Phase-1

Student Name: DHIVYATHARSHINI D

Register Number: 712523243008

Institution: PPG INSTITUTE OF TECHNOLOGY

Department: 2 nd YEAR AI&DS

Date of Submission: 25.04.2025

1.Problem Statement

Air pollution poses significant health risks and environmental challenges globally. In urban areas like Coimbatore, increasing levels of pollutants such as PM2.5, PM10, NO_2, SO_2, CO, and O_3 contribute to respiratory diseases, environmental degradation, and reduced quality of life. Traditional methods of monitoring air quality often lack the predictive capabilities necessary for proactive measures. Therefore, developing a predictive model using advanced machine learning techniques can aid in forecasting air quality levels, enabling timely interventions and policy-making to mitigate adverse effects.

2.Objectives of the Project

- Develop a robust machine learning model to predict Air Quality Index (AQI) based on historical pollutant data.
- Identify key pollutants contributing to poor air quality in the region.
- Provide actionable insights for environmental agencies to implement preventive measures.
- Enhance public awareness by visualizing air quality trends and forecasts.

3.Scope of the Project

Inclusions:

- Analysis of pollutants: PM2.5, PM10, NO_2, SO_2, CO, O_3.
- Utilization of machine learning algorithms such as Random Forest, Support Vector Machines, and Neural Networks.
- Development of a user-friendly interface for visualizing predictions

Exclusions:

- Real-time data acquisition and processing.
- Deployment on mobile platforms.

Constraints:

- Dependence on the availability and quality of historical air quality data.
- Computational limitations for training complex models.

4.Data Sources

- **Source:** ● Public datasets from Kaggle related to customer support (e.g., "Air pollution dataset")

- **Type:** Publicly available datasets.

- **Nature:** Static datasets comprising historical air quality measurements.

- **Features:** Concentrations of various pollutants, meteorological data, timestamps, and geographical information.

5.High-Level Methodology

- **Data Collection** –To predict air quality levels, we obtain data from:

⁕Historical air quality datasets from sources such as Kaggle.

- **Data Cleaning** –⁕Handling missing values in pollutant concentrations using interpolation, mean imputation, or dropping depending on severity.

⁕Removing duplicates to avoid skewed analysis.

⁕Standardizing units and formats across datasets

⁕Time alignment for hourly or daily aggregation, especially if combining multiple data sources.

- **Exploratory Data Analysis (EDA)** – To understand air quality behavior:

⁕Time-series visualizations to observe pollutant trends over time.

⁎Heatmaps and correlation matrices to examine relationships between pollutants and weather variables.
⁎Boxplots to identify pollution outliers across regions or times.
⁎Geospatial plots using libraries like Folium or Plotly to visualize regional variations in air quality.
⁎Pollutant distribution histograms to detect skewness or anomalies.

- **Feature Engineering** –Beyond the pollution content features:

⁎Temporal Features: The Date column was converted to datetime format, allowing extraction of features like month, day, and year if needed JETIR

⁎Categorical Encoding: Categorical variables, such as City, can be encoded using techniques like Label Encoding or One-Hot Encoding to be utilized in machine learning models.

- .**Model Building** – A Random Forest Classifier was employed to predict air quality categories:

⁎Model Initialization: RandomForestClassifier was instantiated with 20 estimators and a fixed random state for reproducibility.

⁎Training: The model was trained on the training dataset (X_train, y_train).

⁎Prediction: Predictions were made on both training (X_train) and testing (X_test) datasets.

- **Model Evaluation** — The model's performance was assessed using:

⁎Confusion Matrix: To visualize the performance of the classification model and identify misclassifications.

⁎Accuracy Score: Calculated to determine the percentage of correct predictions made by the model.

⁎Classification Report: Provides precision, recall, f1-score, and support for each class, offering a comprehensive view of model performance.

- **Visualization & Interpretation** — To effectively present key findings, insights, and predictions from your air quality analysis, consider the following visualization strategies:

⁎Time-Series Plots:Purpose: Illustrate trends and patterns over time for pollutants and AQI levels.

⁎Bar Charts for AQI Distribution:Purpose: Compare the distribution of air quality categories across different cities.

⁎Confusion Matrix and Accuracy Score:
Purpose: Evaluate the performance of your classification model.

⁎Feature Importance Visualization:Purpose: Identify which features most significantly influence the model's predictions.

- **Deployment** –Colab Notebooks

  ⁎Overview: For sharing purposes, you can use the Colab notebook itself as a report.

  ⁎ Documentation: Ensure your notebook is well-documented with markdown cells explaining each step.

  ⁎Sharing: Share the notebook link with stakeholders, allowing them to view or run the notebook as needed.

6.Tools and Technologies

- **Programming Language** –Python.
- **Notebook/IDE** – *Visual studio,Google Colab,Jupyter Notebook.*
- **Libraries** –*Numpy,Pandas,Seaborn,Matplotlib,Streamlit,Scikit Learn,SK Learn.*
- **Optional Tools for Deployment** – *Flask, Streamlit cloud*

7.Team Members and Roles

Team Member	Role	Responsibilities
Banumathi .v	Team Lead & Coordinator	Oversees the project, coordinates between teams, and ensures timely delivery.

Dhivyatharshini.D	Data analyst	Responsible for data collection, cleaning, and preprocessing.
Sakina .E	Frontend developer	Designs and implements the user interface
SKamala kaviya J	Machine Learning developer	Designs, builds, and tunes predictive models (Random Forest , XGBoost).
Yogadharshini R	EDA Specialist	Performs exploratory data analysis (EDA) using visualization tools.