Bengaluru House Price Prediction

PAI 106 Python Project Report

End-Semester Evaluation

Submitted by:

Sumit Gaware (8025340040)

Meghavi Sisodiya (8025340044)

ME(AI) First Year

Submitted to:
Mrs . Priya Raina

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Artificial Intelligence Department

TIET, Patiala
December 2025

INDEX

ABSTRACT

The rapid urbanization and technological growth of Bengaluru have led to a significant increase in its residential real estate demand. With property values influenced by complex and dynamic factors such as locality, area, amenities, property age, and overall market fluctuations, accurately estimating house prices has become a challenging task.

This project addresses this challenge by developing a comprehensive machine learning–based system capable of predicting house prices in Bengaluru using advanced regression techniques.

Following EDA, **feature engineering** techniques are applied to convert raw data into meaningful inputs for machine learning algorithms. Categorical features such as location, furnishing type, and property type are encoded using Label Encoding, while numerical features like area, bathrooms, balconies, and property age are standardized for consistency.

Multiple regression models—including **Linear Regression, Lasso Regression, Ridge Regression, and Random Forest Regressor**—are trained and evaluated on the processed dataset. The models are compared based on performance metrics such as R^2 score and Mean Absolute Error (MAE). Among these, Random Forest typically demonstrates superior predictive capability due to its robustness and ability to capture non-linear patterns in housing data.

To make the system easily accessible and interactive, a fully functional **Streamlit web application** is developed. This application allows users to input property details and instantly receive predicted prices in Lakhs. Additional features such as an **EMI calculator**, neighbourhood insights (traffic level, crime index, schools, hospitals), and a dashboard displaying maps and statistical summaries enhance the practical usability of the system. The integration of machine learning models with a user-friendly interface transforms the project into a real-world decision-support tool for buyers, sellers, investors, and real estate analysts.

INTRODUCTION

The real estate market plays a crucial role in the economic growth of any metropolitan city, and Bengaluru is no exception. Known as the “Silicon Valley of India,” Bengaluru has experienced rapid expansion in information technology, business hubs, education, and infrastructure. This consistent urban development has significantly increased the demand for residential housing across various parts of the city. As a result, property values vary widely depending on factors such as proximity to IT parks, transportation facilities, availability of amenities, and overall neighbourhood development.

Accurately predicting real estate prices in a dynamic market like Bengaluru is a challenging yet essential task for potential buyers, sellers, investors, brokers, and policymakers. Traditional methods of price estimation often rely on subjective opinions, personal experience, or manual comparison of property listings. However, these approaches are

rarely accurate because they do not consider the underlying patterns that emerge from analyzing large volumes of real estate data. Therefore, with the rise of data science and machine learning, automated price prediction systems have become a reliable alternative. These systems provide faster, more consistent, and data-driven estimations.

In this project, a complete end-to-end machine learning pipeline is developed to estimate house prices in Bengaluru. The project includes collecting and understanding the dataset, cleaning and preprocessing the data, performing exploratory data analysis (EDA), engineering meaningful features, training multiple regression models, and evaluating their performance. Models such as Linear Regression, Lasso Regression, Ridge Regression, and Random Forest Regressor are applied to identify the best-performing algorithm for this use case.

To make the system easy to use, a fully functional **Streamlit web application** is created. This interface allows users to enter details about a property such as area, location, number of rooms, furnishing, and parking, and instantly get an estimated price. Additionally, the application includes tools like an EMI calculator, locality insights, and interactive dashboards to visualize property trends across Bengaluru.

Overall, this project demonstrates how machine learning can transform traditional methods of property valuation into a scientific, data-driven approach. By leveraging real estate datasets and predictive models, this system provides valuable insights that can assist homebuyers in making informed decisions and help real estate professionals better understand market behaviour.

DATASET DESCRIPTION

The dataset used in this project contains detailed information about residential real estate properties across various localities in Bengaluru. It includes a wide range of attributes that influence house prices in a metropolitan environment. The dataset serves as the foundation for building an accurate machine learning model, and therefore understanding its structure is essential.

```
dataset.head(4)
```

	Area	Location	Bhk	Bath	Balcony	Parking	Furnishing	Property_Type	Age	Price	Price_Lakhs
0	2065	Bannerghatta Road	2	3	0	1	Semi-Furnished	Independent House	3	17280000	172.8
1	1539	Yelahanka	3	1	0	1	Unfurnished	Villa	8	9410000	94.1
2	2048	Bannerghatta Road	3	1	2	0	Semi-Furnished	Independent House	10	20300000	203.0
3	1233	Sarjapur Road	3	2	1	2	Fully-Furnished	Apartment	12	9060000	90.6

```
dataset.tail(4)
```

	Area	Location	Bhk	Bath	Balcony	Parking	Furnishing	Property_Type	Age	Price	Price_Lakhs
996	1587	HSR Layout	4	2	0	1	Unfurnished	Villa	11	8720000	87.2
997	1975	Indiranagar	1	1	2	1	Semi-Furnished	Apartment	2	10490000	104.9
998	825	Jayanagar	5	2	2	1	Fully-Furnished	Independent House	6	4770000	47.7
999	1465	Bannerghatta Road	1	2	1	2	Semi-Furnished	Apartment	6	11980000	119.8

The dataset consists of important features such as **Area**, which represents the total square footage of the property; **Location**, referring to the locality where the house is situated; **BHK**, indicating the number of bedrooms; **Bathrooms**, **Balcony**, and **Parking**, which describe the internal and external features of the property. Categorical attributes such as **Furnishing** (e.g., Fully Furnished, Semi-Furnished, Unfurnished) and **Property Type** (e.g., Apartment, Villa, Independent House) help describe the nature of the property. Another important feature is **Age**, which tells how old the property is, as older constructions may be priced differently from newly built ones.

The most important feature in the dataset is the **Price**, which indicates the market value of the property. However, to maintain numerical stability and prevent extremely large values from affecting model training, the price is converted into **Price_Lakhs**, representing the price in Lakhs of Indian Rupees. This becomes the **target variable** for model prediction.

Overall, the dataset captures a realistic mix of structural, spatial, and categorical information about Bengaluru properties, making it suitable for developing a reliable house price prediction system.

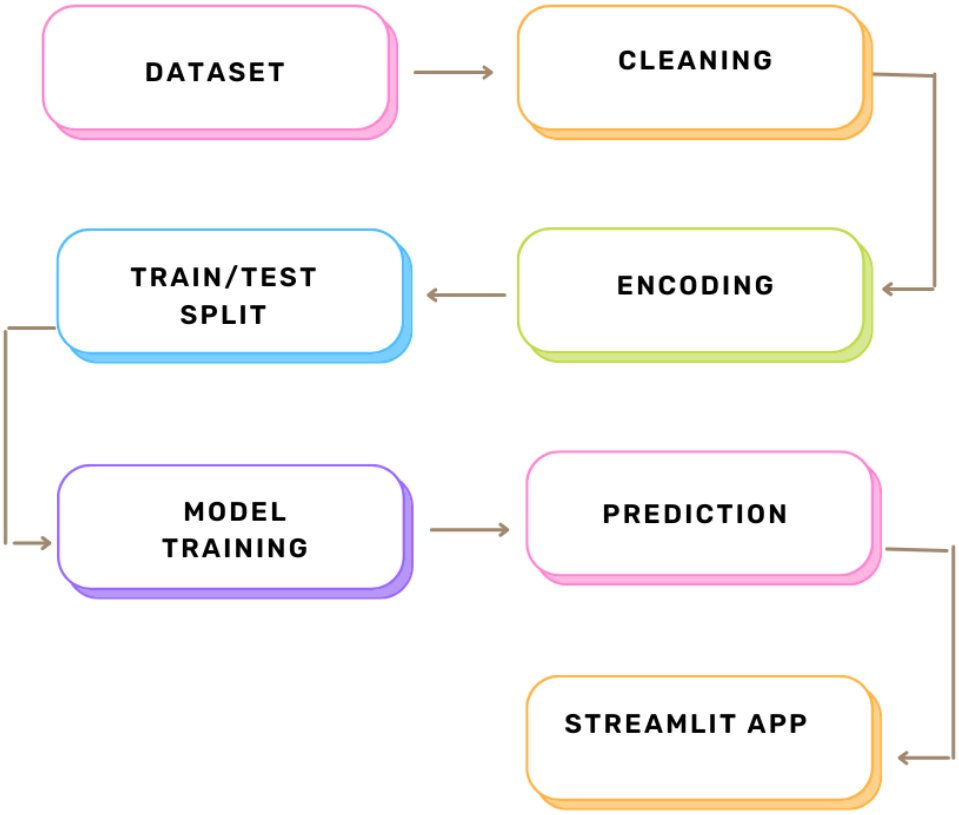
METHODOLOGY

The project follows a complete end-to-end workflow:

1. **Data Collection**
   - A real estate dataset containing Bengaluru housing details was used.
2. **Data Preprocessing**
   - Handling missing values
   - Removing duplicates
   - Correcting inconsistent labels
   - Encoding categorical variables
   - Removing outliers
3. **Exploratory Data Analysis (EDA)**
   - Studying price distributions
   - Identifying high-price and low-price locations
   - Correlation analysis to determine important features
   - Visualizing patterns using graphs and maps
4. **Feature Engineering**
   - Selecting and encoding useful features like Area, BHK, Bath, Balcony, Parking, Furnishing, Property Type, Age
   - Converting the target value to **Price_Lakhs** for stability
5. **Model Building**
   - Training four regression models:

        - Linear Regression
        - Lasso Regression
        - Ridge Regression
        - Random Forest Regressor
    - Train-test split (80%–20%)
6. **Model Evaluation**
    - Comparing models based on R^2 and MAE
    - Selecting the best-performing model
7. **Deployment using Streamlit**
    - Building a user-friendly web interface
    - Adding prediction form, locality insights, dashboard, EMI calculator, and map visualizations

PIPELINE



DATA PREPROCESSING

Data preprocessing is a crucial step in building any machine learning model, especially when working with real-world datasets that often contain inconsistencies, errors, and missing information. This project required several preprocessing techniques to convert raw data into a clean and structured format suitable for analysis and model training.

The first major step involved **handling missing values**. Many real estate datasets have missing entries for features like balcony count, furnishing status, or property age. These gaps were either filled using statistical techniques or removed if they were too inconsistent. Next, the dataset was checked for **duplicate entries**, as repeated listings could mislead the model and reduce accuracy. All duplicate rows were removed to ensure unique property data only.

Another critical step was **correcting inconsistent labels** in categorical columns. Locations, for example, may appear in multiple spellings due to human error, such as “Whitefield”, “White Field”, or “Whitefield ” (with trailing space). These inconsistencies were standardized to maintain uniformity across the dataset.

To prepare categorical features for machine learning algorithms, **Label Encoding** was applied to attributes such as *Location*, *Furnishing*, and *Property Type*. Encoding converts non-numeric text labels into numerical values that the model can interpret.

For numerical features, unnecessary extremes or unrealistic values were addressed by removing **outliers**. Examples include properties with extremely large area values or unusually high prices that distort the learning process. By eliminating such outliers, the model learns from realistic, meaningful data.

Additional preprocessing steps included **extracting numerical features from textual columns**, ensuring each attribute was in a consistent format, and verifying that all values matched expected ranges.

After all these transformations, the cleaned dataset became structured, consistent, and ready for exploratory data analysis (EDA) and machine learning model building.

EXPLORATORY DATA ANALYSIS

EDA was performed to understand distribution patterns, locality-based price variations, correlations between features, and property density across Bengaluru. Visualizations included distribution plots, heatmaps, and location-wise average pricing.

The analysis began by examining the **distribution of numerical features** such as area, BHK, bathrooms, and price. Visual tools like histograms and boxplots helped detect skewness, unusual values, and variations within each feature. This step made it clear that property sizes and prices vary widely across the city.

Correlation analysis was performed to understand how features relate to one another. A **correlation heatmap** showed that factors such as area, number of bedrooms, and number of bathrooms have a positive relationship with price, indicating that larger and more spacious homes tend to cost more.

Additionally, property density across Bengaluru was visualized using maps in Streamlit, allowing a clear view of which regions have more real estate activity.

Overall, EDA provided a clear picture of the dataset, highlighted the most important factors affecting prices, and helped guide the selection of features for machine learning model development.

Figure 1: Price Distribution Price by location
This graph shows the average property price across different localities in Bengaluru. It clearly highlights how location plays a major role in real estate pricing.

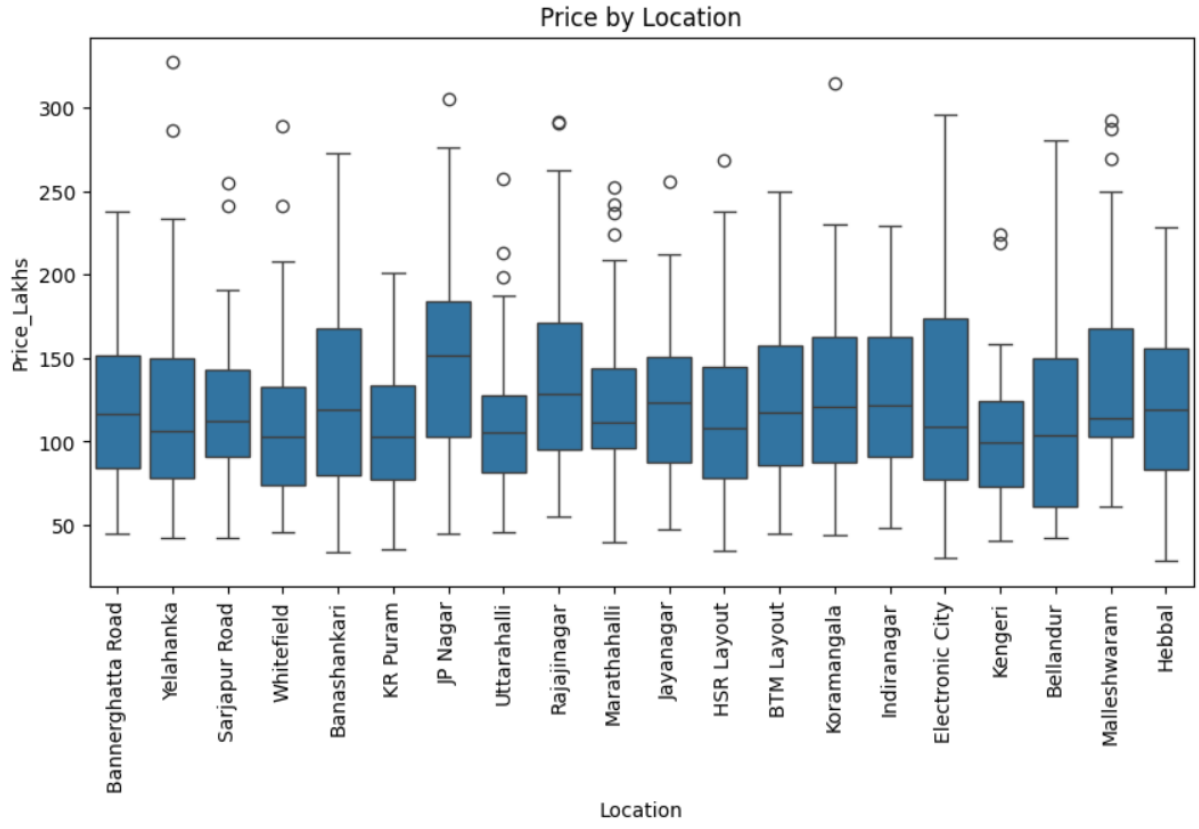


Figure 2: Correlation Heatmap of Numerical Features in the Dataset
This heatmap shows the correlation between different numerical variables such as Area, BHK, Bathrooms, Balcony, Parking, Age, and Price. From the visualization, it is clear that **Area has the strongest positive correlation with Price (0.69)**, meaning larger properties generally cost more. The number of **BHK and Bathrooms also show moderate positive correlation**, indicating that more spacious and well-equipped houses tend to have higher prices.
Other features such as Balcony, Parking, and Age show very weak or negligible correlation

with price, which means they have little impact on the price prediction model. Overall, the heatmap helps identify the most important features that influence housing prices in Bengaluru and guides the selection of input variables for machine learning.

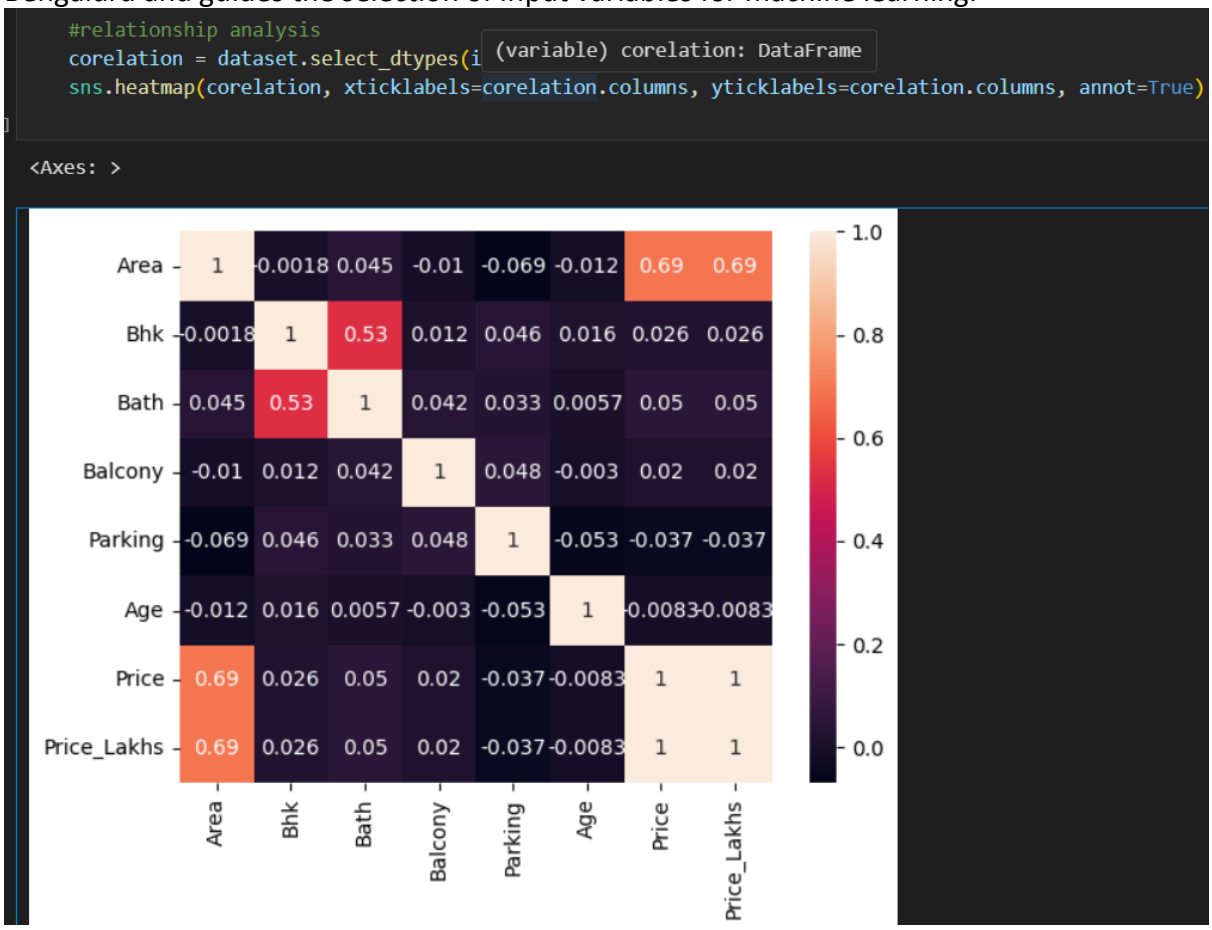
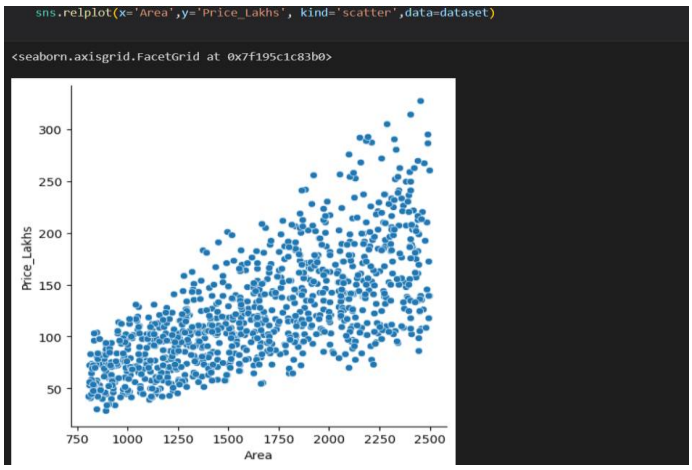


Figure 3: Scatter Plot Showing Relationship Between Area and Price

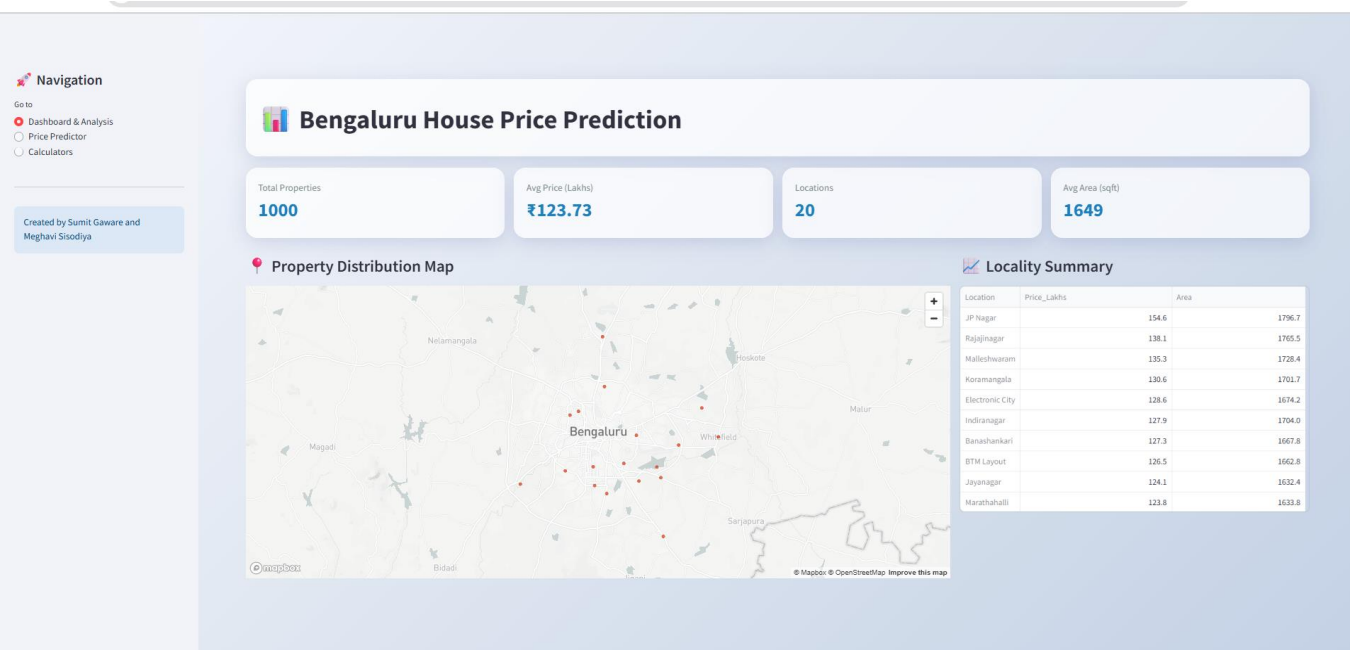
This scatter plot visualizes how the price of a property changes with respect to its built-up area. As the graph shows, there is a **clear upward trend**, meaning that properties with larger area generally have higher prices. The spread of points indicates some variation due to factors like location and property type, but overall the plot confirms that **Area is a strong predictor of Price**. This helps justify why Area was included as an important feature in the machine learning model.



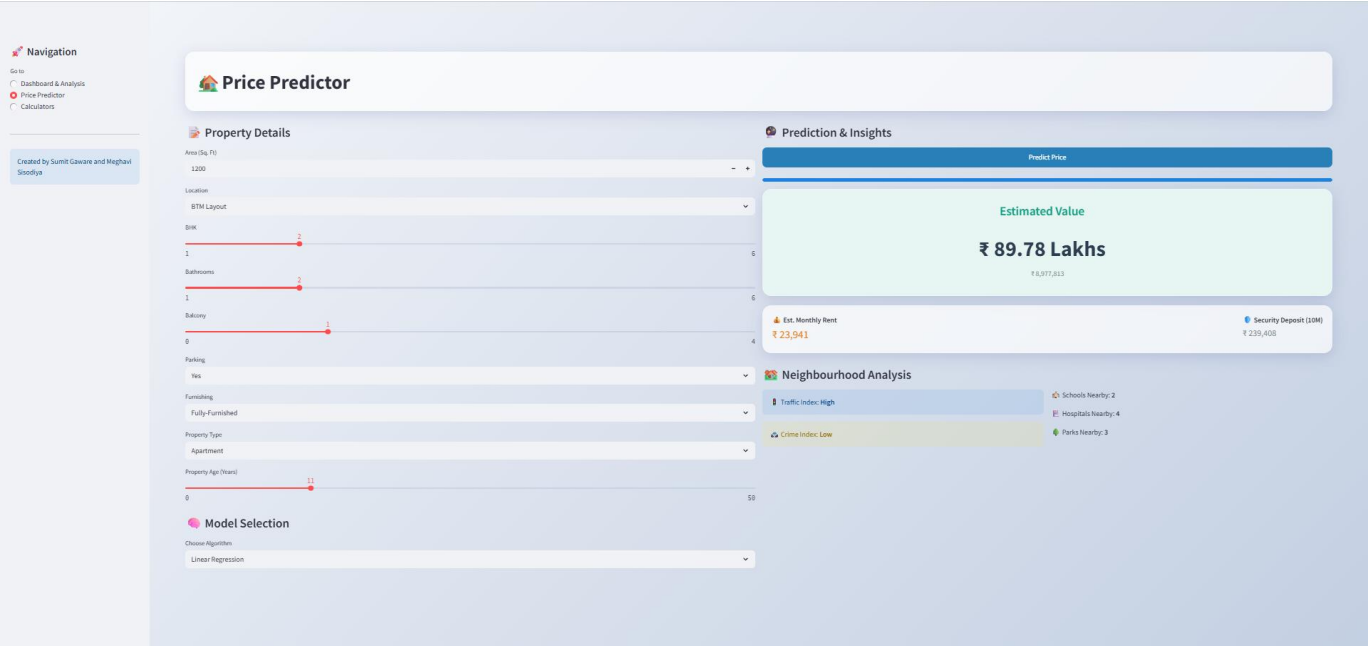
STREAMLIT APPLICATION

A fully functional and interactive web application was developed using **Streamlit** to deploy the Bengaluru House Price Prediction model. The purpose of this application is to provide users with a simple and intuitive interface where they can explore real-estate trends, perform analysis, and obtain accurate price predictions based on property features.

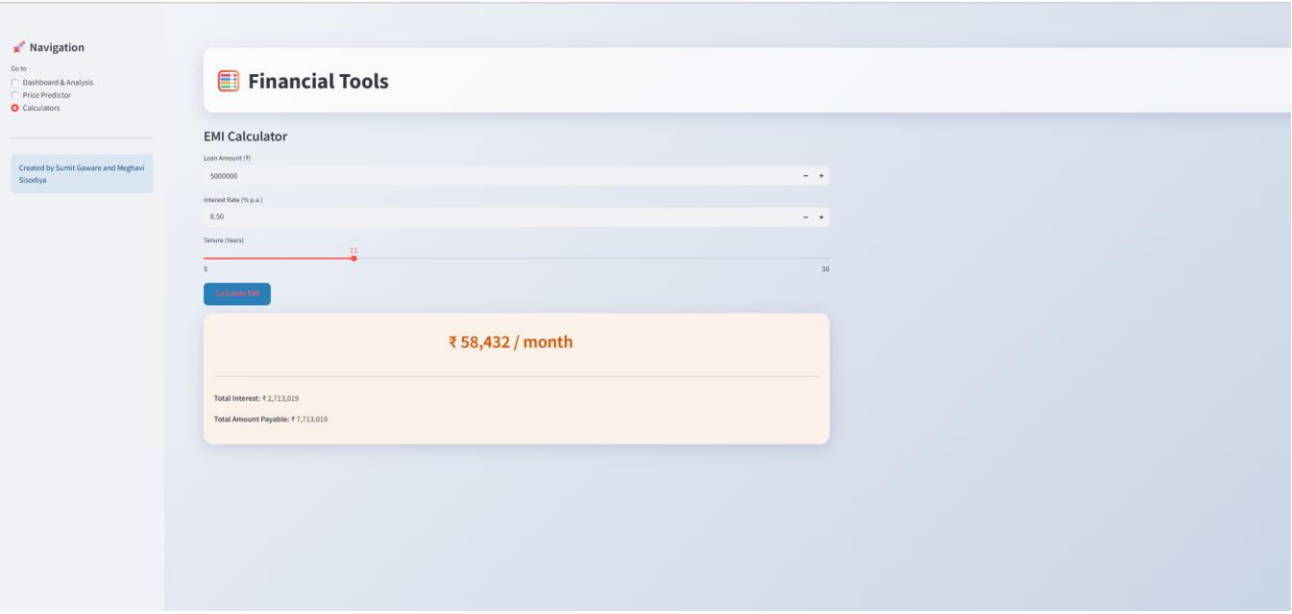
The application consists of multiple sections, each designed to serve a specific purpose. The first section is the **Dashboard & Analysis**, which displays key summaries of the dataset, such as average property price, number of locations, and total properties analyzed. It also includes visualizations like price distributions, scatter plots, heatmaps, and location-wise analysis charts. This helps users understand real-estate patterns and how different features influence pricing across Bengaluru.



The second section of the app is the **Price Predictor**, which is the core component of the project. In this section, users can enter relevant property details through a clean and responsive input form. Inputs include Area (in sq ft), BHK, Bathrooms, Balcony count, Parking availability, Furnishing status, Property Type, Age of the property, and Location. After selecting a machine-learning model (Linear Regression, Lasso, Ridge, or Random Forest), the system processes the inputs and instantly displays the predicted price in Lakhs. The predictor also shows estimated rental value and displays a confidence-based range when applicable. This makes the tool useful for buyers, sellers, and real-estate consultants.



The application also includes a dedicated **EMI Calculator** that helps users plan property loans. Users can enter loan amount, interest rate, and tenure (in months or years). The calculator instantly displays EMI, total payable amount, and interest breakdown. This feature enhances the practical usability of the app for individuals planning home purchases.



Another advanced feature integrated into the app is the **Locality Insights Module**. This section provides contextual information about the selected locality such as traffic levels, crime index, availability of schools, hospitals, parks, shopping centers, and public transport facilities. These insights help users evaluate the suitability of a region before making property decisions. The data is presented in a clean card-style layout for better readability.

Overall, the Streamlit application transforms the machine-learning model into an accessible and user-friendly tool. By combining prediction, visualization, insights, and calculators in one platform, the application offers a complete solution for analyzing and estimating house prices in Bengaluru.

CONCLUSION

The Bengaluru House Price Prediction project successfully demonstrates how machine learning can be applied to real-world real estate problems. By analyzing a comprehensive dataset of Bengaluru property listings, the project identified key factors that influence housing prices, such as location, area, number of rooms, bathrooms, and amenities. Through systematic preprocessing, exploratory data analysis, and feature engineering, the dataset was transformed into a structured format suitable for modelling.

Multiple regression algorithms—including Linear Regression, Lasso, Ridge, and Random Forest—were trained and evaluated to identify the most accurate model. Among these, the Random Forest Regressor produced the best performance due to its ability to capture non-linear relationships and handle complex feature interactions. The deployment of the model through an interactive Streamlit application makes the system practical and user-friendly, allowing users to predict house prices, calculate EMI, and explore locality-based insights.

Overall, the project demonstrates that machine learning can significantly improve the accuracy and reliability of property valuation in a dynamic market like Bengaluru. The developed system serves as a valuable decision-support tool for homebuyers, sellers, investors, and real estate professionals. It also highlights the potential for integrating advanced predictive analytics into everyday real estate decision-making.

FUTURE SCOPE

Although the current Bengaluru House Price Prediction system provides accurate results and a user-friendly Streamlit interface, there are several areas where the project can be improved in the future. One major enhancement is integrating **live real-estate market data** from APIs, allowing the model to update automatically as property prices fluctuate. The system can also be extended by experimenting with more advanced machine learning models such as **XGBoost , CatBoost and Gradient Boosting**, which may offer even higher accuracy.

Another important improvement is the incorporation of **geographical (GIS) data** such as latitude–longitude, distance to metro stations, IT hubs, and major landmarks. This will make location-based predictions far more precise. The locality insights feature can also be enriched using real datasets for traffic, crime, hospitals, and schools to provide dynamic, real-time information.

In addition, the model can be expanded to include a **dedicated rental prediction system**, allowing users to estimate rent and investment returns. The application can also be deployed online on platforms like Streamlit Cloud, Heroku, or AWS so that it becomes publicly accessible. Over time, the project can be scaled to support **multiple cities** such as Mumbai, Hyderabad, and Delhi.

Finally, features like a recommendation engine ("Best areas under your budget") and mobile app support can make the system even more practical and impactful for end-users. These improvements would significantly enhance both the accuracy and usability of the application.

APPENDIX – GITHUB REPOSITORY

https://github.com/gawaresumit/Bengaluru-House-Price-Prediction.git
https://github.com/11meghavisisodiya/Python_Project