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

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.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 populationgrowthincitieshasincreasedandwillincreasefurther 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.**

ThisstudyservesasacomprehensiveAI-drivenframework 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.

Thisresearchpresentsanend-to-enddatasciencepipeline thatincorporatesrigorouspreprocessingincludingIQR-based outlier treatment andadvancedfeatureengineeringtoextract therequiredfeaturessuchasluxuryamenities,servantroom availability, floornumber, andsoon.Toensurehighavailabil- ity,amulti-modelcomparisonwasalsoconducted,evaluating 11algorithms, including linearregression,randomforest,and XGBoost, across various encoding strategies such as one-hot and targetencoding. ThefinalR-squaredscoreisapproximately0.90. Inaddition,arecommendationmodulewasdevelopedusing cosinesimilarity onfeaturevectorsandlocationadvantagesto provideuserswithrelevantpropertyalternatives.Thissystemhas been deployed using a Streamlit-based dashboard. Furthermore, ananalyticsmodulefeaturinginteractivegeo-mapsprovides real-timeinsights. Wehavecreatedafunctionalrealestate modulefromrawwebdatathatprovidesstakeholderswithbetterinsights.

***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 incred- iblycomplex.Buyersstruggletodeterminewhethertheprice is fair or not, and sellers also face difficulties in justifying the value of their properties against so many others.

Previousmethodsfordeterminingpricesorotherfactors 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]. Therewas a clear need for a smarter, data-driven approach to reactto 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:

- PricePredictionModule:Predictspricesusingadvanced 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,itintegratesacontent-basedrecommendationsystemthat suggests the best property alternatives using user prefer- ences [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.

©IJRASET: All Rights are Reserved

II. RESEARCH OBJECTIVES

The main goal of this study is to create a smart platformthatunderstands,assists,andnavigatestheGurgaonrealestate 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 92acres for properties (flats, apartments, houses) in Gurgaon across different sectorsto 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 eval- uate different machine learning algorithms, such as Ran-domForest,LinearRegression,andothers,toidentifythe algorithm that handles complex real estate data with the highest accuracy.
- Easy-to-Use Tool for Users: Create a Streamlit dash- board 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 im- portance 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, empha- sizing that Random Forest outperforms traditional regression techniques due to its ability to capture nonlinear feature relationships.However,thestudydoesnotincluderecommen- dation systems or real-time analytical dashboards, providing scope for an integrated platform.

Zhao [2] evaluates various machine learning algorithms, alongwithensemblemethods,forpredictinghouseprices.The study proves that ensemble methods yield more stable and accurate results over linear regression models.

Cellmer and Kobylińska[3] employed a hybrid approach integrating machine learning and geostatistical methods to an- alyze the complex relation between spatial nuances and house prices. By modelling geographic variables, they achieved bet- ter 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 estatepricepredictionsystemusingmachinelearning,evaluatedwith standard metrics. The study focused on prediction but doesnot 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 recom- mendations represents a fresh and innovative step.

Malpni [7] acknowledges the IQR method as a depend-able approach for detecting deviations, mainly because of its straightforwardandflexiblenature.Thoughproveneffectivein healthcare and financial analysis, using IQR within property recommendationandpredictionmodelsisstillarelativelynew 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 abletofindcomplex,nonlinear relationshipsinrealestatedatasets. Logarithmic transformations (e.g., 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 re- liedonalimiteddatasetandlackeddeploymentconsiderations for real-world applications.

Geertsetal.[10]exploredtheapplicationoflargelanguage 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 deploy- ment.

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 frameworkthat 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. DataAcquisitionandPre-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.

- DataCollection:Thedatasetincluded1,044independent homes and approximately 3,000 flats, with 21 features including sector location, built-up area, furnish type, and more.
- MissingValueImputation:Featureswithmissingvalues were handled using mode and median imputation strate- gies 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 appliedto the areaWithTypecolumn to extract Super Built- up Area, Built-up Area, and Carpet Area.

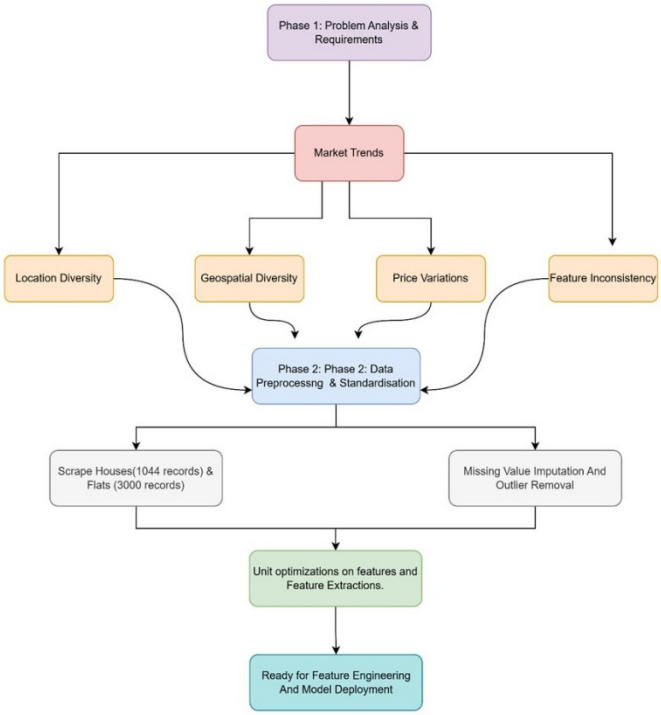


Fig.1.DataAcquisitionandPre-processingPipeline

B. ModelDevelopmentandAnalyticalModules

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

- ExploratoryDataAnalysis:Univariateandmultivariate analyses were performed to understand market trends, revealing hidden facts such as that 75% of residentialarea in Gurgaon relies on flats.
- Geospatial Analytics: A customized lat-long scraperwas developed to fetch coordinates of Gurgaon sectors, enabling interactive sector-based price-per-sqft trend de- tection on a geo-map.

- Price Prediction Module: Eleven different regression algorithmswereapplied,includingXGBoost,Random Forest, and Extra Trees, to handle the skewed pricing data. A log-normal transformation (np.log1p) was ap- plied to the target variable, which significantly improved model convergence [8] [9].
- Recommendation Module: A similarity engine using cosine similarity was built to recommend relevant prop- erties. Multi-vector factors such as facility similarity, amenities, property specifications, and location advan- tages were integrated [5] [6].

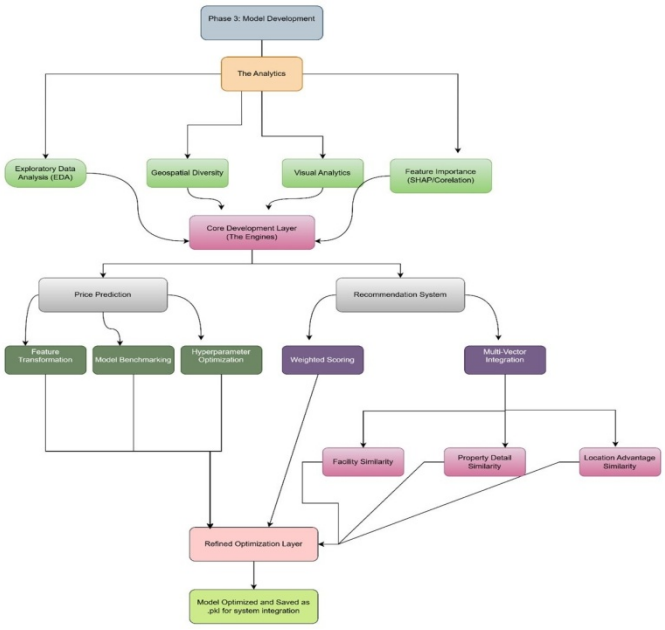


Fig.2.ModelDevelopmentOverview

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,pricepredictor,andrecommendationsystem.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 dash-board,pipeline.pklandsimilarity.pklare instantly deserialized, providing a real-time interface without recalculating the entire logic.

3) Real-Time Data Flow: The integration layer manages userinputssuchaspropertysector,BHK,andarea, passingthem through the samepreprocessingscripts used during training to ensure valid results.

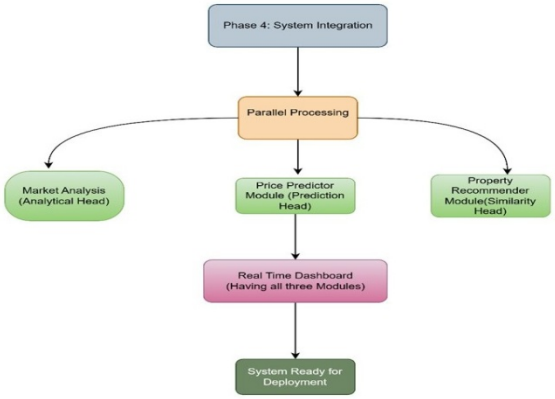


Fig.3.SystemIntegrationArchitecture

D. TestingandValidation

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

- Lab-Based Performance Assessment: The system was evaluated using a reverse 20% split of the Gurgaon prop- ertydataset,ensuringthemodelwastestedonunseendata to provide an unbiased measure of real-world predictive performance.
- QuantitativeEvaluationMetrics:
  - R-Squared (R^2) Score: The XGBoost model achieved a significant R^2score of 0.88–0.90.
  - Mean Absolute Error (MAE): A low MAE of approximately0.44onthelog-transformedscalewas maintained.
  - Cosine Similarity Score: A recommendation simi- larity score of 0.85 was maintained, validating sug- gestions based on shared location and luxury fea- tures.
- Comparative Analysis: Initial tests were performed us- inglinearregressionandsupportvectorregression(SVR) to establish a performance baseline. Ensemble methods suchasRandomForest,ExtraTrees,andXGBoost were then evaluated. The comparison confirmed that the XGBoost pipeline achieved the best performance across Gurgaon's housing sectors [11].

E. RefinementandOptimization

- Hyperparameter Tuning: In-depth tuning of model pa- rameterswasperformedwithGridSearchCV,workingon

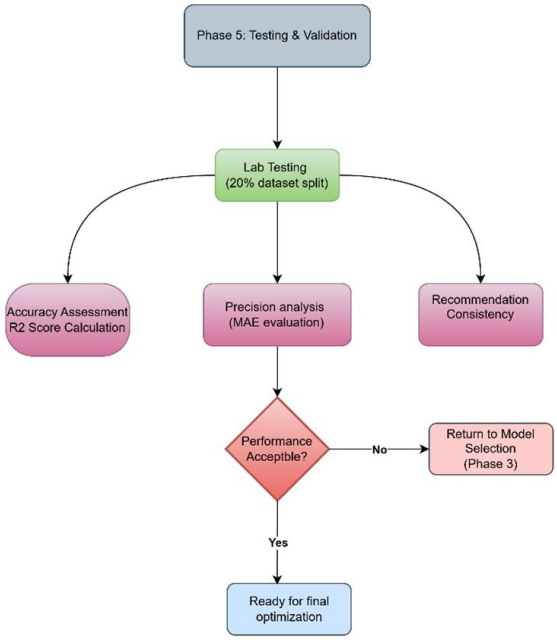


Fig.4.TestingandValidationFlowchart

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

- Data Transformation Refinement: Log-normal transfor- mation(np.log1p)wasappliedtoadjustfortheskewed distribution of real estate prices, decreasing the error rate across all price ranges.
- Model Serialization and Deployment Efficiency: Op- timized 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 dashboardwas developed including real-time price anomaly alerts, historical market trend studies using geospatial analysis, and a smart recommendation platform.

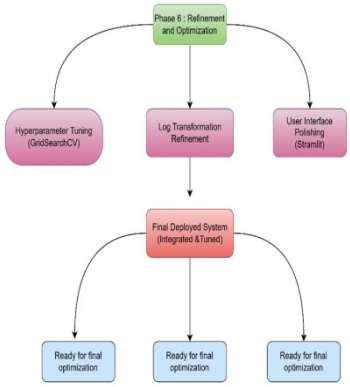


Fig.5.RefinementandOptimizationPhase

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 pro- vides relevant recommendations using multi-factor anal- ysis, 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^2values as shown in Table I.
4) Real-Time Efficiency: The serialized pipeline ensures lowsystemresponsetimes,confirmingsuitabilityforreal- 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.

TABLEI
PERFORMANCEMETRICSOFTHEPROPOSEDSYSTEM

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

TABLEII
COMPARATIVEANALYSISOFREGRESSIONMODELS

Algorithm	R-Squared(R^2)
XGBoost	0.90
RandomForest	0.88
ExtraTrees	0.87
LinearRegression	0.82
SupportVectorRegressor(SVR)	0.79

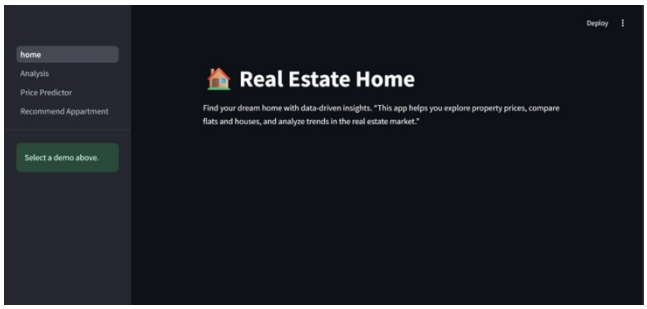


Fig.6.UserInterfaceoftheStreamlitDashboard

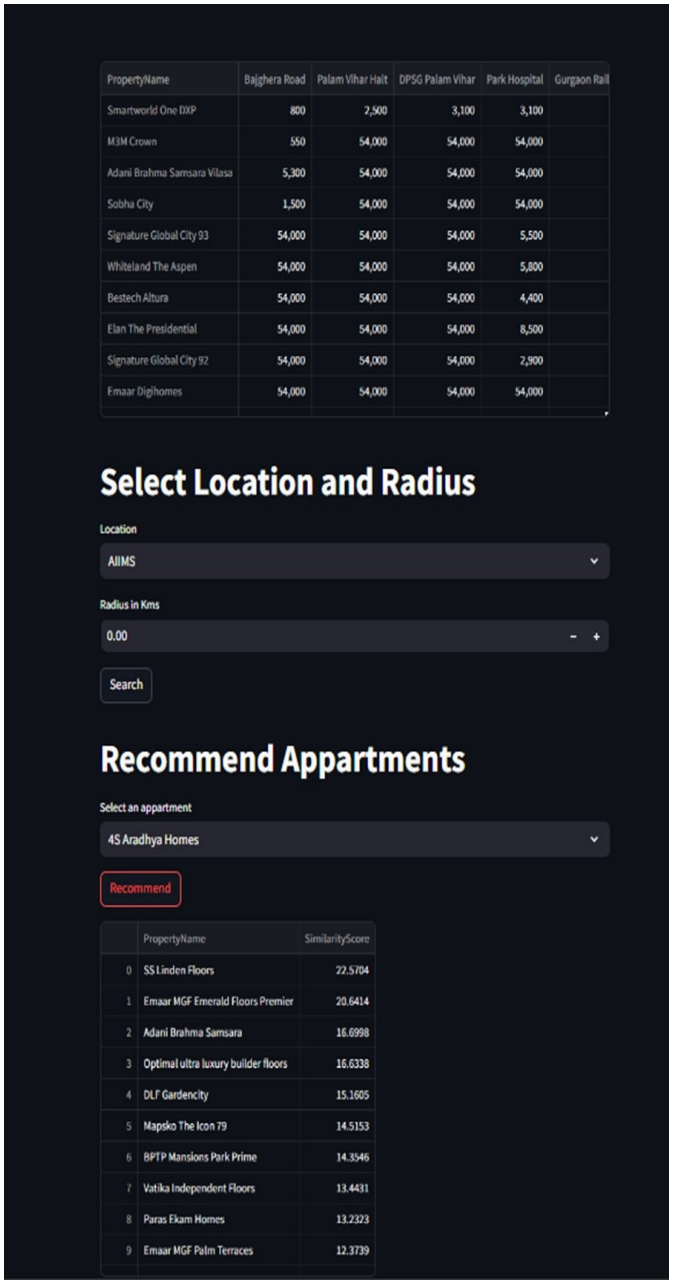


Fig.7.RecommendationModule:LocationSearchandCosineSimilarityRankings

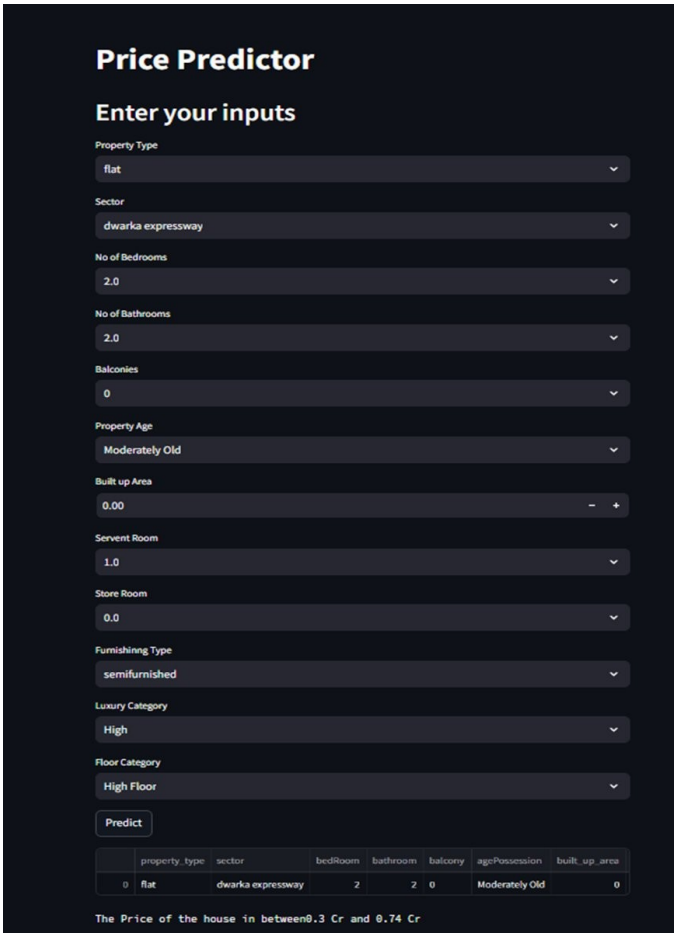
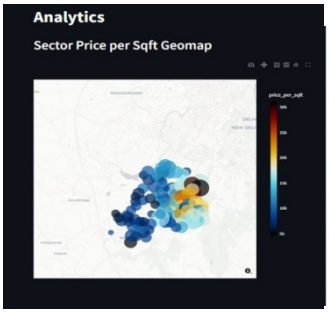


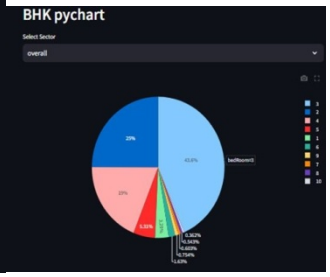
Fig.8.PricePredictionModule:InputFormandPredictedPriceOutput



(a)SectorPriceperSqftGeo-map

(b)AmenitiesWordCloud

(c)AreavsPriceScatterPlot



(d)BHKDistributionPieChart

(e)PriceDensitybyPropertyType (f)BHKPriceRangeBoxPlot

Fig.9.AnalyticalModule:InteractiveVisualizationsforGurgaonRealEstateMarket

VI. CONCLUSIONS

This research reports the development of an AI-based plat- form used to analyze, predict, and recommend properties inthe Gurgaon real estate market, derived from a large 99acres dataset. Moving beyond usual guesswork, data-based valua- tions are presented. The XGBoost model, when paired with log-normal transformation, significantly outperforms linear regressionandSVM,achievingahighaccuracyofR^2=0.90[1][2][4][9].

Theplatformprovidescompetitivepriceforecastsaswellas relevant property recommendations. Complex machine learn- ing models have been integrated into a user-friendly dash- board,makingitequallyaccessibletotheaveragehomebuyer and industry professionals. This constitutes a practical toolthat 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 propertiesandrapidlyadaptedfordifferentregionssuchas Delhi, Noida, Mumbai, and others. Incorporating time series forecastingwithhistoricalpricedatawillenablepredictionsof future property rates, helping investors identify the best times to buy or sell.

Amobileapplicationwithlocation-basednotificationsalert- ing users when prices drop and displaying new similar prop- erty listings in real time is planned for development. Future versions will incorporate deep learning for property photo analysis,enablingautomaticratingofluxuryandqualitybased 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 allowuserstoobtainprofessional-qualityvaluationreportsfor use in bank loans and legal documents.

REFERENCES

[1] S. Nagula, "Real Estate Price Prediction Using Machine LearningModels," 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.CellmerandK.Kobylin´ska,"HousingPricePrediction–Ma-chine 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 MachineLearningPredictRealEstatePrices,"arXiv:2504.04303,2025.https: //doi.org/10.48550/arXiv.2504.04303

[5] Singaravelu, Muthuselvan, et al., "Real Estate Price Prediction SystemUsing Machine Learning Algorithm," AIP Conference Proceedings,vol.3175,no.1,AIPPublishingLLC,2025.https://doi.org/10.1063/5.0254265

[6] K. Singh, M. Mishra, and Er. S. Singh, "Content-based RecommenderSystem Using Cosine Similarity," IJRASET, ISSN: 2321-9653, vol. 12,

[7] IssueV,May2024.https://doi.org/10.22214/ijraset.2024.61835

[8] Dr. K. Malpni, "Detecting Outliers for Single Dimensional Data UsingInterquartileRange,"JournalofEngineeringResearchandApplication,ISSN:2248-9622,vol.9,Issue9,pp.31–35,September2019.https://doi.org/10.9790/9622-0909013135

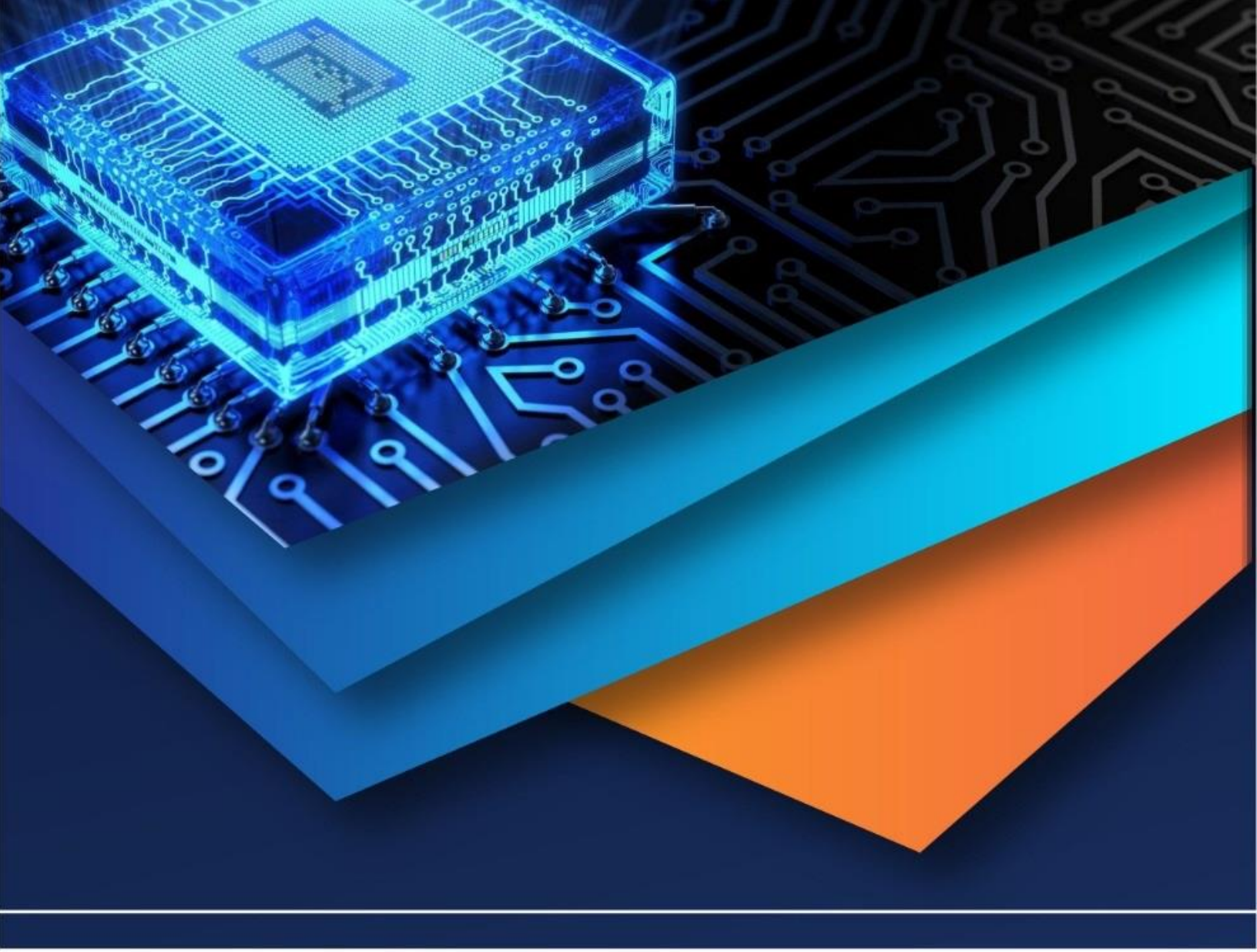
[9] H. Sharma, H. Harsora, and B. Ogunleye, "An Optimal House PricePrediction Algorithm: XGBoost," Analytics, vol. 3, pp. 30–45, 2024.https://doi.org/10.3390/analytics3010003

[10] H. Li, "House Price Prediction and Analysis Based on Random ForestandXGBoostModels,"HighlightsinBusiness,EconomicsandManage-ment,vol.21,pp.934–938,2023.https://doi.org/10.54097/hbem.v21i.14837

[11] M. Geerts, M. Reusens, B. Baesens, S. vanden Broucke, and J. DeWeerdt, "On the Performance of LLMs for Real Estate Appraisal," inECMLPKDD2025,LectureNotesinComputerScience,vol.16021,

[12] Springer,Cham,2026.https://doi.org/10.48550/arXiv.2506.11812

[13] P. Gu¨mmer, J. Rosenberger, M. Kraus, P. Zschech, and N. Hambauer,"UnveilingLocation-SpecificPriceDrivers:ATwo-StageClusterAnal-ysisforInterpretableHousePricePredictions,"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)