Project Report: Bengaluru House Price Prediction with Streamlit Integration

# Project Title

Bengaluru House Price Prediction Using Machine Learning and Streamlit

# Objective

To develop a predictive model that estimates housing prices in Bengaluru using historical housing data and integrate the model into a user-friendly web application using Streamlit.

# Dataset Description

- Source: Kaggle or similar real estate data sources

- File Name: Bengaluru\_House\_Data.csv

- Features: Includes location, size, total\_sqft, bath, price, etc.

- Target Variable: Price (in lakhs)

# Tools and Technologies Used

- Programming Language: Python

- Libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, Joblib, Streamlit

# Data Preprocessing

- Dropped Irrelevant Columns: 'area\_type', 'availability', 'society', 'balcony'

- Handling Missing Values:

- 'location': filled with 'Sarjapur Road'

- 'size': filled with '2 BHK'

- 'bath': filled with the median value

- Feature Engineering:

- Extracted 'bhk' from 'size'

- Converted ranges in 'total\_sqft' to average values

- Created 'price\_per\_sqft' for better analysis

# Outlier Detection and Removal

- Removed entries with unrealistic bedroom sizes (e.g., bhk > 20)

- Removed entries with total\_sqft per bhk < 300

- Removed price\_per\_sqft outliers using mean and standard deviation method by location

- Removed BHK-based price anomalies

# Data Cleaning Summary

- Grouped locations occurring less than 10 times under 'other'

- Ensured numerical data consistency

- Saved cleaned dataset as "Cleaned\_data.csv"

# Exploratory Data Analysis (EDA)

- Visualized feature correlations using heatmaps

- Analyzed feature distributions and relationships

# Model Building

- Input Features (X): location, total\_sqft, bath, bhk

- Target (y): price

- Split: 80% training, 20% testing

- Transformations:

- OneHotEncoder for 'location'

- StandardScaler for numerical scaling

# Models Applied

- Linear Regression:

- Interpretable, baseline model

- Performance: R² Score and MAE

- Random Forest Regressor:

- Higher accuracy through ensemble learning

- Better R² Score and lower MAE than Linear Regression

# Evaluation Metrics

- R² Score (Goodness of Fit)

- Mean Absolute Error (MAE)

- Visual comparison of actual vs predicted prices for both models

# Model Export

- Random Forest model saved using Joblib as "RandomForestRegressionModel.joblib"

# Streamlit Web App Integration

- Developed a Streamlit app to interactively predict housing prices

- Features include user input for location, total square feet, number of bathrooms and bedrooms

- Real-time prediction output displayed on the app

- Includes optional visualization: correlation heatmap of features

- Uses the trained Random Forest model and cleaned dataset

# Conclusion

The project demonstrates effective data preprocessing, outlier removal, and machine learning to predict housing prices. The model was successfully deployed via a Streamlit web app to make the solution accessible and interactive.

# Future Scope

- Deploy the app using cloud platforms (e.g., Streamlit Cloud or Heroku)

- Incorporate more features like amenities, property age, or local market trends

- Test other advanced models like Neural Networks