## 

## **Phase-2 Submission Template – Data Analytics**

**Student Name:** BALAJI.K  
**Register Number:** 621523121009  
**Institution:**Mahendra College Of Engineering  
**Department:** B.E Biomedical Engineering  
**Date of Submission:** [Insert Date]  
**GitHub Repository Link:** [Update the project source code repository link]

### **1. Problem Statement**

### Air pollution negatively impacts human health and ecosystems. Predicting air quality levels accurately can help in issuing warnings and planning preventive actions.The project aims to build a machine learning model to predict the Air Quality Index (AQI) based on environmental factors like pollutant levels and weather conditions.

### **2. Project Objectives**

* Collect and prepare air quality and weather data.
* Analyze data patterns and relationships.
* Train machine learning models for accurate AQI prediction.
* Compare different algorithms and select the best model.
* Visualize results and provide actionable insights.

**3. Flowchart of the Project Workflow**

Start

↓

Data Collection

↓

Data Preprocessing

↓

Exploratory Data Analysis (EDA)

↓

Feature Engineering

↓

Model Training and Evaluation

↓

Prediction

↓

Visualization and Insights

↓

End

### **4. Data Description**

### **Example Dataset:**

**Columns**:

Date, PM2.5, PM10, NO2, SO2, CO, O3, Temperature, Humidity, WindSpeed, AQI

**Source**: Kaggle public datasets, UCI Repository, CPCB India, etc.

Sample of data (first 5 rows):

|  |  |  |
| --- | --- | --- |
| Da PM PM N  te 2 10 O  5 2 | S C O Te  O O 3 m  2 pe  ra  tu  re | Hu Wi A  mi nd Ql  di Sp  ty ee  d |
| 20 85 16 45  23 5  -0  1-  01 | 10 0. 25 22  8 | 60 2. 15  5 2 |
| 20 92 18 48  23 0  -0  1-  02 | 11 0. 28 23  7 | 65 2. 16  0 0 |
| **. . . . . . . . . . . .** | **. . . . . . . . . . . .** | **. . . . . . . . .** |

### **5. Data Preprocessing**

### **Libraries**

**import pandas as pd**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn.preprocessing import StandardScaler**

**# Load Data**

**df = pd.read\_csv('air\_quality\_data.csv')**

**# Check for missing values**

**print(df.isnull().sum())**

**# Fill missing values (simple approach)**

**df.fillna(method='ffill', inplace=True)**

**# Feature and Target Split**

**X = df[['PM2.5', 'PM10', 'NO2', 'SO2', 'CO', 'O3', 'Temperature', 'Humidity', 'WindSpeed']]**

**y = df['AQI']**

**# Scaling features**

**scaler = StandardScaler()**

**X\_scaled = scaler.fit\_transform(X)**

**# Train-Test Split**

**X\_train, X\_test, y\_train, y\_test = train\_test\_split(X\_scaled, y, test\_size=0.2, random\_state=42)**

### **6. Exploratory Data Analysis (EDA)**

import matplotlib.pyplot as plt

import seaborn as sns

# Histogram of AQI

plt.figure(figsize=(8,6))

sns.histplot(df['AQI'], bins=30, kde=True)

plt.title('Distribution of AQI')

plt.show()

# Correlation Heatmap

plt.figure(figsize=(10,8))

sns.heatmap(df.corr(), annot=True, cmap='coolwarm')

plt.title('Correlation between Variables')

plt.show()

# Time Series Trend (optional)

df['Date'] = pd.to\_datetime(df['Date'])

plt.figure(figsize=(12,6))

plt.plot(df['Date'], df['AQI'])

plt.title('AQI Over Time')

plt.xlabel('Date')

plt.ylabel('AQI')

plt.show()

**Insights Example:**

* High correlation between PM2.5, PM10 and AQI.
* Seasonal trends: AQI rises in winter due to lower dispersion.

### **7. Tools and Technologies Used**

**Linear Regression:**

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error, r2\_score

# Model

lr = LinearRegression()

lr.fit(X\_train, y\_train)

# Prediction

y\_pred = lr.predict(X\_test)

# Evaluation

print('MSE:', mean\_squared\_error(y\_test, y\_pred))

print('R² Score:', r2\_score(y\_test, y\_pred))

### **8. Team Members and Contributions**

[List the members and their responsibilities clearly.]

| **Name** | **Contribution** |
| --- | --- |
| Balaji.K | Task 1 |
| Jeevanandham.A | Task 2 |
| Dhanush.B | Task 3 |
| Vignesh.M |  |