**Phase-2 Submission Template**

**Student Name:** A.Narendra reddy

**Register Number :** 9347401266

**Institution :** Akshaya college of engineering and technology

**Department:** Electronics and communication engineering

**Date of Submission:**

**Github Repository Link:** [https://github.com/<your-username>/air-quality-prediction](https://github.com/%3cyour-username%3e/air-quality-prediction)

# Problem Statement

Air pollution poses a significant threat to public health, climate, and ecosystems. Rapid urbanization and industrialization have led to a surge in harmful pollutants like PM2.5, PM10, NO₂, SO₂, CO, and O₃ in the atmosphere. Real-time monitoring and accurate prediction of air quality are essential for timely public health advisories, urban planning, and environmental policymaking.

This project aims to develop a machine learning-based solution to predict air quality levels using historical environmental and pollutant data. By learning from structured time-series and/or sensor datasets, the model can forecast Air Quality Index (AQI) or specific pollutant concentrations, enabling proactive measures to mitigate exposure risks.

* **Problem Type**: This is primarily a regression problem if predicting continuous AQI values or individual pollutant concentrations. It can also be formulated as a classification problem if the goal is to predict categorical AQI levels (e.g., Good, Moderate, Unhealthy).
* **Relevance**: Accurate air quality prediction can inform public health decisions, reduce pollution exposure, and support smart city planning. It can empower local governments and citizens with actionable environmental insights.

# Project Objectives

1. **Objective**: To develop a machine learning-based system to predict air quality (AQI level) using historical environmental data.
2. **Key Goals**:
   1. Improve prediction accuracy and model robustness.
   2. Create interpretable models that highlight key pollutants.
   3. Ensure practical applicability for public health and city planning.
3. **Updated Insight**: Based on EDA, we decided to classify AQI levels into categories instead of predicting continuous AQI values, to better inform public health alerts.

**3. Flowchart of the Project Workflow**

Data Collection

↓

Data Cleaning & Preprocessing

↓

Exploratory Data Analysis (EDA)

↓

Feature Engineering

↓

Model Building (Training & Testing)

↓

Model Evaluation

↓

Visualization & Interpretation

↓

Final Reporting

# 4. Data Description

 **Source**: OpenAQ Dataset via Kaggle

 **Type**: Structured, time-series

 **Records**: ~50,000 rows

 **Features**: 12+ (including PM2.5, PM10, NO₂, SO₂, CO, O₃, Temperature, Humidity)

 **Static/Dynamic**: Static snapshot of sensor data

 **Target Variable**: AQI Category (Good, Moderate, Unhealthy, etc.)

# 5. Data Preprocessing

* **Missing Values** : Handled using mean imputation for numerical features.
* **Duplicates** : Removed based on timestamp and location IDs.
* **Outliers** : Detected via IQR and treated.
* **Data Types** : Cleaned and converted (e.g., datetime parsing).
* **Encoding** : One-hot encoding for categorical columns (if any).
* **Scaling** : StandardScaler applied to continuous variables.

# Exploratory Data Analysis (EDA)

**Univariate**: Histograms for pollutant levels, boxplots for outliers.

**Bivariate**: Correlation heatmap to identify key predictors of AQI.

**Multivariate**: Scatterplots of AQI vs PM2.5, NO₂.

**Insights**:

* 1. PM2.5 and NO₂ are strong predictors of AQI.
  2. AQI levels spike during winter and evening hours.

# Feature Engineering

* Created new features:
  + **Pollutant Ratios**: PM2.5/PM10
  + **Datetime Decomposition**: Hour, day, month extracted.
* **Dimensionality Reduction**: PCA explored but not implemented due to high v interpretability loss.

# 8. Model Building

* **Algorithms**: Random Forest Classifier, XGBoost Classifier
* **Train-Test Split**: 80/20 with stratified sampling on AQI category.
* **Metrics**:
  + Accuracy: ~89%
  + Precision, Recall, F1-score (macro avg): ~87%
* **Best Performing Model**: XGBoost with highest F1-score.

# 9. Visualization of Results & Model Insights

* **Confusion Matrix**: Visual validation of classification.
* **ROC Curve**: AUC > 0.90
* **Feature Importance Plot**: PM2.5 and NO₂ as most impactful features.
* **Residual Analysis**: Misclassification mostly between neighboring AQI categories.

# 10. Tools and Technologies Used

* **Language**: Python
* **IDE**: Google Colab
* **Libraries**: pandas, numpy, seaborn, matplotlib, scikit-learn, xgboost
* **Visualization**: matplotlib, seaborn, Plotly

# 11. Team Members and Contributions

M.Naga yashwanth - Data cleaning &pre- processing

A.Narendra reddy - Exploratory data analysis

B.Kathik - Model building &evaluation

G.Mahendra - Documantation and reporting