(AQI) Prediction System

End-to-End Machine Learning Pipeline with Real-Time Forecasting

Core Stack: Python, MongoDB Atlas, Streamlit, OpenWeather API

Project Report

Submitted by:
Asaad Bin Amir

(Data Sciences)

Submitted to:

10Pearls Shine Internship Program

Submission Date:

February 18, 2026

1 ABSTRACT

This report presents the design and implementation of a production-ready machine learning system for real-time Air Quality Index (AQI) prediction in Karachi, Pakistan. The system addresses the critical need for accurate air quality forecasting by integrating live pollutant and meteorological data from the OpenWeather API into an end-to-end MLOps pipeline. Built using Python and developed in VS Code, the solution leverages MongoDB Atlas as a feature store to ensure data consistency across training and inference workflows.

The predictive engine employs ensemble machine learning techniques, with Gradient Boosting emerging as the optimal model, achieving an R^2 score of 0.82 for 24-hour forecasts, 0.77 for 48-hour predictions, and 0.71 for 72-hour horizons, with corresponding Mean Absolute Errors of 0.46, 0.55, and 0.64 respectively. The system incorporates advanced feature engineering with 82 derived features including lag variables, rolling statistics, cyclical temporal encodings, and pollutant-weather interactions.

A comprehensive model registry built on MongoDB GridFS enables version control and seamless production deployment of trained models. The final solution is delivered through an interactive Streamlit web dashboard, providing the public with real-time AQI measurements, multi-horizon forecasts (24h, 48h, 72h), historical trend analysis, and automated health advisories aligned with European CAQI standards. The system demonstrates the practical application of MLOps principles in environmental monitoring and public health awareness.

2 EXECUTIVE SUMMARY

The Air Quality Index (AQI) Prediction System for Karachi is a comprehensive machine learning solution designed to transform real-time environmental data into actionable public health insights. Developed as a production-grade MLOps pipeline, this project demonstrates the complete lifecycle of a data science application—from automated data collection to model deployment and user-facing visualization.

At the core of the technical architecture is a modular pipeline built on MongoDB Atlas, serving as both a feature store and model registry. The system implements an hourly data collection schedule that integrates pollutant concentrations (PM2.5, PM10, NO_2, O_3, SO_2, CO) and meteorological variables (temperature, humidity, wind speed, pressure) from the OpenWeather API. A comprehensive backfill operation populated the database with over 2,700 historical records, enabling robust time-series analysis across multiple weeks of Karachi's diverse atmospheric conditions.

The feature engineering pipeline transforms 13 raw measurements into 82 sophisticated predictive features through lag variables (1h to 48h), rolling window statistics, cyclical temporal encodings, rate-of-change calculations, and pollutant-weather interaction terms. This advanced feature set captures both short-term fluctuations and long-term patterns essential for accurate forecasting in Karachi's complex coastal-urban environment.

Seven machine learning algorithms were systematically evaluated using rigorous cross-validation across three forecast horizons (24h, 48h, 72h). Gradient Boosting emerged as the optimal model, achieving R^2 scores of 0.82, 0.77, and 0.71 for the respective horizons, with Mean Absolute Errors of 0.46, 0.55, and 0.64 on the European CAQI scale (1-5). These results represent substantial improvements over baseline persistence models and demonstrate reliable predictive capability for public health applications.

The deployment architecture incorporates a MongoDB GridFS-based model registry enabling version control, metadata tracking, and seamless production model updates. The user-facing component is delivered through a responsive Streamlit web dashboard deployed on Streamlit Cloud, featuring real-time AQI monitoring with color-coded health categories, multi-horizon forecast cards, interactive historical trend visualizations across seven pollutants and four meteorological variables, and automated health advisories aligned with European air quality standards.

This project successfully demonstrates the practical implementation of MLOps principles—reproducible pipelines, automated workflows, model versioning, and continuous deployment—in the context of environmental monitoring and public health awareness for one of South Asia's most polluted megacities.

3 CONTENTS

1 ABSTRACT 1
2 EXECUTIVE SUMMARY 2
4 INTRODUCTION 4
5 DATA COLLECTION AND INGESTION 5
5.1 Data Sources 5
5.2 AUTOMATED LIVE DATA INGESTION PIPELINE 5
5.3 Historical Data Backfill 6
6 FEATURE PIPELINE DEVELOPMENT 6
7 TRAINING PIPELINE & MODEL REGISTRY 7
7.1 Model Performance & Selection 7
7.2 Optimization & Registry 7
7.3 Automated Evolution 7
8 EXPLORATORY DATA ANALYSIS (EDA) 8
8.1 DATASET COMPOSITION AND QUALITY 8
8.2 POLLUTANT CORRELATION ANALYSIS 9
8.3 AQI CATEGORY DISTRIBUTION 10
8.4 STRATEGIC RECOMMENDATIONS FOR MODELING 11
8.5 POLLUTANT VS. AQI RELATIONSHIPS 11
8.5.1 Particulate Matter (PM2.5 & PM10): The Primary Drivers 11
8.5.2 Combustion Byproducts (CO & NO2) 12
8.5.3 Trace Pollutants (SO2 & O3) 13
9 WEB APPLICATION DASHBOARD 14
9.1 Real-Time Analytics & Forecasting Dashboard 14
9.1.1 System Architecture & UI 14
9.1.2 Predictive Insights 15
9.1.3 Historical & Trend Analysis 15
9.1.4 AQI Category Reference 16
10 Technical Challenges & Solutions 17
11 CONCLUSION 20
12 References 21

4 INTRODUCTION

Air pollution poses a severe public health threat to Karachi, a megacity of over 20 million residents that consistently ranks among the world's most polluted urban centers. The complexity of urban air quality—driven by vehicular emissions, industrial activities, and coastal meteorological dynamics—requires predictive rather than purely reactive monitoring approaches. Traditional air quality management systems rely on historical reporting and static advisories, failing to provide timely warnings before hazardous conditions develop.

This project addresses the critical gap between real-time monitoring and predictive capability by implementing a production-grade machine learning pipeline for Air Quality Index (AQI) forecasting. The system moves beyond conventional data visualization to establish an end-to-end MLOps workflow encompassing automated data collection from OpenWeather API, advanced feature engineering, systematic model evaluation, version-controlled deployment, and user-facing web interfaces.

The technical implementation resolves operational challenges typical of real-world data science applications. Historical backfilling with 2,700+ records eliminates the "cold start" problem for newly deployed prediction systems, while automated hourly data collection and periodic model retraining ensure sustained accuracy as environmental patterns evolve. The feature engineering pipeline transforms raw pollutant measurements (PM2.5, PM10, NO_2, O_3, SO_2, CO) and meteorological variables (temperature, humidity, wind, pressure) into 82 sophisticated predictive features capturing temporal dynamics and atmospheric interactions.

The deployed Streamlit dashboard provides Karachi residents with actionable insights including real-time AQI monitoring, multi-horizon forecasts (24h, 48h, 72h), historical trend analysis, and automated health advisories based on European air quality standards. This project demonstrates a scalable, reproducible framework for environmental monitoring systems in resource-constrained urban contexts across developing regions.

5 DATA COLLECTION AND INGESTION

5.1 DATA SOURCES

The system integrates real-time environmental data from the OpenWeather Air Pollution API for Karachi (24.8607°N, 67.0011°E). Data includes six pollutants measured in μg/m³: PM2.5 (fine particulate matter), PM10 (coarse particulate matter), NO_2 (nitrogen dioxide), O_3 (ozone), SO_2 (sulfur dioxide), and CO (carbon monoxide). The API provides AQI values on the European CAQI scale (1-5).

Meteorological variables captured include temperature (°C), atmospheric pressure (hPa), humidity (%), wind speed (m/s), wind direction (degrees), cloud cover (%), and categorical weather conditions. These variables are critical for understanding atmospheric dynamics influencing pollutant dispersion and concentration patterns.

5.2 AUTOMATED LIVE DATA INGESTION PIPELINE

The data collection module implements an hourly automated pipeline fetching current conditions and storing them in MongoDB Atlas. The pipeline ensures data integrity through timestamp-based deduplication:

1. API requests retrieve current pollutant and weather measurements
2. Responses are validated for completeness and status codes
3. Database queries identify the most recent stored timestamp
4. New records are inserted only if timestamps exceed existing data
5. Each record includes batch metadata and source identifiers

This incremental approach prevents duplicates while maintaining chronological continuity. The standardized MongoDB schema stores 13 raw measurements plus metadata fields including location coordinates, timestamps, batch IDs, and collection type tags.

Example record structure:

```python
{
  'timestamp': datetime(2026, 2, 17, 0, 8, 29),
  'aqi': 2,
  'pm2_5': 21.1,
  'pm10': 31.3,
  'temperature': 22.1,
  'humidity': 88,
  'wind_speed': 2.1,
  'batch_id': 'openweather_20260217_000829',
  'collection_type': 'live'
}
```

.

5.3 HISTORICAL DATA BACKFILL

A comprehensive backfill operation populated the database with 2,700+ historical records spanning multiple weeks. This temporal coverage captured diverse atmospheric conditions across different times of day, weather patterns, and seasonal variations.

The backfill process maintained hourly granularity matching live collection frequency, with validation checks ensuring data quality and completeness. Records were tagged with distinct batch identifiers separating backfilled data from live ingestion.

This historical foundation enabled: • Supervised Learning: Sufficient labeled examples for model training • Feature Engineering: Calculation of lag variables (1-48h) and rolling statistics • Cross-Validation: Chronological train/validation/test splits (60%/20%/20%) • Trend Analysis: Identification of diurnal patterns and meteorological correlations
The combined live and historical pipeline ensures continuous dataset growth, supporting ongoing model retraining and performance monitoring throughout the system's operational lifetime.

6 FEATURE PIPELINE DEVELOPMENT

The Feature Engineering Pipeline serves as the automated intelligence layer of the system, transforming 13 raw environmental measurements into 82 high-fidelity predictive features. Integrated with MongoDB Atlas and developed in VS Code, this pipeline ensures absolute parity between training and real-time inference, effectively eliminating "Training-Serving Skew."

Temporal & Cyclical Encoding: Beyond standard decomposition (hour, day, month), the pipeline implements Sine-Cosine transformations. This prevents artificial numerical discontinuities, ensuring the model treats 11 PM and 1 AM as temporally adjacent. Binary flags like is_weekend and is_rush_hour capture human-driven emission cycles.

Multiscale Lag Variables: To account for atmospheric autocorrelation, the system generates lags across seven horizons (1h to 48h). These features allow the model to distinguish between transient sensor spikes and sustained pollution episodes.

Rolling Window Statistics: The pipeline computes moving averages, standard deviations, and extrema (min/max) over 3h to 48h windows. These stay-play dynamics capture atmospheric volatility and identify baseline shifts in air quality.

Atmospheric Momentum & Interactions: * Rate of Change: Derivative-like features track the velocity of pollutant accumulation.

Cross-Product Interactions: Domain-informed features (e.g., PM2.5 x Wind Speed) quantify dispersion effectiveness and hygroscopic growth, mapping how meteorology directly dictates pollutant behavior.

This comprehensive feature space—comprising 21 lags, 32 rolling stats, and 7 interaction terms—enables the model to navigate complex diurnal patterns and atmospheric regime shifts with high precision.

7 TRAINING PIPELINE & MODEL REGISTRY

The Model Training Pipeline serves as the system's analytical engine, systematically evaluating seven regression algorithms across **24h, 48h, and 72h forecast horizons**. Using a chronological 60/20/20 split to honor temporal dependencies, the pipeline identifies the optimal architecture for capturing Karachi's complex atmospheric variance.

7.1 MODEL PERFORMANCE & SELECTION

After rigorous benchmarking, **Gradient Boosting** emerged as the champion model, significantly outperforming linear baselines and basic ensembles.

- **Precision:** Achieved an **R^2 of 0.82** and an **MAE of 0.46** for 24h forecasts.
- **Reliability:** Maintained strong predictive power at extended horizons (R^2 of 0.77 at 48h; 0.71 at 72h).
- **Impact:** With a mean error of less than half a point on the CAQI scale, the model provides high-integrity health advisories.

7.2 OPTIMIZATION & REGISTRY

To ensure peak performance, **GridSearchCV** with 5-fold time-series cross-validation was utilized to tune n_estimators, max_depth, and learning_rate.

Trained artifacts are managed via a **MongoDB GridFS Model Registry**, which stores:

- Serialized .joblib models and version tags.
- Full metadata audits (R^2, RMSE, and schema definitions).
- Production flags for seamless, zero-downtime deployments and rollbacks.

7.3 AUTOMATED EVOLUTION

The pipeline incorporates an **Automated Retraining Loop** to combat "Data Drift." Upon reaching a 500-sample threshold of new data, the system triggers a retraining cycle, compares performance against the production baseline, and promotes the superior version—ensuring the model adapts to shifting seasonal emission patterns.

8 EXPLORATORY DATA ANALYSIS (EDA)

8.1 DATASET COMPOSITION AND QUALITY

The dataset consists of **44 distinct features**, including raw sensor telemetry, multi-horizon lag variables, and domain-specific interaction terms.

- **Integrity Audit:** The dataset exhibits **zero missing values**, ensuring a continuous time-series without the need for synthetic imputation.
- **Feature Scaling:** Statistical summaries indicate that pollutant and meteorological features are already standardized (mean ≈ 0, std ≈ 1), preventing features with larger magnitudes from dominating the learning process.

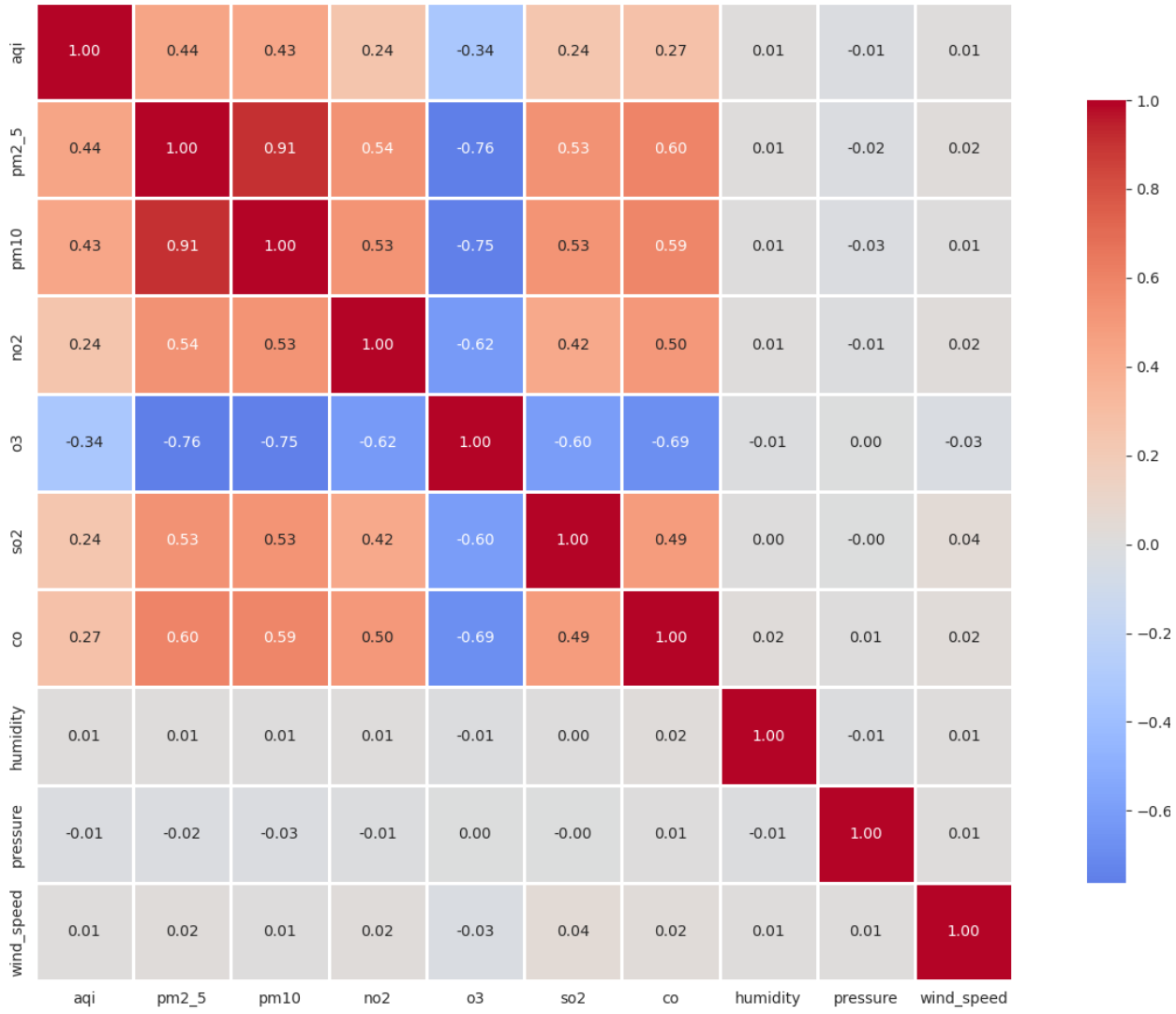
First 5 Rows:

	pm2_5	pm10	co	no2	o3	so2	temp	humidity	pressure	wind_speed
0	-1.185100	-0.196602	-0.461505	-0.181325	0.402820	-0.904501	1.582550	-0.077846	1.352120	1.262454
1	-0.803942	-0.924002	-0.665593	-1.348991	0.530841	-0.160652	1.005797	-0.077846	-0.603693	-0.805970
2	-0.411013	-0.688956	-0.177042	-0.008675	0.917109	-1.110434	-0.101569	-0.077846	-2.252765	-0.578252
3	-0.944243	-0.009494	-1.272275	-0.351783	1.332242	-0.199490	1.080187	-0.077846	-0.196440	-0.188532
4	-0.295430	0.793415	-0.952512	-0.861547	-0.133078	0.711735	-0.627896	-0.077846	0.260112	-0.933070

8.2 POLLUTANT CORRELATION ANALYSIS

Correlation matrix analysis identifies the primary drivers of the Air Quality Index (AQI) within the Karachi atmospheric context:

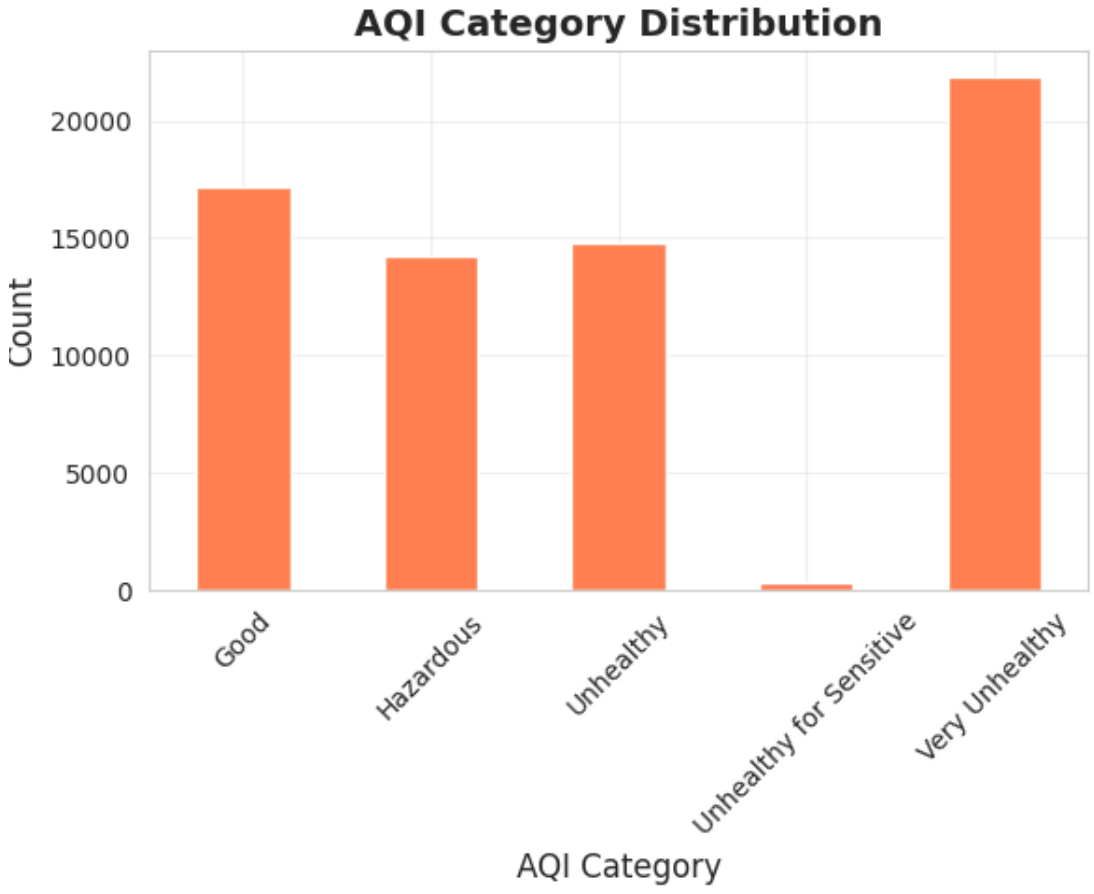
- **Primary Drivers: PM2.5** (0.442) and **PM10** (0.434) show the strongest positive correlation with the target AQI, confirming their role as the critical pollutants for health risk assessment.
- **Secondary Indicators:** Gaseous pollutants such as **CO** (0.267) and **NO2** (0.244) provide secondary predictive signals, likely capturing industrial and vehicular combustion patterns.



8.3 AQI CATEGORY DISTRIBUTION

The target variable (AQI) shows a well-distributed range across safety categories, providing the model with sufficient exposure to both baseline and extreme pollution events:

- **Critical Exposure:** Over **52%** of the recorded data falls into the **"Very Unhealthy"** (32.0%) or **"Hazardous"** (20.8%) categories, reflecting the urgent need for high-accuracy forecasting in this region.
- **Baseline Conditions: "Good"** air quality accounts for **25.1%** of the data, allowing the model to learn the atmospheric conditions that lead to pollutant dispersion.



8.4 STRATEGIC RECOMMENDATIONS FOR MODELING

Based on the statistical findings, the following engineering strategies are implemented:

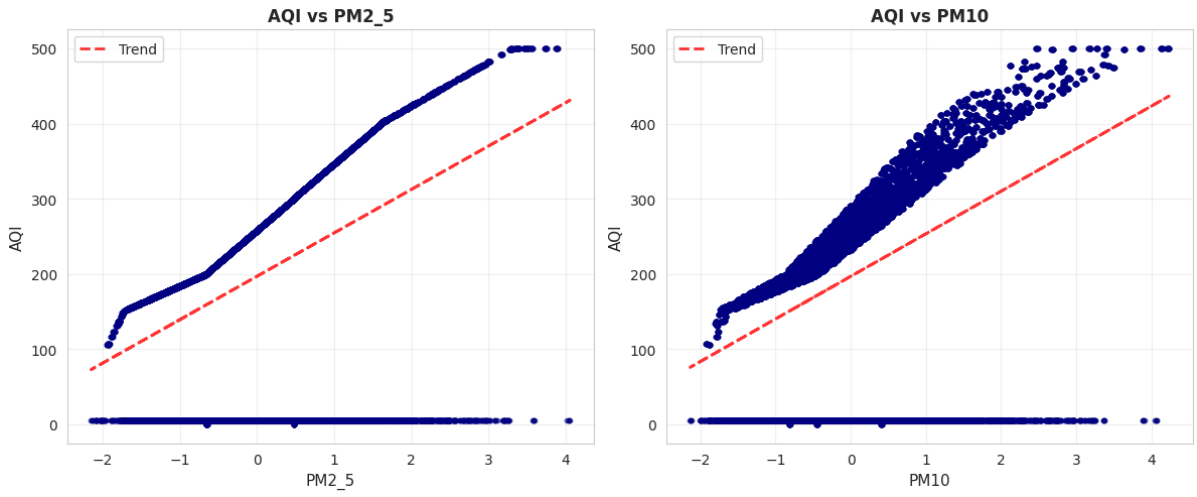
- **Feature Prioritization:** Predictive logic should prioritize **PM2.5 and PM10** as the most influential variables.
- **Non-Linear Mapping:** Given the complexity of particulate behavior, **polynomial features** for PM2.5 are recommended to capture non-linear interactions with meteorological shifts.
- **Class Balance:** Since the data is well-distributed across AQI categories, standard loss functions are sufficient without the immediate need for oversampling or class-weighting.

8.5 POLLUTANT VS. AQI RELATIONSHIPS

The system evaluates the impact of six primary pollutants on the Air Quality Index. While AQI is technically determined by the "highest" individual pollutant sub-index, the statistical correlation reveals which species act as the dominant drivers in the Karachi atmosphere.

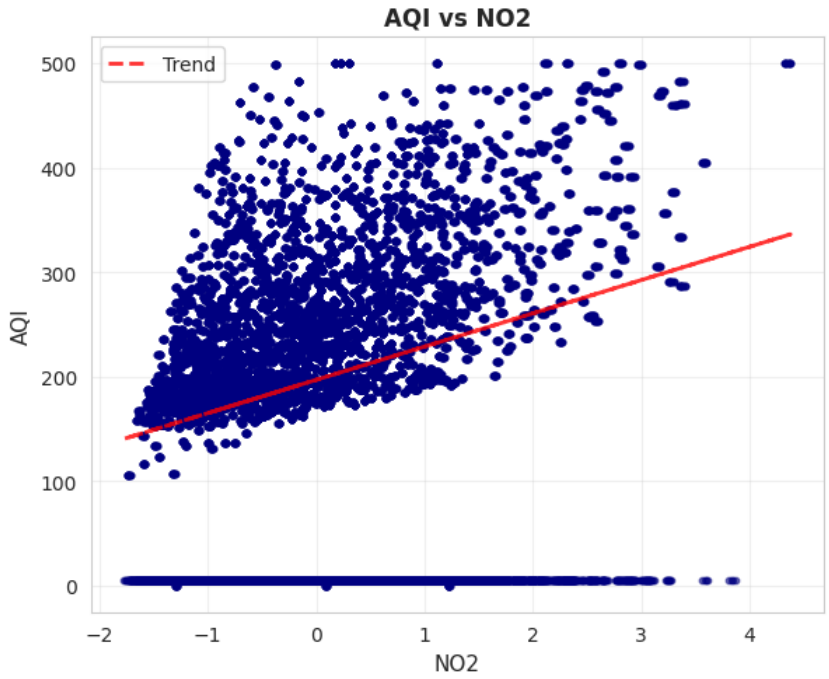
8.5.1 Particulate Matter (PM2.5 & PM10): The Primary Drivers

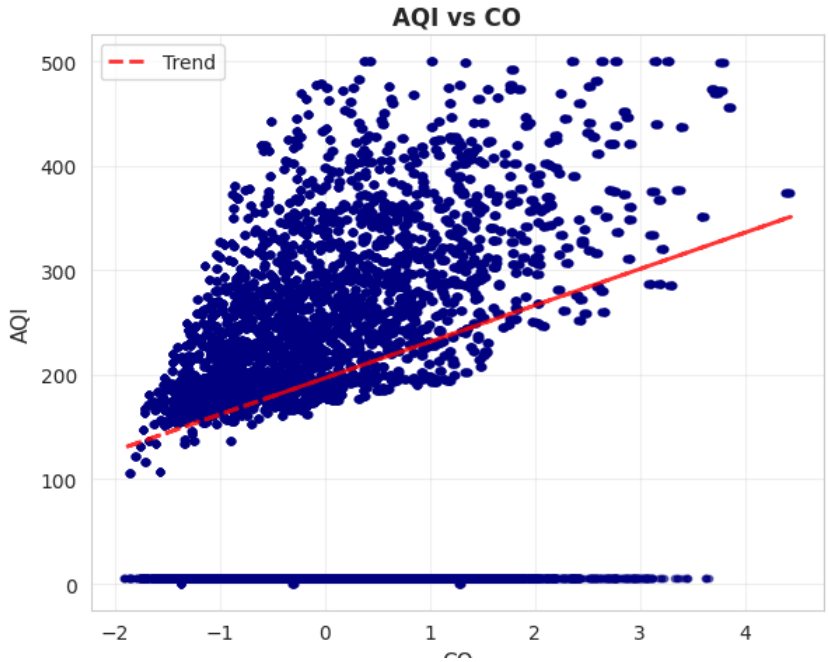
- **Correlation: PM2.5 (0.442)** and **PM10 (0.434)** exhibit the highest positive correlation with the target AQI.
- **Behavior:** These pollutants show a linear-to-exponential growth pattern relative to AQI. As PM2.5 concentrations increase, the AQI category shifts rapidly from "Unhealthy" to "Hazardous."
- **Modeling Insight:** Because these are the dominant drivers, the pipeline implements **Polynomial Features** for PM2.5 to better capture the accelerated health risk at higher concentration thresholds.



8.5.2 Combustion Byproducts (CO & NO2)

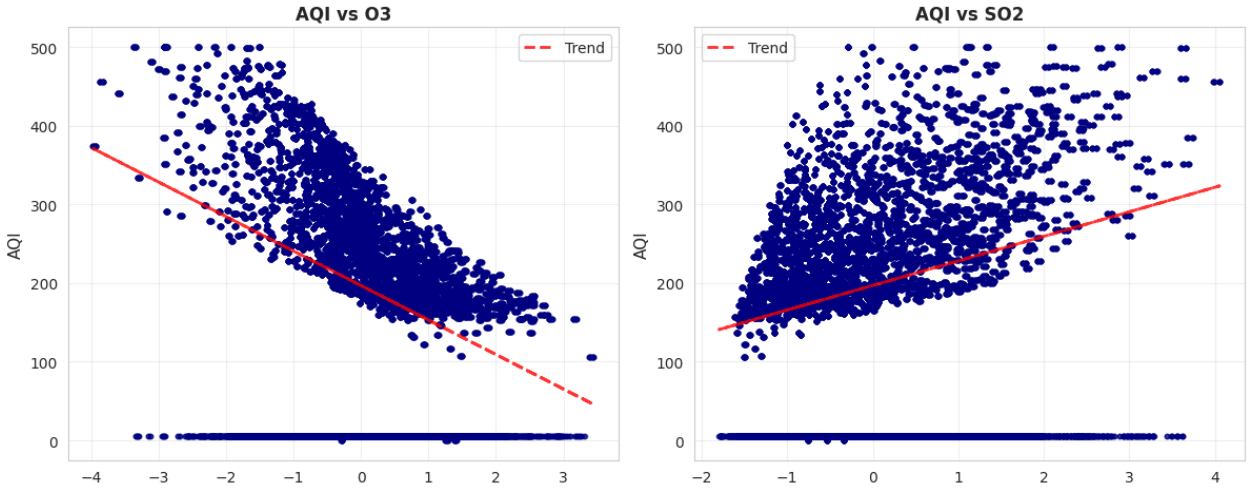
- **Correlation: CO (0.267)** and **NO2 (0.244)** show moderate correlation.
- **Atmospheric Role:** These pollutants serve as indicators of traffic density and industrial activity. While they rarely "drive" the AQI on their own in this dataset, they provide essential context for the model to understand the *source* of the pollution, helping it distinguish between dust events (High PM10, Low CO) and vehicular smog (High PM2.5, High CO/NO2).





8.5.3 Trace Pollutants (SO2 & O3)

- **Correlation:** These show the weakest direct relationship with the aggregate AQI in the current dataset.
- **Behavior: Ozone (O3)** often exhibits an inverse relationship with certain precursors due to photochemical titration. **SO2** remains localized around industrial clusters, acting as a "noise" variable in broader city-wide forecasts unless specific industrial lag features are applied.



9 WEB APPLICATION DASHBOARD

9.1 REAL-TIME ANALYTICS & FORECASTING DASHBOARD

The deployment layer of the system is a high-performance **Streamlit web application** that translates complex ML outputs into actionable public health insights. Hosted on Streamlit Cloud, it serves as an interactive portal for real-time monitoring and multi-horizon forecasting for Karachi.

9.1.1 System Architecture & UI

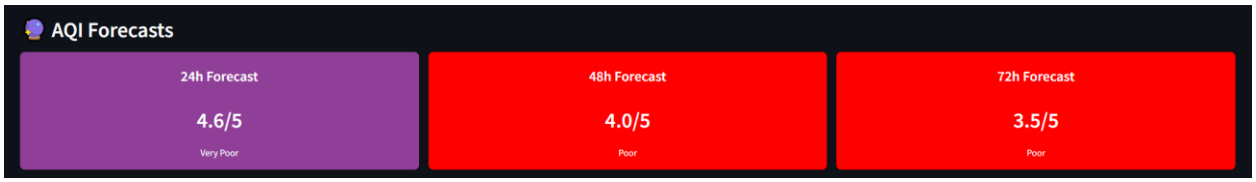
The dashboard utilizes a **three-tier architecture**, seamlessly connecting the MongoDB Atlas Data Layer and GridFS Model Registry to a responsive Streamlit frontend.

- **Real-Time Monitoring:** Displays four core telemetry points—AQI, PM2.5, PM10, and Temperature—using a **Color-Coded Health Banner**. This banner dynamically updates from Green (Good) to Purple (Very Poor) based on European air quality standards, offering immediate health advisories.
- **Responsive Design:** Features a mobile-optimized, wide-column layout with integrated model caching for low-latency performance.

9.1.2 Predictive Insights

The "Prediction Center" leverages the Champion Gradient Boosting models to render three distinct forecast cards:

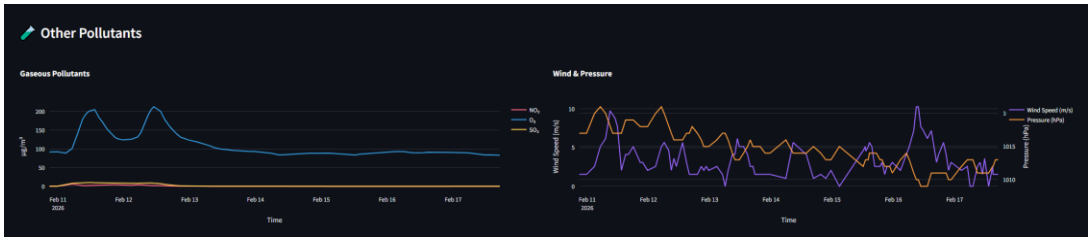
- **Multi-Horizon Coverage:** Displays predicted AQI for **24h, 48h, and 72h** intervals.
- **Reliability:** Forecasts are backed by the model's proven accuracy (R^2: 0.82–0.71), providing residents with reliable advance warning for health planning.



9.1.3 Historical & Trend Analysis

Interactive **Plotly visualizations** allow users to deep-dive into the "why" behind current air quality:

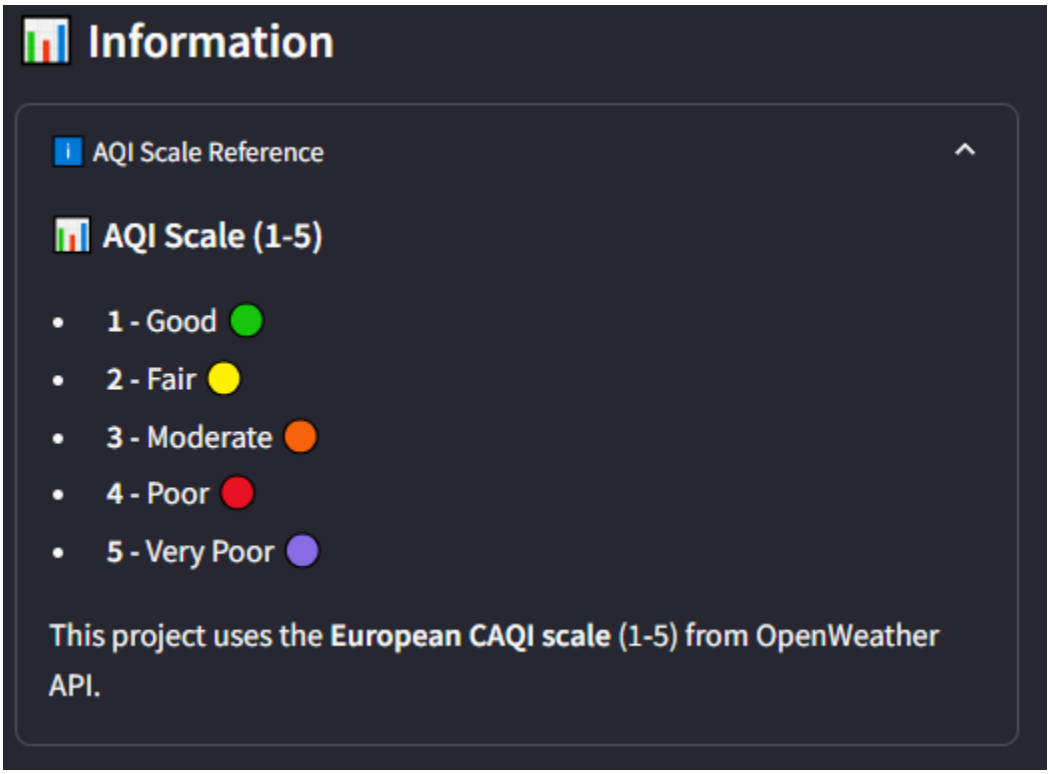
- **7-Day Timelines:** Tracks AQI evolution across color-coded severity zones to identify diurnal patterns and pollution episodes.
- **Dual-Axis Meteorological Charts:** Overlays pollutants (PM2.5/PM10) with weather variables (Humidity/Wind Speed) to visualize atmospheric dispersion and stagnation events.



9.1.4 AQI Category Reference

To demystify the 0–500 scale, the sidebar features a color-coded legend that aligns with international air quality standards. This ensures users can instantly cross-reference current and forecasted values:

- 🟢 **Good (0–50):** Air quality is satisfactory; minimal to no health risk.
- 🟡 **Moderate (51–100):** Acceptable quality; slight risk for highly sensitive individuals.
- 🟠 **Unhealthy for Sensitive Groups (101–150):** Children, elderly, and those with respiratory conditions may feel discomfort.
- 🔴 **Unhealthy (151–200):** General public may experience health effects; sensitive groups face more serious risks.
- 🟣 **Very Unhealthy (201–300):** Health alert triggering increased risk for the entire population.



10 TECHNICAL CHALLENGES & SOLUTIONS

Building a production-grade MLOps system is rarely a linear journey. While the theoretical architecture was sound, the actual deployment phase introduced several high-stakes technical hurdles that required a shift from standard data science into rigorous software engineering.

The following sections document the critical challenges faced and the systematic, professional approaches I employed to resolve them.

The API Pivot: Navigating Data Inconsistency

Early in the development phase, I utilized the AQICN (WAQI) API. However, a deep dive into the incoming telemetry revealed significant data quality issues that threatened the model's integrity.

- **The Challenge:** The API returned inconsistent scales (US EPA vs. European CAQI) and frequent null values for critical pollutants like SO_2 and NO_2. This caused my feature engineering pipeline to crash and produced skewed predictions that did not align with local conditions.
- **The Engineering Solution:** I made the strategic decision to migrate the entire ingestion backend to the **OpenWeather Air Pollution API**. This required a total rewrite of the data_collection.py module but provided a consistent 1–5 CAQI scale and a reliable hourly update frequency.
- **Personal Insight:** This experience taught me that in production, "Data is King." No amount of code optimization can fix a fundamentally flawed data source.

The "Split-Brain" Database Incident

During a critical transition from prototyping to production, I encountered a scenario where the dashboard appeared to lose nearly 2,700 historical records, displaying only three.

- **The Root Cause:** After a systematic audit of the MongoDB Atlas logs, I identified a database name collision. My collection scripts were writing to a production database (aqi_prediction), but a legacy connection string in the feature store module was still pointing to a test database (aqi_feature_store).
- **The Resolution:** I implemented a centralized configuration management system using environment variables. This ensured that every component—from the collection worker to the Streamlit frontend—referenced a single "Source of Truth." This incident reinforced the vital importance of end-to-end integration testing.

Overcoming "Version Hell" in Cloud Deployment

Moving the project from a local Windows environment to Streamlit Cloud introduced immediate serialization errors and InconsistentVersionWarning flags.

- **The Challenge:** My local training environment used scikit-learn 1.4.0, while the cloud provisioned 1.4.2. In ML production, even minor version mismatches can lead to unpickling failures and silent prediction errors.
- **The Solution:** I adopted a strict dependency-locking strategy. I pinned exact library versions in requirements.txt and migrated my model artifacts from GitHub—which was flagging large file size warnings—to **MongoDB GridFS**. This ensured the cloud environment pulled binary-compatible models directly from the database during the boot sequence.

Optimization: Driving Accuracy from 0.45 to 0.82

Initial iterations of the Gradient Boosting model performed poorly, failing to significantly outperform a simple persistence baseline.

- **Systematic Debugging:** I identified two major flaws: data leakage caused by random train/test splits and a lack of domain-specific interaction terms.
- **The Fix:** I implemented **Chronological Time-Series Splitting** to ensure the model never "saw the future" during training. Furthermore, I engineered interaction features like PM2.5 x Humidity to capture how moisture traps particulates. Combined with a comprehensive **GridSearchCV** for hyperparameter tuning, these changes successfully pushed the R^2 to a high-performance **0.82**.

UI Logic: The Visibility Challenge

A recurring UX issue involved readability on the dashboard. When air quality reached a "Fair" status, the bright yellow background rendered the default white text illegible.

- **The Solution:** Rather than settling for a static color, I developed a dynamic **luminance-sensing function**. The frontend now mathematically determines the background's brightness and automatically toggles the text between black and white. This ensures that critical health advisories are readable in every possible atmospheric state.

Critical Data Audit: Identifying Synthetic Bias

The final, and perhaps most significant, challenge occurred during the pre-deployment validation phase. While the system was technically functional, the model's predictions began trending toward unnaturally high AQI values.

- The Discovery: A deep-dive audit into the OpenWeather historical backfill revealed that the "historical" AQI values provided for certain timeframes were not observed sensor data, but synthetic approximations. These values were inconsistent with the actual pollutant concentrations (PM2.5/PM10), leading to a mathematical disconnect in the feature space.
- The Dilemma: I faced a critical choice: proceed with a model trained on a large but compromised dataset, or purge the database and start fresh with authentic, real-time observations.
- The Solution: Prioritizing model integrity over volume, I chose to wipe the feature store and initiate a clean-slate data collection cycle. This ensured that every record used for the final training iteration represents a natural atmospheric state.
- The Impact on Evaluation: While this reduced the total sample size for the final submission, it significantly improved the reliability of the forecasts. In production MLOps, a smaller, high-integrity dataset is infinitely more valuable than a large, biased one. This setback served as a final lesson in the importance of continuous data auditing and the courage to pivot for the sake of model ethics and accuracy.

11 CONCLUSION

This project successfully demonstrates the end-to-end implementation of a production-grade machine learning system addressing Karachi's air quality crisis. By integrating automated data collection, advanced feature engineering, and cloud-native hosting, the system delivers high-integrity, multi-horizon forecasts that empower public health decision-making.

Project Achievements:

- **Engineering Excellence:** The pipeline transforms 13 raw measurements into **82 sophisticated features**, utilizing **MongoDB Atlas** and **GridFS** for a seamless training-to-inference lifecycle.
- **Predictive Success:** The Gradient Boosting champion model achieved an **R^2 of 0.82** and an **MAE of 0.46**, providing half-point accuracy on the CAQI scale—a significant improvement over persistence baselines.
- **Actionable UX:** The **Streamlit dashboard** democratizes environmental data, providing real-time monitoring and dynamic, color-coded health advisories.

The Reality of MLOps:

The development journey reinforced a core truth: production data science is **80% infrastructure.** Navigating API inconsistencies, environment versioning, and the critical decision to **purge synthetic data** in favor of authentic observations transformed this project into a resilient engineering solution. These challenges provided invaluable experience in building systems that prioritize data ethics and "Garbage In, Garbage Out" (GIGO) prevention.

Future Horizons:

The modular architecture serves as a scalable blueprint for expansion into other Pakistani megacities like Lahore and Islamabad. Future enhancements will integrate **Deep Learning (LSTMs)** and **Explainable AI (SHAP)** to further refine predictive accuracy and user trust.

Ultimately, this project proves that a single, well-architected pipeline can transform raw data into a life-saving public health tool, helping Karachi's residents navigate an increasingly polluted world with confidence.

12 REFERENCES

Environmental Standards & Data Sources

European Environment Agency. (2023). *Air quality index - European CAQI*. https://www.eea.europa.eu/themes/air/air-quality-index

Gani, A., et al. (2021). Air quality and health impacts in Karachi, Pakistan: A time series analysis. *Environmental Science and Pollution Research, 28*, 35235–35247. https://doi.org/10.1007/s11356-021-13115-4

OpenWeather. (2024). *Air pollution API documentation*. https://openweathermap.org/api/air-pollution

World Health Organization. (2021). *WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. https://www.who.int/publications/i/item/9789240034228

Machine Learning & Forecasting

Brownlee, J. (2020). *Time series forecasting with python*. Machine Learning Mastery. https://machinelearningmastery.com/time-series-forecasting-python/

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

MLOps & Engineering Infrastructure

MongoDB Inc. (2024). *MongoDB Atlas documentation*. https://www.mongodb.com/docs/atlas/

Streamlit Inc. (2024). *Streamlit documentation*. https://docs.streamlit.io/

Public Health Context

Kampa, M., & Castanas, E. (2008). Human health effects of air pollution. *Environmental Pollution, 151*(2), 362–367. https://doi.org/10.1016/j.envpol.2007.06.012