Phase-1

Predicting Air Quality Levels Using Advanced Machine Learning Algorithms

Student Name: KAMALI.M

Register Number:620123106046

Institution: AVS ENGINEERING COLLEGE

Department: ECE

Date of submission: 30.04.2025

Problem Statement

- Reason: Rising air pollution levels affect human health, climate, and ecosystems.

- Importance: Predicting air quality helps in early warnings, health advisories, and policy planning for pollution control.

Objective

- Predict future air quality levels (e.g., AQI).

- Identify key pollutants and factors affecting air quality.

- Provide actionable insights to stakeholders (public, government, researchers).

Scope of the Project

- Focused on urban environments with available pollution data.

- Predict short-term (hourly/daily) air quality.

- Scalable to other regions with similar data structures.

Data source

- Government datasets: CPCB (India), EPA (USA), OpenAQ.

- Weather data: OpenWeatherMap, NOAA.

- Satellite data (optional): NASA EarthData.

High- Level Methodology

- Data Collection - Collect historical air quality and weather data.

- Data Cleaning - Handle missing values, remove outliers, normalize formats.

- Exploratory Data Analysis (EDA) - Understand pollutant trends, correlation with weather, seasonal variation.

- Feature Engineering - Create time-based features, lag variables, pollutant ratios.

- Model Building - Use algorithms like Random Forest, XGBoost, LSTM (for time series).

- Model Evaluation - Use RMSE, MAE, R^2 Score to assess performance.

- Visualization & Interpretation - Use heatmaps, line plots, and SHAP values for interpretability.

- Deployment - Build a simple web app/dashboard (Streamlit/Flask + cloud hosting).

Tools and Technologies

- Programming Language: Python

- IDE/Notebook: Jupyter Notebook, VS Code

- Libraries:

  - Data Handling: pandas, numpy

  - Visualization: matplotlib, seaborn, plotly

  - Modeling: scikit-learn, xgboost, keras (if using LSTM)

  - EDA: pandas-profiling, sweetviz

  - Deployment: Streamlit, Flask, Heroku, Render

Team Members and Roles

1.Gomathi.T project Manager

Gomathi.T was responsible for collecting and preparing the data needed for air quality prediction. I gathered pollution and weather data from trusted sources like CPCB and OpenWeatherMap. I cleaned the data to remove inconsistencies and performed Exploratory Data Analysis (EDA) to understand patterns and trends. This helped our team identify key variables and decide how to engineer features for better model performance.

2.Kamali.M : Machine Learning Engineer

Kamali .M main responsibility was to build and evaluate machine learning models for predicting air quality levels. I experimented with different algorithms like Random Forest, XGBoost, and LSTM. I fine-tuned the models using techniques such as cross-validation and grid search. I also used metrics like RMSE and R^2 to assess performance and selected the best-performing model for deployment.

3.Kokila.S : Software Developer

Kokila.S focused on the user interface and deployment of the model. I developed a simple and interactive web application using Streamlit to visualize the results and predictions. This made our model accessible and user-friendly for environmental researchers and policymakers. I also ensured that the system could be deployed.