AQI Multi-Day Forecasting System

End-to-End MLOps Project with Streamlit Deployment

Project Overview

The project covers the entire ML lifecycle:

- Data ingestion
- Feature engineering
- Multi-horizon model training
- Experiment tracking
- Model registry & versioning
- Automated inference using GitHub Actions
- Cloud database storage
- Real-time visualization using Streamlit

1. Problem Statement

Air quality has a direct impact on:

- Public health
- Urban planning
- Environmental policy

Traditional AQI dashboards show forecasts.

Objective

Build a system that:

- Predicts AQI for the next 1–3 days
- Runs automatically
- Stores predictions persistently
- Visualizes results in real time

2. Dataset & Data Flow

Data Source

- Hourly air-pollution measurements stored in **MongoDB**
- Pollutants used:
    - PM2.5
    - PM10
    - CO
    - NO_2
    - O_3
    - SO_2
    - NH_3

Data Flow Architecture

MongoDB (Raw AQI Data)

↓

Feature Engineering

↓

Model Training (MLflow)

↓

Model Registry (DagsHub)

↓

Scheduled Inference (GitHub Actions)

↓

Predictions Stored in MongoDB

↓

Streamlit Dashboard

3. Feature Engineering (Time-Series Optimized)

To capture AQI temporal behavior, the following features were engineered:

Time Features

- Hour of day
- Day of month
- Month
- Day of week

Lag Features (Historical Dependency)

- PM2.5 lagged by 1–6 hours

Rolling Statistics

- Rolling mean (3h, 6h)
- Rolling standard deviation
- Rolling maximums

Derived Features

- PM2.5 change rate
- PM10 change rate
- PM2.5 / PM10 ratio
- AQI volatility & momentum

Multi-Day Targets

Assuming hourly data:

Target t+1 → AQI after 24 hours

Target t+2 → AQI after 48 hours

Target t+3 → AQI after 72 hours

4. Multi-Horizon Modeling

Instead of one model predicting multiple days, separate models were trained for:

- Day +1
- Day +2
- Day +3

5. Models Evaluated

- Random Forest Regressor
- Gradient Boosting Regressor
- Linear Regression (baseline)

6. Evaluation Metrics

- MAE (Mean Absolute Error)
- RMSE (Root Mean Squared Error)
- R^2 Score

7. Best Models Selected

Horizon	**Best Model**
t+1	Random Forest
t+2	Random Forest
t+3	Gradient Boosting

Each best model was:

- Logged to MLflow
- Registered in the MLflow Model Registry
- Promoted to Production stage

8. Experiment Tracking & Model Registry

MLflow + DagsHub Integration

- All experiments tracked remotely
- Metrics, parameters, and artifacts logged
- Full reproducibility of training runs

Model Versioning

Each horizon has its own registered model:

  - aqi_t_plus_1
  - aqi_t_plus_2
  - aqi_t_plus_3

This enables:

- Safe model upgrades
- Rollbacks
- Production-grade inference

9. Automated Inference Pipeline

GitHub Actions Workflow

- Runs on a schedule (hourly)
- Loads latest Production models
- Builds features from newest data
- Generates forecasts
- Stores predictions in MongoDB

10. Deployment

- Deployed on Streamlit Cloud
- Connected directly to MongoDB

Summary

- End-to-end ML pipeline
- Multi-day AQI forecasting
- Feature engineering for time-series
- MLflow experiment tracking
- Model registry & production staging
- Automated inference via CI/CD
- Cloud deployment with Streamlit