Phase-2

Student Name: ANUSIYA.S

Register Number: 620123106005

Institution: AVS Engineering College

Department: ECE

Date of Submission: 10/05/2025

Github Repository Link:
https://github.com/Anusiya1903/Anusiya-s-naan-mudhalvan-project-.git

1. Problem Statement

- *Accurately forecasting house prices is crucial for buyers, sellers, and real estate investors to make informed financial decisions. The challenge lies in capturing the complex, non-linear relationships among numerous variables like location, size, amenities, and economic conditions.*
- *Type of Problem: Regression (predicting a continuous variable house price).*
- *Why It Matters: Enhances decision-making in real estate markets, supports financial institutions in loan processing, and aids urban planning initiatives.*

2. Project Objectives

- *Primary Goal: Develop a robust, interpretable, and accurate regression model for house price prediction.*
- *Technical Objectives:*
- *Analyze the dataset to identify significant predictors.*

- *Compare multiple regression techniques.*
- *Optimize performance using feature engineering and hyperparameter tuning.*
- *Updated Goal: After initial EDA, emphasis shifted to improving model interpretability while retaining accuracy due to multicollinearity in features.*
- *Assess model fairness and bias, ensuring that the model does not systematically under predict or over predict based on location or house type.*
- *Updated Focus: After initial EDA, emphasis shifted toward improving model interpretability while maintaining accuracy, due to multicollinearity observed among features.*

3. Flowchart of the Project Workflow

1. Data Collection
↓
2. Data Preprocessing
↓
3. Exploratory Data Analysis (EDA)
↓
4. Feature Engineering
↓
5. Model Building
↓
6. Evaluation
↓
7. Visualization
↓
8. Conclusion

4. Data Description

- *Source:Kaggle – House Prices: Advanced Regression Techniques*
- *Type: Structured data (tabular)*
- *Records & Features: ~1460 rows, 80+ features*

- *Dataset Nature: Static*
- *Target Variable: Sale Price*

5. Data Preprocessing

- *Handled missing values using mean/median or domain-specific logic.*
- *Removed duplicate records and verified unique identifiers.*
- *Detected outliers using IQR and visual methods (boxplots).*
- *Converted categorical columns to numerical using one-hot encoding.*
- *Standardized numeric features using StandardScaler.*
- *Ensured data types were consistent across columns.*

6. Exploratory Data Analysis (EDA)

- *Univariate Analysis:*
  - *Used histograms and boxplots for numeric features.*
    Bivariate/Multivariate Analysis:
    - *Correlation matrix and pair plots for key variables vs. Sale Price.*
- *Insights Summary:*
  - *Strong positive correlation with features like Overall Qual, GrLivArea.*

- *Location (Neighborhood) plays a major role.*
- *Some features are highly skewed and need transformation.*

7. Feature Engineering

- *Created new features such as "Total Bathrooms", "House Age", and "Is Remodeled".*
- *Applied log transformation on skewed features.*
- *Binned continuous variables (e.g., Year Built into decades).*
- *Removed features with high collinear or low variance.*
- *Created interaction terms (e.g., OverallQual * GrLivArea) to capture combined effects.*
- *Introduced polynomial features for important variables like GrLivArea to model non-linear patterns.*

8. Model Building

- *Choice of Models: Selected Linear Regression for baseline interpretability and Random Forest Regressor for handling non-linear relationships and feature interactions.*
- *Data Split: Divided the dataset into 80% training and 20% testing sets to evaluate model generalization performance. Stratification was not required for continuous target variables.*

- *Evaluation Metrics: Used MAE, RMSE, and R^2 Score to objectively compare models' accuracy and reliability for regression tasks.*
- *Performance Observation: Random Forest outperformed Linear Regression by achieving lower error values and a higher R^2 score, showing better ability to model complex patterns in the data.*
    - *E.g., Logistic Regression, Decision Tree, Random Forest, KNN, etc.*
- *Applied cross-validation (k-fold) to validate the robustness of model performance.*
- *Performed Grid Search for Hyperparameter Tuning (e.g., tuning number of trees, max depth in Random Forest).*

9. Visualization of Results & Model Insights

- *Feature Importance Plot: Identified top predictors like OverallQual, GrLivArea, and GarageCars.*
- *Residual Plots: Random Forest had more uniformly distributed residuals.*
- *Model Comparison: Visualized performance using bar plots for RMSE and R^2.*
- *Actual vs Predicted Plot:*

  °*Scatter plot comparing predicted SalePrice vs. actual SalePrice, highlighting model accuracy visually.*
- *Distribution of Prediction Errors:*

  °*Plotted histogram of residuals to check for any bias or skewness in predictions*

10. Tools and Technologies Used

- *Language: Python*
- *IDE: Google Colab*
- *Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, XGBoost* ●

  Visualization Tools: matplotlib, seaborn, Plotly

11. Team Members and Contributions

1) **J.Ayisha banu:** Data cleaning and documentation.
2) **S.Anusiya**: EDA and problem objective.
3) **M.Dharani**: Feature engineering and reporting.
4) **M.Kaviya:** Model development and visualization of results.