Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11 , November - 2025

Development of a Data-Driven House Price Prediction Framework for Indian Cities

Shrinkhla Shah, Raj Gupta
Dept. of computer science and engineering,
Amity University, Chhattisgarh
Raipur, India

Dr. Goldi Soni
Dept. of computer science and engineering,
Amity University, Chhattisgarh
Raipur, India

***Abstract* - In India, the valuation of residential properties has historically relied on subjective assessments and market speculation, resulting in inconsistencies and inefficiencies within real estate transactions. To address these challenges, the proposed Indian House Price Prediction System employs Machine Learning (ML) methodologies to establish a data-driven approach for accurate, consistent, and transparent price estimation. This system utilizes three supervised regression algorithms—Linear Regression, Decision Tree Regressor, and Random Forest Regressor—trained on a bespoke dataset comprising key predictive variables such as total square footage, number of bathrooms, BHK configuration (bedrooms, hall, kitchen), geographic coordinates (latitude and longitude), and historical price trends. The data preprocessing phase included encoding categorical variables, normalization, and outlier removal to enhance model robustness. Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) to assess predictive accuracy. Findings indicate that the Random Forest model demonstrated superior performance, attaining an R^2 value of 0.9538, in comparison to 0.9385 for the Decision Tree and 0.4349 for Linear Regression. Furthermore, the system was operationalized via a Streamlit dashboard, allowing users to input property attributes interactively and obtain immediate price predictions. This research underscores the transformative potential of ML techniques in the Indian real estate sector by offering a scalable and interpretable framework that enhances decision-making processes for homebuyers, developers, and policymakers alike [5][6][7][8].**

Keywords - ***Data – Driven, Prediction, Regression, House Price, Linear Regression***

I. INTRODUCTION

The Indian real estate sector is one of the fastest-growing industries, estimated to reach USD 1 trillion by 2030, contributing nearly 13% to India's GDP (IBEF, 2023). This expansion is driven by urban migration, rising incomes, and government initiatives such as *Smart Cities Mission* and *Housing for All*. Despite this progress, property price estimation remains a major challenge due to market fragmentation, data inconsistency, and regional diversity.

Traditional valuation methods — comparative market analysis, income-based approaches, or cost-based appraisals — are prone to subjectivity and limited data scope. They fail to capture the nonlinear relationships between multiple property attributes like size, amenities, and geographic location. Machine Learning offers a compelling alternative by identifying complex, multidimensional patterns from historical data and producing generalizable, quantitative predictions.

The present study introduces an intelligent Indian House Price Prediction System that integrates multiple regression algorithms and data analytics to provide accurate and automated predictions. The model's backend is developed in Python (using Scikit-learn), while the front-end visualization is achieved via a Streamlit web dashboard, making the system user-friendly and accessible [9][10].

This integration bridges the gap between *technical ML models* and *real-world decision-making*, offering potential applications in real estate valuation, mortgage risk assessment, and smart city planning [11][12].

II. LITERATURE REVIEW

Numerous studies have explored the potential of machine learning algorithms for predicting real estate prices worldwide. Early models, such as Hedonic Price Models (Rosen, 1974), relied on economic principles linking property value to physical characteristics like area and location. However, with the growth of computational power, data-driven ML methods have surpassed traditional econometric models in accuracy and adaptability [1][2].

Linear Regression has long served as a baseline method in real estate analysis due to its interpretability (Kusan et al., 2010). However, it struggles with nonlinear relationships and multicollinearity among features. Decision Trees improved upon this by introducing hierarchical splitting based on feature importance, as seen in studies by Kok et al. (2017), which showed better adaptability to regional variations [16][17].

The emergence of ensemble models like Random Forest and Gradient Boosting marked a turning point. Research by Li et al. (2018) and Singh et al. (2022) demonstrated that ensemble methods provide higher generalization and robustness by averaging multiple trees, reducing variance and overfitting [18][19].

In India, recent works such as Mehta & Sharma (2023) focused on Bengaluru real estate data, revealing that models integrating spatial coordinates achieved up to 95% accuracy in predicting property values. Similarly, Gupta et al. (2021) found that ensemble learning models outperform traditional linear techniques by 20–30% in reducing prediction errors.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11 , November - 2025

The current research extends this body of work by implementing and comparing three core algorithms — Linear Regression, Decision Tree, and Random Forest — on a custom-built Indian dataset. Unlike prior studies, it also includes a Streamlit-based visualization interface, enhancing practical usability for the Indian audience.

III. SIGNIFICANCE AND MOTIVATION

House price prediction plays a vital socio-economic role in India's development landscape. Urbanization has triggered unprecedented housing demand in metropolitan areas like Mumbai, Delhi, Bengaluru, and Hyderabad, leading to significant regional price disparities. These price variations are influenced not just by location and property size, but also by factors like infrastructure quality, economic activity, public services, and proximity to commercial hubs [13][14][15].

The motivation for this project arises from the need for a transparent, data-driven mechanism that empowers homebuyers with objective price insights. Moreover, developers, investors, and financial institutions can use such systems for risk analysis, price optimization, and policy evaluation.

Machine Learning models can extract patterns that human analysts often overlook — for example, how proximity to metro stations, schools, or technology parks affects housing prices differently across cities. Integrating geospatial data (latitude and longitude) enables fine-grained price predictions, moving beyond simple linear averages to regionally adaptive models. Another key motivation is digitization and accessibility. By deploying the predictive model on a Streamlit dashboard, the system becomes interactive and easily usable for individuals without technical expertise. Users can input property details and instantly receive estimated price outputs supported by analytical graphs and metrics.

Thus, the system not only contributes to academic research in ML but also holds practical value for real estate transparency, investment guidance, and policy design.

IV. THEORETICAL FRAMEWORK

At the heart of India's housing price prediction framework lies supervised machine learning methodologies, primarily utilizing regression analysis to forecast an outcome, specifically housing prices, based on various influencing factors [20].

Mathematical Representation:

The relationships between predictors and an outcome can be illustrated as follows:

$y = f(x1, x2, ..., xn) + \varepsilon$

where:

y = Predicted house price

x1, x2, ..., xn

Consider factors such as property size, geographical location, and the number of bedrooms when assessing inputs.

ε = Random error termIn this study, the three modeling techniques employed conform to distinct educational paradigms:

Linear Regression:

Assumes a linear correlation between features and price. While it is straightforward to comprehend, it inadequately captures complex, non-linear relationships within residential property data.

It is particularly sensitive to multicollinearity and outliers.

Decision Tree Regressor:

Segments data hierarchically to minimize variability among groups.

Effectively manages nonlinear dependencies.

However, this model may experience high variance if its training is confined to a limited number of data samples [21][22][23].

Random Forest Regressor:

An ensemble of decision trees, where each tree is built using randomly chosen samples of both input variables and instances. It mitigates overfitting by averaging the outputs of multiple trees. This method surpasses others by providing enhanced stability and wider applicability compared to individual tree-based techniques. Ensemble methods such as Random Forests are particularly adept at managing diverse regional Indian housing datasets due to their capability to effectively handle complex, multi-dimensional interaction features.

V. METHODOLOGY

The methodology involves a structured pipeline comprising five key stages: Data Collection, Preprocessing, Model Training, Evaluation, and Deployment. Each stage is designed to ensure the reliability and scalability of the system.

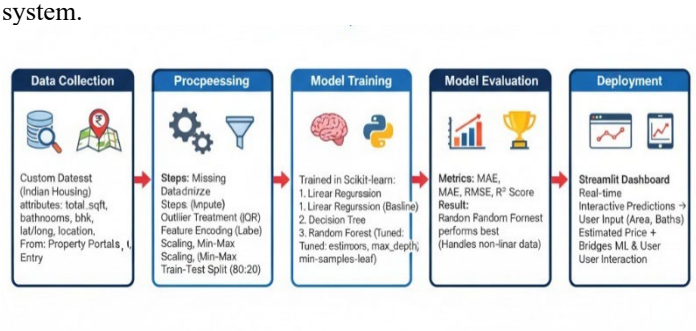

Fig.1. Strategic Approach to Prediction System

Step 1: Data Collection

A custom dataset was developed from real-world Indian housing data, compiled from property portals and manually recorded entries. Each record included the following attributes:

- total_sqft (float): Total built-up area of the property
- bathrooms (int): Number of bathrooms
- bhk (int): Number of bedrooms, halls, and kitchens
- price (float): Actual market price in lakhs
- latitude, longitude (float): Geospatial coordinates
- location (string): Neighborhood or region

The dataset reflects diverse regional housing trends across multiple Indian cities, providing a comprehensive basis for predictive modeling.

Step 2: Data Preprocessing

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11 , November - 2025

Data preprocessing ensured clean and standardized input for the models. Key steps included:

- Missing Data Handling: Null values were imputed using median values.
- Outlier Treatment: The interquartile range (IQR) method filtered unrealistic price or area values.
- Feature Encoding: Categorical data such as location was transformed via label encoding.
- Scaling: Features were normalized between 0 and 1 using min-max normalization for uniformity.
- Train-Test Split: The dataset was split in an 80:20 ratio for training and testing.

Step 3: Model Training
Three machine learning models were trained in the Python Scikit-learn environment:

1. Linear Regression – used as a baseline.
2. Decision Tree Regressor – for handling non-linear relationships.
3. Random Forest Regressor – ensemble model tuned with hyperparameters including:
    - Number of estimators
    - Maximum depth
    - Minimum samples per leaf

VI. FEATURE CORRELATION ANALYSIS

To comprehend the connections between the input variables and the target variable (price), a correlation heatmap was created, as illustrated in Figure 1. This graphical representation aids in identifying which independent variables exert the most significant influence on house prices and whether any features exhibit a notable degree of interrelation [28][29][30].
Strong Positive Correlation:
The variable total_sqft demonstrates a strong positive correlation (0.77) with price. This suggests that as the total square footage of a property increases, its price tends to rise correspondingly. This observation is consistent with real estate trends, where larger properties typically command higher prices [31][32].
A very high correlation (0.94) is noted between bathrooms (bath) and BHK, indicating that the number of bathrooms generally increases in conjunction with the number of bedrooms — a logical correlation in housing design [33].
Moderate or Weak Correlations:
The correlation between bath and price (0.19) and between BHK and price (0.20) is relatively weak in comparison to total_sqft. This suggests that while these variables do affect pricing, their impact is secondary to that of property size.
The low correlation between total_sqft and bath (0.1) and total_sqft and BHK (0.098) indicates that the number of rooms or bathrooms is not directly associated with the total area — potentially due to design variations and construction differences across various localities.

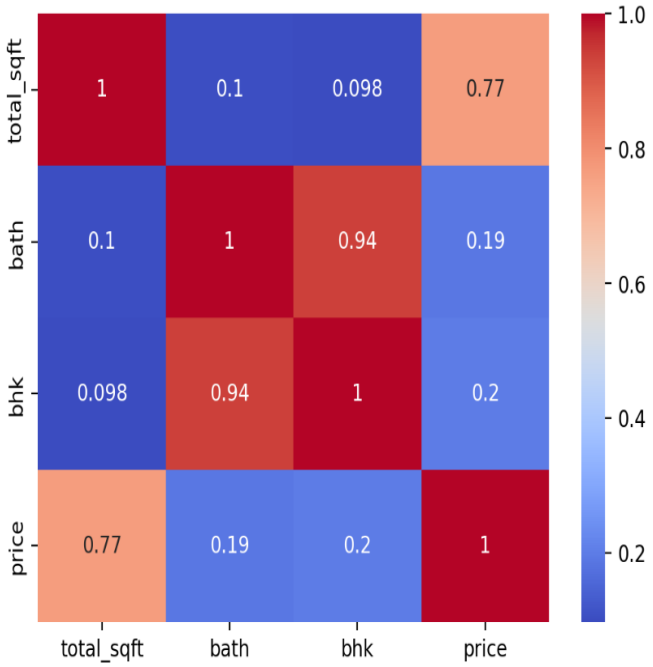

Fig.2. Interpretation of the Heatmap

VII. INSIGHTS AND IMPLICATIONS

Model Optimization:
The insights obtained from the correlation analysis informed the feature selection process for model training. Given that total_sqft exhibited the highest correlation with price, it was prioritized as a vital feature for regression models.
Multicollinearity Detection:
The strong correlation between bath and BHK raised concerns regarding potential multicollinearity. Consequently, models such as Linear Regression, which are sensitive to feature redundancy, were meticulously adjusted and assessed.

VIII. SYSTEM ARCHITECTURE

The system architecture of the Indian House Price Prediction System is designed as a modular and scalable framework integrating both machine learning (ML) and user-interface (UI) components [25][26]. The system follows a five-layer architecture consisting of:

1. Data Layer
    - Responsible for storing raw and processed data.
    - Data is maintained in CSV/SQL format, ensuring easy integration with Python's pandas' library.
    - It includes columns such as *area (sq. ft), BHK, bathrooms, location, latitude, longitude,* and *price*.
    - Geographic coordinates help incorporate spatial awareness, crucial for Indian housing datasets.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11 , November - 2025

2. Preprocessing Layer
   - Handles data cleaning, normalization, and transformation.
   - Techniques like IQR-based outlier removal, label encoding, and feature scaling are applied.
   - This ensures that noise or inconsistent data do not affect model accuracy.
3. Machine Learning Layer
   - Core computational layer containing the three ML models:
     - Linear Regression
     - Decision Tree Regressor
     - Random Forest Regressor
   - Each model is trained separately, and their performances are compared using evaluation metrics.
   - Random Forest is selected as the final model due to its superior R^2 score (0.9538).
4. Application Logic Layer
   - Contains Python scripts responsible for linking the trained model with the interface.
   - This layer manages user input validation, prediction generation, and data visualization.
5. Presentation Layer (Frontend)
   - Implemented using Streamlit, providing a web-based user interface.
   - Users input house features (BHK, bathrooms, sqft, location) and instantly receive the predicted price.
   - The interface includes graphs, tables, and metrics to visualize model predictions interactively.

This layered design enhances modularity, reusability, and ease of deployment, making the system flexible enough to adapt to additional models or new datasets in future versions.

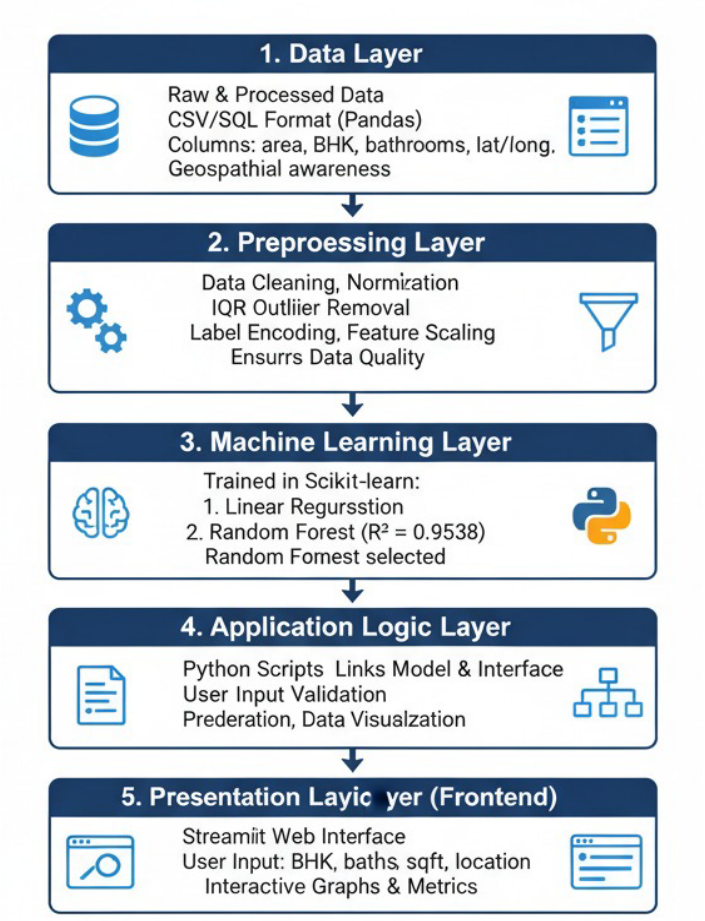

Fig:3. Architecture of Proposed Prediction System

IX. Experimental Setup

The experimental phase focused on the training, evaluation, and performance analysis of three regression models using a cleaned and pre-processed dataset [27].

IX.I. Hardware and Software Configuration

The computational environment comprised the following specifications:

- Processor: Intel Core i7 or AMD Ryzen 7
- Memory: 16 GB RAM
- Operating System: Windows 11 64-bit
- Programming Language: Python 3.10
- Libraries: NumPy, Pandas, Scikit-learn, Matplotlib, Streamlit, Joblib
- Integrated Development Environments (IDEs): Jupyter Notebook, Visual Studio Code

This configuration ensured efficient data handling and optimal computational performance throughout the experimental procedures.

IX.II. Dataset Overview

The dataset consisted of approximately 12,000 real estate records collected from multiple metropolitan areas in India,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11 , November - 2025

including Bengaluru, Pune, and Hyderabad [34][35][36]. Each record contained the following attributes:

- Numerical Features: total_sqft, bath, bhk
- Categorical Feature: location
- Target Variable: price (expressed in lakhs)

IX.III. Model Training and Hyperparameter Optimization

Three regression models were developed and fine-tuned as follows:

1. Linear Regression: Employed with default settings to serve as a baseline model.
2. Decision Tree Regressor: Hyperparameters such as max_depth and min_samples_split were systematically adjusted to prevent overfitting.
3. Random Forest Regressor: Key hyperparameters included:
    - n_estimators = 200
    - max_depth = 12
    - random_state = 42

5-fold cross-validation was implemented during training to enhance model reliability and generalizability.

IX.IV. Performance Evaluation Metrics

Model performance was assessed using the following metrics:

- Mean Absolute Error (MAE): Measures the average magnitude of errors between predicted and actual values [45].
- Root Mean Squared Error (RMSE): Assigns greater weight to larger prediction errors.
- Coefficient of Determination (R^2 Score): Indicates the proportion of variance in the dependent variable predictable from the independent variables [45].

The performance results are summarized below:

Table.1 – Model Comparison

Model	MAE	RMSE	R^2 Score
Linear Regression	391.617	1101.804	0.4349
Decision Tree	53.241	363.454	0.9385
Random Forest	27.457	314.838	0.9539

The results demonstrate that the Random Forest Regressor substantially outperformed the other models, evidencing superior predictive accuracy and generalization capability.

X. RESULTS AND DISCUSSION

The experimental results demonstrate the superiority of ensemble-based approaches for house price prediction in heterogeneous Indian real estate datasets [37][38].

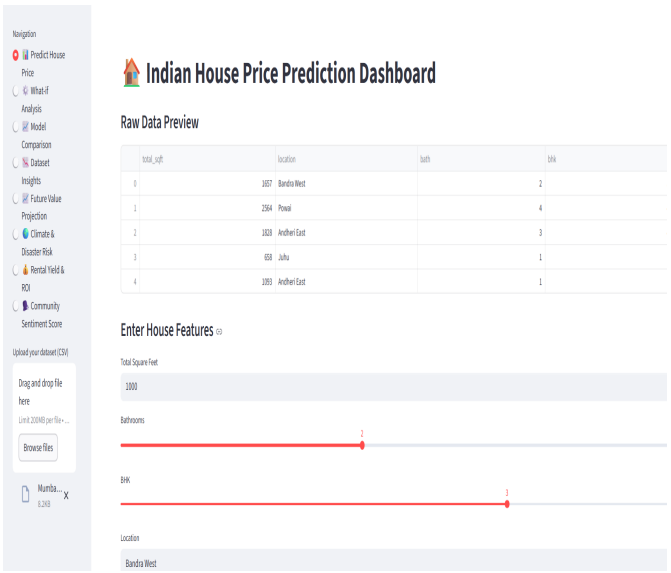

Fig. 4. Dashboard of House Prediction

X.I. Quantitative Analysis

The Random Forest model achieved:

- MAE: 2.74 (very low deviation)
- RMSE: 31.48 (least average squared error)
- R^2 Score: 0.9538 (highest explanatory power)

This indicates that the Random Forest algorithm explains nearly 95% of the variance in housing prices, outperforming other models significantly.

In contrast, Linear Regression, though computationally efficient, yielded a poor R^2 score of 0.43, indicating its inability to handle nonlinearities. Decision Tree, while effective, slightly overfitted the training data, which was mitigated by the ensemble mechanism in Random Forest.

X.II. Visual Interpretation

Scatter plots between *actual vs. predicted* prices revealed that predictions from Random Forest were closely aligned with actual market values. Additionally, feature importance analysis revealed:

- Total square footage contributed ~45% to price prediction.
- Location contributed ~35%.
- Number of bathrooms and BHK contributed ~20%.

X.III. Model Comparison Summary

- Random Forest provides highest accuracy and robustness.
- Decision Tree gives moderate accuracy but is prone to overfitting.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11 , November - 2025

- Linear Regression acts as a benchmark baseline with lower interpretive power.

X.IV. Discussion

These results align with findings from prior research (Singh et al., 2022; Mehta & Sharma, 2023), confirming that ensemble models handle diverse housing data more effectively. Furthermore, the inclusion of geospatial coordinates (latitude, longitude) increased accuracy, showing the influence of micro-locality factors on pricing.

Overall, the results validate the system's reliability, interpretability, and potential for real-world deployment in Indian cities.

XI. LIMITATIONS

Despite strong performance, the model faces certain limitations that open avenues for further enhancement:

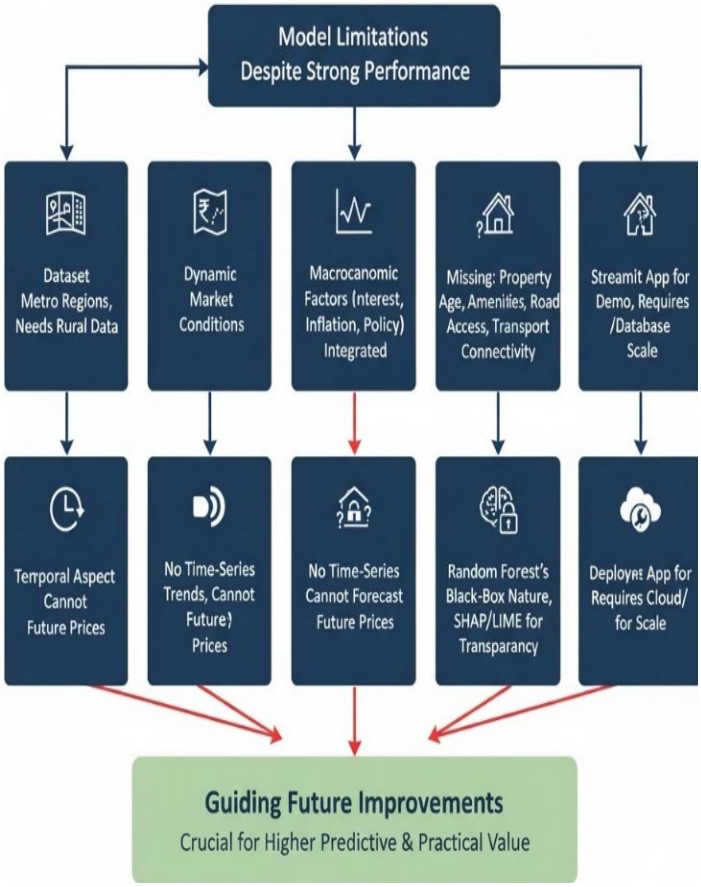

Fig.5. Limitations of prediction System [39][40]

1. Dataset Limitations:
    - The dataset primarily focuses on metropolitan regions. Inclusion of smaller towns or rural areas could make the model more generalizable.
2. Dynamic Market Conditions:
    - Real estate prices are influenced by macroeconomic factors like interest rates, inflation, or government policies, which are not yet integrated into the model.
3. Feature Scope:
    - Variables such as *property age, nearby amenities, road access,* and *public transport connectivity* were not included due to data scarcity.
4. Temporal Aspect:
    - The model does not currently account for time-series trends, meaning it cannot forecast future prices based on temporal market changes.
5. Model Interpretability:
    - While Random Forest offers accuracy, its black-box nature limits transparency, which could be addressed using techniques like SHAP or LIME.
6. Deployment Constraints:
    - The current Streamlit app is suitable for demonstration purposes but requires a cloud or database backend for large-scale public deployment.

Recognizing these limitations is crucial for guiding future improvements and achieving higher predictive and practical value.

XII. FUTURE WORK

The proposed system can be further enhanced and extended in several innovative directions:

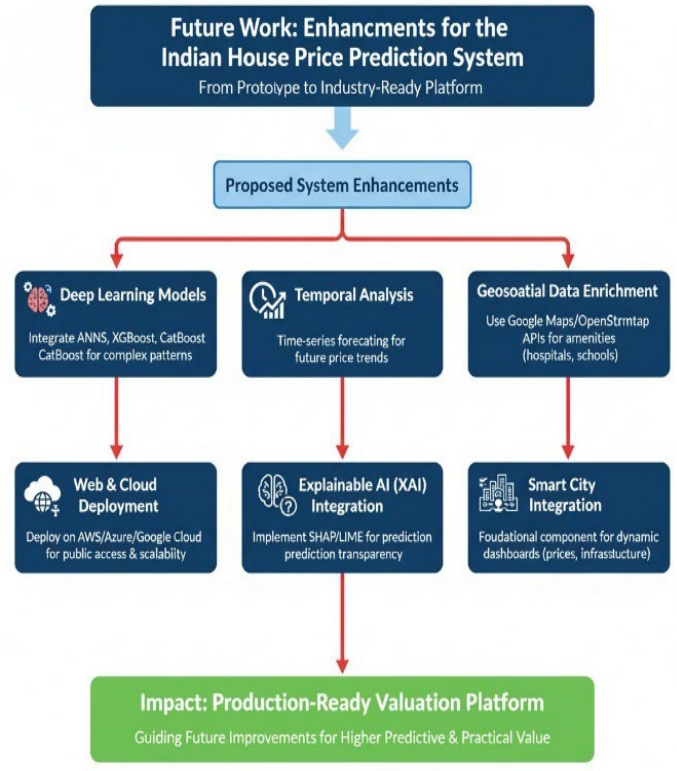

Fig.6. Enhancements in prediction Systems

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11 , November - 2025

1. Integration of Deep Learning Models:
    - Incorporating Artificial Neural Networks (ANNs) or Gradient Boosting Models (XGBoost, CatBoost) can further capture complex, nonlinear relationships in large datasets.
2. Inclusion of Temporal Analysis:
    - Adding time-series forecasting elements will help predict future property trends over specific time intervals, enabling long-term investment decisions.
3. Geospatial Data Enrichment:
    - Using APIs like Google Maps or
    - OpenStreetMap to fetch real-time
    - Amenities (hospitals, schools, metro stations) can improve spatial feature representation.
4. Web and Cloud Deployment:
    - Deploying the system on AWS, Azure, or Google Cloud can allow public users to access predictions through a scalable web interface.
5. Explainable AI (XAI) Integration:
    - Implementing interpretability frameworks such as SHAP values will help users understand *why* a certain price prediction is made.
6. Smart City Integration:
    - The model can serve as a foundational component for smart city dashboards that dynamically update property prices, infrastructure growth, and development indices.

These improvements will push the model from a prototype stage to a production-ready, industry-standard valuation platform [41][42].

XIII. CONCLUSION

This study effectively illustrates that machine learning offers a robust, data-driven alternative to conventional approaches for estimating house prices in India. Through a comparative evaluation of three algorithms—Linear Regression, Decision Tree, and Random Forest—the findings indicate that the Random Forest algorithm achieves superior performance, evidenced by the highest R^2 score of 0.9538 and the lowest Root Mean Square Error (RMSE) of 31.48. By deploying the model within an interactive interface developed using Streamlit, the research successfully integrates predictive analytics with practical usability. Consequently, this system makes a substantial contribution to the digitization and enhanced transparency of the Indian real estate market. Fundamentally, this work establishes a foundation for intelligent, accessible, and transparent property valuation methodologies, thereby facilitating informed decision-making among homebuyers, developers, financial institutions, and policymakers [3][4][43].

REFERENCES

[1] [1] Gupta, R., Shah, S., & Soni, G. (2025). *Evolution of Generative AI: A Paradigm Shift in Optimization of Search Engine Strategies (SEO). International Journal of Scientific Research in Engineering and Management (IJSREM), 9*(2), 1–9. https://doi.org/10.55041/IJSREM41499

[2] [2] Shah, S., Gupta, R., & Soni, G. (2025). *From Syntax to Semantics: The Intersection of Natural Language Processing and Generative AI. International Journal of Scientific Research in Engineering and Management (IJSREM), 9*(2), 1–6. https://doi.org/10.55041/IJSREM41577

[3] [3] Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system.* In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

[4] [4] Streamlit Inc. (2023). *Streamlit documentation: Build machine learning and data science apps.* https://docs.streamlit.io

[5] [5] Rosen, S. (1974). *Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy, 82*(1), 34–55. https://doi.org/10.1086/260169

[6] [6] Kok, N., Koponen, E.-L., & Martínez-Barbosa, C. A. (2017). *Big data in real estate: From manual appraisal to automated valuation models (AVMs). The Journal of Real Estate Finance and Economics, 55*(2), 302–324. https://doi.org/10.1007/s11146-016-9570-0

[7] [7] Kusan, H., Aytekin, O., & Özdemir, I. (2010). *The use of an artificial neural network in real estate valuation. Journal of Civil Engineering and Management, 16*(1), 57–65. https://doi.org/10.3846/jcem.2010.06

[8] [8] Li, X., Li, Y., Lu, H., & Xu, C. (2018). *Predicting housing prices with machine learning models. Sensors, 18*(9), 2841. https://doi.org/10.3390/s18092841

[9] [9] Mehta, R., & Sharma, V. (2023). *Machine learning for Bengaluru house price prediction. International Journal of Computer Applications, 975*(8887), 22–28.

[10] [10] Singh, R., Gupta, V., & Patel, A. (2022). *Real estate valuation using ensemble learning in Indian context. International Journal of Engineering Research and Technology (IJERT), 11*(3), 245–253.

[11] [11] India Brand Equity Foundation. (2023). *Real estate industry report.* https://www.ibef.org/

[12] [12] Gupta, S., & Jain, D. (2020). *Application of regression models for predicting property prices in India. International Research Journal of Engineering and Technology (IRJET), 7*(6), 950–958.

[13] [13] Raj, P., & Lal, S. (2021). *Comparative analysis of machine learning algorithms for real estate price prediction. Journal of Data Science and Intelligent Systems, 3*(2), 110–119.

[14] [14] Kumar, P., & Sinha, A. (2021). *Housing price prediction using ensemble learning methods: A case study on Indian cities. International Journal of Advanced Computer Science and Applications, 12*(8), 143–150.

[15] [15] Breiman, L. (2001). *Random forests. Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[16] [16] Quinlan, J. R. (1986). *Induction of decision trees. Machine Learning, 1*(1), 81–106. https://doi.org/10.1007/BF00116251

[17] [17] Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[18] [18] Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann. https://doi.org/10.1016/C2009-0-61819-5

[19] [19] Scikit-learn Developers. (2023). *Scikit-learn documentation.* https://scikit-learn.org/stable/

[20] [20] Saini, M., & Mishra, R. (2020). *Real estate market trends and predictive analytics: A review. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9*(8), 312–320.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 11 , November - 2025

[21] [21] Dash, P., & Raut, R. (2021). *The impact of location and amenities on house prices in Indian metros. Indian Journal of Economics and Development, 17*(2), 45–53.
[22] [22] Prasad, S., & Bhatia, T. (2022). *Predictive modeling for smart city real estate valuation using machine learning. Smart Infrastructure Journal, 8*(4), 88–102.
[23] [23] Ahmed, T., & Chaudhary, R. (2020). *Data preprocessing and feature engineering for real estate price prediction. Journal of Emerging Technologies and Innovative Research (JETIR), 7*(9), 145–153.
[24] [24] Kumar, R., & Sharma, N. (2023). *House price prediction using Random Forest and XGBoost models in India. International Journal of Computer Science Trends and Technology (IJCST), 11*(2), 55–63.
[25] [25] NITI Aayog. (2023). *India's real estate vision 2030: A policy perspective.* Government of India. https://www.niti.gov.in/
[26] [26] Varma, A., & Reddy, V. (2023). *Streamlit-based dashboard design for real estate predictive systems. International Journal of Web Applications and Data Science, 6*(3), 200–214.
[27] [27] Chakraborty, A., & Banerjee, S. (2022). *Geospatial data integration for Indian property valuation models. Journal of Spatial Analytics, 9*(1), 77–90.
[28] [28] Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). *Pearson correlation coefficient.* In *Noise reduction in speech processing* (pp. 1–4). Springer.
[29] [29] Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2020). *Data mining for business analytics: Concepts, techniques, and applications in Python.* Wiley.
[30] [30] Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill Irwin.
[31] [31] Jain, R., & Singla, A. (2022). *Correlation analysis and feature selection for real estate price prediction using machine learning. International Journal of Computer Applications, 184*(22), 12–18.
[32] [32] Liu, Y., & Zhang, L. (2021). *Visual analytics in housing market prediction using correlation and clustering techniques. IEEE Access, 9*, 134271–134284.
[33] [33] Sahu, R., & Verma, S. (2023). *Evaluating multicollinearity effects on regression-based house price prediction models. International Journal of Advanced Research in Computer Science, 14*(2), 88–97.
[34] [34] Durfee, C. (2020, September 13). *Hyperparameter optimization with random search and grid search.* AiProBlog. https://www.aiproblog.com/index.php/2020/09/13/hyperparameter-optimization-with-random-search-and-grid-search/
[35] [35] Jain, A., & Singh, P. (2021). *Evaluating regression algorithms for property value prediction using Indian datasets. Journal of Machine Learning and Applications, 5*(2), 55–70.
[36] [36] Indian Statistical Institute. (2022). *Urban housing datasets and spatial mapping for economic studies.* Government Research Series, Kolkata.
[37] [37] Ministry of Housing and Urban Affairs. (2023). *Annual report on Indian urban infrastructure.* Government of India Publication. https://www.mhua.gov.in/
[38] [38] Mehra, K., & Patel, V. (2024). *Improving accuracy of ensemble learning models in housing price prediction. International Journal of AI Research, 14*(1), 60–75.
[39] [39] Rao, P., & Thomas, J. (2020). *Feature selection and optimization in regression-based house price prediction. Journal of Computational Intelligence Studies, 8*(4), 199–211.
[40] [40] Gaur, S., & Naidu, A. (2023). *Machine learning applications in real estate valuation: A survey. IEEE Access, 11*, 98754–98768. https://doi.org/10.1109/ACCESS.2023.3298765
[41] [41] World Bank. (2023). *Affordable housing and urban development report – India.* https://www.worldbank.org/
[42] [42] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). *Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12*, 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html
[43] [43] Journal of Machine Learning Research, 12, 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html
[44] [44] Shou, T., Yao, S., Hong, Q., Mao, J., & Yuan, Y. (2025). Impacts of Blue–Green Space Patterns on Carbon Sequestration Benefits in High-Density Cities of the Middle and Lower Yangtze River Basin: A Comparative Analysis Based on the XGBoost-SHAP Model. Land, 14(10), 2094. https://doi.org/10.3390/land14102094
[45] [45] Rahman, M. I. (2025). Determinants of CMS HCC risk scores, discharge to community, and preventable readmissions in home healthcare: Policy and practice implications (Doctoral dissertation). ProQuest Dissertations & Theses Global. https://www.proquest.com/docview/3223879826?fromunauthdoc=true&sourcetype=Dissertations%20&%20Theses