House Price Prediction:

What I Implemented

In this assignment, I implemented a house price prediction model using two machine learning algorithms: Linear Regression (LR) and Random Forest Regressor (RF).

I first loaded the clean_house_dataset.csv file and split it into features (X) and target (y), where the target was the Price column and all other columns except LogPrice were used as features. The dataset was split into 80% training and 20% testing using train_test_split with a random_state of 42 to ensure reproducibility.

I trained a LinearRegression model and a RandomForestRegressor (with n_estimators=100 and random_state=42) using the training data. After training, I made predictions on the test set and evaluated the performance of both models using the metrics: R^2, MAE, MSE, and RMSE.

Comparison of Models

To compare their real-world prediction behavior, I selected three different rows from the test set (using .iloc) for a sanity check.

•For some samples, Linear Regression produced predictions that were too high or too low compared to the actual prices, showing that it struggled when the house features were very different from the average pattern.

•Random Forest predictions were closer to the actual prices for most of the chosen samples. It captured more variation and handled outlier-like data points better.

This showed that while both models learned general trends, Random Forest was more accurate and consistent in individual predictions.

Understanding Random Forest

Random Forest is an **ensemble machine learning method** that builds **many decision trees** during training and **averages their predictions** for regression tasks (or takes the majority vote for classification).

- Each decision tree is trained on a random subset of the data and random subsets of features.
- This randomness reduces overfitting and improves generalization.
- When predicting, each tree outputs a price, and the forest takes the **average** of all these outputs as the final prediction.

This ensemble approach makes Random Forest more stable and accurate compared to a single decision tree or a simple linear model.

Metrics Discussion

When comparing the performance metrics:

- **Random Forest achieved higher R² scores** (closer to 1), meaning it explained more variance in the target prices.
- **Random Forest had lower MAE and RMSE** compared to Linear Regression, indicating it made smaller prediction errors on average.
- **Linear Regression had higher errors** and struggled when relationships between features and price were non-linear.

This shows that **Linear Regression works well if the relationship is strictly linear and simple**, while **Random Forest is stronger when relationships are complex and non-linear**.

My Findings

Based on my results, I prefer **Random Forest** for house price prediction.
It produced **more accurate predictions (higher R², lower error metrics)** and handled the diversity of house features much better than Linear Regression. It is more reliable for real-world datasets where patterns are complex and not perfectly linear.

Although Random Forest is more computationally expensive and less interpretable than Linear Regression, the **improved accuracy and stability make it the better choice** for this problem.