Model Comparison for House Price Prediction

Task 2 – AI & Machine Learning Internship

Name: Aryan Marota
Internship Organization: Maincraft Technologies
Task Title: Model Comparison for House Price Prediction
Dataset: California Housing Dataset (Scikit-learn)

---

1. Introduction

House price prediction is a regression problem where the goal is to estimate the value of a house based on various socio-economic and geographical features. While a single model can provide reasonable predictions, different machine learning models often perform differently on the same dataset.

The objective of this task is to **train multiple regression models**, **compare their performance**, and **identify the best-performing model** for house price prediction using the California Housing dataset. This task builds upon the previous work by extending the analysis from a single model to a comparative evaluation.

---

2. Dataset Description

The dataset used for this task is the **California Housing Dataset**, which is available in the Scikit-learn library. It contains data collected from the 1990 California census and represents housing information for different districts.

Dataset characteristics:

- Total records: **20,640**
- Input features: **8**
- Target variable: **HousePrice**

Key features include:

- Median Income
- House Age
- Average Rooms and Bedrooms
- Population
- Average Occupancy
- Latitude and Longitude

All features are numerical, and the dataset does not contain missing values. This makes it suitable for applying regression models without additional data cleaning.

---

3. Methodology

3.1 Feature Scaling

Since the input features exist on different numerical scales, **StandardScaler** was applied to normalize the feature values. Feature scaling is especially important for linear and regularized models to ensure stable and fair learning.

3.2 Train-Test Split

The dataset was split into:

- **80% training data**
- **20% testing data**

This ensures that model performance is evaluated on unseen data.

---

4. Models Used

The following regression models were trained and evaluated:

1. **Linear Regression** – Used as a baseline model
2. **Ridge Regression** – A regularized linear model to reduce overfitting
3. **Decision Tree Regressor** – A non-linear model capable of capturing complex relationships

Each model was trained using the same training data and evaluated using the same test set for fair comparison.

---

5. Model Evaluation Metrics

The models were evaluated using the following metrics:

- **Mean Absolute Error (MAE):** Measures average absolute prediction error
- **Root Mean Squared Error (RMSE):** Penalizes larger errors more heavily
- **R^2 Score:** Indicates how well the model explains the variance in house prices

---

6. Model Comparison Results

The performance of the models is summarized below:

Model	MAE	RMSE	R^2 Score
Linear Regression	0.5332	0.7456	0.5758

Model	MAE	RMSE	R^2 Score
Ridge Regression	0.5332	0.7456	0.5758
Decision Tree	**0.5223**	**0.7242**	**0.5997**

Observation

- Linear Regression and Ridge Regression showed very similar performance.
- The **Decision Tree model achieved the lowest RMSE and the highest R^2 score**, indicating better predictive performance compared to linear models.

---

7. Best Model Selection and Visualization

Based on the evaluation metrics, the **Decision Tree Regressor** was selected as the best-performing model.

The **Actual vs Predicted** scatter plot for the Decision Tree model shows that predictions generally follow the ideal diagonal line, especially for mid-range house prices. However, some dispersion is observed, which is expected due to the complexity of housing data.

The improved performance of the Decision Tree suggests that non-linear relationships exist in the dataset that linear models are unable to capture effectively.

---

8. Conclusion

In this task, multiple regression models were trained and compared for house price prediction using the California Housing dataset. Among the models tested, the **Decision Tree Regressor** performed the best, achieving the lowest RMSE and highest R^2 score.

This comparison highlights the importance of evaluating multiple models rather than relying on a single approach. While linear models provide simplicity and interpretability, tree-based models can better handle non-linear patterns in real-world data.

---

9. Future Scope

- Experiment with deeper or ensemble tree-based models such as Random Forest
- Perform hyperparameter tuning to improve model performance
- Compare results with regularized and non-linear regression techniques