House Price Prediction Report

1.Introduction

The objective of this project was to classify house price prediction using various features from a dataset obtained from kaggle. The aim is to build a predictive model that accurately estimates house prices based on a set of input features like area, number of bedrooms, location and condition etc.

2. Data Loading and Preprocessing

The dataset is loaded and necessary preprocessing steps are applied to handle missing value, normalize features, and encode categorical variables. Exploratory Data Analysis (EDA) is conducted to understand the distribution of the data, and outliers that might affect model performance.

Exploratory Data Analysis

- No missing values
  - All columns are complete with no null entries.
- Feature Distribution
- ➢ Location:
  - Downtown: 558
  - Urban: 485
  - Suburban: 483
  - Rural: 474
- ➢ Condition:
  - Fair: 521
  - Excellent: 511
  - Poor: 507
  - Good: 461

- Garage:
  - No: 1038
  - Yes: 962
- Bedrooms:
  - Range: 1 to 5 (fairly evenly)
- Bathrooms:
  - Range: 1 to 4
- Correlation Heatmap:
  - Shows relationship between features and Price
  - Strong positive correlation between Area and Price

Data Preprocessing:

- Encoding:
  - Categorical variables (Location, Condition, Garage) encoded using LabelEncoder.
- Feature Scaling:
  - All features normalized using StandardScale.
- Train-Test Split:
  - 80% training, 20% testing (random state = 0)

3. Models Applied:

Several machine learning models are implemented, including: Linear Regression: As a baseline model to understand linear relationship between features and price. Random Forest: A more complex model that captures non-linear relationship and feature interactions. Gradient Boosting: A powerful ensemble method that iteratively improves the model by focusing on the errors of previous models. The models are evaluated using metrics like Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

Lasso Regression:

- **Train Accuracy:** 99.9999%
- **Test Accuracy:** 99.9999%

Errors:

- **Mean Squared Error (MSE):** ~0.88
- **Mean Absolute Error (MAE):** ~0.80

These extremely low error values suggest that the model is **overfitting,** possibly due to label leakage or overly clean/synthetic data.

4. Model Evaluation:

Cross-validation is performed to ensure that the model generalizes well to unseen data. Feature importance analysis is conducted to determine which features contribute most to the prediction of house prices.

Hyperparameters for models like Random Forest and Gradient Boosting are tuned using techniques like GridSearchCV to improve performance.

5. Conclusion:

The best performing model is selected based on evaluation metrics. Prediction are made on the test dataset, and the result are analyzed to determine the model's success in predicting house prices.