***Phase-1 Submission***

***Student Name:*** *IPSITHA.G*

***Register Number:*** *712523104027*

***Institution:*** *PPG INSTITUTE OF TECHNOLOGY*

***Department:*** *BE/CSE*

***Date of Submission:***

***1. Problem Statement***

*The goal is to accurately forecast house prices using advanced regression techniques in data science. House prices are influenced by a wide range of factors such as location, size, number of rooms, proximity to schools, and more.*

***2. Objectives of the Project***

* *Build a predictive model that accurately forecasts house prices based on multiple features.*
* *Analyze and identify key features that significantly influence house pricing.*
* *Evaluate the performance of different regression models and select the most effective one.*

***3.Scope of the Project***

* *Square footage, number of bedrooms and bathrooms, location, year built, etc*
* *Use of only regression-based models (e.g., Linear, Ridge, Lasso, XGBoost).*
* *Static dataset (not real-time data).*

***4.Data Sources***

***Dataset:*** *Ames Housing Dataset (available on Kaggle)*

***Source:*** *Public Dataset* [*https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset*](https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset)

***Type:*** *Static*

***Description:*** *Contains detailed information on residential homes in Ames, Iowa with 80+ features.*

***5.High-Level Methodology***

* ***Data Collection:*** *Download dataset from Kaggle.*
* ***Data Cleaning:*** *Handle missing values, remove duplicates, and fix inconsistencies.*
* ***Exploratory Data Analysis (EDA):*** *Use histograms, correlation heatmaps, and pair plots to discover patterns.*
* ***Feature Engineering:*** *Create new features (e.g., age of house), apply transformations (log/square root), and encode categorical variables.*
* ***Model Building:*** *Test Linear Regression, Ridge, Lasso, Random Forest Regressor, and XGBoost.*
* ***Model Evaluation:*** *Use RMSE, MAE, R² score, and cross-validation.*
* ***Visualization & Interpretation:*** *Use Seaborn and Matplotlib to visualize key insights and model predictions.*
* ***Deployment (Optional):*** *Deploy a web app using Streamlit to allow users to input home features and predict prices.*

***6.Tools and Technologies***

* ***Programming Language:*** *Python*
* ***Notebook/IDE:*** *Google Colab / Jupyter Notebook*
* ***Libraries:*** *pandas, numpy, seaborn, matplotlib, scikit-learn, xgboost*
* ***Deployment Tools (Optional):*** *Streamlit*

***7.Team Members and Roles***

|  |  |
| --- | --- |
| ***Team Members*** | ***Roles*** |
| *Athira. S* | *Project lead, EDA and Visualization* |
| *Ipsitha. G* | *Data Engineer* |
| *Harivarshini. R* | *Model selection* |
| *Devipriya. S* | *Model evaluator* |
| *Tyson Roy* | *Presentation and Deployment* |