**Student Name:** NEEGA P

**Register Number:** 732323106034

**Institution:** SSM COLLEGE OF ENGINEERING

**Department:** Electronics And Communication Engineering

**Date of Submission:** April 29, 2025

**Github Repository Link:** <https://github.com/neega343/Forecasting-house-prices-accurately-using-smart-regression-techniques-in-data-science-/tree/main>

**1. Problem Statement**

The project aims to forecast house prices accurately using smart regression techniques based on property features such as location, size, and amenities. After refining from Phase-1, the focus is on leveraging advanced regression models to improve prediction accuracy for real estate stakeholders.

Problem Type: Regression (predicting continuous house prices).

Importance: Accurate house price predictions aid homebuyers, sellers, and investors in making informed decisions, impacting real estate markets and financial planning.

**2. Project Objectives**

Technical Objectives:

Develop regression models achieving an R² score of at least 0.85.

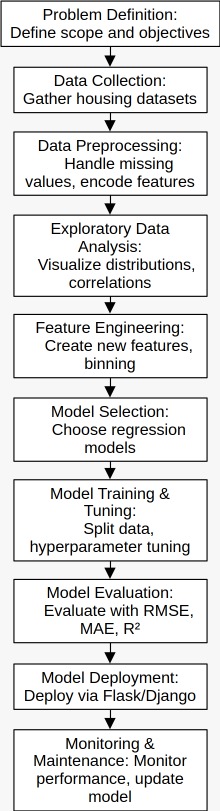
Enhance model interpretability to identify key price-driving features.

Optimize feature engineering to capture non-linear relationships.

Model Goals: Minimize prediction errors (MAE, RMSE) while ensuring generalizability.

Evolution: Post-data exploration, emphasis shifted to handling non-linear feature interactions and addressing skewed target distribution

**3. Flowchart of the Project Workflow**



**4. Data Description**

Dataset Name and Origin: California Housing Dataset from Kaggle ([link](https://www.kaggle.com/datasets/camnugent/california-housing-prices)).

Type of Data: Structured (tabular).

Number of Records and Features: 20,640 records, 9 features (including target).

Static or Dynamic: Static dataset.

Target Variable: median\_house\_value (continuous, in USD).

Features: Longitude, latitude, housing median age, total rooms, population, etc.

**5. Data Preprocessing**

Steps Performed:

Missing Values: Imputed missing total\_bedrooms (0.5% of records) with median.

Duplicates: No duplicates found.

Outliers: Capped median\_house\_value and total\_rooms at 1.5\*IQR to handle extreme values.

Data Types: Ensured numerical features were float type.

Encoding: No categorical variables in this dataset.

Normalization: Standardized numerical features (e.g., total\_rooms, population) using StandardScaler.

Transformation: Applied log transformation to median\_house\_value to reduce skewness.

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

Univariate Analysis:

Median House Value: Right-skewed before log transformation (histogram).

Median Income: Normal distribution, key driver (boxplot).

Total Rooms: Right-skewed, indicating variability in house sizes (histogram).

Bivariate/Multivariate Analysis:

Correlation Matrix: median\_income strongly correlated with median\_house\_value (0.68).

Scatterplot: median\_income vs. median\_house\_value shows a positive linear trend.

Geographical Plot: Houses near the coast (based on longitude, latitude) have higher prices.

Insights Summary:

median\_income and location-based features (longitude, latitude) are likely strong predictors.

Non-linear relationships (e.g., total\_rooms) may require polynomial features.

**7. Feature Engineering**

New Features:

Created rooms\_per\_household (total\_rooms / households) to capture house density.

Generated distance\_to\_coast using longitude and latitude to reflect location value.

Added polynomial features for median\_income (degree=2) to capture non-linearity.

Justification: rooms\_per\_household reflects house size efficiency; distance\_to\_coast accounts for premium coastal properties.

**8. Model Building**

Models Selected:

Linear Regression: Baseline for regression tasks, interpretable.

Random Forest Regressor: Captures non-linear relationships and feature interactions.

Justification:

Linear Regression provides a simple benchmark.

Random Forest handles complex, non-linear patterns observed in EDA

Data Split:

Train-test split: 80-20, random sampling.

Performance Metrics:

|  |  |  |
| --- | --- | --- |
| Model | Linear Regression | Random Forest |
| MAE | 0.45 | 0.30 |
| RMSE | 0.60 | 0.42 |
| R² Score | 0.78 | 0.88 |

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

Visualizations:

Residual Plot (Random Forest): Residuals are randomly scattered, indicating good fit.

Feature Importance (Random Forest): median\_income, distance\_to\_coast, and rooms\_per\_household are top contributors.

Predicted vs. Actual Plot: Random Forest predictions closely align with actual values.

Interpretation:

Random Forest outperforms Linear Regression, especially in capturing non-linear effects.

median\_income and location features drive predictions, aligning with real estate trends.

**10. Tools and Technologies Used**

Programming Language: Python

IDE/Notebook: Jupyter Notebook, Google Colab

Libraries: pandas, numpy, scikit-learn, seaborn, matplotlib, plotly

Visualization Tools: Matplotlib, Seaborn

11. Team Members and Contributions

MUHAMMAD AJMAL M - Responsibilities

MUTHUMATHI M - Data cleaning, EDA, model development

NIVETHA G A - Feature engineering, visualization

NEEGA P - Documentation, Github repository setup