Linear Regression on California Housing Dataset

Introduction

This report presents the implementation and evaluation of a Linear Regression model to predict house prices using the California Housing dataset. Linear Regression is a supervised machine learning algorithm used for predicting a continuous dependent variable based on one or more independent variables. In this task, a Linear Regression model is used to predict median house prices and understand the machine learning workflow including data loading, exploration, training and evaluation.

Dataset Description

The California Housing dataset is a built-in dataset available in the scikit-learn library. It contains 20,640 records with 9 columns (8 features and 1 target variable) is displayed. such as median income, house age, average rooms, population, and geographical coordinates. No missing or null values are observed in the dataset.

Exploratory Data Analysis

- Statistical summaries (mean, median, standard deviation) for all numerical features are generated.
- A correlation heatmap is produced showing relationships between features.
- Median Income is observed to have the highest positive correlation with median house value.
- Visualization confirms the importance of location-based features such as latitude and longitude.

Model & Evaluation

The performance of the Linear Regression model was evaluated using the following metrics:-

- Mean Absolute Error (MAE) ≈ Indicates average error in house price prediction
- Root Mean Squared Error (RMSE) ≈ penalizes large prediction errors
- R-squared (R^2) ≈60% variance explained by model

The obtained results show that the model explains approximately 60% of the variance in house prices, indicating a reasonable performance for a basic linear model.

Metric	Value
MAE	~0.53
RMSE	~0.74
R^2 Score	~0.57

Conclusion

The expected outputs confirm that the Linear Regression model is correctly implemented, evaluated, and visualized. The results demonstrate that Linear Regression provides a reasonable baseline performance for predicting housing prices and serves as a foundation for further model improvement.