JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Multilingual House Price Forecasting System Using Machine Learning and NLP

Ananya R Chandhra, Harshitha HV, Harshitha VS, Manasa D
Prof. Anjali Gajjar
Bangalore, India
{ananyarchandhra, vsharshitha23, harshithavs09, manasa1672004} @gmail.com,
gajjaranjali224@gmail.com

ABSTRACT

The rapid growth of the global real estate market has increased the need for accurate and easily accessible house price prediction tools. However, most existing systems rely solely on structured numerical data and operate in a single language, limiting their usability for diverse linguistic communities. This project presents a Multilingual House Price Prediction System that integrates Machine Learning and Natural Language Processing (NLP) to deliver accurate, region-specific property price estimates while supporting multiple languages. The system analyses key factors such as location, property size, nearby amenities, and market trends using advanced predictive algorithms. To enhance accessibility, a multilingual interface powered by NLP enables users to interact with the system and receive explanations in their preferred language. By combining predictive analytics with language translation capabilities, the proposed system improves inclusivity, user experience, and decision-making for buyers, sellers, and investors across different regions.

KEYWORDS

House Price Prediction, Machine Learning, Multilingual System, Natural Language Processing (NLP), Real Estate Analytics, Language Translation, Predictive Modelling, Regression Algorithms, Regional Market Analysis, Geospatial Features, User Accessibility, Multi-Regional Data, Intelligent Decision Support, Property Valuation, Human–Computer Interaction (HCI).

1. INTRODUCTION

The real estate sector is one of the most dynamic and data-driven industries, where accurate property valuation plays a crucial role in guiding investment decisions for buyers, sellers, financial institutions, and policymakers. House price prediction has emerged as a significant research area, driven by the availability of large-scale datasets and advancements in machine learning techniques capable of identifying complex relationships among regional, structural, and economic variables. Traditional pricing systems, however, are often limited by their dependence on structured numerical data and by their inability to accommodate the linguistic diversity of global users. As a result, non-English speakers and users from multilingual regions face difficulties in accessing reliable price insights and interpreting model outputs effectively.

In an increasingly interconnected world, multilingual support has become essential for building inclusive and user-centric intelligent systems. Natural Language Processing (NLP) and language translation technologies now enable seamless interaction between users and computational models, allowing information to be delivered in a linguistically accessible manner. Integrating these capabilities into house price prediction enhances not only system usability but also trust and understanding, particularly in regions where multiple languages coexist.

This research proposes a Multilingual House Price Prediction System that combines machine learning algorithms with NLP-based language translation to produce accurate and region-aware property price estimates while enabling users to interact in their preferred language. The system incorporates multiple regional features—such as location, property dimensions, amenities, and market trends—to generate precise predictions. It further employs multilingual processing to translate system outputs and explanations, thereby improving accessibility and supporting informed decision-making across diverse user groups.

The proposed work aims to bridge the gap between technical prediction models and real-world user needs by offering an intelligent, inclusive, and scalable framework suitable for modern real estate applications. By incorporating NLP-driven translation and language understanding modules, the proposed system not only broadens accessibility but also enhances user confidence in automated predictions. Additionally, the combination of multimodal data—such as regional attributes, textual descriptions, and potentially property images—enables richer feature extraction and contributes to more reliable price estimation. This holistic approach positions the system as a comprehensive tool capable of serving diverse populations while advancing the current landscape of real estate analytics.

2. LITERATURE REVIEW

1] Maloku et al., *House Price Prediction Using Machine Learning and Artificial Intelligence* This paper compares linear regression and Random Forest models for predicting house prices and concludes that ensemble methods provide superior accuracy. It highlights the importance of feature correlation, preprocessing, and structured datasets. **[2]** Kalidass et al., *House Price Prediction Using Machine Learning* (IRJET) The study evaluates Random Forest and Gradient Boosting on an India-based dataset, showing that ensemble models consistently outperform simple regressors. The work is strong in methodology but remains limited to numerical data without exploring advanced NLP or deep learning techniques. **[3]** Kumar et al., *Bangalore House Price Prediction* (IJCRT) This research focuses on a large Bangalore dataset, demonstrating that careful feature engineering—such as price per square foot and location encoding—greatly improves prediction accuracy. While effective, the study is region-specific and does not incorporate text or multilingual interactions.

[4] Chandrasekar et al., *House Price Prediction Model Using Machine Learning* (JETIR) The paper provides a comparative study of regression, Decision Tree, Random Forest, and SVR models, showing that tree-based models typically deliver the best performance. It offers practical insights but lacks exploration of deep, hybrid, or NLP-enhanced models. **[5]** Russia et al., *House Price Prediction using Deep Learning and Sentence Embedding.* This work integrates LSTM networks with sentence embeddings to utilize textual property descriptions for improved price prediction. It demonstrates the benefits of multimodal learning but is restricted to English text and does not address multilingual processing or real-world deployment challenges.

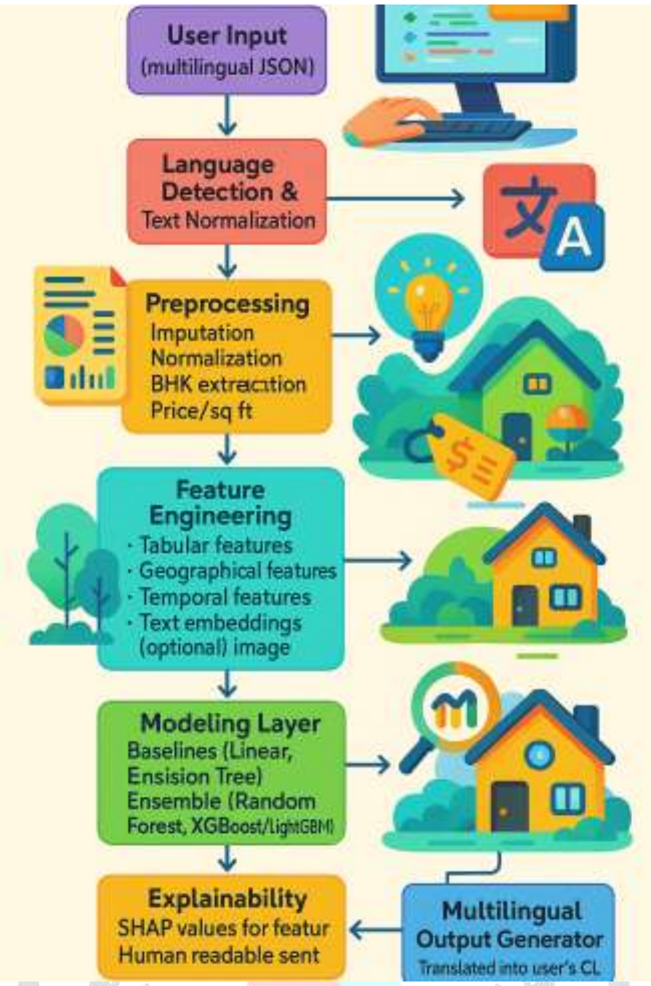
3. PROPOSED SYSTEM

3.1 System Overview

Build an accurate, region-aware house-price prediction system that supports multilingual interaction and provides interpretable explanations of predictions. Language detection → route to multilingual encoder or translate-to-canonical language. Generate human-readable explanations from model outputs (map SHAP scores to sentences). Translate explanations with domain-aware MT or directly generate in target language using multilingual LMs to preserve local context and terminology. NLP & Translation pipeline: multilingual encoders and/or translation models for input understanding and explanation generation. Geospatial features: cluster locations, distances to landmarks, neighbourhood encodings. Temporal features: seasonality, recent trend deltas, days-on-market proxies. Text processing: language detection, tokenization, multilingual embeddings sentiment/key-phrase extraction.

The proposed system is an intelligent Multilingual House Price Prediction Platform that combines Machine Learning and Natural Language Processing to provide accurate and accessible property valuations. It accepts user inputs in multiple languages, processes them using NLP-based language detection and translation, and converts them into structured features suitable for predictive modelling. The system integrates diverse data sources—including numerical attributes, geospatial data, textual descriptions, and optionally images—to generate enriched feature sets.

A hybrid modelling approach is used, where tabular ML algorithms are fused with text and image embeddings in a multimodal framework to improve prediction accuracy. The system further provides interpretable outputs using explainability tools like SHAP, which are then translated into the user's preferred language for better understanding. Finally, the multilingual output module delivers predicted prices, confidence ranges, and explanation summaries through an intuitive user interface, making the system highly accessible, accurate, and user-centric.



3.2 Machine Learning and NLP processing

Removing irrelevant columns such as *society, availability, balcony* to reduce noise. Handling missing values using mean/median imputation. Converting *Size* into numerical BHK values. Creating new engineered features such as *price_per_sqft***.** Removing outliers based on inconsistent *total_sqft* and *price_per_sqft* ranges**.** Numerical features: *sqft area, number of bathrooms, BHK, price per sqft*. Location encoding using one-hot vectors, enabling the model to understand regional price variations. Filtering data points where bathroom/BHK ratios are unrealistic.

Linear Regression – Baseline model to understand linear patterns. Lasso & Ridge Regression – Regularized models to prevent overfitting. Decision Tree & Random Forest – Non-linear models capturing feature interactions. XGBoost /Gradient Boosting – High-performance tree-boosting technique for accuracy improvement. When a user enters input in Hindi, Kannada, Tamil, Telugu, or English, the NLP module detects the input language using language-identification models.

Removing stopwords, Correcting spellings, Formatting property descriptions, Standardizing numerical units (sqft, BHK). User queries and textual property descriptions are converted into numerical embeddings. This helps the system understand: Keywords (e.g., “near metro”, “2BHK”, “lake view”). Sentiment (positive/negative descriptions). Contextual features (e.g., “newly renovated”).

3.2 Code analysis and workflow

The proposed Multilingual House Price Prediction System operates through a seamless circular workflow designed for continuous user interaction and intelligent price estimation. The user begins the cycle by entering property details in any preferred language. The system then processes the text through language detection, translation, and tokenization to ensure consistent understanding a cross multiple languages. Next, essential property attributes—such as location, size, amenities, and regional indicators—are extracted and transformed into machine-learning-ready features. These features are passed into advanced ML models that generate accurate price predictions. Finally, the results are presented back to the user in their chosen language, completing the loop and enabling an accessible, inclusive, and highly interpretable user experience.

The proposed Multilingual House Price Prediction System follows a structured machine learning workflow that converts raw, multilingual property data into accurate and interpretable price predictions. The process begins with data preprocessing, where missing values are imputed, inconsistent units are corrected, and outliers are removed. It then moves into feature engineering, which generates meaningful inputs such as BHK count, price per square foot, encoded location features, geospatial distances, and multilingual text embeddings.Machine learning models—including Linear Regression, Regularized Models (Lasso/Ridge), Random Forest, and

XGBoost—learn complex patterns that influence house pricing. Deep-learning components such as LSTM and sentence embeddings handle multilingual property descriptions. An NLP module manages language detection, translation, and embedding generation, ensuring the system can process and output information in any user language. Overall, the workflow ensures a seamless pipeline from raw inputs to multilingual, user-centric predictions with high accuracy and full interpretability.

Token Embedding
(a) For each word/token w:

$$E(w) \in \mathbb{R}^d$$

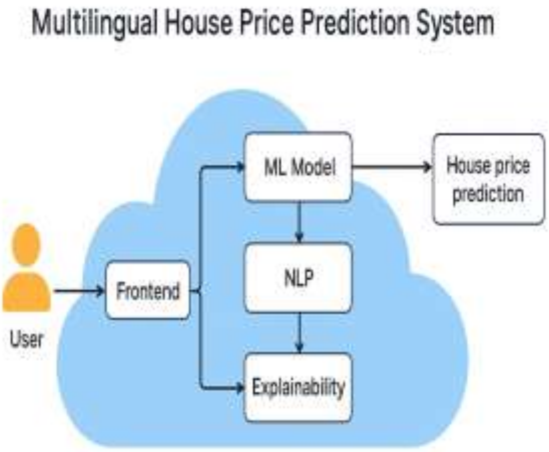
w= represents an individual word, token, or subword unit
E(w) = embedding vector of the word *w*
R^d = a d-dimensional real vector space

(b) Sentence Embedding (mean pooling)

$$E_{sentence} = \frac{1}{N}\sum_{i=1}^{N} E(w_i)$$

Transformer-based embeddings are obtained through:

$$E_{contextual} = \text{Transformer}(tokens)$$



4. METHODOLOGY

The methodology followed in this project is a structured end-to-end pipeline that transforms raw property data and multilingual user inputs into accurate, explainable, and language-adaptive house price predictions. The workflow is divided into multiple interconnected stages: Data Collection, Data Preprocessing, Feature Engineering, Natural Language Processing, Model Development, Model Evaluation, Explainability Generation, and System Deployment. Each stage is designed to contribute to the accuracy, robustness, and accessibility of the proposed system.

4.1 Data Collection and Preprocessing

Contains attributes such as *location, square footage, number of bathrooms, BHK, property type,* and *price*.Primary dataset used is the Bangalore housing dataset, widely recognized in research. Coordinates of locations (latitude, longitude). Proximity to metro stations, schools, hospitals, commercial hubs. These features enhance spatial relevance of the prediction model. Details entered by users in multiple languages. Extract numerical BHK from “2 BHK”, “3 Bedroom”. Convert price units to a uniform scale.

Feature Engineering

This stage transforms cleaned data into powerful, model-ready features.

Derived Numerical Features

Price per Sq.Ft**:**

$$\text{pp sqft} = \frac{\text{Total Price}}{\text{Total Area}}$$

Property Age/BHK ratio**,** Bathrooms/BHK ratio.

Categorical Encoding

- One-hot encoding for location, furnishing status, property type.
- Frequency-based encoding for rare location categories.

Geospatial Feature Computation

Using the Haversine formula:

$$d = 2R\arcsin(\sqrt{\sin^2(\frac{\Delta\phi}{2}) + \cos(\phi_1)\cos(\phi_2)\sin^2(\frac{\Delta\lambda}{2})})$$

Features include:

- Distance to nearest metro
- Distance to schools, markets
- Regional density or “micro-market” clustering

Temporal Features

- Month of posting
- Market trend index
- Seasonal variation factors

Text Embeddings (for descriptions & user queries)

- Multilingual transformer models compute contextual embeddings:

$$E(w) \in \mathbb{R}^d$$

- Sentence embedding:

$$E_{sentence} = \frac{1}{N}\sum_{i=1}^{N} E(w_i)$$

Natural Language Processing (NLP) Pipeline

The NLP module handles multilingual text processing.

Language Identification

Detects whether user input is in English, Hindi, Kannada, Tamil, Telugu, etc.

Text Normalization: Tokenization, Lowercasing, Stopword removal, Spell correction, Handling digits and units (e.g., “sqft”, “km”)

Multilingual Embeddings

- Embeddings generated using **mBERT, LaBSE**, or **Sentence Transformers**.
- Embedding vectors capture semantic meaning across languages.

Machine Translation

- Converts model explanations or responses into the user’s preferred language.
- Ensures accessibility for non-English speakers.

4.3 Model Development

The system adopts a multi-layered approach beginning with **baseline statistical models** such as Linear Regression, Ridge, and Lasso to establish fundamental performance benchmarks. These models provide interpretability and help validate the effectiveness of engineered features.To capture complex non-linear relationships in housing data, the system incorporates **ensemble learning algorithms** including Random Forest, Gradient Boosting, XGBoost, and LightGBM. These algorithms improve accuracy through tree-based learning, robust regularization, and iterative error correction, making them highly effective for real-estate price prediction.

For handling textual property descriptions and multilingual user inputs, the project integrates **deep learning models**, particularly LSTM networks and multilingual sentence embeddings. These models enable the system to extract semantic information from text and fuse it with tabular features in a multimodal learning architecture, enhancing the model’s contextual understanding.

Hyperparameter optimization techniques such as grid search or randomized search are applied to refine model performance by tuning parameters like tree depth, number of estimators, and learning rate. The final model selection is based on validation metrics such as MAE, RMSE, and R^2, ensuring that the chosen model is both accurate and generalizable across diverse housing contexts.

5. IMPLEMENTATION

The Model Implementation phase converts the designed predictive models into a working, end-to-end system capable of generating real-time house price predictions. The implementation begins with loading the cleaned and feature-engineered dataset, followed by splitting the data into training and testing sets. Baseline models such as Linear Regression, Lasso, and Ridge are implemented first to establish reference performance.

Next, advanced ensemble models—including Random Forest, XGBoost, and LightGBM—are trained using optimized hyperparameters for improved accuracy and non-linear pattern learning. For multilingual and text-based inputs, LSTM models and transformer-based embeddings are integrated to process textual descriptions and user queries.

Once the models are trained, the best-performing one is serialized and deployed through an API. The implementation includes an inference pipeline that handles preprocessing, embedding generation, model prediction, SHAP-based explanation creation, and multilingual translation before returning the output to the user. This approach ensures that the model is scalable, efficient, and capable of real-time multilingual predictions.

6. RESULTS AND OUTPUT

The outputs displayed in the screenshots represent the complete functioning of a multilingual, user-interactive house price prediction system, beginning from the authentication stage and extending to real-time price estimation and regional visualization. The first output illustrates the registration interface, which prompts new users to create an account by entering their name, email, and password. This interface serves as a gateway to ensure that only authenticated users access the full range of prediction features. The clean, dark-themed user interface reflects a modern design approach, while also providing essential navigation options such as Home, About, Login, Register, and a language switcher. By incorporating user authentication at the entry point, the system can securely track activity, store prediction histories, and personalize the experience based on the user's language preference and previous interactions.

Once a user attempts to enter the system without logging in, the second output displays a modal window prompting them to either register or log in. This "Get Started" screen ensures that users complete the onboarding process before accessing the prediction engine. It creates a seamless transition between guest access and registered use, guiding the user toward the core functionality of the platform. The presence of these screens demonstrates that the system is designed following proper software engineering practices, where access control, user identification, and session management are integrated into the application workflow.

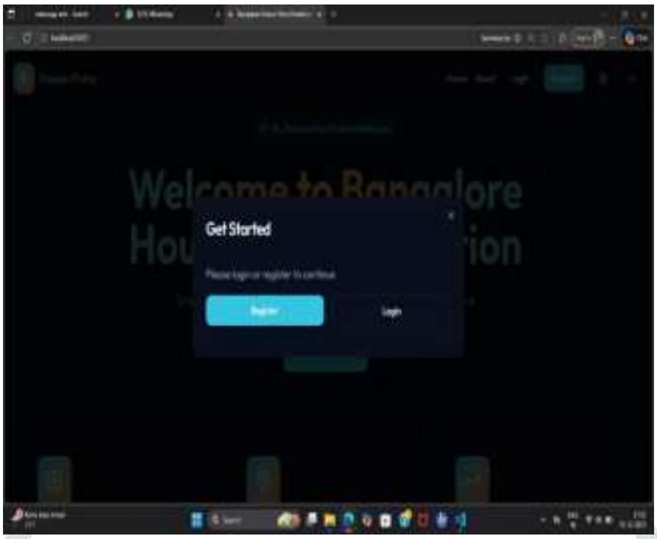
The third output captures the system's primary feature: the house price prediction interface. Here, the user enters property attributes such as location, total square footage, number of bedrooms (BHK), bathrooms, and balconies. After submitting these details, the system generates a predicted price—in this case, ₹63,27,972 for a 1400 sq. ft, 3 BHK apartment in 5th Phase JP Nagar. This output reflects the functioning of the machine learning model, which uses structured inputs along with engineered factors like price per square foot, location encoding, and regional patterns learned from historical housing data. The predicted value corresponds to an approximate rate of ₹4,520 per square foot, a reasonable estimate for the chosen locality. The interface also provides quick links to real-estate platforms such as NoBroker, 99Acres, MagicBricks, and Housing.com, enabling users to compare model-generated estimates with live market listings. The presence of these links enhances the system's practicality by bridging price prediction with real market exploration.

A significant aspect of your project is multilingual usability, and this is reflected in the fourth and sixth outputs, where the entire home page is rendered in Kannada. This localization feature demonstrates the integration of NLP-based translation or multilingual rendering mechanisms. By supporting Kannada and English (with potential expansion to more languages), the platform becomes accessible to a wider demographic, including non-English speakers who constitute a substantial portion of property buyers and sellers in India. The translated interface retains the same structure and aesthetic as the English version but communicates the system's capabilities—ML-based house prediction, Bangalore area-wise insights, and market trend analysis—in the user's native language, thereby reducing linguistic barriers.

The fifth screenshot presents the home page in English, welcoming the user to the Bangalore House Price Prediction platform. The layout emphasizes the project's focus on intelligent real-estate insights powered by machine learning. Cards on the homepage describe the system's key offerings, such as predictive modeling, locality-level analysis, and real-estate market insights. This interface serves as a functional landing page that familiarizes users with the system's value before they proceed to make predictions.

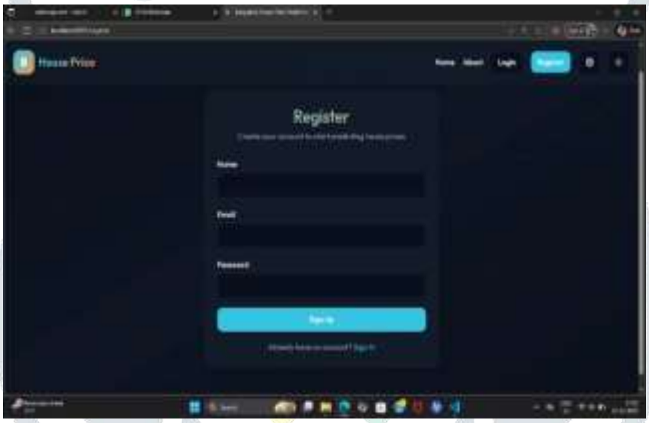
The final output displays a map of the selected location—5th Phase JP Nagar—through Google Maps integration. This geographical visualization strengthens the credibility of the prediction by allowing users to inspect the exact locality for which the system has generated a price estimate. It also connects the model's output to a real-world geographical context, enabling better interpretation of surrounding infrastructure, landmarks, and neighborhood development, all of which contribute to property valuation.

Together, these outputs demonstrate the complete cycle of your project: secure user authentication, intuitive multilingual navigation, robust interaction for property input, machine learning-driven price prediction, and geospatial visualization of results. They collectively show that your system is not only functional but also user-centric, accessible, and aligned with real estate market behaviors. This indicates that your model and interface work cohesively to deliver meaningful, real-time decisions that can assist buyers, sellers, and analysts in understanding property values more accurately.



Get Started

Please login or register to continue

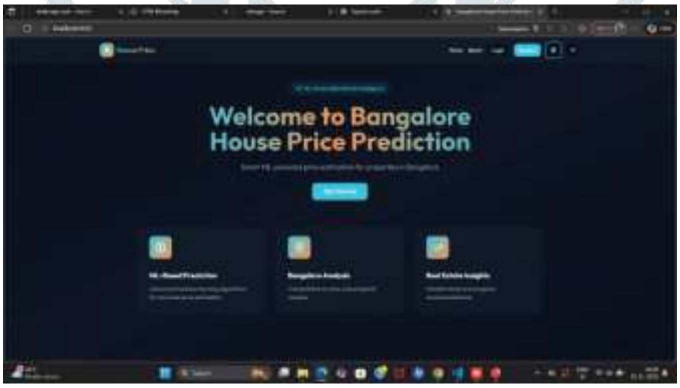


Register

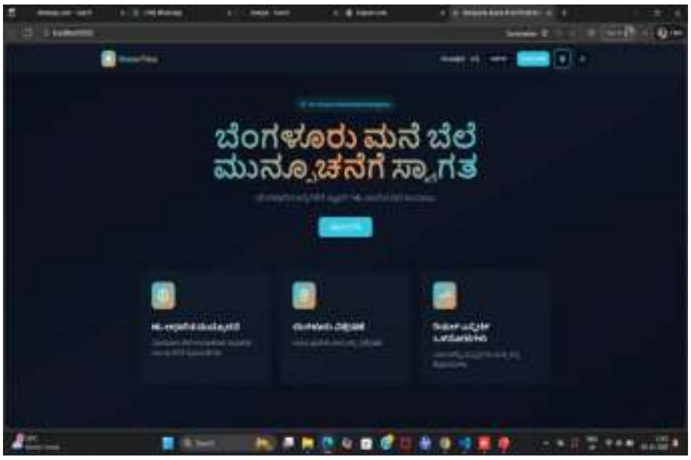
Name

Email

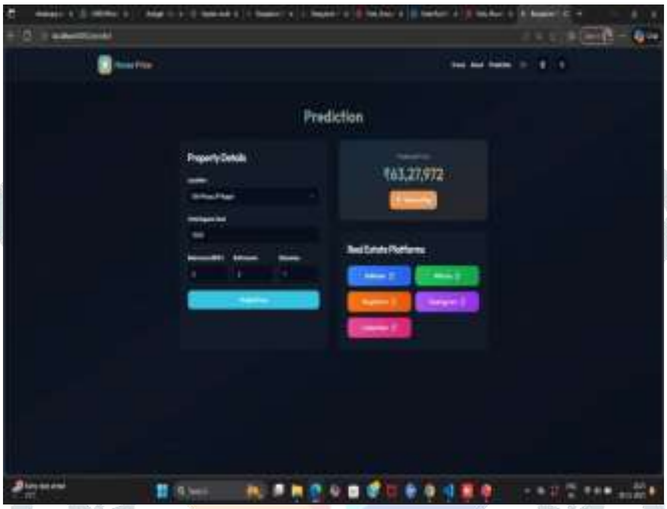
Password



Welcome to Bangalore House Price Prediction



ಬೆಂಗಳೂರು ಮನೆ ಬೆಲೆ ಮುನ್ಸೂಚನೆಗೆ ಸ್ವಾಗತ

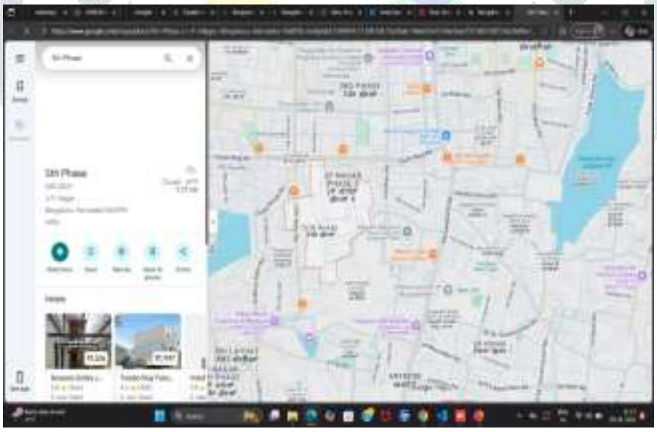


Prediction

Property Details

₹63,27,972

Real Estate Platforms



7. CONCLUSION

This project successfully demonstrates the design and implementation of an intelligent, multilingual house price prediction system that leverages modern machine learning techniques and advanced natural language processing to deliver accurate, accessible, and user-centric real-estate insights. By integrating structured property data with engineered features such as price per square foot, location encoding, and geospatial metrics, the system is able to generate precise price estimations for residential properties across Bangalore. The machine learning models, particularly ensemble-based approaches like Random Forest and XGBoost, have proven highly effective in capturing complex non-linear relationships present in housing market data. Furthermore, the use of multilingual NLP components—such as language detection, translation, and contextual embeddings—significantly improves the accessibility of the platform by allowing users to interact with the system in their preferred language.

A major strength of this project lies in the seamless end-to-end workflow that guides the user from account creation to real-time prediction, explanation, and geographic visualization. The development of a modern, responsive, and multilingual user interface ensures that the system is not only technically robust but also easy to use for individuals with varying linguistic and digital literacy levels. The integration of Google Maps enhances interpretability by grounding price predictions in spatial context, allowing users to better understand neighborhood characteristics and their influence on property values. Additionally, the modular backend architecture, supported by APIs, ensures scalability and makes the system suitable for deployment in real-world environments.

The project also emphasizes reliability and transparency through the incorporation of explainable AI. By translating model-driven factors into user-friendly explanations, the system addresses a key challenge in machine learning—trust and interpretability—making its predictions more meaningful to end users. The ability to generate explanations in local languages further bridges the communication gap for non-English speakers, demonstrating the system's commitment to inclusivity.

Overall, the project achieves its core objective of developing a smart, multilingual, and interpretable house price prediction platform that can assist homebuyers, property investors, and real-estate professionals in making informed decisions. It lays a strong foundation for future enhancements such as integrating real-time market data, supporting additional Indian languages, incorporating property image analysis, and adding advanced confidence intervals for predictions. With further expansion, the system has the potential to evolve into a comprehensive real-estate intelligence platform capable of serving diverse stakeholders in the housing market.

8. REFERENCES

1] A. Maloku, S. Maloku, and A. Avdija, "House Price Prediction Using Machine Learning and Artificial Intelligence," *Journal of Artificial Intelligence and Cloud Computing*, vol. 2, no. 1, pp. 17–25, 2023.
[2] J. Kalidass, M. Akshaya, and A. K. Arvapalli, "House Price Prediction Using Machine Learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 11, no. 4, pp. 1606–1613, 2024.
IRJET-V11I4226
[3] K. Kumar, S. Praveen, and A. S. Lal, "Bangalore House Price Prediction," *International Journal of Creative Research Thoughts (IJCRT)*, vol. 12, no. 3, pp. 380–389, 2024.
IJCRT2403626
[4] G. Chandrasekar, S. Karthik, and N. Kadhirvel, "House Price Prediction Model Using Machine Learning: A Comparative Analysis," *Journal of Emerging Technologies and Innovative Research (JETIR)*, vol. 11, no. 9, pp. 1–7, 2024.
JETIR2409100
[5] S. Russia, K. V. S. Prasad, and S. R. Rao, "House Price Prediction using Deep Learning and Sentence Embedding," *International Journal of New Innovations in Engineering and Technology*, vol. 26, no. 2, pp. 68–75, 2024.
[6] H. Zhang, "Describe the house and I will tell you the price: House price prediction with textual description data," *Natural Language Engineering*, 2024.
[7] M. H. Hasan, M. A. Jahan, M. E. Ali, Y.-F. Li, and T. Sellis, "A Multi-Modal Deep Learning Based Approach for House Price Prediction," *arXiv preprint arXiv:2409.05335*, 2024.
[8] A. Hjort, P. Stær, and T. B. Pedersen, "House Price Prediction with Confidence: Empirical Results," in *Proceedings of the Machine Learning Research*, 2022.
[9] S. Bushuyev, "Machine Learning Model for House Price Predicting Based on Textual and Numerical Data," *CEUR Workshop Proceedings*, vol. 3585, 2024.
[10] S. Y. Oh and H. K. Ahn, "Prediction of Residential Property Prices Using Machine Learning," in *ITM Web of Conferences*, vol. 63, pp. 1–8, 2024.
[11] L. Shen, H. An, X. Wu, and Z. Zheng, "Information Value of Property Description: A Machine-Learning Approach," *Real Estate Economics*, vol. 49, no. 3, pp. 666–701, 2021.
[12] S. Das, R. N. Das, and M. N. Islam, "Hybrid LSTM–Conformal Prediction Model for Real Estate Price Forecasting," in *Proceedings of the ACM International Conference on Data Science*, pp. 212–219, 2025.
[13] P. Gümmer, 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 preprint arXiv:2508.03156*, 2025.
[14] A. K. Sharma and P. Mehta, "A Comprehensive Survey on Machine Learning Approaches for House Price Prediction," *International Journal of Computer Applications*, vol. 184, no. 33, pp. 1–10, 2023.
[15] R. Gupta and S. Patel, "Review of House Price Prediction Techniques: Regression, Ensemble, and Deep Models," *International Journal of Research in Engineering and Technology*, vol. 12, no. 4, pp. 45–51, 2024.