

DATA SCIENCE PROJECT: HOUSE PRICE PREDICTION & FUTURE FORECASTING

Assuming you were a data analyst in a real estate firm, your task is to analyze past housing data, build a predictive model for house prices, and use it to forecast prices for the next 40 years. In this project, you will train and evaluate machine learning models, then apply the best-performing model to predict the prices of 500 future houses.

The purpose of this project is to assess your data science abilities, including data exploration, visualization, model development, and forecasting.

You are provided with a housing dataset containing the following columns:

- **area** (size of the house in sq. ft)
- **bedrooms**
- **bathrooms**
- **stories**
- **parking**
- **Year** (year the house was sold)
- **price**

Project Tasks

Your task is to analyze the dataset, build a predictive model, evaluate it, and finally **predict the prices of houses for the next 40 years** using a second dataset (provided separately) that contains **500 future samples without prices**.

1. Load the housing dataset using **pandas** and display the first 10 rows.
2. Check for:
 - Missing values
 - Summary statistics (mean, median, std, etc.)
3. Explore the relationship between features and price using:
 - Correlation matrix
 - Scatter plots (e.g., area vs price, bedrooms vs price)
4. Make Use of Cross Validations
 - `cross_val_score`,
 - GridSearch CV
 - RandomizedSearchCV
5. Split the dataset into:
 - Feature matrix **X**
 - Target vector **y**
 - Perform **train-test split** (80% training, 20% testing).
6. Train a regression model using **scikit-learn** (choose *at least one*):
 - Linear Regression
 - Random Forest Regressor
 - Gradient Boosting Regressor
7. Evaluate the model using:
 - MAE
 - MSE
 - RMSE

- R^2 score
8. Plot:
 - Actual vs Predicted prices (scatter plot)
 9. Load the **future housing dataset (2026–2065)** containing 500 samples without prices.
 10. Use the trained model to predict the prices of all 500 future houses.
 11. Add the predicted prices to the new dataset and save the results.
 12. Plot:
 - Price forecast distribution
 - Yearly average predicted price trend for 2026–2065.