PHASE-2

GitHub link :

Project Title: Forecasting house prices accurately using smart regression techniques in data science

---

1. Problem Statement

Accurately forecasting house prices is critical for buyers, sellers, real estate investors, and policy-makers. However, the housing market is influenced by numerous complex factors such as location, property characteristics, and economic indicators. The goal of this project is to develop a data-driven solution that utilizes smart regression techniques to predict house prices based on various features.

---

2. Project Objectives

- To analyze and understand the factors influencing house prices.
- To preprocess and prepare data suitable for predictive modeling.
- To apply advanced regression techniques for accurate price prediction.
- To evaluate model performance using appropriate metrics.

- To present insights through visualizations and provide a reliable prediction model.

---

3. Flowchart of the Project Workflow

Flow:
Data Collection → Data Cleaning & Preprocessing → EDA → Feature Engineering → Model Selection & Training → Evaluation → Visualization → Deployment

---

4. Data Description

The dataset used for this project includes features such as:

- SalePrice (Target)
- LotArea, YearBuilt, OverallQual, OverallCond
- Neighborhood, GarageType, TotalBsmtSF, GrLivArea
- Categorical & numerical features related to house attributes and sale conditions.

Source: Kaggle's Ames Housing Dataset (or similar)

---

5. Data Preprocessing

- Handling missing values (imputation strategies)
- Encoding categorical variables (Label Encoding / One-Hot Encoding)
- Scaling numerical features (StandardScaler / MinMaxScaler)

- Outlier detection and removal
- Feature transformation (e.g., log transformation of skewed data)

---

6. Exploratory Data Analysis (EDA)

Univariate Analysis:

- Histograms of key numerical features
- Boxplots to detect outliers
- Frequency plots of categorical variables

Bivariate & Multivariate Analysis:

- Correlation heatmaps
- Pairplots and scatter plots (e.g., GrLivArea vs SalePrice)
- ANOVA or bar charts for categorical features vs SalePrice

Key Insights:

- OverallQual, GrLivArea, and YearBuilt show strong correlation with SalePrice
- Neighborhood has a significant influence on price
- Skewness observed in several features such as SalePrice and LotArea

---

7. Feature Engineering

- Creation of new features (e.g., TotalSF = GrLivArea + TotalBsmtSF)
- Combining similar categorical levels
- Log-transform of skewed target (SalePrice)
- Polynomial features or interaction terms (if applicable)

---

8. Model Building

Algorithms Used:

- Linear Regression
- Ridge & Lasso Regression
- Decision Tree Regressor
- Random Forest Regressor
- XGBoost Regressor

Model Selection Rationale:

- Linear models for baseline
- Regularization to handle multicollinearity
- Tree-based models for non-linear patterns
- XGBoost for performance optimization

Train-Test Split:

- Typical 80/20 or 70/30 split
- K-Fold Cross-Validation used for robust evaluation

Evaluation Metrics:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R-squared (R^2)

---

9. Visualization of Results & Model Insights

Feature Importance:

- Visualized using bar plots (Random Forest/XGBoost feature importances)

Model Comparison:

Model	MAE	RMSE	R^2
Linear Regression	...	...	...
Ridge	...	...	...
Random Forest	...	...	...
XGBoost	...	...	...

Residual Plots:

- Distribution of residuals
- Residuals vs predicted value plots

User Testing:

- Simulated user input interface (optional)
- Predicted vs Actual comparisons for test samples

---

10. Tools and Technologies Used

Programming Language: Python

Notebook Environment: Jupyter Notebook / Google Colab

Key Libraries:

- pandas, numpy (data manipulation)
- matplotlib, seaborn (visualization)
- scikit-learn (modeling)
- xgboost, lightgbm (advanced models)
- statsmodels (regression diagnostics)

---

11. Team Members and Contributions

Name	Contribution
Karthik	Data Collection, Cleaning, Preprocessing
Eswar	EDA, Visualization, Insights
Boopalan	Feature Engineering, Model Selection
Harshath	Model Training, Tuning, Evaluation
Jayakaran	Report Compilation, Testing, Final Presentation