Fuzzy Systems and Soft Computing
ISSN : 1819-4362

ANALYSIS OF REAL ESTATE PRICE PREDICTION USING REGRESSION MODELS

Dr. M. P. Sukassini Assistant Professor, Department of MCA, DDGD Vaishnav College, Arumbakkam, Chennai, India Email: sukassini.dgvc@gmail.com
Dwarakesh S, Student, Department of MCA, DDGD Vaishnav College, Arumbakkam, Chennai, India. Email :dwarakeshsrinivasan2002@gmail.com
Abhishek. M Student, Department of MCA, DDGD Vaishnav College, Arumbakkam, Chennai, India. Email: venkatakrishnanabhi2001@gmail.com

ABSTRACT:
Real estate is the most popular and one of the profitable businesses in many countries. People looking to buy a new home tend to be more conservative with their budgets and marketing strategies. Machine learning algorithms help to predict the sales price of the house depending upon several factors. Machine learning has become integral in various fields, including image detection, spam recognition, speech command processing, product recommendation, and medical diagnosis. It enhances security alerts, public safety, and medical advancements while providing better customer service and safer automobile systems. This paper focuses on predicting future housing prices using machine learning algorithms. By comparing and exploring various prediction methods, the research work identifies the most effective model for this task. The dataset used in this work is housing data taken from the Kaggle repository. Initially the dataset was preprocessed using the Standard methods. Several machine learning algorithms such as Linear Regression, Polynomial Regression, Ridge Regression, Elastic Net Regression, and Gradient Boosting were used to predict the prices. The study aims to develop a reliable and accurate housing price prediction model that can provide valuable insights for real estate agents. This research highlights the importance of advanced machine learning techniques in real estate analytics and contributes to the development of robust prediction models for housing prices. The performance of the models was analyzed using Regression Metrics.

Keywords:
Multiple Linear Regression (MLR), Polynomial Regression(PR), Ridge Regression(RR), Elastic Net Regression (ENR), Gradient Boosting Regressor(GBR), Housing Price Prediction.

I. INTRODUCTION

Machine learning algorithms are gaining popularity in various fields, similar to how animals like rats learn to avoid toxic baits through experience. This paper uses a comprehensive dataset of housing information from Chennai, India. It applies regression models like Multiple Linear Regression, Polynomial Regression, Ridge Regression, Elastic Net Regression, and Gradient Boosting to predict housing prices. The study highlights the importance of choosing appropriate machine learning models and assessment metrics, with Gradient Boosting showing advanced performance. The study also discusses the application of machine learning systems in analyzing large and complex datasets, particularly in fields like astronomy, medical diagnostics, climate prediction, genomics, and e-trade. By leveraging the big memory capability and processing speed of modern computers, machine learning opens new horizons in data analysis and prediction. The study demonstrates the effectiveness of various machine learning models in predicting housing prices and provides insights into their practical implementation for real estate analytics. Accurate housing price predictions can help consumers, sellers, real estate marketers, and policymakers make informed decisions, such as identifying undervalued homes, charging competitively, and implementing effective guidelines. Integrating machine learning into real estate analytics can improve efficiency, transparency, and accuracy in pricing models

The paper is organized as follows. Section II discusses the related work of previous papers in the same topic.Section III The procedures and materials employed in this investigation. The experimental results of the work carried out for the sales price are explained in section IV. Finally, section V summarizes the research findings.

II. LITERATURE SURVEY

Predicting house prices is challenging and requires various systems for accurate and accurate analysis [1]. Housing prices are influenced by factors like bedrooms, bathrooms, crime rate, land proportion, property tax rate, and pupil-teacher ratio. The location also affects prices. Market analysis is crucial in valuation processes, determining market characteristics and pricing features. However, modern technologies and easy access to large databases have led to a shift in property valuation, particularly when using traditional data selection methods [2]. Many researchers have contributed various algorithms in predicting the prices and the related works were discussed below.

Dr. Sonia Juneja and Neha Choudhary[3] analyzed the real estate market dataset, which went through preprocessing to clean the data and handle missing values. Using multiple regression analysis, they predicted that 77% of the results were accurate. Hemalata Sharma et al [4] focused on analyzing the housing dataset of Ames City in Iowa, USA, and predicted that 93% of the results were accurate using a linear regression model, with a mean squared error (MSE) of 0.017 and a mean absolute error (MAE) of 0.075. Boyapati Sai Venkat[5] utilized the Gradient Boosting Algorithm on the same housing dataset, providing results including RMSE: 0.3323, MSE: 0.1105, and MAE: 0.2315.Samkit Saraf et al. [6] performed a comparative analysis of the performance of the multiple regression algorithm on the Melbourne city house price dataset.

Seng Jia Xin and Kamil Khalid [7] compared results with a Ridge regression model on the housing dataset of Ames City in Iowa, USA, providing the results of RMSE: 0.1333 and R2: 0.889. Weinan Weng [8] utilized the Gradient Boosting Algorithm on a housing price dataset with 1460 rows and 80 features.Qingqi Zhang[9] analyzed a housing price dataset with 12 independent variables and 1 dependent variable using a multiple regression algorithm. David Emmanuel Aniobi [10] also analyzed a housing price dataset using the Linear regression algorithm and ridge regression, providing various results for Linear regression MSE:3.12, ridge regression MSE:25365.49 .

Junjie Liu [11] provided important insights for further exploration while analyzing a housing price dataset using the linear regression algorithm. Kunal Sapkal [12] analyzed the housing dataset of Bangalore with 9 columns and 11200 records to compare results with the Ridge regression model, predicting 93% accuracy with MSE as 0.017 and MAE as 0.075. Chenxi Li [13] performed analysis on the King County USA dataset with linear regression, providing results including R2: 0.706, RMSE: 210649.77, and MSE: 44373326009.81.

Table 1: Summary of Literature Survey

Paper Ref.No.	Author	Merits	Demerits
[3]	Dr. Sonia Juneja and Neha Choudhary	- Comprehensive data preprocessing. -Achieved 77% accuracy with multiple regression analysis.	- Accuracy is relatively lower compared to other methods.
[4]	Hemalata Sharma et al	- High accuracy (93%) with linear regression. - Low MSE and MAE values.	- Specific to Ames City, limiting generalizability.
[5]	Boyapati Sai Venkat	- Effective use of Gradient Boosting Algorithm. - Detailed error metrics provided (RMSE, MSE, MAE).	- Higher complexity in the model. - May require more computational resources.
[6]	Samkit Saraf et al	- Comparative analysis of multiple regression on Melbourne dataset.	- Lack of error metrics and accuracy details.
[7]	Seng Jia Xin and Kamil Khalid	- Application of Ridge regression with detailed metrics (RMSE, R2).	- Limited to Ames City dataset. - Less detailed preprocessing

			steps mentioned.
[8]	Weinan Weng	- Utilized Gradient Boosting Algorithm on a large dataset (1460 rows, 80 features).	- No accuracy or detailed error metrics provided.
[9]	Qingqi Zhang	- Analyzed dataset with multiple regression, considering12 independent variables.	- No error metrics or accuracy details provided. - Limited number of variables.
[10]	David Ennamuel Aniobi	- Linear regression and ridge regression analysis. - Provided various results.	- Lack of detailed error metrics and comparative accuracy.
[11]	Junjie Liu	- Insights for further exploration using linear regression.	- No specific results or metrics provided. - Limited information on preprocessing.
[12]	Kunal Sapkal	- Analysis of Bangalore dataset with Ridge regression. - High accuracy (93%), low MSE and MAE.	- Limited to Bangalore dataset. - Limited number of variables analyzed.
[13]	Chenxi Li	- Analysis on King County USA dataset with linear regression. - Detailed metrics (R2, RMSE, MSE).	- R2 value indicates moderate model performance. - High RMSE and MSE values.
[14]	Eva Ostertagova	- Polynomial Regression	- Lack of detailed error metrics and comparative accuracy.
[15]	Xiao Tianli and Raymond	- Important preprocessing step (StandardScalar) highlighted.	- No specific results or accuracy details provided. - Focus on preprocessing rather than model.
[16]	Pinguang Ren	- Ridge and Elastic Net Regression for predicting students' teaching quality.	- Focused on teaching quality rather than housing prices. - Results not directly comparable.

Eva Ostertagova[14] conducted an analysis on the dataset using Polynomial Regression, providing the following results RMSE: 7.876, R2: 0.9233. Xiao Tianli and Raymond [15] standardized their dataset using the StandardScalar method, an important preprocessing step before applying machine learning algorithms. Pinguang Ren's research paper [16] predicted students' teaching quality achievement using Ridge Regression and Elastic Net Regression, providing various results for Ridge Regression MAE:1.348, R2:0.7804 and Elastic Net Regression: for MAE:1.3194, R2:0.7855. Table 1 shows the summary of work done by various researchers.

II. METHODOLOGY

Machine learning is a subfield of Artificial Intelligence (AI)[17] that gives machines the capacity to autonomously learn new things and develop as a consequence of their experiences, even without being specifically programmed to do so. Technological improvements have led to an increase in machine learning (ML). Because of its effectiveness in a variety of areas, including prediction, defect identification, pattern identification, and so forth, machine learning has become more and more popular worldwide.

Figure 1 represents the workflow implemented in this paper work. The pre-processing was done to remove noisy data such as duplicates, missing values and outliers. After preprocessing, the dataset was split into train and test sets in the ratio of 90:10. The data that has been prepared goes into the machine learning model. Each algorithm mentioned below was trained on a portion of the preprocessed dataset and then tested on a dataset to assess its predictive accuracy. Evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute error(MAE) were utilized to compare and assess the accuracy and predictive power of the trained model. The materials and models used in this work were discussed below.

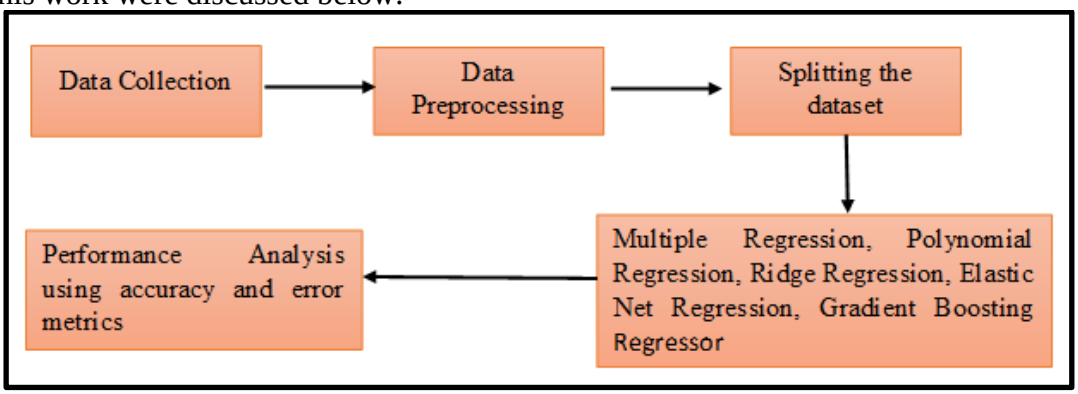


Figure 1: Architecture of the work

A. DATASET

The dataset used for this work is taken from kaggle [18]. The dataset contains information on 5000 houses in Chennai. It has eight attributes such as price in lakhs, area in square feet, status in ready to move or it is under construction, BHK represents number of bedrooms, halls and kitchens, number of bathrooms, age of the house in years, location, and builder name.

B. MULTIPLE LINEAR REGRESSION ALGORITHM

Multiple linear regression is an extension of simple linear regression that models the relationship between a single dependent variable and multiple independent variables. It considers multiple predictors to understand their combined effect on the dependent variable. The model is expressed as

$$Y = b0 + b1 * x1 + b2 * x2 + b3 * x3 + \ldots\ldots bn * xn \quad \textbf{(1)}$$

Where Y is the Dependent variable, while x1, x2, x3, ……xn are the Independent variables.

C. POLYNOMIAL REGRESSION ALGORITHM

Polynomial Regression acts as a regression algorithm. It is how a dependent variable (y) relates to an independent variable (x) by using an nth-degree polynomial. This method is often seen as a special case of Multiple Linear Regression in ML. The Polynomial Regression equation is given below

$$y = b0 + b1x1 + b2x12 + b2x13 + \ldots\ldots bnx1n \quad \textbf{(2)}$$

So to transform the Multiple Linear Regression equation into Polynomial Regression, we must add certain polynomial terms to it.

D. RIDGE REGRESSION ALGORITHM

Ridge Regression is the one of the most pervasive techniques in machine learning, and it helps to reduce overfitting and improves the generalization performance of a model.. It involves adding a penalty term to the loss function, penalizing the sum of the squared values of the model's coefficients.

By imposing this penalty on the squared sum of the coefficients, Ridge Regression discourages large coefficient values while still allowing all coefficients to remain non-zero. This property results in smoother and more stable models compared to Lasso regularization, making Ridge Regression particularly suitable for situations where all features are potentially relevant and the goal is to reduce overfitting without discarding any features completely. In Ridge Regression, the regularization term added to the loss function is expressed as

$$\beta\hat{} = (X^T X + \lambda I)^{-1} X^T Y \quad \textbf{(3)}$$

Where β is the vector of coefficients, X is the matrix of independent variables, Y is the vector of values of dependent variables, $\boldsymbol{\lambda}$ is the penalty parameter and I is the identity matrix.

E. ELASTIC NET REGRESSION ALGORITHM

Elastic Net Regression is a robust model that combines the penalties of ridge regression and lasso regression. It uses independent variables (predictor variables) and dependent variables (response variables). The model uses L1 and L2 regularization terms and is expressed as.

$$\left(\frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda P_\alpha(\beta)\right) \quad \textbf{(4)}$$

When α = 1, elastic nets behave similarly to lasso. Elastic net approaches ridge regression as α approaches zero. For other values of α, the penalty term Pα(β) interpolates between the L1 norm of β and the squared L2 norm of β. It is suitable for datasets with many correlated predictors and helps select important features. Applications include predicting significant events.

F. GRADIENT BOOSTING REGRESSOR ALGORITHM

Gradient Boosting Regressor technique used for regression and classification problems, combining weak prediction models. Iteratively adding models to correct errors, it combines multiple weak learners, like decision trees, to form a strong predictor. The model aims to minimize loss function by adding weak learners trained on residuals of previous learners. It is widely used in machine learning competitions and practical tasks.

G. MEAN ABSOLUTE ERROR (MAE)

Within the domains of machine learning and statistics, the Mean Absolute Error (MAE) is an often utilized indicator. It's a measurement of the typical absolute discrepancies between a dataset's actual values and predicted values.

$$\textbf{MAE} = \frac{1}{n} \sum_{i=1}^{n} |x\,i - y\,i| \quad \textbf{(5)}$$

x_i represents the actual or observed values for the i^{th} data point. y_i represents the predicted value for the i^{th} data point.

H. MEAN SQUARED ERROR (MSE)

A frequently used metric in statistics.& machine learning is called Mean Squared, or MSE.

It measures the square root of the average discrepancies between a dataset's actual values and predicted values. People often use MSE to look at regression problems. It helps check how well predictive models are working.

$$\textbf{MSE} = \frac{1}{n} (x\,i - y\,i)^2 \quad \textbf{(6)}$$

xi represents the actual or observed value for the i-th data point. yi represents the predicted value for the i-th data point.

I. ROOT MEAN SQUARED ERROR (RMSE)

Root Mean Squared Error, or RMSE, is a commonly used accuracy metric in machine learning and regression analysis. It shows how good a model is at making predictions.

$$\textbf{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x\,i - y\,i)^2} \quad \textbf{(7)}$$

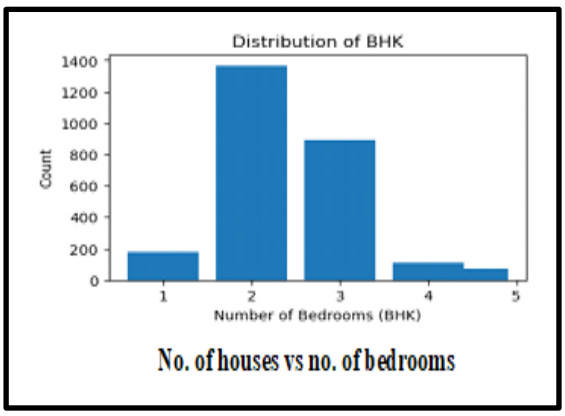
xi represents the actual or observed value for the i-th data point. yi represents the predicted value for the i-th data point.

IV.RESULTS AND DISCUSSIONS

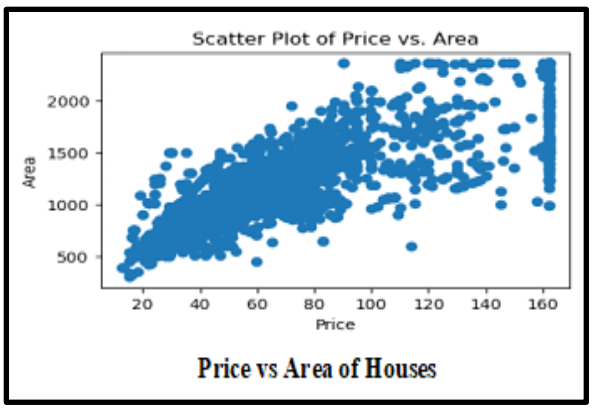
The analysis and interpretation of the dataset helps in understanding the attributes and their correlation. This gives wide scope in retaining the strongly correlated attributes and removing the remaining. Figure 2 showcases the analysis of the features. Two BHKs have the highest count than the one, three, four and five. This shows that two BHK are the most common, likely representing compact apartments suitable for individuals or couples.

Figure 2(a) shows the distribution of the number of bedrooms (BHK) in a dataset of houses. The most common number of bedrooms is 2, followed by 3, while 1, 4, and 5 bedrooms are less common. The count of houses decreases significantly for properties with either fewer or more than 2-3 bedrooms. Figure 2(b) illustrates the relationship between price and area, with blue dots representing data points. Prices vary within each area cluster, with some exhibiting a wider spread and others showing more consistent values.

The results of applying various machine learning algorithms to the chennai housing dataset were discussed below. Each model's performance metrics, including Accuracy, Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) have been evaluated and compared. The accuracy of various regression models was evaluated and tabulated in table 2. It reveals that Multiple Linear Regression achieved an accuracy of 80.83%. Polynomial Regression performed better with an accuracy of 84.48%. Ridge Regression showed similar results to Multiple Linear Regression, with an accuracy of 80.83%. Elastic Net Regression had a slightly lower accuracy of 80.65%. Gradient Boosting outperformed all other models, achieving the highest accuracy of 90.16%. The performance of various regression models was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) is shown in table 3. Figure 3 and 4shows the respective pictorial representations of table 2 and 3. The Multiple Linear Regression model had an MAE of 12.9407, an MSE of 327.5503, and an RMSE of 18.0984. The Polynomial Regression model performed better with an MAE of 11.4095, an MSE of 265.2322, and an RMSE of 16.2860.



2(a)



2(b)

Figure 2: Sample Analysis of Features

Table 2: Accuracy Results of Algorithms

Models	MLR	PR	RR	ENR	GBR
Accuracy	80.8293	84.4766	80.8290	80.6533	**90.1581**

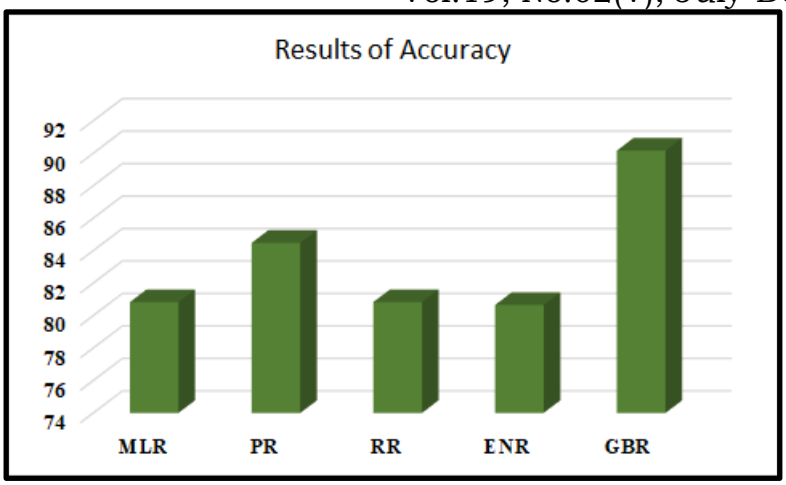


Figure 3: Results of Accuracy

Table 3: Standard Error Metrics of Algorithms

Models	MLR	PR	RR	ENR	GBR
MAE	12.9407	11.4095	12.9408	12.9505	**8.8137**
MSE	327.5503	265.2322	327.5504	330.5582	**168.1587**
RMSE	18.0984	16.2860	18.0984	18.1813	**12.9676**

Ridge Regression showed similar results to Multiple Linear Regression, with an MAE of 12.9408, an MSE of 327.5504, and an RMSE of 18.0984. Elastic Net Regression had an MAE of 12.9505, an MSE of 330.5582, and an RMSE of 18.1813. Gradient Boosting outperformed all other models, achieving an MAE of 8.8137, an MSE of 168.1587, and an RMSE of 12.9676.

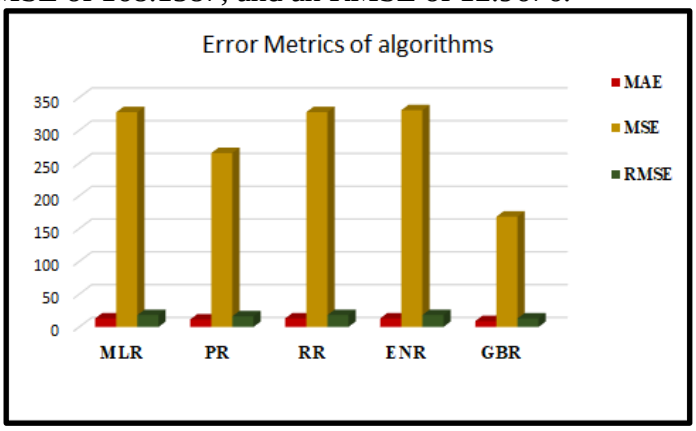


Figure 4: Results of Error Metrics

Table 4: Key Insights of the Models

Factors	Observations
Model Complexity	More complex models like Polynomial Regression and Gradient Boosting performed better than simpler models like Multiple Linear Regression and Ridge Regression. This suggests that capturing non-linear relationships and interactions is crucial for accurate house price predictions.
Regularization	The impact of regularization (Ridge and Elastic Net) was minimal on this dataset. This could indicate that the dataset did not suffer from severe overfitting issues, or the regularization parameters were not optimal.

Ensemble Methods	The superior performance of Gradient Boosting demonstrates the power of ensemble methods in improving predictive accuracy. By combining multiple weak learners, Gradient Boosting effectively reduces both bias and variance, leading to more precise predictions.

Table 4 represents the insights observed during the implementation of models on the dataset.

V. CONCLUSION

Improvement in computing technology has made it possible to examine social information that cannot previously be captured, processed and analyzed.This paper is an exploratory attempt to analyze the regression algorithms in house price estimation and further compare their results. The algorithms used were Multiple Linear Regression, Polynomial Regression, Ridge Regression, Elastic Net Regression, Gradient Boosting Regressor. In this work, the data set was used from online repository kaggle. Exploratory Data Analysis was done on the data to study features. After preprocessing, machine learning models were trained on the training data. The performance of the models was evaluated using accuracy and standard error metrics like Mean Absolute Error, Mean Square Error, Root Mean Square Error. The gradient boosting regressor outperformed with accuracy of 90.16% and Multiple Linear Regression with 80.83%. Polynomial Regression with 84.48%, Ridge Regression with 80.83% and Elastic Net Regression with 80.65%. The error metrics also showed that the gradient boosting regressor is having minimal error than all the other models. Thus, the study shows that a gradient boosting regressor performs excellent than all other regression models. To differentiate itself, the model should consider additional parameters like tax and air quality. This helps people buy houses on a budget and minimizes financial loss, utilizing multiple algorithms for accurate selling prices.

REFERENCES

[1]Amena Begum, Nishad Jahan Kheya, Md. Zahidur Rahman, “Housing Price Prediction with Machine Learning”, International Journal of Innovative Technology and Exploring Engineering, Vol. 11, No. 3, pp. 42–46, Jan. 2022, doi: 10.35940/ijitee.C9741.0111322

[2]Iwona Foryś , “Machine learning in house price analysis: regression models versus neural networks” Procedia Computer Science, Elsevier, pp. 435–445.2022. doi: 10.1016/j.procs.2022.09.078.

[3]Dr. Sonia Juneja, Neha Chaudhary, Ritul Gupta, Ojasvi Kaushik, Mohd Ishan, Ayush Sharma, “House Price Prediction Using Machine Learning Algorithms”, International Journal of Research in Applied Science and Engineering Technology, Vol. 11, No. 6, pp. 3156–3164, Jun. 2023, doi: 10.22214/ijraset.2023.54259.

[4]Hemlata Sharma , Hitesh Harsora and Bayode Ogunleye, “An Optimal House Price Prediction Algorithm: XGBoost”, Analytics, Vol. 3, No. 1, pp. 30–45, Jan. 2024, doi: https://doi.org/10.3390/analytics3010003

[5]Boyapati Sai Venkat , Maddirala Sai Karthik, Konakanchi Subrahmanyam, B Ramachandra Reddy,“An Analysis of House Price Prediction Using Ensemble Learning Algorithms”, Research Reports on Computer Science, pp. 87–96, May 2023, doi: https://doi.org/10.37256/rrcs.2320232639

[6]Samkit Saraf, Aryan Verma, Shashwat Tare, Varun Vasani, Dr. Kamal Mehta, “House Price Prediction Using Linear Regression”, International Journal of Research in Applied Science and Engineering Technology, Vol. 9, No. 10, pp. 1811–1815, Oct. 2021, doi: 10.22214/ijraset.2021.38715.

[7]Seng Jia Xin, Kamil Khalid, “Modelling House Price Using Ridge Regression and Lasso Regression”, 2018. [Online]. Available: www.sciencepubco.com/index.php/IJET

[8]Weinan Weng, “Research on the House Price Forecast Based on machine learning algorithm”, 2022.

[9]Qingqi Zhang, “Housing Price Prediction Based on Multiple Linear Regression”, Science Program, Vol. 2021, 2021, doi: https://doi.org/10.1155/2021/7678931.

[10] David Emmanuel Aniobi, Chukwuemeka Oluebube Ochuba, Saater Benedicta Nguideen, “House Price Prediction: Comparative Analysis of Regression-Based Machine Learning Algorithms”, International Journal of Research in Applied Science and Engineering Technology, Vol. 11, No. 10, pp. 1550–1557, Oct. 2023, doi: 10.22214/ijraset.2023.56232.

[11] Junjie Liu, “Dataset Analysis and House Price Prediction”, 2024.

[12] Kunal Sapkal, Pratik Nikam, Rahul Rasal , Tilakram Yadav , Manoj Shelar, "Machine Learning based Predicting House Prices using Regression Technique", International Journal Of Scientific Research In Engineering And Management, Vol. 08, No. 04, pp. 1–5, Apr. 2024, doi: 10.55041/IJSREM30682.
[13] Chenxi Li, "House price prediction using machine learning", Applied and Computational Engineering, Vol. 53, No. 1, pp. 225–237, Mar. 2024, doi: 10.54254/2755-2721/53/20241426.
[14] Eva Ostertagová, "Modelling using polynomial regression", Procedia Engineering, Elsevier, pp. 500–506, 2012, doi: https://doi.org/10.1016/j.proeng.2012.09.545.
[15] Xiao Tian Li and Raymond Y. Huang, "Standardization of imaging methods for machine learning in neuro-oncology", Neurooncol Advances, Vol. 2, pp. IV49–IV55, Dec. 2020, doi: https://doi.org/10.1093/noajnl/vdaa054
[16] Pinguang Ren, "Comparison and analysis of the accuracy of Lasso regression, Ridge regression and Elastic Net regression models in predicting students' teaching quality achievement", Applied and Computational Engineering, Vol. 51, No. 1, pp. 313–319, Mar. 2024, doi: https://doi.org/10.54254/2755-2721/51/20241625.
[17] Bharati Panigrahi, Krishna Chaitanya Rao Kathalab, M. Sujatha, "A Machine Learning-Based Comparative Approach to Predict the Crop Yield Using Supervised Learning with Regression Models", Procedia Computer Science, Elsevier, pp. 2684–2693,2022 doi: https://doi.org/10.1016/j.procs.2023.01.241.
[18] "Kaggle Chennai House Dataset." Accessed: Jul. 28, 2024. [Online]. Available: https://www.kaggle.com/datasets/amaanafif/chennai-house-price