# Phase-1 PROJECT OVERVIEW

**Student Name:** TAMILSELVAN.J

**Register Number:**511523205058

**Institution:**P.T.Lee Chengalvaraya Naicker College Of Engineering & Technology

**Department:**Information Technology

**Date of Submission:** 26-04-2025

**Ebpl-DS-Forecasting house prices accurately using smart regression techniques in data science**

## 1. Problem Statement

Accurately forecasting house prices is a vital task in real estate, finance, and urban development. Traditional valuation methods often fall short in capturing the complex dynamics of the housing market. This project seeks to address these challenges by applying advanced regression techniques from data science to deliver high-accuracy predictions. These insights can benefit stakeholders such as investors, developers, and homebuyers.

## 2. Objectives of the Project

- Analyze historical housing data to uncover key patterns.

- Preprocess and engineer features that significantly impact pricing.

- Experiment with a variety of regression algorithms to predict house prices.

- Evaluate and compare model performances to identify the most effective approach.

- Highlight the most influential features affecting housing prices.

## 3. Scope of the Project

\*\*Features:\*\*

- Location

- Bedrooms, bathrooms

- Square footage

- Year built

- Additional property features

\*\*Limitations:\*\*

- Static dataset; no real-time updates

- Restricted to structured regression models (no deep learning)

- Deployment optional depending on model complexity

## 4. Data Sources

- \*\*Source:\*\* Kaggle public dataset on housing prices

- \*\*Type:\*\* Public, static dataset

- \*\*Content:\*\* Attributes include location, size, amenities, and sale price

- \*\*Format:\*\* CSV file

## 5. High-Level Methodology

\*\*Data Collection:\*\*

Downloaded from Kaggle (static housing dataset).

\*\*Data Cleaning:\*\*

Handled missing values, corrected inconsistencies, removed duplicates.

\*\*Exploratory Data Analysis (EDA):\*\*

Used histograms, boxplots, correlation heatmaps to explore distributions and relationships.

\*\*Feature Engineering:\*\*

Created new features (e.g., age of the house), normalized data, encoded categorical variables.

\*\*Model Building:\*\*

Tested multiple regression models:

- Linear Regression

- Ridge & Lasso

- Decision Tree

- Random Forest

- Gradient Boosting (XGBoost, LightGBM)

\*\*Model Evaluation:\*\*

Used MAE, RMSE, and R² to assess model performance.

\*\*Visualization & Interpretation:\*\*

Presented insights using bar plots, scatter plots, and feature importance graphs.

\*\*Deployment (Optional):\*\*

Considered a Streamlit web app for interactive predictions (pending further development).

## 6. Tools and Technologies

- \*\*Programming Language:\*\* Python

- \*\*Notebook/IDE:\*\* Jupyter Notebook

- \*\*Libraries:\*\* pandas, numpy, matplotlib, seaborn, scikit-learn, XGBoost, LightGBM

- \*\*Optional Deployment Tools:\*\* Streamlit, Flask

## 7. Team Members and Roles

-**RAGULGANDHI.K**– Data Collection & Cleaning

-**TAMILSELVAN.J**– EDA & Feature Engineering

- **SAKTHIVEL.R**– Model Building & Evaluation

-**JAGAN.S**– Visualization & Documentation