Model Validation, Overfitting & Hyperparameter Tuning

1. Dataset and Tools

The California Housing dataset (1990 census) is a regression dataset used to predict median house prices. It includes features like median income, house age, number of rooms, population, and location details.

Python was used with Pandas, NumPy, Matplotlib, and Scikit-learn. StandardScaler was applied for feature scaling. Three models were compared: Linear Regression, Ridge Regression, and Decision Tree Regressor (with tuning).

---

2. Overfitting and Tuning

The initial Decision Tree showed overfitting, as training error was lower than testing error. Cross-validation using RMSE was applied to check generalization.

GridSearchCV was used to tune parameters such as maximum depth and minimum samples split, which reduced overfitting and improved stability.

---

3. Model Performance

Model	RMSE	R^2 Score
Linear Regression	0.745581	0.575788
Ridge Regression	0.745554	0.575819
Tuned Decision Tree	0.645430	0.682099

---

4. Final Model

The tuned Decision Tree achieved the lowest RMSE (0.645430) and highest R^2 (0.682099), making it the best model.

This project covers preprocessing, overfitting detection, cross-validation, hyperparameter tuning, and model comparison in a complete ML workflow.