Phase-1 Submission:
Air Quality Prediction Project

Student Name: K.SRITHIKA

Register Number: 621123205053

Institution: Idhaya Engineering College For Women

Department: B.Tech Information Technology

Date of Submission: 28-04-2025

1.Problem Statement

Air pollution is a critical issue in both urban and rural areas, causing serious health problems such as asthma, lung cancer, and cardiovascular diseases. According to the World Health Organization (WHO), air pollution leads to millions of premature deaths annually. Despite various governmental regulations, pollution levels often remain dangerously high. This project aims to leverage advanced machine learning techniques to predict air quality levels in advance. Timely and accurate predictions will enable authorities and individuals to take proactive measures to reduce exposure and improve public health.

2. Objectives of the Project

- To develop a predictive model for air quality levels using machine learning algorithms.

- To identify key pollutants contributing to poor air quality.

- To analyze seasonal and temporal patterns in air pollution data.

- To present findings through interactive visualizations and dashboards.

- To explore the feasibility of deploying the model as a web-based application.

3. Scope of the Project

Features to Analyze:

- Real-time and historical data of pollutant concentrations (PM2.5, PM10, CO, NO2, SO2, O3)

- Meteorological parameters (temperature, humidity, wind speed, pressure)

- Temporal features (time of day, day of week, season)

Limitations/Constraints:

- The scope is limited to selected cities or regions depending on dataset availability.

- The project will be a proof of concept and may not be deployed in a production environment.

- Only publicly available or open datasets will be used for this study.

4. Existing System

Currently, air quality monitoring relies heavily on physical monitoring stations operated by government and private environmental agencies. These stations collect data from sensors that track various pollutants and report AQI values on a daily or hourly basis. While this system provides accurate real-time data, it has limitations such as:

- High setup and maintenance costs

- Limited geographical coverage, especially in rural or remote areas

- Delayed response in issuing health advisories

- Lack of predictive capabilities for future air quality trends

5. Proposed System

Our proposed system addresses the limitations of the current setup by introducing a machine learning-based predictive model. The system will:

- Use historical and real-time data to predict future AQI levels

- Incorporate meteorological data to improve prediction accuracy

- Provide dynamic, location-specific predictions

- Offer a web-based interface for real-time user access

- Help authorities and individuals make informed decisions in advance

This approach aims to supplement existing systems with predictive insights that enhance preparedness and response.

6. Data Sources

- **Kaggle Datasets:** e.g., “Delhi Air Quality”, “Beijing Multi-Site Air-Quality Data”

- **UCI Machine Learning Repository:** e.g., “Air Quality Data Set” from Italy

- **OpenAQ API:** Provides real-time air quality data from cities around the world

- **Type of Data:** Public and mostly static (can be updated manually)

- **Formats:** CSV, JSON, or API endpoints with structured data

7. High-Level Methodology

- **Data Collection:** Acquire datasets from Kaggle or access real-time data through OpenAQ API.

- **Data Cleaning:** Detect and handle missing or null values, remove duplicate entries, convert timestamps, and standardize units across data columns.

- **Exploratory Data Analysis (EDA):** Use visualizations like heatmaps and line plots to understand trends and correlations.

- **Feature Engineering:** Create pollution categories, extract datetime components, normalize pollutant readings.

- **Model Building:** Use algorithms like Linear Regression, Decision Trees, Random Forest, Gradient Boosting, Neural Networks.

- **Model Evaluation:** Metrics like MAE, RMSE, R-squared, and cross-validation.

- **Visualization & Interpretation:** Charts for pollutant trends, AQI distribution, feature importance.

- **Deployment (Optional):** Use Streamlit or Gradio for a web interface.

8. Tools and Technologies

- Programming Language: Python

- Notebook/IDE: Google Colab, Jupyter Notebook

- Data Analysis Libraries: pandas, numpy

- Visualization Libraries: matplotlib, seaborn, plotly

- Machine Learning Libraries: scikit-learn, xgboost, lightgbm, keras (optional)

- Deployment Tools (Optional): Streamlit, Gradio, Flask

- Version Control: Git and GitHub

9. Team Members and Roles

1. K.SRITHIKA – Data Collection and Integration: Responsible for sourcing datasets, connecting APIs, and preparing the initial dataset for analysis.

2. A.SHIFANA – Data Cleaning and EDA: Cleans and preprocesses data, performs exploratory analysis, and generates initial insights.

3. P.SWETHA – Feature Engineering and Modeling: Works on feature extraction and selection; develops and trains machine learning models.

4. B.SUBASHINI – Evaluation and Optimization: Tunes hyperparameters, validates models, and documents performance metrics.

5. J.SHOBANA – Documentation and Presentation: Compiles reports, prepares visualizations, and handles presentation and optional deployment.