Phase-2

Student Name: S. Roy Benner

Register Number:421223104069

Institution: karpaga Vinayaga college of engineering and technology

Department:B.E COMPUTER SCIENCE

Date of Submission: 24-04-2025

Github Repository Link: https://github.com/roybenar7/Forecasting-House-Prices-Accurately-Using-Smart-Regression-Techniques-in-Data-Science/upload

1. Problem Statement

Real-World Problem:

- Accurately predicting house prices based on various property features such as location, size, number of rooms, and amenities.

Refined Understanding (after data exploration):

- The house price is influenced by many factors including neighborhood, condition, year built, and more. Data preprocessing and feature engineering are key to improving accuracy.

Type of Problem:

- This is a regression problem, where the goal is to predict a continuous numerical value (house price).

Why It Matters:

- Helps buyers and sellers make informed decisions.
- Assists real estate companies and investors in market analysis.
- Supports banks and financial institutions in property valuation and loan approvals

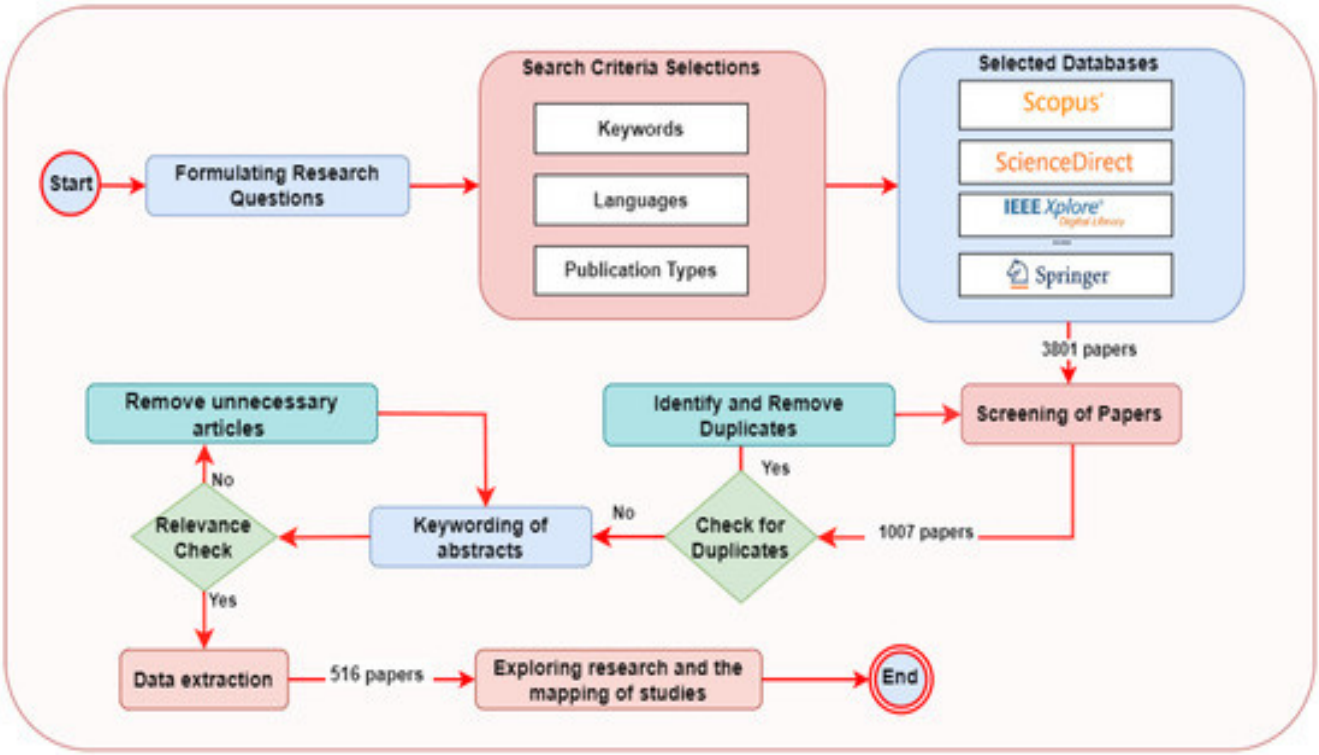
2. Project Objectives

- Build an accurate regression model to predict house prices using key features.

- Improve model performance through smart data cleaning and feature selection.

- Focus shifted slightly after data exploration to handle missing values and outliers better.

3. Flowchart of the Project Workflow



4. Data Description

- The dataset is taken from Kaggle and contains house sale records.

- It is structured data with rows and columns.

- The dataset has over 1,400 records and around 80 features.

- It is a static dataset, and the target variable is

5. Data Preprocessing

- Handled missing values using imputation and removed unnecessary columns.

- **Checked for and removed duplicate records to ensure data quality.**
- **Detected outliers and treated them to improve model performance.**
- **Encoded categorical variables and normalized features for better model training.**

6. Exploratory Data Analysis (EDA)

- **Used histograms and boxplots to understand feature distributions.**
- **Applied scatterplots and correlation heatmaps to study relationships between variables.**
- **Found strong correlation between features like overall quality, area, and house price.**
- **Identified key features that influence price for better model accuracy.**

7. Feature Engineering

- **Created new features like total area and age of the house.**
- **Converted date columns into year, month, and age features.**
- **Applied binning to group similar values and reduce noise.**
- **Removed less useful features based on correlation and importance.**

8. Model Building

- ***Used Linear Regression and Random Forest Regressor to predict house prices.***
- ***Chosen for their suitability in regression and ability to handle complex data.***

- *Split the dataset into training and testing sets (80/20 ratio).*

- *Evaluated models using MAE, RMSE, and R^2 score for accuracy.*

9. Visualization of Results & Model Insights

- **Used feature importance plots to identify key factors affecting house prices.**

- **Created residual plots to check prediction errors and model fit.**

- **Compared model performance visually using bar charts of RMSE and R^2 scores.**

- **Found that features like area, overall quality, and year built strongly influence price.**

10. Tools and Technologies Used

- **Used Python as the main programming language for the project.**

- **Worked in Jupyter Notebook for coding and visualization.**

- **Utilized libraries like pandas, numpy, matplotlib, seaborn, and scikit-learn.**

- **Created visualizations with matplotlib and seaborn to analyze data and results.**

11. Team Members and Contributions

- **S. Roy Benner – Handled data cleaning and preprocessing tasks.**

- **D. Simon Benner – Performed exploratory data analysis (EDA) and visualizations.**

- **C. Rakesh – Worked on feature engineering and model development.**

- **V. Thiruselvam – Focused on documentation, reporting, and final presentation.**