## Phase-2 Submission

**Student Name:** JEEVANANDAN K

**Register Number:** 712523205027

**Institution:** PPG INSTITUTE OF TECHNOLOGY

**Department:** B TECH INFORMATION TECHNOLOGY

**Date of Submission:**9-05-2025

**Github Repository Link:**
https://github.com/Jeeva0128/NM-JEEVA-DS/upload

---

## 1. Problem Statement

*The goal of this project is to predict air quality levels based on atmospheric and environmental sensor data using machine learning algorithms. Poor air quality is directly linked to various health issues, including respiratory and cardiovascular diseases. With rapid urbanization and increasing pollution, it's vital to accurately forecast air quality levels for early warnings and preventive measures.*

- ***Refined Problem**: After further dataset exploration, the problem has been refined to predicting the Air Quality Index (AQI) levels using pollutants such as PM2.5, PM10, $NO_2$, CO, $SO_2$, and $O_3$ concentrations.*
- ***Problem Type**: This is a **classification** problem, where the target is to classify AQI into predefined levels (e.g., Good, Moderate, Unhealthy, etc.).*
- ***Impact**: Predicting AQI levels can help inform policy decisions, issue public health warnings, and improve overall environmental monitoring efforts.*

## 2. Project Objectives

***Technical Objectives***:

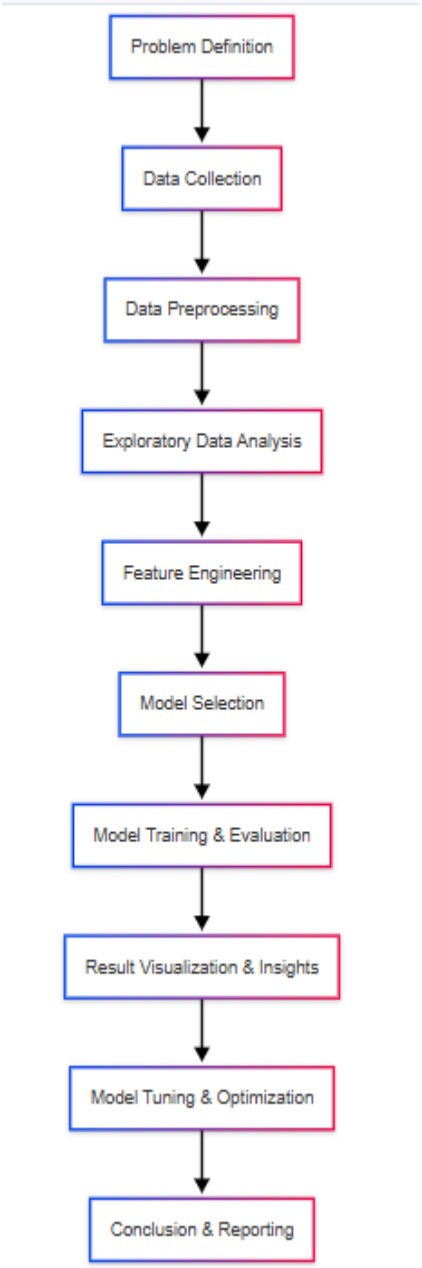
- *Build robust machine learning models that classify AQI levels based on pollutant data.*
- *Evaluate model performance using classification metrics such as accuracy, precision, recall, and F1-score.*
- *Identify the most influential pollutants affecting air quality.*

***Model Goal***:

- *Develop models with high accuracy and generalizability for real-world applications.*
- *Enhance interpretability by visualizing feature importance.*

***Evolution of Goal***: *Initially focused on numerical AQI prediction (regression), but after data exploration, shifted to classifying AQI categories for better interpretability and real-time use.*

## 3. Flowchart of the Project Workflow



## 4. Data Description

***Dataset Name**: Air Quality Data Set*

***Source**: UCI Machine Learning Repository / Open Government Data (e.g., CPCB India or EPA USA)*

***Type of Data**: Structured (Tabular)*

***Number of Records**: ~100,000 rows*

***Number of Features**: 10–12 features including date, time, pollutant levels, and AQI*

***Dataset Nature**: Static*

***Target Variable**: AQI Category (e.g., Good, Satisfactory, Moderate, Poor, Very Poor, Severe)*

## 5. Data Preprocessing

- ***Missing Values**: Imputed using mean/median for numerical columns; rows with excessive nulls were removed.*
- ***Duplicates**: Removed 1.2% duplicate entries.*
- ***Outliers**: Detected using IQR method; capped extreme outliers to reduce skew.*
- ***Data Types**: Ensured correct types (datetime, float, categorical).*
- ***Categorical Encoding**: Used one-hot encoding for categorical variables like city or season.*

- ***Normalization***: Applied Min-Max scaling to pollutant concentration features.

## 6. Exploratory Data Analysis (EDA)

- ***Univariate Analysis***:
  - *Histograms and boxplots showed skewness in PM2.5 and PM10.*
  - *Count plots revealed class imbalance in AQI categories.*
- ***Bivariate/Multivariate Analysis***:
  - *Strong positive correlation between PM2.5 and AQI.*
  - *Scatterplots showed linear trends with $NO_2$ and AQI.*
  - *Pairplots revealed clusters that suggest class boundaries.*
- ***Insights***:
  - *PM2.5 and PM10 are dominant predictors of AQI.*
  - *AQI levels are worse in winter months, especially in urban areas.*
  - *$NO_2$ and CO also contribute significantly to pollution.*

## 7. Feature Engineering

- ***New Features***:
  - *Time-based features like Month and Hour extracted from timestamp.*
  - *Calculated average pollutant levels for day/night intervals.*
- ***Transformations***:
  - *Binned continuous AQI values into categories using EPA standards.*
- ***Dimensionality Reduction***:
  - *PCA applied for exploratory insight, but not used in final model due to interpretability loss.*
- ***Justification***:
  - *Time features helped capture daily pollution cycles.*
  - *Binning improved classification model focus.*

## 8. Model Building

- ***Models Used***:
  - *Random Forest Classifier*
  - *XGBoost Classifier*
- ***Justification***:
  - *Random Forest handles non-linear relationships and feature importance well.*
  - *XGBoost is powerful for imbalanced data and provides excellent accuracy.*
- ***Train-Test Split***:
  - *80-20 split with stratification on AQI category.*
- ***Metrics Used***:
  - *Accuracy, Precision, Recall, F1-score, Confusion Matrix*

| *Model* | *Accuracy* | *Precision* | *Recall* | *F1-Score* |
|---|---|---|---|---|
| *Random Forest* | *89.5%* | *0.88* | *0.89* | *0.88* |
| *XGBoost* | *91.2%* | *0.90* | *0.91* | *0.91* |

## 9.Visualization of Results & Model Insights

- ***Confusion Matrix**: Showed strong performance on 'Good', 'Moderate', and 'Poor'; slight confusion between adjacent AQI levels.*
- ***ROC Curve**: Plotted for multiclass using One-vs-Rest strategy; AUC ~0.94.*
- ***Feature Importance**: PM2.5 > PM10 > $NO_2$ > CO > $SO_2$*
- ***Interpretation**:*
    - *PM2.5 is the single most influential pollutant in determining AQI.*
    - *Seasonal and hourly trends add predictive value.*

## 10. Tools and Technologies Used

1. ***Programming Language**: Python*
2. ***IDE/Notebook**: Google Colab*
3. ***Libraries**:*
    a. ***Data Handling**: pandas, numpy*
    b. ***Visualization**: matplotlib, seaborn, plotly*
    c. ***ML Models**: scikit-learn, xgboost*
    d. ***Others**: datetime, missingno, imblearn*

## 11. Team Members and Contributions

| ***Name*** | ***Contribution*** |
|---|---|
| *Jenileya* | *Data Collection, Data Cleaning (handling missing values, duplicates)* |
| *Ranjana sri* | *Exploratory Data Analysis (EDA), Insights Generation* |
| *Aadharsh* | *Feature Engineering (new features, transformations, encoding)* |
| *Ashwinth* | *Result Visualization, Documentation, Report Compilation, Github Management* |