Task 2: Feature Engineering, Model Optimization & Performance Comparison

1. Introduction

This task builds upon Task 1 by extending the machine learning workflow to include feature engineering, model optimization, and performance comparison. The goal is to understand how preprocessing techniques and different regression models impact prediction accuracy. The California Housing dataset is used for all experiments.

2. Dataset Overview

The California Housing dataset contains information collected from the 1990 census in California. It includes numerical features such as median income, house age, average number of rooms, population, average occupancy, and geographical coordinates. The target variable is the median house value. The dataset does not contain missing values, making it suitable for regression tasks.

3. Feature Engineering & Scaling

Feature engineering in this task focuses primarily on feature scaling. StandardScaler was used to normalize numerical features so that they have zero mean and unit variance. Feature scaling is essential for models such as Linear Regression and Ridge Regression because these algorithms are sensitive to the scale of input features.

4. Model Training

Three regression models were trained and evaluated: Linear Regression, Ridge Regression, and Decision Tree Regressor. Linear Regression serves as a baseline model. Ridge Regression introduces regularization to reduce overfitting by penalizing large coefficients. Decision Tree Regression is a non-linear model capable of capturing complex relationships without requiring feature scaling.

5. Model Evaluation & Comparison

Model performance was evaluated using Root Mean Squared Error (RMSE) and R-squared (R^2) score. RMSE measures the average prediction error magnitude, while R^2 indicates how well the model explains variance in the target variable. The results showed that Ridge Regression performed slightly better than basic Linear Regression, while Decision Tree Regression achieved competitive performance but showed signs of potential overfitting.

6. Visualization

An Actual vs Predicted scatter plot was created to visually inspect model performance. A strong linear trend between actual and predicted values indicates good model fit. This visualization helps in identifying systematic prediction errors.

7. Conclusion & Future Improvements

This task demonstrates the importance of feature engineering and model comparison in machine learning. Ridge Regression provided the best balance between simplicity and performance for this dataset. Future improvements may include hyperparameter tuning, cross-validation, and experimenting with advanced ensemble models such as Random Forests or Gradient Boosting.