Optimized Real Estate Price Prediction Model using GridSearchCV & Machine Learning

Pratik Nagare
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune 411044
pratik.nagare22@pccoepune.org

Suyash Kolte
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune 411044
suyash.kolte21@pccoepune.org

Chaitanya kshirsagar
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune 411044
chaitanya.kshirsagar21@pccoepune.org

Dr. K Rajeswari
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune 411044
kannan.rajeswari@pccoepune.org

Mrs. Sushma Vispute
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune 411044
sushma.vispute@ pccoepune.org

Abstract - **Real Estate industry is dynamic in terms of the prices being fluctuated regularly. It's one of the main area to apply the machine learning concepts to predict the prices of real estate depending upon the current situations and make out maximum accuracy for the same. The research paper mainly focus on to predicting the real valued prices for the places and the houses by applying the appropriate ML algorithms. The proposed article considers some essential aspects and parameters for calculating the prices of real estate property Also some more geographical and statistical techniques will be needed to predict the price of a house. The paper consist how the house pricing model works after using some machine learning techniques and algorithms. The use of the dataset in the proposed system from the reputed website helps to get the detailed analysis of the data points. Algorithms like Linear regression and sklearn are used to effectively increase the accuracy. During model structure nearly all data similarities and cleaning, outlier removal and feature engineering, dimensionality reduction, gridsearchcv for hyperparameter tuning, k fold cross-validation, etc. are covered.**

Keywords - Linear regression model, Python, Machine Learning, House Price, Decision Tree, Lasso, Ridge, KNN.

I. INTRODUCTION

The proposed research paper refers to the predictions on the recent trends and for the plans of economy. The main drive behind the article is prediction of the real estate prices to build best of the house price prediction systems using the machine learning algorithms with maximum accuracy. Under the domain of ML and Data Science the designing of the real estate price prediction along with the full-fledged website is done. According to the census of 2011 only 80 percent of people own their houses. And only people based in rural areas own maximum houses but people in urban sector only about 69 % own a house. This is due to the raising prices of the properties and vague house prices. The main aim to design and develop this model is to produce price prediction system along with a user-friendly front end that will facilitate the users to choose the desired destination and get an idea about the price rates. The Analysis that has been made in the paper is mainly using the dataset from the trusted website that gives ample of sample points for better analysis. One must be aware of the exact price of house before concluding the deal. As the price of house depends on many factors like Area, location, population, size and number of bedrooms & bathrooms given, parking space, elevator, style of construction, balcony space, condition of building, price per square foot etc. The proposed model aims to create an accurate result by taking into consideration all different factors. For House price prediction one can use

various prediction models (Machine Learning Models) like support vector regression, Support vector machine (SVM), Logistic regression, k-means, artificial neural network etc. House- pricing model is beneficial for the buyers, property investors, and house builders. This model will be informative and knowledgeable for the entities related to the real estate and all the stakeholders to evaluate the current market trends and budget friendly properties. Studies initially concentrated on analysis of the attributes which influence prices of the houses based on which model of ML is used and still this article brings together both predicting house price and attributes together. For this paper, Bangalore city is taken as an example because it is Asia's fastest-growing city. The city's growth has already slowed its own economic growth rate and it has gone through various changes that have contributed to its growth over the last few decades, one of which is the IT industry. Bangalore has an excellent social infrastructure, also excellent educational institutions and a rapidly changing physical infrastructure. These factors have led to an increase in migration from other states to Bangalore, but the cost of living as increased, making it difficult for or people to manage their households effectively [5]. The model building starts with the dataset from a reliable source that is simple to use. For a dataset was chosen for our house price prediction, which contains 13320 records of data and 9 features for training our model. There are various machine learning procedures that can be used to forecast future values. In any case, it is required a model that can forecast future property estimations with greater accuracy and less error. With a specific end goal of preparing the model, a significant amount of memorable dataset is required. Generally one wants to create a framework because there is little research on forecasting land property in India. This can forecast the cost of a property by taking into account the various parameters that influence the target value. In addition, the prediction accuracy is measured by taking into account various error metrics [5].

II. Literature Survey

Every common man's first desire and need is for real estate property. Investing in the real estate appears to be very profitable as the property rates do not fall steeply. Investing in real estate appears to be difficult task for investors when one has to select a new house and predict the price with minimum difficulty for this there are several factors which affect the price of a house and all these factors are needed to be taken into consideration to predict the price effectively. Also building such models for prediction needs much research and data analysis as many researchers are already working on it to get the better results.

S. Rana, J. Mondal, A. Sharma and I. Kashyap 2020 [5] have used various regression algorithms to predict the house prices, like XG Boost, Decision Tree Regression, SVR, and Random forest. After applying all these algorithms on to thedataset a comparison for the accuracy is done at the end. From which the maximum accuracy of 99% given by the decision tree algorithm followed by the XG Boost of 63%, this was purely the experimental analysis by testing various algorithms models.

T. D. Phan, 2018 [1] is House Price Prediction using machinelearning algorithms: A case study of Melbourne city, Australia. This is a through case study for analyzing the dataset to give some useful insights on to the housing industry of Melbourne city in Australia. They have used variousregression models. Starting with the data reduction to applying PCA (Principal Component Analysis) steps to get the optimal solution from the dataset. Then they have applied SVM (Support Vector Machine) for the competitive approach. Thus how several methods are implemented to getthe best results out of it.

M. Jain, H. Rajput, N. Garg and P. Chawla 2020 [2] is also ahouse price prediction system using some techniques. In this they have used the simple process of machine learning from data cleaning, visualization, pre-processing and using k-fold cross validation for the output results. Finally they have displayed the graph that shows close resemblance with actual price and the predicted price showing decent accuracy through their working model.

N. N. Ghosalkar and S. N. Dhage 2018 [4], Real Estate Pricevalue using Linear Regression are using simple Linear Regression technique to give the price value for the houses. Through this paper they have tried to have best fitting line (relationship) between the factors of the real estate taken into consideration and used various mathematical techniques likeMSE (Mean Squared Error), RMSE(Root Mean SquaredError) etc.

After reviewing various articles and research papers about machine learning for housing price prediction the article now focus is on understanding current trends in house prices and homeownership. The proposed system uses a machine learning model to predict prices with high accuracy.

III. Proposed System

The main end or focus of our design is to prognosticate the accurate price of the real estate parcels present in India for thecoming forthcoming times through different

Algorithms used in the model building are:

Linear Regression- It's a supervised literacy fashion andresponsible for prognosticating the value of variable(Y) relying on variable(X) which is not dependent [4]. It's the relationship between the input(X) and (Y) [5].

The formula for linear regression equation is given by:
y = a + bx
where y is the predicted value,
a is Y-intercept of the line,
b is Slope of the line,
x is the input value

Least Absolute Shrinkage and Selection Operator- Lasso is direct regression that considers loss. Loss is a point where data values are diminished towards a central point, like the mean. The selection operator is an LR technique that also regularizes functionality, and LASSO stands for least absolute shrinkage. It is similar to ridge regression, but it differs in the values of regularization. The absolute values ofthe sum of the regression coefficients are considered. It evensets the coefficients to zero to eliminate all errors. As a result,lasso regression is used to select features. The lariat procedure encourages simple, sparse models (i.e. models with smaller parameters) [6] [7].

The formula for computing the Lasso regression coefficient can be expressed as:

$\hat{\beta}_{lasso} = \text{argmin} (RSS + \alpha \sum_{j=1}^{p} |\beta_j|)$

Where:

$\hat{\beta}_{lasso}$ represents the estimated coefficients for Lasso regression.

RSS is the residual sum of squares, which measures the difference between the observed and predicted target values.

α is the regularization parameter (tuning parameter), controlling the strength of regularization.

β_j denotes the coefficients of the predictor variables.

Ridge regression is a regularization technique in linear regression that minimizes the residual sum of squares (RSS) between observed and predicted target values while penalizing large coefficients. Ridge regression employs a penalty term proportional to the squared magnitude of coefficients (L2 regularization). This penalty term, controlled by a regularization parameter α helps prevent overfitting by shrinking the coefficients towards zero, particularly useful in handling multicollinearity. Overall, Ridge regression provides stable estimates of coefficients and helps mitigate the issues of overfitting, particularly in situations with highly correlated predictors.

The formula for computing the Ridge regression coefficient can be expressed as:

$\hat{\beta}_{ridge} = (X^TX+\alpha I)^{-1}X^Ty$

Where:

X is the design matrix of predictor variables.

y is the target variable vector.

I is the identity matrix.

α is the regularization parameter.

Decision Tree- It is like linear regression, which is one of the data mining methods of analyzing multiple variables. It is a tree that consist of root node which is also called as decision node and forms a tree with leaf nodes at the end which helps to take the appropriate decision. A sub node is a node with outgoing edges. All other nodes with no outgoing edges are known as child nodes or terminal nodes. Each sub node is parted into two or more sub trees based on the values of the input attributes [8]. Decision tree regression helps to predict the data using trained model in the form of a tree structure to generate the meaningful output and continuous affair which is nothing but non separable result/affair [9].

K-Nearest Neighbors (KNN): K-Nearest Neighbors is a non-parametric algorithm that classifies data points based on the majority class of their nearest neighbors in the feature space. The distance metric (e.g., Euclidean, Manhattan) is used to determine the nearest neighbors.

For a new input sample x:

1. Calculate Distance: Compute the distance between x and all instances in the training dataset using a distance metric (e.g., Euclidean distance, Manhattan distance).

2. Find Neighbors: Select the k instances (neighbors) with the smallest distances to x.

3. Majority Voting: Assign the class label to x based on the majority class among its k nearest neighbors. For classification tasks, this can be achieved through majority voting, where the class with the highest frequency among the neighbors is assigned to x.

Initially feature engineering is applied on the raw data whichincludes cleaning, outlier removal to make the data ready forthe model building. From the fig 1, the dataset is divided intotwo sets i.e. training which is 80% and testing which is 20%.To find the accuracy k-fold cross validation technique is usedwhere value of k is 5 due to which accuracy of model comesout to be around 82% to 85%. The training set is passed through machine learning algorithms to generate trained model also the hyperparameters passed by the k-fold cross validation are helpful to take decision based on best score andbest parameters of the models which are considered here. After evaluating test set and trained model

obtain from a training set is passed on to the artifacts where pickle file contain the model and the json file contain the column details. The back-end is supported by the python flask server which take input as set of values and provide output as predicted values.

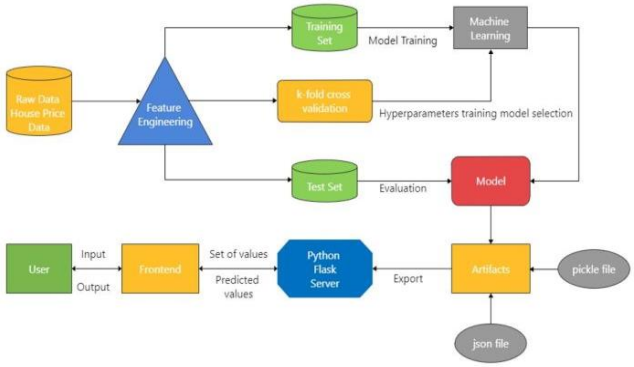

Fig 1 Architecture

Technology used-

Data Science- Data wisdom is the first stage in which we takethe dataset and will do the data drawing on it. We'll do the data drawing to make sure that it provides dependable prognostications.

Machine Learning- The gutted data is fed into the machine literacy model, and we do some of the algorithms like directretrogression, retrogression trees to test out our model.

Front End (UI) - The frontal end is principally the structure or a figure up for a website. In this to admit an information for prognosticating the price. It takes the form data entered by the stoner and executes the function which employs the prediction model to calculate the predicted price for thehouse.

IV. Data Visualization

Visualization gradually makes complex data more accessible,reasonable, and usable as shown in Fig 2 and Fig 3. Dealing with, analyzing, and transmitting this data presents good and orderly challenges for data representation. This test is addressed by the field of data science and experts known as data scientists.

In Fig 2 below shows the scatterplot of price_per_sqft vs Total Square feet of the random place from the dataset Hebbalwhere blue dot represents 2BHK and green plus represents 3BHK. This plot is with the outliers present in the dataset.

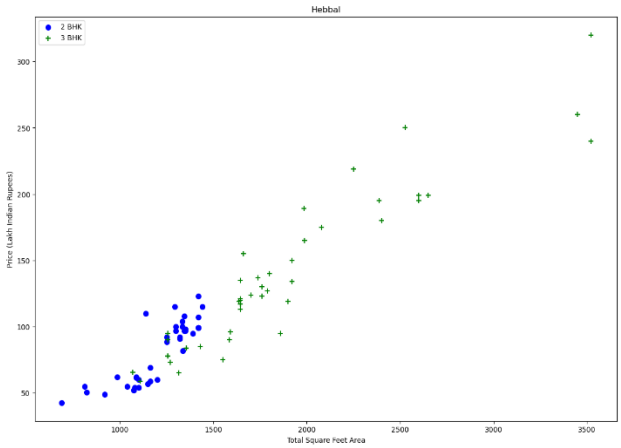

Fig 2 Price Outliers for a place (Hebbal)

In Fig 3 below shows the scatterplot of price_per_sqft vs Total Square feet of a random place from the dataset Hebbalwhere blue dot represents 2BHK and green plus represents 3BHK. This plot is after removing the outliers present in thedataset by using the function. Also in the above fig we can find one or two green plus which is 3BHK and still shows asoutlier after the function is applied. But that is a minordifference where is has come due to the place and its area where the house is present.

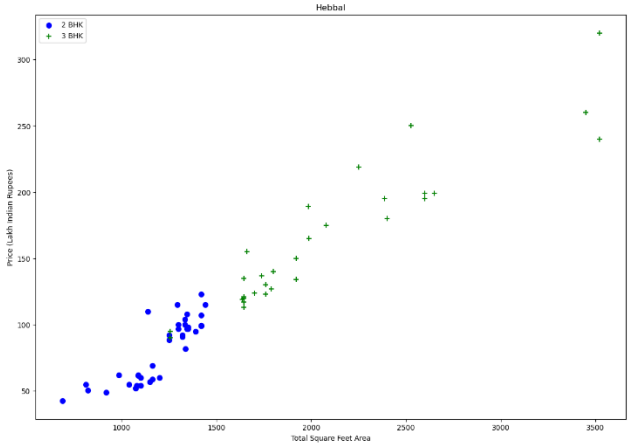

Fig 3 Price after outliers removed (Hebbal)

A correlation matrix is just a simple visual representation table that gives correlation between the different variables of the table. The matrix gives almost all the possible correlation between the variables possible. Whenever the large datasets are considered it is best option to display the summary of thedifferent patterns of the data. The correlation matrix has the value ranging between -1 to +1. Thus the positive number shows the positive links among the variables while the negative number shows the negative link between the variables that are considered. In the Fig 4 below five variables (features- total_sqft, bath, price, bhk, and price_per_sqft) areplotted and the correlation among them

is displayed. For Heatmap the Python library sns is used for data visualizationthat is based on matplotlib.

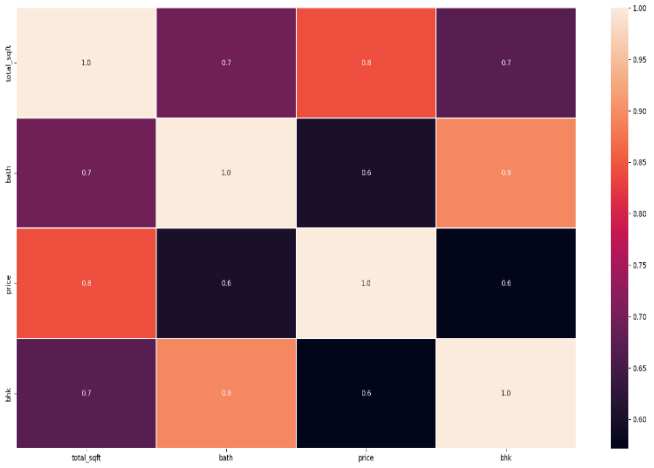

Fig 4 Correlation Matrix

Clustering is an unsupervised learning technique used to group similar objects or data points together based on their characteristics or features. The goal of clustering is to partition a dataset into groups, or clusters, where data points within the same cluster are more similar to each other than they are to data points in other clusters.

Fig. 5 appears to be a scatter plot showing price per square foot for different clusters of properties. The x-axis is labeled "Total Square Feet Area" and goes from 0 to 50,000. The y-axis is labeled "Price per Square Feet" and goes from 0 to 40,000. There are five clusters labeled 0, 1, 2, 3, and 4. Without more data points it is difficult to say anything about the relationship between price and square footage.

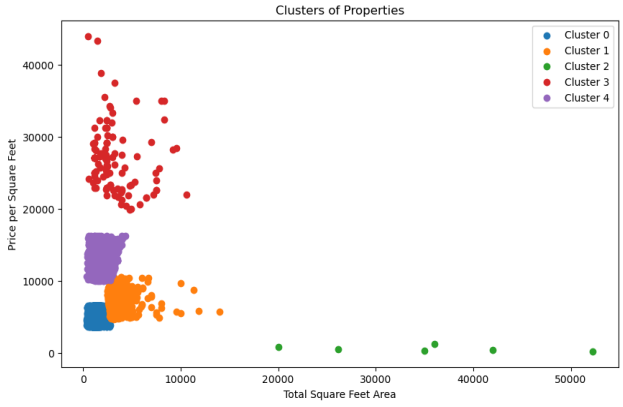

Fig 5 Scatterplot of Clustering Properties

Fig. 6 shows a histogram visualizing the distribution of prices per square foot in my dataset.

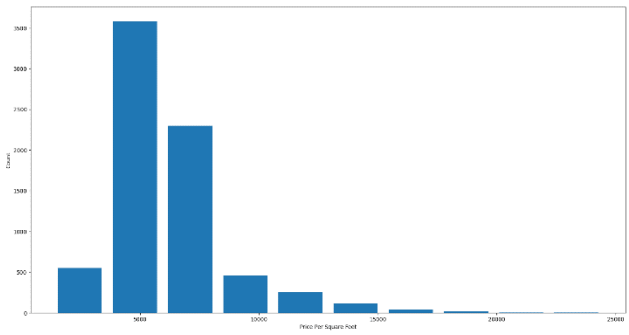

Fig. 6 Histogram

Fig.7 appears to be a bar chart showing the explained variance ratio of principal components. The x-axis is labeled "Principal components" and goes from 1 to 10. The y-axis is labeled "Explained variance ratio" and goes from 0 to 0.5.

Based on the chart, the first principal component explains the most variance in the data, followed by decreasing amounts of variance explained by subsequent components. This is a typical pattern in PCA; the first few components capture most of the important information in the data.

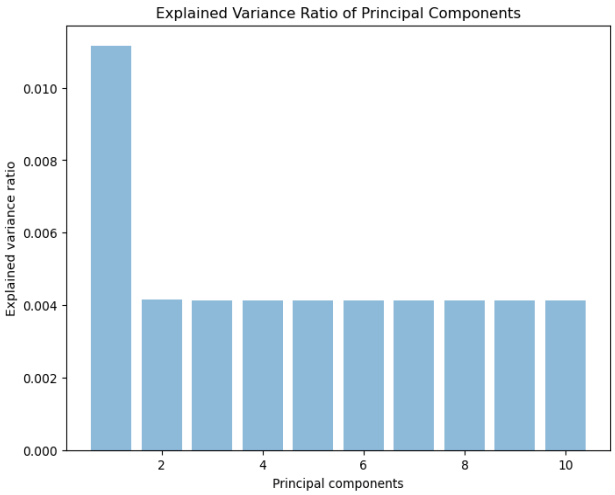

Fig. 7 Bar chart of Variance ratio of PCA

V. Comparative Analysis

In this section, we conduct a comparative analysis of the five classification algorithms namely K-Nearest Neighbors (KNN), Decision Tree, Lasso Regression, Linear Regression and Ridge Regression for the task of real estate price prediction. We evaluate the performance of each algorithm using various evaluation metrics on the test data. Performance Metrics:

	model	best_score	best_params
0	linear_regression	0.847796	{'normalize': False}
1	knn	0.690847	{'metric': 'manhattan', 'n_neighbors': 7, 'wei...
2	lasso	0.726738	{'alpha': 2, 'selection': 'cyclic'}
3	ridge	0.846798	{'alpha': 1, 'solver': 'svd'}
4	decision_tree	0.719068	{'criterion': 'mse', 'splitter': 'random'}

Fig. 1 Comparative Performance Analysis

The above Fig 1 shows the comparison between the various algorithms used to build the price prediction model, where it is found out that the Linear Regression gives the maximum accuracy of about 84.77 percent. While other algorithms KNN, Lasso, Ridge and Decision Tree gives 69.08, 72.67, 84.68 and 71.9 percent respectively.

Regression Evaluation Matrix:

1. **Mean Squared Error (MSE)**: Mean Squared Error is a commonly used metric to measure the average squared difference between the actual and predicted values in regression analysis.

$$MSE=1/n\sum(yi-y\hat{}i)^2$$

Where:

n is the number of samples.

yi is the actual (observed) target value for the i-th sample.

y^i is the predicted target value for the i-th sample.

2. Mean Absolute Error (MAE): Mean Absolute Error measures the average absolute difference between the actual and predicted values in regression analysis.

$$MSE=1/n\sum|yi-y\hat{}i|$$

Where:

n is the number of samples.

yi is the actual (observed) target value for the i-th sample.

y^i is the predicted target value for the i-th sample.

3. R-squared (Coefficient of Determination): R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model.

$$R^2 = 1 - \frac{\sum (yi-y\hat{}i)^2}{\sum (yi-\bar{y}i)^2}$$

Where:

n is the number of samples.

yi is the actual (observed) target value for the i-th sample.

y^i is the predicted target value for the i-th sample.

$\bar{y}$ is the mean of the actual target values.

Fig. 2 Shows the Regression Evaluation Matrix of each model's Mean Squared Error, Mean Absolute Error, R-squared (Coefficient of Determination).

	Algorithm	Mean Squared Error	Mean Absolute Error	R-squared
0	LinearRegression	711.056386	16.155431	0.862913
1	KNeighborsRegressor	1467.781743	22.769244	0.717022
2	Lasso	1460.219445	23.157751	0.718480
3	Ridge	732.795193	16.074983	0.858722
4	DecisionTreeRegressor	1456.610025	19.624036	0.719176

Fig. 2 Regression Evaluation Matrix

Based on the table, Decision TreeRegressor appears to have the lowest MSE (1456.61) and MAE (19.62), suggesting it might be the best performing algorithm in terms of minimizing errors.
Linear Regression has the highest R-squared (0.8629) which indicates it explains the most variance in the data. However, it also has a higher MSE compared to Decision TreeRegressor.

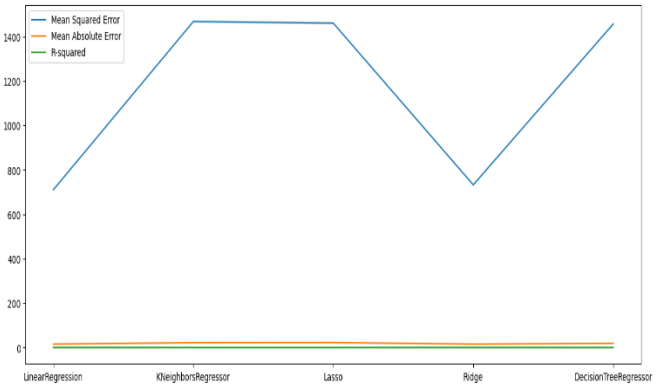

Fig 3: Model Evaluation Metrics Comparison with graph

VI. Conclusion

In this study, various machine learning algorithms are used to estimate house prices. All of the methods were described in detail, and then the dataset is taken as input, applied the various models to give out the results of the prediction. The presentation of each model was then compared based on features where it is found that linear regression gives maximum accuracy of about 84 to 85% after a proper comparison with decision tree and Lasso

regression. The correlation matrix also displays the visualization of the largerdata into compact pattern. Thus the model can work with decent efficiency giving the required features to the customer.

REFERENCES

[1] T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*, Sydney, NSW, Australia, 2018, pp. 35-42, doi: 17.1109/iCMLDE.2018.00017.

[2] M. Jain, H. Rajput, N. Garg and P. Chawla, "Prediction of House Pricing using Machine Learning with Python," *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 2020, pp. 570-574, doi: 10.1109/ICESC48915.2020.9155839.

[3] Nihar Bhagat, Ankit Mohokar and Shreyash Mane. House Price Forecasting using Data Mining. *International Journal of Computer Applications* 152(2):23-26, October 2016.

[4] N. N. Ghosalkar and S. N. Dhage, "Real Estate Value Prediction Using Linear Regression," *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Pune, India, 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697639.

[5] V. S. Rana, J. Mondal, A. Sharma and I. Kashyap, "House Price Prediction Using Optimal Regression Techniques," *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Greater Noida, India, 2020, pp. 203-208, doi: 10.1109/ICACCCN51052.2020.9362864.

[6] J. Manasa, R. Gupta and N. S. Narahari, "Machine Learning based Predicting House Prices using Regression Techniques," 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 2020, pp. 624-630, doi: 10.1109/ICIMIA48430.2020.9074952.

[7] N. S. R H, P. R, R. R. R and M. K. P, "Price Prediction of House using KNN based Lasso and Ridge Model," *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, Erode, India, 2022, pp. 1520-1527, doi:10.1109/ICSCDS53736.2022.9760832.

[8] Z. Zhang, "Decision Trees for Objective House Price Prediction," 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 2021, pp. 280-283 doi; 10.1109/MLBDBI54094.2021.00059.

[9] R. Sawant, Y. Jangid, T. Tiwari, S. Jain and A. Gupta, "Comprehensive Analysis of Housing Price Prediction in Pune Using Multi-Featured Random Forest Approach," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018,
pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697402.

[10] C. R. Madhuri, G. Anuradha and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," 2019 International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 2019, pp. 1-5, doi: 10.1109/ICSSS.2019.8882834.