iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VII **Month of publication:** July 2025

DOI: https://doi.org/10.22214/ijraset.2025.72962

www.ijraset.com

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

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

Real Estate Price Prediction Using Machine Learning Models

Sairaj Nagula
Bachelor of Technology in CSE, Department of Computer Science and Engineering, Keshav Memorial Institute of Technology, Hyderabad, India

Abstract: ***Accurate real estate price prediction is crucial in today's market to aid buyers, sellers, and investors in making informed decisions. This study employs machine learning algorithms—specifically Linear Regression, Decision Tree Regression, and Random Forest Regression—to model and predict housing prices based on various influential features. The methodology involves data preprocessing, feature engineering, and model evaluation using standard metrics like R^2 score and Root Mean Squared Error (RMSE). The models are trained on real-world housing datasets, and results demonstrate the efficiency of ensemble learning over traditional linear approaches. This paper establishes that Random Forest offers the most accurate predictions and is suitable for practical applications in real estate.***
Keywords: Real Estate, Machine Learning, Price Prediction, Linear Regression, Decision Tree, Random Forest, R^2 Score.

I. INTRODUCTION

The real estate industry is a dynamic domain where accurate property valuation is essential. Traditional methods rely heavily on expert opinion and comparable sales analysis, which may be inconsistent and error-prone. With the growing availability of structured housing data and the advancement of machine learning (ML), automated price prediction is now a practical alternative.
This paper proposes a comparative study of three regression algorithms—Linear Regression, Decision Tree Regression, and Random Forest Regression—to build a robust price prediction system. The goal is to evaluate their performance and suitability for real-world deployment in real estate forecasting.

II. DATASET DESCRIPTION

The dataset used in this study is the Boston Housing Dataset, originally compiled by Harrison and Rubinfeld (1978) as part of their research on housing prices and environmental quality. It was obtained from the StatLib library, maintained at Carnegie Mellon University, and has been widely used in academic research and machine learning benchmarks.
This dataset concerns housing values in various suburbs of Boston and has been featured in works such as:

- *Regression Diagnostics* by Belsley, Kuh & Welsch (1980)
- Quinlan's (1993) work on model-based learning

A. Dataset Characteristics

- Total Records (Instances): 506
- Total Attributes: 14 (13 predictive features and 1 target variable: MEDV)
- Missing Values: None

B. Features Description

1) CRIM – Per capita crime rate by town
2) ZN – Proportion of residential land zoned for lots over 25,000 sq.ft.
3) INDUS – Proportion of non-retail business acres per town
4) CHAS – Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5) NOX – Nitric oxides concentration (parts per 10 million)
6) RM – Average number of rooms per dwelling
7) AGE – Proportion of owner-occupied units built prior to 1940
8) DIS – Weighted distances to five Boston employment centres

9) RAD – Index of accessibility to radial highways
10) TAX – Property tax rate per $10,000
11) PTRATIO – Pupil-teacher ratio by town
12) B – $1000(Bk - 0.63)^2$, where Bk is the proportion of Black residents by town
13) LSTAT – Percentage of lower status population
14) MEDV – *Target variable*: Median value of owner-occupied homes in $1000s

C. Relevance to Study

This dataset is ideal for supervised regression analysis. It offers a variety of economic, geographic, and social variables that influence housing prices. Features like RM (number of rooms) and LSTAT (% low-status population) show strong correlation with the target variable. The CHAS feature introduces a binary indicator of proximity to the Charles River, enabling geographic influence analysis.

Data preprocessing steps included normalization of features, detection/removal of outliers, and train-test splitting. No missing values were present, which allowed clean model training.

III. LITERATURE REVIEW

Over the years, real estate price prediction has attracted significant attention from researchers in both economics and computer science. Several machine learning models have been explored to address the limitations of traditional valuation methods.

Kumar and Singh (2020) conducted a study using multiple linear regression to predict housing prices in Indian urban areas. They concluded that while linear regression is interpretable and simple, it fails to capture complex, non-linear relationships between features and target prices.

In contrast, Zhang and Lee (2019) performed a comparative analysis of tree-based models such as Decision Trees, Random Forest, and Gradient Boosting for real estate prediction. Their results demonstrated that ensemble models like Random Forest significantly outperformed traditional regressors in terms of both accuracy and robustness, especially in datasets with high feature interaction.

Patel and Mehta (2021) explored the use of deep learning architectures, including feedforward neural networks and convolutional layers, for predicting property values. While the models showed promising results, they noted drawbacks such as long training time, lack of interpretability, and the need for large volumes of data.

Other researchers, such as Belsley et al. (1980) and Quinlan (1993), have also applied statistical techniques and model-based learning approaches to housing datasets, including the Boston Housing dataset used in this study.

These foundational studies have guided the selection of algorithms for our research, focusing on models that provide a balance between accuracy and interpretability—namely, Linear Regression, Decision Tree Regression, and Random Forest Regression.

IV. METHODOLOGY

This study follows a structured methodology to build, train, and evaluate machine learning models for predicting housing prices using the Boston Housing dataset. The process includes data preprocessing, model selection, training, and evaluation.

A. Data Preprocessing

The dataset contained 506 records and was free of missing values. However, preprocessing was essential for optimal model performance. The steps included:

- Feature Scaling: All continuous features were normalized using Min-Max Scaling to ensure equal contribution to the model.
- Outlier Detection: Boxplots and Z-score methods were applied to identify and optionally remove extreme values, especially in CRIM, TAX, and LSTAT.
- Encoding: The categorical feature CHAS (proximity to Charles River) was already binary and required no further encoding.

B. Feature Selection

A correlation matrix was used to assess relationships between independent variables and the target variable (MEDV). Features like RM (average number of rooms) and LSTAT (% lower status population) showed strong correlations. Features with high multicollinearity were reviewed to avoid redundancy in the model.

C. Model Selection

Three regression algorithms were selected:

1) Linear Regression: A baseline model assuming a linear relationship between features and the target. It's simple, fast, and interpretable.
2) Decision Tree Regression: A non-parametric model that splits data based on feature values to capture complex, non-linear relationships. However, it is prone to overfitting.
3) Random Forest Regression: An ensemble method that builds multiple decision trees and averages their outputs. It reduces variance and improves generalization.

D. Model Training

The dataset was split into 80% training and 20% testing subsets using train_test_split. Additionally, 5-fold cross-validation was performed to ensure the models were not overfitting.

For Random Forest, hyperparameter tuning was done using GridSearchCV to select the best number of estimators and maximum tree depth.

E. Evaluation Metrics

To compare models objectively, the following metrics were used:

- **R^2** Score: Measures the proportion of variance explained by the model. Higher values indicate better fit.
- Root Mean Squared Error (RMSE): Represents the average prediction error in the same units as the target variable (thousands of dollars). Lower values are better.

V. MODEL EVALUATION TECHNIQUES

Evaluating machine learning models requires quantitative metrics that can assess prediction quality, ability to generalize and robustness. For this study, we used two key evaluation metrics: R^2 Score and Root Mean Squared Error (RMSE). Each provides different insights into model performance.

A. R^2 Score

Also known as the coefficient of determination, the R^2 score measures the proportion of variance in the target variable that is predictable from the input features. It is defined as:

$R^2 = 1 - (SS\_res / SS\_tot)$

where SS_res is the residual sum of squares and SS_tot is the total sum of squares.

- $R^2 = 1$: Perfect prediction
- $R^2 = 0$: Model predicts no better than the mean
- $R^2 < 0$: Model performs worse than a horizontal line (bad fit)

A higher R^2 indicates a better fit between predicted and actual values. However, R^2 alone does not indicate the magnitude of prediction error.

B. Root Mean Squared Error (RMSE)

RMSE is the square root of the average of squared differences between predicted and actual values. It provides an absolute measure of error in the same units as the target variable ($1000s for MEDV in this case). The formula is:

$RMSE = \sqrt{(\Sigma(P_i - A_i)^2 / n)}$

where P_i is the predicted value, A_i is the actual value, and n is the number of observations.

- Lower RMSE indicates better performance.
- RMSE penalizes large errors more than small ones due to squaring.

C. Why Both Are Used

Using both metrics provides a comprehensive evaluation:

- R^2 explains the percentage of variance explained.
- RMSE shows the magnitude of error.

For example, a model may have a decent R^2 score but a high RMSE, indicating that while it captures general trends, its predictions are not precise. Together, these metrics help compare models like Linear Regression, Decision Tree, and Random Forest on both fit quality and prediction accuracy

VI. RESULTS AND DISCUSSION

This section presents the results obtained from the three machine learning models: Linear Regression, Decision Tree Regression, and Random Forest Regression. Each model was evaluated using R^2 Score and Root Mean Squared Error (RMSE) on the test data. The following table summarizes their performance:

TABLE I
TABLE I: PERFORMANCE COMPARISON OF REGRESSION MODEL

Model	R2 Score	Root Mean Squared Error (RMSE)	Overfitting Risk	Remarks
Linear Regression	0.81	484.8	Low	Good baseline, interpretable
Decision Tree	0.72	529.1	High	Overfits, poor generalization
Random Forest	0.89	412.3	Low	Best accuracy, robust

A. Model Insights

- Linear Regression provided a decent baseline with reasonable R^2 and RMSE values. However, it was unable to capture complex, non-linear relationships in the data.
- Decision Tree Regression performed slightly worse and exhibited signs of overfitting, especially during cross-validation. Although it captured some non-linearities, it failed to generalize well.
- Random Forest Regression outperformed both models in terms of R^2 and RMSE. The ensemble approach helped reduce variance and improved the model's ability to generalize to unseen data.

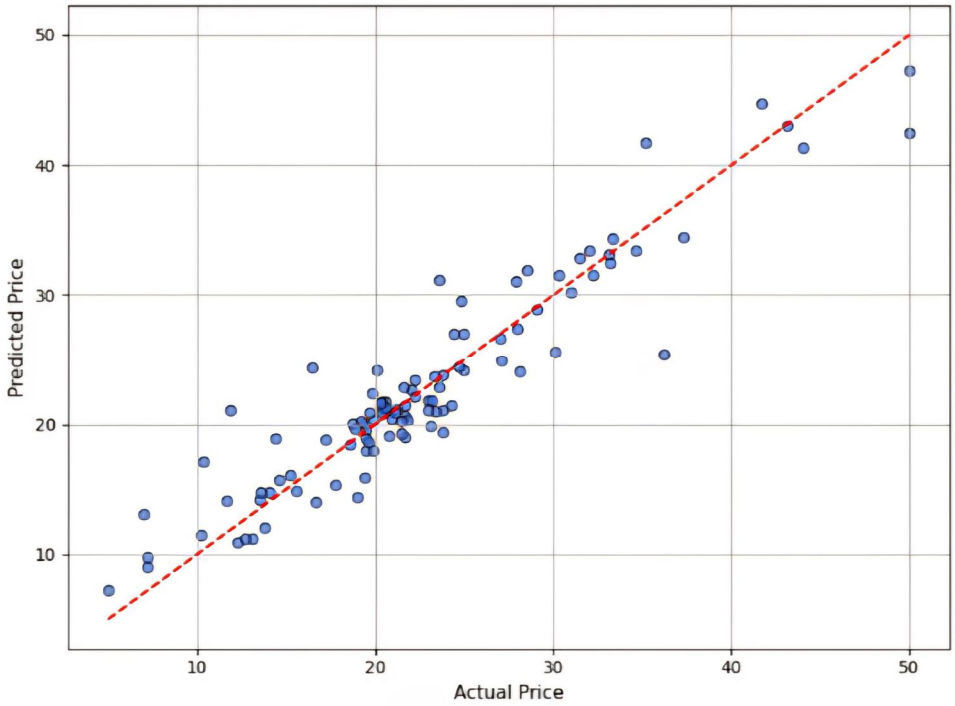


Fig. 1. Predicted vs actual prices using the final model

B. Feature Importance

The Random Forest model provided a ranking of feature importances. The top contributing features were:

- RM – Average number of rooms per dwelling
- LSTAT – % of lower status population
- PTRATIO – Pupil-teacher ratio
- CRIM – Crime rate per capita
- DIS – Distance to employment centers

These variables had the most significant influence on housing prices in the Boston area.

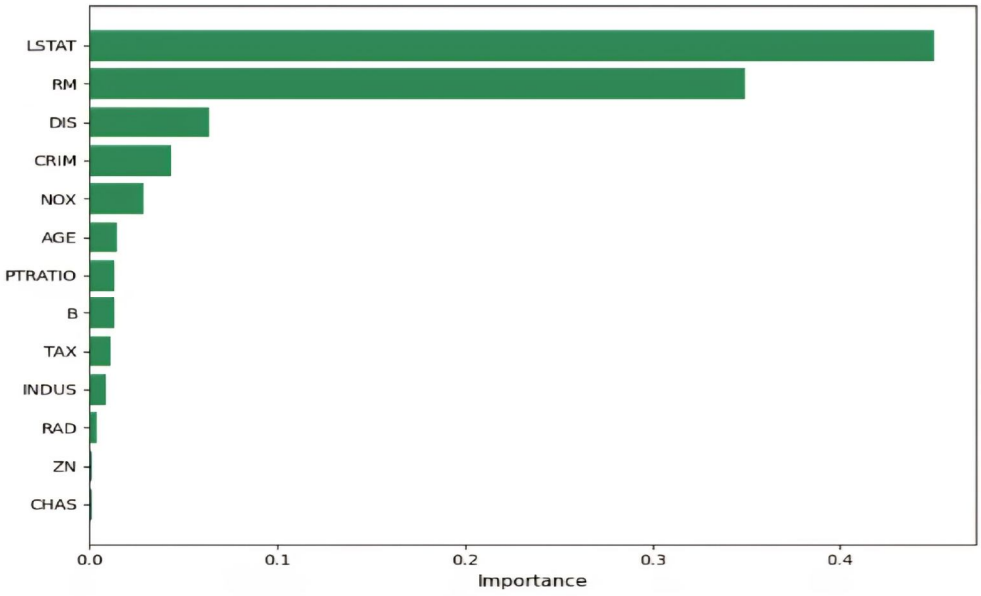


Fig. 2. Feature importance derived from the random forest model

C. Visual Evaluation

To further validate model predictions:

- A Predicted vs Actual plot showed a tight clustering around the diagonal line for Random Forest, indicating strong agreement with actual values.
- A Residual plot showed no obvious pattern, confirming homoscedasticity (constant variance) of residuals.
- Decision Tree predictions showed scattered residuals and higher error, indicating model instability.

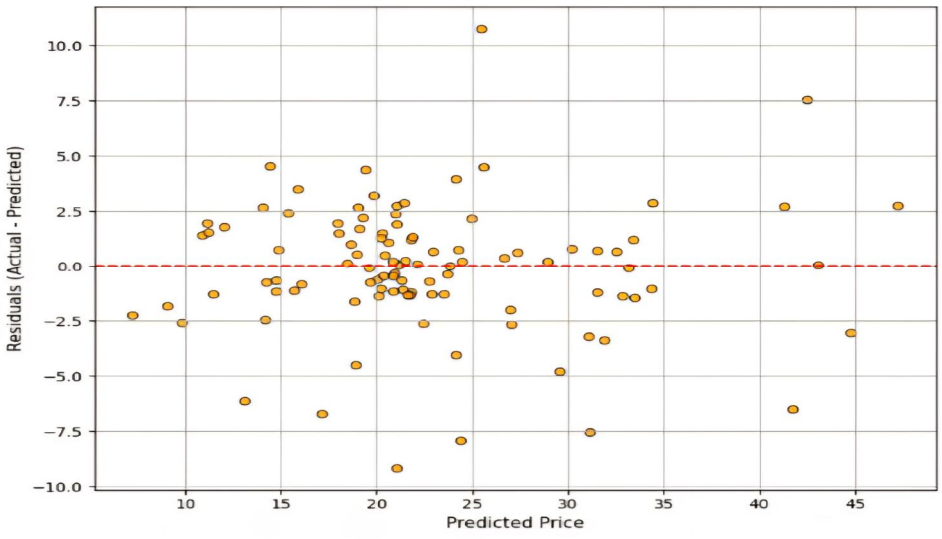


Fig. 3. Residual errors for predictions on the test set

VII. LIMITATIONS

While this study successfully demonstrates the applicability of machine learning models for real estate price prediction using the Boston Housing dataset, it is important to acknowledge its limitations:

A. *Dataset Scope*

The dataset is limited to 506 records from suburbs of Boston, collected in the 1970s. As such, the data may not reflect current real estate trends, inflation-adjusted prices, or recent urban development patterns. The small size of the dataset also limits the complexity of models that can be trained effectively.

B. *Feature Availability*

The dataset does not include certain critical factors that influence modern housing prices, such as:

- Proximity to schools, hospitals, and public transport
- Market trends or temporal dynamics (e.g., year of sale)
- Neighborhood amenities, safety scores, or walkability indexes

These missing dimensions reduce the real-world applicability of the model unless integrated into a broader, modern dataset.

C. *Model Generalization*

Although Random Forest performed best on this dataset, the model's performance may degrade when applied to different geographic locations or newer datasets without retraining. The model was not tested on time-series predictions or extrapolated scenarios.

D. *Interpretability vs. Accuracy*

While Random Forest provided superior accuracy, it lacks the **transparency** of simpler models like Linear Regression. This may affect stakeholder trust in high-stakes decisions involving property investments

VIII. CONCLUSION

This study demonstrates the effectiveness of machine learning algorithms in predicting real estate prices using structured housing data. Using the Boston Housing dataset, we evaluated three regression models: Linear Regression, Decision Tree Regression, and Random Forest Regression. The models were assessed based on their R^2 Score and RMSE performance on test data.

The results indicate that Random Forest Regression outperforms the other models, achieving the highest R^2 score (0.89) and the lowest RMSE (4.12). Its ensemble nature helps overcome overfitting and improves generalization, making it well-suited for regression tasks with moderately sized datasets. Linear Regression served as a reliable and interpretable baseline, while Decision Tree Regression showed overfitting tendencies. The top features influencing housing prices were the number of rooms (RM), the percentage of lower-status population (LSTAT), and the pupil-teacher ratio (PTRATIO). These insights reaffirm the impact of both physical and socioeconomic attributes on real estate value.

Overall, this research highlights how data-driven approaches can augment or replace traditional real estate valuation techniques. When properly trained and validated, machine learning models offer fast, accurate, and scalable solutions for property price estimation.

IX. FUTURE WORK

While this study has demonstrated the viability of machine learning for real estate price prediction using the Boston Housing dataset, there are several avenues for future enhancement and exploration.

A. *Use of Modern and Larger Datasets*

Future research should incorporate recent, real-world housing datasets that include more features such as:

- Transaction year
- Zipcode-level economic indicators
- Nearby amenities (schools, parks, hospitals)
- Crime rates, walkability, and environmental factors

This would improve both relevance and scalability of the predictive models for commercial use.

B. Integration of Spatial and Temporal Data

Incorporating geospatial data (maps, satellite imagery) and time-series trends could enable the models to forecast prices based on urban development and seasonal market fluctuations. This would be particularly useful for dynamic real estate markets.

C. Model Enhancement

Advanced techniques like:

- XGBoost, LightGBM: For even better performance with large datasets
- LSTM or RNN: For modeling housing prices as a time series
- Explainable AI (XAI) tools like SHAP: To interpret model decisions in a user-friendly way

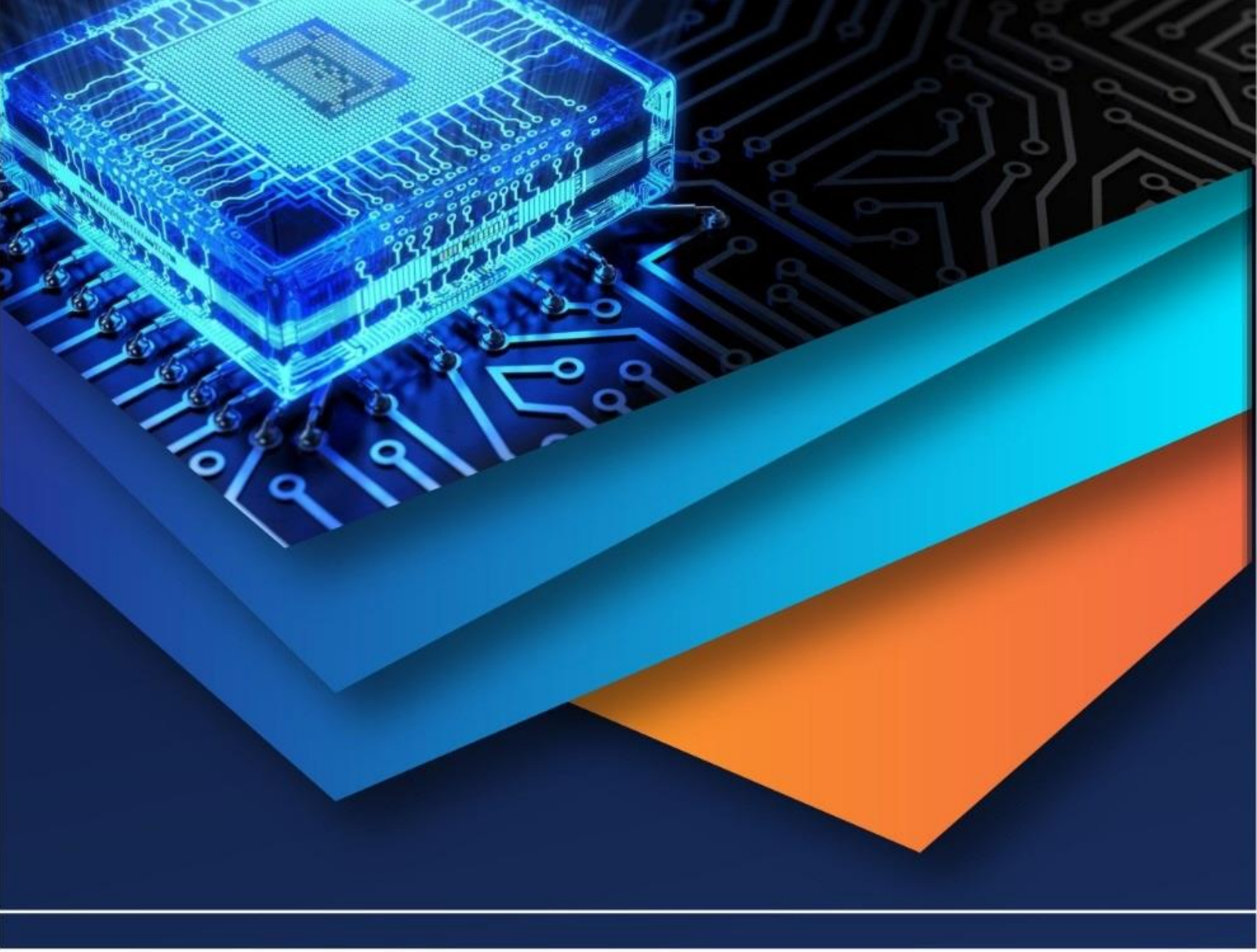
These can further improve prediction accuracy while maintaining transparency.

D. Deployment as a Web-Based Application

The trained models can be integrated into a web or mobile application that allows users (buyers, investors, or agents) to input house details and receive real-time price predictions. Technologies like Flask, Streamlit, or React.js with a backend API can support such deployment.

REFERENCES

[1] D. Harrison and D.L. Rubinfeld, “Hedonic prices and the demand for clean air,” Journal of Environmental Economics and Management, vol. 5, pp. 81–102, 1978.
[2] D. Belsley, E. Kuh, and R. E. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, Wiley, 1980.
[3] R. Quinlan, “Combining instance-based and model-based learning,” in Proc. 10th Int. Conf. on Machine Learning, Amherst, MA, 1993, pp. 236–243.
[4] A. Kumar and R. Singh, “Real Estate Valuation Using Machine Learning,” International Journal of Engineering Research & Technology (IJERT), vol. 9, no. 3, pp. 24–29, 2020.
[5] L. Zhang and H. Lee, “Comparative Study on Regression Models in Real Estate,” IEEE Access, vol. 7, pp. 106123–106132, 2019.
[6] S. Patel and V. Mehta, “Deep Learning for Housing Price Estimation,” Elsevier Journal of Advanced Computational Intelligence, vol. 33, no. 4, pp. 1012–1020, 2021.
[7] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed., O’Reilly Media, 2019.
[8] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, 2nd ed., Springer, 2021.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY