**Phase-1 Submission**

**Student Name:** AADHARSH K

**Register Number:** 712523205001

**Institution:** PPG INSTITUTE OF TECHNOLOGY

**Department:** INFORMATION TECHNOLOGY

**Date of Submission:** Insert Date

**1. Problem Statement**

*Air pollution is a growing concern worldwide, especially in urban regions. Its adverse effects on human health, ecosystems, and climate make accurate forecasting of air quality a vital task. This project aims to leverage advanced machine learning algorithms to predict air quality levels (e.g., AQI), enabling authorities and citizens to make proactive decisions.*

**2.Objectives of the Project**

* *Predict the Air Quality Index (AQI) based on environmental and meteorological parameters.*
* *Identify key contributors (e.g., pollutants) to air pollution.*
* *Generate insights to help government agencies, researchers, and the public.*
* *Optionally develop a web-based dashboard for real-time AQI forecasting.*

**3. Scope of the Project**

* ***Features:*** *PM2.5, PM10, NO2, SO2, CO, O3, temperature, humidity, wind speed.*
* ***Algorithms:*** *Regression models, Ensemble methods (e.g., Random Forest, XGBoost), possibly time-series models like LSTM.*
* ***Limitations:*** *Real-time data availability, data quality from sensors, and computational constraints.*

**4.Data Sources**

* ***Source:*** *Kaggle, UCI, OpenAQ API, Central Pollution Control Board (India).*
* ***Type:*** *Public datasets; both static (historical data) and dynamic (API-fed real-time data).*
* ***Format:*** *CSV, JSON.*
* ***Data source:*** *https://www.kaggle.com/datasets/waqi786/global-air-quality-dataset?resource=download*

**5. High-Level Methodology**

***a. Data Collection***

* *Download datasets from platforms like Kaggle/UCI.*
* *Use APIs (e.g., OpenAQ) for dynamic updates.*

***b. Data Cleaning***

* *Handle missing/null values.*
* *Remove duplicates and irrelevant features.*
* *Normalize pollutant concentrations.*

***c. Exploratory Data Analysis (EDA)***

* *Visualize pollutant trends, seasonal patterns, and location-wise variations using heatmaps, line plots, and box plots.*

***d. Feature Engineering***

* *Create pollutant ratios, categorize AQI levels, and add time-based features.*

***e. Model Building***

* *Test with Linear Regression, Random Forest, XGBoost, LSTM (if time-series).*
* *Optimize using GridSearchCV or RandomizedSearch.*

***f. Model Evaluation***

* *Use MAE, RMSE, and R² Score.*
* *Employ cross-validation.*

***g. Visualization & Interpretation***

* *Use graphs and dashboards to show pollutant levels, AQI trends, and predictions.*
* *Tools: Matplotlib, Seaborn, Plotly.*

***h. Deployment (Optional)***

* *Use Streamlit or Flask to create a web app for AQI prediction.*
* *Integrate with real-time API for live monitoring.*

**6.Tools and Technologies**

|  |  |
| --- | --- |
| ***Category*** | ***Tools & Technologies*** |
| *Programming Language* | *Python* |
| *IDE/Notebook* | *Jupyter Notebook / Google Colab* |
| *Libraries* | *pandas, numpy, seaborn, matplotlib, scikit-learn, XGBoost, TensorFlow (optional)* |
| *Deployment Tools* | *Deployment Tools* |

**7.Team Members and Roles**

|  |  |
| --- | --- |
| ***Team Member Name*** | ***Role/Responsibility*** |
| *Jeniliya* | *Data Collection, Model Building,* |
| *Aadharsh* | *Visualization, Deployment* |
| *Ashwinth* | *Evaluation, Presentation* |
| *Jeevanandan* | *Documentation, EDA* |
| *Ranjana sri* | *Reporting, , Feature Engineering* |