**Phase-2 Submission**

**Student Name:** Sivabalan V

**Register Number:** 712523104060

**Institution:** PPG Institute of Technology

**Department:** B.E. Computer Science and Engineering

**Date of Submission:** 25/04/2025

**Github Repository Link: [AirPreQ](https://github.com/Siva-Balan-V/NM_Sivabalan_AirPreQ)**

### **1. Problem Statement**

Accurately predicting **Air Quality Index (AQI)** using machine learning models can significantly enhance public awareness, assist governments in environmental policy making, and help prevent health risks due to pollution. This project refines the Phase-1 objective by focusing on supervised **regression** to estimate AQI levels based on pollutant and meteorological data

**Type of Problem**: Supervised Regression

**Why It Matters**: Real-time AQI prediction enables timely public advisories, health risk mitigation, and supports data-driven urban planning.

### **2. Project Objectives**

* To build an accurate AQI prediction model using pollution and weather features.
* To identify key variables contributing to poor air quality.
* To implement real-time prediction capabilities through a deployed interface.
* To enhance interpretability using feature importance and model visualization.

**Updated Goal**: After data exploration, the focus narrowed to regression (rather than classification or clustering), and some limitations were noted in model complexity due to time and resource constraints.

### **3. Flowchart of the Project Workflow**

### **Flowchart**

### **4. Data Description**

· **Sources**:

[Github](https://github.com/AbiAyshwariya/Aitr_Quality_Prediction/blob/master/city_day%20(1).csv)

**Type**: CSV(Comma Seperated values)

**Features**: ~15 features including PM2.5, PM10, NO₂, SO₂, CO, O₃, temperature, humidity, pressure

**Target Variable**: AQI

**Dataset Nature**: Dynamic (real-time from APIs), integrated into a static training set

### **5. Data Preprocessing**

**Missing Values**: Imputed using mean/median

**Duplicates**: Removed using Pandas’ .drop\_duplicates()

**Outliers**: Detected using IQR; capped/fixed extreme pollution values

**Data Types**: Timestamps converted to datetime; all numerics coerced to float

**Encoding**: Categorical location data one-hot encoded

**Scaling**: StandardScaler used on pollutant concentration values

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

**Univariate**

Boxplots revealed outliers in PM2.5 and CO

AQI distribution was right-skewed

**Bivariate**:

Correlation matrix: PM2.5 and PM10 had high positive correlation with AQI

Time series plots: Seasonal variations visible, with spikes during winter months

**Key Insights**:

PM2.5, NO₂, and O₃ are strong contributors to poor air quality

Humidity inversely correlated with AQI in some cases (likely due to weather dispersion effects)

### **7. Feature Engineering**

Extracted **hour**, **day**, **month** from timestamps

Created **pollution index** by averaging key pollutants

Added **lag features** for short-term time-series awareness

Applied **log transformation** to highly skewed features

Optional:

PCA applied for dimensional reduction (retained 95% variance in 6 components)

### **8. Model Building**

**Models Used**:

**Linear Regression**: Used as baseline

**Random Forest Regressor**: For robustness and non-linear relationships

**Train-Test Split**: 80/20  
**Evaluation Metrics**:

**Linear Regression**: RMSE = 45.6, R² = 0.65

**Random Forest**: RMSE = 29.3, R² = 0.83

Random Forest outperformed the baseline significantly.

### **9. Visualization of Results & Model Insights**

**Feature Importance**: Random Forest showed PM2.5, PM10, and NO₂ as top contributors

**Residual Plot**: Random Forest residuals randomly distributed around zero (good fit)

**Prediction vs Actual Plot**: Close diagonal fit in test set

**Interactive Streamlit Dashboard**: Real-time input → AQI prediction displayed alongside chart.

### **10. Tools and Technologies Used**

* *Programming Language: Python or R.*
* *IDE/Notebook:* ***VS Code.***
* *Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, streamlit.*
* *Version control:* ***git***

### **11. Team Members and Contributions**

|  |  |  |
| --- | --- | --- |
| Name | Role | Description |
| Sivabalan V | Project Manager | Led the team during model implementation phase, conducted detailed EDA, and derived key insights. |
| Dhyanesh V | Model Integration & Application Testing Lead | Integrated trained ML models into the Streamlit app, tested end-to-end functionality, handled input validation, and prepared the codebase for future deployment. |
| Semmozhiyan NS | Machine Learning Engineer | Built and fine-tuned Linear Regression and Random Forest models, handled training and evaluation |
| Sri Sabarish U | Data Preprocessing Lead | Cleaned and transformed the dataset (handled missing values, outliers, and encoding). |
| Chandru M | UI Developer + Visual Analyst | Developed the Streamlit dashboard, created interactive plots for AQI visualization and prediction |