DATA ANALYSIS AND DATA SCIENCE WITH PYTHON

Task 4: Regression Analysis

Objective: Build a regression model to predict house prices based on various features using linear regression.

Steps to Complete the Project

---

1. Dataset Selection

- **Dataset Name**: `house_prices.csv`
- **Key Columns**:
  - **Size**: Numeric (e.g., in square feet).
  - **Location**: Categorical (e.g., urban, suburban, rural).
  - **Number of Rooms**: Numeric.
  - **Price**: Numeric (target variable).

---

2. Tasks to Perform

1. Load and Explore

- **Inspect the Dataset**:
  - Check for missing values and handle them appropriately.
  - Analyze distributions of numerical variables (e.g., `Size`, `Price`).
  - Identify potential outliers that might skew results.

2. Data Preprocessing

- **Normalize Numerical Data**:
  - Scale features like `Size` and `Number of Rooms` to bring them to a comparable range using methods like Min-Max Scaling or Standardization.
- **Encode Categorical Features**:
  - Convert `Location` into numerical values using methods such as:
    - One-Hot Encoding for non-ordinal categories.
    - Label Encoding for ordinal categories (if any).

3. Feature Selection

Main Flow Services and Technologies Pvt. Ltd.
Contact Us. +91 9389641586, +91 97736 99074
Email-Add. contact.mainflow@gmail.com
www.mainflow.in

- **Analyze Predictors**:
    - Use correlation analysis to identify relationships between features and the target variable (`Price`).
    - Consider removing low-impact predictors to improve model performance.

4. Model Training

- **Train-Test Split**:
    - Divide the dataset into training and testing sets (e.g., 80% train, 20% test). Ensure the split is random but reproducible.
- **Train a Linear Regression Model**:
    - Use libraries like `scikit-learn` or similar tools to fit the regression model.

5. Model Evaluation

- **Evaluation Metrics**:
    - Calculate **Root Mean Square Error (RMSE)** to measure prediction accuracy.
    - Determine **R^2 (Coefficient of Determination)** to evaluate how well the model explains variability in the data.

---

3. Deliverables

1. **Trained Regression Model**:
    - A fitted linear regression model capable of predicting house prices.
2. **Predictions**:
    - Outputs for the test data including predicted vs. actual prices.
3. **Evaluation Metrics**:
    - RMSE and R^2 values for model performance.
4. **Feature Insights**:
    - Summary of the most important predictors influencing house prices.

Expected Insights

- How features like `Size` or `Number of Rooms` correlate with house prices.
- The effect of `Location` on pricing.
- Accuracy and reliability of the regression model for predicting real-world house prices.

Would you like a more detailed breakdown of any specific step?

Main Flow Services and Technologies Pvt. Ltd.
Contact Us. +91 9389641586, +91 97736 99074
Email-Add. contact.mainflow@gmail.com
www.mainflow.in

Deadline Compliance

- **Restriction**: **Submit the project within 7 days** from the start date.
- **Reason**: Meeting deadlines is crucial in the real-world software development environment. This restriction helps students practice **time management** and **task prioritization**. In professional settings, tight deadlines are often the norm, and learning to meet them without compromising quality is an essential skill.
- **Learning Outcome**: Students will learn to manage their time effectively, complete projects under pressure, and **deliver results on time**, which are all important skills in the workplace.