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

Performance analysis of machine learning classifiers: A Case study of House Price Prediction

[1] **Agrima Singh,** [2] **Aditya Maurya,** [3]**Spriha Singh,** [4]**Vandana Bhattacharjee**
[1,2,3] Kalinga Institute of Industrial Technology,
Bhubaneswar, Odisha
[4]Birla Institute of Technology, Mesra

Abstract: Accurate house price prediction is crucial for real estate analytics, as traditional methods often cannot capture the complex interplay of property features. This study compares Linear Regression, K-Nearest Neighbors (KNN) and Decision Tree Regression on a comprehensive housing data set with variables including area, bedrooms, bathrooms, stories and amenities such as main road access, guestroom, basement, hot water heating, air conditioning, parking and furnishing status. After systematic data cleaning, label encoding and feature scaling, models are evaluated using R^2 score, mean squared error and by binarizing prices at the median, accuracy, precision, recall and F1- score. Results show Linear Regression achieves the highest predictive accuracy while Decision Tree and KNN effectively capture non-linear patterns.

Keywords: *House Price Prediction, Machine Learning, Linear Regression, Decision Tree, K-Nearest Neighbors, Regression Metrics, Real Estate Data.*

1. INTRODUCTION

Predicting the value of residential properties is a key concern in real estate, as accurate price estimates support informed decisions for buyers, sellers, investors and policymakers. The challenge lies in the fact that house prices are influenced by a wide range of factors, including structural features, amenities and location-specific attributes. Traditional valuation techniques often fall short in capturing the nuanced relationships among these variables, motivating the adoption of machine learning [1], approaches for more reliable and automated predictions [8].

Machine learning [4],[9] provides powerful tools for uncovering complex patterns in data by leveraging mathematical models [3] trained on historical records. In this study, we utilize the Housing data set, which includes detailed information on property characteristics such as area, number of bedrooms and bathrooms, stories and amenities like main road access, guestroom, basement, hot water heating, air conditioning, parking and furnishing status. The target variable for prediction is the market price of each property.

Our research undertakes a comparative evaluation of three supervised machine learning algorithms: Linear Regression [5][7][10] K-Nearest Neighbors [6] and Decision Tree Regression [11][12]. Each of these models brings unique strengths - Linear Regression is valued for its simplicity and interpretability, KNN excels at capturing local trends and non-linearities and Decision Tree Regression is effective at modeling complex, hierarchical feature interactions.

To ensure robust and fair model comparison, the data set undergoes systematic preprocessing steps, including handling missing values, encoding categorical variables and scaling features. The models are trained and tested using both regression metrics (R^2 score, mean squared error) and by transforming prices into binary categories at the median, classification metrics such as accuracy, precision, recall and F1-score. This comprehensive evaluation framework is inspired by established methodologies in recent machine learning performance studies.

Through this analysis, we aim to identify which algorithm offers the best predictive performance for house price estimation [2] and to provide practical insights for applying machine learning in real estate analytics.

2. OVERVIEW OF THE ALGORITHMS

2.1 Linear Regression

Linear Regression is a widely used statistical and machine learning technique for modeling the relationship between a dependent variable and one or more independent variables. In the context of this housing data set, the dependent variable is the house price, while the independent variables include features such as area, number of bedrooms and bathrooms, number of stories, etc. The goal of Linear Regression is to find the best-fitting linear equation that can predict the house price based on the given features.

The model's coefficients provide direct insight into the relative importance of each feature in determining house prices. However, it assumes a linear relationship between the features and the target variable, which may not always capture the complexity of real-world housing markets.

2.2 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning algorithm that predicts the value of a target variable by considering the 'K' most similar instances in the training data. For regression tasks such as house price prediction, KNN identifies the K houses in the data set that are most similar to the property being evaluated, based on all available features.

Various distance metrics are used to calculate distance between query points and other data points such as Euclidean, Manhattan, Minokowski and Hamming distance. This approach does not make any assumptions about the underlying relationship between features and the target variable, allowing it to capture complex, non-linear patterns in the data.

KNN's primary advantage is its flexibility and ability to model non-linear relationships. However, its performance can be sensitive to the choice of K and the method used to measure similarity between instances.

2.3 Decision Tree Regression

Decision Tree Regression uses a tree-like structure to model decision rules based on feature values. At each node, the algorithm

splits the data according to the feature that results in the greatest reduction in prediction error, recursively partitioning the data set until the leaves represent final price predictions. Decision Trees can model complex, non-linear interactions between features and are easy to interpret, but they can be prone to overfitting if not properly pruned.

3. EVALUATION OF PARAMETERS

3.1 Accuracy

Ratio of accurately predicted prediction of learning model is called accuracy. Accuracy answers the question "Out of all the predictions we made, how many were true?"

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

3.2 Precision

This ratio gives the value of true positive among all positive predictions. It answers the question "Out of all the positive predictions we made, how many were true?"

$$Precision = \frac{TP}{TP + FP}$$

3.3 Recall

Same as True Positive Rate (TPR) that shows the ratio of true positive in total positive. It answers the question "Out of all the data points that should be predicted as true, how many did we correctly predict as true?"

$$Recall = \frac{TP}{TP + FN}$$

3.4 F1 Score

F1 Score is a measure that combines recall and precision. Harmonic mean of precision and recall which performs well in an imbalanced data set.

$$F1 = \frac{2.Precision.Recall}{Precision + Recall}$$

4. EXPERIMENTAL SETUP

The house price data set includes features such as area, bedrooms, bathrooms, stories and amenities (main road access, guestroom, basement, hot water heating, air conditioning, parking, furnishing status) with price as the target variable. Categorical features were label-encoded and all features were standardized for consistency. The data was split into training and test sets, typically using a 60:40, 70:30 and 80:20 ratio.

We implemented Linear Regression, K-Nearest Neighbors (K=3, K=5, K=7) and Decision Tree Regression using Python's scikit-learn library. Each model was trained on the training set and evaluated on the test set. Performance was measured with R^2 score and mean squared error for regression and by binarizing prices at the median, accuracy, precision, recall and F1-score for classification.

This setup ensures fair comparison of all models under the same conditions and provides reliable results for house price prediction.

5. RESULTS AND ANALYSIS

All three models—Linear Regression, K-Nearest Neighbors (KNN) and Decision Tree Regression—were tested on the house price data set. Linear Regression gave the best results, with the highest R^2 score and lowest mean squared error. Decision Tree and KNN performed moderately, but their predictions were less accurate than Linear Regression.

Table 1. Performance Analysis of Linear Regression for different Train/Test splits.

Models	**LR (60/40 split)**	**LR (70/30 split)**	**LR (80/20 split)**
Accuracy	**0.817**	0.817	0.789
Precision	**0.806**	0.778	0.794
Recall	0.874	**0.875**	0.833
F1 score	**0.839**	0.824	0.813

R squared error	**0.649**	0.616	0.633
Mean squared error	**16291705**	16532580	18527458

Table 1 shows the performance of the linear regression algorithm using different train/test splits. The results indicate that the 60/40 split delivered the best overall performance, achieving the highest F1 score (0.839) and the lowest mean squared error. Although accuracy remained consistent across splits, the 60/40 ratio offered the most balanced and dependable outcomes.

Table 2. Performance Analysis of K-Nearest Neighbour for different Train/Test splits.

Models	**KNN (K=3) 60/40 split**	**KNN (K=5) 70/30 split**	**KNN (K=7) 80/20 split**
Accuracy	0.784	0.780	**0.798**
Precision	0.810	0.768	**0.852**
Recall	**0.790**	0.787	0.767
F1 score	0.800	0.778	**0.807**
R squared error	0.476	**0.563**	0.547
Mean squared error	24342086	**18810398**	22890328

Table 2 shows the performance of the KNN algorithm with different train/test splits. The results indicate that the model with K=7 and an 80/20 split achieved the best overall performance, with the highest accuracy (0.798), precision (0.852) and F1 score (0.807). Although the K=5 model with a 70/30 split recorded the lowest mean squared error, the K=7 configuration provided the most balanced and effective results.

Table 3. Performance Analysis of Decision Tree for different Train/Test splits.

Models	**Decision Tree 60/40 split**	**Decision Tree 70/30 split**	**Decision Tree 80/20 split**
Accuracy	**0.729**	0.713	0.716
Precision	**0.746**	0.681	0.738
Recall	0.765	**0.775**	0.750
F1 score	**0.755**	0.725	0.744
R squared error	0.386	0.316	**0.480**

Mean squared error	28480088	29470551	**26278534**

Table 3 shows the performance of the Decision Tree algorithm across different train/test splits. Among the tested configurations, the 60/40 split achieved the highest accuracy (0.729) and F1 score (0.755), while the 80/20 split recorded the lowest mean squared error. Despite some variation in precision and recall, the 60/40 split demonstrated the most balanced and consistent overall performance.

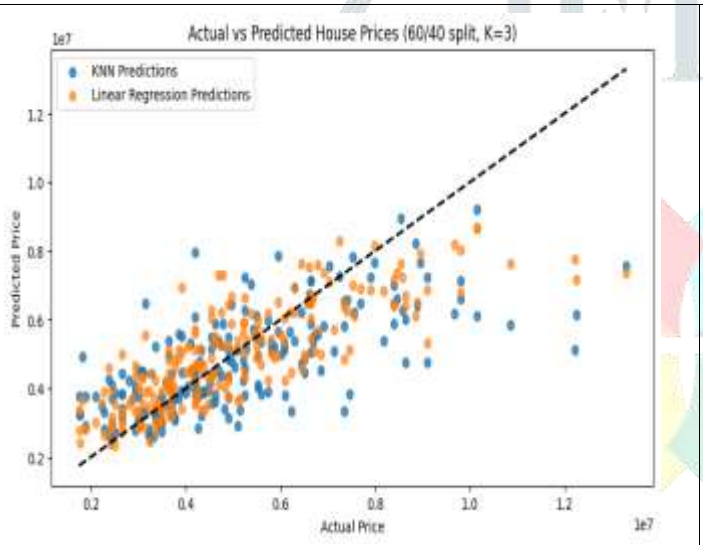

Figure 1: KNN (K=3) vs LR

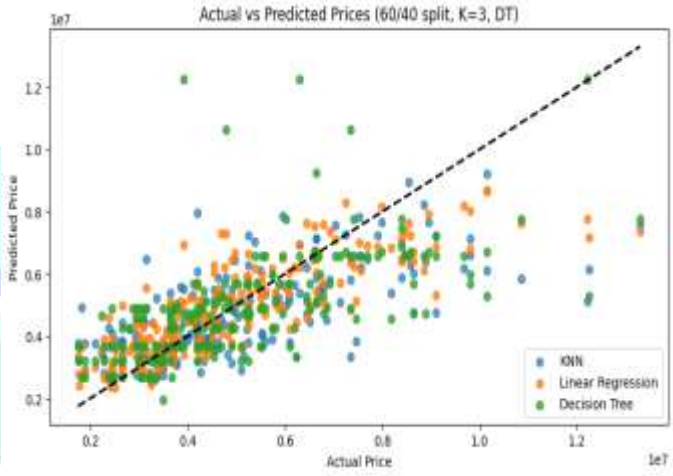

Figure 4: KNN (K=3) vs LR vs DT

Figures 1 - 6 present the graphs of predictions versus the actual values for all algorithms. Each point represents a prediction; the closer the point is to the dashed diagonal line, the more accurate the prediction.

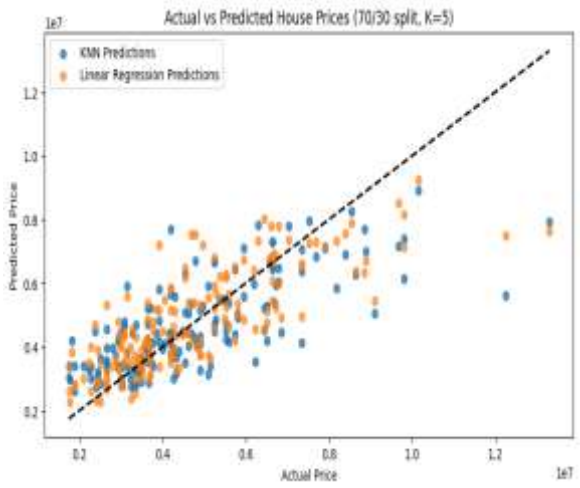

Figure 2: KNN (K=5) vs LR

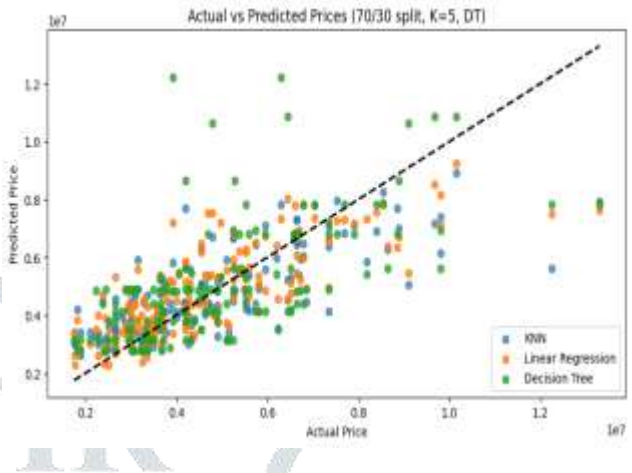

Figure 5: KNN (K=5) vs LR vs DT

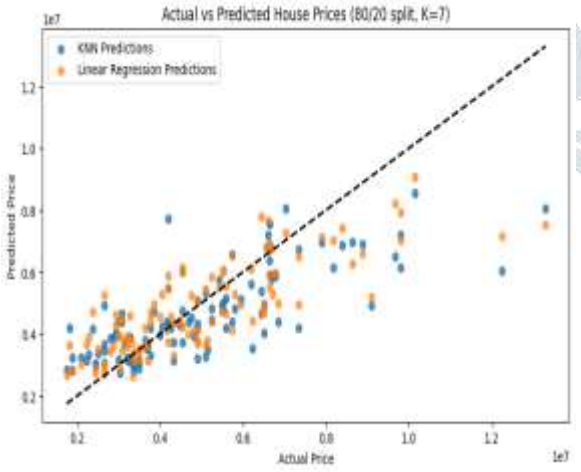

Figure 3: KNN (K=7) vs LR

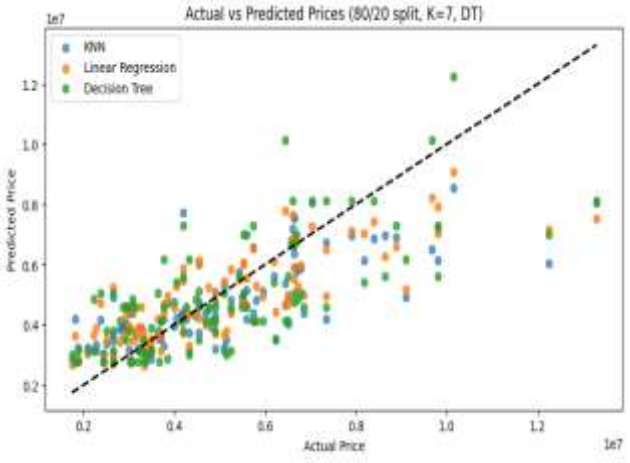

Figure 6: KNN (K=7) vs LR vs DT

Figures 1 - 3 compares predictions from K-Nearest Neighbors and Linear Regression.
Figures 4 - 6 compare predictions from K-Nearest Neighbors, Linear Regression and Decision Tree models.

Model	R^2 Score	MSE	Accuracy	Precision	Recall	F1-Score
LR	Highest	Lowest	Highest	Highest	Highest	Highest
Decision Tree	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate
KNN	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate

Table 4. Summary of all algorithms

From table 4 it is seen that Linear Regression has achieved the highest accuracy, precision, recall and F1-score; Decision Tree and KNN had similar classification results, but both were outperformed by Linear Regression.

6. CONCLUSION

This study presented a comparative analysis of three supervised machine learning algorithms - Linear Regression, K-Nearest Neighbors (KNN) and Decision Tree Regression - for house price prediction using the Housing data set. The data set included a range of structural features and amenities relevant to residential property valuation. All models were evaluated using regression metrics (R^2 score, mean squared error) and for classification purposes, accuracy, precision, recall and F1-score after binarizing prices at the median.

The experimental results indicate that Linear Regression consistently achieved the best predictive performance, with the highest R^2 score and the lowest mean squared error among the three algorithms. In the classification task, Linear Regression also led in accuracy, precision, recall and F1-score. Both Decision Tree and KNN models provided moderate results, capturing some non-linear patterns but generally showing less accuracy and higher error compared to Linear Regression.

7. REFERENCES

[1] Bhagat, A., Gosavi, M., Shahasane, A., Mishra, N., & Nerurkar, A. (2023). House Price Prediction Using Machine Learning. SSRN. https://ssrn.com/abstract=4413863

[2] Li, Z. (2024). A Comparative Study of Regression Models for Housing Price Prediction. *Transactions on Computer Science and Intelligent Systems Research*, 5, 810–816. https://doi.org/10.62051/qjs7y352

[3] Ghosh, K., & Bhattacharjee, V. (2024). Lung cancer prediction: A performance analysis of machine learning classifiers. *International Journal of Statistics and Applied Mathematics, 9*(5), 28–33. https://doi.org/10.22271/maths.2024.v9.i5a.1799

[4] E. Eze, S. Sujith, M. S. Sharif, and W. Elmedany, "A Comparative Study for Predicting House Price Based on Machine Learning," in *Proceedings of the 2023 4th International Conference on Data Analytics for Business and Industry (ICDABI)*, Bahrain, 2023, pp. 75–81. https://doi.org/10.1109/ICDABI60145.2023.10629399

[5] Parekh, J. (2023). House Price Prediction Using Linear Regression Model. *International Journal for Multidisciplinary Research (IJFMR)*, 5(6), 1–7.

https://www.ijfmr.com

[6] Kanadiya, P. J., & Chawan, P. M. (2024). A KNN-Linear Regression Fusion Approach for Improved Real Estate Price Estimation. *International Research Journal of Engineering and Technology (IRJET)*, 11(8), 688–695. https://www.irjet.net

[7] Burse, S., Anjaria, D., & Balaji, H. (2021). Housing Price Prediction Using Linear Regression. *Journal of Emerging Technologies and Innovative Research (JETIR)*, 8(10), d9–d12. https://www.jetir.org

[8] Ariyanti, N. P., Triayudi, A., & Sari, R. T. K. (2024). Analysis of K-NN Algorithm and Linear Regression to Predict House Prices in Jabodetabek. *SaNa: Journal of Blockchain, NFTs and Metaverse Technology*, 2(1), 65–71. https://doi.org/10.58905/sana.v2i1.265

[9] Banerjee, D., & Dutta, S. (2017). *Predicting the housing price direction using machine learning techniques*. In 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI) (pp. 2998-3000). IEEE. DOI: 10.1109/ICPCSI.2017.8392275

[10] Laishram, S., Kumar, R. S., & Priyadarshani, P. (2024). House Price Prediction Using Linear Regression In Machine Learning. *Journal of Artificial Intelligence Research & Advances*, 11(2), 92–100. https://journals.stmjournals.com/joaira

[11] Zhang, Z. (2021). Decision Trees for Objective House Price Prediction. In *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)* (pp. 280–283). Taiyuan, China: IEEE. https://doi.org/10.1109/MLBDBI54094.2021.00059

[12] Darshini, E. V. P., Vinuthna, I., Gayathri, G. B. S., Rani, G., & Roy, I. G. A. (2023). Prediction of House Price Using Machine Learning Algorithms. *International Research Journal of Modernization in Engineering Technology and Science*, 5(3), 906–911. https://doi.org/10.56726/IRJMETS34307