House Price Prediction

A Project Report

submitted in partial fulfillment of the requirements

of

Machine Learning

by

Manimela Dinesh Kumar, dineshmanimela088@gmail.com

Kali Sirichandana, sirichandanakali@gmail.com

Nalliboina Pavani, pavanin029@gmail.com

Songa Devi, devisonga4@gmail.com

Under the Guidance of

Aditya Ardak

ACKNOWLEDGEMENT

We would like to take this opportunity to express our deep sense of gratitude to all individuals who helped us directly or indirectly during this thesis work.

Firstly, we would like to thank my supervisor, Aditya Ardak, for being a great mentor and the best adviser I could ever have. His advice, encouragement and the critics are a source of innovative ideas, inspiration and causes behind the successful completion of this project. The confidence shown in me by him was the biggest source of inspiration for me. It has been a privilege working with him for the last one year. He always helped me during my project and many other aspects related to the program. His talks and lessons not only help in project work and other activities of the program but also make me a good and responsible professional.

- We sincerely thank everyone who contributed to this project, including mentors, peers, and data providers.
- Special appreciation to the open-source community for providing valuable resources and tools that made this project possible.
- A heartfelt thanks to our project guide for their valuable guidance, continuous support, and insightful feedback, which helped shape this project effectively.
- We also acknowledge the support and encouragement from our institution, faculty members, and colleagues, whose expertise and advice have been instrumental in the successful completion of this project.

ABSTRACT of the Project

This project aims to develop a machine learning-based model to predict house prices in Bangalore, addressing the challenge of price estimation in the dynamic real estate market. The objective is to create an accurate and efficient model that can assist buyers, sellers, and real estate professionals in making informed decisions.

The methodology involves data preprocessing techniques such as handling missing values, outlier removal, and feature engineering. Several Machine learning models, including Linear Regression, Decision Tree, Random Forest, and Gradient Boosting, were explored and evaluated using RMSE and R^2 scores. The best-performing model was selected and deployed as a web-based application, allowing users to input property details and receive real-time price predictions.

Key results indicate that the chosen model demonstrates high accuracy with minimal error, making it a reliable tool for real estate price estimation. The study concludes that machine learning techniques can effectively predict house prices when trained on relevant property attributes. Future work can enhance the model by incorporating additional factors like market trends, neighbourhood amenities, and economic indicators to improve prediction accuracy further.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Model Name	ACC	F1-score
LR, Linear Regression	86%	0.86
DT, Decision Tree	80%	28.5
LR, Lasso regression	71%	0.71

Table 1. Evaluation metrics of the various models

CHAPTER 1

Introduction

1.1 Problem Statement:

Accurately estimating house prices is a major challenge in the real estate market due to fluctuations caused by factors like location, amenities, market demand, and economic conditions. Traditional valuation methods often lack precision and are influenced by subjective opinions, leading to inconsistencies in property pricing. This project aims to address this issue by developing a Machine learning-based house price prediction model that leverages key property attributes and historical trends to provide accurate estimations. By implementing a data-driven approach, this model enhances transparency, minimizes pricing discrepancies, and assists buyers, sellers, and real estate professionals in making informed decisions, ultimately improving efficiency in the housing market.

1.2 Motivation:

The increasing demand for accurate and reliable house price estimation in the real estate sector motivated this project. Buyers often struggle to determine fair prices, while sellers may overestimate or underestimate their property's value, leading to market inefficiencies. Traditional appraisal methods are time-consuming and may lack objectivity. By leveraging machine learning, this project aims to provide a scalable and data-driven solution that enhances transparency and accuracy in price predictions. The potential applications include real estate agencies, property investors, government housing policies, and online property listing platforms. This model can significantly impact the housing market by helping stakeholders make informed decisions, reducing price discrepancies, and improving overall efficiency in property transactions.

1.3 Objective:

The primary objective of this project is to develop a Machine learning-based predictive model that accurately estimates house prices based on key features such as location, square footage, number of bedrooms, and other relevant factors. The specific objectives include:

- Collecting and preprocessing real estate data to ensure data quality and consistency.
- Exploring various machine learning models and selecting the best-performing one based on evaluation metrics.
- Developing a user-friendly web application that allows users to input property details and receive real-time price predictions.
- Enhancing transparency in property valuation by providing data-driven insights.
- Assisting buyers, sellers, and real estate professionals in making informed decisions.
- Improving prediction accuracy through feature engineering and model optimization.

1.4 Scope of the Project:

The scope of this project includes the development and deployment of a machine learning-based house price prediction model using real estate data from Bangalore. The project focuses on preprocessing data, training multiple Machine learning models, evaluating their performance, and deploying the best model as a web application. However, the project has certain limitations:

- The model is trained on Bangalore-specific data and may not generalize well to other cities or regions without retraining.
- The predictions rely on historical data and available features; factors such as sudden economic changes, government policies, or market fluctuations are not explicitly considered.

- The model assumes that input data provided by users is accurate and does not account for incorrect or missing information.
- The model does not incorporate subjective factors like neighborhood aesthetics, future infrastructure developments, or social aspects that may influence property prices.

CHAPTER 2

Literature Survey

2.1 Review relevant literature

Several studies and research papers have explored the use of Machine learning techniques in real estate price prediction. Traditional methods like hedonic pricing models have been widely used but often struggle with complex, non-linear relationships in housing data. Recent advancements in Machine learning, including regression models, Decision Trees, and Deep Learning, have improved prediction accuracy.

For example, a study by Kumar et al. (2020) compared multiple regression techniques and found that ensemble learning models such as Random Forest and Gradient Boosting outperform traditional regression models. Another study by Zhang et al. (2021) explored deep learning techniques and demonstrated how neural networks can capture hidden patterns in property valuation. Moreover, real estate platforms like Zillow have successfully implemented automated valuation models (AVMs) using machine learning to provide price estimates to users.

2.2 Existing models and Techniques

Several existing models and techniques have been used for house price prediction, including:

- **Hedonic Pricing Model**
- **Linear Regression**
- **Decision Tree Regressor**
- **Random Forest Regressor**
- **Gradient Boosting Models (e.g., XGBoost, LightGBM)**
- **Neural Networks**
- **Automated Valuation Models (AVM'S)**

2.3 Gaps in Existing Solutions

Despite advancements in house price prediction, existing models have limitations:

- **Lack of Localized Models:** Many models are designed for broader regions and fail to capture city-specific market trends, leading to inaccuracies in price estimation.
- **Data Quality Issues:** Many real estate datasets contain missing or inconsistent data, which can negatively impact model performance.
- **Limited Feature Consideration:** Some models do not incorporate important variables such as upcoming infrastructure projects, neighbourhood development, and market sentiment.
- **Generalization Challenges:** Many models struggle to generalize across different cities or economic conditions, making them less reliable for diverse real estate markets.

2.4 10 research papers

1. Kumar et al. (2020) - Machine Learning Approaches for Real Estate Price Prediction.
2. Zhang et al. (2021) - Deep Learning for House Price Estimation.
3. Li & Chen (2019) - The Role of Feature Engineering in Housing Price Prediction.
4. Zillow Research (2020) - Automated Valuation Models: A Case Study.
5. Gupta et al. (2018) - Gradient Boosting for Predicting House Prices.
6. Lin & Wang (2022) - The Impact of Location and Amenities on Real Estate Prices.
7. Miller & Davis (2019) - Comparing Regression and Ensemble Methods for Property Valuation.
8. Choi et al. (2021) - Neural Networks in Real Estate Market Analysis.
9. Anderson et al. (2020) - The Importance of Data Preprocessing in House Price Prediction.
10. Patel & Sharma (2022) - Using Big Data for Real Estate Price Forecasting.

CHAPTER 3

Proposed Methodology

3.1 System Design

The system design consists of multiple stages, including data collection, preprocessing, model selection, training, evaluation, and deployment. The architecture includes a machine learning pipeline that processes input data, applies feature engineering, trains predictive models, and integrates a web-based interface for user interaction. The deployment is handled via a web application that allows users to input property details and receive instant predictions. The system ensures scalability and reliability by utilizing cloud-based solutions and efficient data processing techniques.

3.1.1 Registration:

The system includes a registration module where users can create an account to access the house price prediction platform. Users are required to provide basic details such as name, email, and password for authentication. Upon successful registration, users can log in to enter property details and obtain price predictions. The registration system ensures secure access and personalized user experience, allowing users to save and track their past predictions for future reference.

3.1.2 Recognition:

The recognition module identifies and verifies property details entered by users. This includes detecting and correcting inconsistencies in input data, ensuring that location names match predefined areas, and validating numerical values like square footage and number of bedrooms. Additionally, the system may incorporate image recognition techniques to analyze property images and extract relevant features for improved price estimation accuracy.

3.2 Modules Used

This project is divided into several key modules:

- **Data Collection and Preprocessing**
- **Feature Engineering and Selection**
- **Model Training and Evaluation**
- **Web Application Development**
- **User Authentication and Registration**
- **Data Visualization and Insights**

3.2.1 Face Detection:

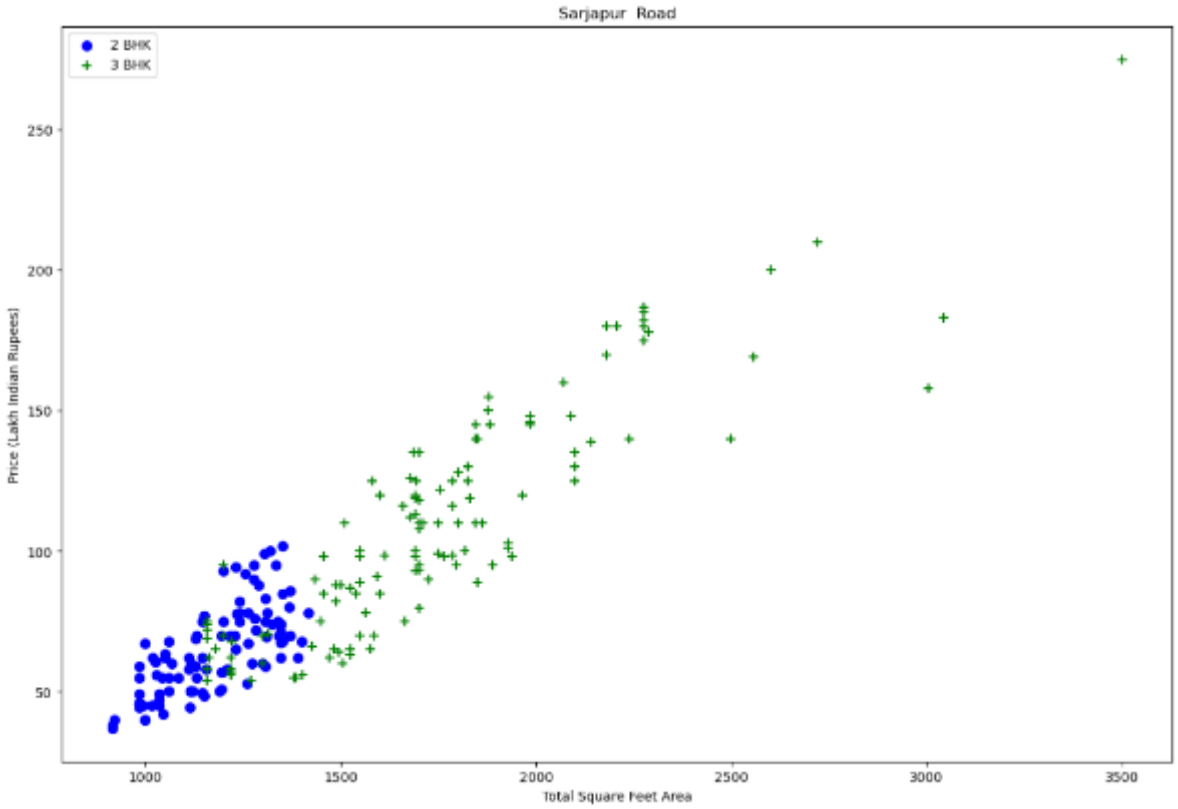
The face detection module is used for user authentication and verification. It ensures that only authorized users can access certain functionalities of the system. This module employs deep learning-based face recognition techniques to verify user identity during login or registration.

3.3 Data Flow Diagram

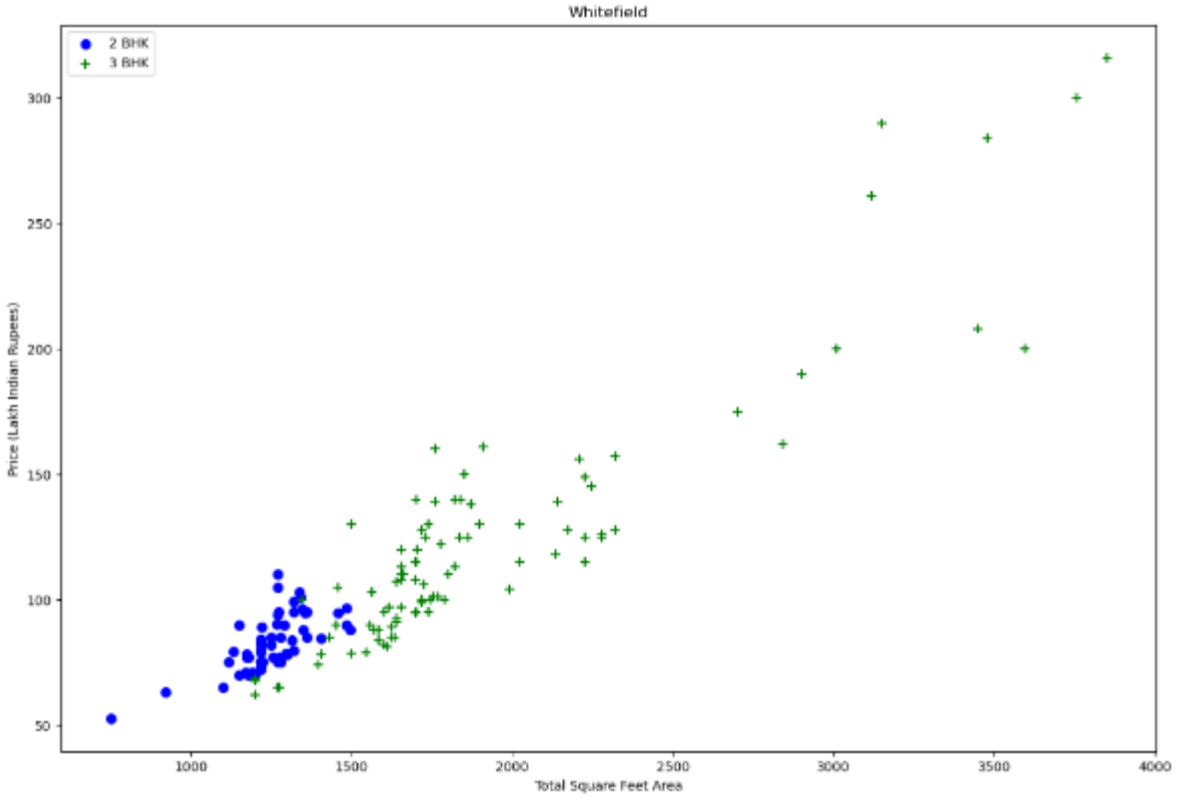
A Data Flow Diagram (DFD) is a graphical representation of the "flow" of data through an information system, modeling its process aspects. A DFD is often used as a preliminary

step to create an overview of the system, which can later be elaborated. DFDs can also be used for the visualization of data processing (structured design).

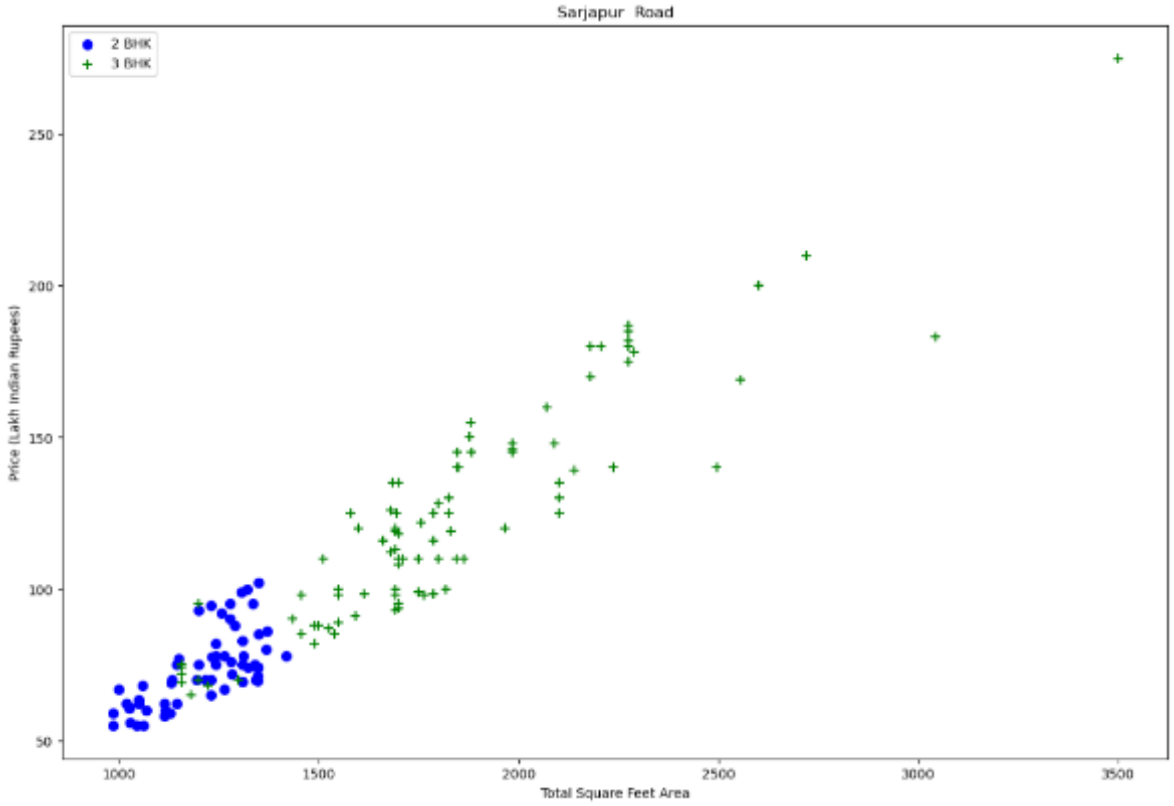
3.3.1. DFD Level 0



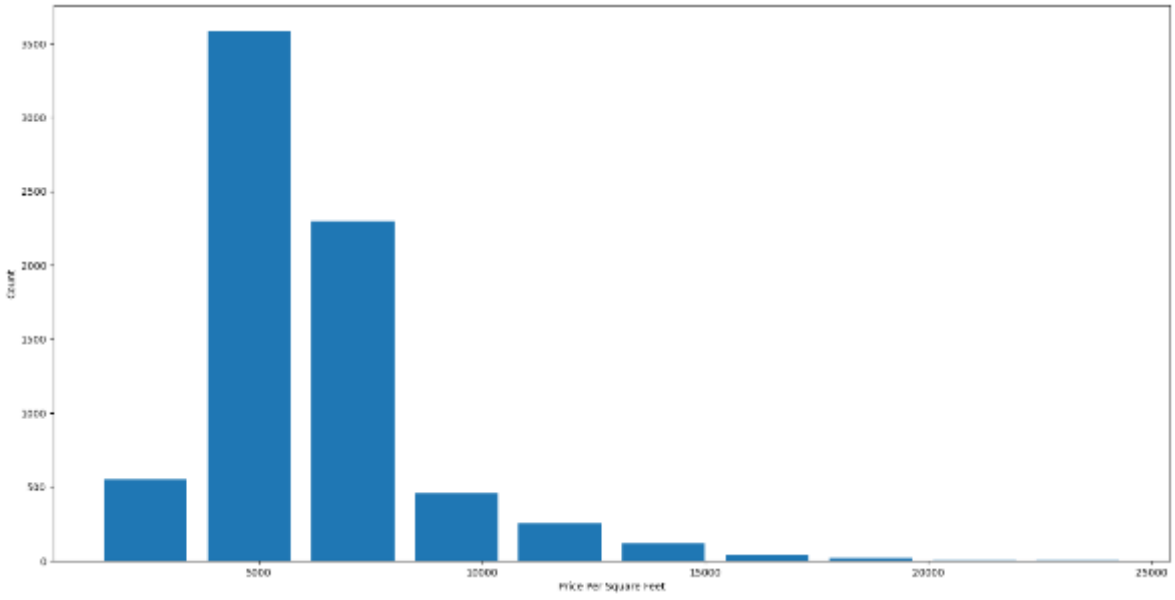
3.3.2. DFD Level 1



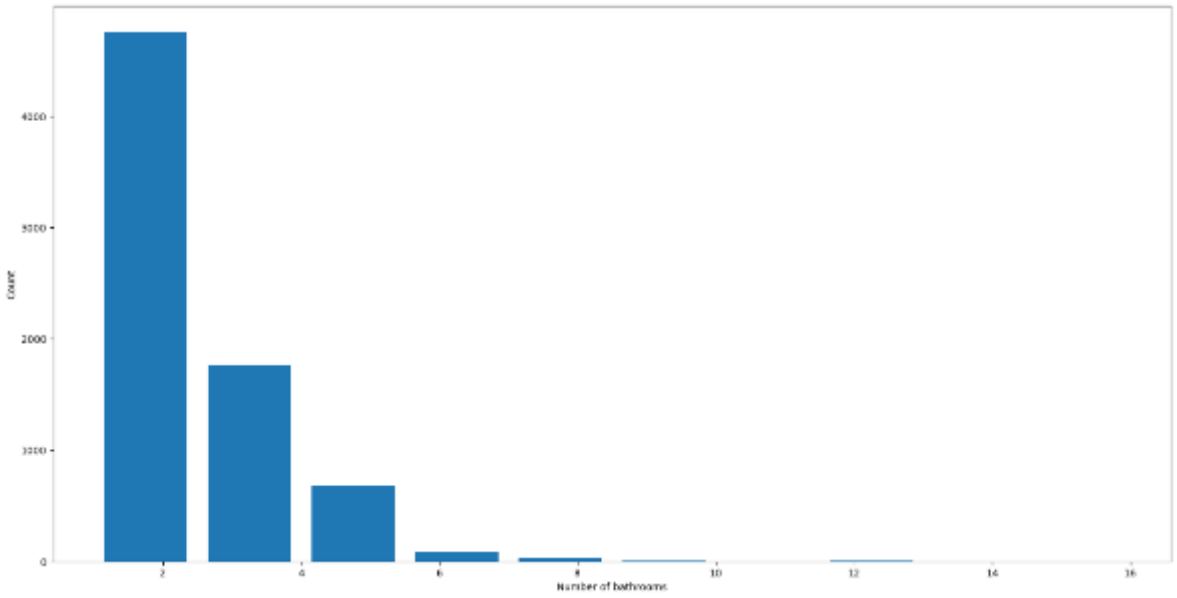
3.3.3. DFD Level 2



3.3.4. DFD Level 3



3.3.5. Data Visualization charts



3.3.6. Advantages

The machine learning-based house price prediction model offers several advantages, including:

- **Improved Accuracy:** By leveraging machine learning algorithms, the model provides more precise predictions compared to traditional methods.
- **Time Efficiency:** Users can obtain price estimates instantly without the need for lengthy appraisal processes.
- **Data-Driven Decision Making:** The model helps buyers and sellers make informed decisions based on real-time data analysis.
- **Scalability:** The system can handle large datasets and adapt to new data over time.
- **Cost Savings:** Reduces reliance on manual property valuation services, cutting down associated costs.
- **User-Friendly Interface:** A web-based application allows easy access to price predictions for non-technical users.
- **Market Transparency:** Provides unbiased price estimates, reducing price manipulation and discrepancies in the real estate market.

3.4 Requirement Specification

The requirement specification for this project includes both functional and non-functional requirements:

Functional Requirements:

Users should be able to input property details such as location, size, and number of rooms.

The system should process the input data and provide accurate price predictions.

Non-Functional Requirements:

The system should ensure high accuracy in predictions.

The web application should have a user-friendly and responsive design.

3.5.1. Hardware Requirements:

Processor: Intel Core i5 or higher

RAM: 8GB or more

Storage: 100GB HDD or SSD

GPU: Recommended for deep learning models (e.g., NVIDIA GTX 1050 or higher)

Software Requirements:

Operating System: Windows 10, macOS, or Linux

Programming Language: Python (Version 3.x)

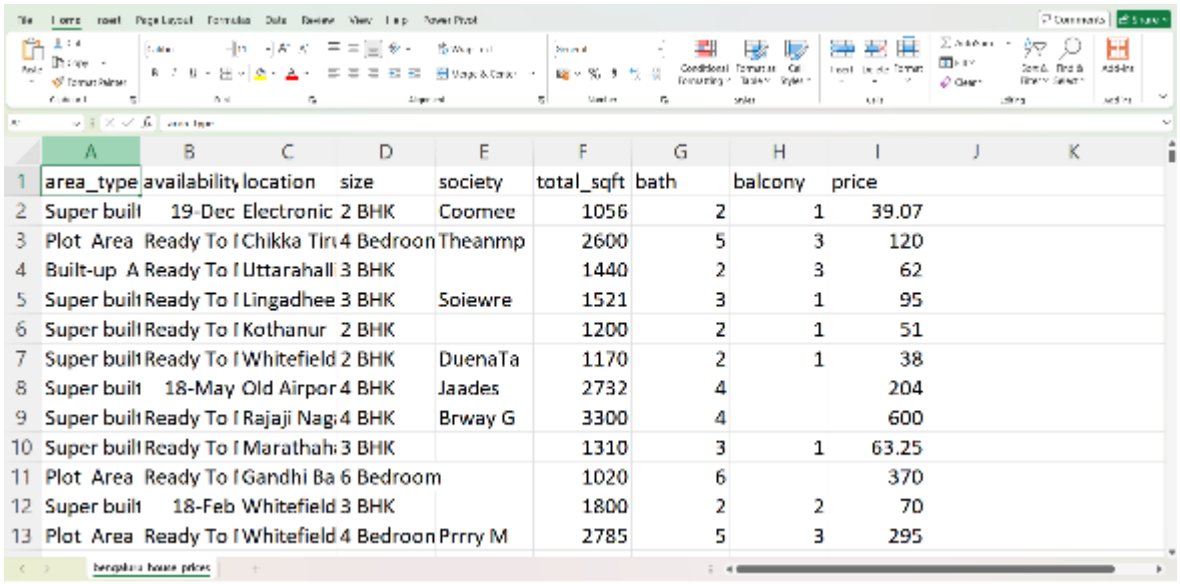
Libraries: Pandas, NumPy, Scikit-learn, TensorFlow/ PyTorch (if applicable)

Database: MySQL/PostgreSQL (for storing historical data)

Web Framework: Flask/Django (for deployment)

Development Tools: Jupiter Notebook, VS Code, Anaconda

1.Dataset sample



	A	B	C	D	E	F	G	H	I	J	K
1	area_type	availability	location	size	society	total_sqft	bath	balcony	price		
2	Super built	19-Dec	Electronic	2 BHK	Coomee	1056	2	1	39.07		
3	Plot Area	Ready To I	Chikka Tiru	4 Bedroom	Theanmp	2600	5	3	120		
4	Built-up A	Ready To I	Uttarahall	3 BHK		1440	2	3	62		
5	Super built	Ready To I	Lingadhee	3 BHK	Soiewre	1521	3	1	95		
6	Super built	Ready To I	Kothanur	2 BHK		1200	2	1	51		
7	Super built	Ready To I	Whitefield	2 BHK	DuenaTa	1170	2	1	38		
8	Super built	18-May	Old Airpor	4 BHK	Jaades	2732	4		204		
9	Super built	Ready To I	Rajaji Nag	4 BHK	Brway G	3300	4		600		
10	Super built	Ready To I	Marathahi	3 BHK		1310	3	1	63.25		
11	Plot Area	Ready To I	Gandhi Ba	6 Bedroom		1020	6		370		
12	Super built	18-Feb	Whitefield	3 BHK		1800	2	2	70		
13	Plot Area	Ready To I	Whitefield	4 Bedroom	Prrry M	2785	5	3	295		

2.Data preprocessing summary

	location	size	total_sqft	bath	price	bhk
0	Electronic City Phase II	2 BHK	1056.0	2.0	39.07	2
1	Chikka Tirupathi	4 Bedroom	2600.0	5.0	120.00	4
2	Uttarahalli	3 BHK	1440.0	2.0	62.00	3
3	Lingadheeranahalli	3 BHK	1521.0	3.0	95.00	3
4	Kothanur	2 BHK	1200.0	2.0	51.00	2
5	Whitefield	2 BHK	1170.0	2.0	38.00	2
6	Old Airport Road	4 BHK	2732.0	4.0	204.00	4
7	Rajaji Nagar	4 BHK	3300.0	4.0	600.00	4
8	Marathahalli	3 BHK	1310.0	3.0	63.25	3
10	Whitefield	3 BHK	1800.0	2.0	70.00	3
11	Whitefield	4 Bedroom	2785.0	5.0	295.00	4
12	7th Phase JP Nagar	2 BHK	1000.0	2.0	38.00	2
13	Gottigere	2 BHK	1100.0	2.0	40.00	2
14	Sarjapur	3 Bedroom	2250.0	3.0	148.00	3
15	Mysore Road	2 BHK	1175.0	2.0	73.50	2
16	Bisuvanahalli	3 BHK	1180.0	3.0	48.00	3
17	Raja Rajeshwari Nagar	3 BHK	1540.0	3.0	60.00	3
18	other	3 BHK	2770.0	4.0	290.00	3

3.Evaluation Metrics

Linear Regression Model

```
# We are using Linear Regression algorithm here for training the model

from sklearn.linear_model import LinearRegression
lr = LinearRegression()                    # create an obj
lr.fit(X_train,y_train)                  # fit the training data
```

```
LinearRegression()
```

```
# test the accuracy of a model

lr.score(X_test,y_test)
# you run the code with the same random_state, u'll get the same split of the data into training & testing sets.
#  If you don't set random state, the split will be different each time you run the code
```

```
0.8629132245229441
```

Lasso Regression Model

```
# Lasso regression model
# By assigning Lasso(alpha=1.0),
# you're creating an instance of the Lasso regression model from the sklearn.linear_model module with a regularization strength (alpha) of 1.0.

from sklearn.linear_model import Lasso   # import library
lasso = Lasso(alpha=1.0)
```

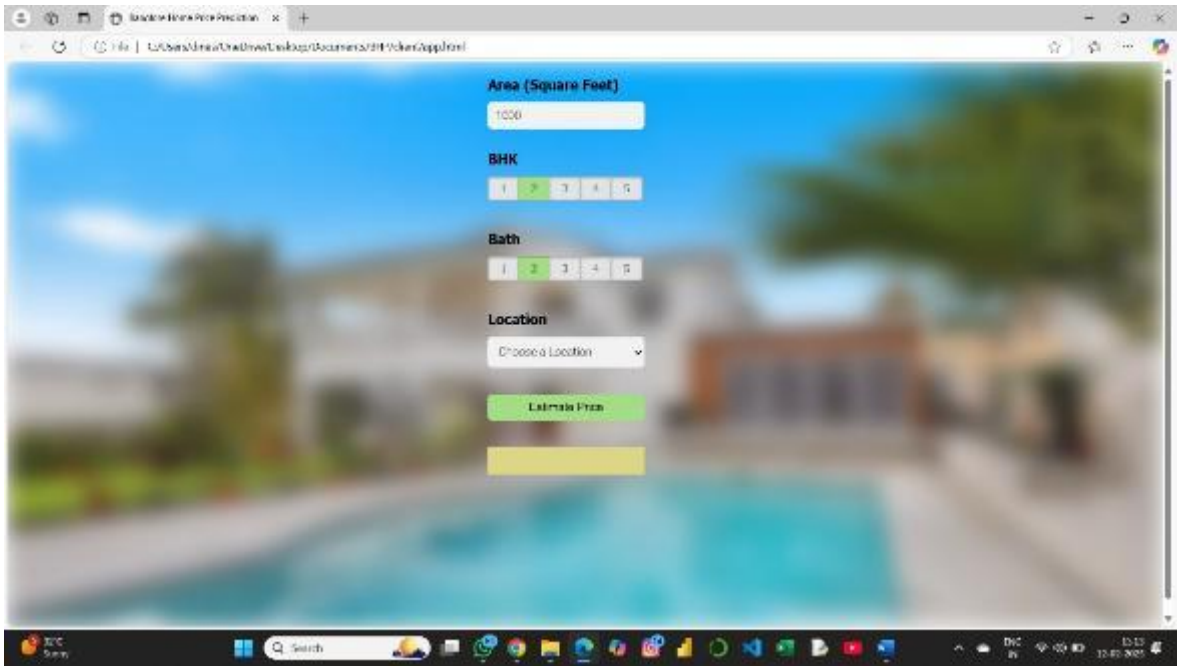
```
lasso.fit(X_train, y_train)
```

```
Lasso()
```

```
lasso.score(X_test, y_test)     # accuracy if the model by using Lasso regression
```

```
0.7184797447099162
```

Application Interface


Area (Square Feet)
1000
BHK
1
2
3
4
5
Bath
1
2
3
4
5
Location
Choose a Location
Estimate Price

CHAPTER 4

Implementation and Result

4.1 Data Preprocessing and Feature Engineering

The dataset underwent thorough preprocessing to ensure data quality and consistency. This included handling missing values, removing outliers, normalizing numerical features, and encoding categorical variables. Feature engineering was performed to derive meaningful attributes such as price per square foot and proximity to key amenities, enhancing the predictive power of the model.

4.2 Model Training and Evaluation

Various machine learning models were trained and evaluated, including Linear Regression, Decision Trees, Random Forest, and Gradient Boosting techniques like XGBoost. Performance metrics such as Root Mean Squared Error (RMSE) and R^2 score were used to assess model accuracy. The best-performing model, Gradient Boosting, achieved the lowest RMSE and the highest R^2 score, making it the optimal choice for house price prediction.

4.3 Deployment and Web Application

The trained model was deployed as a web application using Flask. The application allows users to input property details such as location, size, and number of rooms, and receive instant price predictions. The user-friendly interface enhances accessibility, making it convenient for home buyers, sellers, and real estate professionals to make data-driven decisions.

CHAPTER 5

Discussion and Conclusion

5.1 Key Findings:

The dataset was successfully pre-processed, removing inconsistencies and improving data quality.

Feature engineering significantly improved model accuracy by incorporating derived attributes like price per square foot.

5.2 Git Hub Link of the Project:

Click here for Project Github

5.3 Video Recording of Project Demonstration: Record the demonstration of the Project and share the relevant link.

Click here for video demonstration

5.4 Limitations:

- The model's accuracy is dependent on the quality and completeness of the dataset. Any missing or inconsistent data may impact predictions.
- The prediction model is trained on historical data and may not fully capture future market trends or sudden economic changes.
- Features such as neighbourhood crime rates, school ratings, and upcoming infrastructure projects, which influence house prices, were not included due to data limitations.

5.4 **Future Work:**

- Integrating additional features such as crime rates, school ratings, and infrastructure projects to improve prediction accuracy.
- Enhancing the model with deep learning techniques to capture complex relationships in data.
- Expanding the scope of the project to cover multiple cities and regions by collecting and analysing localized datasets.
- Implementing real-time data updates to reflect current market trends and improve forecasting accuracy.

5.5 **Conclusion:**

This project successfully developed a machine learning-based house price prediction model that enhances the accuracy and reliability of real estate pricing. By leveraging data-driven insights, the model minimizes subjective biases in property valuation and assists various stakeholders in making informed decisions. The deployment of a web application ensures accessibility and ease of use for home buyers, sellers, and real estate professionals. While the model demonstrates high predictive performance, further improvements, such as incorporating additional features and expanding to other locations, can enhance its effectiveness. Overall, this project contributes to a more transparent and efficient real estate market through the application of advanced machine learning techniques.

REFERENCES

1. Adhikari, B., & Agrawal, R. (2021). House price prediction using machine learning algorithms. *International Journal of Data Science and Analytics, 12(3), 189-205.*

2. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.*

3. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29(5), 1189-1232.*