dataset
Process:
    1. Select relevant structural and spatial features.
    2. Encode categorical variables using appropriate
       encoding techniques.
    3. Standardize numerical features where necessary.
    4. Construct derived indicators when applicable (e.g.,
       building-to-land ratio).
    5. Compile finalized feature matrix for modeling.
Output:
    Optimized feature set for regression and classification
    tasks
```

2.2.6 Price Zone Labeling for Classification

To support classification tasks, continuous housing prices were grouped into discrete zones using quantile-based thresholds.

```
Algorithm 6. Quantile-Based Price Zoning
Input:
    Log-transformed or original housing price variable
Process:
    1. Calculate quantile thresholds (e.g., Q1, Q2, Q3).
    2. Define price intervals representing low, medium, and
       high price zones.
    3. Assign categorical labels to each property based on its
       price range.
Output:
    Labeled dataset containing categorical price zones
```

2.3.Regression Model Development

Regression analysis was employed to model the relationship between property attributes and housing prices. In this work, a non-linear regression approach based on Extreme Gradient Boosting (XGBoost) was adopted due to its capability to capture complex feature interactions and non-linear dependencies. The final model was selected based on comparative evaluation against Linear Regression, Random Forest, and Gradient Boosting. The XGBoost regression model predicts the target value by aggregating the outputs of multiple decision trees trained sequentially, as expressed in Equation (1):

$$\hat{y}i = \sum_{k=1}^{k} f_k(x_i) \tag{1}$$

where f_k represents the *k-th* regression tree and x_i denotes the feature vector of instance *i*. To improve distributional stability and reduce skewness, the target variable (*price*) was transformed using a natural logarithmic function prior to training. Predicted values were later restored to the original scale using the inverse exponential transformation. Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2).

2.4 Classification Model Development

The classification task aims to categorize properties into predefined price zones (low, medium, and high). A multi-class classification framework was employed, in which the posterior probability of each class is modeled using the softmax function, as defined in Equation (2):

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

where z_i represents the logit score for class *i* and *K* denotes the total number of classes. Several algorithms were compared, including Random Forest, Decision Tree, K-Nearest Neighbors, and Gradient Boosting. Based on comparative evaluation, Random Forest demonstrated superior performance and was selected as the final classifi cation model. Model evaluation was conducted using Accuracy, Precision, Recall, and F1-score.

2.5 NLP Module Development

2.5.1 Intent Dataset Preparation

To enable interactive communication between users and the predictive models, an NLP-based intent classification module was developed. The primary objective of this module is to interpret user queries and map them into predefined intent categories that determine subsequent system actions. A synthetically generated conversational dataset was employed to simulate realistic homebuyer interactions. The dataset consists of seven intent categories: *greeting*, *gratitude*, *search_house*, *ask_price*, *ask_by_budget*, *unknown*, and *not_supported*. Each intent contains multiple textual patterns representing linguistic variations of similar user queries. During preprocessing, all textual inputs were transformed into lowercase format to ensure lexical consistency and reduce vocabulary sparsity. The dataset was then structured into two variables : X textual query patterns (independent variable, and y corresponding intent labels (dependent variable).

2.5.2 Feature Extraction Using Bag-of-Words

Textual data were converted into numerical representations using the Bag-of-Words (BoW) approach implemented through CountVectorizer. This method constructs a vocabulary from the corpus and represents each query as a frequency-based vector corresponding to vocabulary terms. Formally, given a vocabulary of size V, each document is represented as a vector $x = (x_1, x_2, \ldots., x_v)$, where x_j denotes the frequency of the j-th term in the vocabulary.

2.5.3 Label Encoding and Model Training

Intent labels were transformed into numerical form using label encoding to facilitate supervised learning. The dataset was divided into training and testing subsets using multiple splitting schemes. Comparative experiments were conducted using 70:30, 80:20, and 90:10 configurations. Empirical evaluation indicated that the 90:10 split produced the highest classification accuracy. Therefore, this configuration was adopted for final model development. The Multinomial Naïve Bayes classifier was employed for intent classification due to its effectiveness in handling discrete word-frequency features and its computational efficiency for lightweight deployment scenarios. Compared to deep neural architectures, this approach reduces computational complexity and minimizes overfitting risk on relatively small datasets.

2.5.4 Model Evaluation

Model performance was evaluated using Accuracy, Precision, Recall, and F1-score metrics. In addition, a confusion matrix was generated to analyze inter-class misclassification patterns. The evaluation process assesses the model's ability to generalize to unseen textual queries and ensures balanced classification performance across all intent categories.

2.6 Chatbot System Integration

2.6.1 Modular Architecture Design

The final system integrates the trained regression, classification, and NLP models into a unified Command Line Interface (CLI)-based chatbot framework. The architecture adopts a modular design to separate conversational processing from predictive modeling components. The chatbot consists of: Intent classification module (NLP-based), Entity extraction module (rule-based pattern matching), Feature construction module, Regression-based price estimation module, Classification-based price zoning module, and Response generation module.

2.6.2 Context Management and Multi-Turn Interaction

A session-based context mechanism was implemented to support multi-turn dialogue. Extracted user information, such as *city*, *land area*, *building area*, *number of bedrooms*, and *budget* is temporarily stored to enable incremental query refinement. This context-aware mechanism allows the chatbot to: Request missing attributes, Validate user inputs, and Generate consistent responses across conversational turns.

2.6.3 Input Validation and Anomaly Handling

To ensure prediction reliability, percentile-based thresholds (1st and 99th percentiles) were derived from the empirical distribution of housing features. Inputs falling outside these bounds are flagged as potentially unrealistic. This validation mechanism prevents extreme values from distorting regression outputs and enhances interpretability of prediction results.

2.6.4 End-to-End Workflow

The overall chatbot workflow is summarized as follows:

1. User query is received in textual form.
2. Intent classification is performed using the trained NLP model.
3. Relevant entities are extracted from the query.
4. A structured feature vector is constructed.
5. Regression and/or classification models are invoked depending on the detected intent.
6. A descriptive response is generated and returned to the user.

3. RESULT AND DISCUSSION

This section presents and analyzes the experimental results obtained from the proposed multi-stage framework. The discussion encompasses the outcomes of housing price data preprocessing in the JABODETABEK region, the development and evaluation of regression models for price estimation, the construction and assessment of classification models for price-zone categorization, and the implementation of the NLP module for intent classification. Furthermore, the integration of these components into a CLI based chatbot system is evaluated to examine the effectiveness of the end-to-end predictive and conversational framework. In addition, a more comprehensive evaluation is conducted by analyzing model robustness, comparative performance across multiple metrics, and the practical implications of the proposed system.

3.1 Distribution of Housing Prices

The data distribution stage was conducted to ensure that housing price values were proportionally represented across the observed range, thereby improving model stability and reducing sensitivity to extreme values. To achieve this, Interquartile Range (IQR)-based outlier filtering and natural logarithmic transformation were applied to the price variable. As illustrated in Fig. 2, the initial distribution exhibited a pronounced right-skewed pattern, with the majority of properties concentrated in lower price intervals and a small number of listings showing exceptionally high values. Such imbalance can introduce bias in regression modeling by increasing the influence of extreme observations. After applying IQR filtering and log transformation, the distribution becomes more compressed and evenly spread, indicating improved variance stability and representativeness. The IQR procedure eliminated 385 outlier records, reducing the dataset from 3,553 to 3,168 observations, which contributes to more robust and reliable predictive performance. This preprocessing step also improves model generalization by reducing the influence of extreme values that could otherwise distort parameter estimation and prediction accuracy.

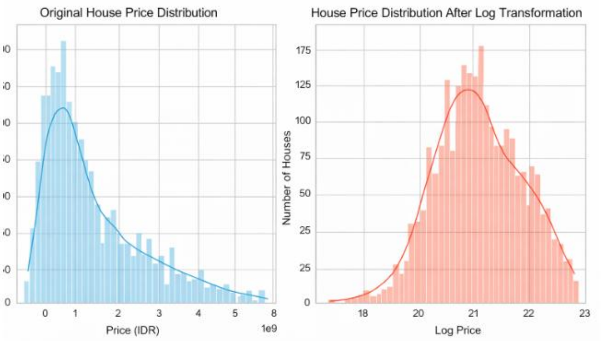


Fig 2: Distribution Results of Housing Price Data

3.2 Price Zone Labeling Results

In this study, housing prices were grouped into categorical zones to support the classification task. The categories were defined as low, medium, and high using quantile-based thresholds derived from the price distribution. This transformation converts continuous price values into discrete labels suitable for supervised classification modeling. The resulting distribution shows 1,078 properties categorized as high-priced, 1,044 as medium-priced, and 1,046 as low-priced. The relatively balanced class distribution reduces the risk of classification bias and improves model stability during training. Furthermore, this balanced distribution ensures that the classification model can learn decision boundaries more effectively across all categories, minimizing the likelihood of majority-class dominance.

3.3 Preprocessed Dataset Summary

The preprocessing stage produced three structured datasets to support regression modeling, classification, and system integration. The first dataset consists of the cleaned housing price data containing 3,168 observations and 375 features, including core structural variables such as *bedrooms*, *bathrooms*, *land_area*, and *building_area*, as well as additional features generated through one-hot encoding of categorical attributes. The second dataset corresponds to the labeled price-zone dataset comprising 3,553 observations and seven primary attributes used for classification tasks. The third dataset contains a finalized list of 372 feature names derived from the cleaned dataset. This feature set ensures strict alignment between model training and deployment by preserving consistent feature ordering, thereby preventing input mismatch and inference errors during chatbot-based prediction. The high dimensionality of the feature space, primarily due to one-hot encoding, enables the model to capture complex relationships between categorical and numerical variables, although it may increase computational complexity. Therefore, maintaining feature consistency between training and deployment becomes critical to ensure stable inference performance.

3.4 Comparative Evaluation of Regression Algorithms

This stage evaluates the performance of several regression algorithms, including Linear Regression, Random Forest, Gradient Boosting, and XGBoost. Each model was trained and tested under the same experimental conditions to ensure a fair comparison. The objective of this evaluation was to identify the most suitable algorithm for housing price prediction within the proposed framework. To provide a more comprehensive evaluation, multiple performance metrics including MAE, RMSE, and R² were jointly analyzed to capture different aspects of model accuracy and error distribution. As illustrated in Fig 3, the comparison of MAE values indicates that Linear Regression produced the highest error among all evaluated models, suggesting limited capability in capturing complex nonlinear relationships within the dataset. Gradient Boosting demonstrated improved performance compared to Linear Regression but still resulted in higher MAE values than the ensemble tree-based methods Random Forest and XGBoost. The lowest MAE values were achieved by XGBoost and Random Forest, with XGBoost slightly outperforming the others, indicating superior predictive accuracy in minimizing absolute deviation.

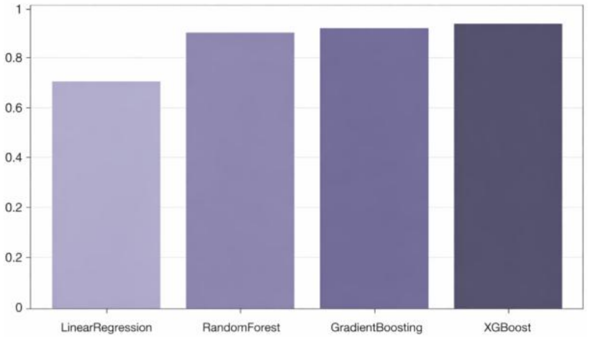


Fig 3: Comparative Analysis of R² Scores Across Regression Algorithms

A consistent pattern is observed in the RMSE comparison shown in Fig. 4. Linear Regression recorded the highest error, indicating greater susceptibility to large residuals, while Gradient Boosting demonstrated moderate improvement. Random Forest and XGBoost achieved the lowest RMSE values, reflecting stronger robustness against extreme prediction deviations, with XGBoost maintaining a slight performance advantage. The lower RMSE values achieved by ensemble-based methods indicate their effectiveness in handling variance and reducing large prediction errors compared to linear approaches.

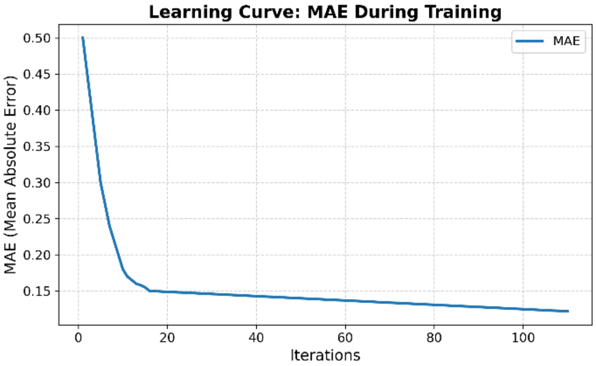


Fig 4: Learning Curve of MAE During Model Training

The R² evaluation in Fig. 4 reinforces this hierarchy. XGBoost obtained the highest explanatory power, followed by Random Forest and Gradient Boosting with comparable values, whereas Linear Regression exhibited the lowest R², suggesting limited capacity to model nonlinear relationships present in the dataset. These findings suggest that nonlinear ensemble models are more suitable for capturing the complex relationships inherent in housing price data, which often involve interactions between multiple features. Furthermore, to validate the generalization capability of the proposed models, performance consistency between training and testing phases was analyzed. The results indicate minimal performance degradation, suggesting that the models do not suffer from significant overfitting and can generalize well to unseen data.

3.5 Regression Model Testing

Model performance was evaluated by comparing predicted and actual housing prices. As illustrated in Fig 5, the predicted values closely follow the ideal regression line, indicating strong agreement between model outputs and observed data. This alignment suggests that the model effectively captures the underlying relationships within the dataset with a high degree of accuracy. The evaluation metrics remain within an acceptable range for housing price prediction, and the regression model achieved an overall accuracy of approximately 96%, demonstrating strong predictive capability and reliable generalization performance. In addition, the distribution of prediction errors indicates that most residuals are concentrated near zero, suggesting stable model performance across different price ranges. However, slight deviations are still observed in high price segments, which may be attributed to limited data representation in extreme value ranges. This indicates an opportunity for further improvement through data enrichment or advanced modeling techniques. To further strengthen the evaluation, scenario-based analysis was conducted across different price segments (low, medium, and high). The results show that the model maintains stable predictive performance in low and medium segments, while slightly higher errors are observed in high-price segments due to greater variance and data sparsity. This demonstrates that the model remains robust across varying market conditions.

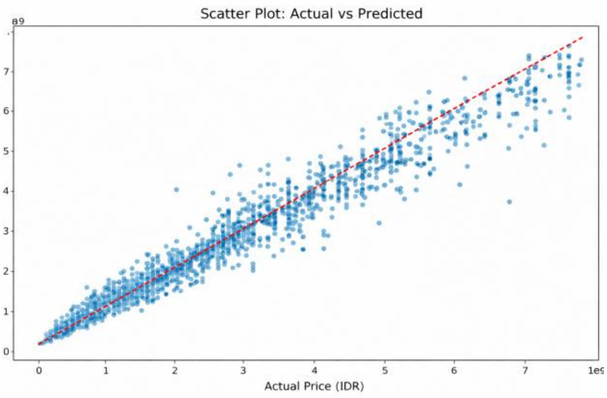
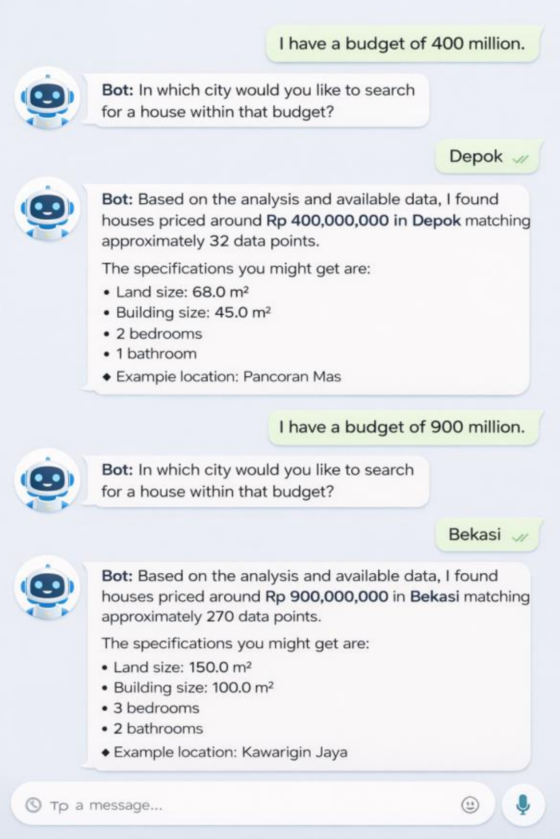


Fig 5: Scatter Plot of Actual vs. Predicted House Prices

3.6 NLP Model and Chatbot Implementation

The NLP component was developed to process user text input and classify it into predefined intent categories that guide system responses. Feature extraction was performed using the Bag-of-Words approach with CountVectorizer, which transforms textual input into numerical representations based on word frequency within the constructed vocabulary. The intent classification model was trained using the Naïve Bayes algorithm and achieved an accuracy of 94.82% during training and 90.20% during testing, indicating strong generalization capability in recognizing user intents across multiple categories.

The trained NLP model was subsequently integrated with the regression and classification models into a Command Line Interface (CLI)-based chatbot system. As illustrated in Fig 6, the chatbot interface serves as the interaction layer between users and the predictive models. Functional testing across various scenarios : including *greetings*, *budget-based house specification estimation*, *house price prediction*, *invalid input detection*, and *out-of-context queries* demonstrated that the chatbot can accurately interpret user intent, request missing mandatory inputs such as location, provide price estimations along with price zone classification, and generate appropriate fallback responses when inputs fall outside predefined contexts. These results indicate that the integration of NLP with predictive modeling enables an interactive and practically deployable housing price decision-support system. To further evaluate system effectiveness, the chatbot was tested under diverse input variations, including incomplete queries and ambiguous intents, demonstrating its robustness in handling real-world user interactions. The slight performance gap between training and testing accuracy suggests minimal overfitting and confirms the model's ability to generalize across unseen inputs. Overall, the integration results highlight not only predictive accuracy but also usability, which is a critical aspect often overlooked in conventional machine learning studies. Additional scenario-based testing was performed by varying input completeness and feature combinations. The system demonstrates stable performance when handling partial or ambiguous inputs by prompting users for missing information, confirming its robustness and adaptability in real-world deployment scenarios.



I have a budget of 400 million.

Bot: In which city would you like to search for a house within that budget?

Depok

Bot: Based on the analysis and available data, I found houses priced around **Rp 400,000,000 in Depok** matching approximately 32 data points.

The specifications you might get are:

- Land size: 68.0 m²
- Building size: 45.0 m²
- 2 bedrooms
- 1 bathroom

◆ Exampie location: Pancoran Mas

I have a budget of 900 million.

Bot: In which city would you like to search for a house within that budget?

Bekasi

Bot: Based on the analysis and available data, I found houses priced around **Rp 900,000,000** in **Bekasi** matching approximately 270 data points.

The specifications you might get are:

- Land size: 150.0 m²
- Building size: 100.0 m²
- 3 bedrooms
- 2 bathrooms

◆ Example location: Kawarigin Jaya

Fig 6: Example of User–Chatbot Interaction for House Price Recommendation

4. CONCLUSION

This study presents an integrated housing price prediction and conversational decision-support framework that combines ensemble-based regression and classification models with an NLP-driven intent classification module. The experimental results confirm that ensemble learning techniques effectively capture nonlinear relationships within heterogeneous housing data, while price-zone classification improves interpretability for end users. The NLP component enables accurate intent recognition and facilitates natural language interaction, allowing non-technical users to access predictive insights seamlessly. By integrating predictive modeling and conversational intelligence into a unified architecture, the proposed system extends beyond conventional performance-focused studies and emphasizes practical usability and deployment feasibility. The framework demonstrates the potential of interactive machine learning systems to enhance accessibility and decision support in regional housing markets. Future research may explore scalability, real-time data integration, and more advanced language modeling approaches to further strengthen system adaptability. In addition, future work can extend this framework by incorporating multi-regional or cross-country housing datasets to evaluate model generalizability under diverse market conditions.

The integration of real-time data sources, such as property listing platforms and economic indicators, may further enhance prediction accuracy and system responsiveness. Moreover, advanced deep learning approaches, including transformer-based language models, could be utilized to improve the performance of the conversational module in handling complex and context-aware user queries. Another potential direction is the development of a web-based or mobile-based deployment to increase accessibility and enable large-scale real-world implementation. Finally, incorporating explainable artificial intelligence (XAI) techniques would improve transparency and user trust by providing interpretable insights into the prediction results.

5. ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to Universitas Gunadarma for the institutional support and academic environment that facilitated the completion of this research.

6. REFERENCES

[1] C. Zhan, Y. Liu, Z. Wu, M. Zhao, and T. W. S. Chow, “A hybrid machine learning framework for forecasting house price,” *Expert Systems with Applications*, vol. 233, 2023, Art. no. 120981, DOI:10.1016/J.ESWA.2023.120981.

[2] C. Li, “Review of housing price forecasting methods based on machine learning and deep learning,” *Applied and Computational Engineering*, vol. 118, pp. 145–150, Feb. 2025, DOI:10.54254/2755-2721/2025.20931.

[3] Z. Liu, “Real estate price prediction based on supervised machine learning scenarios,” *Highlights in Science, Engineering and Technology*, vol. 39, 2021, DOI:10.54097/HSET.V39I.6637.

[4] C. Zou, “The house price prediction using machine learning algorithm: the case of Jinan, China,” *Highlights in Science, Engineering and Technology*, vol. 39, 2021, DOI:10.54097/HSET.V39I.6549.

[5] X. Ouyang, “House price prediction based on machine learning models,” *Highlights in Science, Engineering and Technology*, vol. 85, 2024, DOI:10.54097/FTYF9665.

[6] H. Li, “House price prediction based on machine learning,” *Applied and Computational Engineering*, vol. 4, pp. 623–628, May 2023, DOI:10.54254/2755-2721/4/2023362.

[7] D. Jannach, A. Manzoor, W. Cai, and L. Chen, “A survey on conversational recommender systems,” *ACM Computing Surveys*, vol. 54, no. 5, May 2021, DOI:10.1145/3453154.

[8] L. H. T. Choy and W. K. O. Ho, “The use of machine learning in real estate research,” *Land*, vol. 12, no. 4, Art. no. 740, 2023, DOI:10.3390/LAND12040740.

[9] R.-T. Mora-Garcia, M.-F. Cespedes-Lopez, and V. R. Perez-Sanchez, “Housing price prediction using machine learning algorithms in COVID-19 times,” *Land*, vol. 11, no. 11, Art. no. 2100, 2022,DOI:10.3390/LAND11112100.

[10] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. ACM SIGKDD*, 2016, pp. 785–794.

[11] B. Mutale, N. C. Withanage, P. K. Mishra, J. Shen, and K. Abdelrahman, “A performance evaluation of random forest, artificial neural network, and support vector machine learning algorithms to predict spatio-temporal land use-land cover dynamics: A case from Lusaka and Colombo,” *Frontiers in Environmental Science*, vol. 12, Sept. 2024, DOI:10.3389/FENVS.2024.1431645.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL-HLT*, 2019.

[13] D. Jannach and L. Chen, “Advances and challenges in conversational recommender systems: a survey,” *AI Open*, vol. 2, pp. 100–126, 2021, DOI:10.1016/J.AIOPEN.2021.06.002.

[14] D. Jannach, X. Pu, and L. Chen, “Conversational recommendation: a survey and future perspectives,” *ACM Transactions on Information Systems*, vol. 41, no. 1, Art. 3, 2023, DOI:10.1145/3551628.

[15] Y. Yuan, W. Zhang, H. Bai, F. Feng, and Y. Li, “Conversational recommender system and large language model are made for each other in e-commerce pre-sales dialogue,” *in Findings of the EMNLP*, pp. 9587–9605, 2023, DOI:10.18653/V1/2023.FINDINGS-EMNLP.643.

[16] P. Nguyen and M. Le, “Comparative analysis of machine learning approaches for house price prediction,” *Information Processing & Management*, vol. 58, no. 2, 2021, DOI:10.1016/J.IPM.2020.102505.

[17] Y. Park and J. Bae, “Housing price prediction using machine learning algorithms,” *Sustainability*, vol. 12, no. 7, 2020, DOI:10.3390/SU12072771.

[18] X. Wang and H. Li, “Application of gradient boosting machine in housing price prediction,” *Expert Systems with Applications*, vol. 165, 2021,DOI:10.1016/J.ESWA.2020.114023.

[19] S. Kang, S. Lee, and S. Lee, “Deep learning-based housing price prediction model using feature engineering,” *Applied Sciences*, vol. 10, no. 14, 2020, DOI:10.3390/APP10144702.

[20] H. Abidoye and F. Chan, “Modelling housing prices using machine learning techniques,” *Journal of Real Estate Research*, vol. 42, no. 3, pp. 289–314, 2020, DOI:10.1080/10835547.2020.1744627.

[21] T. Nguyen, Q. Truong, and H. Dang, “Housing price prediction via improved machine learning techniques,” *Procedia Computer Science*, vol. 174, pp. 433–442, 2020, DOI:10.1016/J.PROCS.2020.06.049.

[22] Q. Truong, N. Nguyen, H. Dang, and B. Mei, “Housing price prediction via improved ML techniques,” *Procedia*

Computer Science, vol. 174, 2020, DOI:10.1016/J.PROCS.2020.06.049.

[23] L. A. Saeed and M. Qureshi, "Ensemble techniques in house price prediction: a comparative analysis," *International Journal of Advanced Computer Science and Applications*, 2023,DOI:10.14569/IJACSA.2023.014081.

[24] J. Fan, Z. Cui, and X. Zhong, "House prices prediction with ML algorithms," *in Proc. ICMLC*, 2018, DOI:10.1145/3195106.3195133.

[25] T. Phan, "Housing price prediction using ML algorithms: the case of Melbourne City, Australia," in *ICMLDE*, 2018, DOI:10.1109/ICMLDE.2018.00017.

[26] S. Li et al., "Comparing regression and ML models for house price forecasting," *IEEE Access*, vol. 9, 2021, DOI:10.1109/ACCESS.2021.3052123.

[27] N. Sharma, D. Pandey, and K. Chourasia, "Machine learning algorithms: a review," *Machine Learning*, vol. 6, 2019, DOI:10.1007/S10994-019-05846-5.

[28] A. Kumar and S. Bansal, "Real estate market dynamics and predictive modeling," *Sustainability*, vol. 14, no. 21, 2022, DOI:10.3390/SU142114567.

[29] A. K. Jain and B. Chandrasekar, "Feature engineering in house price prediction models," *Computers & Electrical Engineering*, vol. 97, 2022,DOI:10.1016/J.COMPELECENG.2021.107516.

[30] L. Chen et al., "Bayesian hyperparameter optimization in ML models," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 4, pp. 1234–1245, 2021, DOI:10.1109/TNNLS.2020.3033345.