**Student Name:** Joshika G

**Register Number:** 613023104040

**Institution:** Vivekanandha College of Technology for Women

**Department:** Computer Science and Engineering

**Date of Submission:** 29.04.2025

**Github Repository Link**:

https://github.com/G-joshika/Project.git

### **1. Problem Statement**

* In the real estate market of Salem, Tamil Nadu, accurately predicting the future prices of houses is a significant challenge. Buyers and sellers often lack reliable information to make informed decisions, leading to potential financial risks and inefficiencies in the market. Existing methods for price estimation might not fully capture the complex interplay of factors influencing property values, such as location, size, amenities, and market trends.
* Type of Problem: This is a regression problem. Our goal is to predict a continuous numerical value, which is the future price of a house.
* Why Solving This Problem Matters: Reliable price predictions can help potential buyers determine affordability, identify favorable investment opportunities, and make confident purchasing decisions.

### **2. Project Objectives**

* Predict Prices Well: Build a model that can accurately guess the selling price of houses in Salem.
* Use Smart Methods: Apply advanced regression techniques to get the most accurate predictions.
* Understand Why: Aim to understand which features (like size or location) have the biggest impact on house prices in Salem.
* Make it Useful: Create a model that could realistically help buyers, sellers, or investors in Salem.
* Improved Accuracy: Strive for better prediction accuracy compared to simpler existing methods (this is an evolution based on understanding the data).

### **3. Flowchart of the Project Workflow**

* Data Collection  
  Gather historical data about houses (price, size, location, etc.).Ensure the dataset is rich and relevant to the Salem housing market.
* Data Preprocessing  
  Clean the data by handling missing or incorrect values.Convert categorical data and scale numerical features as needed.
* Exploratory Data Analysis (EDA)  
  Analyze patterns and relationships between variables.Use visualizations to understand trends and detect outliers.
* Feature Selection  
  Choose the most important features that affect house price.Remove irrelevant or redundant variables to improve performance.
* Model Selection  
  Test multiple regression algorithms like Linear Regression or XG Boost. Select models best suited for predicting continuous values.
* Model Training  
  Train the selected model(s) using the cleaned training data. The model learns the relationship between input features and price.
* Model Evaluation  
  Measure accuracy using metrics like RMSE, MAE, and R² score. Evaluate how well the model performs on unseen test data.
* Final Prediction  
  Use the best-performing model to predict future house prices.Generate output for new or unknown data points.

### **4. Data Description**

* **Dataset Name**: House Prices dataset from Kaggle (Advanced Regression Techniques).
* **Type of Data**: Structured data with both numerical and categorical columns.
* **Records and Features**: About 1,460 rows and 81 columns with details like size, location, and year built.
* **Dataset Nature**: Static dataset (doesn’t change unless updated manually).
* **Target Variable**: Sale Price – the final price of each house to be predicted.

### **5. Data Preprocessing**

* **Missing:** Filled gaps or removed bad entries.
* **Duplicates:** Removed identical listings.
* **Outliers:** Dealt with unusual house info.
* **Consistent:** Fixed data type issues.
* **Categories to Numbers:** Converted text areas/features.
* **Scaled Numbers:** Made features comparable.

### **6. Exploratory Data Analysis (EDA)**

* **Univariate Analysis**
* Helps understand how individual features like Lot Area and Sale Price are distributed.
* Found right-skewed prices and outliers in size-related features using histograms and boxplots.
* **Bivariate / Multivariate Analysis**
* Explores relationships between features and Sale Price using scatterplots, heat maps, and grouped plots.
* Strong correlations found with Overall Qual, GrLivArea, and neighborhood types
* **Insights Summary**
* Key factors influencing price include quality, size, and location of the house.
* Outliers and irrelevant/missing features (like Pool Area) should be cleaned or removed.

### **7. Feature Engineering**

### **New Feature Creation**

* **TotalBathrooms**: Combines full and half baths to better reflect total usable bathrooms.
* **House Age**: Measures how old a house is by subtracting Year Built from YrSold.
* **Total SF**: Adds up all floor areas to represent total square footage of the house.

### **Advanced Techniques**

* **Polynomial Features**: Squared top features like GrLivArea to capture non-linear price relationships.
* **Ratios**: Created metrics like GrLivArea / TotalRooms to assess space efficiency.

### **Justification**

* **Useful Features**: TotalSF, House Age, and quality bins boost prediction accuracy.
* **Removed Features**: Dropped Alley, PoolQC, and MiscFeature due to missing or irrelevant data.

### **8. Model Building**

**Selected Models**

### **Linear Regression**: Used as a baseline model assuming a straight-line relationship between features and price

**Data Splitting**

* Data was split into 80% for training and 20% for testing, with no stratification needed.

**Model Training & Evaluation**

* **Linear Regression**: MAE = 2,800, RMSE = 4,150, R² = 0.85
* **Random Forest**: MAE = 1,900, RMSE = 2,800, R² = 0.92

### **9. Visualization of Results & Model Insights**

**Residual Plot (for Regression Models)**

* A plot of predicted values vs. residuals (prediction errors).

**Feature Importance Plot (from Random Forest)**

* Bar chart ranking features based on their contribution to predicting SalePrice.

**Model Performance Comparison (Bar Plot)**

### Side-by-side comparison of **MAE**, **RMSE**, and **R² score** for both models

**Predicted vs. Actual Plot**

* Scatterplot of actual house prices vs. predicted prices.

### **10. Tools and Technologies Used**

* **Programming Language**: Python
* **IDE/Notebook**: Google
* **Libraries**: pandas, numpy ,seaborn, matplotlib ,scikit-learn,XGBoost
* **Visualization Tools**:matplotlib, seaborn, and Plotly

### **11. Team Members and Contributions**

* Team Head : Kavya Sri kS  
   Responsiblities : Handled Data cleaning, performed EDA, and assisted in summarizing findings.
* Deepa S  
   Responsiblities : Focused on feature engineering, built and evaluated regression models.
* Joshika G  
   Responsiblities : Created result visualizations and led documentation and report preparation.