



# IJRASET

International Journal For Research in  
Applied Science and Engineering Technology



---

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

---

**Volume:** 14    **Issue:** IV    **Month of publication:** April 2026

**DOI:** <https://doi.org/10.22214/ijraset.2026.80678>

[www.ijraset.com](http://www.ijraset.com)

Call:  08813907089

E-mail ID: [ijraset@gmail.com](mailto:ijraset@gmail.com)

# Real Estate Ultimate Price Prediction and Recommendation System

Yashasvi Pundir, Tushar Pundir, Rishabh, Nilansh Garg, Shivani Pandey

Dept. of Data Science Meerut Institute of Engineering and Technology, Meerut, India

**Abstract**—There has been a rapid urbanization, due to which population growth in cities has increased and will increase further in the developing cities of India. Gurgaon, one of India's fastest-growing cities, has emerged as a major real estate hub today, thanks to its proximity to the capital, growing infrastructure, corporate presence and improved connectivity, due to which real estate decision making has become very complex for buyers, investors and people working there.

This study serves as a comprehensive AI-driven framework as it provides property price predictions, market analysis, and personalized recommendations using a dataset of near 4,000 residential listings which is web-scraped from an online website 92acres.com.

This research presents an end-to-end data science pipeline that incorporates rigorous preprocessing including IQR-based outlier treatment and advanced feature engineering to extract the required features such as luxury amenities, servant room availability, floor number, and soon. To ensure high availability, a multi-model comparison was also conducted, evaluating 11 algorithms, including linear regression, random forest, and XGBoost, across various encoding strategies such as one-hot and target encoding. The final R-squared score is approximately 0.90. In addition, a recommendation module was developed using cosine similarity on feature vectors and location advantages to provide users with relevant property alternatives. This system has been deployed using a Streamlit-based dashboard. Furthermore, an analytics module featuring interactive geo-maps provides real-time insights. We have created a functional real estate module from raw web data that provides stakeholders with better insights.

**Keywords**—Real Estate, Price Prediction, XGBoost, Random Forest, Cosine Similarity, Recommendation System, Machine Learning, Streamlit, Gurgaon, Data Science

## I. INTRODUCTION

A perfect home at the right price – that is something everyone is thinking about, especially in fast-growing cities like Gurgaon. As the city fills up with new apartments and independent houses, the real estate market has become incredibly complex. Buyers struggle to determine whether the price is fair or not, and sellers also face difficulties in justifying the value of their properties against so many others.

Previous methods for determining prices or other factors in houses were dependent on word of mouth of humans and visiting multiple brokers, which was insufficient. It was slow, prone to human error, and inadequate for keeping up with thousands of listings across different sectors [1] [2]. There was a clear need for a smarter, data-driven approach to react to the market in real time.

Artificial intelligence and machine learning are well-suited to this problem. By analyzing a large dataset from sites like 92acres, patterns can be identified that humans might miss or fail to recognize, as demonstrated by recent studies [1] [3][4]. Factors such as the number of bedrooms, luxury score, amenities, and even the sector location play a significant role in buying a house. In this project, we have built a complete system that works in three parts:

- **Price Prediction Module:** Predicts prices using advanced models like XGBoost and Random Forest that have outperformed traditional regression methods [2] [4] [5]. Accurate price estimates are generated based on property features.
- **Analyzer Module:** Analyses the market through various visual tools like charts, heatmaps, and other tools that help users understand price trends across Gurgaon.
- **Recommendation Module:** Based on cosine similarity, it integrates a content-based recommendations system that suggests the best property alternatives using user preferences [5] [6].

These three modules are integrated into a single Streamlit dashboard. This study demonstrates that next-generation AI can move out of the lab and help people make better financial decisions in the real world.

## II. RESEARCH OBJECTIVES

The main goal of this study is to create a smart platform that understands, assists, and navigates the Gurgaon real estate market using data science. Specifically, this project aims to achieve the following objectives:

- Detailed Study of the Real Estate Market: Scrape and clean real estate data from 92 acres for properties (flats, apartments, houses) in Gurgaon across different sectors to understand property fluctuations.
- Price Prediction Module with Accuracy: Develop a machine learning pipeline that estimates the fair market value of houses and flats by analyzing features such as location, area, number of rooms, luxury features, and other property characteristics.
- Recommendation System: Design a module based on cosine similarity that suggests the most similar and relevant housing options to stakeholders, based on their specific interests and property criteria.
- Comparing Machine Learning Models: Test and evaluate different machine learning algorithms, such as Random Forest, Linear Regression, and others, to identify the algorithm that handles complex real estate data with the highest accuracy.
- Easy-to-Use Tool for Users: Create a Streamlit dashboard combining all complex models that is accessible not just for experts but also for regular users, enabling instant price predictions and property suggestions.
- Data Reliability: Apply technical steps such as IQR outlier removal to ensure that SHAP-based feature importance is derived from clean, high-quality data.

## III. LITERATURE REVIEW

Nagula [1] provides a comparative analysis of multiple machine learning models for house price prediction, emphasizing that Random Forest outperforms traditional regression techniques due to its ability to capture nonlinear feature relationships. However, the study does not include recommendation systems or real-time analytical dashboards, providing scope for an integrated platform.

Zhao [2] evaluates various machine learning algorithms, along with ensemble methods, for predicting house prices. The study proves that ensemble methods yield more stable and accurate results over linear regression models.

Cellmer and Kobylńska [3] employed a hybrid approach integrating machine learning and geostatistical methods to analyze the complex relation between spatial nuances and house prices. By modelling geographic variables, they achieved better prediction accuracy. However, the study remains confined to prediction performance and does not provide personalized house recommendations or address challenges of deploying a real-time model.

Pastukh and Khomyshyn [4] experimented with different ensemble methods, such as Random Forest and Gradient Boosting, for predicting house prices. This study shows that these methods are very robust across diverse datasets.

Singaravelu et al. [5] presented a complete real estate price prediction system using machine learning, evaluated with standard metrics. The study focused on prediction but does not provide an advanced recommendation mechanism or an interactive analytical dashboard.

Singh et al. [6] extend the practice of using content-based recommender systems driven by cosine similarity, originally demonstrated in areas such as movie suggestions and online shopping. Applying the same technique to property recommendations represents a fresh and innovative step.

Malpni [7] acknowledges the IQR method as a dependable approach for detecting deviations, mainly because of its straightforward and flexible nature. Though proven effective in healthcare and financial analysis, using IQR within property recommendation and prediction models is still a relatively new research direction.

Sharma et al. [8] highlight the effectiveness of ensemble regression algorithms such as XGBoost, Random Forest, and Extra Trees in house price prediction. These models are able to find complex, nonlinear relationships in real estate datasets. Logarithmic transformations (e.g.,  $\text{np.log1p}$ ) help improve model convergence, reduce variance, and increase accuracy significantly.

Li [9] conducted a comparative analysis of Random Forest and XGBoost models on housing price datasets and reported that XGBoost achieved superior performance due to efficient handling of feature interactions and overfitting control. The study validated the effectiveness of ensemble learning but relied on a limited dataset and lacked deployment considerations for real-world applications.

Geertset al. [10] explored the application of large language models (LLMs) in real estate appraisal, highlighting their potential for explainability and interactive valuation. While promising, the approach is still in its early stages and lacks large-scale validation for real-time residential market deployment.

Gu`mmeretal.[11]usedaclusteringmethodtofind out what specifically influences real estate rates in different localities, making an interpretable model. While the study proves that a “one-size-fits-all” pricing model does not work, it remains experimental and does not include a user-facing platform to help buyers make real-world decisions.

#### IV. PROPOSED METHOD

The proposed method applies a data science framework that transforms raw, unstructured real estate datasets into predictions and market analyses. The framework consists of six main phases ranging from initial problem analysis to final model deployment.

##### A. Data Acquisition and Pre-processing

Data for this study was scraped from 99acres.com. The resulting dataset comprised approximately 3,900 listings of Gurgaon real estate, including independent houses, flats, and apartments, with a significant amount of noise and missing values.

- **Data Collection:** The dataset included 1,044 independent homes and approximately 3,000 flats, with 21 features including sector location, built-up area, furnish type, and more.
- **Missing Value Imputation:** Features with missing values were handled using mode and median imputation strategies to maintain data integrity.
- **Outlier Treatment:** The Interquartile Range (IQR) method was applied to eliminate extreme values and improve model robustness [7].
- **Feature Extraction:** Regular expressions were applied to the areaWithType column to extract Super Built-up Area, Built-up Area, and Carpet Area.

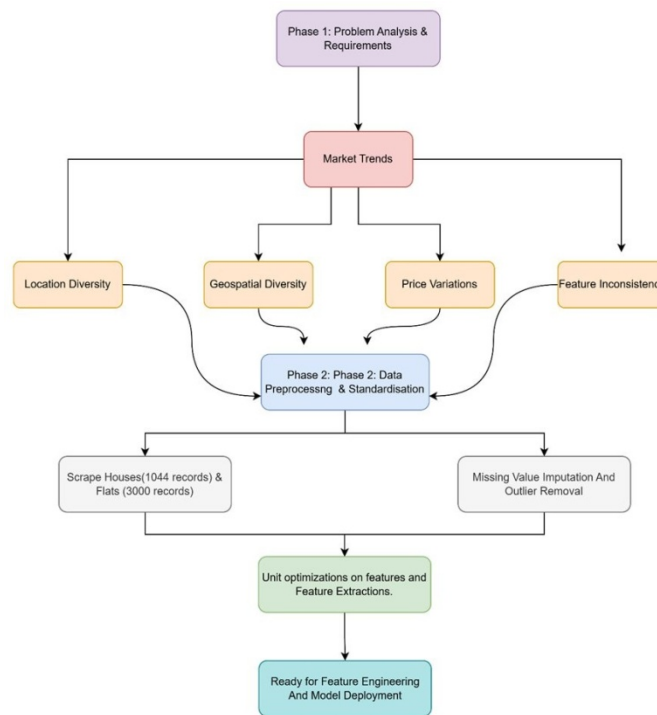


Fig. 1. Data Acquisition and Pre-processing Pipeline

##### B. Model Development and Analytical Modules

In this phase, advanced exploratory analysis and predictive engine construction were performed. The system handles price classification, market trends, and localized spatial trends through specialized models.

- **Exploratory Data Analysis:** Univariate and multivariate analyses were performed to understand market trends, revealing hidden facts such as that 75% of residential area in Gurgaon relies on flats.
- **Geospatial Analytics:** A customized lat-long scraper was developed to fetch coordinates of Gurgaon sectors, enabling interactive sector-based price-per-sqft trend detection on a geo-map.

- Price Prediction Module: Eleven different regression algorithms were applied, including XGBoost, Random Forest, and Extra Trees, to handle the skewed pricing data. A log-normal transformation ( $\text{np.log1p}$ ) was applied to the target variable, which significantly improved model convergence [8] [9].
- Recommendation Module: A similarity engine using cosine similarity was built to recommend relevant properties. Multi-vector factors such as facility similarity, amenities, property specifications, and location advantages were integrated [5] [6].

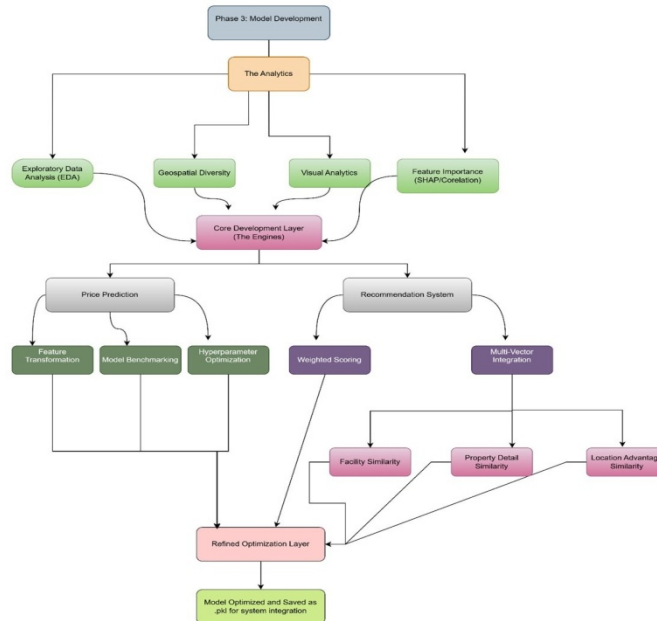


Fig.2. Model Development Overview

### C. System Integration

Using scientific design principles, the individual models were integrated into a web dashboard using the Streamlit framework. A multi-page application was designed for the system architecture.

- 1) Integrated Multi-page Architecture: The application is divided into three core functional pages: the analytics module, price predictor, and recommendation system. A modular design allows the system to efficiently handle parallel user requests.
- 2) Model Persistence and Serialization: To ensure low-latency performance, the pre-trained machine learning pipeline and similarity matrices were serialized into Pickle (.pkl) files. When a user launches the dashboard, pipeline.pkl and similarity.pkl are instantly deserialized, providing a real-time interface without recalculating the entire logic.
- 3) Real-Time Data Flow: The integration layer manages user input such as property sector, BHK, and area, passing them through the same preprocessing scripts used during training to ensure valid results.

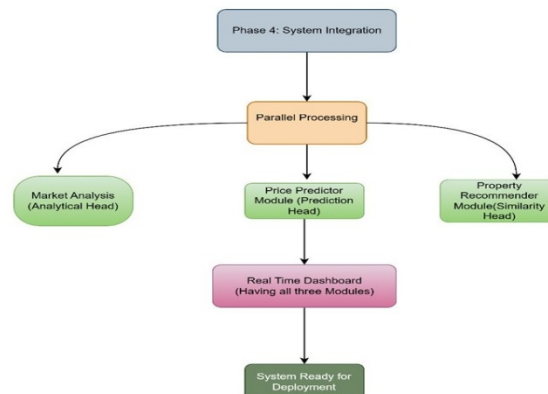


Fig.3. System Integration Architecture

**D. Testing and Validation**

The evaluation phase plays a critical role in determining the reliability and accuracy of the developed system before deployment. This involves evaluating the machine learning model and recommendation engine using standard performance metrics.

- **Lab-Based Performance Assessment:** The system was evaluated using a reverse 20% split of the Gurgaon property dataset, ensuring the model was tested on unseen data to provide an unbiased measure of real-world predictive performance.
- **Quantitative Evaluation Metrics:**
  - **R-Squared ( $R^2$ ) Score:** The XGBoost model achieved a significant  $R^2$  score of 0.88–0.90.
  - **Mean Absolute Error (MAE):** A low MAE of approximately 0.44 on the log-transformed scale was maintained.
  - **Cosine Similarity Score:** A recommendation similarity score of 0.85 was maintained, validating suggestions based on shared location and luxury features.
- **Comparative Analysis:** Initial tests were performed using linear regression and support vector regression (SVR) to establish a performance baseline. Ensemble methods such as Random Forest, Extra Trees, and XGBoost were then evaluated. The comparison confirmed that the XGBoost pipeline achieved the best performance across Gurgaon’s housing sectors [11].

**E. Refinement and Optimization**

- **Hyperparameter Tuning:** In-depth tuning of model parameters was performed with GridSearchCV, working on

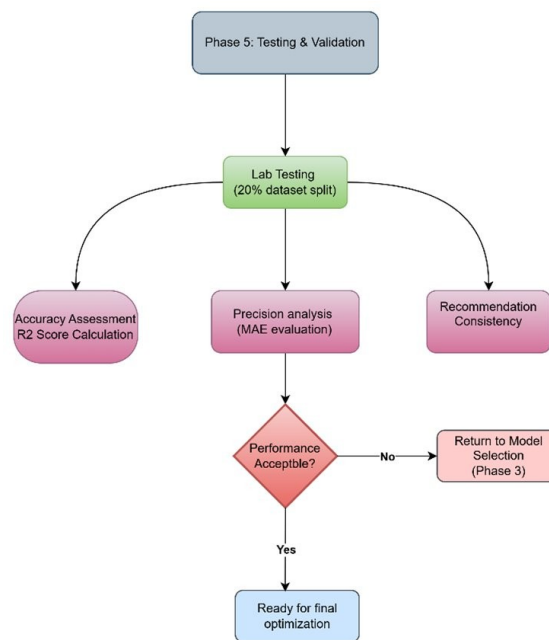


Fig.4. Testing and Validation Flowchart

variable parameters including learning rate, max depth, and number of estimators for XGBoost and Random Forest models. This resulted in a stable  $R^2$  value of 0.90 across the diverse sectors of Gurgaon.

- **Data Transformation Refinement:** Log-normal transformation ( $\text{np.log}1p$ ) was applied to adjust for the skewed distribution of real estate prices, decreasing the error rate across all price ranges.
- **Model Serialization and Deployment Efficiency:** Optimized pipelines were saved as pickled (.pkl) files to greatly reduce server load and provide almost immediate response to user queries without constant retraining.
- **Enhanced User Interface:** An interactive dashboard was developed including real-time price anomaly alerts, historical market trend studies using geospatial analysis, and a smart recommendation platform.

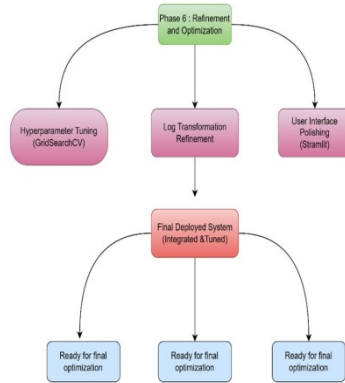


Fig.5.RefinementandOptimizationPhase

### V. RESULTS

The system was trained on a tuned machine learning pipeline using a custom dataset of 1,044 independent house records and approximately 3,000 flat listings from Gurgaon, covering a large range of sectors, property ages, and luxury features.

- 1) Price Prediction Performance: The final XGBoost model demonstrated high performance in the Gurgaon market, performing well even in large growth areas with high price fluctuation [4] [9].
- 2) Recommendation Accuracy: The similarity engine provides relevant recommendations using multi-factor analysis, going beyond price alone. For instance, a high-end flat in Sector 65 is matched based on luxury features rather than just price.
- 3) Performance Metrics: The system reports high and consistent performance across all property types, with strong MAE and  $R^2$  values as shown in Table I.
- 4) Real-Time Efficiency: The serialized pipeline ensures low system response times, confirming suitability for real-time deployment and immediate user queries.
- 5) Comparative Analysis: Across 10 different regression models, XGBoost performed best as shown in Table II, with Random Forest showing strong performance on nonlinear data.

TABLE I  
PERFORMANCE METRICS OF THE PROPOSED SYSTEM

Metric	Result
R-Squared ( $R^2$ )	0.90
Mean Absolute Error (MAE)	0.44 (LogScale) Model
Training Size	3,900+ Records

TABLE II  
COMPARATIVE ANALYSIS OF REGRESSION MODELS

Algorithm	R-Squared ( $R^2$ )
XGBoost	0.90
Random Forest	0.88
Extra Trees	0.87
Linear Regression	0.82
Support Vector Regressor (SVR)	0.79

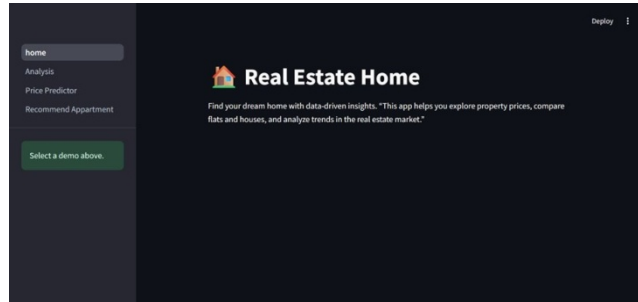


Fig.6. User Interface of the Streamlit Dashboard

PropertyName	Bajghera Road	Palam Vihar Halt	DPSG Palam Vihar	Park Hospital	Gurgaon Rail
Smartworld One DXP	800	7,500	3,100	3,100	
M3M Crown	550	54,000	54,000	54,000	
Adani Brahma Samsara Vilasa	5,300	54,000	54,000	54,000	
Sobhu City	1,500	54,000	54,000	54,000	
Signature Global City 93	54,000	54,000	54,000	5,500	
Whiteland The Aspen	54,000	54,000	54,000	5,800	
Bestech Altura	54,000	54,000	54,000	4,400	
Elan The Presidential	54,000	54,000	54,000	8,500	
Signature Global City 92	54,000	54,000	54,000	2,900	
Emaar Dighomes	54,000	54,000	54,000	54,000	

### Select Location and Radius

Location: AIIMS

Radius in Kms: 0.00

Search

### Recommend Apartments

Select an apartment: 4S Aradhya Homes

Recommend

	PropertyName	SimilarityScore
0	SS Linden Floors	22.5704
1	Emaar MGF Emerald Floors Premier	20.6414
2	Adani Brahma Samsara	16.6998
3	Optimal ultra luxury builder floors	16.6338
4	DLF Gardencity	15.1605
5	Mappso The Icon 79	14.5153
6	BPTP Mansions Park Prime	14.3546
7	Vatika Independent Floors	13.4431
8	Paras Ekam Homes	13.2323
9	Emaar MGF Palim Terraces	12.3739

Fig.7. Recommendation Module: Location Search and Cosine Similarity Rankings

### Price Predictor

#### Enter your inputs

Property Type: flat

Sector: dwarka expressway

No of Bedrooms: 2.0

No of Bathrooms: 2.0

Balconies: 0

Property Age: Moderately Old

Built up Area: 0.00

Servent Room: 1.0

Store Room: 0.0

Furnishing Type: semifurnished

Luxury Category: High

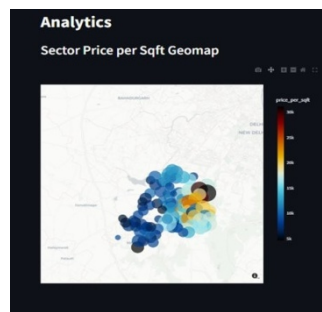
Floor Category: High Floor

Predict

property_type	sector	bedRoom	bathroom	balcony	agePossession	built_up_area	
0	flat	dwarka expressway	2	2	0	Moderately Old	0

The Price of the house in between 0.3 Cr and 0.74 Cr

Fig.8.PricePredictionModule:InputFormandPredictedPriceOutput



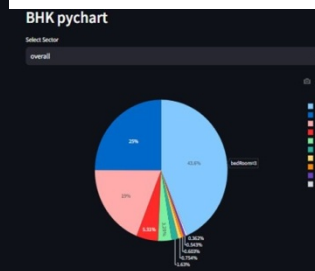
(a)SectorPriceperSqftGeo-map



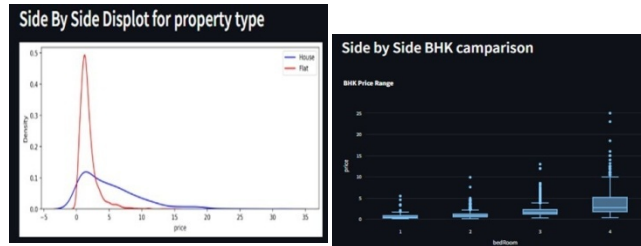
(b)AmenitiesWordCloud



(c)Area vs Price Scatter Plot



(d)BHK Distribution Pie Chart



(e) Price Density by Property Type (f) BHK Price Range Box Plot

Fig.9. Analytical Module: Interactive Visualizations for Gurgaon Real Estate Market

## VI. CONCLUSIONS

This research reports the development of an AI-based platform used to analyze, predict, and recommend properties in the Gurgaon real estate market, derived from a large 99 acres dataset. Moving beyond usual guesswork, data-based valuations are presented. The XGBoost model, when paired with log-normal transformation, significantly outperforms linear regression and SVM, achieving a high accuracy of  $R^2=0.90$  [1][2][4][9].

The platform provides competitive price forecasts as well as relevant property recommendations. Complex machine learning models have been integrated into a user-friendly dashboard, making it equally accessible to the average homebuyer and industry professionals. This constitutes a practical tool that helps navigate India's most complex property markets by connecting analysis with real-world financial decisions. The solution is also scalable and may be deployed in other smart cities [11].

## VII. FUTURE SCOPE

The present model can be extended to include commercial properties and rapidly adapted for different regions such as Delhi, Noida, Mumbai, and others. Incorporating time series forecasting with historical price data will enable predictions of future property rates, helping investors identify the best times to buy or sell.

A mobile application with location-based notifications alerting users when prices drop and displaying new similar property listings in real time is planned for development. Future versions will incorporate deep learning for property photo analysis, enabling automatic rating of luxury and quality based on interior finishes.

Investment risk assessment can be enhanced by connecting the system to city infrastructure data, presenting risk scores related to future metro growth, flood risk, and industrial development. Integration with automated PDF generation will allow users to obtain professional-quality valuation reports for use in bank loans and legal documents.

## REFERENCES

- [1] S. Nagula, "Real Estate Price Prediction Using Machine Learning Models," IJRASET, ISSN: 2321-9653, vol. 13, July 2025. <https://doi.org/10.22214/ijraset.2025.72962>
- [2] T. Zhao, "Predicting House Prices Using Machine Learning Models," Transactions on Computer Science and Intelligent Systems Research, ISSN: 2960-1800, vol. 9, AIDML 2025.
- [3] R. Cellmer and K. Kobylin'ska, "Housing Price Prediction - Machine Geostatistical Methods," Real Estate Management and Valuation, vol. 33, no. 1, 2025. <https://doi.org/10.2478/remav-2025-0001>
- [4] O. Pastukh and V. Khomyshyn, "Using Ensemble Methods of Machine Learning to Predict Real Estate Prices," arXiv:2504.04303, 2025. <https://doi.org/10.48550/arXiv.2504.04303>
- [5] Singaravelu, Muthuselvan, et al., "Real Estate Price Prediction System Using Machine Learning Algorithm," AIP Conference Proceedings, vol. 3175, no. 1, AIP Publishing LLC, 2025. <https://doi.org/10.1063/5.0254265>
- [6] K. Singh, M. Mishra, and Er. S. Singh, "Content-based Recommender System Using Cosine Similarity," IJRASET, ISSN: 2321-9653, vol. 12, Issue 4, May 2024. <https://doi.org/10.22214/ijraset.2024.61835>
- [7] Dr. K. Malpni, "Detecting Outliers for Single Dimensional Data Using Interquartile Range," Journal of Engineering Research and Application, ISSN: 2248-9622, vol. 9, Issue 9, pp. 31-35, September 2019. <https://doi.org/10.9790/9622-0909013135>
- [8] H. Sharma, H. Harsora, and B. Ogunleye, "An Optimal House Price Prediction Algorithm: XGBoost," Analytics, vol. 3, pp. 30-45, 2024. <https://doi.org/10.3390/analytics3010003>
- [9] H. Li, "House Price Prediction and Analysis Based on Random Forest and XGBoost Models," Highlights in Business, Economics and Management, vol. 21, pp. 934-938, 2023. <https://doi.org/10.54097/hbem.v21i.14837>
- [10] M. Geerts, M. Reusens, B. Baesens, S. vanden Broucke, and J. DeWeerd, "On the Performance of LLMs for Real Estate Appraisal," in ECML PKDD 2025, Lecture Notes in Computer Science, vol. 16021, Springer, Cham, 2026. <https://doi.org/10.48550/arXiv.2506.11812>
- [11] P. Gu'mmer, J. Rosenberger, M. Kraus, P. Zschech, and N. Hambauer, "Unveiling Location-Specific Price Drivers: A Two-Stage Cluster Analysis for Interpretable House Price Predictions," arXiv:2508.03156, 2025. <https://doi.org/10.48550/arXiv.2508.03156>



10.22214/IJRASET



45.98



IMPACT FACTOR:  
7.129



IMPACT FACTOR:  
7.429



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24\*7 Support on Whatsapp)