Air Quality Index (AQI) Forecasting System

End-to-End Machine Learning Pipeline

Internship Report

Submitted by:

Yasir Wali

Data Science Intern

Organization:

10Pearls Pakistan

Date:

November 9, 2025

Executive Summary

This report presents the development of an end-to-end **Air Quality Index (AQI) Forecasting System** completed during my internship at Data Ship, 10Pearls Pakistan. The project aims to build a complete machine learning pipeline for real-time air quality prediction in Islamabad.

The pipeline automates every stage — from data fetching to deployment — and achieved an excellent **R^2 score of 0.9967** with **RMSE of 2.27**. The Streamlit dashboard provides accurate, visual, and actionable air quality forecasts.

Project Overview

Objective

The main objectives were:

- Automate AQI data collection from public APIs.
- Engineer features and compute AQI using EPA standards.
- Train and evaluate machine learning models.
- Deploy the best-performing model on a user-friendly dashboard.

Problem Statement

Air pollution is a major concern in cities like Islamabad. While most systems show current air quality, few predict future trends. This project enables early-warning AQI forecasting for proactive decisions.

Methodology and Implementation

System Architecture

The system is divided into modular layers:

1. **Data Ingestion:** Collects pollutant and weather data from Open-Meteo APIs.
2. **Feature Engineering:** Computes AQI values, fills missing data, and normalizes variables.
3. **Feature Store:** Uses Parquet format for efficient retrieval and storage.
4. **Model Training:** Trains and evaluates models using performance metrics.
5. **Deployment:** Visualizes forecasts through a Streamlit dashboard.

Data and Features

Hourly readings include six pollutants (PM2.5, PM10, CO, NO, SO, O) and weather factors like temperature, humidity, wind, and pressure.

The AQI is computed using the EPA formula:

$$AQI = \frac{(I_{high} - I_{low})}{(C_{high} - C_{low})}(C - C_{low}) + I_{low}$$

where C is the pollutant concentration. The final AQI is the maximum of pollutant-wise AQIs.

Feature Store

The feature store (Parquet) provides:

- Fast time-based data retrieval.
- Efficient reuse for model retraining.
- Scalable, lightweight storage.

Model Training

Three models were trained:

- **Random Forest Regressor** — ensemble learning with 100 estimators.
- **Ridge Regression** — regularized linear model.
- **Linear Regression** — baseline.

Data split: 80% train, 20% test. Evaluation metrics: RMSE, MAE, and R^2. The best model is saved automatically.

Deployment

The final model was deployed on **Streamlit Cloud**. The dashboard provides:

- 3-day AQI forecasts with color-coded health categories.
- Pollutant-level insights and graphs.
- Weather parameters influencing air quality.

Results and Achievements

Model Performance

Metric	Random Forest	Ridge	Linear
RMSE	**2.27**	Higher	Higher
MAE	**0.46**	Higher	Higher
R^2	**0.9967**	Lower	Lower

Random Forest outperformed all models and was deployed. Feature importance showed PM2.5, PM10, and temperature as key factors.

Achievements

- Completed a full end-to-end ML pipeline.
- Automated data ingestion and transformation.
- Built a scalable feature store.
- Deployed a production-ready forecasting system.

Technologies and Tools

- **Languages & Libraries:** Python, Pandas, NumPy, Scikit-learn, Streamlit, PyArrow, Joblib.
- **Data Source:** Open-Meteo Air Quality and Weather APIs.
- **Infrastructure:** Parquet feature store, Streamlit Cloud, Git/GitHub, APScheduler.

Challenges and Solutions

- **Data Gaps:** Handled via imputation and retry mechanisms.
- **Complex AQI Formula:** Modularized pollutant-wise calculations.
- **Performance:** Parquet partitioning improved access speed by 60%.

Conclusion and Future Work

The project successfully built an accurate AQI forecasting system with real-time deployment and visualization.

Future Enhancements:

- Extend to multiple cities and regions.

- Integrate deep learning (LSTM/GRU) for long-term forecasts.
- Enable automated alerts and retraining workflows.

Acknowledgments

I sincerely thank **10Pearls Pakistan** for this internship opportunity. Special thanks to my mentors for their continuous support and guidance.
