Infosys Springboard

.

Batch-3

Project title

Air quality forecasting project

Submitted by –Payal Rawat

Mentor - **Ms. Umme Asma S**

PROJECT STATEMENT

Air pollution is a growing concern in urban environments, directly impacting human health and Environmental quality. Accurate forecasting of air quality can help authorities take preventive Measures, issue warnings, and inform the public.

Air pollution is a growing issue due to:

Vehicular Emissions

Release of NO_2, CO, PM from cars, buses, and trucks.

Industrial Activities

Factories and power plants emit smoke, chemicals, SO_2, and PM.

Waste Burning & Landfills

Burning garbage releases toxic gases like methane and dioxins.

Burning of Fossil Fuels

Coal, diesel, petrol, and biomass burning release toxic gases.

Key Pollutants

The harmful pollutants monitored include:

- **PM2.5 - PM2.5 (Fine Particulate Matter)**

   Sources: Vehicle exhaust, industrial emissions, stubble burning, construction dust, burning of coal/wood.

- **PM10 (Coarse Particulate Matter)**

. **Sources**: Construction sites, road dust, mining activities, agricultural fields, pollen.

- **Nitrogen Dioxide (NO_2)**
  Sources: Vehicle emissions, power plants, industrial combustion.
- **Sulphur Dioxide (SO_2)**
  Sources: Burning of coal and fossil fuels, power plants, refineries, industrial smelting.
- **Carbon Monoxide (CO)**
  Sources: Vehicle exhaust, gas stoves, biomass burning, incomplete combustion of fuel.
- **Ozone (O_3) – Ground-Level**
  Sources: Formed when NOx and VOCs react in sunlight (secondary pollutant), mainly from vehicles and industries.
- **Ammonia (NH_3)**
  Sources: Fertilizers, livestock waste, agricultural activities.

IMPACTS

Health Impacts

- **Respiratory Diseases**
  Air pollution significantly increases the risk of **respiratory diseases**, such as asthma and chronic bronchitis, as pollutants irritate the airways and lungs, leading to long-term health issues.

- **Cardiovascular Issues**
  Exposure to poor air quality can lead to **cardiovascular issues**, including heart attacks and strokes, as harmful particles enter the bloodstream and affect heart function and overall health.

- **Other Ailments**
  In addition to respiratory and cardiovascular problems, air pollution contributes to various **other ailments** like allergies and neurological disorders, impacting quality of life and increasing healthcare costs.

Environmental Impacts

- **Crop Damage**
  Air pollution can lead to significant declines in crop yields, affecting food supply and economic stability for farming communities around the world.
- **Climate Change**
  The contribution of air pollutants to climate change intensifies global warming, resulting in altered weather patterns and increased frequency of extreme weather events.
- **Ecosystem Degradation**
  Pollutants disrupt ecosystems, affecting biodiversity, soil health, and water quality, which can lead to long-lasting negative impacts on wildlife and natural habitats.

Project Objectives

This project aims to build a time series forecasting system that predicts **Air Quality Index (AQI)** and pollutant levels using historical pollution data. The goal is to empower environmental agencies, city planners, and the public with timely insights for healthier decision-making

Importance of Air Quality Forecasting

Predictive Capabilities

- Predict pollutant levels in advance
- Enable authorities to implement preventive measures

Preventive Measures Include:

- Restricting industrial emissions
- Controlling traffic
- Advising the use of masks and air purifiers
- Providing timely warnings for outdoor activity modifications

Additional Benefits:

- Supports policy-making and urban planning
- Assists environmental research
- Helps cities develop long-term strategies for sustainable living

Expected Project Outcomes

✅ Time series forecasting model to predict AQI and key pollutants

📊 Visual dashboard showing historical trends and future predictions

⚠ Alert system when predicted AQI crosses safety thresholds

📈 Analysis capabilities for pollutant contributions and seasonal trends

🔧 Admin functionality to upload datasets and retrain models

The internship is divided into four milestone 2 weeks of each milestone

<u>Milestone 1</u>:

<u>data preprocessing and eda</u>

Duration: Weeks 1-2

<u>DATA SOURCES AND AGENCIES</u>

- **<u>CPCB (Central Pollution Control Board)</u>**
- **<u>Type</u>**: Government agency under Ministry of Environment, Forest & Climate Change, India
- **<u>Role:</u>** National monitoring and regulation body Data Source: Continuous Ambient Air Quality Monitoring Stations (CAAQMS)
- **<u>SAFAR (System of Air Quality and Weather Forecasting)</u>**
  Type: IITM Pune project, funded by Ministry of Earth Sciences
  Role: Real-time monitoring and 1-3 day forecasts for major cities
  Data Source: Specialized SAFAR stations with meteorological data

- **<u>OpenAQ</u>**
  Type: Non-profit, open-data platform (global)
  Role: Aggregates worldwide air quality data
  Data Source: APIs from official sources including CPCB and EPA4

- **<u>World Air Quality Index (WAQI/AQICN)</u>**
  Type: International air quality visualization initiative
  Role: Real-time AQI maps from 130+ countries
  Data Source: Official monitoring networks with unified global scale

<u>Exploratory Data Analysis (EDA)</u>

EDA is the first and most important step in data analysis, involving:

Key EDA Activities

1 **Load data and Understand structure**
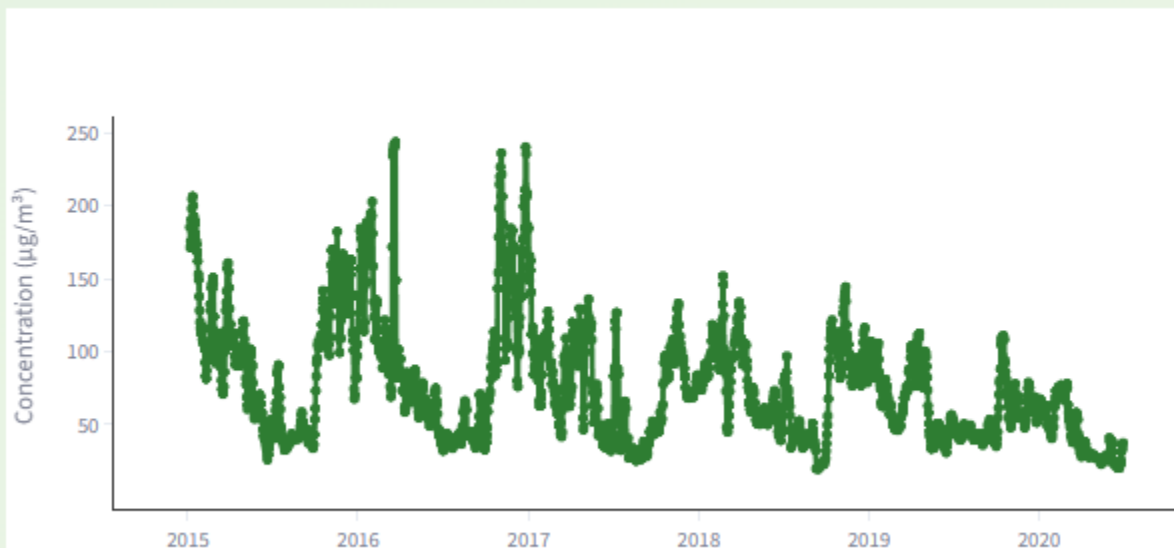
2 **handle missing value**

3 **Summary statistics**

4 **Detect outliers**

5 **time series pattern**

6 **Correlation analysis**

<u>Visualization Techniques</u>

1. Time Series Plots

Purpose:

PM2.5 Time Series

- **Display pollutant concentration changes over time**
- **Shows pollution patterns peaks, and drops**
- **Detects daily cycles**
- **Supports forecasting model development**

(ARIMA, Prophet, LSTM)

2. Pollutant Correlation Analysis

Purpose: Shows how pollutants influence each other

Pollutant Correlations

- Identifies related pollutants
- Improves model accuracy
- Helps select useful prediction features

3. Statistical Summary

Key Metrics: Mean, median, max, min, standard deviation, data points

Statistical Summary

Mean (µg/m³)	Median (µg/m³)
76.9	60.4

Max (µg/m³)	Min (µg/m³)
894.8	3.0

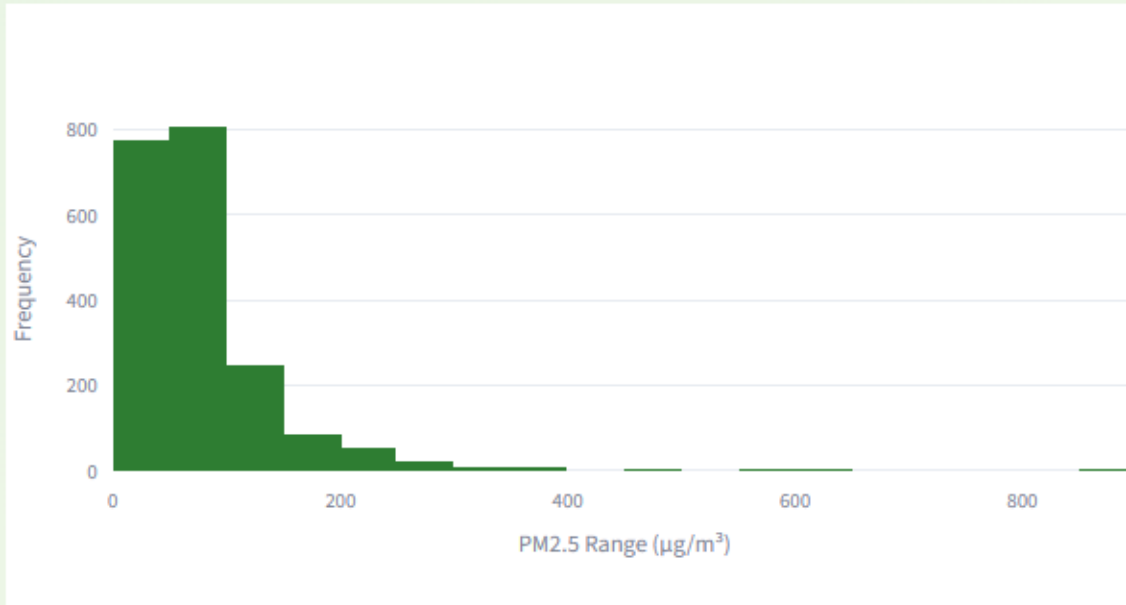| Std Dev	Data Points
60.4	2009

- Provides quick snapshot of central tendency

- Shows extreme pollution events
- Guides model selection and tuning

4. Distribution Analysis (Histograms)

Purpose: Shows frequency of pollutant values in different ranges

- Understands pollution level patterns
- Detects outliers
- Checks data skewness for model improvement

MILESTONE 2:

Model training and evaluation

Duration: Weeks 3-4

Machine Learning Overview

Machine Learning (ML) enables computers to learn from data and improve performance without explicit programming, identifying patterns to make predictions.

ML Types for Air Quality Forecasting

1. **Supervised Learning**
   - Definition: Trained on labeled data with input-output pairs
   - Goal: Learn mapping for predictions on new data
2. **Unsupervised Learning**
   - Definition: Works with unlabeled data to find patterns
   - Goal: Identify groupings without predefined labels
3. **Reinforcement Learning**
   - Definition: Learns through environment interaction and rewards
   - Goal: Develop strategy maximizing cumulative reward
   -

Forecasting Models

1. **ARIMA (AutoRegressive Integrated Moving Average)**
   - Type: Statistical time series model
   - Best for: Short-term, linear, stationary data
   - AQI Application: Captures trends and gradual pollutant changes
   -
2. LSTM (Long Short-Term Memory)
   - Type: Deep learning (Recurrent Neural Network)
   - Best for: Sequential data with complex patterns
   - AQI Application:
     - Learns long-term pollution effects
     - Handles multiple features (PM2.5, temperature, wind)
     - Manages non-linear relationships

3. **XGBoost (Extreme Gradient Boosting)**
   - Type: Machine learning ensemble (decision trees)
   - Best for: Tabular data with many features
   - AQI Application:
   - Includes external variables (weather, holidays, traffic)
   - Fast and accurate for structured data
4. **Prophet (by Meta)**
   Type: Additive time series model
   Best for: Data with seasonal patterns
   AQI Application: Captures daily/weekly/yearly seasonality Simple and interpretable
   Works with missing data

Model Persistence

Pickle

Purpose: Save and load ML models without retraining Benefits: Time-saving, easy deployment

- Faster and more memory-efficient than pickle for big models.
- With the help of train model visualize data

Joblib

- Purpose: Efficient saving/loading of large models
- Advantages: Faster and more memory-efficient than pickle

Model Evaluation Metrics

1. RMSE (Root Mean Squared Error)
   Definition: Average magnitude of errors with emphasis on large errors Use:
   Evaluates prediction closeness, penalizes extreme pollution spikes
2. MAE (Mean Absolute Error)
   Definition: Average absolute difference between predicted and actual values
   Use: Simple accuracy measure, less sensitive to outliers
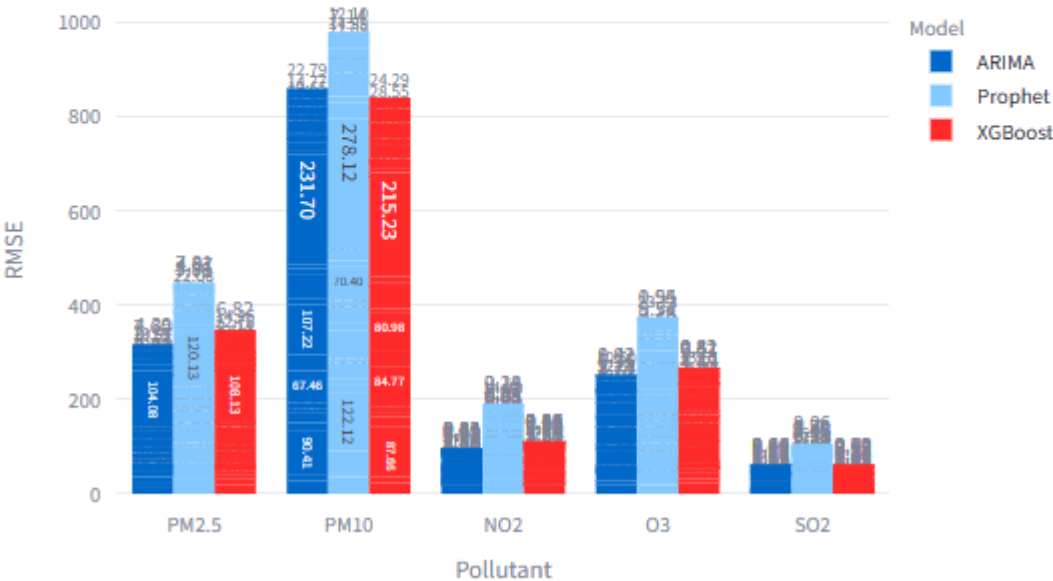
Visualization Components

1.Model Comparison Chart

Type: Grouped bar chart

Model Performance

⦿ RMSE ◯ MAE

Model Performance (RMSE)

Purpose: Compare RMSE/MAE across models for each pollutant
Insight: Identifies best-performing models
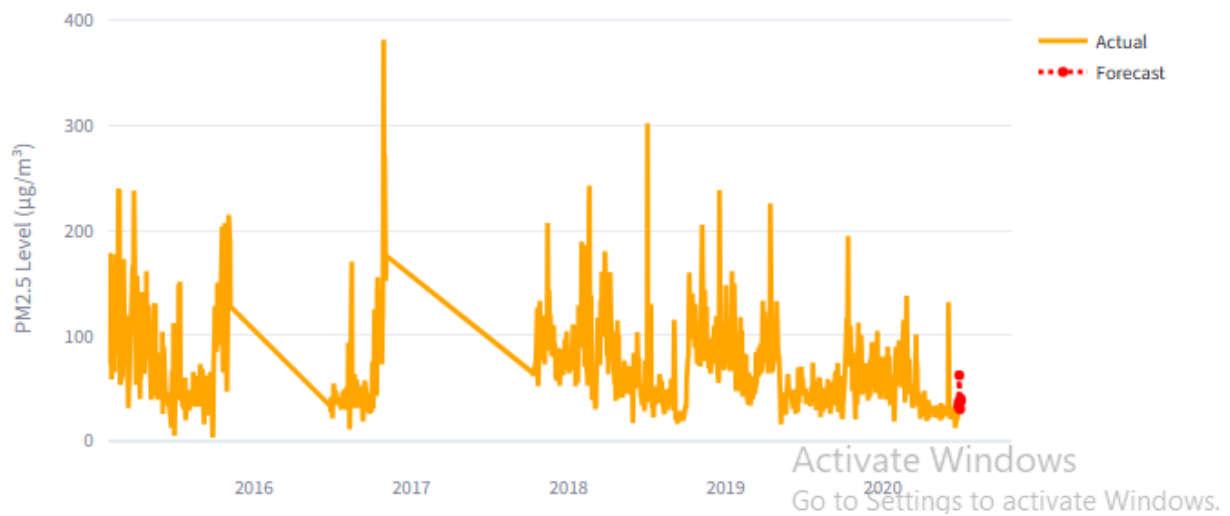
2.Forecast Graph

Type: Time series line chart

PM2.5 Forecast

City

Ahmedabad ⌄

Pollutant

PM2.5 ⌄

Purpose: Visualize actual vs predicted values with confidence intervals
Insight: Shows model trend-following capability

3. Best Model Summary Table
   Type: Summary table

Best Model by Pollutant

|	City	Pollutant	Best_Model	RMSE
0	Ahmedabad	PM2.5	ARIMA	16.7354
1	Ahmedabad	PM10	XGBoost	22.8277
2	Ahmedabad	NO2	Prophet	17.0911
3	Ahmedabad	O3	ARIMA	18.2229
4	Ahmedabad	SO2	ARIMA	1.8653
5	Chennai	PM2.5	Prophet	5.3932
6	Chennai	PM10	XGBoost	12.6622
7	Chennai	NO2	ARIMA	1.7933
8	Chennai	O3	XGBoost	8.8046
9	Chennai	SO2	XGBoost	1.1915

Purpose: Lists optimal model for each pollutant with performance scores
Insight: Guides model selection and deployment

4. Forecast Accuracy Chart
   Type: Multi-line chart

Forecast Accuracy by Horizon 🔗
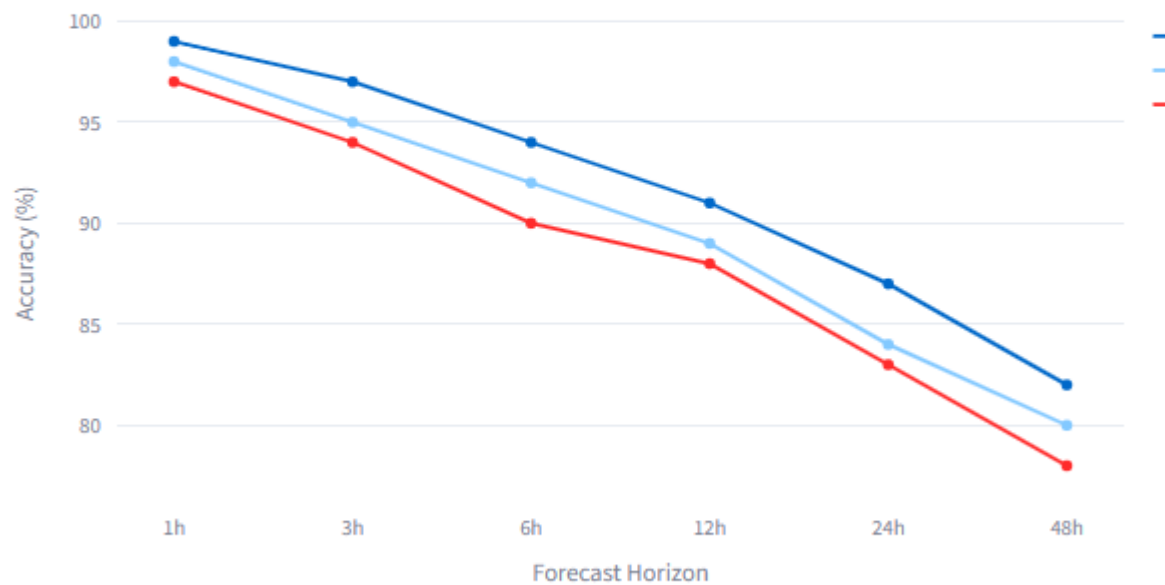
Purpose: Shows accuracy changes over forecast horizons

Insight: Reveals trade-off between prediction length and accuracy

Milestone 3: Alert Logic and Visualization

Duration: Weeks 5-6

Dashboard Components

1. Current Air Quality Gauge

Reason:This shows real-time overall air quality at a specific location.

AQI is a combined score calculated from pollutants (PM2.5, PM10, NO2, O3, etc.).

People quickly understand:

Is the air safe?
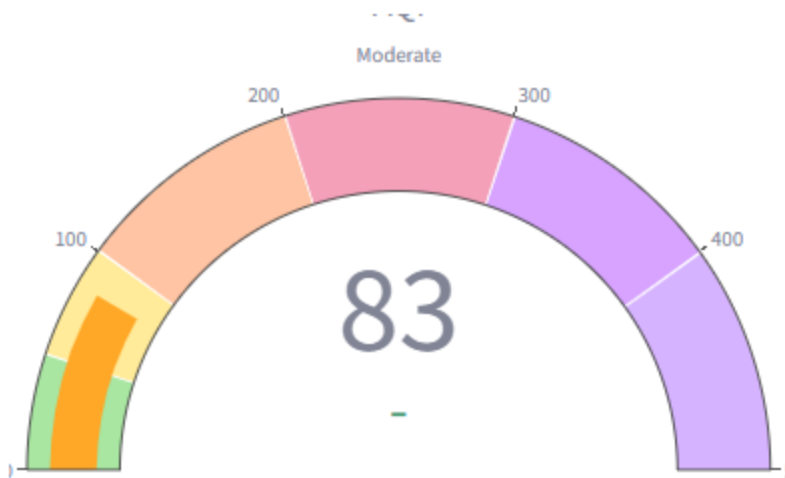
Should sensitive groups avoid outdoor activity?

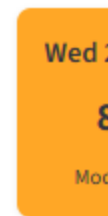Features:

Circular AQI gauge (e.g., AQI = 78)

Air Quality Alert System

Milestone 3: Working Application — 7-day forecasts, AQI, and pollutant trend

Current Air Quality

7-Da

Moderate

200   300

100   400

83

-

Wed 2

8

Mod

Color-coded levels: Green → Yellow → Orange → Red

Status indicator ("Moderate")

Purpose:

 Real-time air quality assessment

Quick safety evaluation

Public health guidance

2. 7-Day Forecast Cards

Features:

Daily AQI predictions (Monday → Sunday)

 Color-coded severity levels

 Trend visualization

Purpose:

Activity planning assistance

Pollution trend awareness

Proactive health measures

3.Pollutant Concentration Chart

Features: Hourly concentration tracking Multiple pollutant lines (PM2.5, PM10, O_3) WHO safety limit indicators

Purpose: Detailed pollutant analysis Limit exceedance identification Comprehensive air quality understanding

4.Active Alerts Panel

Features:A notification box showing real-time alerts based on predicted pollutant levels and AQI thresholds.
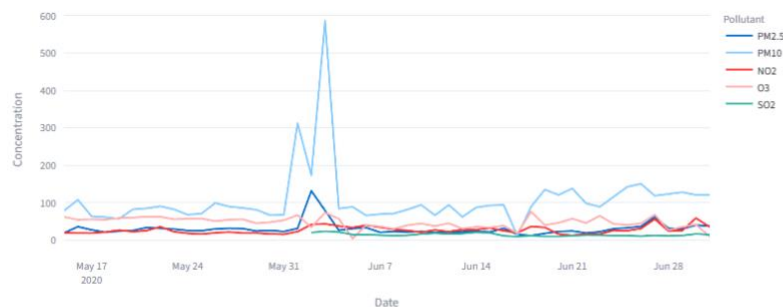
Purpose :

Warns users when air quality may become harmful or risky

Helps health-sensitive communities prepare in advance.

Pollutant Analysis & Alerts

Pollutant Concentrations (48 hrs)

Active Alerts (Next 7 Days)

Unhealthy for Sensitive — Mon, 22 Jun
Predicted AQI: 121

Unhealthy for Sensitive — Tue, 23 Jun
Predicted AQI: 121

Unhealthy for Sensitive — Thu, 25 Jun
Predicted AQI: 121

Unhealthy for Sensitive — Fri, 26 Jun
Predicted AQI: 121

Unhealthy for Sensitive — Sat, 27 Jun
Predicted AQI: 121

Unhealthy for Sensitive — Mon, 29 Jun
Predicted AQI: 121

Unhealthy for Sensitive — Tue, 30 Jun
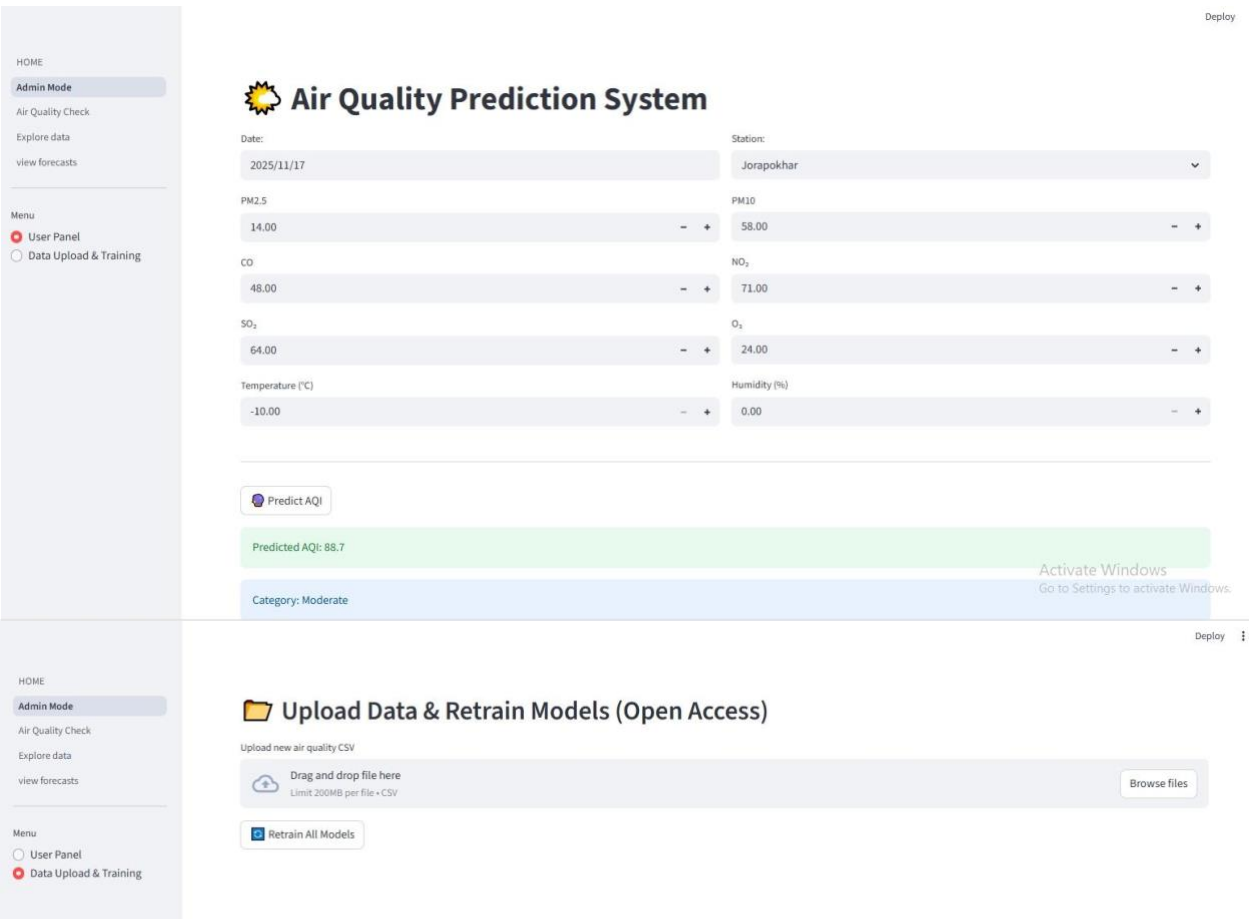Predicted AQI: 121

Milestone 4: Streamlit Application Development

Duration: Weeks 7-8

User Interface Features

Core Functionality

- Interactive dashboard with real-time data visualization
- CSV upload capability for new datasets
- Model retraining interface for updated forecasting
- Alert system with customizable thresholds

🌊 Air Quality Monitoring Dashboard

Explore, Analyze, and Forecast Air Quality Data

📊 Explore Data

View city-wise pollutant levels, summary, and statistics.

Go → Explore Data

📈 View Forecasts

Check time-series predictions for PM2.5 and other pollutants.

Go → View Forecasts

🌍 Air Quality Check

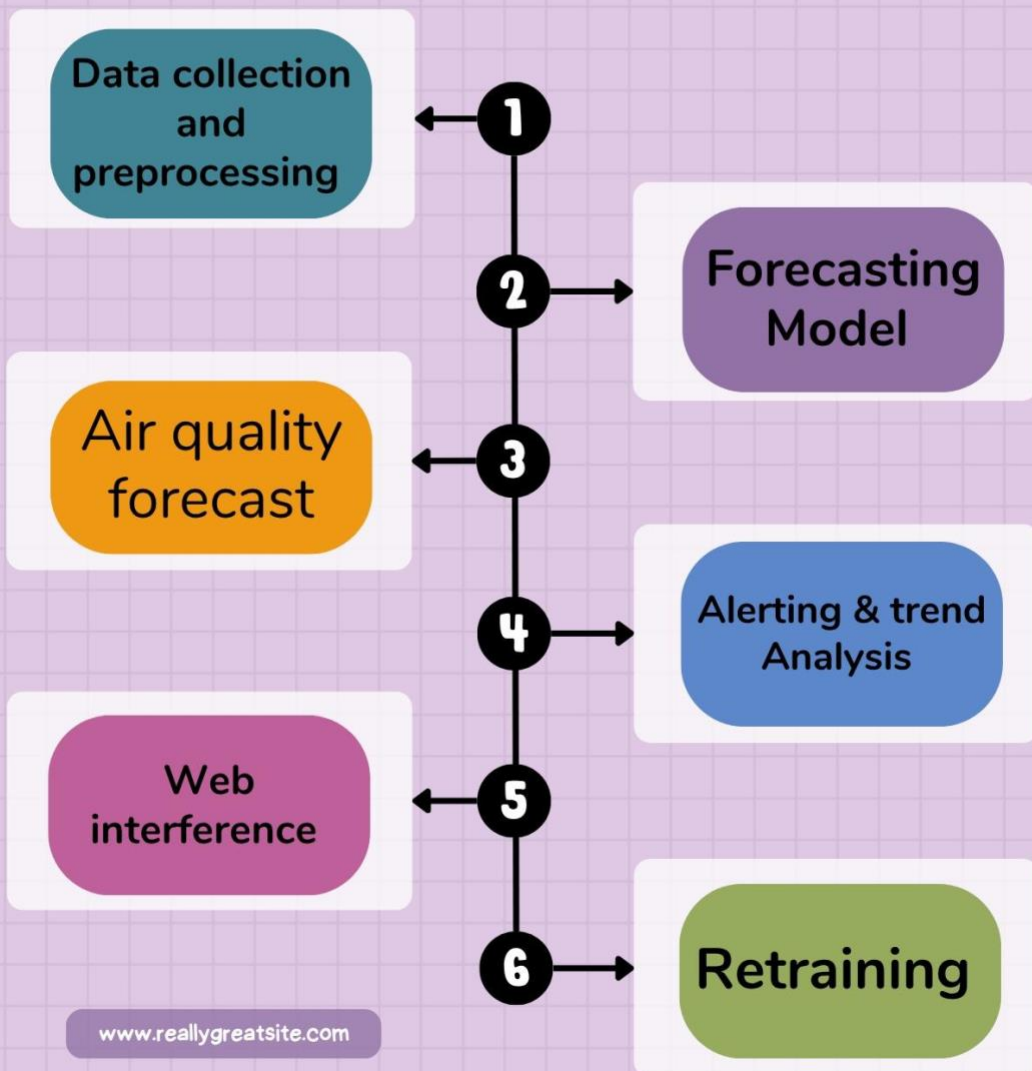AQI insights and pollution category classification.

Go → Check AQI

🔧 Retrain Model

check AQI,upload files.

Go → Retrain Model

Developed for Air Quality Analysis Project

TECHNICAL ARCHITECTURE

Data collection and preprocessing — 1

2 → Forecasting Model

Air quality forecast — 3

4 → Alerting & trend Analysis

Web interference — 5

6 → Retraining

PROJECT IMPACT AND BENEFITS

ENVIRONMENTAL AGENCIES

- Proactive pollution management
- Evidence-based policy decisions
- Resource optimization

CITY PLANNERS

- Urban development guidance
- Traffic management insights
- Industrial zone planning

FOR THE PUBLIC

- Health protection through early warnings
- Activity planning assistance
- Environmental awareness enhancement

FOR RESEARCHERS

- Data-driven environmental studies
- Pattern recognition in pollution trends
- Climate change impact assessment