Artificial Intelligence & Machine Learning – Task 2

Feature Engineering, Model Optimization & Performance Comparison

---

1. Introduction

The purpose of this task is to design and evaluate a machine learning solution for predicting house prices using a real-world dataset. The task emphasizes the importance of data preprocessing, feature scaling, model selection, and performance evaluation. By implementing and comparing multiple regression models, the most suitable model for accurate prediction is identified.

This task provides practical exposure to the complete machine learning workflow, from dataset handling to final model selection, while focusing on performance comparison using standard evaluation metrics.

---

2. Dataset Description

The **California Housing Dataset** is used in this task. It contains housing-related information collected from the California census. Each data entry represents aggregated information for a specific district.

The dataset includes features such as median income, house age, average number of rooms, population, and geographical attributes. The target variable is the **median house value**, which the models attempt to predict based on the given features.

This dataset is widely used for regression problems and is suitable for evaluating the performance of different machine learning models.

---

3. Data Preprocessing and Feature Scaling

Before training the models, the dataset was prepared to ensure accurate and reliable results. The data was first divided into input features and the target variable.Feature scaling was applied using **StandardScaler** to normalize the feature values. This step is essential because the dataset contains features with different numerical ranges. Without scaling, models may become biased toward features with larger values.After scaling, the dataset was split into training and testing sets using an 80:20 ratio. The training set was used for model learning, while the testing set was reserved for unbiased evaluation.

4. Machine Learning Models Implemented

Three regression models were implemented and compared:

Linear Regression was used as a baseline model due to its simplicity and interpretability. It assumes a linear relationship between input features and the target variable.

Ridge Regression was implemented to address potential overfitting by introducing regularization. This helps control large coefficient values and improves model generalization.

Decision Tree Regressor was used to capture non-linear relationships in the data. To avoid overfitting, the depth of the tree was limited during training.

---

5. Model Evaluation and Comparison

The performance of each model was evaluated using two standard metrics:

- **Root Mean Squared Error (RMSE):** Measures the average prediction error. Lower RMSE values indicate better accuracy.

- **R^2 Score:** Indicates how well the model explains the variance in the target variable. Higher values represent better performance.

After evaluation, the models were compared based on these metrics. The results showed that Linear Regression achieved low prediction error and consistent performance, while Ridge Regression showed similar but slightly less effective results. The Decision Tree model showed signs of overfitting despite parameter control.

---

6. Final Model Selection and Conclusion

Based on the comparative analysis, **Linear Regression** was selected as the final model for house price prediction. It demonstrated stable performance, lower error, and good generalization on the test dataset.In conclusion, this task successfully demonstrated the importance of proper data preprocessing, feature scaling, and model evaluation. The comparison highlighted that simpler models can perform effectively when trained and evaluated correctly. This task provided valuable hands-on experience in applying machine learning techniques to real-world data.