Real Estate Price Prediction using Machine Learning

1. Introduction

The real estate market involves complex patterns influenced by multiple factors such as property type, location, and number of bedrooms/bathrooms. Predicting property prices accurately is crucial for buyers, sellers, and investors. In this project, we analyze a dataset of Gurgaon real estate listings, perform data cleaning and preprocessing, and develop machine learning models to predict housing prices. Both Random Forest Regression and Lasso Regression are used to compare performance.

2. Objectives

- To explore and analyze the Gurgaon real estate dataset.
- To handle missing values, duplicates, and outliers effectively.
- To preprocess categorical and numerical data using encoding, scaling, and imputation.
- To train and evaluate predictive models for housing prices.
- To compare Random Forest and Lasso Regression performance using Mean Squared Error (MSE).

3. Dataset Description

The dataset Gurgaon_RealEstate.csv includes:
- Features: Property details (property_type, society, bedrooms, bathrooms, etc.), numerical attributes (square footage, etc.)
- Target Variable: price – continuous numeric variable representing property price.

4. Data Exploration & Cleaning

- Removed duplicate rows to reduce noise.
- Performed exploratory data analysis (EDA) on property types, societies, and price distribution.
- Handled missing values: bathrooms with median, bedrooms with mode.
- Detected and capped outliers in price using the IQR method.

5. Methodology

- Preprocessing: Numerical features scaled with StandardScaler; categorical features encoded with OneHotEncoder.
- Train-test split: 80% training and 20% testing.

- Models: Random Forest Regressor (n_estimators=100) and Lasso Regression (alpha=0.01).
- Evaluation metric: Mean Squared Error (MSE).

6. Results

- Random Forest achieved lower MSE, showing better predictive performance.
- Lasso Regression showed higher MSE, indicating weaker performance on nonlinear patterns.

Model Comparison Table:

- Random Forest – Lower MSE
- Lasso Regression – Higher MSE

7. Tools & Libraries Used

- Python 3
- pandas, numpy – Data handling
- matplotlib, seaborn – Visualization
- scikit-learn – Preprocessing, RandomForestRegressor, Lasso, Pipeline

8. Conclusion

This project demonstrates the use of machine learning models for real estate price prediction. Random Forest outperformed Lasso Regression due to its ability to capture nonlinear relationships. The project highlights the importance of preprocessing and model comparison in predictive analytics.

Future Enhancements:

- Try Gradient Boosting models (XGBoost, LightGBM).
- Perform hyperparameter tuning for better accuracy.
- Include location-based/geospatial features.
- Apply log transformation on price to reduce skewness.