PHASE-2 SUBMISSION

Student Name: Shamila. D

Register Number: 410723104080

Institution: Dhanalakshmi college of Engineering

Department: Computer Science and Engineering

Date of Submission: 07.05.2025

Github Repository Link:
https://github.com/Shamiladas/nm_shamila.git

Forecasting house prices accurately using smart regression techniques in data science

1. Problem Statement

- *Forecasting house prices accurately is a difficult task in the real estate industry, affecting buyers, sellers, investors, and policymakers. With the rise of data science, this task has evolved from basic estimations to worldly predictive modeling using smart regression techniques. These methods enable the analysis of large, complex datasets containing features like location, square footage, number of rooms, age of the property, and market trends and school nearest.*

- *Smart regression techniques include traditional models such as Linear, Ridge, and Lasso Regression, as well as advanced machine learning algorithms like Random Forest, Gradient Boosting (e.g., XG boost), and*

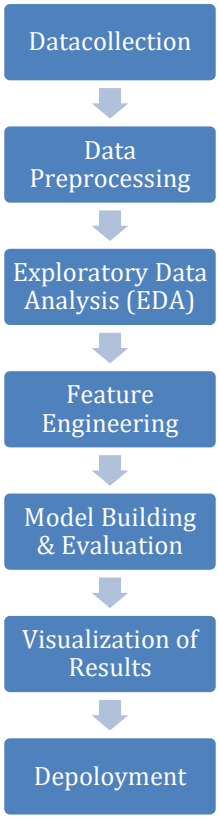
Neural Networks. These models are capable of capturing nonlinear relationships and interactions between variables that influence house prices.

- *The forecasting process involves crucial steps such as data cleaning, feature engineering, model selection, and evaluation using metrics like RMSE and R^2. Proper preprocessing and model tuning are essential to ensure accuracy and avoid overfitting.*
- *By applying these intelligent regression methods, data scientists can deliver highly accurate price predictions, supporting more informed and data-driven decision-making in real estate.*

2. Project Objectives

- *The primary objective of this project is to develop a robust, data-driven framework for accurately forecasting house prices using advanced regression techniques in data science.*
- *The predictive power of modern machine learning algorithms to model the complex, nonlinear relationships that influence residential real estate prices. Which are based on housing features, location-based variables, economic indicators, and historical event, the goal is to build regression models that can provide precise price estimates.*
- *Regression techniques including, but not limited to, Linear Regression, Ridge and Lasso Regression, Decision Tree Regression, Random Forests, Gradient Boosting Machines and Artificial Neural Networks. Each model will be evaluated using appropriate performance metrics such as RMSE, MAE, and R^2 to assess accuracy.*
- *Special attention will be given to mitigating issues such as overfitting and data imbalance. The ultimate objective is not only to achieve high predictive accuracy but also to provide actionable insights into the key factors driving house prices, thereby aiding stakeholders like buyers, sellers, real estate investors, and policy makers in making informed decisions.*

3. Flowchart of the Project Workflow



4. Data Description

- ***Dataset name:*** *Housing price dataset*
- ***Source****: Kaggle*
- ***Type of data****: structured, tabular data*
- ***Number of records and features****: 546 Price and 12 Features*

- ***Static or dynamic****: Dynamic Dataset*
- ***Target variable****: House price, Sale price*
- ***Source****: https://www.kaggle.com/datasets/yasserh/housing-pricesdataset*

5. Data Preprocessing

- ***Missing values:*** *no missing values were found in dataset*
- ***Duplicate Records:*** *Duplicate rows were checked and removed if present.*
- ***Outliers:*** *Detected using boxplots; outliers in amount were handled using transformation.*
- ***Data Types:*** *All features are numeric, No conversion needed.*
- ***Encoding Categorical Variables:*** *Not required as all features are already numerical.*
- ***Normalization:*** *Amount and Time were scaled using standard scaler to bring them on the same as V1-V2*

6. Exploratory Data Analysis (EDA)

- ***Univariate Analysis:***
    - *A histogram is useful for continuous numerical features like price, square, footage etc.*
    - *A boxplot for house price will show the median, quartiles, and outliers.*
    - *Count plot is useful for understanding the distribution of categorical variables.*
- ***Bivariate/Multivariate Analysis:***

- *Correlation matrix identifies which variable have strong relationships with the house price.*
- *Can plot the relationship between house price and feature like square feet or number of bedrooms.*
- *Grouped bar plots shows the comparison of neighbourhood, house style and conditions.*

- ***Insights Summary**:*

  - *Square feet, bathrooms and grade show the strong positive correlation.*
  - *Neighborhood greatly influences average price and location matter.*
  - *Scatter plots confirm a clear trend between size and price.*

7. Feature Engineering

- *New features like total = bedrooms + bathrooms +other rooms, house age = year sold – year built.*

- *Split data column into year, month, day and combine latitude and longitude into a "location cluster" using K means.*

- *Bin house age into categories: "new", "mid-age", "old" and polynomial features like $(area)^2$ or area * number of rooms.*

- *Use PCA on geographical or neighbourhood features if they are numerous and correlated.*

- *Based on correlation analysis domain relevance, and model performance (e.g., cross-validation scores)*

8. Model Building

- ***Machine Learning Models:***
    - *Model 1: Linear Regression (Baseline)*
    - *Model 2: Random forest Regressor (Advance ensemble model)*

- ***Model Selection:***
    - *Linear Regression*
    - *Ridge/Lasso Regression*
    - *Decision Tree Regression*
    - *Random Forest Regression*

- ***Model Evaluation Metrics:***
    - *RMSE (Root Mean Square Error)*
    - *MAE (Mean Absolute Error)*
    - *R^2 (Coefficient of determination)*

9. Visualization of Results & Model Insights

- ***Feature importance plots**:*
    - *Visualized using bar plots from Random Forest or XG Boost*
    - *Residual plots (actual vs. predicted prices)*
    - *Prediction error plots*
    - *Distribution of predicted vs. actual prices*

- ***Model Comparison**:*
    - *Compare models using bar charts of metrics like RMSE, MAE, and R^2.*
    - *Line or scatter plots showing model predictions vs. actual prices.*
- ***Residual Plots**:*

- *Use SHAP (SHapley Additive exPlanations) or feature importance plots to show which variables (e.g., number of rooms, square footage, location) have the most influence.*

10. *Tools and Technologies Used*

- ***Programming Language*** *–Python is a main programming language*
- ***Notebook/IDE*** *–The platform we used to work in are Google Collab, Jupyter Notebook, VS Code*
- ***Libraries*** *–The key libraries are used for data processing, visualization, and modeling are pandas, numpy, seaborn, matplotlib, scikit-learn, TensorFlow*
- ***Optional Tools for Deployment*** *–Some tools or frameworks that might use for deployment are Streamlit, Flask and fast API*
- ***Version Control*** *- Git/GitHub*

11. Team Members and Contributions

Name	***Roles***	***Responsibility***
Shamila .D	*Team leader*	*Data cleaning and EDA*
Saranya .S	*Team member*	*Feature Engineering*
Sharumathi .M	*Team member*	*Model development*
Reena .R	*Team member*	*Documentation and reporting*