Phase-2 Submission Template

Student Name: [M.BARATH]

Register Number: [422223106007]

Institution: [SURYA GROUP OF INSTITUTIONS]

Department: [2ND YEAR -ECE]

Date of Submission: [08/05/2025]

Github Repository Link: [https://github.com/barath26-nish/Phase-2]

---

1. Problem Statement

[Real estate pricing is influenced by numerous complex factors, making it challenging to estimate property values accurately. This project focuses on predicting house prices using machine learning techniques that analyze various features like location, size, amenities, and market trends. The goal is to assist buyers, sellers, and agents in making informed decisions through accurate price forecasting, formulated as a regression problem..]

2. Project Objectives

[Predict housing prices using regression-based machine learning models

Identify key features that influence price trends

Improve model performance using smart suggestion techniques (feature selection, regularization, ensemble models)

Enhance prediction accuracy using tuning and validation strategies

Provide interpretable and actionable insights to stakeholders]

3. Flowchart of the Project Workflow

[1. Data Collection

2. Data Preprocessing

3. Exploratory Data Analysis (EDA)

4. Feature Engineering

5. Model Building (Linear, Tree-Based, Ensemble)

6. Evaluation & Smart Suggestion Techniques

7. Visualization & Result Interpretation

8. Deployment-Ready Output

.]

4. Data Description

[Dataset: Ames Housing Dataset

Source: Kaggle

Type: Structured tabular data

Size: ~1,460 records, 80+ features

Target Variable: SalePrice

Notable Features: Lot Area, Year Built, Overall Quality, Location, Garage Type, etc.]

5. Data Preprocessing

[Handled missing values using median, mode, or predictive imputation

Encoded categorical features using one-hot and label encoding

Scaled numerical features using StandardScaler and MinMaxScaler

Detected and removed outliers using z-score and IQR

Verified data types, fixed anomalies, and standardized formats.]

6. Exploratory Data Analysis (EDA)

[Explored price distribution and feature correlations

Visualized how features like size, location, and quality affect price

Identified multicollinearity using heatmaps

Detected non-linear relationships in certain variables

Used scatter plots, box plots, and bar charts for insights

.]

7. Feature Engineering

[Created new features like total square footage, age of house, and neighborhood value

Combined multiple variables to improve prediction (e.g., bathrooms + bedrooms)

Applied log transformation to skewed features

Selected top features using correlation, mutual information, and feature importance

Removed irrelevant or highly collinear features.]

8. Model Building

[Regression models used: Linear Regression, Ridge, Lasso, Random Forest, XGBoost

Smart techniques: Cross-validation, Grid Search, Regularization

Metrics used: RMSE, MAE, R^2 Score

XGBoost yielded the best accuracy with optimized hyperparameters

Linear models provided interpretability, tree-based models captured non-linear effects].

9. Visualization of Results & Model Insights

[RMSE & R^2 comparison across models

Feature importance bar charts

Actual vs Predicted scatter plot

Residual plot to check model stability

Insights: Neighborhood, size, and quality are top predictors.]

10. Tools and Technologies Used

[Programming: Python

IDE: Jupyter Notebook

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

Visualization: matplotlib, seaborn, Plotly

Techniques: Regularization, Ensemble Learning, Cross-Validation]

11. Team Members and Contributions

[kumran.s]: Data Cleaning, Feature Engineering

[G.dhinesh]: EDA, Model Training

[M.Barath]: Model Evaluation, Tuning

[Bharathan]: Visualization, Documentation]