AI/ML Entry-Level Project: House Price Prediction

Project Goal

Build a machine learning model to predict house prices based on different features such as size, location, and number of rooms. The project should demonstrate data preprocessing, model training, evaluation, and reporting of insights**.**

Dataset

https://www.kaggle.com/datasets/lespin/house-prices-dataset/data

- Target variable: SalePrice
- Features

Project Intstructions

1. Data Understanding & Cleaning
    a. Load dataset, inspect rows/columns.
    b. Handle missing values (drop/fill).
    c. Convert categorical variables (e.g., Neighborhood) into numeric (One-Hot Encoding).
    d. Check for outliers (e.g., huge lot size, unrealistic prices).
    e. Normalise / scale features where needed.

2. Exploratory Data Analysis (EDA)
    a. Summary statistics of numeric features.
    b. Correlation heatmap (which features affect price).
    c. Visualisations: scatter plot of LotArea vs SalePrice, boxplot of Neighborhood vs SalePrice.
    d. Identify most important factors (at least 3).

3. Model Building
    a. Train/Test Split (e.g., 80/20).
    b. Train at least 3 models:
        i. Linear Regression (baseline)
        ii. Decision Tree Regressor
        iii. Random Forest Regressor
    c. Evaluate models using:
        i. RMSE (Root Mean Square Error)
        ii. MAE (Mean Absolute Error)
        iii. R^2 Score (variance explained)

d. Compare results and select the best model.

4. Model Tuning & Feature Importance
    a. Tune hyperparameters of the best model (e.g., tree depth, number of estimators).
    b. Analyse feature importance (which features drive prices most).
    c. Visualise top features (bar chart).

5. Reporting
    a. Write a short project report
        i. Introduction (project goal).
        ii. Dataset description.
        iii. Data preprocessing.
        iv. EDA findings (with visuals).
        v. Model performance comparison table.
        vi. Insights (e.g., "Location and house quality explain 70% of price variance").
        vii. Conclusion.

Deliverables

1. Cleaned dataset (CSV).
2. Jupyter Notebook (or Google Colab) with code and outputs.
3. Visualisations (charts, heatmaps, feature importance).
4. Project report (Word/PDF).