# Phase-II

# Predicting Air Quality Levels Using Advanced Machine Learning Algorithms

Student Name: S.Kokila

Register Number :630123106055

Institution: AVS ENGINEERING COLLEGE

Department: ECE

Date of submission:10.05.2025

GitHub Repository Link:

http://https://github.com/kokila2410/Kokila.S-Naan-Mudhalvan-project-.git

## 1.Problem Statement

Air pollution poses significant health and environmental risks worldwide. Traditional methods of monitoring air quality can be limited in spatial and temporal coverage. The goal is to leverage advanced machine learning techniques to accurately predict air quality index (AQI) levels using environmental data, enabling early warnings and informed decision-making.

## 2. Project Objective

- Develop a machine learning model to predict AQI based on real-time and historical environmental data.
- Identify key factors influencing air quality.
- Provide visual insights into pollution trends to support policy makers and citizens.
- Improve accuracy and reliability compared to conventional statistical models.

## 3. Project Workflow (Flowchart)

Data Collection
↓
Data Preprocessing
↓
Exploratory Data Analysis (EDA)
↓
Feature Engineering
↓
Model Building & Evaluation
↓
Result Visualization & Interpretation
↓
Deployment (Optional)

## 4. Data Description

The dataset includes air quality measurements and environmental factors:
- Features: PM2.5, PM10, NO2, SO2, CO, O3, temperature, humidity, wind speed, pressure, etc.
- Target: Air Quality Index (AQI) or categorized levels (Good, Moderate, Poor, etc.)
- Source: OpenAQ, UCI Machine Learning Repository, Kaggle Datasets, Government APIs.

## 5. Data Preprocessing

- Handling missing values using interpolation or imputation techniques.
- Outlier detection using IQR or Z-score.
- Normalization or standardization of features.
- Encoding categorical data (if any).
- Time-based aggregation (hourly/daily).

## 6. Exploratory Data Analysis (EDA)

- Distribution plots for pollutant concentrations.
- Correlation heatmap between variables.
- Time series trends for pollutants.
- AQI variation by time of day, month, and location.
- Boxplots and histograms for pollutant levels.

## 7. Feature Engineering

- Creation of new features (e.g., pollutant ratios, moving averages).
- Time-based features (hour, day, month).
- Encoding AQI levels into categories for classification models.
- Dimensionality reduction (e.g., PCA) if required.

## 8. Model Building

Algorithms used:

- Linear Regression / Logistic Regression
- Random Forest
- Gradient Boosting Machines (XGBoost, LightGBM)
- Support Vector Machines
- Neural Networks (optional)

Model Evaluation Metrics:

- Regression: RMSE, MAE, $R^2$ Score
- Classification: Accuracy, Precision, Recall, F1-Score

## 9. Result Visualization & Model Insights

- Predicted vs actual AQI plot.
- Confusion matrix (for classification).
- Feature importance graphs (SHAP, permutation importance).
- Residual plots.
- Interactive dashboards using Plotly or Streamlit.

## 10. Tools and Technologies Used

- Programming Language: Python
- Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, XGBoost, LightGBM, SHAP, Plotly
- Environment: Jupyter Notebook, Google Colab
- Version Control: Git/GitHub

## 11. Team Members and contribution

| Name | Role | Contributions |
|---|---|---|
| T.Gomathi | Data Engineer | Data collection, cleaning And preprocessing |
| S.Kokila | Data Analyst | EDA, Visualization,and statistical analysis |
| M.Kamali | Project lead & developer | workflow Design, visualization set up |