Forecasting Air Quality Using Historical Pollution Data

A Data-Driven Approach to Environmental Health

> “Forecasting Tomorrow’s Air, Protecting Today’s Lives”

Name: **Neha Vithob Nayak**

Department: **Computer Science & Engineering**

Year: **2025**

Guide: **Asma**

Project Overview: The Predictive Power of Data

Air pollution stands as one of the most critical public health and environmental challenges in modern urban environments. This project addresses this challenge head-on by developing an advanced system designed to **predict future Air Quality Index (AQI)**.

Core Objective

To predict future AQI using comprehensive historical pollution data.

Key Deliverable

An interactive dashboard visualizing these predictions for easy access.

Impact

Empowering citizens, policymakers, and environmental agencies with real-time, data-driven insights.

Defining the Problem and Project Purpose

Our mission is driven by the urgent need for proactive air quality management.

The Problem

- Air pollution is a major health crisis, causing severe respiratory and cardiovascular illnesses and impacting millions annually across the globe.
- A significant gap exists in the market: a lack of accessible, real-time AQI forecasting systems readily available to the general public.
- Decision-makers need predictive tools, not just current readings, to implement timely preventive strategies.

Our Solution

This project directly addresses this gap by creating a robust **machine learning-based forecasting platform** that accurately predicts AQI trends and specific pollutant concentration levels in advance.

This allows for better preparedness and improved air quality awareness across communities.

Key Outcomes and Deliverables

The successful execution of this project yields several critical functional components designed to deliver comprehensive air quality intelligence.

1

Time Series Forecasting Model

Predictive model for AQI and major pollutants: PM2.5, PM10, NO_2, SO_2, CO, and O_3.

2

Interactive Visual Dashboard

A user-friendly dashboard for visualizing historical data and future air quality trends over various time horizons.

3

Automated Alerts & Notifications

Systematic alerts delivered when forecasted AQI levels are projected to cross predefined unsafe thresholds.

4

Analysis Tools

Functionality to analyze and report on seasonal, regional, and temporal variations in pollutant concentrations.

Tools and Technologies Utilized

Our forecasting platform is built upon a modern, flexible technology stack, leveraging powerful data science libraries and scalable deployment tools.

Category	Tools/Technologies
Programming Language	Python (Primary language for data processing and model building)
Key Libraries	Streamlit (for front-end dashboard), Prophet, ARIMA (for time series modeling), Pandas, NumPy, Plotly (for interactive visualizations)
Data Source	Kaggle dataset (leveraging CPCB & NASA references for comprehensive geographical coverage and historical depth)
Deployment	Streamlit Cloud (for fast, accessible web deployment)
Development Environment	VS Code / Jupyter Notebook (for iterative development and experimentation)

The integration of **Prophet** and **ARIMA** models allows us to capture both linear trends and complex seasonal patterns inherent in pollution data.

Understanding the Air Quality Index (AQI)

The AQI is the universal language for communicating air pollution levels. It transforms complex pollutant data into a single, easy-to-understand number.

1

What is AQI?

A numerical scale, typically ranging from 0 to 500, used by government agencies to communicate how clean or polluted the air is and what associated health risks might be.

2

Main Pollutants Tracked

Carbon Monoxide (CO), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), Particulate Matter (PM10 and PM2.5), Lead (Pb), Ammonia (NH_3), and Ozone (O_3).

Range	Category	Health Implications
0–50	**Good**	Minimal impact; air quality is considered satisfactory.
51–100	**Satisfactory**	Minor breathing discomfort possible for sensitive people.
101–200	**Moderate**	Breathing discomfort to people with lung/heart disease; moderate impact for general public.
201–300	**Poor**	Breathing difficulty for most people; prolonged exposure can cause severe respiratory effects.
301–400	**Very Poor**	Respiratory illness upon prolonged exposure; serious health effects for vulnerable groups.
401–500	**Severe**	Air quality is hazardous. Affects healthy people and seriously impacts those with existing diseases.

Global Agencies and Platforms for Monitoring AQI

A variety of reputable organizations and platforms provide current and historical air quality data, which serve as foundational resources for our project.

CPCB

The Central Pollution Control Board provides comprehensive national data and standards for India, crucial for regulatory context.

SAFAR-India

System of Air Quality and Weather Forecasting and Research; a dedicated system for providing location-specific information.

AQICN.org

The World Air Quality Index Project, aggregating data from hundreds of monitoring stations globally.

IQAir & BreezoMeter

Leading technology companies that use proprietary algorithms to provide real-time, global air quality intelligence.

Major Contributors to Air Pollution

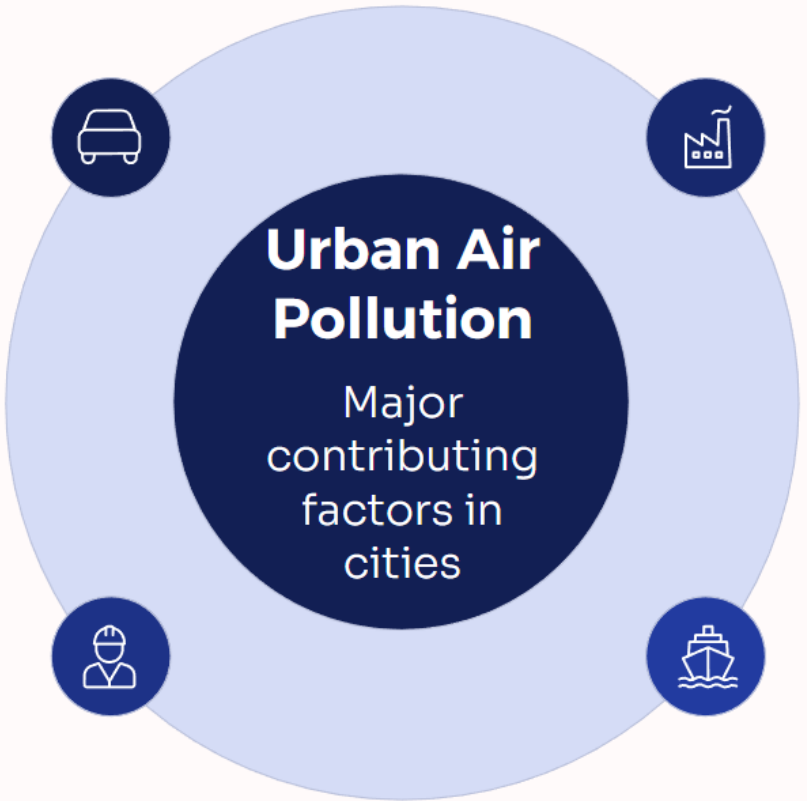
Air pollution is a complex issue driven by numerous anthropogenic and natural sources. Addressing the problem requires understanding the root causes.

Vehicular Emissions

Exhaust from cars, trucks, and buses

Industrial Exhausts

Factory stacks and manufacturing emissions



Construction Dust

Particulate matter from building sites

Fossil Fuel Burning

Power plants and heating-related emissions

Impact on Human Health and the Global Environment

The consequences of poor air quality extend beyond localized smog; they pose fundamental threats to public health and ecological stability.

Human Health Crisis

Exposure leads to chronic respiratory diseases (asthma, bronchitis), cardiovascular illnesses, fatigue, and can even impair cognitive development in children.

Environmental Degradation

Pollutants lead to acid rain, reduced visibility (smog), and direct damage to sensitive ecosystems, including soil quality and agricultural crops.

Global Climate Change

Many air pollutants (like black carbon and ground-level ozone) are short-lived climate pollutants that significantly increase greenhouse gas concentrations, accelerating global warming.

Our forecasting model provides an essential tool for mitigating these impacts by allowing governments and individuals to implement proactive protective measures during high-risk periods.

Empowering Action: Tools for Awareness and Prevention

Tree Plantation Drives

Implementing large-scale afforestation and urban greening projects, as trees are vital natural filters for absorbing CO2 and capturing particulate matter.

Real-time AQI Monitoring Apps

Providing accessible mobile applications that offer current Air Quality Index (AQI) data for user locations, enabling immediate, informed choices.

Predictive Early Warning Systems

Leveraging advanced modeling (like our AirAware project) to forecast future air quality, giving communities and authorities lead time to prepare and mitigate risks.

AirAware Project Impact: Driving Public Health and Policy

The AirAware project translates complex environmental data into actionable insights, providing value across multiple user segments—from individuals to governmental agencies.

Safe Outdoor Activities

Empowers citizens to plan their daily routines, especially outdoor exercise or prolonged exposure, based on forecasted air quality, minimizing health risks.

Promoting Public Awareness

By visualizing air quality trends clearly, the project significantly increases public understanding of air pollution's severity and variability.

Assisting Government Bodies

Provides pollution management teams with data-driven tools for strategic planning, resource allocation, and timely issuance of public advisories or regulatory actions.

Supporting Global Goals

Directly contributes to achieving Sustainable Development Goal (SDG) 13: Climate Action, by providing crucial data for environmental monitoring and accountability.

Diving Deeper into the Workflow Stages

Data Collection

Gathering raw atmospheric data (PM2.5, NO2, O3, SO2, etc.) from reliable public APIs and historical archives.

Preprocessing

Cleaning data, handling missing values, standardizing formats, and feature engineering to prepare inputs for the prediction models.

AQI Calculation

Converting raw pollutant concentrations into the standardized Air Quality Index (AQI) metric for universal understanding.

Model Training

Applying various time-series algorithms (Prophet, ARIMA, and LSTM Deep Learning) to identify patterns and relationships within the historical data.

Forecast Generation

Utilizing the trained models to generate precise short-to-medium-term AQI predictions for selected cities.

Deployment & Visualization

Deploying the interactive application on Streamlit Cloud, allowing users to easily visualize historical trends and future forecasts.

Project Overview: 4-Milestone Journey

Our project progressed through four distinct and critical phases, building from raw data to a deployed, functional application.

Milestone 1: Data Preprocessing & EDA

Establishing a clean, structured foundation of air quality data.

Milestone 2: Model Training & Evaluation

Developing and validating predictive models for accurate AQI forecasting.

Milestone 3: Alert Logic & Trend Visualization

Translating raw forecasts into actionable alerts and clear visual trends.

Milestone 4: Streamlit Dashboard Implementation

Building and deploying the interactive user interface.

Milestone 1: Data Preprocessing & Exploratory Data Analysis (EDA)

The success of any forecasting model hinges on the quality of its input data. This phase focused on rigorous data preparation.

Source & Collection

Datasets were collected from **Kaggle** , containing comprehensive hourly measurements of various pollutants across major Indian cities.

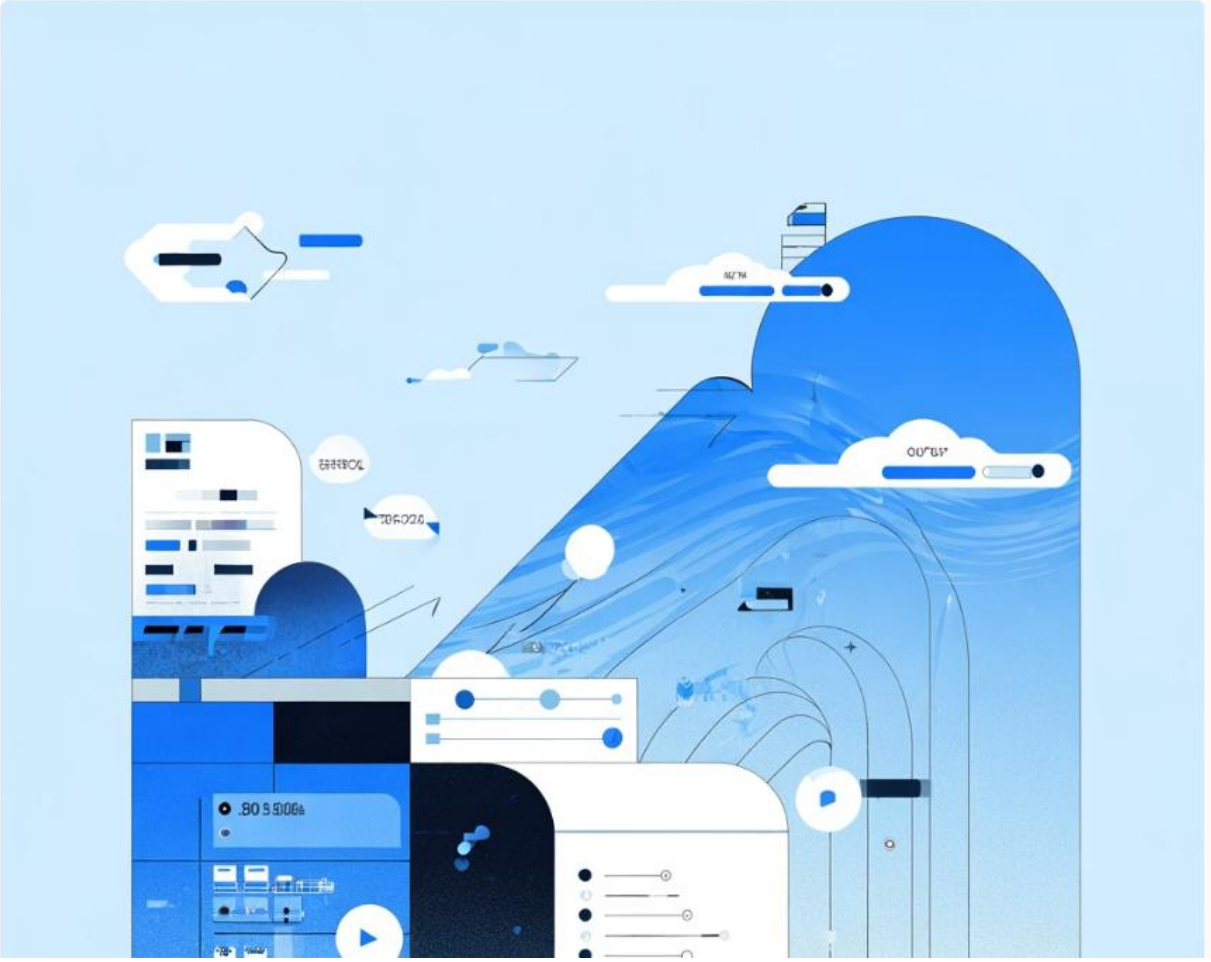
Data Cleaning & Preparation

Key challenges included handling missing timestamps and identifying/treating outliers. Data was then resampled to **daily averages** for consistent time-series analysis.

Exploratory Analysis

Conducted detailed EDA to understand pollutant behavior, including correlation matrices and time-series decomposition to reveal underlying seasonality and trends.

[View Milestone 1 Files on Google Drive]

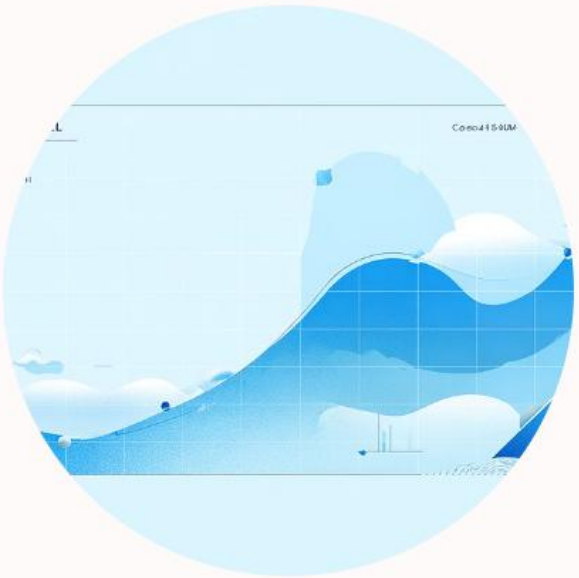


Milestone 2: Model Training and Evaluation

We tested multiple advanced time-series forecasting techniques to find the most accurate prediction method for the complex nature of air quality data.

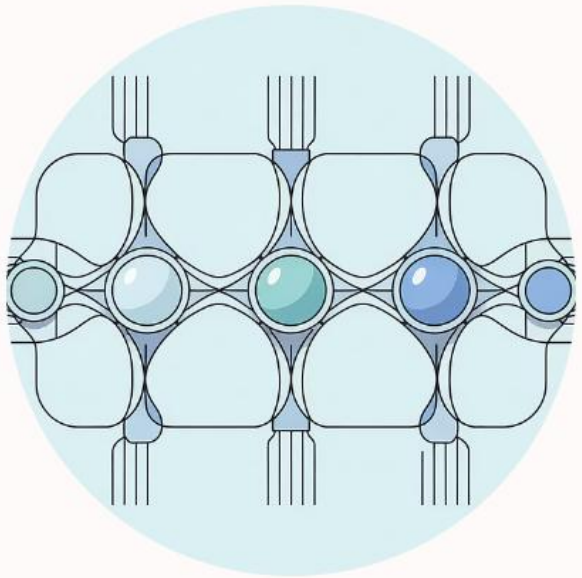
ARIMA (Autoregressive Integrated Moving Average)

A classical statistical approach, robust for stationary time-series data, providing a baseline prediction capability.



Prophet (Developed by Meta)

Excellent for handling data with strong seasonality and holidays, providing reliable long-term forecasts even with missing data.



LSTM (Long Short-Term Memory)

A deep learning model specialized in sequence data, capable of capturing non-linear relationships and long-term dependencies in pollutant levels.

Performance Metrics and Model Selection

Model effectiveness was quantified using standard metrics, leading to an optimized approach where performance was maximized per city.

Mean Absolute Error (MAE)	Measures the average magnitude of errors in a set of forecasts, without considering their direction.	Minimize
Root Mean Square Error (RMSE)	Measures the square root of the average of the squared errors, penalizing large errors more heavily.	Minimize
Model Selection	Selected the best-performing model (based on lowest MAE/RMSE) **individually for each city**, ensuring tailored accuracy.	Select Best

[View Milestone 2 Files on Google Drive]

Milestone 3: Alert Logic and Trend Visualization

Forecasting data is only useful if it is translated into understandable, actionable information for the public. This required sophisticated visualization and alert protocols.

AQI Categorization

Implemented the standard AQI classification system (Good, Satisfactory, Moderate, Poor, Very Poor, Severe) to immediately contextulize the predicted value.

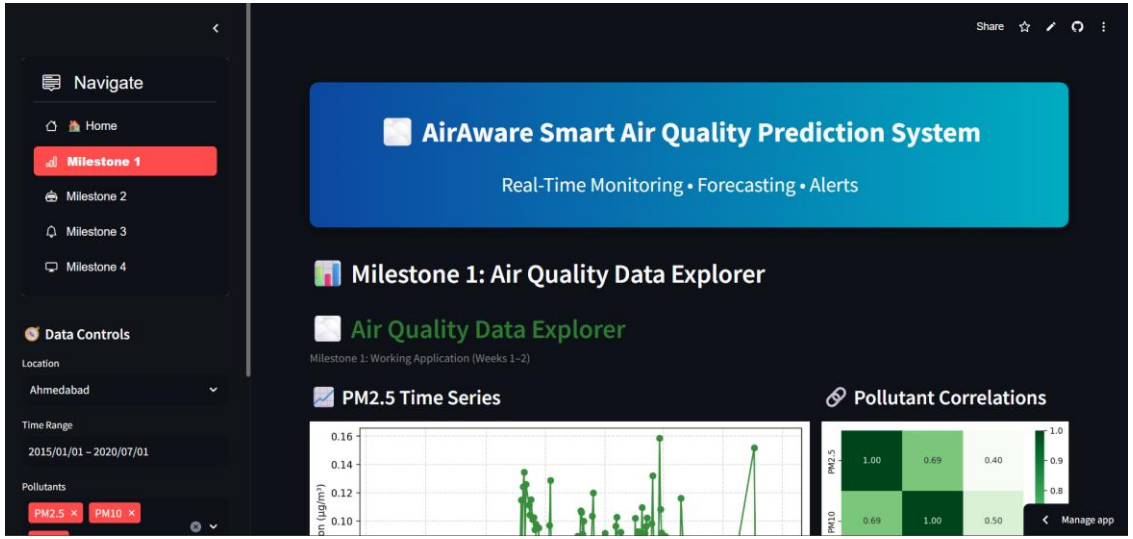
Threshold Alerting

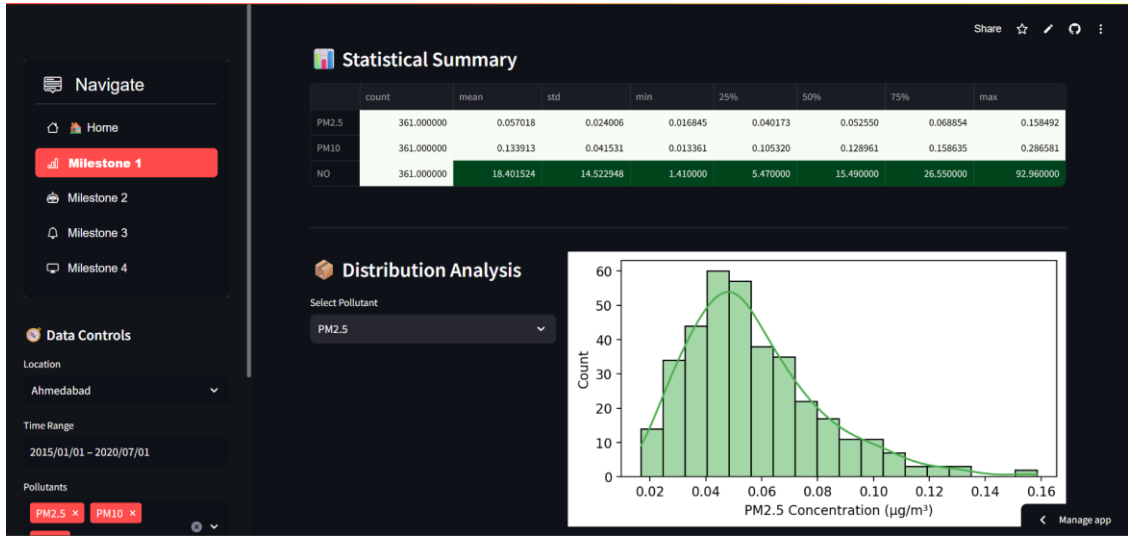
Logic was built to automatically trigger visual alerts (banners and color changes) when the predicted AQI exceeds safe operational thresholds, prompting precautionary measures.

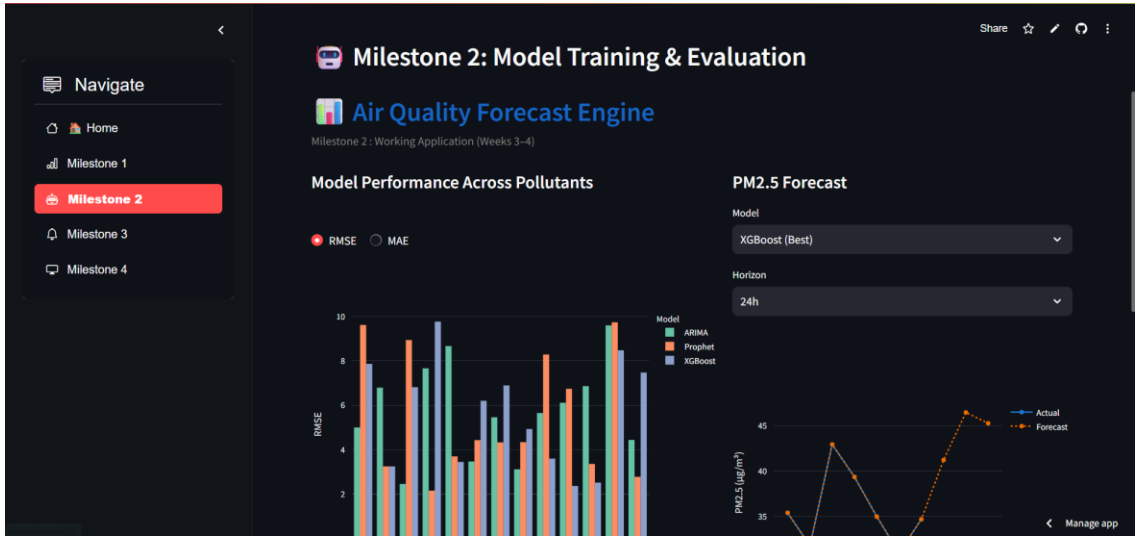
Seasonal Trend Analysis

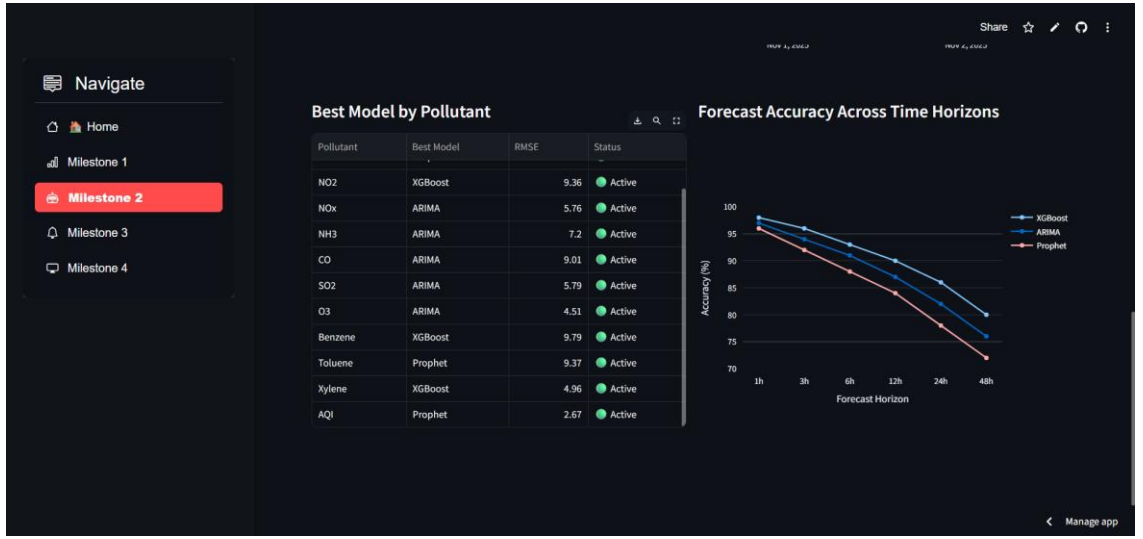
Visualizations of pollutant-wise seasonal trends were added to provide historical context, showing users when pollution is typically highest (e.g., winter smog, monsoon decrease).

Milestone 3 Outcome: Visualizing Pollution Patterns









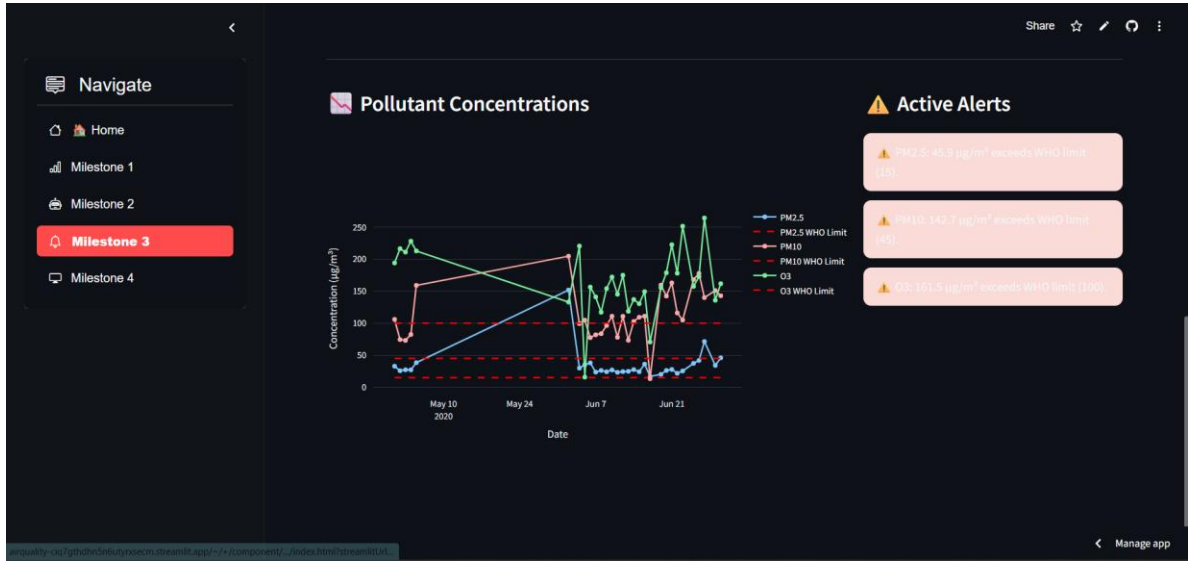
Milestone 4: Streamlit Dashboard Implementation

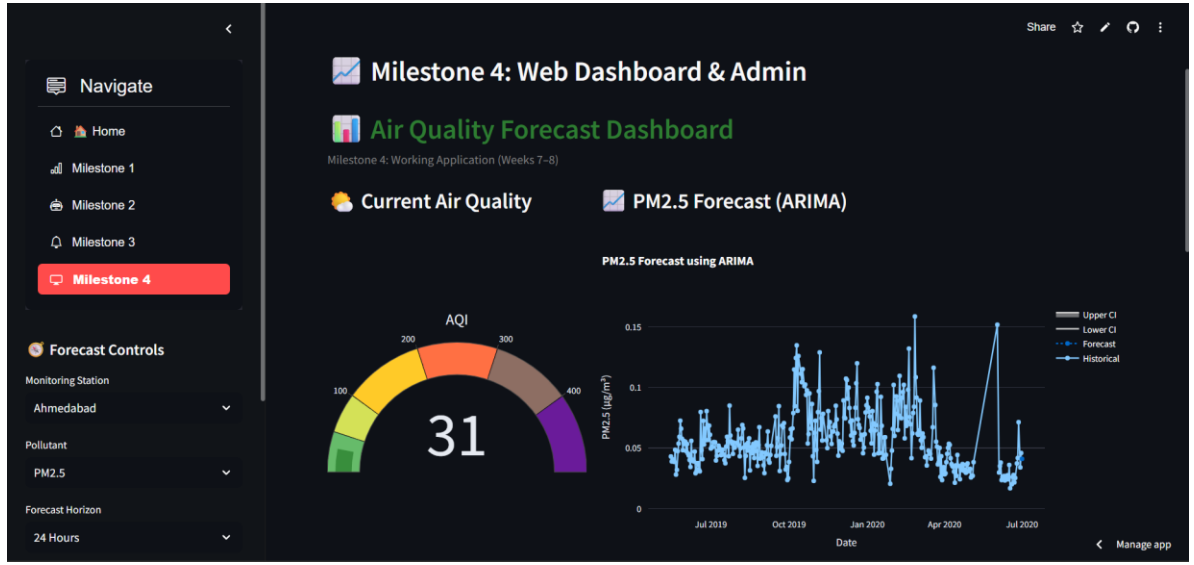
The final phase involved developing a robust and user-friendly web application using Streamlit, enabling live interaction with our predictive models.

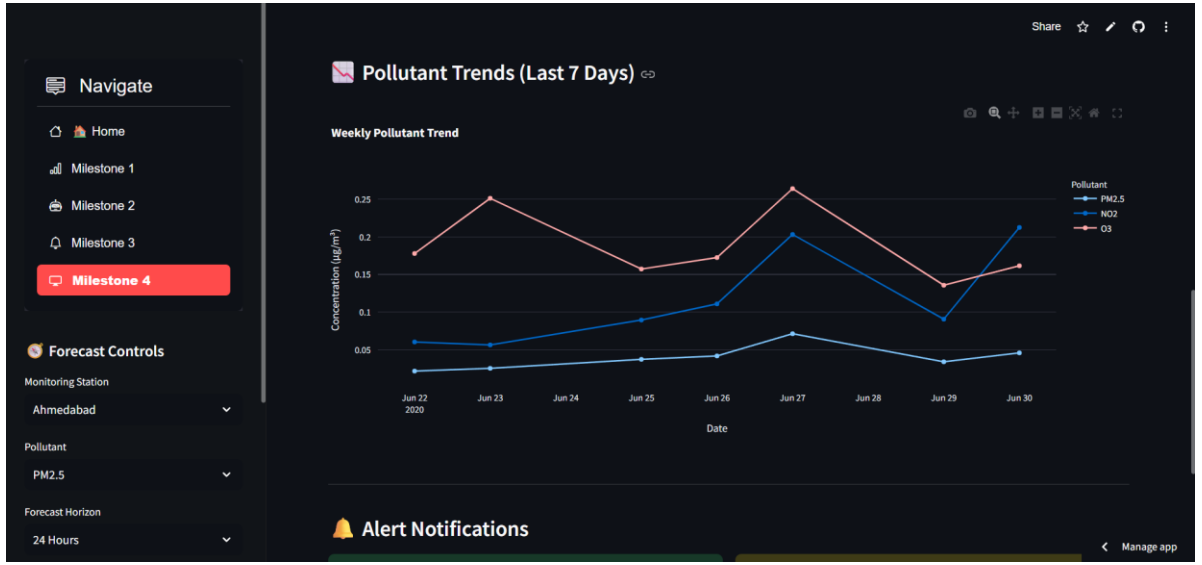
Interactive Features

- Users can easily select their **city, pollutant, and date range** for personalized forecasts.
- Display of **real-time line plots** showing predicted values versus historical data.
- Prominent **AQI gauges** and alert banners for immediate status reporting.

Milestone 4 Outcome







Our Solution is Live: Explore the Deployed Dashboard!

I am excited to announce the successful deployment of our Streamlit application, which seamlessly integrates all project milestones into a powerful, interactive tool for public health and urban planning.

Visit the Live Dashboard Here!

https://airquality-ciq7gthdhn5n6utyrxsecm.streamlit.app/

Getting Started: Running the AirAware Dashboard

Launching the AirAware dashboard is a straightforward process that allows users to quickly access and interact with the air quality forecasts. Follow these simple steps to run the project locally.

Install Dependencies

Execute `pip install -r requirements.txt` to ensure all necessary Python libraries and packages are installed in your environment.

Run the Application

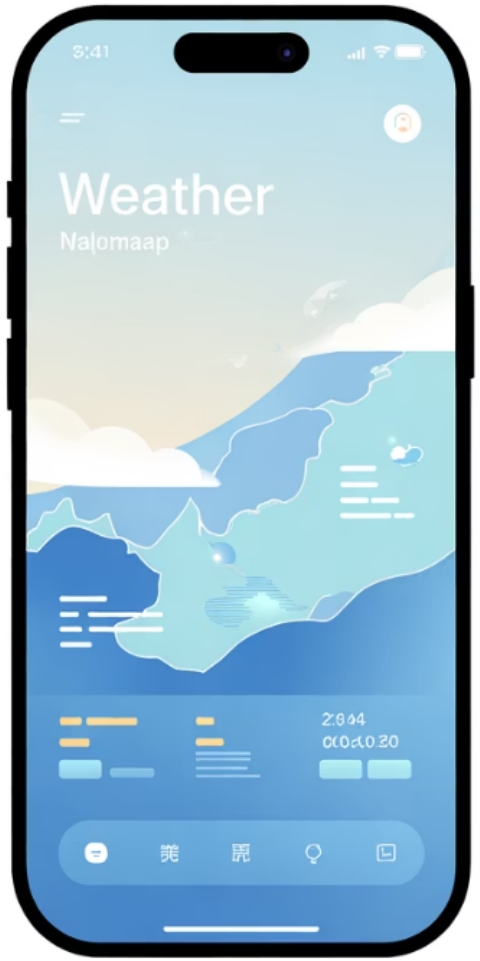
Start the Streamlit web application by running the command: `streamlit run main_dashboard.py` in your terminal.

Select Parameters

Once the dashboard loads, use the interactive sidebar to select your desired city and the specific date range for the forecast.

View Results

Analyze the historical AQI trends and examine the forecasted AQI results presented interactively through clear charts and visualizations.



Enhancing AirAware: Our Roadmap for Future Development

We plan to continuously expand the capabilities and reach of the AirAware system by integrating more data sources and developing user-centric features. This iterative development ensures the platform remains cutting-edge and highly relevant.

Mobile Application for Alerts

Developing a native mobile app to provide push notifications for dangerous AQI levels and interactive air quality maps.

Integration with IoT Sensors

Connecting the platform directly with localized Internet of Things (IoT) air quality sensors for micro-level, hyper-accurate readings.

Advanced ML Model Hybrids

Exploring and implementing advanced machine learning techniques, such as combining Prophet with Deep Learning (LSTM/GRU), for superior predictive accuracy.

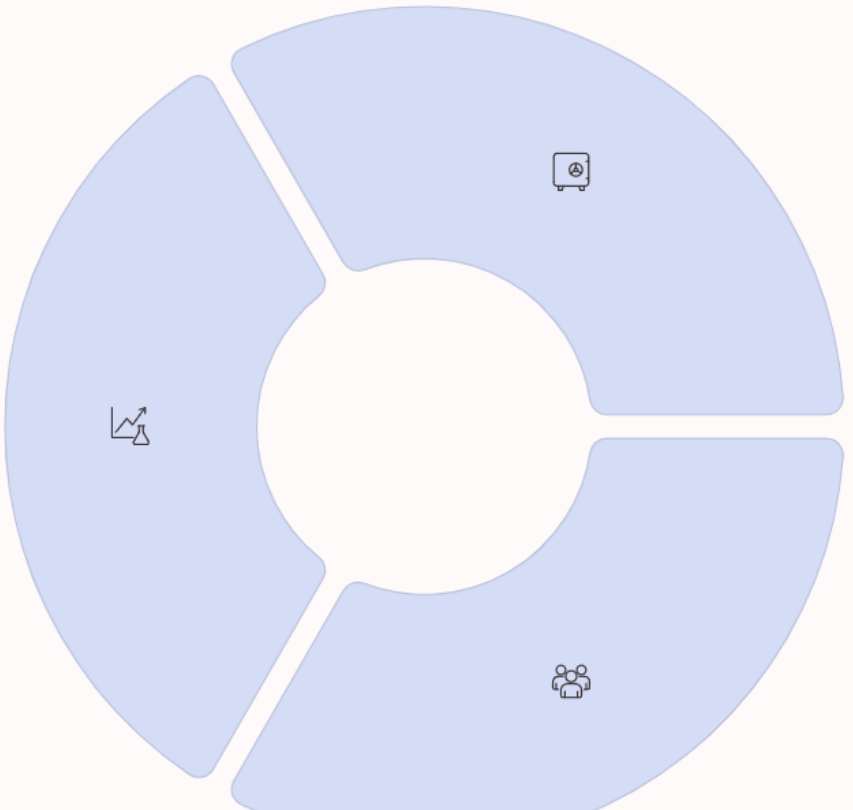
Conclusion: AirAware—Breathing Smarter, Not Harder

> **“Let’s breathe smarter — not harder.”** 🌿

AirAware successfully provides an accessible and robust data-driven solution for forecasting and visualizing air quality. By demystifying complex environmental data, the project promotes increased public awareness, enables healthier personal decision-making, and fosters greater environmental consciousness within the community.

Data-Driven Insights

Delivering accurate and reliable AQI predictions.



Informed Decisions

Empowering citizens to protect their health proactively.

Environmental Advocacy

Building a foundation for positive community action on pollution.