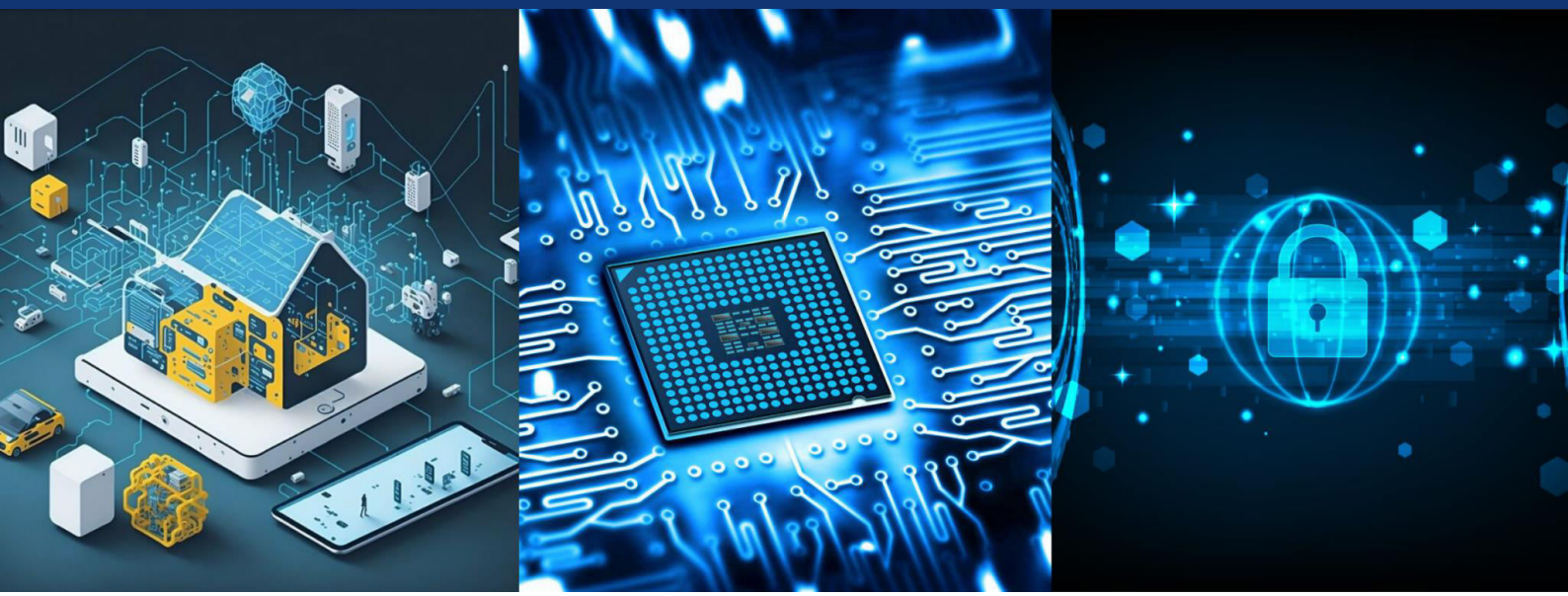
IJIRCCE

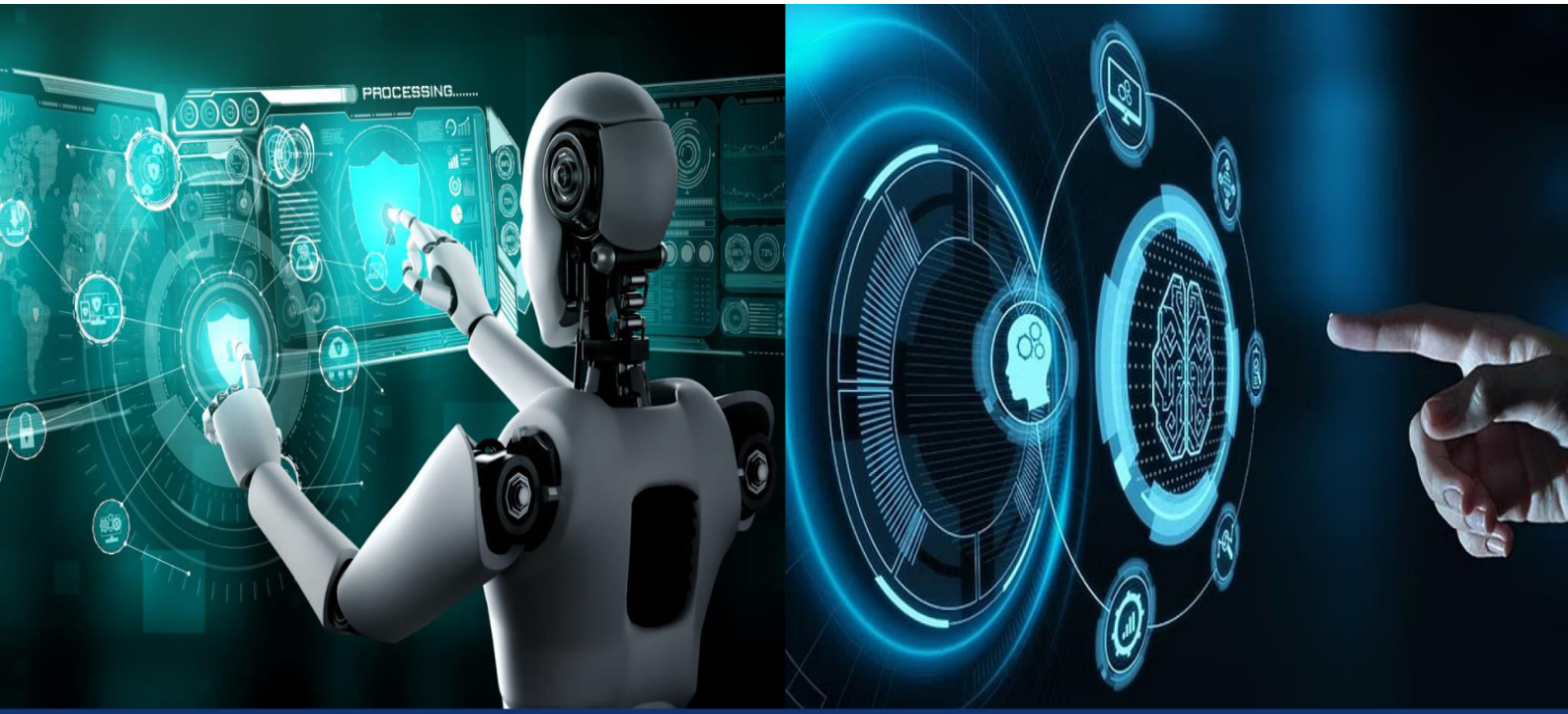
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 4, April 2025

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304311

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Data-Driven Real Estate Valuation: A Case Study Integrating Linear Regression and Random Forest Models

Preethika S K, Srinivi P, Shenbaha S

PG Student, Department of Mathematics, Dr. N.G.P. Arts and Science College, Coimbatore, India

PG Student, Department of Mathematics, Dr. N.G.P. Arts and Science College, Coimbatore, India

Assistant Professor, Department of Mathematics, Dr. N.G.P. Arts and Science College, Coimbatore, India

ABSTRACT: House price is the fundamental part of residential property. According to the 2024 census, 7.3% of the Indian GDP is contributed by the real estate sector, and one of the significant components of real estate is the residential property. People who buy, sell, and invest must be aware of the accurate house price, and if they didn't know that, then they would face a huge financial crisis. House prices are influenced by the locality, bathrooms, bedrooms, etc. This study focuses on forecasting the house price using the factors influencing the house, and this is being done by using machine learning methods such as k-means, linear regression, and random foresting. The result is analyzed using MSE, RMSE, and R^2. Finally, Random foresting shows the best result in predicting the house price.

KEYWORDS: House Price Prediction, Linear Regression, Random Forest, Machine Learning.

I. INTRODUCTION

House Price prediction is a great way to predict the price of a real estate property. Indian people are fond of making assets and one such asset creation is buying a house and it is a dream for everyone, but it shouldn't remain as a dream and people those who are planning to buy their own house, it's important to know the price range in future. Home price prediction helps people, whether the property they are going to buy is worth the money or not, and people who are going to sell; by optimizing those factors, they can set the right price range. By knowing which factors affect the price of a house, they can make a profit by increasing those factors. In olden days, predictions were made by hand, but nowadays they are done with the help of machine learning methods. Machine learning is the part of artificial intelligence that uses algorithms and statistical methods that enable computers to learn and make predictions or decisions. In this study we have used two machine learning methods to make the house price prediction. The paper consists of the following sections: literature survey, methodology, and results and discussion.

II. LITERATURE REVIEW

Recent trends have been developed in predicting the house price, and it became a need for a common person. To fulfill this need, so many remarkable works have been done using machine learning algorithms. Kevin and Hieu[12] have developed a model for predicting house prices using XGBoost, which shows significant accuracy before and after COVID-19 in Chicago. Jaykumar[9] has done a study on predicting the house price using a linear regression model with higher accuracy. Siddharth[18] made use of linear regression and gradient descent in predicting the house price. Anirudh and Achyut[3] have developed a dashboard using multiple linear regression which helps in predicting the house price. Fatbardha[6] also used linear regression along with the random forest method in order to find the house price.

III. METHODOLOGY

The study mainly focuses on retrieving a good accuracy model for an existing dataset, and to do so, the data should be preprocessed before stepping into evaluation. This section clearly mentions the data preprocessing and emphasizes data enhancement. The dataset is trained with the help of a machine learning algorithm, and then it is tested. The architecture of the paper is mentioned in the figure.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304311

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

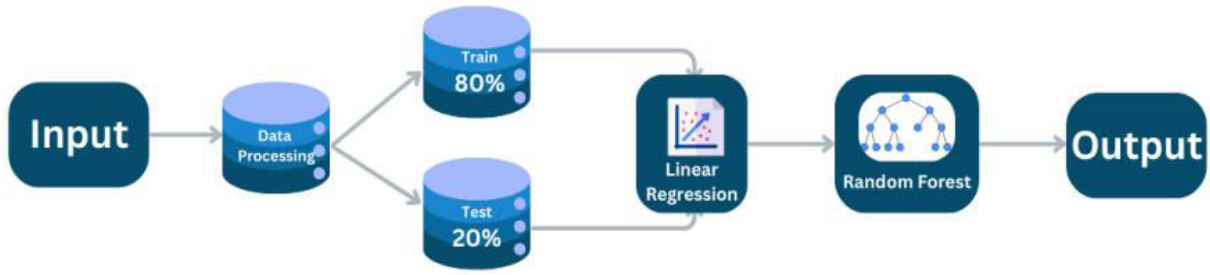


Figure: 1 Proposed Architecture

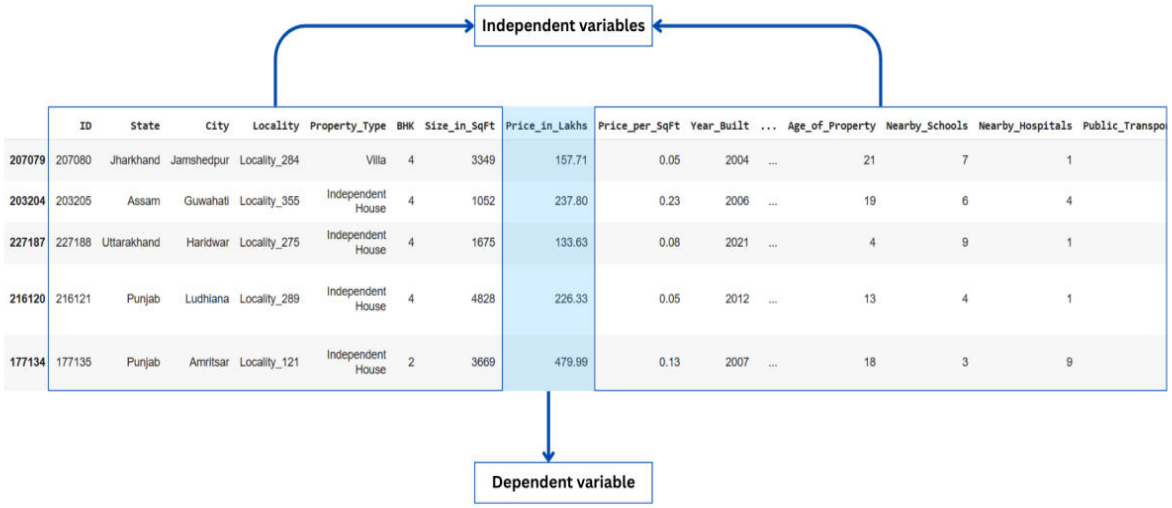
3.1 Data Collection

Data collection plays a remarkable role in predicting the house price. The data set used in this paper is taken from Kaggle which consists of 2.5 lakh rows and 23 columns. The column consists of the factors that affect the house price and the data set is in CSV file format. Here the dependent variable is the Price in Lakhs and all other remaining entries are independent variable.

Figure: 2 Dependent and Independent variables

3.2 Data cleaning

Once the data are loaded into the Colab environment, the first thing is to check for the Null value (NA) values, and it is done with the help of the method called “isnull”, and after analyzing the dataset, the column and rows need to be dropped, which can be done with the help of the "drop" method.



	ID	State	City	Locality	Property_Type	BHK	Size_in_SqFt	Price_in_Lakhs	Price_per_SqFt	Year_Built	...	Age_of_Property	Nearby_Schools	Nearby_Hospitals	Public_Transpo
207079	207080	Jharkhand	Jamshedpur	Locality_284	Villa	4	3349	157.71	0.05	2004	...	21	7	1	
203204	203205	Assam	Guwahati	Locality_355	Independent House	4	1052	237.80	0.23	2006	...	19	6	4	
227187	227188	Uttarakhand	Haridwar	Locality_275	Independent House	4	1675	133.63	0.08	2021	...	4	9	1	
216120	216121	Punjab	Ludhiana	Locality_289	Independent House	4	4828	226.33	0.05	2012	...	13	4	1	
177134	177135	Punjab	Amritsar	Locality_121	Independent House	2	3669	479.99	0.13	2007	...	18	3	9	

	State	City	BHK	Size_in_SqFt	Price_in_Lakhs	Furnished_Status	Nearby_Schools	Nearby_Hospitals	Public_Transport_Accessibility	Parking_Space	Amenities
84125	Assam	Guwahati	5	4985	493.36	Unfurnished	3	7	High	No	Playground, Gym, Pool, Clubhouse, Garden
207372	Uttarakhand	Dehradun	2	4429	86.13	Furnished	3	10	Low	No	Gym, Garden
192234	Kerala	Trivandrum	4	4046	198.84	Unfurnished	5	6	High	Yes	Gym, Garden, Playground
216637	Punjab	Ludhiana	1	2519	85.60	Semi-furnished	5	1	Low	Yes	Garden, Gym, Pool, Clubhouse, Playground
122472	Tamil Nadu	Coimbatore	1	1635	22.38	Unfurnished	5	6	Low	No	Clubhouse, Pool, Garden

Figure: 3 Dataset after cleaning

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304311

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3.3 Feature engineering

In order to perform regression, the data in the string format should be converted into the integer format. In this dataset, the column "Amenities" consists of a gym, clubhouse, garden, playground, and pool, and for each amenity a new column is being generated. In the dataset, columns like Parking_Space, Furnished_Status, and Public_Transport_Accessibility have input data as strings, but they should be converted into integers. In Parking_Space the input data are 'yes' and 'no'; these string values are converted into numerical values as 1 and 0, i.e., 1 is assigned for 'yes' and 0 is assigned for 'no'. Similarly, the column Furnished_Status consists of entries like furnished, semi-furnished and unfurnished, and they are assigned to the values 1, 0.5 and 0, respectively. In the same way for Public_Transport_Accessibility the entries are High, Medium and Low, and the values are allocated as 1, 0.5 and 0, respectively. The columns 'State' and 'City' have the entries in text. Using the one-hot encoding method, the string entries are converted into binary values. That is, the column 'State' consists of 20 state names, and 'City' consists of 42 city names, and these are converted into each separate column. For example, the State column contains an entry as 'Tamil Nadu', and it is converted as a separate column as 'State_Tamil Nadu'. Similarly, the column City contains the entry as 'Chennai', and now it is converted to a column named 'City_Chennai'.

	State	City	BHK	Size_in_SqFt	Price_in_Lakhs	Furnished_Status	Nearby_Schools	Nearby_Hospitals	Public_Transport_Accessibility	Parking_Space	clubhouse	garden	gym	playground
150894	Uttar Pradesh	Noida	3	2920	241.90	0.5	3	5	0.5	0	0	0	0	1
207908	Gujarat	Ahmedabad	5	1606	336.94	0.5	6	7	0.5	1	1	0	1	1
119240	Bihar	Gaya	2	1188	246.81	0.5	5	8	0.5	1	1	1	1	0
193290	Chhattisgarh	Raipur	4	4225	342.89	0.0	3	2	0.5	1	1	1	1	1
24223	Maharashtra	Pune	5	4096	255.86	1.0	10	7	1.0	1	1	1	1	1

Figure: 4 Dataset after using One-hot encoding

	BHK	Size_in_SqFt	Price_in_Lakhs	Furnished_Status	Nearby_Schools	Nearby_Hospitals	Public_Transport_Accessibility	Parking_Space	clubhouse	garden	...	City_Patna	City_Pune	City_Raipur
179351	5	1773	475.09	0.5	1	10	0.5	1	1	1	...	False	False	False
135686	2	2525	331.84	0.5	5	1	1.0	0	1	1	...	False	False	False
154487	4	1026	419.49	1.0	4	7	0.5	1	0	0	...	False	False	False
247711	3	3990	355.73	1.0	8	10	0.0	0	1	1	...	False	False	False
58909	3	1833	390.95	1.0	7	9	0.0	0	0	1	...	False	False	False

Figure: 5 Dataset after using One-hot encoding

3.4 Training and testing the model

The data set is divided into two parts: train and test. 80% of the dataset is for training, and 20% is for testing. The data that is used for testing is distinct from training. Machine learning algorithms are used to train and test the model. Initially the machine learning algorithm is applied in the training model, and then it is implemented in the testing model.

IV. RESULTS AND DISCUSSION

The machine learning algorithms used in this paper are linear regression and random forest. The most commonly used linear regression doesn't work properly because of the complex relationship between variables. On the other hand,

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304311

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

random forest shows a better result. The output of these models and the accuracy graph of linear regression and random forest are given below.

```
Model: Linear Regression
Mean Squared Error (MSE): 10165.917623267707
R-squared (R2): 0.49005970257278475
--------------------
Model: Random Forest
Mean Squared Error (MSE): 81.46323910357698
R-squared (R2): 0.9959136607321327
--------------------
```

Figure: 6 Output of Linear Regression and Random Forest

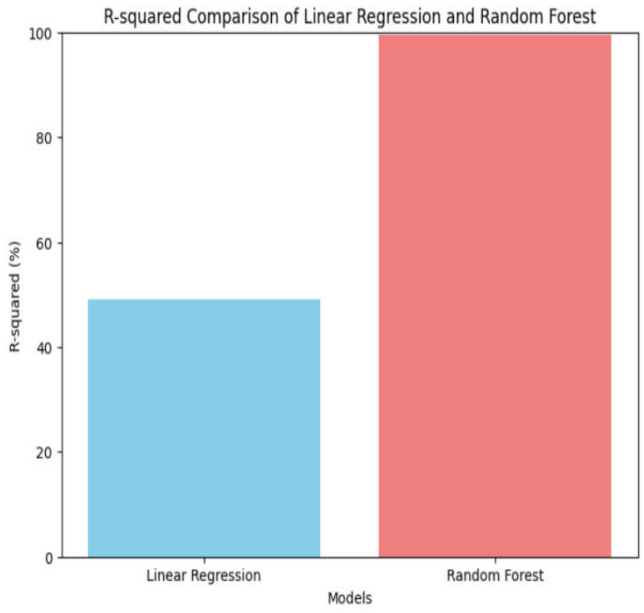


Figure: 7 Accuracy Graphs of Linear Regression and Random Forest

Data visualization plays a prominent role in data analytics because it helps to expose patterns, discover anomalies, and improve the accuracy of the data. There are so many visualization tools and graphs. In this study, the visualization is done with the help of Python, and to do so, a package called Matplotlib is used. Pyplot is used, and graphs like pie charts and line plots are used. Among the states, Karnataka has the highest price range, and Delhi has the lowest price range (Fig. 8). Among the cities, Bangalore has the highest price range, and Cuttack has the lowest price range (Fig. 9). In contribution to the house price, the number of hospitals holds the first position, and the size of the property holds the second position compared to others.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304311

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

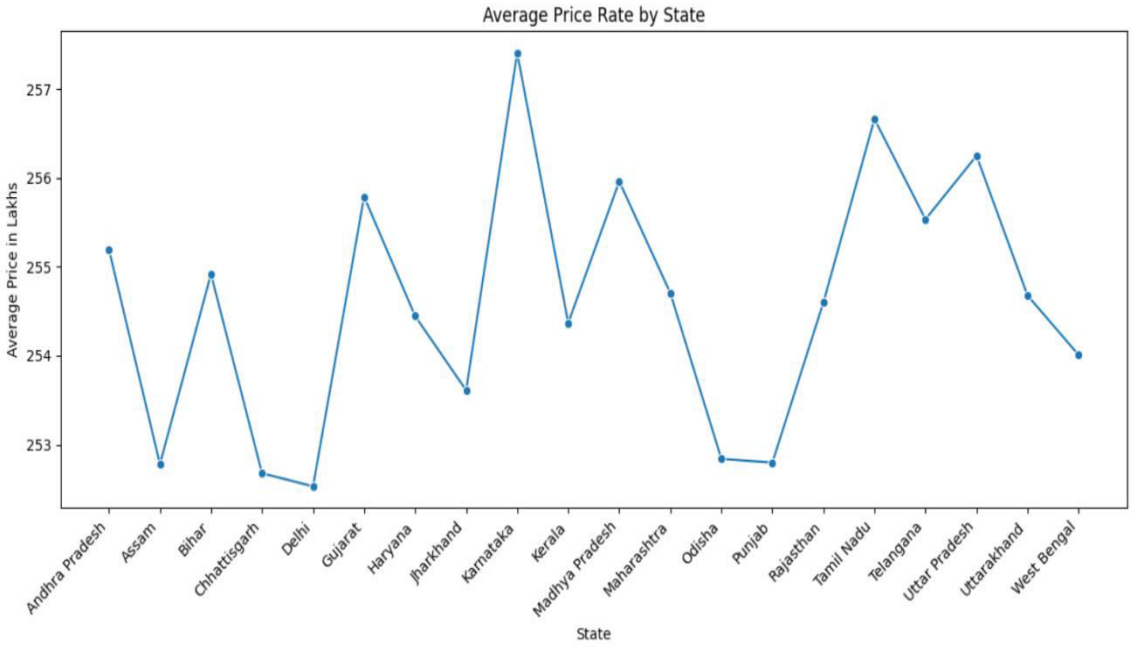


Figure: 8 State Vs Average Price in Lakhs

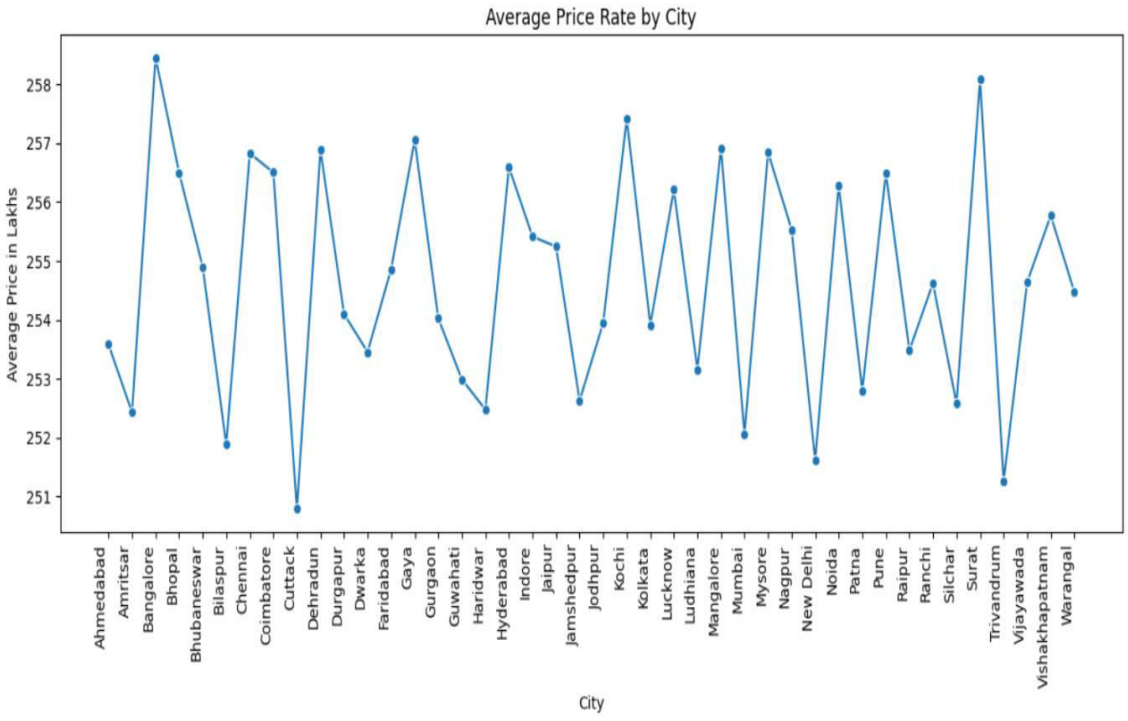


Figure: 9 City Vs Average Price in Lakhs

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI: 10.15680/IJIRCCE.2025.1304311

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

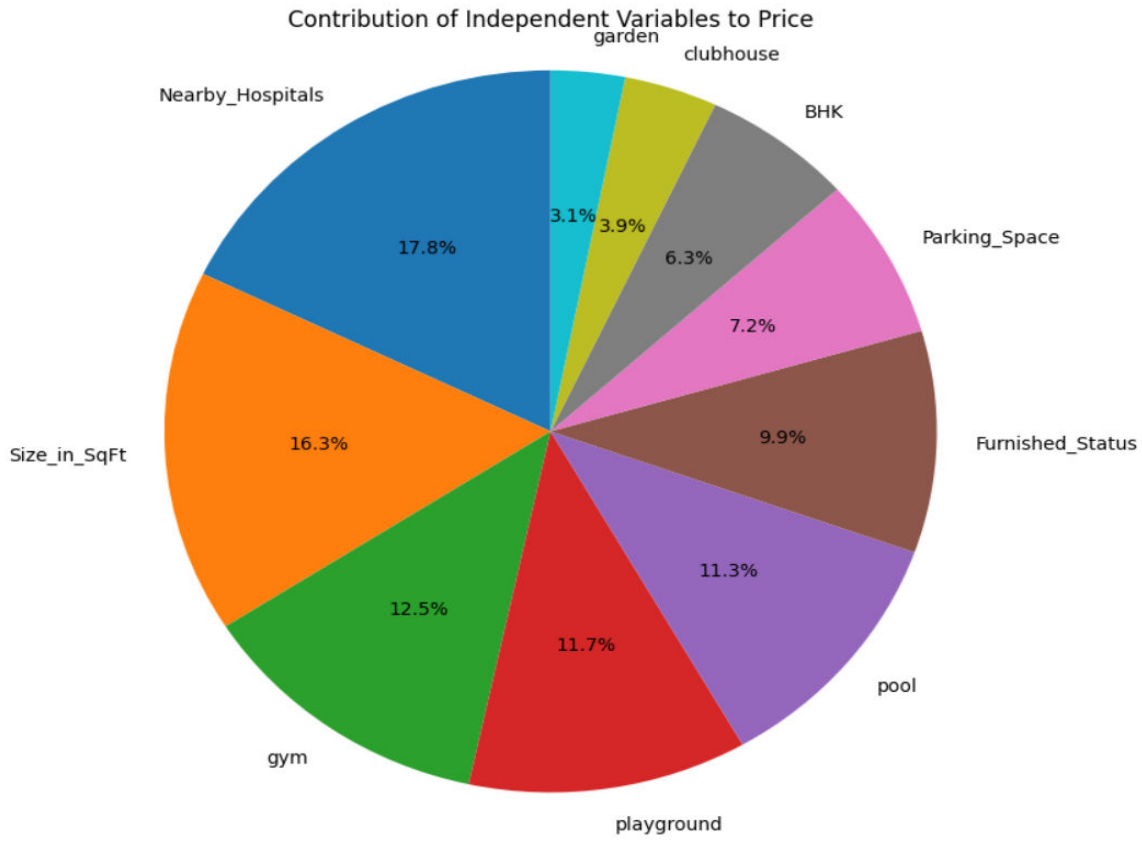


Figure: 10 Contribution of Independent Variable to Dependent variable

V. CONCLUSION

In Conclusion, this study intends to create an accurate model for House price prediction. This starts from data collection to data preprocessing. This preprocessing enhanced the model and helped in production with a better outcome. This study focuses on the two machine learning algorithms that are linear regression and random forest. It is found that random forest shows a better performance since it is one of the effective tools to perform on a large dataset. The outcome of this research will facilitate more accurate house price predictions, benefiting individuals in their decision-making processes.

REFERENCES

[1] Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, Wayan Firdaus Mahmudy, Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization, 2017, IJACSA.
[2] Anand G. Rawool, Dattatray V. Rogye, Sainath G. Rane, Dr. Vinayk A. Bharadi, House Price Prediction Using Machine Learning, 2021, IRE Journals.
[3] Anirudh Kaushal, Achyut Shankar, House Price Prediction Using Multiple Linear Regression, 2021, SSRN.
[4] Arshiya Shaikh, R. Vinayaki, G. Siddhanth, Y. Phanindra Varma, House price prediction using multivariate analysis, 2020, IJCRT.
[5] Dr. N. B. Chaphalkar, Sayali Sandbhor, Use of Artificial Intelligence in Real Property Valuation, 2013, IJET.
[6] Fatbardha Maloku, Besnik Maloku and Akansha Agarwal Dinesh Kuma, House Price Prediction Using Machine Learning and Artficial Intelligence,2024, JAICC.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[7] G. Naga Satish, Ch. V. Raghavendran, M.D. Sugnana Rao, Ch.Srinivasulu, House Price Prediction Using Machine Learning, 2019, IJITEE.
[8] Itedal Sabri Hashim Bahia, A Data Mining Model by Using ANN for Predicting Real Estate Market: Comparative Study, 2013, IJIS.
[9] Jaykumar Parekh, House Price Prediction Using Linear Regression Model, 2023, IJFMR.
[10] John W. Birch and Mark A. Sunderman, Estimating Price Paths for Residential Real Estate, 2003, JRER.
[11] Julius Olufemi Ogunleye, Predictive Data Analysis Using Linear Regression and Random Forest, 2022.
[12] Kevin Xu, Hieu Nguyen, Predicting Housing Prices and Analyzing Real Estate Markets in the Chicago Suburbs Using Machine Learning, 2022, JSR.
[13] Li Li and Kai-Hsuan Chu, Prediction of Real Estate Price Variation Based on Economic Parameters, 2017, IEEE.
[14] Madan Mohan Tito Ayyalasomayajula, Santhosh Bussa, Sailaja Ayyalasomayajula, Forecasting Home Prices Employing Machine Learning Algorithms: XGBoost, Random Forest, and Linear Regression, 2021, ESP-JETA.
[15] Ms. A. Vidhyavani, O. Bhargav Sathwik, Hemanth.T, Vishnu Vardhan Yadav. M, House Price Prediction Using Machine Learning, 2021, IJCRT.
[16] M Thamarai, S P Malarvizhi, House Price Prediction Modeling Using Machine Learning, 2020, DJIEEB.
[17] Quang Truong, Minh Nguyen, Hy Dang, Bo Mei, Housing Price Prediction via Improved Machine Learning Techniques, 2020.
[18] Siddharth Tomar, Sunny Arora, Yatharth Khansali, Rahul Yadav, Shubham Kumar, Prediction of House Price Using Linear Regression,2021,IJCRT.
[19] Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh , A hybrid regression technique for house prices prediction, 2017, IEEE.
[20] Thuraiya Mohd, Suraya Masrom, Noraini Johari, Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia, 2019, IJRTE.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details