**Phase-1 Submission**

**Student Name:** Ahisha J P

**Register Number:** 712523205003

**Institution:** PPG Institute of Technology

**Department:** B tech Information Technology

**Date of Submission:** 02/05/2025

**1.ProblemStatement**

1. *In the real estate market, accurately predicting house prices is crucial for buyers, sellers, and investors.*
2. *However, due to market volatility and the influence of numerous variables such as location, size, condition, and neighborhood amenities, estimating property prices can be challenging.*
3. *This project aims to build a robust regression model to predict house prices with high accuracy, thereby aiding stakeholders in making informed decisions.*

**2.Objectives of the Project**

* *Develop a predictive model using regression techniques to estimate house prices.*
* *Analyze which features have the greatest impact on pricing.*
* *Evaluate the performance of different regression algorithms.*
* *Create an interpretable model that can assist in real-world pricing decisions.*

**3.Scope of the Project**

* *Key features to analyze: number of bedrooms, bathrooms, square footage, location, year built, etc.*
* *Focus on using public datasets (e.g., from Kaggle).*
* *Limited to regression-based models (Linear, Ridge, Lasso, Random Forest, XGBoost).*
* *Project will be implemented in a Jupyter notebook, not deployed.*

**4.Data Sources**

* *Dataset: Ames Housing Dataset from Kaggle.*
* *Public dataset, downloaded once (static).*
* *Contains a wide range of features useful for price prediction.*

Data Source:<https://www.kaggle.com/datasets/samuelcortinhas/house-price-prediction-seattle>

**5.High-Level Methodology**

* *Data Collection: Download from Kaggle.*
* *Data Cleaning: Handle missing values, outliers, and data formatting.*
* *EDA: Use seaborn/matplotlib for visualizations, correlation heatmaps.*
* *Feature Engineering: Encode categorical variables, log transformation, polynomial features.*
* *Model Building: Test Linear Regression, Ridge, Lasso, Decision Tree, Random Forest, XGBoost.*
* *Model Evaluation: Use RMSE, R² Score, Cross-Validation for evaluation.*
* *Visualization & Interpretation: Present feature importance, residual plots.*
* *Deployment: Not deploying in Phase-1; future deployment may use Stream lit.*

**6.Tools and Technologies**

* ***Programming Language:*** *Python*
* ***Notebook/IDE:*** *Jupyter Notebook*
* ***Libraries:*** *pandas, NumPy, seaborn, matplotlib, scikit-learn, xgboost*
* ***Optional Deployment Tools:*** *Stream lit or Flask (for future scope)*

**7.Team Members and Roles**

|  |  |  |
| --- | --- | --- |
| *Name* | *Role* | *Responsibilities* |
| *Manoj M* | *Data Acquisition& Initial Analysis* | *Responsible for data collection and preliminary analyses, ensuring the dataset*  *is clean and ready for exploration.* |
| *John Isaac K.* | *EDA & Visualization Expert* | *Leads the exploratory data analyses (EDA) and assists in visualizing patterns and trends.* |
| *Bharathi Kannan V. K* | *Feature Engineering Lead* | *Incharge of feature engineering and transformation to enhance model performance.* |
| *Ahisha J. P* | *Model Development Tuning* | *Handles model selection, training and fine-tuning of various regression algorithms.* |
| *Madhumitha V.* | *Evaluation & Reporting Specialist* | *Oversees model evaluation, documentation, and presentation of results in a clear and structure format.* |