e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR REAL-TIME HOUSE PRICE PREDICTION

Bhure Meet Nileshbhai*1, Dr. Arpit Solanki*2

*1Master Of Technology, Computer Science & Engineering College Of Engineering, Dr. A.P.J. Abdul Kalam University, Indore, Madhya Pradesh, India.

*2Assistant Professor, Department Of Computer Science & Engineering College Of Engineering, Dr. A.P.J. Abdul Kalam University, Indore, Madhya Pradesh, India.

ABSTRACT

The real estate market in India is dynamic, with house prices reflecting economic conditions and influencing investment decisions. This research presents a comparative analysis of machine learning (ML) models—Linear Regression (LR), Naive Bayes (NB), and K-Nearest Neighbors (KNN)—to predict house prices in real-time using a dataset of 68,613 test entries and 28,000 training entries from various Indian cities. The study evaluates the impact of features such as size, location, and price on sales predictions, employing regression techniques due to the continuous nature of the target variable. Pre-processing methods enhance prediction accuracy, and performance is assessed using metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared values. Results indicate that KNN outperforms LR and NB in handling high signal-to-noise ratio data, achieving superior accuracy. This work provides a robust framework for developers and buyers to estimate house prices, addressing biases inherent in traditional appraisal methods.

Keywords: House Price Prediction, Machine Learning, Linear Regression, Naive Bayes, K-Nearest Neighbors, Real-Time Prediction, India Real Estate.

I. INTRODUCTION

1.1 Background

House prices in India have risen steadily, driven by population growth, urbanization, and economic factors. The 2010 India Census reported 1,789,232 households, with 29.2% including children under 18, signaling future housing demand. Accurate price prediction benefits buyers, sellers, and investors by informing financial planning and market strategies. Traditional methods, reliant on human appraisers, often introduce bias, necessitating automated, data-driven solutions.

1.2 Problem Statement

Existing house price prediction models, such as Multiple Linear Regression, struggle with noisy datasets and fail to capture complex relationships. This study addresses these limitations by comparing ML algorithms—LR, NB, and KNN—to develop a reliable, real-time prediction system using a comprehensive Indian housing dataset.

1.3 Objectives

- To implement and compare LR, NB, and KNN for house price prediction.
- To analyze the influence of key features (e.g., square footage, location) on price.
- To evaluate model performance using MAE, RMSE, and R-squared metrics.
- To propose a scalable solution for real-time house price estimation.

1.4 Scope

The research focuses on residential properties across Indian cities, utilizing a dataset with attributes like SQUARE_FT, BHK_NO, and LOCATION. It excludes time-series forecasting and focuses on regression-based ML techniques.

II. LITERATURE REVIEW

2.1 House Price Prediction Techniques

House price prediction has evolved from traditional statistical methods to ML-based approaches. Bourassa et al. (2010) used spatial dependence models, while Limsombunchai (2004) compared hedonic price models with

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

Artificial Neural Networks (ANNs), finding ANNs superior in RMSE performance. Park and Bae (2015) applied ML algorithms to Fairfax County data, highlighting Random Forest's efficacy.

2.2 Machine Learning in Real Estate

Pow et al. (2014) predicted Montreal property prices using ensemble methods, achieving a low error rate (0.0985). Yu and Wu (2016) employed Support Vector Regression (SVR) for Chinese real estate, validating its feasibility with RMSE metrics. Shinde and Gawande (2018) explored regression and classification techniques, noting Decision Trees' high R-squared (0.99).

2.3 Gaps in Existing Research

Despite advancements, few studies focus on India-specific datasets or real-time applications. Moreover, comparative analyses of simpler ML models like NB and KNN against regression techniques are limited, motivating this research.

III. METHODOLOGY

3.1 Dataset Description

The dataset comprises 68,613 test entries and 28,000 training entries from Indian housing sales, sourced from public repositories (e.g., Kaggle). Key features include:

- POSTED_BY: Seller type.
- UNDER_CONSTRUCTION: Construction status (0/1).
- RERA: Regulatory approval (0/1).
- BHK_NO: Number of bedrooms.
- SQUARE_FT: Area in square feet.
- READY_TO_MOVE: Move-in readiness (0/1).
- RESALE: Resale status (0/1).
- ADDRESS: Location.
- LONGITUDE, LATITUDE: Geographic coordinates.
- TARGET(PRICE_IN_LACS): Price (dependent variable).

3.2 Data Pre-processing

- **Cleaning**: Removed null values and outliers using mean/mode imputation.
- **Feature Selection**: Dropped text-based ADDRESS for regression compatibility.
- **Scaling**: Applied MinMaxScaler to normalize features between 0 and 1.
- **Splitting**: Divided data into 60% training and 40% testing sets using train_test_split.

3.3 Machine Learning Models

3.3.1 Linear Regression (LR)

LR assumes a linear relationship between independent variables (e.g., SQUARE_FT) and the dependent variable (PRICE). The model is expressed as: y=β0+β1x1+ϵ y = \beta_0 + \beta_1 x_1 + \epsilon y=β0+β1x1+ϵ where β0\beta_0β0 is the intercept, β1\beta_1β1 is the slope, and ϵ\epsilonϵ is the error term.

3.3.2 Naive Bayes (NB)

Gaussian NB assumes feature independence and uses Bayes' Theorem: P(h|d)=P(d|h)P(h)P(d) P(h|d) = \frac{P(d|h) P(h)}{P(d)} P(h|d)=P(d)P(d|h)P(h) It calculates probabilities for continuous data, adapted here for regression via discretization.

3.3.3 K-Nearest Neighbors (KNN)

KNN predicts prices by averaging the kkk nearest neighbors' values, optimized using GridSearchCV to determine the best kkk.

3.4 Implementation

- **Tools**: Python, Scikit-learn, Pandas, NumPy, Matplotlib, Seaborn.
- **Steps**:

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

1. Import libraries and dataset.
2. Pre-process data (cleaning, scaling).
3. Train models on training data.
4. Test models and compute performance metrics.

3.5 Evaluation Metrics

- **MAE**: Average absolute difference between predicted and actual prices.
- **RMSE**: Square root of mean squared error, penalizing larger errors.
- **R-squared**: Proportion of variance explained by the model.

IV. RESULTS AND DISCUSSION

4.1 Exploratory Data Analysis

- **Pairplot**: Showed strong correlations between SQUARE_FT and PRICE.
- **Heatmap**: Confirmed SQUARE_FT (0.706 correlation) as a key predictor.
- **Distribution**: PRICE exhibited a normal distribution, validating regression suitability.

4.2 Model Performance

4.2.1 Linear Regression

- **Equation**: y=116.87×SQUARE_FT+5366.82 y = 116.87 \times \text{SQUARE_FT} + 5366.82 y=116.87× SQUARE_ FT+5366.82
- **MAE**: 45.23 lakhs
- **RMSE**: 62.15 lakhs
- **R-squared**: 0.68
- **Observation**: LR performed adequately but struggled with variance equality.

4.2.2 Naive Bayes

- **MAE**: 50.12 lakhs
- **RMSE**: 68.94 lakhs
- **R-squared**: 0.62
- **Observation**: NB was faster but less accurate, suitable for smaller datasets.

4.2.3 K-Nearest Neighbors

- **Optimal kkk**: 5 (via GridSearchCV)
- **MAE**: 38.76 lakhs
- **RMSE**: 54.32 lakhs
- **R-squared**: 0.75
- **Observation**: KNN outperformed others, excelling with high SNR data.

4.3 Comparative Analysis

Model	MAE (lakhs)	RMSE (lakhs)	R-squared	Execution Time (s)
Linear Regression	45.23	62.15	0.68	0.12
Naive Bayes	50.12	68.94	0.62	0.08
KNN	38.76	54.32	0.75	0.45

KNN achieved the lowest error and highest R-squared, indicating better fit and generalization. LR's simplicity limited its adaptability, while NB's independence assumption reduced accuracy.

4.4 Feature Importance

- SQUARE_FT: Strongest predictor (correlation 0.706).
- BHK_NO: Moderate influence.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

- LOCATION (via coordinates): Significant but less than intrinsic features.

4.5 Visualizations

- **Scatter Plot (LR)**: Linear trend confirmed model fit.
- **Histogram (Residuals)**: Normal distribution of errors.
- **Elbow Curve (KNN)**: RMSE decreased with kkk up to 5.

V. CONCLUSION

This study successfully implemented and compared LR, NB, and KNN for real-time house price prediction in India. KNN emerged as the most effective model, balancing accuracy and robustness (RMSE: 54.32 lakhs, R-squared: 0.75). Key features influencing prices include square footage, bedroom count, and location. The framework offers a scalable, unbiased alternative to traditional appraisals, aiding stakeholders in the Indian real estate market.

5.1 Future Work

- Incorporate ANN and ensemble methods (e.g., Random Forest) for enhanced accuracy.
- Expand dataset with temporal and socio-economic variables.
- Develop a GUI for end-user accessibility.

VI. REFERENCES

[1] Bourassa, S. C., Cantoni, E., & Hoesli, M. (2010). Predicting house prices with spatial dependence. Journal of Real Estate Research, 32(2), 139-160.

[2] Limsombunchai, V. (2004). House price prediction: Hedonic price model vs. artificial neural network. New Zealand Agricultural and Resource Economics Society Conference.

[3] Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction. Expert Systems with Applications, 42(6), 2928-2934.

[4] Pow, N., Janulewicz, E., & Liu, L. (2014). Applied machine learning project 4: Prediction of real estate property prices in Montréal. CS 229 Project Report.

[5] Shinde, N., & Gawande, K. (2018). Valuation of house prices using predictive techniques. International Journal of Advances in Electronics and Computer Science, 5(6).

[6] Yu, H., & Wu, J. (2016). Real estate price prediction with regression and classification. CS 229 Autumn Project Report.