AI & ML Task 2

Feature Engineering, Model Optimization & Performance Comparison

1. Introduction

The objective of this task was to enhance a regression-based Machine Learning model for predicting house prices using the California Housing Dataset. Unlike basic model training, this task focuses on applying feature engineering techniques, training multiple algorithms, and performing structured performance comparison.

The goal was not only to build a predictive model but also to analyze how preprocessing and algorithm selection impact performance in real-world ML workflows.

2. Dataset Description

The California Housing Dataset contains real-world housing-related data collected from California districts.

Target Variable:

- **Median House Value**

Input Features:

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

3. Methodology

3.1 Data Preparation

The dataset was loaded using sklearn.datasets.fetch_california_housing.
Features (X) and target variable (y) were separated for supervised learning.

A train-test split (80% training, 20% testing) was applied to ensure that model evaluation was performed on unseen data.

3.2 Feature Scaling

Since regression algorithms such as Linear Regression and Ridge Regression are sensitive to feature magnitude, **StandardScaler** was applied.

StandardScaler transforms features so that:

- Mean = 0
- Standard Deviation = 1

This ensures:

- Equal contribution of all features
- Faster convergence
- Improved numerical stability

3.3 Models Trained

Three regression models were trained and compared:

1️⃣ Linear Regression

- Serves as a baseline model
- Assumes linear relationship between input features and target

2️⃣ Ridge Regression

- Extension of Linear Regression
- Adds L2 regularization
- Helps reduce overfitting
- Improves generalization

3️⃣ Decision Tree Regressor

- Captures non-linear relationships
- Splits data into hierarchical decision nodes
- Can overfit if tree depth is not controlled

4. Evaluation Metrics

To evaluate model performance, the following metrics were used:

Root Mean Squared Error (RMSE)

- Measures average prediction error magnitude

- Lower value indicates better accuracy

R^2 Score (Coefficient of Determination)

- Measures how well the model explains variance in target variable
- Value ranges between 0 and 1
- Higher value indicates better explanatory power

5. Results & Model Comparison

Observations:

- Ridge Regression showed slightly improved performance compared to Linear Regression due to regularization.
- Decision Tree captured non-linear patterns but showed signs of higher variance.
- Feature scaling improved stability for linear models.

6. Visual Validation

An Actual vs Predicted scatter plot was generated to visually validate model performance.

Observations from the plot:

- Points closer to the diagonal line indicate better predictions.
- Some deviation exists at higher house values, suggesting limitations in modeling extreme price ranges.
- Ridge Regression demonstrated consistent prediction spread.

7. Conclusion

This task demonstrates the importance of:

- Proper data preprocessing
- Feature scaling
- Training multiple models
- Objective performance comparison

Among the tested models, Ridge Regression provided the best generalization performance due to regularization. Feature scaling significantly improved learning stability.

This workflow reflects real-world Machine Learning practices where model optimization and systematic comparison are essential before deployment.