PROJECT REPORT

House Price Prediction using Linear Regression

---

1. Introduction

Machine Learning is widely used today to solve real-world prediction problems. One such application is predicting house prices based on various features like income level, house age, population, and location.

In this project, a Linear Regression model was built to predict house prices using the California Housing dataset. The goal of this task was to understand the complete machine learning workflow — from loading data to evaluating model performance.

---

2. Objective

The main objective of this project is to develop a predictive model that can estimate house prices based on given input features.

Through this project, the following concepts were learned:

- Data loading and preprocessing
- Exploratory Data Analysis (EDA)
- Model training using Linear Regression
- Performance evaluation using error metrics

---

3. Dataset Description

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

It contains information collected from different districts of California, including:

- Median Income
- House Age
- Average Number of Rooms
- Average Number of Bedrooms
- Population
- Households
- Latitude
- Longitude

The target variable is:

Median House Value

This represents the price of houses in a particular area.

---

4. Exploratory Data Analysis (EDA)

Before training the model, the dataset was explored to understand its structure and relationships between features.

The following steps were performed:

- Checked for missing values
- Studied statistical summary of data
- Observed feature distributions
- Analyzed correlation between variables

Correlation analysis helped in understanding how strongly different features influence house prices.

---

5. Model Used

For this project, **Linear Regression** was used.

Linear Regression is one of the simplest and most commonly used machine learning algorithms for prediction tasks. It works by finding the best linear relationship between input features and the target variable.

The dataset was divided into:

- Training Data (80%)
- Testing Data (20%)

The model was trained using the training dataset and tested on unseen data.

---

6. Model Evaluation

To measure the performance of the model, the following evaluation metrics were used:

Mean Absolute Error (MAE)

Measures the average difference between actual and predicted values.

Root Mean Squared Error (RMSE)

Gives an idea of how large the prediction errors are.

R^2 Score

Shows how well the model explains the variation in house prices.

MAE : 0.53

RMSE : 0.73

R^2 Score : 0.6

---

7. Visualization

A scatter plot was created to compare:

Actual house prices
Predicted house prices

This helped visualize how close the predictions were to real values.

If the model performs well, the plotted points appear close to a straight line.

(Add your graph screenshot in report)

---

8. Conclusion

The Linear Regression model was successfully implemented to predict house prices using the California Housing dataset.

The model showed reasonable prediction capability based on the evaluation metrics. This project provided a clear understanding of how machine learning models are built and evaluated.

---

9. Future Improvements

The model performance can be improved further by:

- Applying feature scaling
- Using advanced regression models like Ridge or Lasso
- Performing hyperparameter tuning
- Trying non-linear algorithms

---

10. Tools & Technologies Used

- Python
- Pandas
- NumPy
- Scikit-learn
- Matplotlib

- Seaborn
- Jupyter Notebook