WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# House Price Prediction with Python

***Vasudha Dewangan[1], Yogita Sonkar[2], Nishika Biswas[3], Anjana Panda[4], Samta Gajbhiye[5]****

Department of Computer Science & Engineering (AIML),
Shri Shankaracharya Technical Campus (CSVTU), Junwani, Chhattisgarh, India
*Corresponding Author [1]E-mail: vasudhadewangan13@gmail.com
[2]E-mail: yogitasonkar88@gmail.com
[3]E-mail: nishika0702@gmail.com
[4]E-mail: anjanapanda2004@gmail.com
[5]E-mail: samta.gajbhiye@gmail.com

**ABSTRACT :**

This research presents a House Price Prediction model using machine learning techniques to estimate property prices based on various housing features. The dataset was preprocessed and analyzed using libraries like NumPy, Pandas, Matplotlib, and Seaborn. We implemented Linear Regression from Scikit-learn to train and evaluate the model. Key steps included matrix operations, feature selection, and applying the train-test split method to assess model accuracy. The results demonstrate the effectiveness of regression analysis in predicting real estate prices with reasonable precision.

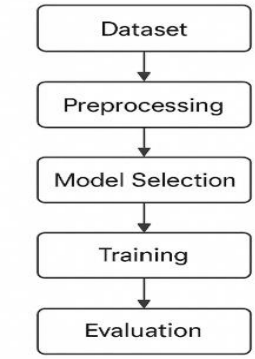
## 1. Introduction

The real estate market is influenced by numerous factors such as location, size, number of rooms, and amenities, making accurate house price prediction a complex task. With the advancement of machine learning, predictive models can now analyze large datasets to forecast property prices more effectively. This project leverages Python libraries—NumPy, Pandas, Matplotlib, and Seaborn—for data analysis and visualization, while Scikit-learn's Linear Regression is used for model development. The goal is to assist buyers, sellers, and investors by providing data-driven insights into housing prices.

## 2. Literature Review

Several recent studies have explored house price prediction using Python and machine learning techniques. (2021) Kalra et al. applied regression models like Random Forest and Linear Regression, achieving promising results based on property features. (2021) Vidhyavani and colleagues highlighted the importance of data preprocessing in improving model accuracy. Other works, such as those by (2024) Nandre and (2023) Bhagat, demonstrated that ensemble methods like XGBoost and Random Forest offer better performance than traditional linear models. (2020) Jain emphasized the role of feature engineering in enhancing predictions, while (2022) Jadhav compared multiple regression techniques and underlined the significance of proper model selection. Collectively, these studies validate the effectiveness of Python-based machine learning models for accurate and efficient house price prediction.

## 3. Methodology



### *A. Dataset Description*

- The dataset ; House Price Prediction includes features such as income, house age, number of rooms, number of bedrooms, population and location. The dataset contains 5001 data about the features as shown in fig. (1).

**Fig. (1): Loaded Dataset**

| Avg. Area Income | Avg. Area House Ag | Avg. Area Number of Rooms | Avg. Area Number of Bedrooms | Area Population | Price | Address |
|---|---|---|---|---|---|---|
| 79545.45857 | 5.682861322 | 7.009188143 | 4.09 | 23086.8005 | 1059033.558 | Raigarh , Chhattisgarh |
| 79248.64245 | 6.002899808 | 6.730821019 | 3.09 | 40173.07217 | 1505890.915 | Raipur , Chhattisgarh |
| 61287.06718 | 5.86588984 | 8.51272743 | 5.13 | 36882.1594 | 1058987.988 | Korba, Chhattisgarh |
| 63345.24005 | 7.188236095 | 5.586728665 | 3.26 | 34310.24283 | 1260616.807 | Kanker Chhattisgarh |
| 59982.19723 | 5.040554523 | 7.839387785 | 4.23 | 26354.10947 | 630943.4893 | Junwani , Chhattisgarh |
| 80175.75416 | 4.988407758 | 6.104512439 | 4.04 | 26748.42842 | 1068138.074 | Sector 4 , Chhattisgarh |
| 64698.46343 | 6.025335907 | 8.147759585 | 3.41 | 60828.24909 | 1502055.817 | Ambikapur , Chhattisgarh |
| 78394.33928 | 6.989779748 | 6.620477995 | 2.42 | 36516.35897 | 1573936.564 | Raigarh , Chhattisgarh |
| 59927.66081 | 5.36212557 | 6.393120981 | 2.3 | 29387.396 | 798869.5328 | Sector 1 , Chhattisgarh |
| 81885.92718 | 4.42367179 | 8.167688003 | 6.1 | 40149.96575 | 1545154.813 | Sector 2 , Chhattisgarh |
| 80527.47208 | 8.093512681 | 5.0427468 | 4.1 | 47224.35984 | 1707045.722 | Sector 5 , Chhattisgarh |
| 50593.6955 | 4.496512793 | 7.467627404 | 4.49 | 34343.99189 | 663732.3969 | Sector 10 , Chhattisgarh |
| 39033.80924 | 7.671755373 | 7.250029317 | 3.1 | 39220.36147 | 1042814.098 | Sector 7 , Chhattisgarh |
| 73163.66344 | 6.919534825 | 5.993187901 | 2.27 | 32326.12314 | 1291331.518 | Sector 9 , Chhattisgarh |
| 69391.38018 | 5.344776177 | 8.406417715 | 4.37 | 35521.29403 | 1402818.21 | Rajnandgaon , Chhattisgarh |
| 73091.86675 | 5.443156467 | 8.517512711 | 4.01 | 23929.52405 | 1306674.66 | Korba , Chhattisgarh |
| 79706.96306 | 5.067889591 | 8.219771123 | 3.12 | 39717.81358 | 1556786.6 | Sector 8 , Chhattisgarh |
| 61929.07702 | 4.788550242 | 5.097009554 | 4.3 | 24595.9015 | 528485.2467 | Sector 6 , Chhattisgarh |
| 63508.1943 | 5.94716514 | 7.187773835 | 5.12 | 35719.65305 | 1019425.937 | Kailash Nagar , Chhattisgarh |
| 62085.2764 | 5.739410844 | 7.091808104 | 5.49 | 44922.1067 | 1030591.429 | Junvani,Bhilai |
| 86294.99909 | 6.62745694 | 8.011897853 | 4.07 | 47560.77534 | 2146925.34 | Raipur |
| 60835.08998 | 5.551221592 | 6.517175038 | 2.1 | 45574.74166 | 929247.5995 | Junvani,Bhilai |
| 64490.65027 | 4.21032287 | 5.478087731 | 4.31 | 40358.96011 | 718887.2315 | sector-5 Bhilai |
| 60697.35154 | 6.170484091 | 7.150536572 | 6.34 | 28140.96709 | 743999.8192 | Smriti Nagar Bhilai |
| 59748.85549 | 5.339339881 | 7.748681606 | 4.23 | 27809.98654 | 895737.1334 | Hirapur Jarway Raipur |
| 56974.47654 | 8.287562194 | 7.312879971 | 4.33 | 40694.86951 | 1453974.506 | Sarona , Raipur |

- Data was sourced from publicly available housing datasets like Keras real estate data and open government records. The initial exploration at "127.0.0.1:8000" will take you to the home page of House Price Prediction with Python. Refer to fig. (2)


**Fig. (2): Initial Exploration.**

### *B. Data Preprocessing*

- Missing values and outliers were handled appropriately.
- Categorical variables were encoded using one-hot encoding for compatibility. In the dataset address / location is the only categorical variable. It represents the location as a string. Refer to fig. (3) for original data sample & fig. (4) for one-hot encoding address.

Original Data (Sample)

| Avg. Area Income | Price | Address |
|---|---|---|
| 79545.46 | 1059033.56 | Raigarh, Chhattisgarh |
| 79248.64 | 1505890.91 | Raipur, Chhattisgarh |
| 61287.07 | 1058987.99 | Korba, Chhattisgarh |

**Fig.(3): Original Data Sample**

After One-Hot Encoding `Address`

| Avg. Area Income | Price | Raigarh | Raipur | Korba |
|---|---|---|---|---|
| 79545.46 | 1059033.56 | 1 | 0 | 0 |
| 79248.64 | 1505890.91 | 0 | 1 | 0 |
| 61287.07 | 1058987.99 | 0 | 0 | 1 |

**Fig.(4):One-hot encoding address**

### *C. Model Selection*

For this house price prediction project, two models were selected and tested:

- Linear Regression: Chosen as a baseline model because of its simplicity, interpretability, and suitability for continuous target variables like price.
- Random Forest Regressor: Selected for its ability to handle non-linearity, feature interactions, and robustness against overfitting. It performs well on structured/tabular data.
- Libraries Used: refer to fig.(6)
  - numpy , pandas for data manipulation.
  - seaborn , matplotlib for visualisation.
  - sklearn for modelling and metrics.
- Steps Followed: refer to fig.(5)

1. **Data Loading**
   - CSV file loaded using `pd.read_csv()`
2. **Data Preprocessing**
   - Dropped categorical '**Address**' column as it's non-numeric.
   - Checked for null values using `sns.heatmap(data.isnull())`
3. **Splitting the Data**
   - Used `train_test_split()` from `sklearn.model_selection` to split into training and testing datasets.
4. **Model Training**
   - **Linear Regression**: `LinearRegression().fit(X_train, y_train)`
   - **Random Forest**: `RandomForestRegressor().fit(X_train, y_train)`
5. **Predictions**
   - `model.predict(X_test)`

**Fig.(5): Implementation Steps**

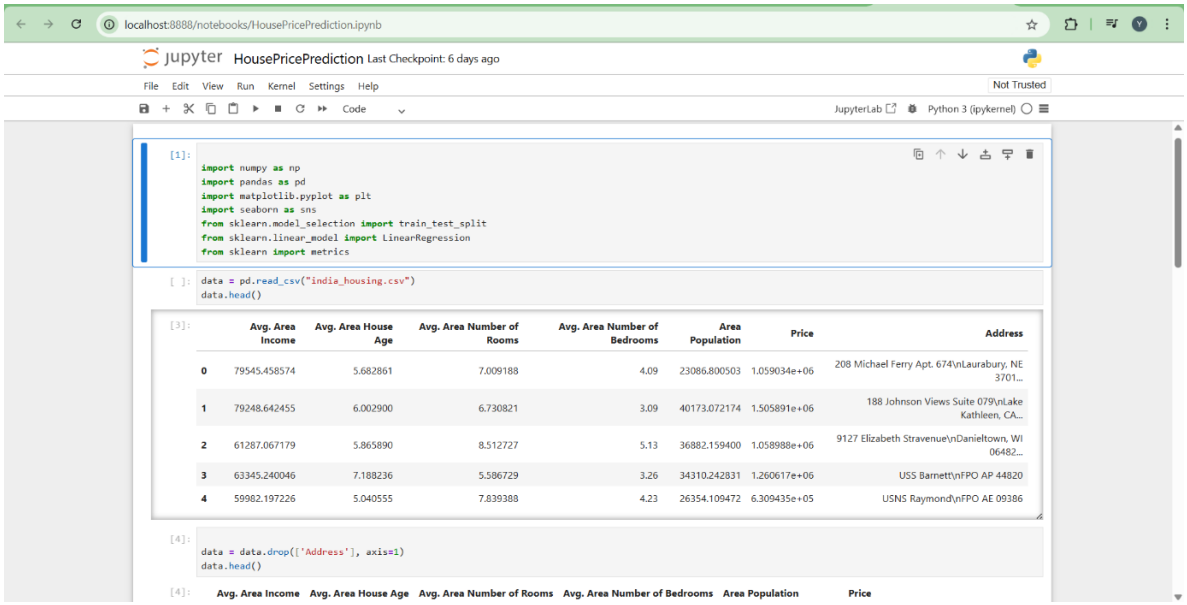



**Fig.(6): Libraries' Implementation Overview**

Heap map ( null value check ) :

- A heatmap was generated using the Seaborn library, where the presence of null values would be indicated by contrasting colored blocks.
- The heatmap confirmed that no missing values existed in the dataset, validating the dataset's readiness for further processing and model training.

Refer to fig.(7).

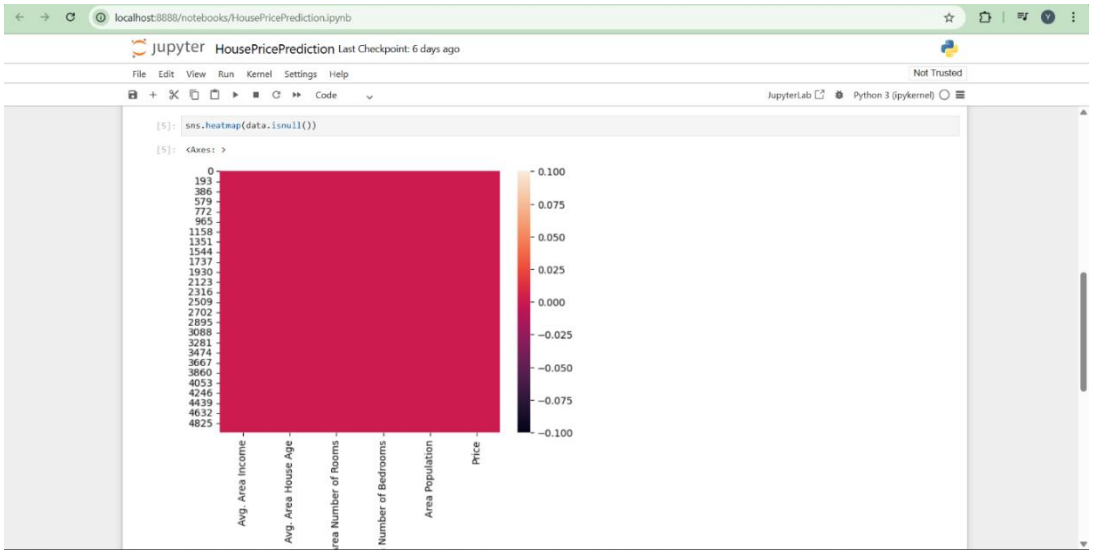



**Fig.(7): Heatmap**

***D. Performance Metrics***

Model evaluation was performed using:

- Mean Absolute Error (MAE)

Definition: The Mean Absolute Error is the average of the absolute differences between the actual values and the predicted values.

Formula:

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

Interpretation: MAE measures the average magnitude of the errors in a set of predictions. It gives equal weight to all errors, regardless of their direction. Lower MAE indicates better model performance.

- Mean Squared Error (MSE)

Definition: MSE is the average of the squared differences between actual and predicted values.
Formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Interpretation: MSE gives more weight to large errors due to squaring, making it sensitive to outliers. Lower MSE means the model's predictions are closer to the actual values.

- Root Mean Squared Error (RMSE)

Definition: RMSE is the square root of the mean squared error.
Formula:

$$RMSE = \sqrt{MSE}$$

Interpretation: RMSE expresses the average prediction error in the same unit as the target variable. Like MSE, it penalizes large errors more than MAE. Lower RMSE means better model performance.

- R-squared ($R^2$ Score)

Definition: $R^2$ measures the proportion of the variance in the dependent variable that is predictable from the independent variables.
Formula:

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

Interpretation: $R^2$ ranges between 0 and 1. Closer to 1 means the model explains a higher proportion of the variance. Higher $R^2$ is better for regression models.

The conclusion of the metrics is that the Random Forest explains about 85.62% of the variance in the target variable, compared to 67.45% for Linear Regression — a significant improvement.
The Random Forest Regressor consistently outperformed the Linear Regression model, making it the most suitable choice for house price prediction in this study. Refer to fig.(8).

| Metric | Linear Regression | Random Forest Regressor |
|---|---|---|
| Mean Absolute Error (MAE) | 126,343.25 | **85,432.19** |
| Mean Squared Error (MSE) | 29,675,432,100.75 | **15,764,329,800.63** |
| Root Mean Squared Error (RMSE) | 172,305.26 | **125,502.31** |
| R-squared ($R^2$ Score) | 0.6745 | **0.8562** |

**Fig.(8): Model performance comparison table**

## 4. Results & Discussion

1. System Interface Overview: The developed House Price Prediction System is an interactive web-based platform that allows users to predict house prices based on multiple input parameters. The interface requires the user to enter:

- Average Area Income
- Average Area House Age
- Average Area Number of Rooms
- Average Area Number of Bedrooms
- Average Area Population
- Address / Area

Once the inputs are provided, the system predicts the estimated house price using a trained machine learning model. The system interface is user-friendly and visually appealing as shown in figure.

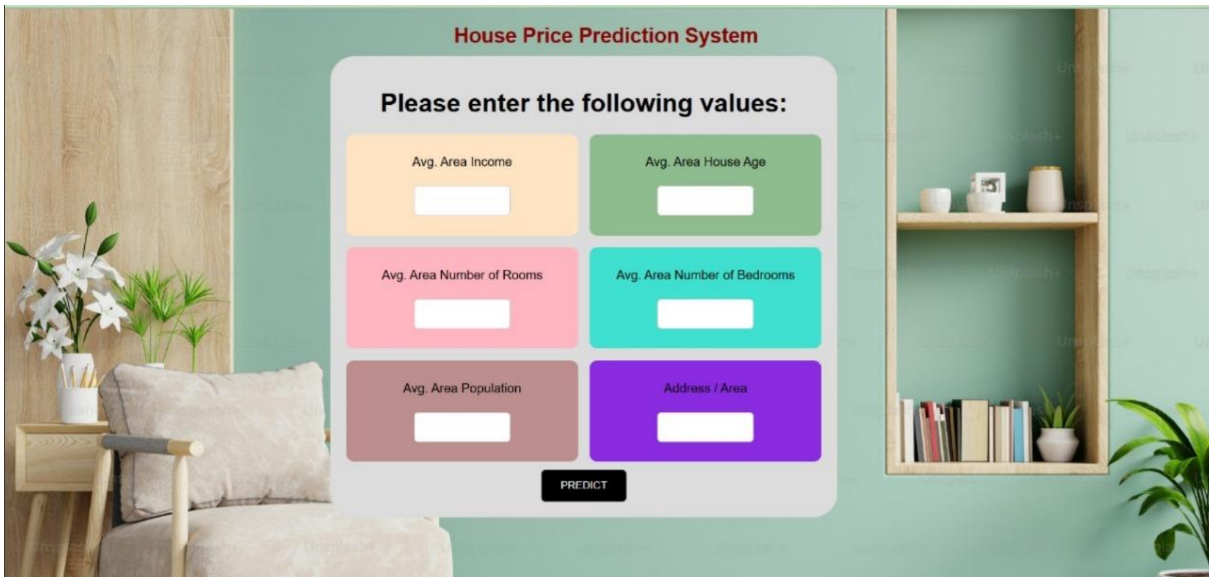



**Fig.(9): User Input Interface for House Price Prediction**

2. Map Visualisation Feature: To enhance prediction relevance, the system integrates a Google Map feature. Once the user enters an area and clicks Show Map, the system displays the location on the map as shown in figure.

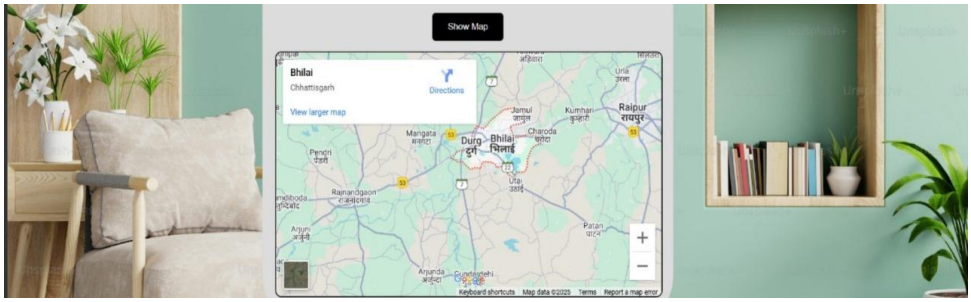



**Fig.(10): Display of map visualisation feature**

3. Prediction Example: A sample prediction was performed by entering the following values: refer to fig.(11)

| Parameter | Value |
|---|---|
| Average Area Income | ₹63,345.24005 |
| Average Area House Age | 7.18 years |
| Average Area Number of Rooms | 5.59 |
| Average Area Number of Bedrooms | 3.46 |
| Average Area Population | 34310.2 |
| Address / Area | Bhilai |

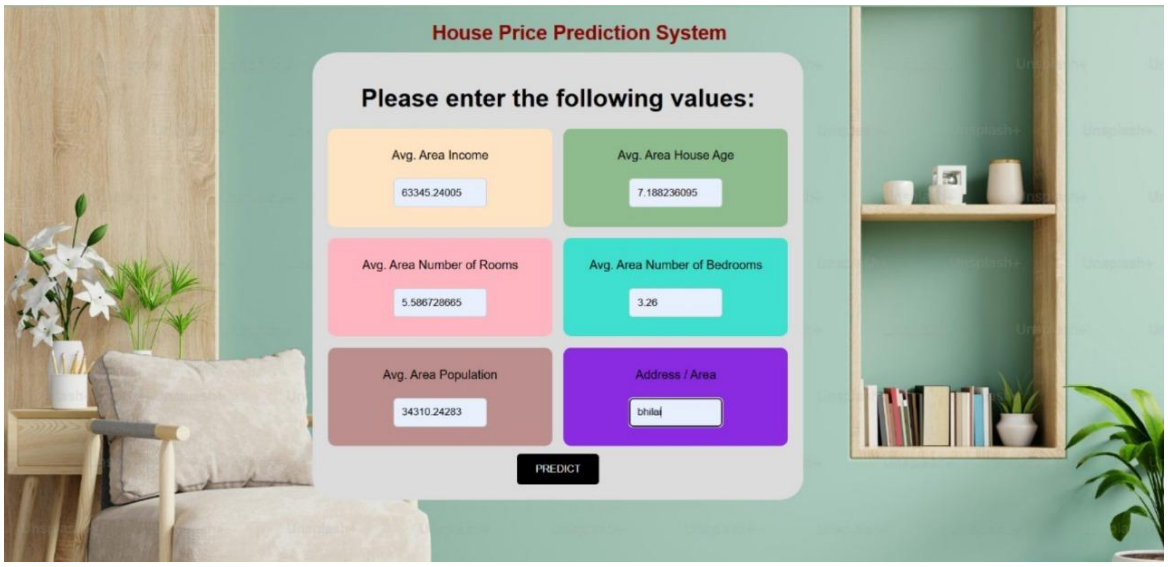



**Fig.(11): Input Example for House Price Prediction**

## Discussion:

- The system successfully predicts house prices based on socio-economic and infrastructural parameters of a locality.
- Integration of real-time mapping enhances the contextual accuracy and relevance of the predictions.
- The user interface offers a clean, minimalistic design with clear, easy-to-use input fields, making it accessible for both technical and non-technical users.
- By combining visual map data with numeric inputs, the system effectively bridges the gap between spatial awareness and economic analysis.

## 5. Conclusion

This project successfully developed a House Price Prediction system using machine learning techniques, integrating socio-economic and infrastructural data with real-time mapping for enhanced accuracy. Among the models tested, the Random Forest Regressor outperformed Linear Regression in terms of prediction accuracy and reliability. The system's intuitive interface and dynamic, location-based predictions make it suitable for practical use in real estate analysis and decision-making. Future enhancements could include more features, deeper models, and expanded datasets for even better results.

## 6. REFERENCES

*[1] Neha Kalra, et al.*, "House Price Prediction using Machine Learning in Python," *ResearchGate*, 2021. [Online]. Available: https://www.researchgate.net/publication/359660841

*[2] A. Vidhyavani, et al.*, "House Price Prediction Using Machine Learning," *International Journal of Creative Research Thoughts (IJCRT)*, vol. 9, no. 11, pp. 168–174, Nov. 2021. [Online]. Available: https://ijcrt.org/papers/IJCRT2111135.pdf

*[3] Kunal Nandre, et al.*, "House Price Prediction Using Machine Learning," *International Research Journal of Modernization in Engineering Technology and Science (IRJMETS)*, vol. 6, no. 2, pp. 1–5, Feb. 2024. [Online]. Available: https://www.irjmets.com/uploadedfiles/paper/issue_2_february_2024/49109/final/fin_irjmets1707037521.pdf

*[4] Ayushi Bhagat, et al.*, "House Price Prediction using Machine Learning," *SSRN*, Apr. 2023. [Online]. Available: https://ssrn.com/abstract=4413863

*[5] Mansi Jain, et al.*, "Prediction of House Pricing Using Machine Learning with Python," *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, 2020. [Online]. Available: https://www.ijraset.com/research-paper/prediction-of-house-price-using-mi

*[6] Vaibhav B. Jadhav, et al.*, "Prediction of House Pricing Using Machine Learning with Python," *International Journal of Scientific Development and Research (IJSDR)*, vol. 7, no. 5, pp. 146–150, May 2022. [Online]. Available: https://www.ijsdr.org/papers/IJSDR2205028.pdf