Task 1: Linear Regression – California Housing Dataset

1. Introduction

This project focuses on building a beginner-level Machine Learning model using Linear Regression. The objective is to understand the complete ML workflow including data loading, exploration, preprocessing, model training, evaluation, and saving the trained model. The California Housing dataset provided by scikit-learn is used to predict median house prices.

2. Dataset Description

The California Housing dataset contains data collected from the 1990 California census. It includes numerical features such as median income, house age, average number of rooms, population, and geographical information like latitude and longitude. The target variable is the median house value.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to understand the structure and distribution of the data. Summary statistics such as mean, standard deviation, and quartiles were examined. A histogram of the target variable showed that house prices are slightly right-skewed. No missing values were found in the dataset.

4. Model Building

The dataset was split into training and testing sets using an 80-20 ratio. A Linear Regression model from scikit-learn was used. This algorithm was chosen due to its simplicity and suitability for beginner-level learning.

5. Model Evaluation

The trained model was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared score. These metrics help in understanding prediction accuracy and how well the model explains variance in the data. The results indicate that the model performs reasonably well for a basic regression approach.

6. Model Saving and Prediction

After training, the model was saved using the pickle library. A separate Python script was created to load the saved model and predict house prices using new input data. This demonstrates how trained models can be reused in real-world applications.

7. Conclusion and Future Improvements

This project successfully demonstrates the complete machine learning workflow using Linear Regression. Future improvements may include feature scaling, trying advanced regression models such as Ridge or Lasso, and using non-linear models to improve prediction accuracy.