Internship Project Report: Karachi AQI Prediction System

Intern: Karan Kumar
Project Title: End-to-End Automated Air Quality Forecasting System For Karachi , Sindh , Pakistan.

---

1. Executive Summary

The **Karachi AQI Prediction System** is a sophisticated machine learning platform designed to provide real-time and predictive insights into Karachi's air quality. The project addresses the critical need for accurate air quality monitoring in high-density urban areas. Key achievements include the deployment of a robust ML pipeline with an **RMSE of 1.75**, automated data ingestion from global APIs, and a professional-grade interactive dashboard.

2. System Architecture

The system follows a modular, scalable architecture implementing a complete Data-to-Inference loop:

- **Data Ingestion Engine**: Fetches hourly weather and air quality parameters from the **Open-Meteo API**.
- **Feature Store (MongoDB Atlas)**: A cloud-native NoSQL database serves as the centralized feature store, ensuring data persistence and enabling time-series analysis.
- **ML Pipeline**: Implements automated preprocessing (interpolation, lag engineering) and model retraining.
- **Inference Layer**: A dark-themed **Streamlit** dashboard provides real-time predictions and 72-hour forecasts.

3. Implementation Details

3.1 Data Ingestion & Storage

- **Source**: Open-Meteo Air Quality API.
- **Parameters**: PM2.5, PM10, CO, NO2, SO2, O3, and US-AQI.
- **Reliability**: Implemented **retry logic** and **exponential backoff** to handle API rate limits and network instability.

- **Persistence**: Data is synchronized to MongoDB Atlas using efficient **Upsert** (Bulk Write) operations to avoid duplication.

3.2 Machine Learning Workflow

- **Features**: Lag features (1h, 2h, 24h), rolling averages (6h, 24h), and temporal components (hour, day).
- **Model Selection**: Evaluated Linear Regression, Random Forest, and **XGBoost**. XGBoost was selected as the production model due to its superior handling of non-linear trends in AQI data.
- **Validation**: 80/20 Time-series split ensuring the model is validated on future data relative to training.

3.3 Dashboard Features

Preview unavailable *Figure 1: Real-time AQI Gauge and 3-Day Forecast Visualization*

- **Real-time Monitoring**: Instant display of the current hour's predicted AQI with color-coded severity badges.
- **Interactive Forecasts**: Graphing 72-hour trends using Plotly for deep dive analysis.
- **Automated Insights**: Context-aware summary boxes indicating whether air quality is improving or worsening.

4. Technical Challenges & Solutions

Challenge	Impact	Solution Implemented
Data Synchronization	Inconsistent hourly data retrieval	Implemented GitHub Actions with cron jobs for hourly sync and daily retraining.
Prediction Drift	Multi-step forecasts tended to explode	Applied **Recursive Multi-Step Forecasting** with dampening and physical limit clipping (0-500).
Database Connectivity	Intermittent cloud connection errors	Integrated robust Mongo client initialization with DNS resilience and connection timeouts.
Model Sensitivity	Outliers in sensor data affecting accuracy	Implemented training-time clipping (1st-99th percentile) during preprocessing.

5. Performance Metrics

Metric	Value	Significance
RMSE	1.75	High precision in AQI point estimation.
R^2 Score	0.99	Explains 99% of variance in the test set.
Inference Time	<50ms	Rapid dashboard response for end-users.

6. Conclusion

This project successfully demonstrates the application of modern ML Ops principles to public health data. By combining automated data pipelines, a centralized cloud feature store, and high-performance regressors, the Karachi AQI Prediction System provides a reliable tool for urban environmental monitoring.