TASK 1

House Price Prediction using Linear Regression

Introduction:

This project demonstrates a basic machine learning workflow using the California Housing dataset.

The goal of this project is to build and evaluate a Linear Regression model to predict median house prices.

The project covers data exploration, model training, and performance evaluation.

Exploratory Data Analysis (EDA):

The California Housing dataset contains information about housing features such as median income, house age, average number of rooms, population, and location details.

Initial exploration showed that the dataset contains no missing values and all features are numerical.

During visualization, the distribution of house prices showed that most values are concentrated in the lower to mid range.

Correlation analysis indicated that median income has a strong positive relationship with house prices.

Model Description:

In this project, a Linear Regression model was used to predict median house prices.

Linear Regression is a simple and widely used machine learning algorithm that models the relationship between input features and a target variable.

It was chosen as a baseline model because it is easy to understand and interpret.

Model Evaluation & Metrics:

The performance of the Linear Regression model was evaluated using three standard metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score.

The model achieved an MAE of approximately 0.53, which indicates the average difference between predicted and actual house prices.

The RMSE value was around 0.75, showing that larger prediction errors are penalized more strongly.

The R^2 score was approximately 0.58, meaning the model explains about 58% of the variation in house prices.

Overall, the results show that the model performs reasonably well for a baseline approach.

Conclusion & Improvement Ideas:

This project demonstrated a complete machine learning workflow using the California Housing dataset.

A Linear Regression model was successfully built and evaluated, achieving reasonable performance as a baseline model.

In the future, the model's performance could be improved by applying feature scaling, experimenting with regularized regression techniques such as Ridge or Lasso, or using more advanced machine learning models.