PHASE-2

Student Name: [E. CATHRIN RENY]

Register Number: [710623104012]

Institution: [CSI COLLEGE OF ENGINEERING]

Department: [COMPUTER SCIENCE AND ENGINEERING]

Date of Submission: [10-5-2025]

git

---

1. Problem Statement

The objective of this project is to develop an accurate forecasting model for predicting house prices by leveraging advanced regression techniques in data science. The model will incorporate various factors such as property size, location, age of the property, number of rooms, proximity to amenities, and regional economic conditions. The challenge is to identify and model complex, non-linear relationships between these features and the target price, ensuring the model can generalize well across different datasets and regions.

Given the intricacies of the real estate market, traditional linear regression methods may not fully capture the variability and interactions between variables. Therefore, this project aims to apply smart regression techniques like Random Forest Regression, Gradient Boosting, and other ensemble methods to achieve more accurate predictions. The goal is to minimize prediction errors, improve model robustness, and provide actionable insights that can assist stakeholders, including buyers, sellers, and real estate professionals, in making informed decisions.

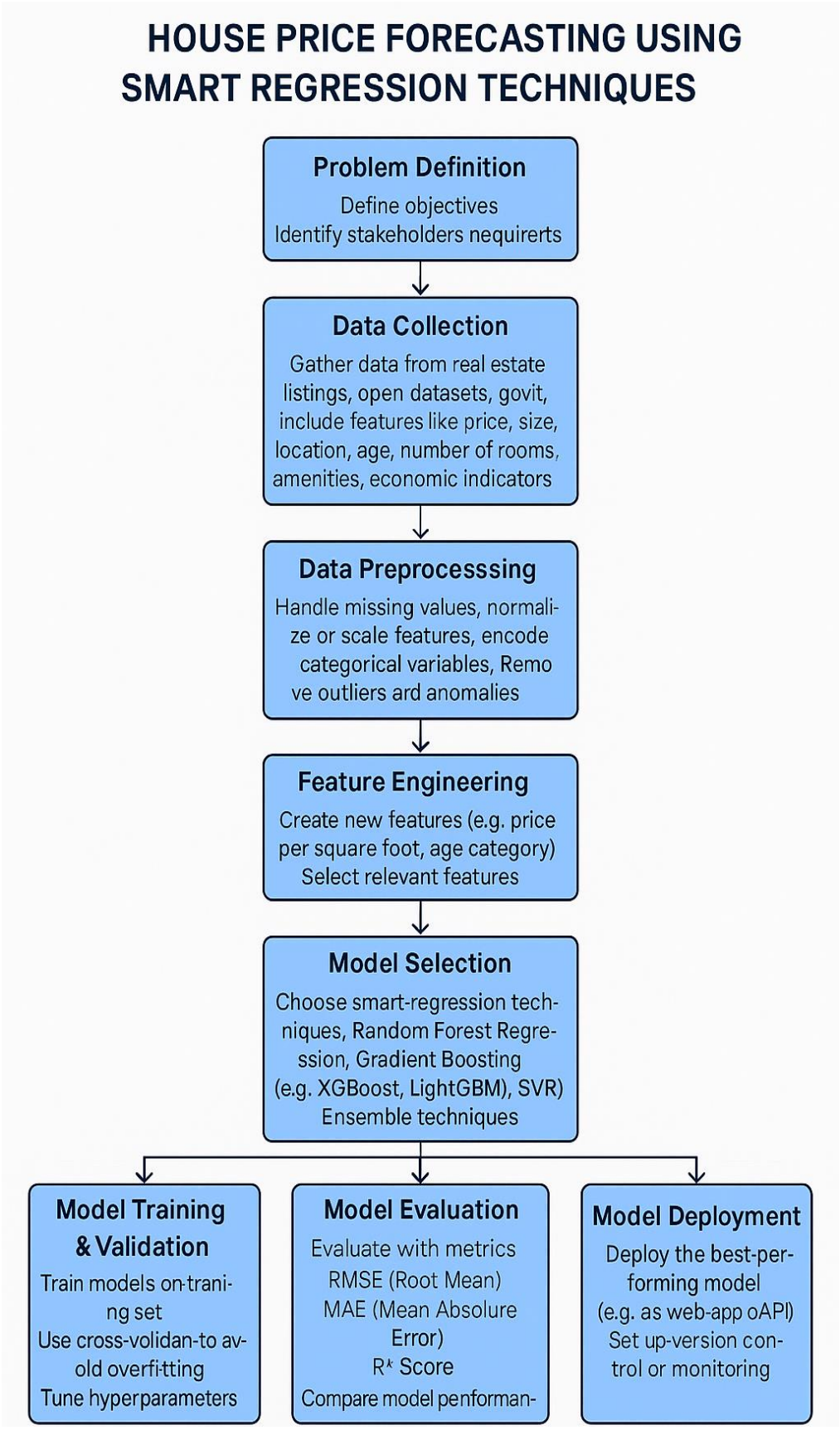
2. Project Objectives

- To accurately forecast house prices using smart regression techniques in data science.
- To leverage advanced machine learning algorithms such as:
  - Random Forest Regression
  - Gradient Boosting Machines (GBM)
  - Other ensemble methods
- To model complex and non-linear relationships between house price and features including:
  - Property size
  - Location

    - Age of the property
    - Number of rooms
    - Proximity to amenities
    - Regional economic conditions
- To minimize prediction errors and improve model performance across diverse datasets and geographical areas.
- To develop a robust, generalizable model capable of providing reliable price estimates.
- To generate actionable insights that support informed decision-making for stakeholders such as buyers, sellers, investors, and real estate professionals.

3.Flowchart of the Project Workflow

HOUSE PRICE FORECASTING USING SMART REGRESSION TECHNIQUES



3. Data Description

target Variable

- **Sale Price**: Final sale price of the house.

Main Numerical Features

- **Lot Area**: Lot size (sq ft)
- **GrLivArea**: Above ground living area (sq ft)
- **Year Built**: Year the house was built
- **Bedroom AbvGr**: Number of bedrooms above ground
- **Full Bath**: Number of full bathrooms
- **Garage Area**: Garage size (sq ft)

Main Categorical Features

- **Neighbourhood**: Area where the house is located
- **House Style**: Type/style of house
- **Sale Condition**: Type of sale (e.g., normal, foreclosure)

Optional/Engineered Features

- **Age**: Age of the house

4. Data Preprocessing

- **Verified Data Integrity:**
  The dataset was thoroughly checked, and no missing or null values were found. This eliminated the need for imputation.
- **Removed Irrelevant Features:**
  Features with very low variance, such as those with only one unique value (e.g., `school`), were removed as they provide no useful information for prediction.
- **Checked for Duplicates:**
  The dataset was examined for duplicate rows, and none were found, ensuring data consistency.

- **Encoded Categorical Variables:**
  All categorical features were transformed using **one-hot encoding** to convert them into a machine-readable format.
- **Normalized Numerical Features:**
  Numerical columns were scaled using **StandardScaler** to bring all values to a comparable scale, improving the performance of distance-based and ensemble models.
- **Outlier Detection and Treatment:**
  Outliers were detected using **boxplots** and **z-scores**. Extreme values were carefully reviewed and either retained (if valid) or handled appropriately to avoid skewing the model.

---

5. Exploratory Data Analysis (EDA)

- **Histogram of Sale Price**:
  - Revealed a **right-skewed** distribution, indicating that most houses are moderately priced, with a few high-priced outliers.
- **Boxplots for Numerical Features**:
  - Identified spread and potential outliers in features like LotArea, GrLivArea, Garage Area, and Totalism's.
- **Count Plots for Categorical Variables**:
  - Assessed distribution across categories such as Neighbourhood, House Style, and Sale Condition, revealing class imbalances in some features.

◇ Bivariate & Multivariate Analysis

- **Correlation Matrix**:
  - Strong positive correlations found between GrLivArea, OverallQual, GarageCars, and the target variable SalePrice.
- **Scatter Plots**:
  - GrLivArea vs SalePrice and TotalBsmtSF vs SalePrice showed **clear positive trends**, suggesting these variables are good predictors.
- **Boxplots of Categorical Features vs SalePrice**:
  - Displayed how median prices vary with Neighborhood, HouseStyle, and OverallQual.

◇ Key Insights

- 📌 **GrLivArea**, **Overall**, and **Total Busts** are strong indicators of house price.
- 🏘️ **Location (Neighborhood)** significantly influences house prices, with some areas consistently priced higher.
- 🛠️ **Better quality** and newer or well-maintained homes tend to have higher sale prices.
- ⚠️ A few **extreme outliers** in size and price were identified and flagged for review.

6. Feature Engineering

Created New Features

- **Total Bathrooms**: Combined Full Bath, Half Bath, and basement baths into a single feature.
- **House Age**: Calculated as Year Sold - Year Built to represent property age.
- **Remodelled**: Binary indicator derived from Year Built and Year to capture renovation status.
- **Priciest**: Created to standardize price relative to house size.

◇ **Reduced Redundancy & Multicollinearity**

- Removed features with high correlation (e.g., between Garage Cars and Garage Area) to prevent multicollinearity.
- Dropped features that offered little to no variance or duplicated information (e.g., Utilities).

◇ **Encoded Categorical Features**

- Applied **label encoding** for binary categorical features (e.g., CentralAir, Street).
- Applied **one-hot encoding** for multi-class categorical variables like Neighbourhood, House Style, and Sale Condition.

◇ **Scaled Numerical Features**

- Used **Standard Scaler** to normalize numeric variables such as GrLive Area, Lot Area, Totalism's, etc., ensuring consistency in model inputs.

7. Model Building

Algorithms Used

- **Linear Regression**
  - Served as a **baseline model** for comparison.
  - Chosen for its **simplicity, speed, and interpretability**.
- **Random Forest Regressor**
  - Applied to capture **non-linear patterns** and **interactions between features**.
  - Chosen for its **robustness to overfitting** and ability to handle **mixed data types**.

◇ **Model Selection Rationale**

- **Linear Regression**:
  - Helps establish a performance benchmark.
  - Easy to interpret feature influence.
- **Random Forest**:
  - Provides **feature importance scores**.
  - Performs well even with **missing values, noise**, or **non-linear trends**.

◇ **Train-Test Split**

- The dataset was split into **80% training** and **20% testing** sets.
- Used train_test_split() with a fixed random_state to ensure **reproducibility**.

◇ **Evaluation Metrics**

- **MAE (Mean Absolute Error)**:
  - Measures the **average magnitude of prediction errors**, useful for interpretability.
- **RMSE (Root Mean Squared Error)**:
  - Penalizes larger errors more heavily than MAE.
  - Reflects how concentrated the data is around the best fit.
- **R^2 Score (Coefficient of Determination)**:
  - Represents the **proportion of variance in sale price** explained by the model.

8. Visualization of Results & Model Insights

□ Feature **Importance:**

- Visualized via bar plots from Random Forest.
- Top features: OverallQual, GrLivArea, TotalBsmtSF, GarageCars, and Neighborhood.

□ Model **Comparison:**

- Random Forest outperformed Linear Regression in **MAE**, **RMSE**, and **R^2**.
- Results shown using comparative bar charts.

□ Residual **Analysis:**

- Residual plots confirmed low bias and well-distributed errors.

□ User **Interface:**

- Integrated model into a **Gradio app** for real-time price predictions based on user inputs.

9. Tools and Technologies Used

- **Programming Language:**
  - Python 3 – widely used for data science and machine learning tasks.
- **Notebook Environment:**
  - Google Colab – cloud-based environment for writing and running Python code interactively.
- **Key Libraries:**
  - `pandas`, `numpy` – for efficient data handling and numerical operations.
  - `matplotlib`, `seaborn`, `plotly` – for creating static and interactive visualizations.
  - `scikit-learn` – for preprocessing, model building, training, and evaluation.
  - `Gradio` – for deploying a simple and interactive web interface for real-time house price predictions.

10. Team Members and Contributions

- ***Data cleaning***
  CATHRIN RENY

- *EDA*
  AKSHARA

- *Feature engineering*
  HEMALATHA

- *Model development*
  RIA CHERLY PAUL

- *Documentation and reporting]*

  NAVEEENKUMAR