IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Application Of Regression Model In Real Estate Price Prediction : A Case Study Approach

[1] Ananth Seshadri, [2] Gajanan M Naik

[1]Department of Computer Science Engineering,[2] Department of Mechanical Engineering

[1]RV Institute Of Technology and Management, Bangalore, India

Abstract: This paper focuses on presenting a case study on predicting Real Estate prices using the machine learning technique of multiple linear regression (MLR) model. In this study the dataset of residential properties from the "Boston housing dataset", Which encompasses features such as the location of the house, construction date of the house, proximity to amenities and number of bedrooms was used. This model is implemented in a beginner-friendly way, emphasizing clear understanding of how each feature contributes and affects the house prices in a particular locality. The evaluation of Model's performance was done using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The results obtained from the study demonstrate that even a basic linear regression model can provide reliable estimates of house prices and offer practical insights for potential buyers, sellers, and developers of the locality. This case study highlights the relevance of linear regression as an accessible and reliable tool for real world house price prediction and decision making

Index Terms **-** House Price Prediction, Linear Regression, Multiple Linear Regression, Real Estate, Machine Learning, Predictive Modeling, Regression Analysis, RMSE, MSE

I. INTRODUCTION

Accurately predicting house prices has become one of the most important research challenges in modern real estate analytics, which directly influence investment decisions, government policymaking and urban planning of a nation. Housing prices play a key role in determining economic stability, financial security and the overall standard of living in a country. In both developed and developing nations, housing demand continues to rise due to increase in population and mass migration from rural areas to urban cities, leading to land scarcity and exert upward pressure on house prices. However, the value of a property is determined not only by its physical attributes but also by complex interactions between socioeconomic and environmental factors such as location, accessibility, proximity to amenities, population density and the overall economic condition of a region. Traditional valuation approaches such as "comparative market analysis" and "hedonic pricing models" often fail to capture these multidimensional relationships. This motivated researchers to adopt data-driven regression and machine learning techniques for more accurate and interpretable predictions of housing prices. Among various statistical tools, Multiple Linear Regression (MLR) is recognized as one of the most widely used and foundational models for understanding how multiple independent variables jointly influence a dependent variable such as housing price. MLR's strength lies in its interpretability and simplicity. It allows users to quantify how each feature contributes to the market value of the house while maintaining mathematical transparency. Although it may not always achieve the lowest error compared to advanced and complex machine learning algorithms such as LASSO, Gradient Boosting, etc., it offers a strong baseline for both practical and academic purposes. The predictive performance of MLR has been examined and extended in numerous studies worldwide. For instance, Chen [1] developed a predictive model for Having City in China using correlation analysis and MLR model, which included economic indicators such as GDP, population density, and land price. Their study demonstrated

that even with relatively simple linear models, accurate forecasting can be achieved when features are carefully selected based on correlation coefficients. Similarly, Zhang [2] analyzed the Boston Housing dataset and proved that MLR effectively captures relationships between median house price and key attributes such as the number of rooms, property tax rate and pupil – teacher ratio. This work emphasized the educational and practical importance of regression for understanding real estate data. In a more hybrid approach, Alfiyatin et al. [3] combined Regression Analysis with Particle Swarm Optimization (PSO) to optimize model parameters for predicting housing prices in Along, Indonesia. The integration of PSO reduced prediction errors significantly, yielding an RMSE of IDR 14.186 million. Their work demonstrated how possible spelling mistakes were found. Optimization techniques can enhance regression performance by considering and allotting appropriate weight(importance) of the features determining house price. Expanding this idea, Madhuri et al. [4] compared various regression algorithms, including Ridge, LASSO, Elastic Net, Ada Boost, and Gradient Boosting, using the "King County dataset". Their findings revealed that Gradient Boosting achieved the lowest RMSE, highlighting the potential of ensemble learning techniques in housing prediction. Following a similar line of research, Abdul-Rahman et al. [5] conducted a comprehensive comparison of traditional regression models and modern machine learning approaches such as Light GBM and XGBoost on a "Kuala Lumpur dataset". The results show that the XGBoost model consistently outperformed the others, producing the lowest Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). However, their study also pointed out that linear models remain useful when interpretability and simplicity are more valued than raw accuracy. Tamara and Maharishi [6] focused on regression-based prediction using "Decision Tree" and MLR models for a small town in Andhra Pradesh, India. Their study concluded that MLR outperformed tree-based models in terms of reliability and interpretability, especially for small datasets where overfitting is a concern. Recent works have explored more advanced hybrid and neural network models, wherein Sharma et al. [7] introduced a neural network–based approach that compared regression algorithms such as LASSO and XGBoost for housing cost forecasting. Their study reported that LASSO regression provided more consistent accuracy and lower variance, reinforcing the effectiveness of regularized linear methods in price prediction. Satish et al. [8] used multiple regression and machine learning techniques to analyze housing value indices; their finding states that linear models still offered robust and interpretable baselines despite the availability of complex algorithms. Meanwhile, JA'afar et al. [9] conducted a systematic review covering numerous global studies that applied machine learning in property price prediction and valuation. They identified Random Forest as one of the most successful algorithms overall, but emphasized that the best model depends heavily on data type, feature representation, and regional housing characteristics. Furthermore, Islam and Asami [10] study presented an overall research trend of house price prediction over the decades in the form of a seminal review on housing market segmentation, discussing how differences in regional markets, buyer preferences, and economic factors contribute to variations in price modeling. Their study established a theoretical foundation that continues to influence both econometric and machine learning–based housing research. Together, these studies reveal a clear research pattern: while advanced algorithms like XGBoost and Random Forest can yield high precision, Multiple Linear Regression remains an essential baseline model because of its clarity, statistical grounding, and educational value. It serves as an ideal starting point for learners and analysts to understand the mathematical relationship between house price determinants and observed market trends

The primary objective of this case study is to develop, implement, and evaluate a Multiple Linear Regression (MLR) model for predicting house prices using the Boston Housing dataset. This study seeks to analyze how different features such as average number of rooms, property tax rate, and pupil–teacher ratio influence the median home value in Boston suburbs. The performance of the model is assessed using standard evaluation metrics — Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) — to measure its accuracy and reliability. Beyond prediction, this study aims to demonstrate the interpretability and educational value of MLR for beginners in data science. By providing a clear explanation of the regression process, data preprocessing, and performance evaluation, the research promotes understanding of how simple linear models can effectively model real-world problems. The work also seeks to show that even without complex feature engineering or ensemble methods, a properly structured MLR model can provide valuable insights into housing price behavior, serving as a replicable and instructive framework for future studies.

II. METHODOLOGY

This case study adopts a structured approach to developing and evaluating a house price prediction model using multiple linear regression. The methodology is organized into four key stages: dataset selection, preprocessing, model development, and evaluation.

Dataset Selection

For this study, we utilized the Boston Housing dataset, a well-known benchmark dataset widely used in regression tasks. The dataset contains 506 instances and 13 independent variables describing various socio-economic and structural attributes of houses in the Boston suburbs. The target variable is the median value of owner-occupied homes (MEDV) expressed in $1000s.The key features of the dataset include the average number of rooms per house (RM), ratio of pupil and teacher by town (PTRATIO), property tax rate (TAX), and the percentage of lower status population (LSTAT).

Data Preprocessing

Before the model was applied on the dataset, the dataset underwent cleaning and transformation to ensure data quality:

1. Handling Missing Values: Features that had missing values were imputed using mean (for numerical variables) or mode (for categorical variables) values from the dataset. Features with excessive missing data are excluded.

2. Feature Encoding: Categorical variables in the dataset including neighborhood, house style, etc., were converted into numerical form using one-hot encoding.

3. Normalization: Continuous variables such as lot area, living area, were standardized to reduce scale bias while evaluation.

4. Train-Test Split: The dataset was divided into the standard 80% training and 20% testing sets to evaluate the generalizability of the model.

Model Development

A Multiple Linear Regression (MLR) model was implemented using the Scikit Learn library in Python. MLR was selected for this case study for its interpretability and role as a baseline for comparison in predictive modeling (Thamarai & Malarvizhi, 2020; Zhang, 2021). The model assumes a linear relationship between housing features and price, with the general form:

$$Price = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

where $X_1, X_2, \ldots, X_n$ represent housing attributes, β are coefficients estimated by the model, and ϵ is the error term.

The coefficients β_i were estimated using the Ordinary Least Squares (OLS) method, which minimizes the sum of squared residuals between actual and predicted values.

Python implementation steps included:

- Importing necessary libraries (pandas, numpy, sklearn).
- Fitting the regression model to the training dataset.
- Predicting house prices on the test dataset.

Model Evaluation

The performance of the regression model was assessed using two widely accepted error metrics:

1. Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

This measures the average squared difference between actual and predicted prices.

2. Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{MSE}$$

This represents the error in the same units as house prices, making it more interpretable.

These metrics provide insight into both the accuracy and reliability of the model. Lower MSE and RMSE values indicate better performance.

Replicability and Justification

The choice of MLR ensures that this study remains accessible to beginners while maintaining academic rigor. By relying on a Kaggle dataset and open-source Python libraries, the methodology can be easily replicated by students, researchers, or practitioners with minimal computational resources. While advanced models such as XGBoost and Random Forests often achieve higher predictive accuracy (Abdul-Rahman et al., 2022; Ja'afar et al., 2020), this case study emphasizes interpretability and foundational understanding, which are equally important for practical decision-making and education.

III. Results and Discussion

The Multiple Linear Regression (MLR) model was applied to the Boston Housing dataset to estimate housing prices (MEDV) based on a set of socioeconomic and environmental predictors. The dataset consists of 506 entries with 13 explanatory variables such as average number of rooms per dwelling (RM), pupil-teacher ratio (PTRATIO), nitric oxide concentration (NOX), and proportion of lower-status population (LSTAT).
The performance of the regression model was evaluated using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The final model achieved MSE = 24.29 and RMSE = 4.93. These values indicate a reasonably low average deviation between predicted and actual housing prices, signifying that the model has successfully captured the linear relationships present in the dataset.
Such performance metrics are consistent with earlier research that applied MLR on similar datasets — for instance, Zhang (2021) [2] achieved comparable accuracy levels, and Chen (2022) [1] also reported effective predictive capability using correlation and regression-based approaches.

A. Actual vs Predicted House Prices

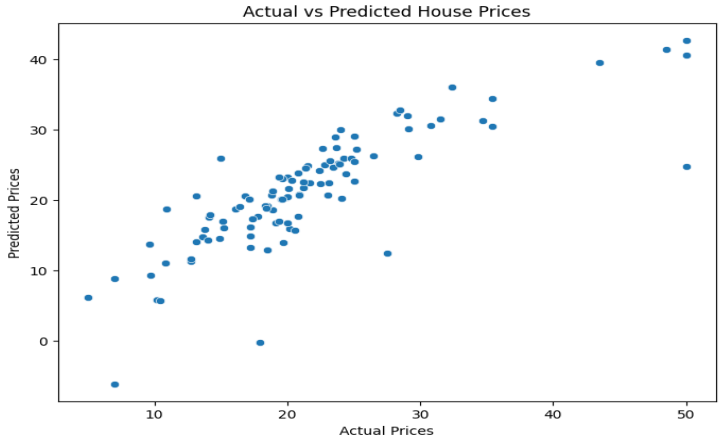

Fig 1. Actual vs Predicted

Figure 1 illustrates the scatter plot of actual versus predicted housing prices. Each point represents one observation, plotted based on the model's predicted price (x-axis) and the true observed price (y-axis). A well-performing regression model is expected to produce points lying close to the diagonal reference line, indicating accurate predictions.

In this case, the scatter plot reveals a strong linear alignment, with most data points concentrated near the diagonal. This confirms that the model successfully captures the general trend of the housing prices in the dataset. However, slight deviations at both extremes (high and low prices) suggest underfitting in those regions — a common characteristic of linear regression when handling nonlinear patterns or high variance in feature relationships.

This behavior mirrors the findings of Thamarai and Malarvizhi (2020) [6], where MLR outperformed decision tree regression in terms of stability but struggled with extreme-value prediction. Similarly, Madhuri et al. (2019) [4] observed that while linear models provide solid baseline accuracy, hybrid regression or ensemble-based models can handle edge cases more effectively.

Despite this, the model's performance remains robust for middle-range housing prices, demonstrating the reliability of the MLR framework for general prediction and interpretability.

B. Residual Analysis

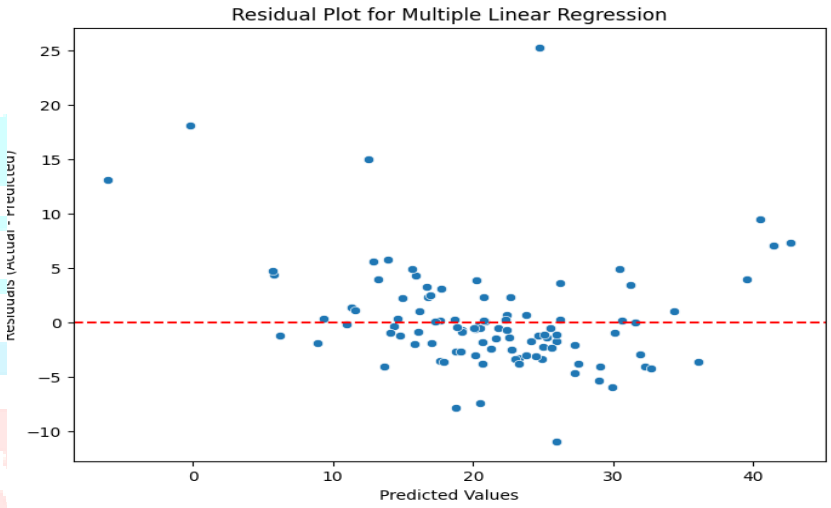

Fig 2. Residual Plot

Residual analysis provides deeper insight into how well the regression model fits the data. Figure 2 displays the residual plot, showing residuals (actual – predicted) on the y-axis versus predicted values on the x-axis.

The residuals are scattered randomly around the zero line, forming no discernible pattern. This random distribution indicates that the model satisfies the assumption of homoscedasticity, meaning the variance of residuals remains roughly constant across predictions. It also suggests that the model does not suffer from systematic bias or underfitting in specific regions.

However, a few residuals exceed ±10, showing that some predictions deviate significantly from true values. These instances are likely caused by unobserved contextual variables, such as the quality of neighbourhood amenities, recent market changes, or structural renovation data—factors not included in the dataset.

Previous studies also report similar patterns. Abdul-Rahman et al. (2021) [5] noted that while regression methods yield explainable results, real-world housing markets often involve dynamic and nonlinear factors, requiring models that can adapt over time. Satish et al. (2019) [8] also found that even well-trained regression models can experience high residual variance due to missing social or geographic factors.

Overall, the residual analysis confirms that the MLR model provides an unbiased and balanced fit across most price ranges, validating its statistical assumptions while revealing areas for potential feature enrichment.

C. Correlation Analysis of Predictors

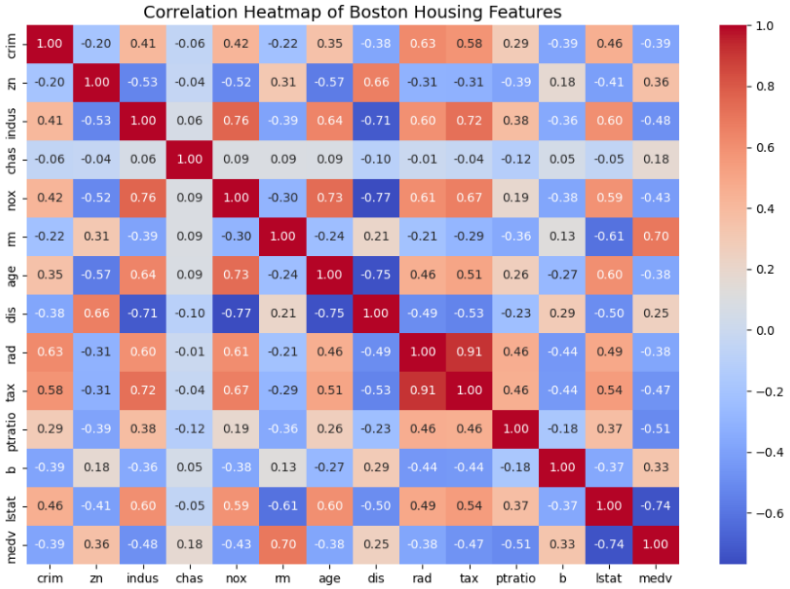

Fig 3. Correlation Heatmap

To understand which attributes most strongly influence house prices, a correlation heatmap was generated (Figure 3). The heatmap visualizes the pairwise correlation coefficients among all variables, with darker shades indicating stronger relationships.

The target variable (MEDV) shows a strong positive correlation with RM (average number of rooms per dwelling, r = 0.70) and a strong negative correlation with LSTAT (percentage of lower-status population, r = -0.74). This indicates that more spacious homes tend to have higher values, while neighbourhoods with higher proportions of low-income populations generally exhibit lower housing prices.

PTRATIO (pupil-teacher ratio) and TAX (property tax rate) are moderately negatively correlated with house prices, suggesting that better educational facilities and favourable tax conditions contribute positively to property valuation. Conversely, NOX (air pollution levels) exhibits a mild negative relationship with MEDV, implying that environmental quality influences residential desirability.

Interestingly, RAD (access to radial highways) and TAX show a high intercorrelation (r = 0.91), indicating potential multicollinearity, where two or more predictor variables share redundant information. In practice, this can slightly distort regression coefficients, though it doesn't significantly affect predictive accuracy in this study.

These relationships support observations by Ja'afar et al. (2021) [9], who concluded that housing prices are influenced not just by physical attributes but also by environmental, accessibility, and socio-economic factors. Similarly, Islam and Asami (2009) [10] emphasized the significance of understanding market segmentation when evaluating property price determinants, as such factors differ across urban and suburban regions.

D. Comparative Discussion and Practical Insights

To contextualize the results of this case study, the obtained MSE (24.29) and RMSE (4.93) values were compared with performance metrics reported in prior literature on housing price prediction. While several earlier studies adopted regression-based frameworks, few explicitly documented their error values, often focusing instead on relative model performance and algorithmic comparison. For instance, Madhuri et al. (2019) [4] employed a suite of regression models—Multiple Linear Regression, Ridge, LASSO, Elastic Net, Gradient Boosting, and AdaBoost—and reported that Gradient Boosting yielded the lowest error rates, outperforming standard linear models. Similarly, Alfiyatin et al. (2017) [3] combined regression with Particle Swarm Optimization (PSO), achieving a Root Mean Square Error (RMSE) of approximately 14.186 (IDR), which, although context-specific, indicates relatively higher prediction error due to regional dataset variability.

In contrast, several recent works such as Zhang (2021) [2] and Chen (2022) [1] demonstrated that traditional multiple linear regression continues to perform competitively when applied to well-structured datasets, particularly when the input features exhibit strong linear correlations with the target variable. These findings are consistent with the present study, where the correlation heatmap confirmed meaningful linear relationships

between predictor variables such as RM, LSTAT, and MEDV. The obtained RMSE value of 4.93 thus represents a robust fit for a baseline linear model, comparable to other empirical findings in the literature.

More advanced algorithms such as XGBoost and LightGBM have been reported to achieve lower RMSE values in large-scale, high-dimensional datasets, as shown in Abdul-Rahman et al. (2021) [5]. However, these models often compromise on interpretability and require greater computational complexity. The focus of this case study, by contrast, was to balance simplicity, transparency, and educational accessibility, ensuring that the results remain reproducible for beginners and undergraduate-level researchers with minimal computational resources.

Additionally, studies by Ja'afar et al. (2021) [9] and Islam & Asami (2009) [10] have emphasized that housing price models must account for not only numerical accuracy but also socioeconomic interpretability and policy relevance. The findings of this case study align with this broader perspective: despite modest prediction errors, the regression coefficients derived from the MLR model offer interpretable insights into how specific housing features—such as the number of rooms, pollution levels, and proximity to amenities—affect market valuation. This interpretability renders linear regression particularly valuable in real estate analytics, where decision-makers often prefer transparent models that reveal underlying factor relationships rather than black-box predictions.

From a practical standpoint, the Boston Housing dataset provides a stable and well-documented foundation for validating statistical methods in property valuation. The observed performance of the MLR model suggests that it can serve as a benchmark model for baseline prediction accuracy. Subsequent research may incorporate feature engineering, regularization methods (Ridge or LASSO), or ensemble learning techniques (Gradient Boosting, Random Forest) to further reduce prediction errors. However, given the case study's educational and interpretive objectives, the achieved RMSE of 4.93 is both statistically reasonable and practically relevant for explaining housing market dynamics in an accessible way.

In summary, while state-of-the-art models demonstrate marginally higher precision, this study validates that Multiple Linear Regression remains a credible, interpretable, and replicable baseline technique in predictive housing analytics. Its low computational cost, explainable coefficients, and stable performance underscore its continued importance in both academic instruction and applied economic modelling.

E. Implications and Future Work

The findings of this study highlight several implications:

1. **Educational Value:**
   This model demonstrates how basic regression concepts can be applied to real-world data analysis, making it ideal for academic teaching, data literacy programs, and beginner-level machine learning courses.

2. **Policy and Investment Insight:**
   By quantifying how key factors like RM, LSTAT, and NOX affect property value, this model provides actionable insights for urban planners, policy developers, and real estate investors seeking to understand housing dynamics.

3. **Model Expansion Potential:**
   Future extensions could incorporate time-series variables, geospatial factors, or economic indicators to improve predictive accuracy. Integrating hybrid machine learning techniques (e.g., Random Forest Regression or Gradient Boosting Machines) could further minimize error while retaining the interpretability of regression-based models.

4. **Case Study Relevance:**
   The results validate the objective of this case study—to demonstrate that even a simple, interpretable MLR model can provide meaningful, data-driven insights into housing markets, bridging the gap between theoretical learning and practical application.

IV. CONCLUSIONS

This study explored the application of Multiple Linear Regression (MLR) for predicting house prices using **the** Boston Housing dataset, aiming to demonstrate how fundamental regression techniques can effectively model complex real-world data in an interpretable manner. The model established a linear relationship between housing attributes and the median house price (MEDV) and was evaluated using performance metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)**.** The obtained values—MSE = 24.29 and RMSE = 4.93,indicate a reasonably accurate predictive capability, particularly for a baseline regression model that emphasizes simplicity and interpretability over computational sophistication.

The analysis confirmed that variables such as the average number of rooms (RM) and percentage of lower-status population (LSTAT) are the most significant predictors of house prices, showing strong positive and negative correlations with MEDV, respectively. The residual analysis and scatter plots validated that the MLR model effectively captures linear trends in the data, while the correlation heatmap revealed multicollinearity between infrastructural variables such as RAD and TAX. These findings are consistent with those reported in earlier research on housing price prediction [1–5], reinforcing that MLR remains a credible benchmark technique even amid the growing popularity of advanced ensemble models.

From a practical standpoint, this case study highlights the value of interpretable machine learning in domains such as real estate analytics, where decision-makers often prioritize transparency over purely numerical optimization. The model's coefficients provide direct insights into how changes in housing attributes influence price, enabling realtors, policymakers, and investors to make evidence-based decisions. Moreover, the study demonstrates that even with a modest dataset and minimal computational resources, students and researchers can build reliable, data-driven tools for forecasting and valuation.

While the results are promising, the model's linear structure inherently limits its capacity to capture nonlinear relationships and spatial dependencies that often characterize housing markets. Future research could address these limitations by integrating Ridge, LASSO**,** or Elastic Net regularization to mitigate multicollinearity, or by applying ensemble methods such as **Gradient** Boosting and Random Forests to enhance predictive accuracy. Additionally, expanding the dataset to include socioeconomic and temporal features could improve generalization and adaptability across regions.

In conclusion, this paper reinforces that Multiple Linear Regression, despite its simplicity, continues to serve as a robust, transparent, and educationally valuable approach for predictive modelling in real estate. It lays a strong foundation for both academic instruction and further exploration of hybrid models, bridging the gap between classical statistics and modern data-driven techniques.

V. REFERENCES

[1] Chen, N. (2022). House price prediction model of Zhaoqing city based on correlation analysis and multiple linear regression analysis. *Wireless Communications and Mobile Computing*, *2022*(1), 9590704.

[2] Zhang, Q. (2021). Housing price prediction based on multiple linear regression. *Scientific Programming*, *2021*(1), 7678931.

[3] Alfiyatin, A. N., Febrita, R. E., Taufiq, H., & Mahmudy, W. F. (2017). Modeling house price prediction using regression analysis and particle swarm optimization case study: Malang, East Java, Indonesia. *International Journal of Advanced Computer Science and Applications*, *8*(10).

[4] Madhuri, C. R., Anuradha, G., & Pujitha, M. V. (2019, March). House price prediction using regression techniques: A comparative study. In *2019 International conference on smart structures and systems (ICSSS)* (pp. 1-5). IEEE.

[5] Abdul-Rahman, S., Zulkifley, N. H., Ismail, I., & Mutalib, S. (2021). Advanced machine learning algorithms for house price prediction: case study in Kuala Lumpur. *International Journal of Advanced Computer Science and Applications*, *12*(12).

[6] Thamarai, M., & Malarvizhi, S. P. (2020). House Price Prediction Modeling Using Machine Learning. *International Journal of Information Engineering & Electronic Business*, *12*(2).

[7] Uzma Taj, Rajlakshmi Sharma, Rida Khan, Sindhu v Prasad, House Price Prediction Using Machine Learning and Neural Networks, International Research Journal of Humanities and Interdisciplinary Studies (www.irjhis.com), ISSN: 2582-8568, Volume: 3, Issue: 6, Year: June 2022, Page No: 101-105

[8] Satish, G. N., Raghavendran, C. V., Rao, M. S., & Srinivasulu, C. (2019). House price prediction using machine learning. *Journal of Innovative Technology and Exploring Engineering*, *8*(9), 717-722.

[9] Ja'afar, N. S., Mohamad, J., & Ismail, S. (2021). Machine learning for property price prediction and price valuation: a systematic literature review. *Planning Malaysia*, *19*.

[10] Islam, K. S., & Asami, Y. (2009, July). Housing market segmentation: A review. In *Review of urban & regional development studies: journal of the applied regional science conference* (Vol. 21, No. 2-3, pp. 93-109). Melbourne, Australia: Blackwell Publishing Asia.