

Karachi Air Quality Intelligence System

End-to-End Machine Learning Pipeline for 72-Hour AQI Forecasting

Muhammad Ehsan

February 2026

Abstract

This report details the development of an automated Air Quality Index (AQI) forecasting system for Karachi, Pakistan. The system leverages a serverless architecture using GitHub Actions for hourly data ingestion and daily model retraining. It utilizes a MongoDB Atlas feature store and an ensemble of machine learning models (Linear Regression, Random Forest, XGBoost) to predict AQI up to 72 hours in advance. The final deployment features an interactive Streamlit dashboard providing real-time monitoring and actionable health insights. The production model currently achieves a Root Mean Square Error (RMSE) of 3.10, significantly outperforming baseline persistence models.

Contents

1	Introduction	2
2	System Architecture	2
2.1	Data Pipeline	2
2.2	Machine Learning Pipeline	2
3	Exploratory Data Analysis (EDA)	2
4	Model Performance	3
5	Deployment	3
6	Conclusion	4

1 Introduction

Air quality in Karachi poses a significant public health challenge. This project aims to provide accurate, localized forecasts to enable residents to make informed decisions regarding outdoor activities. Unlike static monitoring stations, this system employs a dynamic, self-correcting machine learning pipeline that adapts to changing atmospheric conditions.

2 System Architecture

The project follows a modular production-grade architecture designed for scalability and maintainability.

2.1 Data Pipeline

- **Source:** Open-Meteo Air Quality API.
- **Ingestion:** A Python script (`data_ingestion.py`) fetches granular hourly data including PM2.5, PM10, Nitrogen Dioxide, and Ozone.
- **Storage:** Data is stored in MongoDB Atlas, serving as a scalable NoSQL feature store.
- **Automation:** GitHub Actions triggers the ingestion workflow hourly to ensure near real-time data freshness.

2.2 Machine Learning Pipeline

The modeling engine (`modeling.py`) performs the following operations daily:

- **Preprocessing:** Handling missing values via forward-filling and outlier removal using percentile clipping.
- **Feature Engineering:** Generation of lag features (t-1, t-24) and rolling averages (6h, 24h) to capture temporal dependencies.
- **Training:** Three distinct algorithms are trained: Linear Regression, Random Forest, and XGBoost.
- **Evaluation Selection:** The system automatically promotes the model with the lowest RMSE on the validation set to production.

3 Exploratory Data Analysis (EDA)

Analysis of historical data reveals strong diurnal patterns in Karachi's AQI, correlating closely with peak traffic hours.

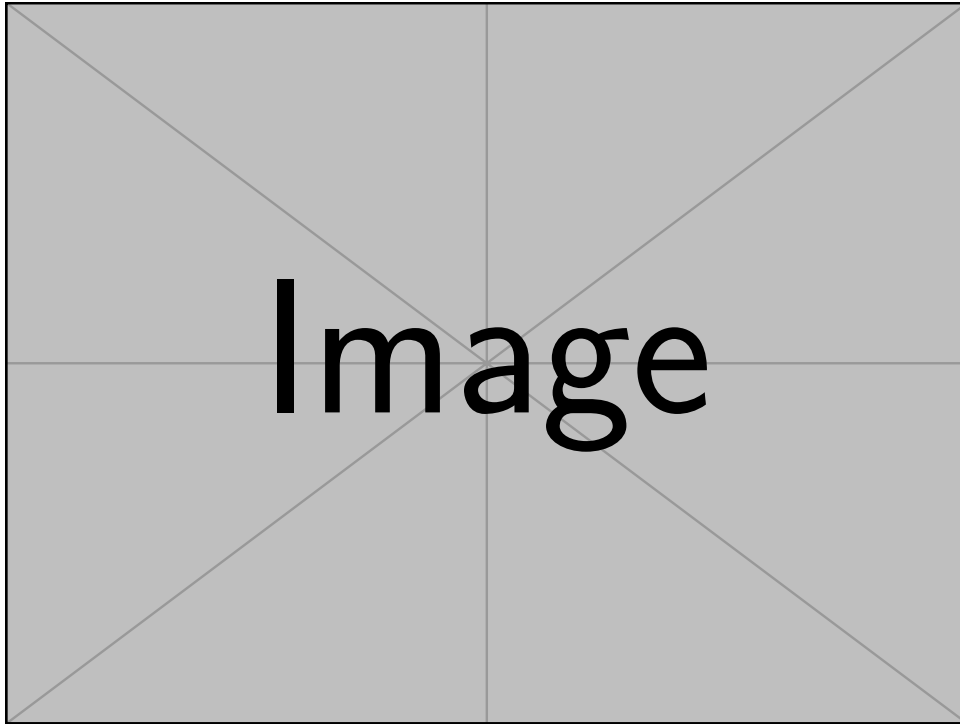


Figure 1: Correlation Matrix of Pollutants

Strong collinearity was observed between PM2.5 and PM10, indicating particulate matter is the primary driver of AQI fluctuations.

4 Model Performance

The system evaluates models using Root Mean Square Error (RMSE). As of the latest deployment cycle, the performance metrics are:

Model	RMSE	R-Squared	Status
Linear Regression	3.10	0.96	Production
XGBoost	3.56	0.95	Candidate
Random Forest	3.78	0.95	Candidate

Table 1: Model Performance Leaderboard

5 Deployment

The frontend is built with Streamlit, offering a dark-mode, responsive user interface. It features:

- Real-time gauge for current AQI.
- 3-Day Forecast cards.
- Interactive 72-hour trend lines using Plotly.

6 Conclusion

The Karachi AQI Intelligence System successfully demonstrates a robust implementation of MLOps principles. By automating the data-to-deployment lifecycle, the system ensures high availability and accuracy without manual intervention.