Kimat: AI-Driven Geospatial and Machine Learning-Based House Price Prediction System

Dr. M.K. Jayanthi Kannan[1], Ayush Shukla[2], Krish Pandey[3], Aditya Anand[4], Arihant Jain[5], Jasmine Chauhan[6], Atreya Malviya[7]

[1]*Professor, SCAI (Supervisor), School of Computing Science Engineering and Artificial Intelligence, VIT Bhopal University, Bhopal-Indore Highway, Kothrikalan, Sehore, Madhya Pradesh - 466114*

[2,3,4,5,6,7]*UG Students, School of Computing Science Engineering and Artificial Intelligence, VIT Bhopal University, Bhopal-Indore Highway, Kothrikalan, Sehore, Madhya Pradesh - 466114*

Abstract*—The real estate market is rapidly evolving with complex pricing dynamics, making manual property valuation subjective and prone to human error. The "Kimat" platform is designed as an AI-driven house price prediction system that transforms historical housing data into accurate real-time price predictions. By leveraging a focused regression workflow, the system analyzes structured property and locality attributes to output estimated prices accompanied by confidence-oriented explanations. Utilizing machine learning models such as Random Forest, which evaluate key property features and market trends, Kimat provides fast, reliable, and data-driven preliminary valuations. The platform compares Multiple Linear Regression against ensemble techniques using R^2, RMSE, and MAE, explains the strongest price drivers, and exposes the selected model through a simple, pincode-aware prediction interface. This empowers homebuyers, real estate agents, and property investors by democratizing real estate valuation and saving time and appraisal costs through an intuitive web interface.

***Index Terms*—House Price Prediction, PropTech, Machine Learning, Random Forest, Multiple Linear Regression, XGBoost, Real Estate Valuation, Predictive Analytics, Kimat.**

I. INTRODUCTION

The increasing complexity of the real estate market, coupled with subjective and error-prone manual property valuation, underscores the necessity of advanced predictive models for housing prices. The AI-driven House Price Prediction System, named "Kimat", is developed to learn patterns from historical housing data to make reliable price predictions. The objective is to accurately estimate residential property prices by training models on real-world datasets encompassing diverse property attributes. By leveraging architectural, temporal, and location-based attributes such as size, age, and neighborhood, the system drives prediction accuracy. The system processes historical housing datasets to ensure data quality, identifies the strongest predictors of property value through feature extraction, and applies multiple regression and ensemble machine learning techniques for accurate modeling. Ultimately, this culminates in a responsive web application that generates real-time property valuations, bridging traditional real estate practices with modern machine learning to reduce subjectivity and improve market transparency in the PropTech domain. The scientific rationale behind Kimat lies in several key aspects. Firstly, supervised learning on structured data allows the model to capture non-linear relationships between property characteristics and sale price. Secondly, ensemble methods such as Random Forest reduce variance and resist overfitting. Thirdly, a pincode-aware interface places every estimate in its local market context, so that a price is never shown without the inventory, demand and price bands that surround it.

II. LITERATURE REVIEW OF EXISTING SYSTEMS

The development of the AI-driven house price prediction system is grounded in a comprehensive review of existing methodologies, comparing multiple regression models against neural networks, boosted trees and data-stream techniques. Six research papers were selected for Review I. For each paper, the objective, technology used, methodology used,

efficiency and issues are summarized below, followed by the review figure prepared by the team member responsible for that paper.

Table 1: Comparison of the six reviewed papers

Paper	Year	Main model(s)	Dataset	Key outcome
1. Adetunji et al.	2022	Random Forest	Boston Housing (506 rows)	Error margin about ±5%
2. Foryś	2022	Regression vs. ANN	Control sample (N=229)	Neither model best; over-estimation
3. Shi et al.	2022	Batch vs. data-stream	Home sales time series	Data-stream beats regression
4. Geerts et al.	2023	Survey (93 studies)	1992-2021 literature	Spatial data is key; ML rising
5. Sharma et al.	2024	XGBoost, RF, SVR, MLP, MLR	Ames, Iowa housing	XGBoost best performer
6. Pita et al.	2026	Systematic review	Published ML studies	No universal best; temporal gap

2.1. Paper 1: "House Price Prediction using Random Forest Machine Learning Technique"

Focuses on predicting individual house prices rather than relying solely on broad House Price Indices (Adetunji et al., Procedia Computer Science, 2022). *Journal: Procedia Computer Science, Vol. 199, 2022, pp. 806-813 | DOI: 10.1016/j.procs.2022.01.100 | Reviewed by: Atreya Malviya (25BAI11420).* The key objectives of the study include: predicting individual house prices from property and location information; investigating whether Random Forest can learn the relationship between housing characteristics and sale price; and evaluating the model by comparing predicted and actual prices.

Paper 1 Atreya Malviya 25BAI11420	Objective	Technology Used	Methodology Used	Efficiency	Issues
Title: House Price Prediction using Random Forest Machine Learning Technique **Journal:** Procedia Computer Science, Vol. 199, 2022, pp. 806–813 **Year:** 2022 **DOI:** 10.1016/j.procs.2022.01.100 **URL:** https://doi.org/10.1016/j.procs.2022.01.100	**Objective:** **1.** Predict individual house prices rather than relying only on a broad House Price Index. The study argues that individual valuation needs property and location-related information. **2.** Investigate whether Random Forest can learn the relationship between housing characteristics and sale price. The aim is to obtain predictions that are close to observed values. **3.** Evaluate the proposed model by comparing predicted and actual house prices. The comparison is used to judge whether the approach is practically acceptable.	**Technologies:** **1.** Random Forest Regression is the main machine-learning technique used for prediction. It combines multiple decision trees to obtain an ensemble prediction. **2.** Supervised learning is used because historical observations contain both explanatory housing attributes and a known target price. **3.** The experiment uses the UCI Boston Housing dataset. It contains 506 observations and 14 features for the reported experiment. **4.** The model is implemented as a machine-learning regression workflow that produces predicted prices for comparison with actual prices.	**Methodology:** **1.** Historical housing observations are treated as supervised training examples, with housing characteristics acting as input variables and price as the target. **2.** A Random Forest model is constructed from multiple decision trees. The ensemble combines tree-level predictions rather than depending on one tree. **3.** The trained model generates a predicted price for each test observation. These predictions are compared with the corresponding actual prices. **4.** The study evaluates the difference between predicted and actual values and reports the resulting error margin as an indication of model acceptability.	**Efficiency / Results:** **1.** The study reports an acceptable predicted value compared with actual values, with an error margin of approximately ±5% on the selected dataset. **2.** Random Forest can represent nonlinear relationships between housing characteristics without requiring a single linear relationship between every feature and price. **3.** The ensemble structure makes the approach less dependent on the errors of one individual decision tree and provides a useful tabular-data baseline. **4.** The result demonstrates that tree-based ensemble learning can be useful for individual house-price prediction when suitable historical data are available.	**Issues / Limitations:** **1.** The experiment is based on the Boston Housing dataset, which is relatively small with 506 observations. A small benchmark cannot represent every housing market. **2.** The reported performance may not transfer directly to another city or country because housing prices are strongly affected by local market conditions. **3.** Location, city and population-related factors can materially influence house prices. A limited feature set may therefore restrict generalization. **4.** An error margin reported on one dataset should not be interpreted as a universal accuracy guarantee for future real-estate markets.



Figure 1: Literature Review of House Price Prediction using Random Forest (feature distributions, correlation matrix, actual vs. predicted)

The technology used in the study includes Random Forest Regression, which combines multiple decision trees into an ensemble prediction, applied as supervised learning on the UCI Boston Housing dataset (506 observations, 14 features). The methodology follows: historical observations are treated as supervised training examples; the trained forest generates a predicted price for every test observation, and the difference between predicted and actual values is reported as the error margin. The study demonstrates efficiency with an acceptable predicted value and an error margin of approximately ±5%.

Random Forest represents non-linear relationships without assuming a single linear link between each feature and price, and the ensemble is less dependent on the errors of any one tree. However, key challenges include: the Boston dataset is small (506 observations), results may not transfer to another city or country, a limited feature set restricts generalization, and an error margin from one dataset is not a universal accuracy guarantee. Relevance to Kimat: Kimat adopts Random Forest as its primary model for the same reason, and treats the Boston result as a baseline expectation rather than a guarantee. Because the paper warns that a small, single-city dataset limits generalization, Kimat also trains on Ames and Kaggle data

2.2. Paper 2: "Machine learning in house price analysis: regression models versus neural networks"

Evaluates automated house-price determination using multiple regression versus artificial neural networks (Foryś, Procedia Computer Science, 2022). *Journal: Procedia Computer Science, Vol. 207, 2022, pp. 435-445 | DOI: 10.1016/j.procs.2022.09.078 | Reviewed by: Arihant Jain (25BAI10246).* The key objectives of the study include: studying automatic house-price determination with multiple regression and machine-learning approaches; comparing regression with neural networks to see whether more complex models are necessarily better; and assessing both on a control group with known prices.

Paper 2 Arihant Jain 25BAI10246	Objective	Technology Used	Methodology Used	Efficiency	Issues
Title: Machine learning in house price analysis: regression models versus neural networks **Journal:** Procedia Computer Science, Vol. 207, 2022, pp. 435–445 **Year:** 2022 **DOI:** 10.1016/j.procs.2022.09.078 **URL:** https://doi.org/10.1016/j.procs.2022.09.078	**Objective:** **1.** Study automatic house-price determination using multiple regression and machine-learning approaches. The work is motivated by the growing use of automated valuation in real-estate appraisal. **2.** Compare regression models with neural-network approaches using property characteristics as predictors. The comparison examines whether more complex ML models are necessarily better. **3.** Assess the models on a control group for which property characteristics and known prices are available. The purpose is to compare estimated values with observed values. **4.** Investigate how machine learning can support the classical valuation process. The study also discusses practical acceptance of automated valuation models.	**Technologies:** **1.** Multiple / multivariate regression is used as a conventional statistical baseline. It estimates price from a set of explanatory property variables. **2.** Artificial Neural Networks are evaluated as a flexible machine-learning alternative. The network can represent nonlinear relationships between inputs and price. **3.** Automated Valuation Model concepts frame the application. The models are intended to support repeated and large-scale property valuation. **4.** Input normalization is applied for neural-network modeling so that variables can be represented on an appropriate numerical scale.	**Methodology:** **1.** The same general property characteristics are prepared for the compared model families. This makes it possible to evaluate differences attributable to the modeling approach. **2.** A multivariate regression model is trained for house-price estimation. Its predictions provide a comparatively interpretable baseline. **3.** Neural-network models are trained after preprocessing and normalization. Their outputs are then compared with the regression predictions. **4.** A control sample with known property characteristics and prices is used to examine prediction behavior. The study specifically observes differences and overestimation in predicted values. **5.** The paper compares model quality rather than assuming that a more	**Efficiency / Results:** **1.** The regression approach provides a useful baseline for automated valuation and is comparatively straightforward to interpret. It can be useful when relationships are sufficiently stable. **2.** The neural-network models provide a flexible way to model nonlinear relationships. However, greater complexity does not automatically translate into the best result for the study sample. **3.** Both approaches can process property information quickly and can support the valuation of many properties. This is one of the practical motivations for AVM systems. **4.** The paper demonstrates the importance of empirical model comparison. Model selection should be based on observed performance for the target dataset rather than model popularity.	**Issues / Limitations:** **1.** The study reports that neither the regression nor neural-network approach produced the highest-quality model for the investigated sample. This shows that model selection remains data dependent. **2.** The models overestimated prices for most properties in the control sample. Such systematic differences can be important in an automated valuation setting. **3.** Performance can change when the property population, market, feature set or sample size changes. Results from one study should therefore not be generalized without validation. **4.** Automated predictions should be interpreted as decision-support information. They should not automatically replace professional appraisal or market-specific judgment.

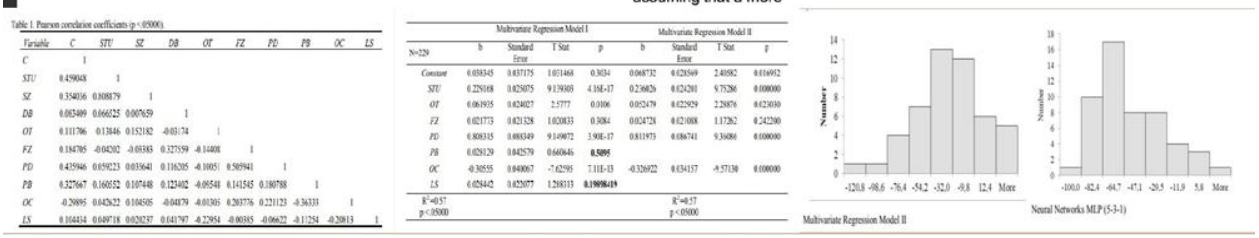


Figure 2: Literature Review of Regression Models versus Neural Networks (Pearson correlations, regression table, residual histograms)

The technology used in the study includes: multivariate regression as a conventional statistical baseline, Artificial Neural Networks (MLP 5-3-1) as a flexible non-linear alternative, input normalization, and Automated Valuation Model (AVM) concepts. The methodology follows: the same property characteristics are prepared for both model families, the regression and neural models are trained after preprocessing and normalization, and their predictions on a control sample are compared, with attention to overestimation. The study demonstrates efficiency with: a regression baseline that is straightforward to interpret and a neural model that captures non-linear relationships; both can value many properties quickly. The study shows that model selection must rest on observed performance, not model popularity. However, key challenges include: neither approach produced the highest-quality model for the sample, both overestimated prices for most control properties, results change with the market and sample size, and automated outputs should support rather than replace professional appraisal. Relevance to Kimat: Kimat keeps Multiple Linear Regression as its interpretable baseline but does not assume that a more complex neural model would be better. The over-estimation finding motivates the likely-range display and the confidence value shown with each valuation.

2.3. Paper 3: "Predicting home sale prices: A review of existing methods and illustration of data stream methods for improved performance"

Addresses the limits of batch learning in dynamic housing markets (Shi, Guan, Zurada & Levitan, WIREs Data Mining and Knowledge Discovery, 2022). *Journal: WIREs Data Mining and Knowledge Discovery, Vol. 12(2), 2022, e1435 | DOI: 10.1002/widm.1435 | Reviewed by: Ayush Shukla (25BAI10340).* The key objectives of the study include: reviewing methods used for residential sale-price prediction; identifying the limitations of conventional batch-learning approaches as markets change over time; and investigating data-stream learning that incorporates new transactions as they arrive.

Paper 3 Ayush Shukla 25BAI10340	Objective	Technology Used	Methodology Used	Efficiency	Issues
Title: Predicting home sale prices: A review of existing methods and illustration of data stream methods for improved performance **Journal:** WIREs Data Mining and Knowledge Discovery, Vol. 12(2), 2022, e1435 **Year:** 2022 **DOI:** 10.1002/widm.1435 **URL:** https://doi.org/10.1002/widm.1435	**Objective:** **1.** Review methods that have been used for residential property sale-price prediction. The paper examines how researchers have approached this problem using different modeling strategies. **2.** Identify limitations of conventional batch-learning approaches. A key concern is that housing transactions and market conditions change over time. **3.** Investigate data-stream learning as an alternative to treating historical housing data as permanently fixed. The approach is intended to incorporate new observations as they become available. **4.** Compare the data-stream perspective with traditional regression. The goal is to understand whether continual updating can improve predictive performance.	**Technologies:** **1.** Traditional multiple regression is used as a reference method for residential price prediction. It represents a conventional batch-learning approach. **2.** Machine-learning prediction methods are reviewed across existing literature. The review considers how different approaches process historical housing information. **3.** Data-stream learning is the central methodological direction. New transactions are treated as a continuing stream rather than a dataset that never changes. **4.** Incremental / evolving-model concepts allow predictions to be updated as additional observations arrive. This is relevant to real-estate markets with changing conditions.	**Methodology:** **1.** The paper first reviews existing home-sale-price prediction methods and their reported characteristics. This establishes the limitations of conventional approaches. **2.** Historical observations are viewed as an evolving stream of housing transactions. The model can therefore incorporate information that arrives after the original training period. **3.** Prediction performance is evaluated while the underlying data distribution changes. This better represents the temporal nature of a real housing market. **4.** The data-stream approach is compared with traditional regression. The comparison examines whether continual adaptation produces better results. **5.** The study uses the results to motivate prediction systems that can update rather than remain permanently tied to an old	**Efficiency / Results:** **1.** The study reports that the data-stream approach outperformed traditional regression in the investigated setting. This supports the value of adapting to incoming housing observations. **2.** Continual updating can make a model more responsive to changing market conditions. This is particularly relevant when prices, buyer behavior and economic conditions evolve. **3.** The paper highlights temporal information as an important part of house-price prediction. A model that ignores time may become less representative as the market changes. **4.** The data-stream perspective provides a useful extension beyond static batch prediction. It can inform future real-time or periodically retrained house-price systems.	**Issues / Limitations:** **1.** Traditional batch models may become outdated when the statistical characteristics of the housing market change. Their historical training data may no longer represent current conditions. **2.** Static datasets do not naturally capture the arrival of new transactions. This creates a gap between an offline model and a continuously changing real-world market. **3.** Data-stream learning requires a reliable flow of new observations. Missing, delayed or biased transactions can affect the updating process. **4.** Continual updating introduces additional implementation and monitoring requirements. The system must control how new data influence the model and its future predictions.



Figure 3: Literature Review of Batch Learning versus Data-Stream Methods for Home Sale Price Prediction

The technology used in the study includes: traditional multiple regression as the batch-learning reference, a review of machine-learning predictors, and incremental / evolving data-stream models that update as new observations arrive. The methodology follows: historical sales are viewed as an evolving stream of transactions, performance is evaluated while the data distribution changes, and the data-stream approach is compared with traditional regression. The study demonstrates efficiency, with the data-stream approach outperforming traditional regression in the investigated setting, showing that continual updating makes a model more responsive to changing prices, buyer behavior, and economic conditions, and that temporal information matters. However, key challenges include: batch models become outdated when the market shifts, static datasets miss newly arriving transactions, streaming needs a reliable flow of unbiased observations, and continual updating adds implementation and monitoring overhead. Relevance to Kimat: The temporal gap identified here is why Kimat is built as a modular pipeline whose Training Module can be re-run on fresh transactions, so the pre-trained model can be refreshed periodically instead of staying tied to an old dataset.

2.4. Paper 4: "A Survey of Methods and Input Data Types for House Price Prediction"

is a comprehensive survey that maps research trends in residential valuation and stresses spatial information (Geerts, vanden Broucke & De Weerdt, ISPRS International Journal of Geo-Information, 2023). *Journal: ISPRS International Journal of Geo-Information, Vol. 12(5), 2023, Article 200 | DOI: 10.3390/ijgi12050200 | Reviewed by: Aditya Anand (25BAI10258).* The key objectives of the study include: systematically reviewing methods and input data types in house-price prediction; mapping research trends; investigating the role of spatial and geospatial information; and identifying opportunities for advanced inputs and models.

Paper 4 Aditya Anand 25BAI10258	Objective	Technology Used	Methodology Used	Efficiency	Issues
Title: A Survey of Methods and Input Data Types for House Price Prediction **Journal:** ISPRS International Journal of Geo-Information, Vol. 12(5), Article 200 **Year:** 2023 **DOI:** 10.3390/ijgi12050200 **URL:** https://doi.org/10.3390/ijgi12050200	**Objective:** **1.** Systematically review methods and input-data types used in house-price prediction. The study provides a broad view of how the field has developed. **2.** Map research trends in residential property valuation by examining model types and the data used as predictors. The review covers studies published across a long time period. **3.** Investigate the role of spatial and geospatial information. House prices depend not only on building characteristics but also on location and neighborhood context. **4.** Identify opportunities for advanced input data and modeling methods. The paper points toward unstructured, complex spatial data and newer learning approaches.	**Technologies:** **1.** Systematic literature review is used to organize and analyze existing house-price prediction research. The study identifies 93 relevant articles published between 1992 and 2021. **2.** Conventional and machine-learning models include Multiple Regression, SVM, ANN, Random Forest and Gradient Boosted Trees. **3.** Spatial and geospatial data are considered important input types. These can represent location, neighborhood and other spatial relationships. **4.** Advanced inputs such as graph, image and textual data are discussed as opportunities for future valuation systems. Deep learning and tailored methods are also identified.	**Methodology:** **1.** The authors search and screen scientific literature using a systematic-review process. Relevant studies are categorized according to prediction methods and input-data characteristics. **2.** Each identified study is examined for model novelty and data novelty. This allows the research landscape to be mapped rather than focusing on one individual prediction model. **3.** Cluster analysis is used to identify trends in the property-valuation domain. The analysis connects modeling approaches with the kinds of data used. **4.** The reviewed studies include structural, temporal, socioeconomic, environmental, point-of-interest and spatial variables. This shows the multidimensional nature of house-price prediction.	**Efficiency / Results:** **1.** The review shows that conventional models and traditional input data remain predominant in house-price research. However, advanced techniques and innovative data sources are gradually increasing. **2.** Spatial information is highlighted as an important part of property valuation. This supports including location-related information when a project has access to reliable spatial features. **3.** The survey identifies growing use of machine learning, deep learning and nontraditional data sources. These developments provide directions for improving predictive systems. **4.** The paper is valuable as a research map because it compares broad trends across 93 studies. It helps justify why multiple algorithms and richer features should be considered.	**Issues / Limitations:** **1.** Many studies still depend on conventional models and traditional inputs. This can limit the ability of a model to represent complex spatial and socioeconomic relationships. **2.** Spatial dependence and spatial heterogeneity make property valuation difficult. Nearby properties may share information, while different regions can follow different price relationships. **3.** Advanced spatial, image, graph and textual data are less commonly used than conventional tabular variables. This leaves opportunities for further research. **4.** The review covers studies published through 2021. Newer algorithms, datasets and applications therefore require continued research and evaluation.

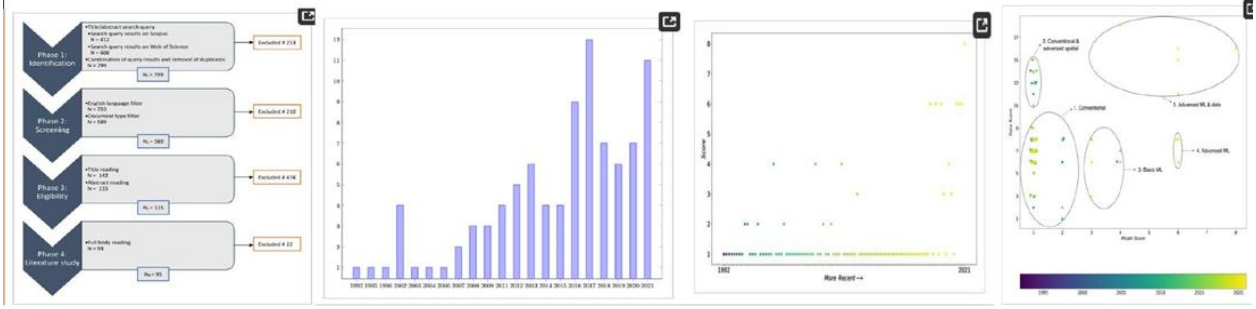


Figure 4: Literature Review of Methods and Input Data Types (data selection workflow, publication trend, model/data novelty clusters)

The technology used in the study includes: a systematic literature review of 93 articles published between 1992 and 2021, covering Multiple Regression, SVM, ANN, Random Forest, and Gradient Boosted Trees, together with spatial, graph, image and textual inputs. The methodology follows four phases: authors search and screen the literature (identification, screening, eligibility, literature study), score each study for model novelty and data novelty, and use cluster analysis to link modelling approaches to the data they use. The study demonstrates efficiency with: conventional models and traditional inputs still dominate, spatial information is an important part of valuation, and machine learning, deep learning, and non-traditional data are growing. This justifies testing several algorithms and richer features. However, key challenges include: many studies depend on conventional inputs, spatial dependence and heterogeneity complicate valuation, advanced data types are rarely used, and the review ends in 2021, so newer methods need continued evaluation. Relevance to Kimat: The survey justifies Kimat's inclusion of locality information: the pincode-level market read (average price, price per sq ft, inventory and price bands) is the project's practical form of spatial context for a non-technical user.

2.5. Paper 5: "An Optimal House Price Prediction Algorithm: XGBoost"

Treats prediction as a supervised regression problem and benchmarks several algorithms (Sharma, Harsora & Ogunleye, Analytics, 2024). *Journal: Analytics, Vol. 3(1), 2024, pp. 30-45 | DOI: 10.3390/analytics3010003 | Reviewed by: Jasmine Chauhan (25BAI11279).* The key objectives of the study include: treat house-price prediction as supervised regression; compare several algorithms rather than XGBoost alone; identify the strongest model; and identify the factors that most influence housing cost.

Paper 5 Jasmine Chauhan 25BAI11279	Objective	Technology Used	Methodology Used	Efficiency	Issues
Title: An Optimal House Price Prediction Algorithm: XGBoost **Journal:** Analytics, Vol. 3(1), 2024, pp. 30–45 **Year:** 2024 **DOI:** 10.3390/analytics3010003 **URL:** https://doi.org/10.3390/analytics3010003	**Objective:** **1.** Treat house-price prediction as a supervised regression problem. The target is to estimate property value from multiple independent housing characteristics. **2.** Compare several machine-learning algorithms rather than evaluating XGBoost in isolation. The study considers both tree-based and other regression approaches. **3.** Identify the algorithm with the strongest predictive performance on the selected housing data. The comparison is intended to guide model selection. **4.** Identify important factors influencing housing costs. Understanding influential variables can complement the numerical prediction itself.	**Technologies:** **1.** XGBoost is the principal boosted-tree algorithm investigated. It builds an ensemble of decision trees sequentially to improve predictive performance. **2.** Random Forest Regressor is included as a tree-ensemble comparison. It provides a useful non-linear baseline. **3.** Support Vector Regressor and Multilayer Perceptron are also evaluated. Multiple Linear Regression provides a traditional statistical baseline. **4.** The experiment uses the Ames City, Iowa housing dataset. The dataset contains detailed property characteristics suitable for supervised regression.	**Methodology:** **1.** The house-price task is formulated as a regression problem. Housing attributes are used as independent variables and sale price is used as the dependent target. **2.** Several algorithms are trained and compared using the same general prediction task. This includes XGBoost, SVR, Random Forest, MLP and Multiple Linear Regression. **3.** Model predictions are evaluated using quantitative performance measures. The comparison is used to identify the strongest-performing algorithm. **4.** The study also examines independent variables associated with housing cost. This connects predictive modeling with interpretation of price drivers. **5.** The final model choice is based on comparative empirical performance rather than assuming one algorithm	**Efficiency / Results:** **1.** The study reports XGBoost as the best-performing model among the algorithms compared. This directly supports considering XGBoost as a candidate model in our project. **2.** Boosted trees can capture nonlinear relationships and interactions among tabular housing features. This is useful because property prices rarely depend on one feature independently. **3.** The use of the Ames Housing dataset makes the study highly relevant to an Ames-based project. It provides a useful benchmark for algorithm comparison. **4.** The paper demonstrates why a project should compare several candidate models before selecting the final predictor. This can make the proposed system more evidence based.	**Issues / Limitations:** **1.** Performance is dependent on the dataset and experimental setup. A model that performs best in Ames, Iowa may not be the best model for another housing market. **2.** XGBoost has hyperparameters that need to be tuned carefully. Poor tuning can lead to underfitting, overfitting or unnecessary computational cost. **3.** High predictive performance can be less directly interpretable than a simple linear model. Feature importance and explanation techniques may therefore be useful. **4.** Strict train-validation-test separation is required to avoid data leakage. Any preprocessing that uses target information must be handled carefully.

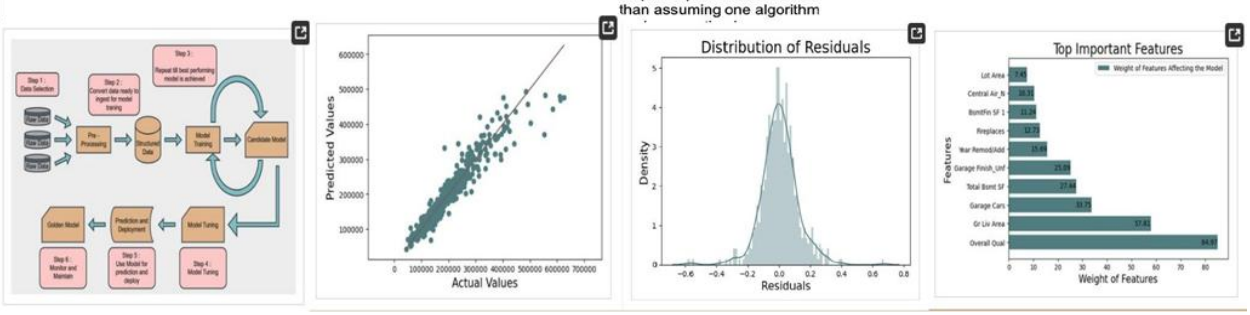


Figure 5: Literature Review of XGBoost for House Price Prediction (workflow, predicted vs. actual, residuals, top features)

The technology used in the study includes: XGBoost (sequential boosted trees), Random Forest Regressor, Support Vector Regressor, Multilayer Perceptron, and Multiple Linear Regression, trained on the Ames City, Iowa housing dataset. The methodology follows: house attributes are the independent variables and sale price the dependent target; all models are trained on the same task and scored with quantitative measures; feature importance (overall quality, living area, garage cars, basement area) then explains the price drivers. The study demonstrates efficiency with XGBoost, reported as the best-performing model, able to capture non-linear relationships and interactions among tabular features. The paper shows why several candidate models should be compared before one is chosen. However, key challenges include: results depend on the dataset and setup, XGBoost needs careful hyperparameter tuning, high accuracy can reduce interpretability, and strict train-validation-test separation is required to avoid data leakage. Relevance to Kimat: Because XGBoost won the benchmark on Ames data, it is included as a candidate model beside Random Forest. The feature-importance chart (overall quality, living area, garage, basement) also supports Kimat's goal of explaining the strongest price drivers.

Paper 6 Krish Pandey 25BAI11066	Objective	Technology Used	Methodology Used	Efficiency	Issues
Title: A Systematic Review of the use of Machine Learning in the Prediction of House Pricing **Journal:** Computational Economics, Vol. 67(5), 2026, pp. 3993–4032 **Year:** 2026 **DOI:** 10.1007/s10614-025-10983-4 **URL:** https://doi.org/10.1007/s10614-025-10983-4	**Objective:** **1.** Systematically review research trends in machine-learning-based house-price prediction. The review synthesizes evidence across published studies. **2.** Identify frequently used algorithms and feature categories. The study examines which approaches and variables appear most often in the literature. **3.** Examine common data sources and possible sources of bias. Understanding the origin of training data is important for judging how broadly results can be applied. **4.** Identify unresolved research gaps. Particular attention is given to the effect of time and market temporality on prediction.	**Technologies:** **1.** Systematic literature review is used to synthesize the research field. The review compares algorithms, features, data sources and reported performance. **2.** Random Forest, Gradient Boosting Machine and XGBoost are among the frequently used models identified by the review. **3.** Linear Regression and Artificial Neural Networks are also common approaches. This supports comparing simple, ensemble and neural models. **4.** Feature categories include locational, financial and physical information. Location, condominium fee and property area are highlighted as frequently used examples.	**Methodology:** **1.** The authors collect and synthesize published research on machine learning for real-estate price prediction. Studies are compared rather than one new predictive model being proposed. **2.** Algorithms, input features and data sources are grouped across the literature. This enables patterns in model usage and feature importance to be identified. **3.** The review examines whether data volume or algorithm choice has a consistent relationship with prediction quality. The analysis therefore considers dataset characteristics as well as algorithms. **4.** Research gaps are identified from the collected evidence. Temporal effects and the need for correction mechanisms that follow market changes are highlighted.	**Efficiency / Results:** **1.** Random Forest, GBM/XGBoost, Linear Regression and ANN are among the frequently used model families. This provides strong literature support for comparing these approaches. **2.** Location, condominium fee and property area are frequently used features in locational, financial and physical categories. These categories show that price depends on several types of information. **3.** The review finds that algorithm performance varies according to database characteristics. Therefore, there is no single algorithm that can be declared universally best. **4.** The findings support building a project with systematic preprocessing and model comparison. Performance should be measured on the project's own dataset rather than copied from another study.	**Issues / Limitations:** **1.** Real-estate websites are common data sources and may introduce significant bias. Listings can differ from actual transactions and may not represent the complete market. **2.** The review does not find a clear universal relationship between dataset size, algorithm choice and reported quality metrics. More data alone therefore does not guarantee a better model. **3.** Temporal effects on property sale value remain an important research gap. A model trained on older market conditions may lose accuracy when the market changes. **4.** Future systems may require temporal correction or updating mechanisms. Such mechanisms can help predictions follow changing market conditions instead of relying only on static historical relationships.

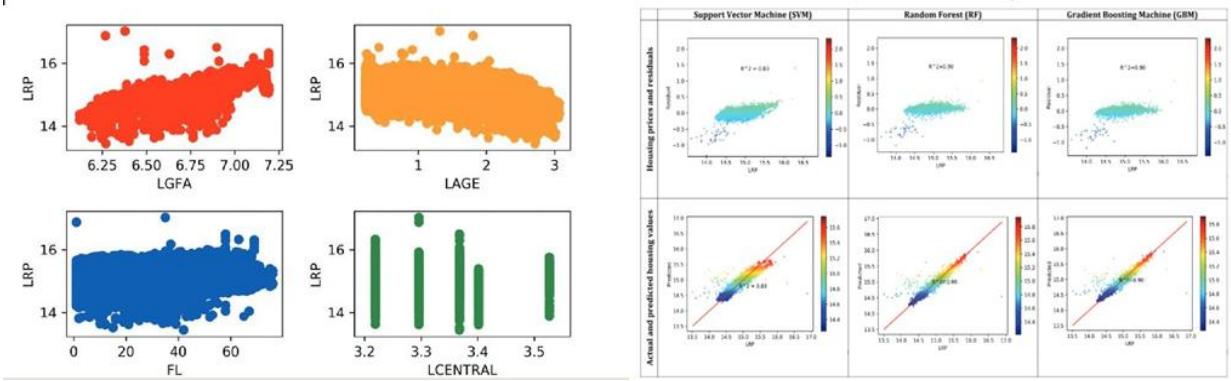


Figure 6: Literature Review of Machine Learning in House Pricing (LRP scatter plots and SVM / RF / GBM actual vs. predicted)

2.6. Paper 6: "A Systematic Review of the use of Machine Learning in the Prediction of House Pricing"

Systematically reviews research trends and unresolved gaps in machine-learning house pricing (Pita et al., Computational Economics, 2026). *Journal: Computational Economics, Vol. 67(5), 2026, pp. 3993-4032 | DOI: 10.1007/s10614-025-10983-4 | Reviewed by: Krish Pandey (25BAI11066)* The key objectives of the study include: review research trends

in machine-learning-based price prediction; identify frequently used algorithms and feature categories; examine common data sources and sources of bias; and identify gaps, especially temporal effects. The technology used in the study includes: a systematic review that compares algorithms, features, data sources and reported performance, finding Random Forest, GBM/XGBoost, Linear Regression and ANN among the most common, with locational, financial and physical feature categories. The methodology follows: published studies are collected and compared rather than proposing a new model; algorithms, input features and data sources are grouped to reveal patterns, and the link between data volume, algorithm choice and quality is examined. Summary of the review. Across the six papers, Random Forest and XGBoost are consistently strong tabular-data learners, regression remains the interpretable baseline, neural networks do not automatically win, and temporal and spatial information are recurring gaps. These findings directly motivate Kimat's dual-model comparison (MLR versus Random Forest, with XGBoost as an extension), its pincode-aware locality features and its design for periodic retraining.

III. PROPOSED SYSTEM DESIGN: HOUSE PRICE PREDICTION

Problem Statement: Develop a supervised machine learning system capable of learning patterns from historical housing data to make reliable price predictions, accurately estimating residential property prices by training models on real-world datasets with diverse property attributes, and leveraging architectural, temporal, and location-based attributes such as size, age, and neighborhood. Objectives: Input Processing: analyze and clean historical housing datasets to ensure data quality. Feature Extraction: identify the strongest predictors of property value from available attributes.

Transformation Algorithm: apply multiple regression and ensemble ML techniques for accurate modeling. Output Generation: deploy a responsive web application for real-time property valuation. Domain Analysis: The system is situated in PropTech (Property Technology) and Predictive Analytics, which merge real estate with advanced data science to modernize how properties are valued and transacted. The focus is on using quantitative attributes such as square footage, location, age, and structural features to forecast financial valuation with high accuracy. The core methodology utilizes supervised machine learning and ensemble techniques, primarily Random Forest Regression, owing to its ability to handle high-dimensional feature spaces and resist overfitting. Hardware Configuration: Minimum 8 GB RAM for efficient data processing and model training. Multi-core CPU for parallel computation on large housing datasets.

A standard GPU is recommended for accelerating model training workflows. Software Configuration: Python 3.x for model development and data processing. Scikit-Learn for regression and ensemble algorithms. Pandas for data manipulation, cleaning, and handling of large housing datasets. Flask / Streamlit for the responsive web application interface; the delivered prototype uses a Node.js (pnpm workspace, TypeScript) API server and a React client. Datasets: Models are trained using the Ames Housing Dataset (79 explanatory variables covering nearly every aspect of residential homes in Ames, Iowa, a widely used regression benchmark) and historical Kaggle Real Estate Data, which provides diverse geographic and temporal pricing benchmarks for generalization and validation. The prototype additionally ships a demo listings dataset (*listings.csv*, *listings. json*) and a locality summary (*pincode-insights. json*).

IV. ARCHITECTURE DIAGRAM PREDICTION

The AI-driven prediction pipeline features a modular architecture to allow independent updates to the model or UI without disrupting the full pipeline. User Input: The user enters home attributes (size, location, age) via a web form. Web App Interface: The frontend collects and validates input data. Input Validation: Checks for missing values, format and data integrity.

Data Preprocessing: The backend handles missing values, scaling, encoding, and feature engineering. Pre-trained ML Model: A Random Forest / Regression engine processes the feature vectors. Price Prediction Engine and Output: The engine generates a predicted property price with a confidence score, returned in real time to the user, closing the real feedback loop.

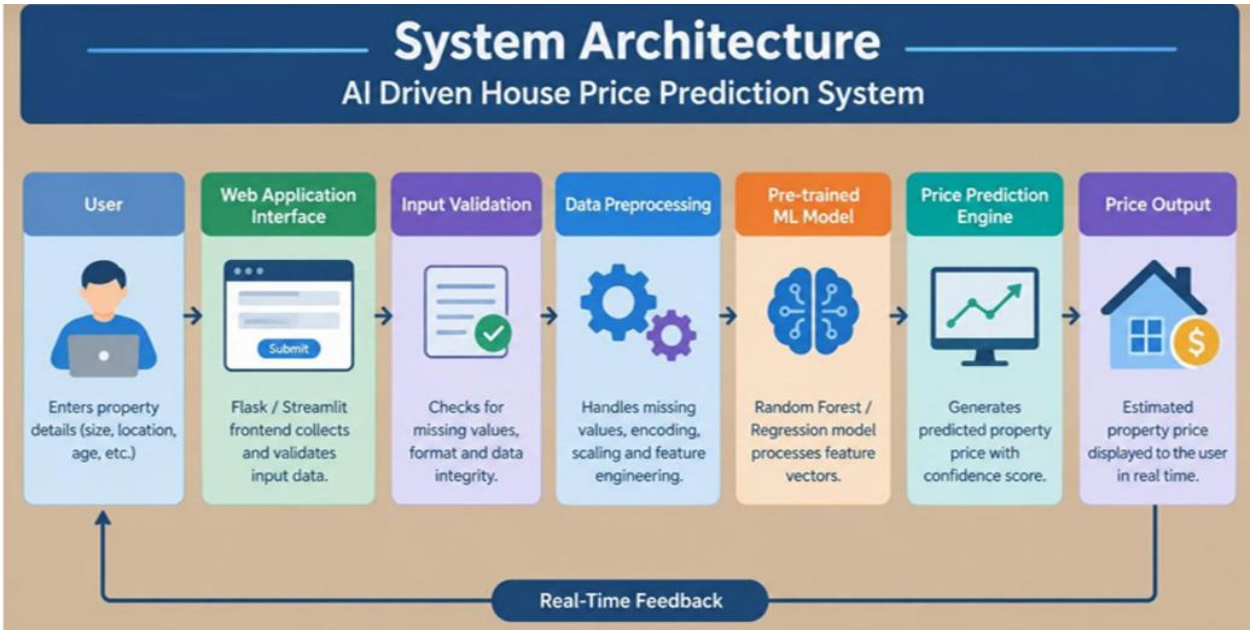


Figure 7: Architecture Diagram of Kimat: AI Driven House Price Prediction System

V. METHODOLOGY AND ALGORITHMS USED HOUSE PRICE PREDICTION

The methodology relies on a supervised learning approach utilizing ensemble techniques. Multiple Linear Regression (MLR) establishes the baseline, while Random Forest Regression is the primary model, selected for robustness against overfitting and its ability to handle high-dimensional feature spaces. The key steps are: Data cleaning and normalization: missing values, outliers (for example ultra-luxury homes) and inconsistent entries are handled.

Evaluation formulas. With n test samples, actual prices y_i, predicted prices $\hat{y}_i$ and mean price $\bar{y}$: $R^2 = 1 - \Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2$; $RMSE = \sqrt{[\ (1/n)\ \Sigma(y_i - \hat{y}_i)^2\ ]}$; $MAE = (1/n)\ \Sigma|y_i - \hat{y}_i|$. RMSE penalizes large valuation errors more heavily, MAE reports the average absolute error in price units, and R^2 reports the share of price variance explained.

Feature selection via importance ranking: the strongest predictors of price are retained. 80/20 train-test split: the cleaned data is divided to measure performance on unseen properties. Hyperparameter tuning using cross-validation: reduces overfitting and improves generalization. The dual-model comparison uses three evaluation metrics: the R^2 score (target greater than 0.85), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Comparing Random Forest against MLR highlights the gains of ensemble learning for non-linear housing data. As the literature review shows, XGBoost is retained as a further candidate model. Evaluation formulas.

Table 2: Comparison of candidate algorithm families considered for Kimat.

Algorithm family	Methodology	Efficiency	Issues
Multiple Linear Regression	Ordinary least squares regression on housing attributes.	Fast training and prediction; highly interpretable results.	Struggles with non-linearity and complex feature interactions; limited accuracy on diverse datasets.
Random Forest and XGBoost	Ensemble tree-based methods using bagging and boosting strategies.	High predictive accuracy; handles missing data and outliers well.	Computationally heavy; longer training time; less interpretable than linear models.
Neural Networks and Gradient Boosting	Deep-learning feature mapping; sequential boosting of weak learners.	Excellent feature extraction; scalable to large datasets.	Neural networks are black-box models; boosting is prone to overfitting without careful tuning.

Risk Analysis: Three principal risks are identified for the project, together with the planned mitigation.

Table 3: Risk analysis and mitigation for Kimat.

Risk	Description	Mitigation
Data quality	Datasets may contain missing values, inconsistent entries, and skewed outliers caused by ultra-luxury home prices, which can distort training and reduce accuracy.	Cleaning and normalization in the Data Ingestion Module; duplicate removal; outlier handling before training.
Overfitting	The model may memorize training patterns rather than generalizing, performing poorly on unseen real-world property data.	80/20 hold-out split, cross-validated hyperparameter tuning, and ensemble averaging.
Deployment latency	The web application must deliver instant predictions in real time for good user experience and usability.	Pre-trained model loaded once; lightweight JSON Prediction API; modular, independently testable layers.

Project Schedule: Phase 1 (Review I) covers problem definition, literature review and dataset selection. Phase 2 (Review II) covers preprocessing, feature engineering, model training, optimization and comparison of regression techniques. Phase 3 (Final Review) covers frontend and backend integration, system testing, performance evaluation and a live demonstration.

VI. PROJECT FUNCTIONAL MODULES IMPLEMENTATION KIMAT

The system pipeline is composed of three interconnected modules, each responsible for a distinct stage of the data-to-prediction pipeline and independently testable: Data Ingestion Module: loads raw CSV datasets, standardizes input features, removes duplicates, and handles missing values for downstream operations. Training Module: splits the cleaned data into an 80/20 train-test ratio, fits the regression model, tunes hyperparameters, and evaluates performance metrics. Prediction API: receives structured JSON input from the web frontend, applies the pre-trained ML model, and returns real-time price estimates to the application. System flow. The modules operate as a structured pipeline from raw data to user-facing predictions: (1) Data Ingestion - CSV datasets are loaded, cleaned and validated for missing values; (2) Model Training - preprocessed data is split 80/20 and fed into the regression model; (3) Prediction API - the trained model is exposed through a REST API that receives JSON input from the frontend; (4) Web App Display - real-time price estimates are returned and rendered on the user interface. Each module is independently testable, which keeps the architecture robust and maintainable.



Figure 8: Functional Modules of Kimat: Data Ingestion, Training and Prediction API

VII. METHODOLOGY FOR DEVELOPING THE KIMAT WEBSITE SCREENSHOTS

The platform addresses a critical research gap: existing AI-based valuation tools are overly technical. Kimat provides an intuitive interface for non-technical users (homebuyers, agents) to dynamically input real-time custom home metrics. The frontend validates these inputs and communicates with the Prediction API to fetch confidence-backed estimates. The development followed a needs assessment, requirement definition, interface design, development, testing and deployment cycle. Unit testing validates the prediction algorithm with known inputs, verifying that regression outputs match expected estimates within acceptable error margins, while UI/UX usability testing evaluates the interface with real users for intuitive navigation, accurate form handling and clear display of valuations.

Table 4: Software testing plan for Kimat.

Test type	What is validated	Pass criterion
Unit testing	Individual model components and the prediction algorithm are run with known inputs.	Regression outputs match expected price estimates within acceptable error margins.
UI/UX usability testing	The web interface is used by real users: navigation, form input handling and display of valuations.	Intuitive navigation, accurate input handling, and clear display of predicted valuations.

Integration / system testing	Frontend, Prediction API and model are tested together on the full pipeline.	Correct end-to-end estimates returned in real time.

VIII. KIMAT PROTOTYPE, ALGORITHM, AND PROGRAM LOGIC

The prototype demonstrates a modern, intuitive user interface that directly translates predictive modeling into actionable intelligence. The landing page lets users search by pincode and plain English; the system turns local signals into a ranked shortlist and a number that can be explained. Figures 9 to 15 present the prototype screens and the source workspace.

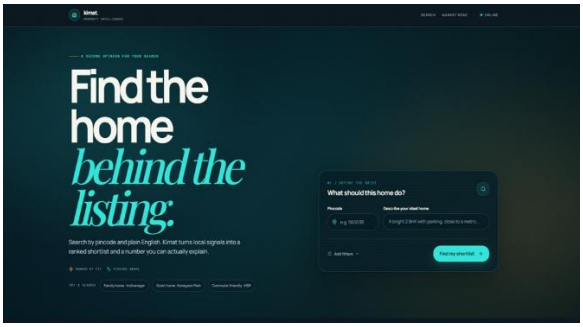


Figure 9: Kimat Landing Interface: search by pincode and a plain-English brief (e.g., "A bright 2 BHK with parking, close to a metro")

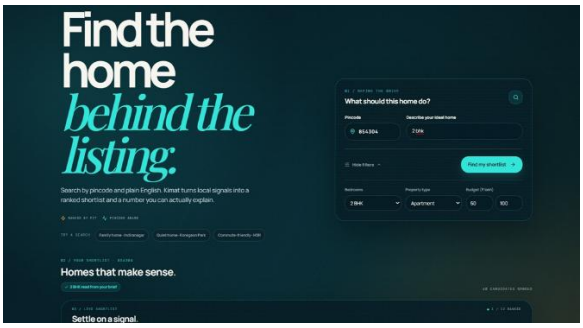


Figure 10: Define the Brief: Pincode 854304, 2 BHK apartment, budget of Rs. 50 to 100 lakhs, with 40 candidates ranked

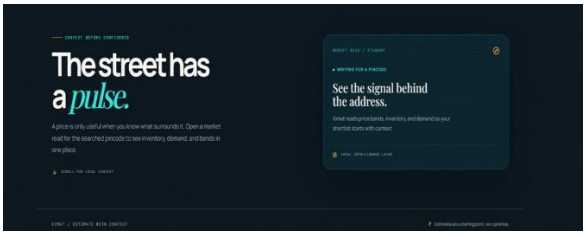


Figure 11: Market Read in standby, awaiting a pincode to provide local market context

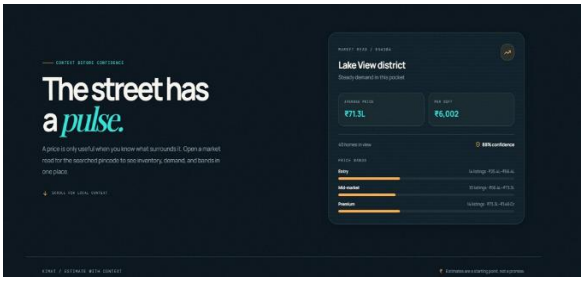


Figure 12: Market Read for Lake View district (854304): average price Rs. 71.3L, Rs. 6,002 per sq ft, 88% confidence, and Entry / Mid-market / Premium price bands

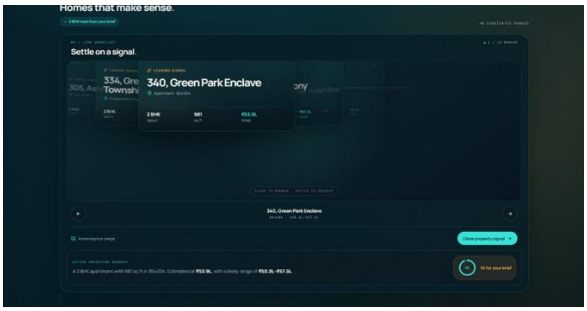


Figure 13: Live Shortlist: 340, Green Park Enclave (2 BHK, 981 sq ft) valued at Rs. 53.9L with a likely range of Rs. 50.3L to Rs. 57.5L and a 93-fit score

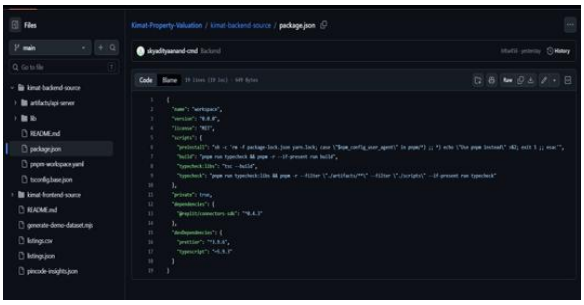


Figure 14: Backend source (*kimat-backend-source/package.json*): pnpm workspace for the Kimat API server

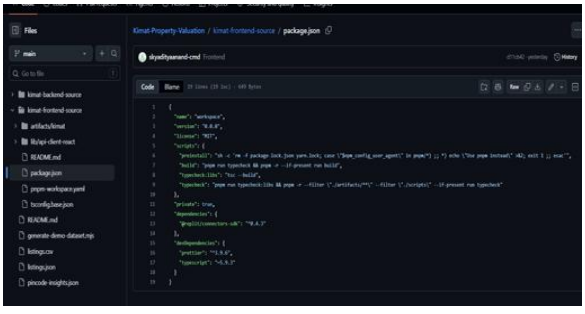


Figure 15: Frontend source (kimat-frontend-source/package.json): workspace for the Kimat React client and API client library

Algorithm and program logic. (1) The user supplies a pincode and a brief; bedrooms, property type and budget are read from the brief or the filter panel. (2) The API loads the listings for that pincode and computes a market read: average price, price per sq ft, inventory count, a confidence value and three price bands. (3) Each candidate listing receives a model valuation with a likely range and a fit score against the brief. (4) Candidates are ranked and returned to the shortlist, where the user can browse and open a property signal. The interface states plainly that estimates are a starting point, not a promise.

IX. CONTRIBUTION AND FINDINGS

Findings: Location and square footage emerged as the highest predictive weight features, with properties in premium neighborhoods and larger floor areas consistently yielding significantly higher estimates across all regression techniques tested. The age of a property was less influential as a standalone predictor; combined with renovation history and condition ratings, temporal attributes contributed moderately but were outweighed by spatial and structural features. Expected Efficiency: The Random Forest ensemble fulfils the project objectives by handling both linear and non-linear housing data and is expected to outperform the MLR baseline. The model targets an R^2 score greater than 0.85, meaning that over 85% of the variance in house prices is explained, and aims to minimize RMSE so that predictions stay close to actual market prices and costly valuation errors are reduced. Justification: Ensemble methods reduce variance and bias relative to single-model baselines. Random Forest processes architectural, temporal and location-based attributes simultaneously, generalizes well to unseen data, resists overfitting, and delivers robust feature-importance rankings, making it a scalable and accurate choice for the valuation engine. End Users: Homebuyers who want data-driven valuations to avoid overpaying; real estate agents and brokers who need fast, reliable estimates to advise clients and set listing prices; and property investors and developers who use predictive analytics to identify undervalued assets and optimize portfolios. Society Contribution: Kimat democratizes real estate valuation, making it accessible to everyday buyers and sellers rather than only industry professionals. Traditional appraisals are costly and time-consuming, often taking days; by providing instant web-based valuations, the system reduces friction and financial burden and contributes to a more transparent, efficient, and equitable real estate market.

X. CONCLUSION

The AI-driven House Price Prediction System successfully bridges the gap between machine learning and PropTech. By combining ensemble methods (Random Forest, XGBoost) and historical housing data, the Kimat system delivers scalable, accurate, and accessible real-time property valuations. The literature review of six studies supports a model-comparison approach, and the working prototype shows how a pincode-aware market read, a ranked shortlist, and an explainable valuation range can be delivered through a simple interface. Future work includes training on larger multi-city datasets, adding temporal updating so the model follows market changes, and incorporating richer spatial features. This project demonstrates the transformative potential of AI in modernizing real estate appraisal for homebuyers, agents, and investors alike.

REFERENCES

[1] Adetunji, A. B., et al., “House price prediction using Random Forest machine learning technique,” *Procedia Computer Science*, vol. 199, pp. 806–813, 2022, doi: 10.1016/j.procs.2022.01.100.

[2] G. D. K., Singh, M. K., and M. Jayanthi, Eds., *Network Security Attacks and Countermeasures*. IGI Global, 2016, doi: 10.4018/978-1-4666-8761-5.

[3] I. Foryś, “Machine learning in house price analysis: Regression models versus neural networks,” *Procedia Computer Science*, vol. 207, pp. 435–445, 2022, doi: 10.1016/j.procs.2022.09.078.

[4] D. Shi, J. Guan, J. Zurada, and A. S. Levitan, “Predicting home sale prices: A review of existing methods and illustration of data stream methods for improved performance,” *WIREs Data Mining and Knowledge Discovery*, vol. 12, Art. no. e1435, 2022, doi: 10.1002/widm.1435.

[5] J. Kannan, M. K., N. Aaijaz, K. G. Mani, and V. Tewari, *The Future of Innovation and Technology in Education: Trends and Opportunities*. S&M Publications, 2025.

[6] B. R. M. and J. K. M. K., "Intrusion detection on AWS cloud through hybrid deep learning algorithm," *Electronics*, vol. 12, Art. no. 1423, 2023, doi: 10.3390/electronics12061423.

[7] M. Geerts, S. vanden Broucke, and J. De Weerdt, "A survey of methods and input data types for house price prediction," *ISPRS International Journal of Geo-Information*, vol. 12, no. 5, Art. no. 200, 2023, doi: 10.3390/ijgi12050200.

[8] H. Sharma, H. Harsora, and B. Ogunleye, "An optimal house price prediction algorithm: XGBoost," *Analytics*, vol. 3, no. 1, pp. 30–45, 2024, doi: 10.3390/analytics3010003.

[9] S. K. Shukla, U. Dwivedi, M. K. Jayanthi Kannan, and C. Sarvani, *Python for Data Analytics: Practical Techniques and Applications*. JSR Publications, 2024.

[10] J. Kannan, M. K., S. Kumar, P. T. Kalaivaani, and G. Tripathi, *Artificial Intelligence and Blockchain Technology for Human Resource Management*. Scientific International Publishing House, 2025.

[11] M. K. Jayanthi, "Strategic planning for information security—DID mechanism to befriend the cyber criminals to assure cyber freedom," in *Proc. 2nd Int. Conf. Anti-Cyber Crimes (ICACC)*, Abha, Saudi Arabia, 2017, pp. 142–147, doi: 10.1109/Anti-Cybercrime.2017.7905280.

[12] R. P. Pita, et al., "A systematic review of the use of machine learning in the prediction of house pricing," *Computational Economics*, vol. 67, no. 5, pp. 3993–4032, 2026, doi: 10.1007/s10614-025-10983-4.