Pearls AQI Predictor – Final Project Report

Author: Danish Ahmed

Domain: Data Sciences | 10Pearls Shine Internship

Date: 9 November 2025

1. Project Overview

The AQI Predictor project develops a production-grade machine-learning pipeline to forecast the Air Quality Index (AQI) over the next **24 to 72 hours** in Karachi. Utilizing the OpenWeather APIs and a Hopsworks Feature Store, the system automates data collection, feature-engineering (29 features), daily model training, and pushes predictions to a live Streamlit dashboard. The goal is to enable stakeholders to monitor and anticipate air quality changes for informed decision-making.

2. Problem & Objectives

Air pollution in Karachi is a growing public-health concern: fine particulate matter (PM2.5) routinely exceeds safe limits.

Objectives:

- Build a forecast system for next-day AQI.
- Engineer meaningful features from weather and pollutant data.
- Compare multiple machine-learning models and select best performer.
- Deploy an interactive, user-friendly dashboard for real-time usage.
- Automate the pipelines (feature ingestion, model training, deployment) for scalability.

3. Architecture & Technology

The system ingests data hourly via the OpenWeather Air Pollution API (current, historical, forecast), processes and stores features in the Hopsworks feature store, trains models daily via GitHub Actions, and deploys the best model to a Streamlit dashboard. Key technologies include Python 3.10, scikit-learn, Hopsworks, Streamlit, SHAP, and GitHub Actions for CI/CD.

4. Data, Features & Modelling

Data Sources: Hourly weather and pollutant measurements for Karachi (e.g., PM2.5, PM10, O_3, NO_2, SO_2, CO; temperature, humidity, pressure, wind speed).

Feature Engineering: 29 engineered features covering time-based variables (hour, day, month, cyclical encodings), lagged AQI values (1h, 3h, 6h, 12h, 24h), rolling statistics (3h, 6h, 12h windows), change/ratio features, pollutant and weather variables.

Models Trained: Ridge Regression, Random Forest, Gradient Boosting, Lasso.

Results: The Ridge Regression model achieved the best performance with RMSE = 9.14, MAE = 7.53, $R^2 = 0.741$.

5. Results & Evaluation

Model	RMSE	MAE	R^2
Ridge Regression	9.14	7.53	0.741
Random Forest	10.32	8.08	0.669
Gradient Boosting	9.70	7.61	0.708
Lasso Regression	10.96	8.70	0.627

The R^2 of 0.741 indicates approximately **74% of variance** in AQI is explained by the model. This result, combined with automated pipelines and live deployment, demonstrates a robust solution ready for production monitoring of air quality.

6. Deployment & Automation

- The dashboard is live and publicly accessible. The pipeline is fully automated: Feature pipeline runs hourly.
- Model training pipeline runs daily via GitHub Actions.
- Model versioning and lineage maintained via Hopsworks.
- Environment variables `OPENWEATHER_API_KEY` and `HOPSWORKS_API_KEY` configured securely in GitHub Secrets.

- **Links:**
  Live Dashboard at https://aqi-predictor-aewykmqnvacleujd4kzero.streamlit.app/
  GitHub repo at https://github.com/Danish-Ahmed-Head/aqi-predictor

7. Limitations

- The model is city-specific (Karachi only); extending to other cities requires new data and retraining.
- Dependent on external data quality and availability: gaps or errors in OpenWeather data will affect predictions.
- Seasonal or extreme pollution events (fires, dust storms) may degrade performance; retraining and additional features may mitigate this.
- API free-tier rate-limits restrict heavy backfill or large-scale usage.

8. Conclusion

This project successfully delivers a scalable, automated, production-ready system for forecasting AQI in Karachi. With strong modelling results, live deployment, and robust engineering, it meets all submission requirements of the Shine programme.

Prepared by: Danish Ahmed
Date: 9 November 2025