Phase-1 Submission

Student Name: Abinaya B

Register Number:410723104003

Institution: Dhanalakshmi College Of Engineering

Department: Computer Science and Engineering

Date of Submission: 30/04/2025

1.Problem Statement

Predicting air quality levels using advanced Machine Learningalgorithms for Environmental Insights.

This project aims to accurately predict air quality levels by leveraging advanced machine learning models on environmental and meteorological data. Key air pollutants such as PM2.5, PM10, NO2, CO, SO2, and O3 are analyzed alongside weather parameters like temperature, humidity, wind speed, and atmospheric pressure. The model is trained to classify air quality into standard AQI categories (e.g., Good, Moderate, Unhealthy) or predict exact pollutant concentrations. Advanced techniques such as ensemble learning (e.g., Random Forest, XGBoost) and deep learning models (e.g., LSTM, CNN) are employed for improved prediction performance.

2.Objectives of the Project

The objective of air quality prediction using machine learning is to accuratelyforecast pollution levels (like PM2.5, PM10, NO2, CO, etc.) in the air based onvarious environmental, weather, and traffic-related features. Estimate future levels of pollutants such as PM2.5, PM10, NO2, SO2, CO, andO3.This helps in early warning systems for public health. Use featureimportance to identify which factors (e.g., temperature, humidity, wind speed, traffic levels) most affect air quality. Provide forecasts to the public via apps or websites so they can take precautions.

3.Scope of the Project

This project focuses on building a predictive system that forecasts air pollutant levels or the Air Quality Index (AQI) using historical and real-time data. The scope includes data collection, preprocessing, model training, evaluation, and deployment. The goal is to assist environmental agencies, policymakers, and the public in making informed decisions to reduce health risks and environmental impact.

4.Data Sources

- Dataset used in this project from the source- KAGGLE. Kaggle is a public and it is dynamic platform.

- Data Source: air quality prediction data

5.High-Level Methodology

Predict specific air pollutants (e.g., PM2.5, NO2, O3).
Identify the prediction horizon (e.g., hourly, daily).

- **Data Collection**–Gathering datasets from public sources like Kaggle, UCI, or government APIs.Handling missing data, normalization, and encoding.

- **Data Cleaning** –My Dataset is Complete with no missing values andchecked for the missing values. And there is no Duplicate values in the dataset.

- **Exploratory Data Analysis (EDA)** –

To uncover patterns, trends, and relationships in the NR-quality dataset, we performed the following EDA steps:

Data Inspection: Reviewed the dataset structure, identified data types,and checked for missing values and anomalies.

Time Series Decomposition: Explored pollutant trends over time using line plots and seasonal decomposition.

Distribution Analysis: Used histograms and boxplots to visualize pollutant distributions and detect outliers.

Correlation Matrix: Plotted a heatmap to identify relationships between different pollutants and meteorological features.

Feature Engineering –To improve model performance, we created and transformed features from the raw dataset.

- **Model Building** –To predict air quality levels or pollutant concentrations accurately, we plan to experiment with the following machine learning models:

1.Logistic Regression:

Useful when predicting continuous values (e.g., pollutant concentration) or classifying AQI levels.

2. Random Forest

A robust ensemble method that handles non-linear relationships well.

3.XGBoost

Known for high performance in tabular datasets .Can efficiently handle missing data and works well with feature-richenvironmental datasets.

4. Support Vector Machine (SVM)

Effective in high-dimensional spaces and with smaller datasets.
Useful for binary or multi-class AQI level classification.

- **Model Evaluation** –To assess the performance and reliability of the models used for predicting air quality levels or pollutant concentrations, the following evaluation metrics will be employed:

For Classification Tasks (predicting AQI categories):

Accuracy: Measures the overall correctness of the model.

Precision, Recall, and F1-Score: Useful for imbalanced AQI categories to evaluate the model's ability to identify specific classes.

Confusion Matrix: Visual tool to observe true vs predicted class performance.

.

- **Visualization & Interpretation** – To better understand the data, model behavior, and results, we employed the following visualizations:

  Histograms & Boxplots: To understand the distribution and identify outliers for each pollutant.

  Time Series Decomposition: To break down seasonal, trend, and residual components of air quality levels.

  Scatter Plot: A scatter plot is a graph that shows the relationship between two variables. Each point on the graph represents one observation.

  Pie Chart: A pie chart is a circular chart divided into slices to show proportions or percentages. Each slice represents a category's contribution to the whole.

6.Tools and Technologies

Tool used here is google collab and programming language is python and the details about the tools and technologies are explained below:

- **Programming Language** –Python is a high-level, interpreted, and general-purpose programming language known for its simplicity, readability, and versatility. It was created by Guido van Rossum and first released in 1991. Python supports multiple programming paradigms, including procedural, object-oriented, and functional programming.Its clean syntax and extensive standard libraries make it ideal for beginners and professionals alike, and it is widely used in fields such as web development, data science, machine learning, automation, and scientific computing.

- **Notebook/IDE** –Google Collab , jupyter notebook

- **Libraries** –Key libraries used in data processing, visualization, and modelling are pandas, numpy , seaborn , matplotlib , scikit-learn.

7.Team Members, Roles and Responsibility

NAME	ROLES	RESPONSIBILTY
Abinaya B	Team Leader	Model building and evaluation
Mohana priya K	Team Member	EDA and Feature Engineering
Narmadha D	Team Member	Data Collection and Preprocessing
Hema Kanitha S	Team Member	Data Visualisation and Interpretation