Linear Regression on California Housing Dataset

Name: Nirmalya Ghosh
Task1: Build & Evaluate a Linear Regression Model (House Price Predictor)
Dataset: California Housing Dataset
Tool: Python, scikit-learn

---

1. Introduction

This project demonstrates an end-to-end machine learning workflow using **Linear Regression** to predict median house values in California.
The objective is to understand the relationship between socio-economic and geographical features and housing prices, perform exploratory data analysis, train a regression model, and evaluate its performance using standard metrics.

Linear Regression was chosen as it is a simple and interpretable algorithm that serves as a strong baseline for regression problems.

---

2. Dataset Description

The dataset used in this project is the **California Housing Dataset**, which is available directly through the scikit-learn library.

- **Source:** scikit-learn
- **Number of instances:** 20,640
- **Number of features:** 8
- **Target variable:** Median House Value (**MedHouseVal)**

Features

- Median Income
- House Age
- Average Rooms
- Average Bedrooms
- Population
- Average Occupancy
- Latitude
- Longitude

The target variable represents the **median house value** for California districts.

---

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to understand the structure and characteristics of the dataset.

Key Observations

- No missing values were found in the dataset.
- The dataset contains both numerical and geographical features.
- Summary statistics were examined using descriptive measures such as mean, standard deviation, minimum, and maximum values.
- A **correlation heatmap** was used to analyze relationships between features.

Insights

- **Median Income** shows a strong positive correlation with median house value.
- Geographical features (latitude and longitude) also influence housing prices.
- Some features show weak correlation, indicating possible non-linear relationships.

---

4. Model and Methodology

Model Used

- **Algorithm:** Linear Regression
- **Library:** scikit-learn

Data Splitting

The dataset was split into:

- **80% training data**
- **20% testing data**

This ensures that the model is evaluated on unseen data and helps prevent overfitting.

Training Process

- The Linear Regression model was trained using the training dataset.
- The model learned coefficients for each feature to minimize prediction error.

Linear Regression was selected because it is easy to interpret and provides a baseline for future improvements.

---

5. Model Evaluation

The model was evaluated using the following metrics:

Metric	Value
Mean Absolute Error (MAE)	0.533
Root Mean Squared Error (RMSE)	0.746
R^2 Score	0.576

Interpretation

- **MAE (0.533):** On average, the model's predictions differ from the actual values by approximately 0.53 units.
- **RMSE (0.746):** Higher than MAE, indicating that some predictions have larger errors.
- **R^2 (0.576):** The model explains approximately **58% of the variance** in house

prices. Overall, the model performs reasonably well for a baseline linear model.

---

6. Visual Analysis

To better understand model performance, the following plots were used:

- **Actual vs Predicted Values Plot:**
  Shows how close the predicted values are to the actual house prices.
- **Residual Plot:**
  Displays the difference between actual and predicted values.
  Residuals were randomly scattered around zero, indicating no strong systematic error.

These visualizations confirm that the Linear Regression model captures general trends in the data.

---

7. Conclusion

This project successfully implemented an end-to-end machine learning pipeline using Linear Regression on the California Housing dataset.
The model achieved reasonable performance and demonstrated the importance of exploratory data analysis and proper evaluation.

Linear Regression serves as a strong baseline, and further improvements can be achieved using regularization techniques and non-linear models.