**Phase-2 Submission Template**

**Student Name:** RANJITHKUMAR S

**Register Number:** 412023106308

**Institution:** SRI KRISHNA ENGINEERING COLLEGE :: 4120

**Department:** B.E. Electronics and Communication Engineering

**Date of Submission:** 04-05-2025

**Github Repository Link:** [Update the project source code to your Github Repository]

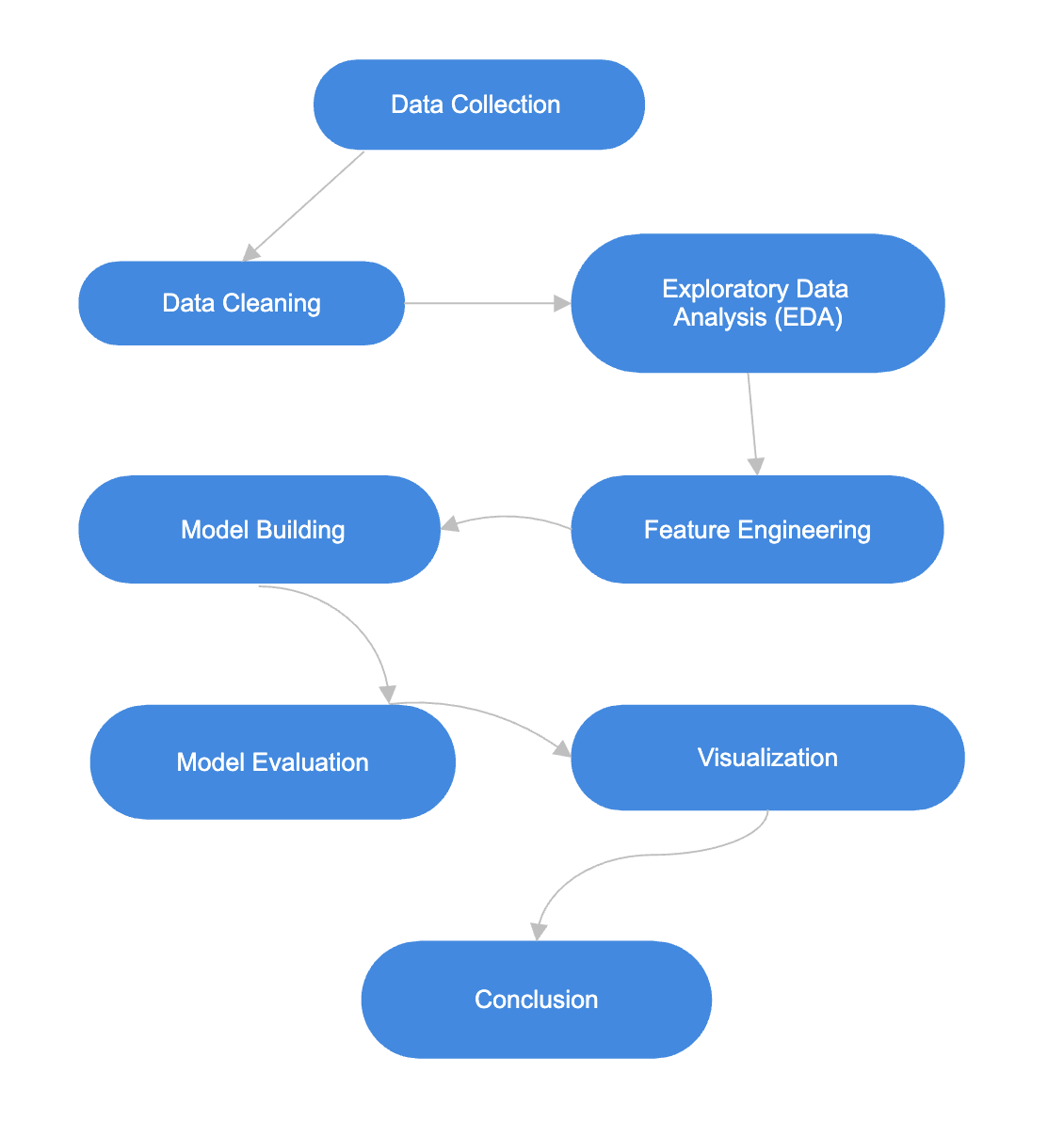
# 1. Problem Statement

* Accurately predicting house prices is a complex task due to various factors such as location, size, amenities, and market trends. Existing models often struggle to capture the intricate relationships between these factors, leading to inaccurate predictions.

# 2. Project Objectives

* **Develop an accurate predictive model :** Create a robust regression model that accurately forecasts house prices based on historical data and relevant features.
* **Identify key predictors:** Determine the most influential factors affecting house prices, such as location, size, amenities, and market trends.
* **Evaluate model performance:** Assess the performance of different regression techniques and identify the best approach for house price prediction.

**3. Flowchart of the Project Workflow**



# 4. Data Description

* **Target Variable:** House Price (continuous variable)
* **Predictor Variables:**

- Location (categorical variable: city, state, zip code)

- Size (continuous variable: square footage)

- Number of Bedrooms (integer variable)

- Number of Bathrooms (integer variable)

- Age of Property (integer variable)

- Amenities (categorical variable: pool, garden, garage, etc.)

- Market Trends (continuous variable: historical price trends)

# 5. Data Preprocessing

* **Handling Missing Values:** Imputing missing values using mean, median, or imputation algorithms.
* **Data Normalization:** Scaling numeric features to a common range (e.g., 0-1) to prevent feature dominance.
* **Encoding Categorical Variables:** Converting categorical variables into numerical representations (e.g., one-hot encoding, label encoding).
* **Removing Outliers:** Identifying and removing or transforming outliers to prevent skewing the model.
* **Feature Scaling:** Scaling features to a common range to improve model performance.
* **Data Transformation:** Applying transformations (e.g., logarithmic, square root) to stabilize variance or improve linearity.

# 6. Exploratory Data Analysis (EDA)

**EDA Objectives**

**1. Understand Data Distribution:** Visualize and summarize the distribution of variables.

**2. Identify Relationships:** Explore relationships between variables and the target variable (house price).

**3. Detect Outliers:** Identify and understand the impact of outliers on the data.

**EDA Techniques**

**1. Summary Statistics:** Calculate mean, median, mode, and standard deviation for numerical variables.

**2. Data Visualization:** Use plots (e.g., histograms, scatter plots, box plots) to visualize data distribution and relationships.

**3. Correlation Analysis:** Calculate correlation coefficients to identify relationships between variables.

**4. Heatmaps:** Visualize correlation matrices to identify patterns and relationships.

# 7. Feature Engineering

**Feature Engineering Ideas**

**1. Location-Based Features:** Create features based on location data (e.g., proximity to amenities, neighborhood characteristics).

**2. Property Characteristics:** Engineer features that capture property characteristics (e.g., age, size, condition).

**3. Market Trends:** Create features that capture market trends (e.g., historical price trends, seasonality).

**4. Interaction Terms:** Create interaction terms between variables to capture complex relationships.

**Benefits of Feature Engineering**

**1. Improved Model Accuracy:** By capturing complex relationships between variables.

**2. Reduced Overfitting:** By selecting relevant features and reducing noise.

**3. Enhanced Interpretability:** By creating features that are meaningful and interpretable.

# 8. Model Building

**Model Building Objectives**

**1. Develop an Accurate Model:** Create a robust regression model that accurately forecasts house prices.

**2. Compare Model Performance:** Evaluate and compare the performance of different regression techniques.

**Regression Techniques**

**1. Linear Regression:** A linear model that assumes a linear relationship between variables.

**2. Ridge Regression:** A linear model with regularization to prevent overfitting.

**3. Lasso Regression:** A linear model with L1 regularization to select relevant features.

**4. Elastic Net Regression:** A linear model with both L1 and L2 regularization.

**5. Gradient Boosting Regression:** An ensemble model that combines multiple weak models to create a strong predictive model.

**6. Random Forest Regression:** An ensemble model that combines multiple decision trees to create a robust predictive model.

**Model Evaluation Metrics**

**1. Mean Absolute Error (MAE):** Measures the average difference between predicted and actual values.

**2. Mean Squared Error (MSE):** Measures the average squared difference between predicted and actual values.

**3. Root Mean Squared Error (RMSE):** Measures the square root of the average squared difference between predicted and actual values.

**4. R-Squared (R2):** Measures the proportion of variance in the target variable explained by the model.

# 9. Visualization of Results & Model Insights

**Visualization Objectives**

**1. Communicate Results:** Effectively communicate model performance and insights to stakeholders.

**2. Identify Trends:** Visualize trends and patterns in the data to gain a deeper understanding.

**Visualization Techniques**

**1. Scatter Plots:** Visualize the relationship between predicted and actual house prices.

**2. Residual Plots:** Visualize the residuals to identify patterns and outliers.

**3. Feature Importance Plots:** Visualize the importance of each feature in the model.

**4. Heatmaps:** Visualize the correlation between variables.

**Benefits of Visualization**

**1. Improved Communication:** Effectively communicate results and insights to stakeholders.

**2. Deeper Understanding:** Gain a deeper understanding of the data and model performance.

**3. Informed Decision-Making:** Provide valuable insights for informed decision-making.

# 10. Tools and Technologies Used

**Programming Languages**

**1. Python:** A popular language for data science and machine learning.

**2. R**: A language for statistical computing and data visualization.

**Data Visualization Tools**

**1. Matplotlib:** A popular data visualization library in Python.

**2. Seaborn:** A visualization library built on top of Matplotlib.

**3. Plotly:** An interactive visualization library.

**11. Team Members and Contributions**

|  |  |  |
| --- | --- | --- |
| **Name** | **Role** | **Responsibilities** |
| DURGASHREE T | Project Lead | Data Collection, Model Building, Report Preparation |
| RANJITH KUMAR S | Data Engineer | Data Cleaning, Text Preprocessing, Visualization |
| SYAMA MURALI P | Data Analyst / Feature Lead | Exploratory Data Analysis (EDA), Feature Engineering |
| RESHMA I M | QA & Other Lead | Model Evaluation, Final Review, Presentation Preparation |