**Phase-3 Submission Template**

**Student Name:** Devipriya S S

**Register Number:** 712523104013

**Institution:** PPG INSTITUTE OF TECHNOLOGY

**Department:** BE.CSE

**Date of Submission:**

**Github Repository Link: <https://github.com/SS-Devipriya/NM_DEVIPRIYA_DS.qit>**

### **1. Problem Statement**

*You are solving a regression problem where the goal is to predict the price of a house based on its features (e.g., area, number of bedrooms, location, etc.). This is important for real estate companies, buyers, and investors to make informed decisions*

### **2. Abstract**

*Your project uses a dataset (e.g., from Kaggle) to forecast house prices. The aim is to preprocess data, explore patterns, build and compare multiple regression models (e.g., Linear Regression, Random Forest, XGBoost), and deploy the best-performing one. The outcome is a web-based tool where users can input house features and get price predictions.*

### **3. System Requirements**

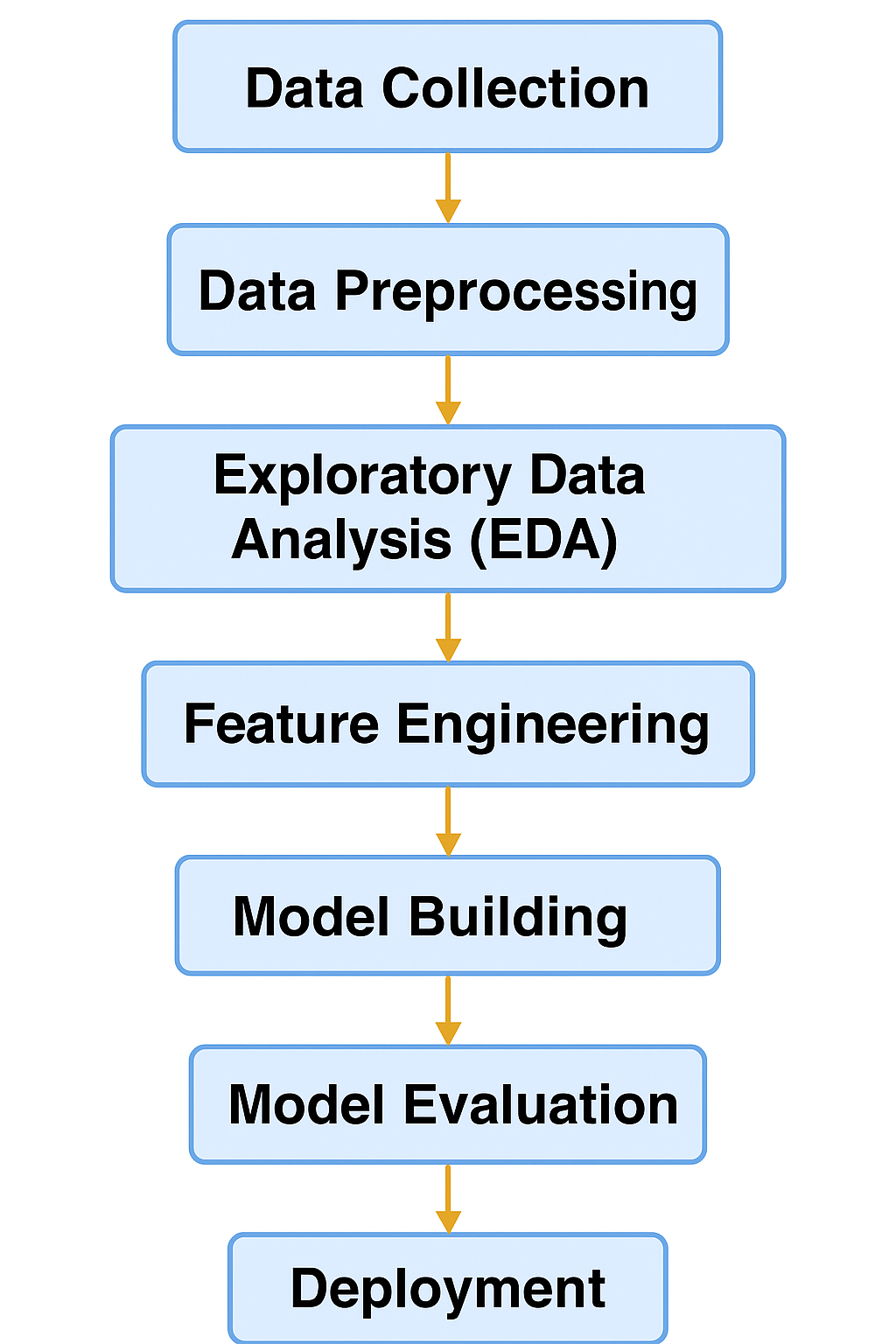
* ***Hardware****: 4 GB RAM, Intel i5 or above*
* ***Software****: Python 3.10, Jupyter Notebook/Colab****Libraries****: pandas, numpy, scikit-learn, matplotlib, seaborn, xgboost, streamlit*

### 

### **4. Objectives**

* *Predict house prices with high accuracy*
* *Identify key features influencing house prices*
* *Create a web-based tool for end users to predict house prices*
* *Deliver actionable insights for real estate stakeholders*

**5. Flowchart of Project Workflow**

*[*

### **6. Dataset Description**

* ***Source****: Kaggle (e.g., "House Prices - Advanced Regression Techniques")*
* ***Type****: Public*
* ***Size****: ~1500 rows × 80 columns (varies)*
* ***Include****: df.head() screenshot*

### 

### **7. Data Preprocessing**

### *Handle nulls using mean/median or drop*

### *Encode categorical data using One-Hot or Label Encoding*

### *Scale numeric values using MinMax or StandardScaler Include before/after screenshots of transformations*

### **8. Exploratory Data Analysis (EDA)**

* *Histograms (to understand distributions)*
* *Boxplots (to detect outliers)*
* *Heatmaps (to see correlation)  
  Write 3–5 insights such as:*
* *"GrLivArea is highly correlated with price"*
* *"Houses with more bathrooms tend to cost more"*

### 

### **9. Feature Engineering**

* *Create new features like Age = YrSold – YearBuilt*
* *Use feature selection (e.g., SelectKBest or Feature Importance from models)*
* *Explain impact of selected features*

### 

### **10. Model Building**

* *Try multiple models (baseline and advanced)*
* *Explain why those models were chosen*
* *Include screenshots of model training outputs*

### 

### **11. Model Evaluation**

* ***RMSE, MAE, R² Score***
* *Visuals: Residual plots, Prediction vs Actual*
* *Table comparing all models  
  Include screenshots*

### **12. Deployment**

*Use* ***Streamlit Cloud*** *or* ***Gradio + HuggingFace***

* *Include UI screenshot and prediction example*
* *Deployment link (*[*e.g., https://house-price-predictor.streamlit.app*](file:///C:\Users\ADMIN\Downloads\e.g.,%20https:\house-price-predictor.streamlit.app)*)*

**13. Source code**

*Upload the .ipynb or .py files to GitHub*

[*https://github.com/SS-Devipriya/NM\_DEVIPRIYA\_DS.qit*](https://github.com/SS-Devipriya/NM_DEVIPRIYA_DS.qit)

**14. Future scope**

* *Use more location-based data (e.g., crime rate, school rating)*
* *Integrate with real-time property listings*
* *Convert into a mobile app or integrate chatbot for queries*

**13. Team Members and Roles**

|  |  |
| --- | --- |
| ***Team Members*** | ***Contribution*** |
| ***Devipriya.SS*** | ***Data cleaning*** |
| ***Harivarshini.R*** | ***EDA*** |
| ***Ipsitha.G*** | ***Feature engineering*** |
| ***Athira.s*** | ***Model development*** |
| ***Ch.Tyson*** | ***Documentation and reporting*** |