**Phase-2 Submission Template**

**Student Name:** Priyadharshini B

**Register Number:** 622023104081

**Institution:** Paavai College Of Engineering

**Department:** Computer Science And Engineering

**Date of Submission:**

**Github Repository Link:** [**https://github.com/priyadharshini1558/Predicting-air-quality-**](https://github.com/priyadharshini1558/Predicting-air-quality-)

# 1. Problem Statement

* **Real-World Problem:**

Industrial emissions significantly contribute to air pollution, which poses serious threats to environmental health and public safety. Tracking and visualizing the concentration of pollutants like PM10, O3, CO, NO2, and SO2 is crucial for understanding pollution levels, identifying trends, and mitigating environmental damage.

* **Refining the Problem:**

Based on the provided dataset (air\_dataset.csv), you aim to identify patterns, visualize distributions, and analyze the behavior of key air pollutants emitted during industrial processes. By refining these insights, you can support environmental policy development or pollution control strategies.

* **Type of Problem:** Classification problem is used for predict air quality level.
* **This analysis falls under:**

1. **Exploratory Data Analysis (EDA):** Visualizing patterns, trends, and distributions.
2. **Potential Regression Problem:** Predicting pollutant levels based on influencing factors like industrial activities or time.
3. **Clustering Problem:** Grouping locations or emissions by pollution severity to focus mitigation efforts.

**Why Solving This Problem Matters:**

* **Impact:** Pollution control directly contributes to reducing health risks such as respiratory and cardiovascular diseases, especially in industrial areas.
* **Relevance:** Insights can shape environmental policies, industrial regulations, and sustainable development plans.
* **Application:** Results can be applied in urban planning, resource allocation for clean-up efforts, or improving industrial emission standards globally.

# 2. Project Objectives

**Key Technical Objectives**

* Analyze air pollution data, detect trends, and visualize pollutant distributions effectively.
* Develop predictive models or clustering techniques to assist in industrial emission management.

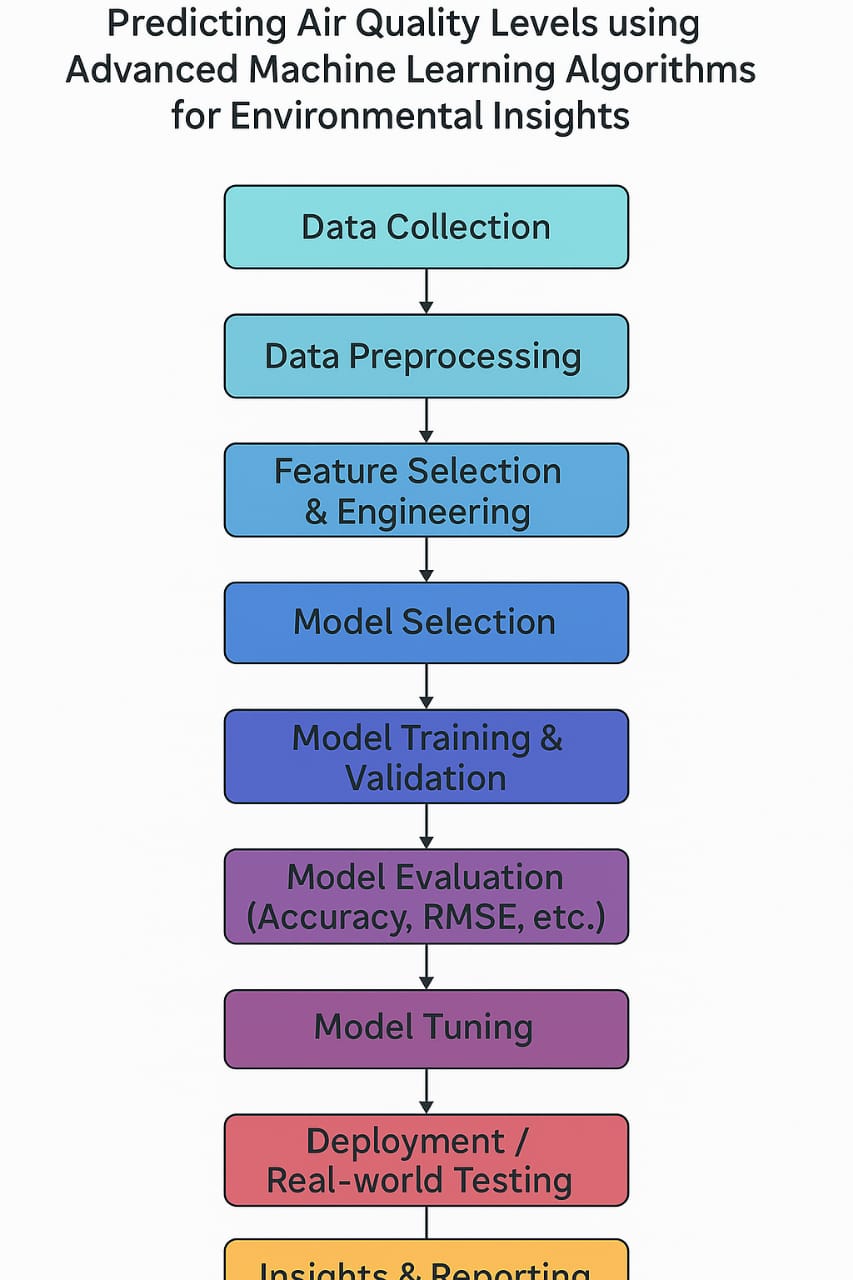
**Model Goals**

* Achieve high accuracy and ensure interpretability for stakeholders.
* Provide real-world applicability by offering actionable insights for emission control policies.

**Evolution of Goals**

After data exploration, the focus may shift to predictive modeling of pollutant levels, identifying key sources, or applying advanced techniques like machine learning for deeper analysis.

**3. Flowchart of the Project Workflow**



**4. Data Description**

* **Dataset Name and Origin**: [air\_dataset.csv](file:///C:\Users\ADMIN\Documents\air_dataset.csv)
* **Type of Data:** Structured tabular data
* **Number of Records and features:** The model has multiple rows with air composition attributes like range and many plots like boxplot, histogram , barchart , piechart .
* **Static or Dynamic Dataset :** It is static dataset loaded from a CSV file.
* **Target Variable:** The air composition values like PM10,O3,CO,NO2 used for predictions.

**5. Data Preprocessing**

Data Preprocessing in the Project "Predicting Air Quality Levels":

In this project, data preprocessing involves cleaning and preparing environmental datasets by handling missing values (e.g., imputation), removing duplicates, treating outliers, converting data types, and encoding categorical variables (e.g., label and one-hot encoding) to ensure the data is suitable for machine learning models.

# 6. Exploratory Data Analysis (EDA)

* **Univariate Analysis:**Explored distribution of individual features using histograms, boxplots, and countplots.
* **Bivariate/Multivariate Analysis:** Used correlation matrix, pairplots, scatterplots, and grouped bar plots.Analyzed relationships between features and the target air quality index.
* **Insights Summary:** Highlighted key patterns and trends.

Identified important features influencing air quality predictions.

**7. Feature Engineering**

* Create new features based on domain knowledge or insights from EDA.
* Combine or split columns, such as extracting parts of date features.
* Apply feature transformation techniques like binning, creating polynomial features, and ratios.
* Use dimensionality reduction methods (optional), e.g., PCA.
* Justify each added or removed feature to ensure model relevance.

# 8. Model Building

* Decision Tree, Random Forest is selected to implement machine learning models.

* Choose this models based on nature of the data of air composition.

* Split data into training and testing sets using stratification need.

* Train models and evaluate initial performance using appropriate metrics.

* For classification: accuracy

* For regression: R² score.

# 9. Visualization of Results & Model Insights

* **Confusion Matrix**: Measures model accuracy (TP, TN, FP, FN).
* **ROC Curve**: Shows sensitivity vs. specificity (AUC close to 1 = better).
* **Feature Importance Plot**: Highlights key variables (e.g., PM2.5, humidity).
* **Residual Plots**: Assess prediction errors (random = good model).
* **Model Comparison**: Visualize performance differences (e.g., bar/box plots).

# 10. Tools and Technologies Used

* Programming Language: Python .
* IDE/Notebook: Google Colab
* Libraries: pandas, numpy, seaborn, matplotlib
* Visualization Tools: Plotly

# 11. Team Members and Contributions

**Soundhraya.E** – Data cleaning,EDA

**Priyadharshini.B –** Model development

**Unnamalai.R**  -Documentation and reporting