International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(15s) 2026 pp. 998-1005
https://doi.org/10.70917/ijcisim-2026-4497
Article

# Predictive Modeling of Real Estate Prices Using Machine Learning Techniques

**Mahesh Mahajan[1], Makrand Shahade[2], Aishwarya Nikam[3], Paresh Wani[4], Bhagyashri Patil[5], Chandrabhan Patil[6]**

[1]Department of Master of Computer Applications, SVKM's Institute of Technology, Dhule, Maharashtra, India.
Email: mahesh.docsnmu@gmail.com
[2]Department of Computer Engineering, SVKM's Institute of Technology, Dhule, Maharashtra, India.
Email: makarand.shahade@svkm.ac.in
[3] Department of Master of Computer Applications, SVKM's Institute of Technology, Dhule, Maharashtra, India.
Email: aishwaryasnikam15@gmail.com
[4] Department of Master of Computer Applications, SVKM's Institute of Technology, Dhule, Maharashtra, India.
Email: pareshwani8828@gmail.com
[5] Department of Master of Computer Applications, SVKM's Institute of Technology, Dhule, Maharashtra, India.
Email: patilbhagyashri742@gmail.com
[6] Department of Master of Computer Applications, SVKM's Institute of Technology, Dhule, Maharashtra, India.
Email: chandrabhanpatilashok@gmail.com

**Abstract:** Accurately estimating residential property prices is a challenging yet essential task in today's changing market environment. Conventional manual valuation methods often result in inconsistencies; therefore, data-driven approaches have increased importance. This research focuses on developing a regression model using machine learning that can predict house prices and assist in informed decision-making within the real estate industry. The Ames, Iowa dataset from Kaggle contains 1,460 houses with 81 features that include lot size, neighborhood, construction year, and overall quality. Data pre processing involved handling missing values in numerical fields with median imputation and in categorical fields with mode imputation, followed by the application of one-hot encoding and standardization of features using the z-score.

The regression algorithms Random Forest, Linear Regression, Decision Tree, Support Vector Regression, and K-Nearest Neighbors were trained on an 80/20 train-test split and evaluated using the metrics R2, RMSE, and MAE. Among these, the Random Forest model performed best with R2 ≈ 0.891, RMSE ≈ 28,872,
and MAE ≈ 17,614. On the other hand, the weakest performance was observed for Support Vector Regression.

The final Random Forest model was deployed as a web application using Streamlit, allowing users to input features and obtain instant predictions of property prices. Overall, the study showed that advanced regression methods and effective feature engineering significantly contribute to improving real estate price forecasts.

**Keywords:** House Price Prediction, Machine Learning, Real Estate, Random Forest Regression, Linear Regression, Decision Tree Regression, K-Nearest Neighbors (KNN), Support Vector Regression (SVR)

## 1. INTRODUCTION

Shelter is one of the most essential needs in human life. Selecting the right house depending on Location, Price, size, and available facilities. [1] House price prediction means using

a computer model to analyze how much a house costs based on its features. Predicting the current house price has become a more challenging task. In the present World, the cost of housing is changing rapidly. [3] Over the years, property prices have gone up and down due to various factors, such as demand, economic conditions, and political events. A house is one of the important needs in a person's life. Some people buy a house to live in, others buy them as an investment. The housing market plays a major role in a country's economy. Demand for houses keeps increasing day by day because of rapid population growth and urbanization. People migrated and are still migrating to cities in search of opportunities. Previously, house prices were estimated by hand, either from experience or rough calculations, both of which often resulted in errors and losses. [10] This calls for a growing urge to understand and predict house prices among buyers, sellers, and investors.

[9] House prices often change due to many elements, such as location, structural features like the number of rooms or total area. These price changes can cause challenges for buyers, sellers, and developers. Then, the house price prediction system's correct price prediction of the house has become an important task. [4] For many buyers, it is tough to say whether the listed price of a house is fair or not. Similarly, for sellers, the prices quoted may not be exactly reflective of prevailing market conditions. Traditional price estimation methodologies fail to present appropriate results because they are inefficient in analyzing big and complex data. [7] Machine learning (ML) can now predict house prices in a much smarter and faster way, learning from a huge amount of past data. The main idea of this research is to create a strong Machine Learning (ML) model that will predict house prices based on several factors. [10] Machine learning provides unique and modern techniques for predicting house prices using historical

data set analysis. Analyze the relationship between house features on market price. Machine Learning (ML) regression algorithms are particularly useful because they can predict house prices continuously. In this study, different Machine Learning algorithms such as Linear Regression, Random For-est Regression, Decision Tree, K-nearest neighbors (KNN) are utilized for House price prediction, and accuracy is selected. [1], [2] Now, in solving this problem, Machine Learning techniques are being used. ML lets computers learn from data to make accurate predictions, without actually being programmed for this. Among several ML algorithms, the Random Forest algorithm has proved to be one of the most reliable and accurate methods for house price prediction; by combining a set of decision trees, Random Forest is likely to provide better and more stable predictions. [7] The Linear Regression model creates a straight-line relationship between the features of the houses and their prices, the Decision Tree model divides of data into smaller parts based on feature values like area, location, or rooms, and decides upon each step accordingly. These methods help in arriving at better accuracy and reliability in the results. [6] By using Random Forest, we can analyze features regarding the house and understand which ones affect its price the most, like location, number of rooms, or nearby facilities. This would improve not only the predictive accuracy but also save time and effort rather than compared to traditional approaches. This project will use the real estate data available-for instance, Kaggle datasets-with features including location, number of bedrooms and bathrooms, and square footage. The system processes this data through cleaning, analysis, and model training steps. [9] These algorithms minimize time, human error, and also increase accuracy. The present study aims to develop an appropriate predictive model of house prices using different techniques of machine learning, including N-nearest neighbors and Support Vector Regression. This model will be trained on real-world data after its preparation through steps such as feature engi-neering and cleaning to enhance its performance and accuracy. The general objective of this Study, therefore, is to come up with a fact-based prediction system for house pricing that will contribute to making better decisions in the real estate industry. This research study analyzes the factors of house pricing and the different machine learning models to choose the best model that can give accurate results in predicting the price. [1], [2], [4] Such a system allows any person - a buyer, a seller, or a real estate company - to easily investigate whether the price of a house has been compared to the market. It gives a clearer view of the present market, promotes affordable pricing, and supports time-saving property. [5] This process of homes becomes smarter, faster, and more solution-oriented, accurate for buying and selling. [3] This study highlights how real-world problems concerning property, predicting house prices, can be resolved using Machine Learning. The pro posed model helps them in decision-making towards better financial outcomes and improves the general efficiency of the real estate market. [6] In short, this paper will look at how machine learning improves the efficiency, accuracy, and reliability of

real estate valuations to enable better decision-making by individuals within a rapidly changing housing market. [11]

## 2. LITERATURE REVIEW

This study [11] represents the initial investigation into house-price prediction, validating three machine-learning models: The Decision Tree Regression model, and Linear Regression model, and Random Forest Regression model. These models have been trained on various housing features such as a number of rooms, location, and overall condition. Among them, the Decision Tree Regression model yielded an RMSE of 4.32%, while the Linear Regression model gave an RMSE of 5.03%. Then, altogether, the Random Forest model performed the best with an RMSE of 3.29%. The paper finds that Random Forest provides a highly accurate forecast amid the verified methods.

The second study [12] states that location, size, and number of bedrooms have major influence on costs when it comes to machine learning research on the prediction of home and rental prices. XGBoost, Random Forest, SVR, LASSO, Linear Regression, and Cat Boost are common algorithms; Random Forest usually produces the best results. In this study, the accuracy of the rent prediction model using linear regression was approximately 89%, whereas the accuracy of the house price model was approximately 90%.

In this paper, [13] found that neighborhood quality, struc-tural features, and location are important predictors of home prices. Multiple Linear Regression model, Support Vector Re-gression (SVR), Gradient Boosting, Artificial Neural Networks (ANN) and the Hedonic Price Model are some of the methods discussed in the paper to model these relationships.according to earlier research, the number of rooms, bathrooms, area, location, and proximity to amenities all affect a home's price. Researchers used machine learning techniques because earlier manual prediction systems had an error rate of almost 25%. Regression models, decision trees, and hybrid approaches were employed in numerous studies to identify intricate patterns in housing data.

This paper [14] as per the document one of the major requirements for the machine learning to be complex for home price prediction to be accurate is that a simple solution like lin-ear regression cannot catch complex relationships. The paper compares the algorithms involved Linear Regression, Decision Trees, Gradient Boosting (XGBoost, Light GBM), Random Forest, ANN, SVM, KNN, and a hybrid ensemble model. The proposed ensemble model outperformed other models, showing higher accuracy and lower errors, as it was able to capture non-linear patterns effectively. Prices are significantly influenced by the location, size, age, economic indicators, and quality of the neighborhood; feature engineering and model optimization help to improve results.

According to the researchers, [15] machine learning al-gorithms such as Random Forest, The Decision Trees, and Linear Regression, and XGBoost are mostly employed, and the prediction of house prices is vital for buyers and sellers. The research reveals that Random Forest which very often attains

90% accuracy is the best compromise between efficiency and accuracy. The location, size of the property, the number of bedrooms, the amenities, and the adjacent facilities are some of the factors that determine the price. Accurate predictions can only be made if feature selection, data pre processing, and model evaluation are done properly. Ensemble methods like, Random Forest model, which is a collection of decision trees, are typically able to achieve better results than support vector models and traditional regression.

The study [16] survey, several Machine Learning (ML) and Deep Learning (DL) methods have been utilized in the research which is aimed at the prediction of housing prices. Location, the number of rooms, the year of construction, and amenities have been the most significant variables that were used as the main predictors. To make the prediction more accurate, DL models based on Convolutional Neural Networks (CNN's) and traditional regression analysis were employed. The model's efficiency was elevated and the resid-ual errors were lessened by means of hyperparameter tuning with optimizers such as Adam, RMSprop, AdaGrad, AdaDelta, Nadam, AdaMax, and FTRL. Adaptive Exploratory Data Analysis (AEDA) was used to recognize patterns, handle missing values, and get the most relevant features from the datasets. Researchers employing this method have shown the effectiveness of the integration feature engineering, DL, and ML models, thus, they have been able to reach the prediction accuracy of around 92–95%.

An article review paper [17] This research applies machine learning, particularly XGBoost and Gradient Boosting, to predict housing prices. Part of the data pre processing involved handling missing values and feature normalization. XGBoost, with an $R2$ of 0.64, was better than Gradient Boosting, which had an $R2$ of 0.61 Byrnes Also, there was mention of different algorithms such as SVM, KNN, Random Forest, and Decision Trees, and Linear Regression. The best performing method was XGBoost, which was able to handle complex data, prevent overfitting, and give the most useful features such as size, location, and bedrooms.

An article [18] reviewing this research implements a stack-ing ensemble of CNN and machine learning models (Random Forest, XGBoost, AdaBoost) enhanced with linear regression to predict the prices of second-hand houses in Thailand. The number of bedrooms, the number of bathrooms, the lot size, the house size, public transportation, neighboring roads, average district prices, and image features are the main attributes. CNN-XGBoost with linear regression produced 22.78% while CNN-RF with linear regression was beaten by CNN-XGBoost with a MAPE of 21.78%. The results reveal that the prediction accuracy is substantially higher when image features are used together with ensemble models and linear regression. Algorithms used: CNN, Random Forest, XGBoost, AdaBoost, and Linear Regression; best MAPE accuracy of 21.78% Regression models including Linear Regression, and Decision Tree, and

Lasso, Random Forest, and XGBoost are employed in the research described in paper to predict the housing prices in big Indian metropolitan areas. The dataset

consists of 35,062 records with nine features.

The trained Random Forest model was integrated into a Streaml it-based web application to make the system accessi-ble and interactive. Users can input property details such as liv-ing area, construction year, and quality rating, and receive an instant predicted price. The application also includes graphical comparisons between all algorithms to display performance metrics such as $R2$, RMSE, and MAE.

## 3. METHODOLOGY

This research work aims at developing an intelligent and data-driven system that is capable of predicting the price of a house by availing the use of Machine Learning(ML) techniques. The methodology followed in this project is very systematic, comprising steps such as gathering of reliable data, preparing the data via cleaning and transformation, selecting suitable algorithms, training and evaluation of the models, and finally deploying the best-performing model through a user-friendly web interface. [1], [3], [4]

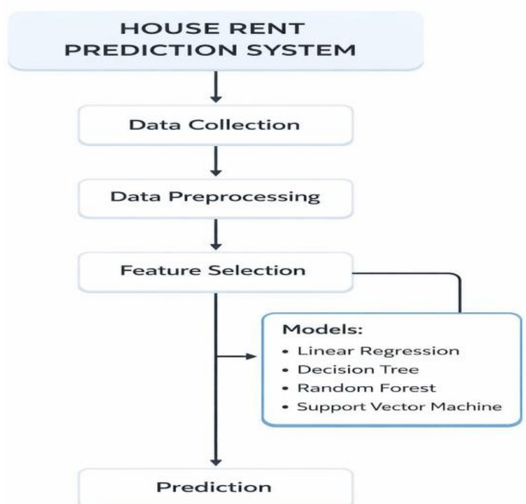


Fig. 1. Proposed architecture of house price prediction model.

### *A. Data Collection*

The data analyzed below was obtained for this research from the repository of the Kaggle [8] competition entitled House Prices-: Regression Techniques Advanced, This dataset composed of 1,460 records with 81 different attributes, each one with Pertinent information about residential properties in Ames , Iowa (USA). This dataset contains both numerical and categorical variables that greatly influence the selling price, including Lot Area, Neighborhood, Year Built, Overall Quality, and Garage Cars. The dependent variable or target attribute is represented by the Sale Price,which means the market value of each house. It is one of the well-accepted datasets in regression analysis and sets a realistic platform for the estimation of real estate prices.

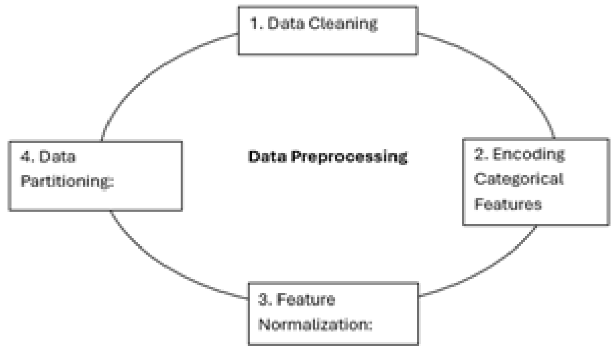


Fig. 2. Proposed architecture of house price prediction model.

## *B. Data Pre processing*

Raw data typically includes missing values, inconsistent formats, and other information that may not make a difference, which all could impact model accuracy. Pre processing the data was therefore necessary to guarantee that the dataset was clean, consistent, and prepared for model training. [4], [8] The data is cleaned up in the Pre processing state as shown in Fig. 2. It is a simple The process of locating and removing mistakes to enhance data efficiency. [1]

*a)* Data Cleaning:: Numerical voids were replaced by the median value while the absent categorical data was replaced by the mode. This avoids bias that might come about when one removes incomplete rows.

*b)* Encoding Categorical Features: : Since machine learning algorithms work best with numerical data, categorical columns were transformed using One-Hot Encoding, creating binary variables for each category.

*c)* Feature Normalization: : To maintain consistency among features of different scales, the Standard Scaler technique was applied:

$$x_{\text{scaled}} = \frac{X - \mu}{\sigma} eq \quad (1)$$

where μ denotes the mean and σ denotes the standard deviation of the feature.

*d)* Data Partitioning: : Here are a few paraphrased versions of that sentence: The 'train-test' split technique so that a model will generalize well on unseen data into 80% with dataset, was divided and 20% for training and testing.

## *C. Model Development*

Multiple supervised learning algorithms were implemented and compared to deter mine which one was the most efficient in making the predictions of house prices.[9,11] The selected algorithms are briefly described below:

*a)* *Linear Regression::* It is an algorithm that assumes the target variable and a linear relationship between the independent features. [4]

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdot\cdot\cdot + \beta_n X_n + \epsilon \quad (2)$$

then, $Y$ is the predicted price, $X_i$ represents the input features, $\beta_i$ are the coefficients, and $\epsilon$ is the error term.

*b)* *Decision Tree Regressor: :* This algorithm segments data into smaller subsets by using conditional splits.

[4]Each leaf node gives a predicted output, and each internal node represents a test on a feature. While representing nonlinear relationships effectively, it may overfit if not controlled properly. [5], [9].

c) *Random Forest Regressor: :* It is an algorithm that assumes the target variable and a linear relationship between the independent features.Random Forest works by taking the outputs of multiple Decision Trees and averaging their predictions:

$$\hat{Y} = \sum_{t=1}^{T} f_t(X) \tag{3}$$

where T represents the number of trees and $f_t(X)$ is the prediction from the $t^{th}$ tree. This ensemble approach reduces variance and increases predictive stability.

d) *K-Nearest Neighbors (KNN): :* KNN predicts the target value by averaging the prices of the K most similar data points (neighbors):

$$\hat{Y} = \frac{1}{K} \sum_{i=1}^{K} y_i \tag{4}$$

where *w* and *b* define the slope and bias of the hyperplane. It works well for smaller datasets.

### *D. Model Evaluation.*

All models were trained on the preprocessed dataset and evaluated using standard regression metrics such as:

a) Coefficient of Determination (R2):: SVR attempts to find a function that fits the data within a tolerance margin ($\epsilon$). The general model is expressed as:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{6}$$

This gives the measure of how well the model describes the variation of the target variable

b) Mean Absolute Error (MAE): :

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{7}$$

Mean Absolute Error is a measure that represents the average size of the errors in a set of predictions, without taking their direction into consideration.

c) Root Mean Square Error (RMSE): :

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{8}$$

It is a measure, in the same units of the target variable, that indicates the magnitude of prediction errors.

TABLE I

PERFORMANCE COMPARISON OF REGRESSION ALGORITHMS

| **Algorithm** | *R2* | **RMSE** | **MAE** | **Result** |
|---|---|---|---|---|
| Linear Regression | 0.6557 | 51392.66 | 20236.41 | Moderate perfor- |
| | | | | mance |
| Decision Tree | 0.7679 | 42194.69 | 26547.49 | Overfitting |

| | | | | |
|---|---|---|---|---|
| | | | | observed |
| KNN | 0.7298 | 45523.11 | 25822.90 | Average |
| | | | | performance |
| SVR | -0.0245 | 88647.43 | 59541.62 | Poor |
| | | | | performance |
| Random Forest | **0.8913** | **28871.66** | **17613.77** | **Best model** |

Random Forest Regressor achieved the highest performance with $R^2 = 0.8913$, RMSE = 28,871.66, and MAE = 17,613.77.

Hence, it was selected as the final model for deployment.

### *E. Model Deployment*

The trained Random Forest model was integrated into a Streamlit-based web application to enhance accessibility and interactivity. Users can input property details such as living area, construction year, and quality rating to receive instant predicted prices. The application also pro-vides graphical comparisons of all algorithms, illustrating performance metrics including $R^2$, RMSE, and MAE.

## 4. RESULT:

After completing all steps of pre-processing and preparing the dataset for model training, the developed house price prediction system was evaluated. The dataset was split into training and testing subsets to measure the performance of the algorithm on unseen data. Standard evaluation metrics, including the Coefficient of Determination ($R^2$), RMSE, and MAE, were used to assess the predictions. These metrics

is attributed to its ensemble approach, where multiple decision trees are combined to improve prediction accuracy while reducing overfitting.

Therefore, the Random Forest model was selected for deployment in the Streamlit-based application. Its excellent evaluation performance ensures that the prediction system provides accurate, fast, and reliable house price estimates for real-world decision-making.

The comparative performance of all models is summarized below:

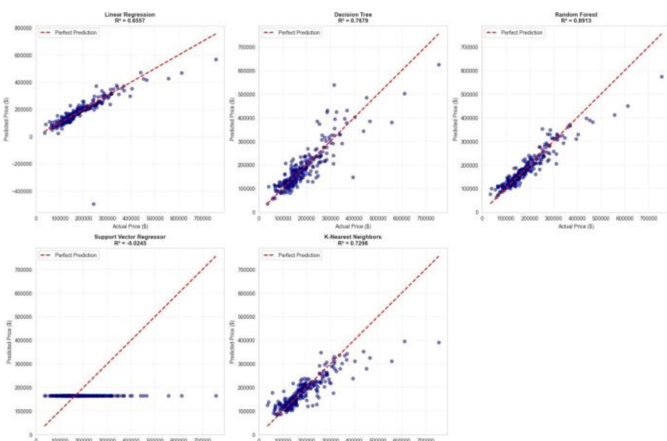


Fig. 3. Accuracy plot. .

## 5. CONCLUSION

The study successfully demonstrates that machine learn-ing can be effectively applied to predict house prices with high accuracy by leveraging structured real-estate datasets. Through systematic preprocessing, feature engineering, and the evaluation of multiple regression algorithms, the Random Forest Regressor emerged as the most reliable model, showing

superior performance across $R2$, RMSE, and MAE metrics. Its ensemble-based learning approach enables it to capture complex, non-linear relationships between property attributes and market values, resulting in more consistent and accurate predictions compared to traditional statistical methods.

The deployment of the final Random Forest model using a Streamlit web application enhances the practical value of this research by allowing users to interact with the prediction system in real time. Users can input property features such as location, number of rooms, size, and structural attributes to instantly obtain estimated market prices. This supports potential buyers, sellers, real-estate consultants, and investors in making informed, data-driven decisions.

The findings of this research highlight the effectiveness of modern machine learning techniques for real-estate valuation. By bridging the gap between traditional appraisal methods and the dynamic nature of today's housing market, this study illustrates how intelligent prediction systems can provide reliable, scalable, and user-friendly solutions for real-world applications.

**References**


1. Sharma, S., Shankar, G., Arora, D., Sharma, P., Motwani, V.:House price prediction using machine learning algorithm.In: 7th International Conference on Computing Methodologies and Communication (ICCMC
2. ), pp. 982–986 (2023).
3. Madhuri, G.V., Kumar, A.A., Hussain, M.A., Girish Kumar, N., Jabbar, M.A.:House price prediction using regression in machine learning.In: In-ternational Conference on Frontier Technologies and Solutions (ICFTS),
4. pp. 1–6 (2025).
5. Dwivedi, R., Gupta, R., Pal, P.K.:House price prediction using regression techniques. In: 4th International Conference on Advances in Comput-ing, Communication Control and Networking (ICAC3N), pp. 165–170 (2022).
6. Saini, K., Kashyap, P. Ramvachan, Yadav, P.: Prediction of house pricing using machine learning.In: 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), pp. 131–135 (2023).
7. Kin, C.C., Salam, Z.A.B.A., Nowshath, K.B.:Machine learning based house price prediction model.In: International Conference on Edge Computing and Applications (ICECAA),pp. 1423–1426 (2022).
8. Gupta, U., Vaidya, H., Chauhan, R., Bhatt, C.:House price prediction us-ing gradient boosting and linear regression.In: International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET), pp. 829–832 (2023).
9. Abubakar, A., Benjamin, I.:Random forest-based house price prediction. In: 3rd International Conference on Information Technology Research and Innovation (ICITRI), pp. 8892 (2023).
10. Simanungkalit, D.G.I., Meylia, B., Salim, J., Edbert, I.S., Suhartono, D.:House base-price prediction with machine learning methods.In: Inter-national Conference on Informatics, Multimedia, Cyber and Information Systems (ICIMCIS), pp. 318–323 (2023).
11. Kumar, A., Shakya, A., Singh, K., Raj, A.:House price prediction using machine learning.In: International Conference on Sustainable Power Energy (ICSPE), pp. 1–10 (2024).
12. Upadhyay, Y., Bhardwaj, D., Agarwal, G., Dutt, A.:Prediction of house price using machine learning.In: 3rd International Conference on Com-munication, Security, and Artificial Intelligence (ICCSAI), pp. 702–705 (2025).
13. Prakash, H., Kanaujia, K., Juneja, S.:Using machine learning to predict housing prices.In: International Conference on Artificial Intelligence and Smart Communication (AISC), pp.1353–1357 (2023).
14. Khandaskar, S. et al.:House and rent price prediction system using regression. In: International Conference on Sustainable Computing and Smart Systems (ICSCSS ) (2023).
15. Kumar, D., Rawat, A.S., Jha, S., Yadav, D.:Machine learning-based prediction of home prices.In: International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) (2023).
16. Kumar, A. et al.: House price prediction system using ensemble learning approach. In: 8th International Conference on Parallel, Distributed and Grid Computing (PDGC) (2024).
17. . Sri, B.U. et al.: Random forest-based house price prediction. In: International Conference on Augmented Intelligence and Sustainable Systems (ICAISS) (2023).
18. Priya, R.P. et al.: Enhancing house price prediction accuracy using deep learning and hyperparameter tuning with AEDA.In: International Conference on Image Information Processing (ICIIP) (2023).
19. Srirutchataboon, G., Prasertthum, S., Chuangsuwanich, E., Pratan-wanich, P.N., Ratanamahatana, C.:Stacking ensemble learning for hous-ing price prediction: A case study in Thailand.In: 13th International Conference on Knowledge and Smart Technology (KST)(2021).
20. Naik, S., Puthran, A.:Metro cities house price prediction using machine learning.In: International Conference on Self Sustainable Artificial In-telligence Systems (ICSSAS) (2023).