**PROJECT 1**

# **Project Title:** House Price Prediction System Using Data Analytics Algorithms

## **Objective:**

The primary objective of this project is to develop a predictive system that can accurately estimate house prices based on various factors like area, number of bedrooms, location, condition, and more, using machine learning techniques and data visualization.

**Dataset Description:**

The dataset used is named **House Price Prediction Dataset.csv**, which contains the following features:

| **Feature** | **Description** |
| --- | --- |
| Id | Unique identifier for each house |
| Area | Total area in square feet |
| Bedrooms | Number of bedrooms |
| Bathrooms | Number of bathrooms |
| Floors | Number of floors |
| YearBuilt | Year the house was built |
| Location | Area where the house is located |
| Condition | House condition (Excellent, Good, etc.) |
| Garage | Whether it has a garage (Yes/No) |
| Price | Target variable: House price in INR |

**Tools & Technologies Used:**

* **Language**: Python
* **IDE**: Google Colab
* **Libraries**:
  + pandas, numpy – for data manipulation
  + seaborn, matplotlib – for visualization
  + scikit-learn – for machine learning (RandomForestRegressor, OneHotEncoder, Pipeline)

**Exploratory Data Analysis (EDA):**

**1. Correlation Heatmap**

* Shows relationships between numeric features (Area, Bedrooms, Price, etc.)
* Helps identify strong predictors of price.

**2. Average Price by Location (Bar Graph)**

* Displays which locations have higher house prices on average.

**3. Boxplot (Price by Bedrooms)**

* Visualizes price range variation for different bedroom counts.

**4. Actual vs Predicted Prices (Scatter Plot)**

* Compares actual vs predicted prices to evaluate prediction accuracy.

**Machine Learning Model:**

* **Model Used:**

**Random Forest Regressor** (ensemble learning technique)

* **Preprocessing:**
* Categorical features (Location, Condition, Garage) were encoded using **OneHotEncoding**.
* Numerical features were passed as-is.
* Data was split into **80% training** and **20% testing**.

**Evaluation Metrics:**

| **Metric** | **Value** | **Meaning** |
| --- | --- | --- |
| **RMSE** | ₹292,410 approx | Average prediction error |
| **R² Score** | -0.099 | Poor fit; model underperformed |

Note: Accuracy can be improved with better models or feature engineering.

**Summary of Steps:**

1. **Data Upload** in Google Colab
2. **Data Preprocessing** using scikit-learn Pipelines
3. **Model Training** with Random Forest
4. **Evaluation** using RMSE and R²
5. **Visualization** using Seaborn and Matplotlib

**Graphical Outputs:**

* **Correlation Heatmap:**

Highlights the linear relationships between features like Area, Price, Bedrooms, etc.

* **Bar Graph:**

Shows average price per location.

* **Boxplot:**

Visualizes how price changes with bedroom count.

* **Scatter Plot:**

Visualizes actual vs predicted prices to show how well the model is performing.

**Conclusion:**

This project demonstrated how machine learning and data analytics can be used to predict house prices. Although the current model has limited accuracy, it lays a strong foundation for further development with better data and advanced techniques.