Artificial Intelligence & Machine Learning – Task 2

Feature Engineering, Model Optimization & Performance Comparison

1. Objective

The objective of this task is to understand how machine learning models are improved in real-world scenarios by applying proper data preprocessing techniques, training multiple models, evaluating their performance, and selecting the best-performing model.

This task focuses on moving beyond basic model training and adopting an industry-level machine learning workflow.

2. Dataset Used

The **California Housing Dataset** was used for this task.

It contains real-world housing data collected from California districts.

- **Target Variable:** Median House Value (HousePrice)
- **Input Features:** Median income, house age, average rooms, population, and location-based attributes

This dataset is suitable for regression-based machine learning models.

3. Data Preprocessing & Feature Scaling

Before training the models, the dataset was preprocessed by separating input features and the target variable.

Since the features were present on different numerical scales, **StandardScaler** was applied to normalize the data. Feature scaling ensures that all features contribute equally to model learning and improves stability and performance.

4. Machine Learning Models Used

To identify the best-performing model, multiple regression algorithms were trained and evaluated:

- **Linear Regression** – Used as a baseline model
- **Ridge Regression** – Helps reduce overfitting by applying regularization
- **Decision Tree Regressor** – Captures non-linear relationships in the data

Training multiple models allows objective comparison instead of relying on a single algorithm.

5. Model Evaluation Metrics

Each model was evaluated using the following metrics:

- **RMSE (Root Mean Squared Error):** Measures prediction error (lower is better)
- **R^2 Score:** Measures how well the model explains the variance in data (higher is better)

These metrics help in comparing model performance accurately.

6. Model Comparison & Results

After evaluation, all models were compared using RMSE and R^2 scores.

A comparison table was generated to clearly analyze performance differences.

Based on the comparison results, the model with the **lowest RMSE** and **highest R^2 score** was considered the best-performing model.

7. Visual Validation

An **Actual vs Predicted House Prices** scatter plot was created to visually validate the model's performance.

Points closer to the diagonal line indicate better prediction accuracy.

This visualization helps in understanding how well the model generalizes to unseen data.

8. Conclusion

In this task, feature scaling was successfully applied, multiple regression models were trained and evaluated, and the best-performing model was selected based on objective metrics.

This task demonstrates a complete, industry-aligned machine learning workflow and highlights the importance of preprocessing, model comparison, and performance evaluation.