Phase-1 Submission Template

Student Name: B. Mathanraj

Register Number: 620123106064

Institution: AVS engineering college

Department: Electrics and Communication Engineering

Date of Submission: 30/04/2025

1. Problem Statement

Forecasting house prices is a critical task in the real estate domain. Inaccurate pricing leads to poor investment decisions, over/under-valuation of properties, and dissatisfaction among stakeholders. Traditional methods lack adaptability to large and diverse datasets. This project aims to solve this problem using data science techniques, especially advanced regression models, to accurately predict house prices based on property features.

2. Objectives of the Project

- Develop predictive models for house price forecasting using smart regression algorithms.
- Identify key features that influence housing prices.
- Compare performance of multiple models to choose the most effective one.
- Deliver a reliable and interpretable model that can assist in real-world applications.

3. Scope of the Project

- Features Analyzed:
  - Property size, location, number of rooms, year built, amenities, condition, etc.
- Limitations/Constraints:
  - Focuses only on structured numerical and categorical data.
  - Relies on publicly available datasets.
  - Image data and real-time web scraping are excluded.
  - Deployment is optional depending on timeline.

4. Data Sources

- Dataset Name: House Prices - Advanced Regression Techniques
- Source: Kaggle (https://www.kaggle.com/c/house-prices-advanced-regression-techniques)
- Type: Public dataset
- Access: Static (downloaded once for analysis)

5. High-Level Methodology

- Data Collection: Download dataset from Kaggle.
- Data Cleaning: Address missing values, drop duplicates, convert data types, encode categorical values.
- Exploratory Data Analysis (EDA): Use visualizations such as histograms, scatter plots, and heatmaps to identify trends and correlations.
- Feature Engineering: Generate new variables (e.g., price per square foot), one-hot encode categories, handle skewed data.

- Model Building: Experiment with multiple regression models—Linear Regression, Lasso, Ridge, Decision Trees, Random Forest, Gradient Boosting (XGBoost, LightGBM).
- Model Evaluation: Use RMSE, MAE, and R^2 Score. Apply k-fold cross-validation to ensure generalization.
- Visualization & Interpretation: Present results using charts, model comparison graphs, and feature importance visualizations.
- Deployment: Optionally deploy using Streamlit to create a user interface for predictions.

6. Tools and Technologies

> Programming Language: Python
> Notebook/IDE: Google Colab / Jupyter Notebook
> Libraries:
  - pandas, numpy for data manipulation
  - matplotlib, seaborn, plotly for visualization
  - scikit-learn, xgboost, lightgbm for modeling
> Optional Deployment Tools: Streamlit, Flask

7. Team Members and Roles

- B. Mathanraj - Primary Analayst & Execution

-S. Rakesh – Data Collection & Cleaning, EDA

-R. Rahul Kumar – Feature Engineering, Model Building & Evaluation

- K. Ramesh – Visualization, Interpretation & Deployment