Real Estate Price Prediction using Machine Learning

---

1. Introduction

This project focuses on building a machine learning model to predict real estate prices based on various features such as location, property type, size, and number of bedrooms. The goal is to assist stakeholders in making informed decisions regarding property pricing using data-driven insights.

---

2. Problem Statement

Property pricing is influenced by numerous factors, and manual estimation often leads to inaccuracies. The objective of this project is to develop an accurate and reliable predictive model using machine learning techniques to estimate property prices.

---

3. Dataset Description

The dataset includes:

- **area_type**: Categorical (e.g., Super built-up Area, Plot Area, etc.)
- **availability**: Categorical (e.g., Ready To Move, Under Construction)
- **location**: Categorical (e.g., Indiranagar, Whitefield, etc.)
- **size**: Categorical (e.g., 2 BHK, 3 BHK)
- **society**: Categorical (name of the housing society, may contain missing values)
- **total_sqft**: Numerical (total area in square feet)
- **bath**: Numerical (number of bathrooms)
- **balcony**: Numerical (number of balconies)
- **price**: Target variable (price in lakhs)

---

4. Data Preprocessing

- **Handling Missing Values**: Rows with null values in critical columns like total_sqft, bath, and location were removed or imputed appropriately.
- **Feature Engineering**: Extracted BHK from the size column, and converted total_sqft to numerical values.
- **Encoding Categorical Variables**: One-Hot Encoding was applied to location, area_type, and availability features.
- **Feature Scaling**: Not necessary for tree-based models like Random Forest.
- **Feature Selection**: Final features used include: location, area_type, availability, total_sqft, bath, balcony, and BHK.

---

5. Model Selection

After experimenting with different models such as Linear Regression and Decision Trees, the **Random Forest Regressor (RFR)** was chosen due to its high accuracy and ability to handle non-linear data efficiently.

---

6. Model Training

- The dataset was split into **training and test sets (80:20)**.
- A **Random Forest Regressor** was trained on the training data.
- Hyperparameters were tuned using cross_val_score for optimal performance.

---

7. Evaluation Metrics

- **R2 Score**

The model achieved:

- R2 Score: **0.994** (close to 1 indicates a good fit)

---

8. Deployment (Optional)

The model was integrated with a **Flask web application** where users can input property details and get the predicted price instantly.

---

9. Conclusion

The Random Forest Regressor model successfully predicts real estate prices with considerable accuracy. With proper input data and preprocessing, it can serve as a valuable tool for real estate professionals and buyers.

---

10. Future Work

- Add more features like amenities, age of property, and proximity to city centers.
- Use advanced ensemble techniques or deep learning models.
- Expand dataset for more locations to improve generalizability.