MAJOR PROJECT–1

HOUSE PRICE PREDICTION USING REGRESSION TECHNIQUES

1. INTRODUCTION

Accurate house price prediction is an important problem in the real estate domain. Property prices depend on several factors such as area, location, number of rooms, and available amenities. Traditional price estimation methods often fail to consider all these factors together, leading to inaccurate predictions.

This project focuses on predicting house prices using **regression models**. Different regression approaches are applied to analyze the relationship between housing features and price, and to identify the model that provides the best prediction performance.

2. OBJECTIVE OF THE PROJECT

The main objectives of this project are:

- To study the relationship between house features and price
- To preprocess housing data for regression analysis
- To apply different regression models for house price prediction
- To compare regression models and select the best-performing one
- To identify the most important factors affecting house prices

3. DATASET DESCRIPTION

- **Dataset Name:** Bangalore Housing Price Dataset
- **Total Records:** 13,320 (before preprocessing)
- **Final Records:** 12,668 (after cleaning)
- **Target Variable:** price

The dataset contains the following attributes:

- **area_type:** Indicates the type of area such as Super built-up area, Built-up area, or Plot area.
- **availability:** Specifies whether the house is ready for occupancy or available at a future date.
- **location:** Represents the locality or area in which the house is situated.
- **total_sqft:** Denotes the total area of the house measured in square feet.
- **bath:** Indicates the number of bathrooms available in the house**.**
- **balcony**: Specifies the number of balconies present in the house.

- **size:** Describes the size of the house in terms of BHK (for example, 2 BHK or 3 BHK).
- **price:** Represents the price of the house in lakhs and is used as the target variable for prediction.

4. DATA PREPROCESSING

The dataset contained missing values, inconsistent formats, and non-numeric entries. To prepare the data for regression modeling, the following steps were performed:

1. Removed columns with excessive missing values
2. Removed rows containing remaining missing values
3. Extracted numerical **BHK** values from the size column
4. Converted total_sqft values given as ranges into numeric values
5. Created a new feature called **price_per_sqft**

After preprocessing, the dataset was cleaned and suitable for regression analysis.

5. EXPLORATORY ANALYSIS

Initial analysis was performed to understand the relationship between input features and house price.

Observations

- House prices increase with an increase in total square feet
- Higher BHK houses generally have higher prices
- Location plays a significant role in price variation
- Some extreme values were observed in price-related features

These observations helped in improving data quality before model building.

6. OUTLIER REMOVAL

Outliers can negatively affect regression models. Two types of outliers were identified and removed:

1. **Price per Square Foot Outliers:**
   Extreme price_per_sqft values were removed using statistical limits.
2. **BHK-Based Anomalies:**
   Houses where higher BHK prices were lower than smaller BHK prices in the same location were removed.

This step improved the reliability of regression results.

7. REGRESSION MODELS USED

The following regression models were implemented to predict house prices:

1. **Linear Regression**
2. **Decision Tree Regression**
3. **Random Forest Regression**

The dataset was divided into **80% training data** and **20% testing data**.

8. MODEL EVALUATION

The regression models were evaluated using the following metrics:

- **Mean Absolute Error (MAE)**
- **Root Mean Squared Error (RMSE)**
- **R^2 Score**

Performance Comparison

Linear Regression showed basic prediction capability. Decision Tree Regression performed better by modeling non-linear patterns. Random Forest Regression achieved the highest accuracy with the lowest error values, making it the best performing model for house price prediction.

9. FEATURE IMPORTANCE

Feature importance analysis was performed using the Random Forest Regression model.

Top Features Influencing Price

1. Total square feet
2. Location
3. Number of BHK
4. Number of bathrooms
5. Price per square foot

10. RESULTS AND DISCUSSION

The results show that regression models can effectively predict house prices when trained on cleaned and well-structured data. Among all models tested, Random Forest Regression performed the best due to its ability to handle non-linear relationships and reduce overfitting.

11. CONCLUSION

This project successfully applied **regression methods** to predict house prices using a real-world dataset. Proper preprocessing and outlier removal significantly improved model accuracy. Random Forest Regression was found to be the most suitable model for this problem.

12. FUTURE SCOPE

- Inclusion of additional location-specific features
- Analysis of time-based price trends
- Use of advanced regression techniques
- Deployment as a prediction system