*Data and Metadata*. 2024; 3:.231
doi: 10.56294/dm2024.231

ORIGINAL

# Predicting Housing Sale Prices Using Machine Learning with Various Data Split Ratios

## Predicción de los precios de venta de la vivienda mediante aprendizaje automático con varios ratios de división de datos

Awais Azam[1], Alimul Haque[2], Sakshi Rai[1]

[1]Department of computer science LNCT University. Bhopal, India.
[2]Department of Computer Science, Veer Kunwar Singh University. Ara-802301, India.

**Cite as:** Azam A, Haque A, Rai SR. Predicting Housing Sale Prices Using Machine Learning with Various Data Split Ratios. Data and Metadata. 2024; 3:231. https://doi.org/10.56294/dm2024.231

**Submitted:** 10-01-2024 **Revised:** 22-08-2024 **Accepted:** 14-12-2024 **Published:** 15-12-2024

**Editor:** Adrián Alejandro Vitón Castillo

**Corresponding author:** Alimul Haque

## ABSTRACT

**Introduction:** recent advancements in technology and data analytics have propelled the rapid growth of artificial intelligence (AI) and machine learning (ML), which are now central to various industries. These technologies have become essential tools in many sectors, especially in predictive modeling for asset pricing.
**Objective:** from stock markets and rental properties to real estate and second-hand goods, AI and ML algorithms are widely applied to estimate values, optimize pricing strategies, and forecast market trends.
**Method:** by analyzing vast amounts of data, these tools enable more accurate predictions and informed decision-making, revolutionizing traditional approaches to pricing and valuation. In this study, the primary goal is to achieve the most accurate price prediction for houses or apartments by experimenting with different data split ratios.
**Result:** RMSE (House Price) 188965,28 is acceptable as best average price for houses.
**Conclusions:** the value of RMSE of this model are relatively low and also the value Squared Correlation is 64 % which is above the threshold of 50 %, so the predicted price of this model is seems appropriate, so I have presented this model and its predicted house price as final acceptable value for my research outcome.

**Keywords:** Data; Computational Methods; House Prediction.

## RESUMEN

**Introducción:** los recientes avances en tecnología y análisis de datos han impulsado el rápido crecimiento de la inteligencia artificial (IA) y el aprendizaje automático (AM), que ahora son fundamentales para diversas industrias. Estas tecnologías se han convertido en herramientas esenciales en muchos sectores, especialmente en la elaboración de modelos predictivos para la fijación de precios de activos.
**Objetivo:** desde los mercados de valores y las propiedades en alquiler hasta los bienes inmuebles y de segunda mano, los algoritmos de IA y ML se aplican ampliamente para estimar valores, optimizar estrategias de fijación de precios y predecir tendencias de mercado.
**Método:** al analizar grandes cantidades de datos, estas herramientas permiten realizar predicciones más precisas y tomar decisiones informadas, revolucionando los enfoques tradicionales de fijación de precios y valoración. En este estudio, el objetivo principal es lograr la predicción de precios más precisa para casas o apartamentos experimentando con diferentes ratios de división de datos.
**Resultado:** RMSE (Precio de la casa) 188965,28 es aceptable como mejor precio medio para casas.
**Conclusiones:** el valor de RMSE de este modelo es relativamente bajo y también el valor de Correlación al

© 2024; Los autores. Este es un artículo en acceso abierto, distribuido bajo los términos de una licencia Creative Commons (https://creativecommons.org/licenses/by/4.0) que permite el uso, distribución y reproducción en cualquier medio siempre que la obra original sea correctamente citada

Cuadrado es del 64 % que está por encima del umbral del 50 %, por lo que el precio predicho de este modelo parece apropiado, así que he presentado este modelo y su precio predicho de la casa como valor final aceptable para el resultado de mi investigación.

**Palabras clave:** Datos; Métodos Computacionales; Predicción del Precio de la Vivienda.

## INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) have gained significant attention in recent years, driven by advancements in technology and improvements in data analytics. These tools have proven instrumental in various domains, including predictive modeling for pricing assets such as stocks, rental properties, homes, and used goods.[1] As the demand for homeownership continues to rise among middle-class families, the real estate industry has responded with a diverse range of housing options, including new constructions and pre-owned properties.[2,3] Accurate property valuation has become essential, allowing potential buyers to make informed decisions while ensuring sellers receive a fair price.[4]

This study focuses on identifying appropriate pricing strategies for homes in urban areas, including state capitals. Factors such as location, floor plan, amenities (like parking spaces or lawns), and additional features (such as terraces) are analyzed to provide reliable estimates tailored to buyers' needs. A linear regression model is employed to predict housing prices, aiming to establish the relationship between property features and their market value.[5] By experimenting with various data splits, such as training and testing ratios, the study determined that an 80/20 division yielded the most reliable results, making it a preferred approach for real estate analytics. Determining the optimal selling price is critical. Overpricing a property can deter potential buyers, prolonging the sale process, while underpricing results in financial losses for sellers. This underscores the need for a balanced pricing approach that meets market demands and aligns with customer expectations. Leveraging the insights provided by machine learning, real estate stakeholders can evaluate property values with greater precision, ensuring competitive pricing strategies.[2]

Machine learning models, including linear regression, are widely used in diverse fields beyond real estate. Examples include pandemic-related diagnostics, product performance analyses, vehicle tracking, and more.[6,7] Linear regression, in particular, is valued for its ability to identify relationships between dependent and independent variables, providing a clear framework for understanding data patterns.[8] In this context, it serves as a robust tool for real estate cost estimation, combining statistical principles with predictive capabilities to meet practical business needs effectively.[3,8]

This paper aims to explore the methods and technologies applied to analyze economic and productivity indicators for various Indian cities, using advanced data analytics. By focusing on indicators such as R&D expenditure, employment rates, and ICT sector contributions, the study highlights how data analytics tools like machine learning, predictive modeling, and visualization techniques are being used to process large datasets and derive meaningful insights. Finally, the paper addresses the challenges posed by incomplete data, regional disparities, and the complexities of integrating diverse indicators.

The evolution of home price prediction has deep roots in the application of mathematics and statistical modeling. Initially, traditional methods primarily relied on basic regression techniques and straightforward numerical computations to estimate property values. However, with the exponential growth of datasets and the increasing availability of computational power, machine learning has emerged as a transformative tool for improving prediction accuracy. This paper builds upon this progression by harnessing advanced machine learning techniques to redefine predictive modeling in real estate. Modern valuation models aim to achieve unparalleled levels of precision, adaptability, and reliability in estimating property prices. By integrating these advanced algorithms, the field has shifted towards more sophisticated approaches that consider the dynamic nature of real estate markets, leading to more robust and flexible predictive systems.

### System design

The system's architecture for estimating real estate costs illustrates the sequential steps required to develop a reliable pricing model. The process begins with data preprocessing, where missing values are addressed, and categorical variables are encoded for compatibility with machine learning algorithms. The cleaned dataset is then split into two subsets: one for training the model and another for testing its performance.[10] Using the training data, separate models are built using random forest and gradient boosting algorithms, both of which are designed to predict property prices based on various input features. Once the individual models are trained, their predictions are combined into a composite model, leveraging the strengths of both approaches to enhance overall accuracy. This integrated model is then deployed as a practical tool for property valuation.

The deployment process includes monitoring the model's performance in real-world scenarios and ensuring its accuracy by periodically updating it with fresh data. This continuous learning approach helps maintain the model's relevance in dynamic real estate markets. A visual representation of this system's design provides an overview of its components, from data preparation to final deployment as a property value estimator.

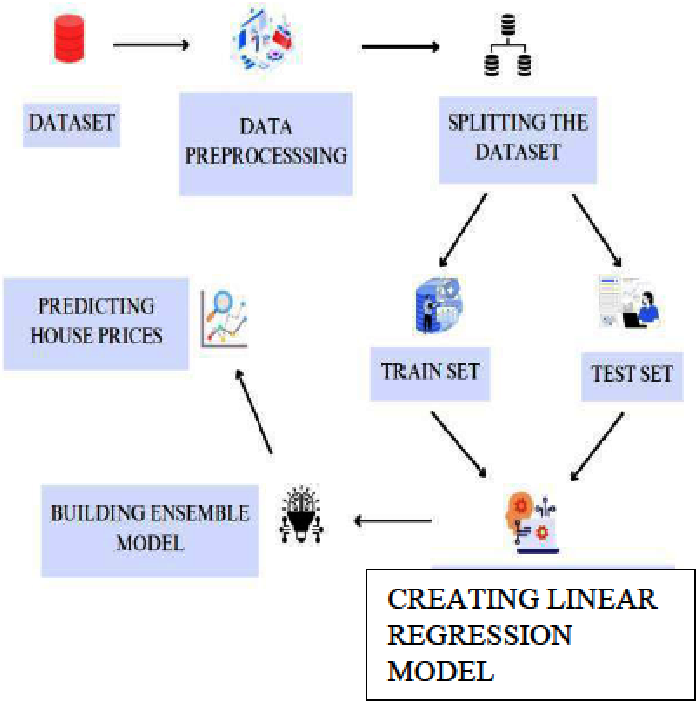

**Figure 1.** System Design

## METHOD

The process of forecasting real estate prices involves multiple phases, including data acquisition, preprocessing, model development, evaluation, and result generation. In this study, the primary goal is to achieve the most accurate price prediction for houses or apartments by experimenting with different data split ratios. To accomplish this, the implementation was divided into three key approaches:

1. Implementation-I: A data split ratio of 60:40, where 60 % of the data is used for training and 40 % for testing.
2. Implementation-II: A data split ratio of 70:30, allowing for a slightly larger training set while retaining a robust testing set.
3. Implementation-III: An 80:20 split, maximizing the training data for better model learning while leaving a smaller portion for evaluation.

Each implementation phase follows a systematic progression, from preparing the dataset to training the predictive models under the specified split ratios. The final stage involves testing the models and analyzing their results to determine the most effective approach for accurate property price prediction. This structured methodology ensures a comprehensive evaluation of the model's performance across different scenarios, leading to reliable and actionable forecasting outcomes.

### Research implementation

To develop an optimal house price prediction using a Multiple Linear Regression Model, the implementation was divided into three phases, each based on distinct data split ratios for the training and testing datasets. These split ratios were chosen to evaluate the model's performance under varying conditions, ensuring a comprehensive analysis of its predictive accuracy. After completing each phase, the results were compiled and integrated into a subsequent section, providing a comparative analysis of the different approaches. This methodology highlights the effectiveness of data partitioning in enhancing the model's reliability and accuracy.

*Implementation Process Phase-I with split ratio 65:35*

In the initial phase of implementation, a data split ratio of 65:35 was applied, where 65 % of the data was allocated to the training set and 35 % to the test set. This approach aimed to balance the amount of data used for model training and validation. Utilizing the house sales dataset, the process was developed and executed using RapidMiner, as illustrated in figure 2. This setup provided a structured foundation for analyzing the model's performance and refining predictive accuracy.

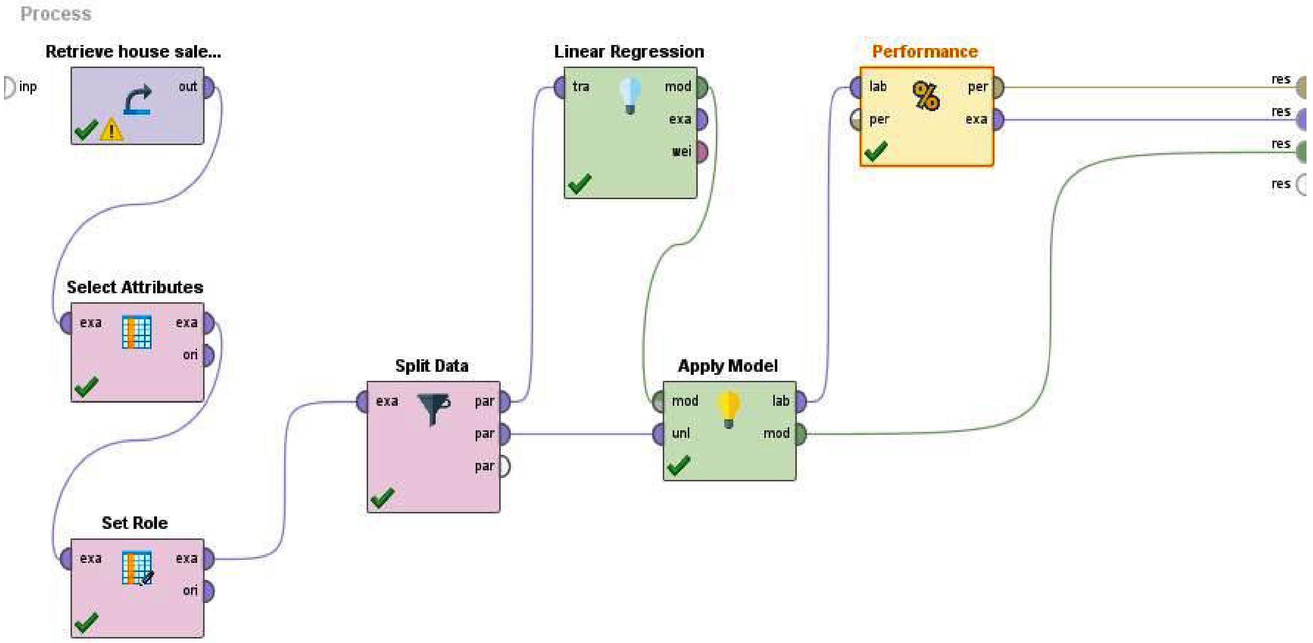

**Figure 2.** Linear regression model Process-1 with split ratio 65:35

*Implementation Process Phase-II with split ratio 70:30*

In the second phase of implementation, a data split ratio of 70:30 was employed, allocating 70 % of the dataset for training and 30 % for testing. This approach aimed to provide a slightly larger training dataset to enhance the model's learning capability while maintaining sufficient data for performance evaluation. The process was designed and executed using RapidMiner, as depicted in figure 3. This phase further refined the predictive model, enabling a more robust analysis of house sales data.

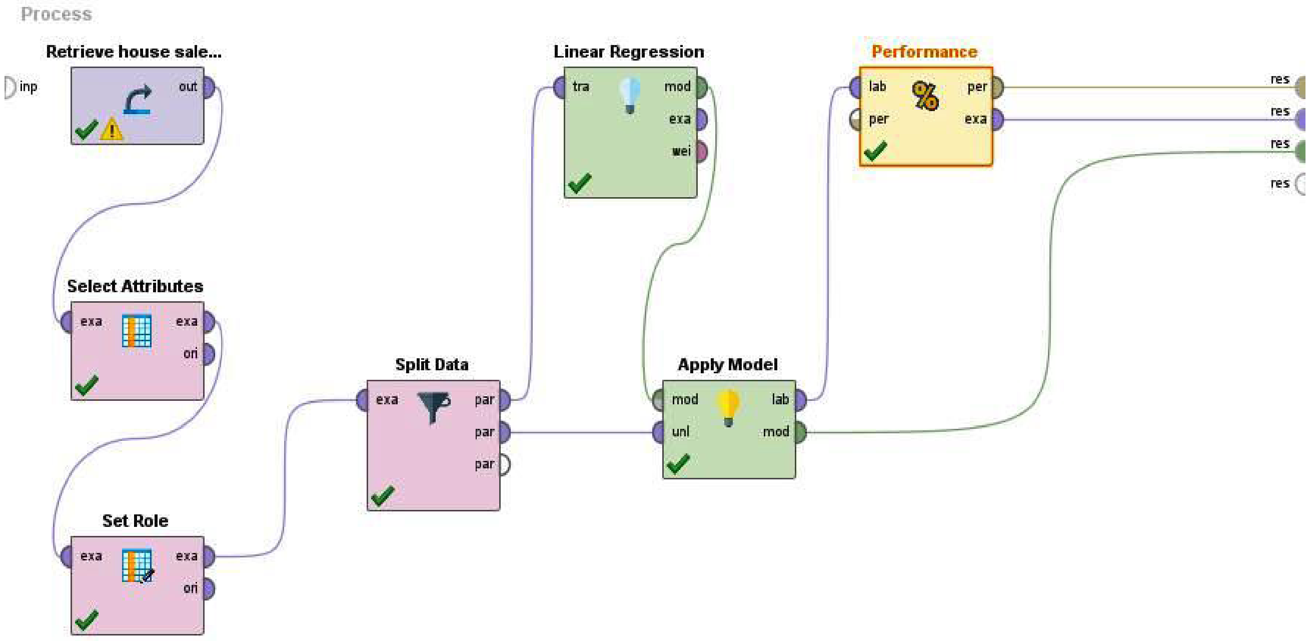

**Figure 3.** Linear regression model Process-II with split ratio 70:30

*Implementation Process Phase-III with split ratio 80:20*

In the third phase of implementation, an 80:20 data split ratio was chosen, with 80 % of the dataset allocated for training and 20 % reserved for testing. This configuration aimed to maximize the data available for model training, enhancing its ability to learn patterns effectively while maintaining a reliable test set for evaluation. The house sales dataset was processed and analyzed using RapidMiner, with the workflow visually depicted in figure 4. This phase represents the final stage of experimentation, focusing on optimizing the predictive performance of the model.

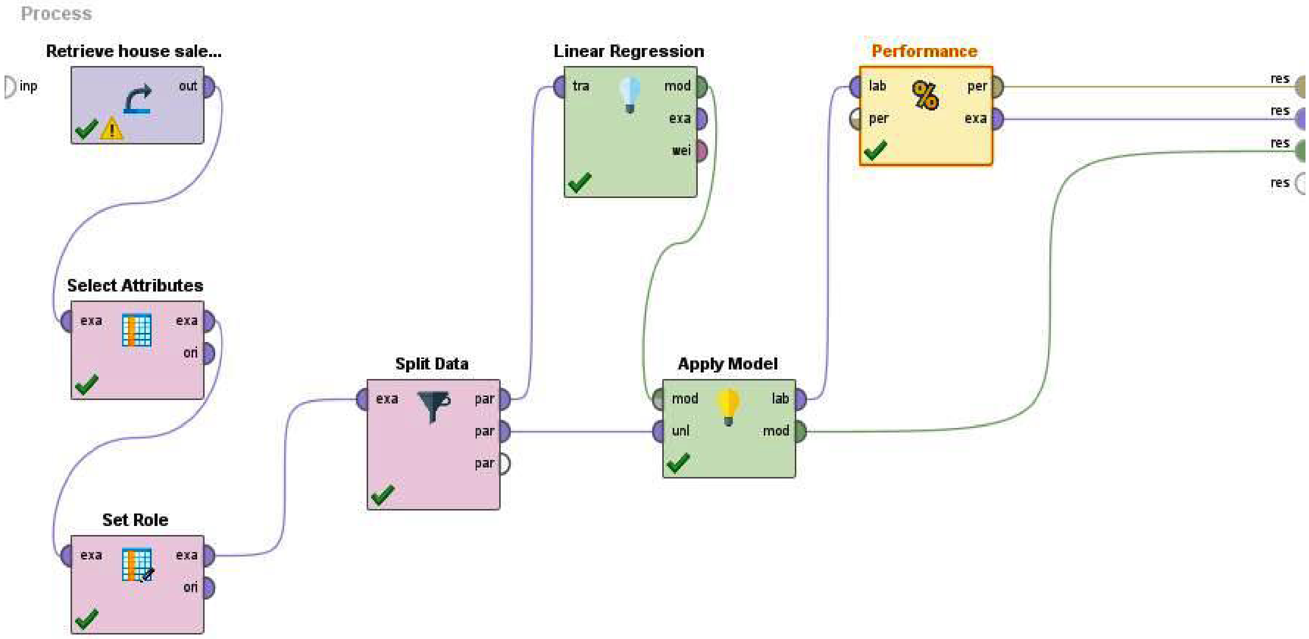

**Figure 4.** Linear regression model Process-III with split ratio 80:20

## RESULT

*Comparative Analysis of the outcome*

Upon completing the three phases of house price prediction testing using multiple linear regression with varying data split ratios, the results were consolidated into a table for comparison. Two key performance metrics were chosen to evaluate the models:

1. Root Mean Squared Error (RMSE): This metric represents the average difference between predicted and actual values, reflecting the model's accuracy in estimating the optimal average house price.
2. Squared Correlation ($R^2$): This metric measures the proportion of variance in the target variable that is explained by the independent variables, indicating the model's effectiveness in capturing relationships within the data.

The combined results from all three implementations are summarized in table 1, providing insights into the comparative performance of each approach and guiding the selection of the most effective model configuration.

**Table 1.** Integrated results of all the three phases implemented processes

| Implemented Process | Data Split Ratio | Value of RMSE | Value of Squared Correlation |
|---|---|---|---|
| Process Phase-I | 65:35 | 242614.52 | 44 % |
| Process Phase-II | 70:30 | 256031.09 | 49 % |
| Process Phase-III | 80:20 | 188965.28 | 64 % |

## CONCLUSION

This research focuses on validating the use of a multiple linear regression model for predicting housing sale prices while addressing limitations of simple linear regression. Simple linear regression, relying on a single independent variable, often fails to capture the complexity of real-world data, leading to suboptimal predictions. To overcome this, multiple linear regression, incorporating multiple predictors, was employed using three data split ratios: 65:35, 70:30, and 80:20. The 65:35 model was discarded due to unmet variance assumptions, while the 70:30 model showed suboptimal RMSE and squared correlation values. The 80:20 model, with an RMSE improvement and 64 % variance explanation, was deemed most accurate and adopted as the final predictive model.

The methodologies and models discussed in this research establish a robust framework that can be expanded for more advanced analytical tasks, such as building recommendation systems or conducting time-series forecasting. Furthermore, future research could explore a comparative analysis of house price predictions using various machine learning techniques like Lasso regression, Support Vector Machines, Artificial Neural Networks (ANN), and XGBoost. These models offer diverse strengths in handling different data complexities and could provide deeper insights into predicting real estate prices with even greater accuracy and reliability. Expanding to these approaches could significantly enhance predictive modeling in this domain.

## REFERENCES

1. Awais Azam M, Rai S, Shams Raza M. Predictive Analytics for Housing Market Trends and Valuation. Management. 2025 Jan 1;3 SE-Or:115. Available from: https://doi.org/10.62486/agma2025115

2. Fourkiotis KP, Tsadiras A. Comparing Machine Learning Techniques for House Price Prediction. In: IFIP International Conference on Artificial Intelligence Applications and Innovations. Springer; 2023. p. 292-303.

3. Choy LHT, Ho WKO. The use of machine learning in real estate research. Land. 2023;12(4):740.

4. Park B, Bae JK. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Syst Appl. 2015;42(6):2928-34.

5. Pai PF, Wang WC. Using machine learning models and actual transaction data for predicting real estate prices. Appl Sci. 2020;10(17):5832.

6. Md Alimul Haque DS, Shameemul Haque MR and, Kumar K. Learning Management System Empowered by Machine Learning. In: AIPCP21-AR-CRSE2021-00085 Recent Trends in Science and Engineering (CRSE2021). 2021.

7. Zeba S, Haque MA, Alhazmi S, Haque S. Advanced Topics in Machine Learning. Mach Learn Methods Eng Appl Dev. 2022;197.

8. Azrar A, Ali Y, Awais M, Zaheer K. Data mining models comparison for diabetes prediction. Int J Adv Comput Sci Appl. 2018;9(8):320-3.

9. Whig V, Othman B, Gehlot A, Haque MA, Qamar S, Singh J. An Empirical Analysis of Artificial Intelligence (AI) as a Growth Engine for the Healthcare Sector. In: 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE; 2022. p. 2454-7.

10. Oh SY, Hang SP, Wang JTW. Prediction of residential property prices using machine learning algorithms. In: ITM Web of Conferences. EDP Sciences; 2024. p. 1042

## AVAILABILITY OF DATA AND MATERIALS

The datasets used in this research are publicly available(Kaggel) and properly cited in our dataset section for transparency and ease of replication.

## FUNDINGS

None.

## CONFLICT OF INTEREST

None.

## AUTHORSHIP CONTRIBUTION

*Conceptualization:* Awais Azam, Alimul Haque and Sakshi Rai.
*Investigation:* Awais Azam, Alimul Haque and Sakshi Rai.
*Methodology*: Awais Azam, Alimul Haque and Sakshi Rai.
*Writing - original draft:* Awais Azam, Alimul Haque, Sakshi Rai.
*Writing - review and editing:* Awais Azam, Alimul Haque, Sakshi Rai.