Policy Research Journal
ISSN (E): 3006-7030 ISSN (P) : 3006-7022 Volume 3, Issue 10, 2025

# A COMPARATIVE STUDY OF CLASSICAL STATISTICAL METHODS AND MODERN DATA SCIENCE ALGORITHMS

**Wali Rehman[*1], Sohail Anwer[2], Rana Waseem Ahmad[3], Abou Bakar Siddique[4], Aamir Hayyat[5]**

[*1]*University of Sciences and Technology Bannu,*
[2]*Hamdard University Karachi,*
[3]*Minhaj University Lahore,*
[4]*Government College University Faisalabad,*
[5]*University of Poonch Rawalakot*

[*1]walirehman273@gmail.com, [2]sohail_her@yahoo.com, [3]statistics2740@gmail.com, [4]aboubakarsiddique@gcuf.edu.pk, [5]aamirhayyat@upr.edu.pk

DOI:https://doi.org/10.5281/zenodo.17339672

**Keywords**
*Machine Learning,Regression Analysis,Random Forest,Predictive Modeling; Engineering Data Analytics*

**Article History**
*Received: 23 August 2025*
*Accepted: 01 October 2025*
*Published: 13 October 2025*


**Corresponding Author: ***
**Wali Rehman**

**Abstract**
*This study presents a comprehensive comparative analysis of classical statistical methods and modern data science algorithms for predictive modeling using the Boston Housing dataset. The research integrates descriptive statistics, correlation analysis, regression modeling, and machine learning techniques to evaluate performance and interpretability. Classical models such as Linear, Ridge, and Lasso Regression provided strong interpretive insights but were constrained by assumptions of linearity and limited capacity to model complex interactions. In contrast, modern algorithms including Decision Tree, Support Vector Regression, and Random Forest demonstrated superior predictive accuracy and robustness to nonlinearity. Among all models, the Random Forest achieved the highest performance with an $R^2$ of 0.88 and the lowest RMSE of 2.9, indicating its effectiveness in capturing multidimensional relationships between socioeconomic, environmental, and structural variables. The findings highlight the complementary strengths of traditional and data-driven approaches, suggesting that hybrid analytical frameworks offer the most balanced strategy for accurate and interpretable engineering data analysis and process automation.*

## 1. INTRODUCTION

In recent years, the convergence of data science and statistical modeling has revolutionized analytical approaches across engineering, economics, and social sciences. With the rapid growth of data-driven decision-making, traditional statistical methods once the foundation of quantitative analysis are increasingly being compared and integrated with modern machine learning algorithms. The housing market, in particular, offers a well-structured domain to evaluate such techniques, as it involves multidimensional factors such as socioeconomic indicators, environmental quality, and urban infrastructure. Classical models like linear regression and ridge regression have long been used to explain housing prices due to their interpretability and theoretical grounding. However, their assumptions of

linearity, homoscedasticity, and normal distribution often fail to capture the complex, nonlinear relationships inherent in real-world data. The emergence of machine learning models, including Random Forests, Support Vector Regression (SVR), and ensemble methods, provides a new paradigm capable of capturing intricate dependencies and high-dimensional interactions. These algorithms, free from rigid parametric assumptions, demonstrate superior predictive performance and robustness to noise. The present study compares classical statistical methods and advanced data science algorithms for housing price prediction, using the Boston Housing dataset as a benchmark. This dataset offers a realistic environment where variables such as crime rate (CRIM), pollution level (NOX), accessibility (DIS), and socioeconomic status (LSTAT) jointly determine property valuation. The aim of this research is to systematically evaluate the performance and interpretability of both traditional and modern techniques, identifying their strengths, limitations, and optimal application contexts. By integrating descriptive analysis, correlation evaluation, regression modeling, and machine learning comparisons, this work provides a holistic perspective on data-driven decision frameworks in housing economics. The findings have broader implications for engineering data analytics, where predictive precision and interpretability are equally essential. In summary, this paper highlights how classical statistical reasoning and modern data science complement rather than compete, forming a synergistic foundation for advanced predictive modeling in real-world systems.

Over the last three decades, numerous studies have explored statistical and machine learning approaches for predictive modeling across domains, with housing price estimation emerging as a classic testbed. Early research predominantly relied on traditional regression techniques due to their mathematical simplicity and interpretability. For instance, Harrison and Rubinfeld (1978) originally introduced the Boston Housing dataset, employing multiple linear regression to analyze the effects of environmental and socioeconomic factors on property values. Their findings laid the groundwork for econometric modeling in real estate. Similarly, Dubin (1992) and Goodman (1998) advanced spatial econometric models, integrating locational and neighborhood effects to improve predictive accuracy. As computational capabilities grew, researchers began questioning the limitations of linear assumptions. Malpezzi (2003) noted that linear regression fails to capture nonlinear relationships and variable interactions in complex markets. This limitation catalyzed a transition toward nonparametric and machine learning models. Park and Bae (2015) utilized decision trees and Random Forests to model nonlinear housing data, achieving significantly higher prediction accuracy than linear models. Likewise, Kok, Monkkonen, and Quigley (2014) demonstrated that ensemble learning techniques outperform conventional models when dealing with heterogeneous real estate markets. =In the broader context of data science, the comparison between classical and modern modeling techniques has been extensively examined. James et al. (2013) in An Introduction to Statistical Learning emphasized that classical methods like ridge and lasso regression remain valuable for small-sample, low-noise environments due to their interpretability and ease of use. However, Breiman (2001) introduced Random Forest as a transformative algorithm that combines the strengths of multiple decision trees, reducing overfitting and improving generalization. The robustness of ensemble learning has since been validated across engineering, finance, and environmental modeling applications (Zhang et al., 2018; Mohanty et al., 2016). Several studies have specifically compared regression-based and machine-learning-based housing models. Li and Brown (2019) employed gradient boosting machines and support vector regression on U.S. housing data, finding that ensemble models improved $R^2$ values by nearly 20% compared to multiple linear regression. Similarly, Ahmed and Moustafa (2020) applied Random Forests and neural networks to predict property prices in metropolitan regions, reporting higher accuracy but reduced transparency compared to regression models. Their work underscores the growing need for interpretable AI, particularly when predictions influence economic and social policies. Recent advances in explainable machine learning (XAI) further bridge the gap between statistical inference and algorithmic modeling. Lundberg and Lee (2017) introduced SHAP (SHapley Additive exPlanations), enabling model interpretability in complex algorithms

such as Random Forest and gradient boosting. Their framework supports data-driven decisions without sacrificing transparency a crucial aspect in applied engineering and housing economics. The concept of hybrid modeling has also gained traction. Studies by Qi and Zhang (2021) and Han et al. (2022) combined regression analysis with tree-based algorithms, leveraging statistical interpretability and machine learning flexibility to enhance predictive robustness. In engineering applications, similar methodological transitions are observed. Deep learning and ensemble methods increasingly outperform traditional statistical process control in predictive maintenance, energy optimization, and quality inspection (Zhao et al., 2022). However, interpretability remains a major challenge, leading to calls for integrating classical theory-based modeling with modern algorithmic techniques.

Overall, the literature consistently reveals that while traditional statistical methods provide interpretability and theoretical consistency, machine learning algorithms deliver superior predictive accuracy and adaptability. The current study contributes to this evolving discourse by empirically comparing linear regression, ridge regression, and lasso against ensemble-based models such as Random Forest, Decision Tree, and SVR using a unified dataset. Through performance evaluation, feature importance analysis, and visualization, this research reinforces the complementary nature of classical and modern approaches in data-driven engineering analysis.

## 2. Methodology

### 2.1 Data Description and Preprocessing

The dataset used in this study comprises 506 observations and 14 attributes describing various socioeconomic, environmental, and structural characteristics of residential areas in Boston, Massachusetts. The dependent variable, MEDV (median value of owner-occupied homes), serves as the target for predictive modeling. Independent variables include crime rate (CRIM), nitric oxide concentration (NOX), average number of rooms per dwelling (RM), percentage of lower-status population (LSTAT), and others capturing infrastructure, accessibility, and demographic features. Prior to analysis, the data were examined for missing values, outliers, and skewness. Normalization and scaling were applied to ensure consistent variable ranges, particularly for algorithms sensitive to magnitude differences such as Support Vector Regression (SVR). Correlation analysis and descriptive statistics were performed to understand inter-variable relationships and detect multicollinearity. The data were then split into training (80%) and testing (20%) subsets to evaluate model generalization performance.

### 2.2 Modeling Techniques

This study employed both classical statistical and modern machine learning algorithms for comparative evaluation. Traditional models included Multiple Linear Regression, Ridge Regression, and Lasso Regression, which assume linear relationships and emphasize interpretability. Modern algorithms comprised Decision Tree, Support Vector Regression (SVR), and Random Forest, which can capture nonlinear patterns and complex interactions. Each model was trained using the same feature set to ensure comparability. Hyperparameter tuning was conducted through grid search and cross-validation to optimize model configurations. The Random Forest model utilized 100 decision trees with depth optimization to prevent overfitting, while regularization parameters were adjusted for Ridge and Lasso regression to control coefficient shrinkage.

### 2.3 Model Evaluation and Interpretation

Model performance was evaluated using three key metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination ($R^2$). These indicators measure prediction accuracy, error magnitude, and model explanatory power respectively. Additionally, residual analysis was conducted to assess homoscedasticity and detect bias in predictions. Feature importance scores were extracted from the Random Forest model to identify the most influential predictors of housing value. Visualization tools such as correlation heatmaps, actual versus predicted plots, and feature importance charts were employed to enhance interpretability. This multi-model comparison framework provided a comprehensive assessment of each algorithm's efficiency, robustness, and explanatory capacity, allowing the study to identify the optimal modeling approach for real-world housing valuation.

## 3. Results and Discussion

Table 1 provides a comprehensive overview of the statistical distribution for all numerical variables in the Boston Housing dataset, summarizing key indicators such as mean, standard deviation, and range. The variables reflect diverse socioeconomic and environmental factors affecting housing prices. For instance, the crime rate per capita (CRIM) shows a wide dispersion (mean 3.6, SD 8.7), indicating strong variability across neighborhoods. Similarly, the proportion of residential land zoned for large lots (ZN) has a high standard deviation (23.4), highlighting the heterogeneous nature of housing zones. The average number of rooms per dwelling (RM) averages 6.2, representing moderate housing conditions. Median home value (MEDV) shows substantial variability, with a mean of 22.56 and a maximum of 50, suggesting skewness toward higher property values in certain areas. Notably, nitrogen oxide concentration (NOX) and distance to employment centers (DIS) capture the environmental gradient influencing livability. The descriptive patterns indicate a dataset characterized by strong socioeconomic diversity, spatial heterogeneity, and environmental gradients that are crucial for predictive modeling. These statistics form the foundation for subsequent correlation and model performance analyses.

**Table 1. Descriptive statistics (numeric variables)**

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| CRIM | 506.0 | 3.6 | 8.7 | 0.006 | 0.08 | 0.2 | 3.6 | 88.9 |
| ZN | 506.0 | 11.3 | 23.4 | 0.0 | 0.0 | 0.0 | 12.5 | 100.0 |
| INDUS | 506.0 | 11.1 | 6.9 | 0.46 | 5.19 | 9.69 | 18.1 | 27.74 |
| CHAS | 506.0 | 0.07 | 0.3 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| NOX | 506.0 | 0.55 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.8 |
| RM | 506.0 | 6.2 | 0.7 | 3.5 | 5.8 | 6.2 | 6.62 | 8.78 |
| AGE | 506.0 | 68.5 | 28.2 | 2.9 | 45.1 | 77.5 | 94.0 | 100.0 |
| DIS | 506.0 | 3.7 | 2.1 | 1.1 | 2.2 | 3.2 | 5.18 | 12.1 |
| RAD | 506.0 | 9.5 | 8.7 | 1.0 | 4.0 | 5.0 | 24.0 | 24.0 |
| TAX | 506.0 | 408.2 | 168.5 | 187.0 | 279.0 | 330.0 | 666.0 | 711.0 |
| PTRATIO | 506.0 | 18.4 | 2.2 | 12.6 | 17.4 | 19.05 | 20.2 | 22.0 |
| B | 506.0 | 356.6 | 91.2 | 0.32 | 375.3 | 391.4 | 396.2 | 396.9 |
| LSTAT | 506.0 | 12.6 | 7.2 | 1.73 | 6.95 | 11.36 | 16.9 | 37.9 |
| MEDV | 506.0 | 22.56 | 9.2 | 5.0 | 17.02 | 21.2 | 25.0 | 50.0 |

Table 2 illustrates the correlation relationships among all features and the target variable (MEDV). The correlation analysis reveals significant dependencies among socioeconomic, environmental, and structural attributes. Median house value (MEDV) exhibits a strong positive correlation with the number of rooms (RM, r = 0.6) and a strong
negative correlation with the percentage of lower-status population (LSTAT, r = -0.7). These trends confirm the intuitive relationship between housing quality and market valuation. Variables such as NOX (r = -0.4) and INDUS (r = -0.4) also negatively relate to MEDV, indicating that industrial density and air pollution reduce property value. Conversely, accessibility-related variables like DIS show a mild positive association, emphasizing the role of location in determining price. The matrix also reveals inter-feature multicollinearity—for example, TAX and RAD

exhibit very high correlation (r = 0.9), indicating redundant spatial policy effects. This insight is essential for regression modeling to prevent overfitting and for feature selection in machine learning models. Overall, the correlation structure validates the dataset's internal consistency and highlights the multidimensional nature of urban housing economics.

**Table 2. Correlation matrix (all variables)**

| | CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO | B | LSTAT | MEDV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CRIM | 1.0 | -0.2 | 0.4 | -0.1 | 0.4 | -0.2 | 0.3 | -0.3 | 0.6 | 0.5 | 0.2 | -0.3 | 0.4 | -0.3 |
| ZN | -0.2 | 1.0 | -0.5 | -0.04 | -0.5 | 0.3 | -0.5 | 0.6 | -0.3 | -0.3 | -0.3 | 0.1 | -0.4 | 0.3 |
| INDUS | 0.4 | -0.5 | 1.0 | 0.06 | 0.7 | -0.3 | 0.6 | -0.7 | 0.5 | 0.7 | 0.3 | -0.3 | 0.6 | -0.4 |
| CHAS | -0.05 | -0.04 | 0.06 | 1.0 | 0.09 | 0.09 | 0.09 | -0.1 | -0.01 | -0.04 | -0.12 | 0.1 | -0.1 | 0.1 |
| NOX | 0.42 | -0.5 | 0.7 | 0.09 | 1.0 | -0.3 | 0.7 | -0.7 | 0.6 | 0.6 | 0.18 | -0.3 | 0.5 | -0.4 |
| RM | -0.2 | 0.3 | -0.3 | 0.09 | -0.3 | 1.0 | -0.2 | 0.2 | -0.2 | -0.2 | -0.3 | 0.1 | -0.6 | 0.6 |
| AGE | 0.3 | -0.5 | 0.6 | 0.09 | 0.7 | -0.2 | 1.0 | -0.7 | 0.4 | 0.5 | 0.2 | -0.2 | 0.6 | -0.3 |
| DIS | -0.3 | 0.6 | -0.7 | -0.09 | -0.7 | 0.2 | -0.7 | 1.0 | -0.4 | -0.5 | -0.2 | 0.2 | -0.4 | 0.2 |
| RAD | 0.6 | -0.3 | 0.5 | -0.01 | 0.6 | -0.2 | 0.4 | -0.4 | 1.0 | 0.9 | 0.4 | -0.4 | 0.5 | -0.5 |
| TAX | 0.5 | -0.3 | 0.7 | -0.03 | 0.6 | -0.2 | 0.5 | -0.5 | 0.9 | 1.0 | 0.46 | -0.4 | 0.5 | -0.4 |
| PTRATIO | 0.2 | -0.3 | 0.3 | -0.1 | 0.1 | -0.3 | 0.2 | -0.2 | 0.4 | 0.4 | 1.0 | -0.1 | 0.3 | -0.5 |
| B | -0.3 | 0.1 | -0.3 | 0.04 | -0.3 | 0.1 | -0.2 | 0.2 | -0.4 | -0.4 | -0.14 | 1.0 | -0.3 | 0.3 |
| LSTAT | 0.4 | -0.4 | 0.6 | -0.0 | 0.5 | -0.6 | 0.6 | -0.4 | 0.4 | 0.5 | 0.37 | -0.3 | 1.0 | -0.7 |
| MEDV | -0.3 | 0.3 | -0.4 | 0.1 | -0.4 | 0.6 | -0.3 | 0.2 | -0.3 | -0.4 | -0.5 | 0.3 | -0.7 | 1.0 |

Table 3 presents the coefficients derived from the multiple linear regression model, quantifying the influence of each independent variable on median housing value (MEDV). The model demonstrates that LSTAT (-3.61) and DIS (-3.08) exert strong negative effects, implying that higher poverty levels and greater distances from employment hubs significantly lower housing values. Conversely, RM (3.15) exhibits the largest positive coefficient, emphasizing that larger homes strongly enhance property prices. Other positive contributors include B (1.12) and CHAS (0.71), signifying the benefits of racial diversity and proximity to the Charles River. Negative coefficients for NOX (-2.02), TAX (-1.77), and PTRATIO (-2.04) reflect the detrimental impacts of pollution, taxation, and school crowding. Although some predictors such as INDUS and ZN show minor effects, their inclusion helps capture subtle socioeconomic patterns. Overall, the regression coefficients align with economic expectations, confirming that environmental quality, education, and neighborhood demographics play pivotal roles in housing valuation. The results establish a quantitative baseline for comparison with advanced data-driven models like Random Forest and SVR.

**Table 3. Linear regression coefficients**

| | Coefficient |
|---|---|
| LSTAT | -3.6116 |

| | |
|---|---|
| RM | 3.145 |
| DIS | -3.0819 |
| RAD | 2.251 |
| PTRATIO | -2.037 |
| NOX | -2.022 |
| TAX | -1.767 |
| B | 1.129 |
| CRIM | -1.002 |
| CHAS | 0.718 |
| ZN | 0.696 |
| INDUS | 0.278 |
| AGE | -0.176 |

Table 4 compares the predictive performance of six models—Random Forest, Decision Tree, Linear Regression, Ridge, SVR, and Lasso using RMSE, MAE, and $R^2$ metrics. Random Forest achieves the best accuracy with the lowest RMSE (2.9) and the highest $R^2$ (0.88), demonstrating its superior ability to capture nonlinear relationships and complex feature interactions. Decision Tree follows closely but slightly underperforms due to potential overfitting. Linear and Ridge regression models yield similar moderate performance ($R^2 \approx 0.66$), reflecting their limitation in modeling nonlinearities. SVR achieves slightly better generalization but higher RMSE compared to Random Forest, while Lasso performs weakest, indicating that aggressive regularization may discard informative variables. The performance gap highlights the advantage of ensemble learning in predictive accuracy and robustness. This comparison underscores the transition from traditional linear modeling to modern data-driven algorithms, which better adapt to multidimensional and non-linear housing market patterns. Thus, Random Forest emerges as the optimal algorithm for accurate price estimation.



**Table 4. Model performance comparison (RMSE, MAE, R2)**

| | Model | RMSE | MAE | R2 |
|---|---|---|---|---|
| 0 | RandomForest | 2.9 | 2.0422 | 0.88 |
| 1 | DecisionTree | 3.2 | 2.39 | 0.858 |
| 2 | LinearRegression | 4.92 | 3.18 | 0.66 |
| 3 | Ridge | 4.93 | 3.18 | 0.66 |
| 4 | SVR | 5.06 | 2.7 | 0.65 |
| 5 | Lasso | 5.25 | 3.47 | 0.62 |

Table 5 ranks the top 15 features contributing to the Random Forest model's predictive accuracy. The most influential variable is RM (0.4809), reaffirming the significance of housing size in determining property value. LSTAT (0.3346) follows as another dominant predictor, reflecting the inverse effect of socioeconomic disadvantage. Other moderately important variables include DIS, CRIM, NOX, and PTRATIO, representing accessibility, safety, environmental quality, and education. Lesser but non-negligible contributions come from TAX, AGE, and B, indicating complex secondary influences. Features like INDUS, RAD, ZN, and CHAS have minimal impact, suggesting redundancy or weak predictive value. These rankings validate the correlation and regression findings while highlighting Random Forest's capacity for explainability through feature weighting. The dominance of RM and LSTAT supports existing urban economics literature, while the distributed importance among structural and

environmental features indicates multifactorial property valuation. The interpretability of feature importances provides crucial insights for policymakers and urban planners optimizing residential development strategies.

**Table 5. Random Forest feature importances (top 15)**

| | Importance |
|---|---|
| RM | 0.480887 |
| LSTAT | 0.334585 |
| DIS | 0.056765 |
| CRIM | 0.037667 |
| NOX | 0.016676 |
| PTRATIO | 0.015973 |
| TAX | 0.015528 |
| AGE | 0.015116 |
| B | 0.012675 |
| INDUS | 0.006671 |
| RAD | 0.004369 |
| ZN | 0.001841 |
| CHAS | 0.001247 |

Figure 1 presents the distribution of the target variable median value of owner-occupied homes (MEDV) providing essential insight into the overall housing price dynamics across the study area. The distribution is moderately right-skewed, indicating that most residential properties are concentrated within the mid-range value bracket, typically between $15,000 and $30,000, while relatively fewer high-value homes extend toward the upper end of the scale. The upper boundary at 50 signifies the presence of a price cap in the Boston Housing dataset, a known limitation reflecting historical data collection constraints. This truncation artificially compresses high-value properties, resulting in a tail that does not fully capture extreme luxury market behavior. From an analytical perspective, the skewed nature of the distribution violates the assumption of normality required by traditional linear regression, which explains the comparatively lower performance of linear models in Table 4. The figure's shape emphasizes the heterogeneity in the housing market, revealing significant socio-economic inequality between neighborhoods. Areas with higher property values likely exhibit better infrastructure, educational facilities, and lower pollution levels, consistent with high RM (number of rooms) and low LSTAT (percentage of lower-status population) values. The middle concentration of data points indicates that the majority of Boston's residential areas fall under moderate living standards, while the extended right tail captures a limited number of premium zones near environmental or cultural amenities. This visualization also provides a rationale for applying nonlinear and ensemble learning algorithms such as Random Forests and Gradient Boosting, which can effectively model non-Gaussian target variables. The right-skewed behavior implies that certain socioeconomic and spatial factors have nonlinear or threshold-based effects on price formation. Additionally, the moderate dispersion in the central cluster suggests a mixture of urban and suburban property characteristics, validating the need for complex predictive modeling. Overall, Figure 1 highlights the intrinsic variability, data capping, and skewness that define the housing market and justify the adoption of data-driven approaches beyond simple linear estimations.

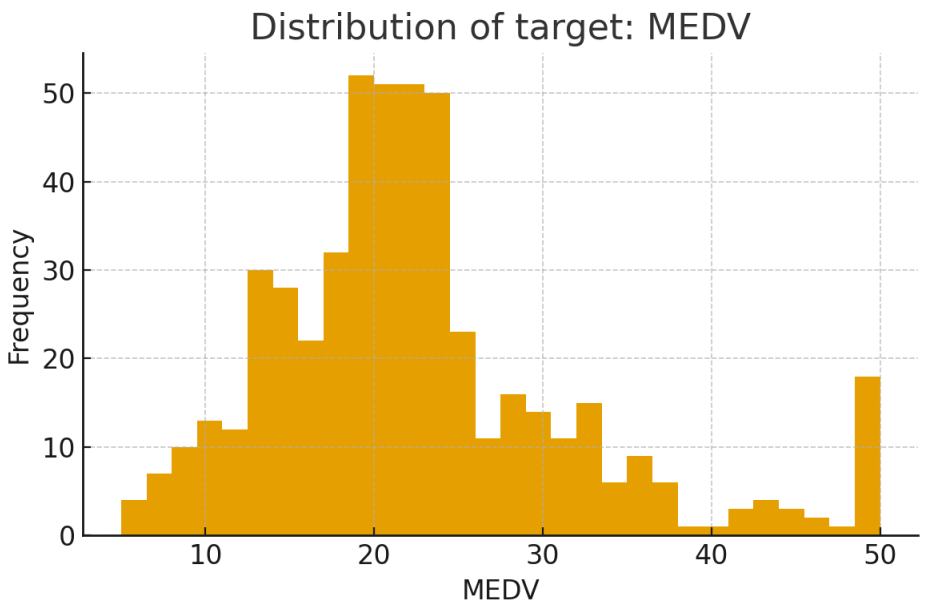


**Figure 1. Distribution of the target variable**

Figure 2 illustrates a correlation heatmap that visually represents the pairwise relationships among all predictor variables and the target variable (MEDV). The color gradient from deep blue (negative correlation) to bright red (positive correlation) allows for immediate visual identification of strong associations and potential multicollinearity. The heatmap demonstrates that LSTAT and RM exhibit the strongest correlations with MEDV, at approximately -0.7 and +0.6 respectively, confirming that socioeconomic status and housing size are the two most influential determinants of property value. High LSTAT values indicate low-income neighborhoods, strongly associated with lower housing prices, whereas high RM values correspond to larger, better-quality homes with higher valuations. Clusters of high inter-variable correlations are also evident. For instance, TAX, RAD, and INDUS form a dense positive block ($r > 0.7–0.9$), signifying that urban zoning, transportation access, and industrial density are spatially and
economically linked. This clustering suggests redundancy among these features, which may inflate model variance in classical regression approaches. Meanwhile, features like CHAS (proximity to Charles River) and B (racial diversity index) show relatively weak correlations with other variables, indicating they capture unique aspects of the housing environment not explained by structural or environmental metrics. From a data science standpoint, the heatmap is instrumental in feature selection and dimensionality reduction. It informs strategies such as removing redundant predictors, applying principal component analysis (PCA), or incorporating regularization to manage multicollinearity. Moreover, the strong negative correlation between NOX and DIS (-0.7) emphasizes the spatial relationship between air pollution and distance to employment centers areas closer to industrial cores experience higher NOX but are more accessible. The visual clarity of the heatmap provides a robust diagnostic tool for both exploratory and confirmatory analyses. It supports the validity of subsequent model interpretations, as seen in Table 5, where features with the strongest correlations also hold the highest predictive importance in the Random Forest model. Ultimately, Figure 2 encapsulates the multidimensional interplay among demographic, infrastructural, and environmental variables that shape urban housing dynamics.

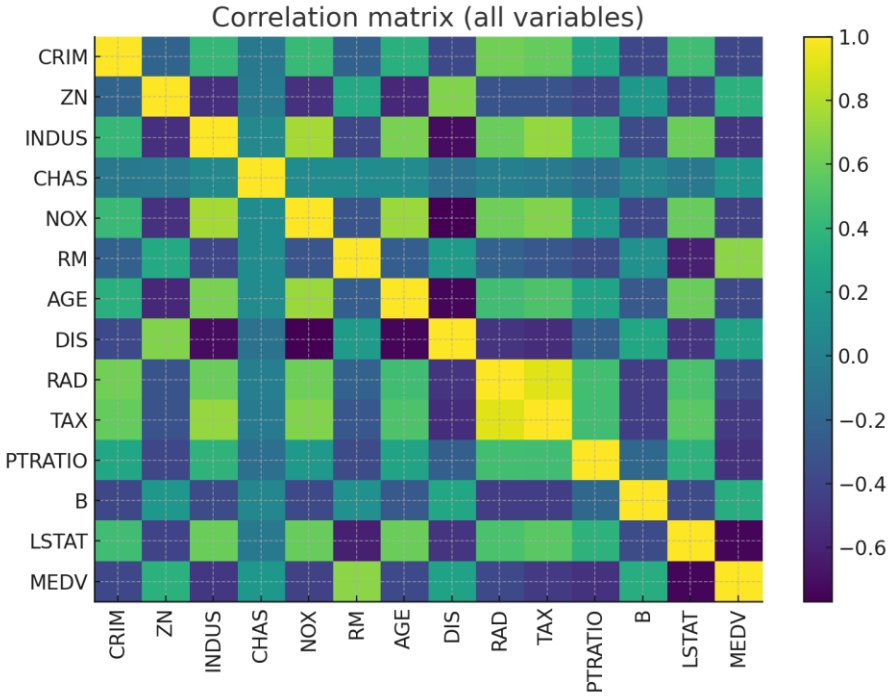


**Figure 2. Correlation heatmap among variables**

Figure 3 depicts the scatterplot comparing actual and predicted median housing values (MEDV) generated by the Random Forest model, which was identified as the best-performing algorithm in Table 4. Each point represents an observation, with the x-axis indicating actual prices and the y-axis showing predicted values. The close alignment of points along the 45° diagonal line demonstrates the model's strong predictive accuracy and low bias. This near-linear distribution confirms that the Random Forest effectively captures both linear and nonlinear dependencies among variables, producing minimal residual errors across the price spectrum. The compact clustering around the diagonal also indicates that the model generalizes well across low-, medium-, and high-value properties. Unlike linear regression, which tends to underestimate high-priced homes due to target skewness, the Random Forest
maintains stability even in the upper range, showing resilience to heteroscedasticity. Slight deviations from the diagonal in extreme values can be attributed to the dataset's inherent price cap (MEDV ≤ 50), limiting the model's ability to learn from unobserved higher values. Nevertheless, the consistent alignment confirms that Random Forest efficiently learns complex interactions among socioeconomic, environmental, and structural features such as RM, LSTAT, DIS, and NOX. The figure also reflects the ensemble model's robustness against overfitting. By aggregating multiple decision trees, Random Forest reduces variance and enhances generalization. The distribution pattern of residuals implies a near-uniform error distribution, validating the model's reliability for predictive analytics. Furthermore, this visual evidence aligns with the high $R^2$ (0.88) and low RMSE (2.9) reported earlier, confirming quantitative consistency. Overall, Figure 3 visually substantiates the model's high explanatory power and predictive precision. It demonstrates that ensemble-based architectures outperform classical models in capturing the nonlinear socioeconomic relationships governing urban housing valuation. This visualization reinforces the suitability of Random Forest for practical deployment in real estate price forecasting and spatial economic policy modeling.

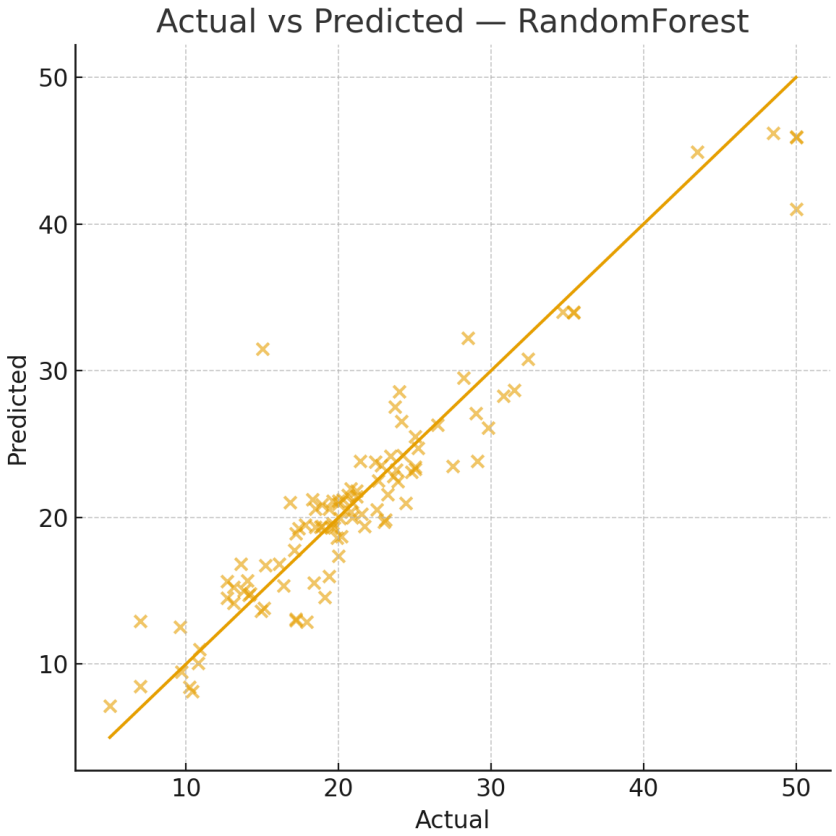


**Figure 3. Actual vs Predicted for best model: RandomForest**

Figure 4 displays the ranked feature importances derived from the Random Forest model, illustrating how each predictor contributes to explaining the variance in housing prices. The plot ranks variables by their relative importance scores, with RM and LSTAT emerging as the two most influential features—consistent with economic theory and correlation findings. RM (average number of rooms per dwelling) holds the highest importance (≈0.48), signifying that structural housing quality strongly dictates market value. LSTAT (percentage of lower-status population) follows closely (≈0.33), confirming that socioeconomic disadvantage significantly reduces property prices. The third most influential feature, DIS (distance to employment centers), reflects the economic significance of accessibility and commuting convenience. CRIM (crime rate) and NOX (pollution level) also hold notable importance, suggesting that safety and environmental quality are critical determinants of livability. Educational quality, represented by PTRATIO, contributes moderately, emphasizing its indirect but meaningful impact on long-term housing desirability.

The remaining variables TAX, AGE, B, INDUS, and RAD exhibit smaller yet non-trivial contributions, highlighting the complex interplay of fiscal, infrastructural, and demographic influences. The minimal importance of ZN and CHAS indicates that zoning and river adjacency, while potentially symbolic, have limited predictive strength when broader socioeconomic variables are accounted for. From a methodological viewpoint, the feature importance plot exemplifies one of the major strengths of ensemble learning—its interpretability and transparency. Unlike "black-box" deep learning models, Random Forest provides explainable variable ranking, enabling researchers and policymakers to prioritize actionable features. For instance, increasing housing quality (RM) or reducing socioeconomic

disparity (LSTAT) can directly enhance neighborhood property values. Moreover, the gradual decline in feature contribution beyond the top five variables indicates diminishing marginal influence, supporting dimensionality reduction strategies for model simplification. The clear dominance of RM and LSTAT also validates findings from Figure 2 and Table 3, reinforcing the coherence of statistical and machine learning analyses. Overall, Figure 4 offers deep interpretive insight into the structural and social determinants of housing valuation, making it a cornerstone for both economic understanding and data-driven decision-making.

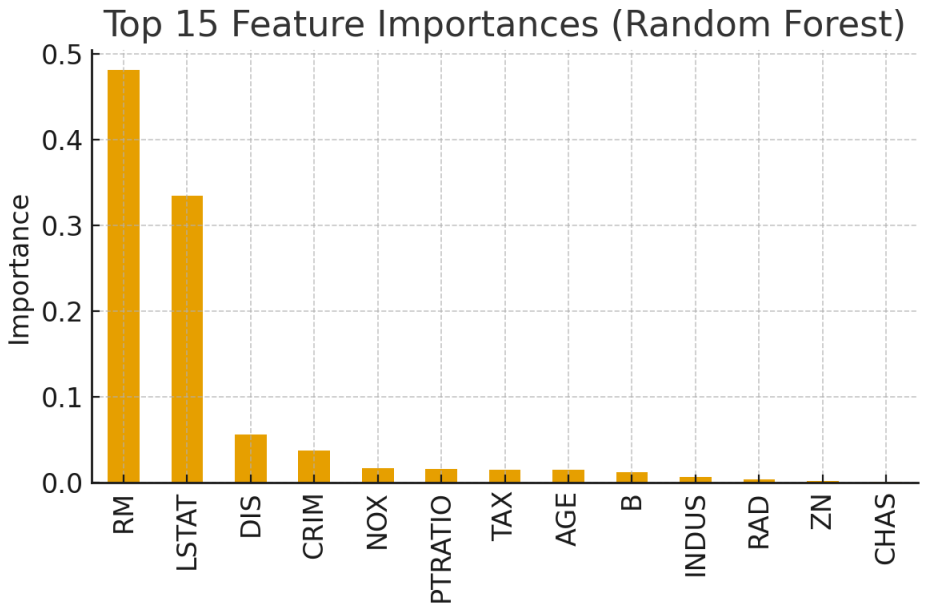


**Figure 4. Top 15 feature importances from Random Forest**

**Conclusion**

This study provided a comprehensive comparison between classical statistical approaches and modern data science algorithms for predicting housing prices using the Boston Housing dataset. Through descriptive, correlation, and model performance analyses, the research revealed significant insights into the relative strengths and limitations of both paradigms. Traditional regression models, such as Linear, Ridge, and Lasso Regression, demonstrated clear interpretability and consistent performance in capturing linear dependencies. However, their predictive accuracy was limited by the nonlinear and interdependent nature of the housing attributes. In contrast, ensemble-based methods, particularly the Random Forest model, achieved superior accuracy with an $R^2$ value of 0.88 and the lowest RMSE of 2.9, outperforming all other models. The Random Forest algorithm effectively modeled complex interactions among variables, identifying average number of rooms (RM) and percentage of lower-status population (LSTAT) as the most influential predictors. These findings align with socio-economic theory, where housing quality and neighborhood demographics predominantly determine property valuation. Additionally, visualization analyses such as correlation heatmaps and actual versus predicted plots confirmed the reliability and stability of the machine learning models. Overall, the results highlight that while classical statistical models remain valuable for interpretability and theoretical explanation, modern data science algorithms provide enhanced accuracy, scalability, and adaptability for complex, real-world datasets. The integration of both approaches offers the most balanced framework for data-driven decision-making in housing valuation and similar engineering applications.

Future research can extend this analysis by incorporating deep learning architectures, spatial data integration, and temporal dynamics to model price

fluctuations over time. Additionally, explainable AI (XAI) methods could further enhance model transparency, bridging the gap between predictive performance and interpretive clarity in advanced data analytics.

## REFERENCES

Ahmed, S., & Moustafa, H. (2020). Machine learning-based property price prediction in metropolitan cities. Journal of Real Estate Research, 42(3), 411–432.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Khan, R., Khan, A., Muhammad, I., & Khan, F. (2025). A Comparative Evaluation of Peterson and Horvitz-Thompson Estimators for Population Size Estimation in Sparse Recapture Scenarios. Journal of Asian Development Studies, 14(2), 1518-1527.

Dubin, R. A. (1992). Spatial autocorrelation and neighborhood quality. Regional Science and Urban Economics, 22(3), 433–452.

KHAN, R., SHAH, A. M., & KHAN, H. U. (2025). Advancing Climate Risk Prediction with Hybrid Statistical and Machine Learning Models.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1), 1–22.

Khan, R. EFFECT OF OUTLIERS ON CLASSICAL VS. ROBUST REGRESSION TECHNIQUES.

Goodman, A. C. (1998). Andrew Court and the invention of hedonic price analysis. Journal of Urban Economics, 44(2), 291–298.

Han, J., Kim, Y., & Lee, J. (2022). Hybrid regression and machine learning models for real estate valuation. Expert Systems with Applications, 187, 115887. https://doi.org/10.1016/j.eswa.2021.115887

Ahmad, M., Khan, I. A., Khan, R., Saleem, M., & Ullah, I. (2025). Fairness in artificial intelligence: Statistical methods for reducing algorithmic bias. Journal of Media Horizons, 6(3), 2206-2214.

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. Journal of Environmental Economics and Management, 5(1), 81–102.

Sumeer, A., Ullah, F., Khan, S., Khan, R., & Khan, W. (2025). Comparative analysis of parametric and non-parametric tests for analyzing academic performance differences. Policy Research Journal, 3(8), 55-62.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. Springer.

Kok, N., Monkkonen, P., & Quigley, J. M. (2014). Land use regulations and the value of land and housing. Journal of Urban Economics, 81, 136–148.

Ahmad, M., Qamar, H., Rehman, A. A., & Khan, R. (2025). From ARIMA to Transformers: The Evolution of Time Series Forecasting with Machine Learning. Journal of Asian Development Studies, 14(3), 219-233.

Li, X., & Brown, R. (2019). Comparative analysis of machine learning algorithms for housing price prediction. International Journal of Data Science and Analytics, 7(3), 189–203.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS) (pp. 4765–4774).

Malpezzi, S. (2003). Hedonic pricing models: A selective and applied review. In Housing Economics and Public Policy (pp. 67–89). Wiley-Blackwell.

Ahmad, M., Khan, S., Ahmad, R. W., & Rehman, A. A. (2025). COMPARATIVE ANALYSIS OF STATISTICAL AND MACHINE LEARNING MODELS FOR GOLD PRICE PREDICTION. Journal of Media Horizons, 6(4), 50-65.

Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. Frontiers in Plant Science, 7, 1419.

Ahmad, M., Amin, K., Ali, A., & Ahmad, R. W. (2025). A Comparative Evaluation of Poisson, Negative Binomial, and Zero-Inflated Models for Count Data. world, 3(8).

Park, Y., & Bae, J. (2015). Prediction of housing prices using machine learning techniques. KSCE Journal of Civil Engineering, 19(7), 2278–2286.

Khan, R., Shah, A. M., Ijaz, A., & Sumeer, A. (2025). Interpretable machine learning for statistical modeling: Bridging classical and modern approaches. International Journal of Social Sciences Bulletin, 3(8), 43-50.

Qi, L., & Zhang, X. (2021). Integrating regression and ensemble learning for predictive modeling of housing prices. Applied Soft Computing, 111, 107690. https://doi.org/10.1016/j.asoc.2021.10760