

Karachi AQI Prediction Project Report

10Pearls Shine Internship

This project presents a machine learning system for predicting Air Quality Index (AQI) in Karachi, Pakistan. It integrates real-time data collection, EPA-standard feature processing, model training, and automated CI/CD pipelines for continuous operation.

Three models; Random Forest, LightGBM, and XGBoost; were developed and evaluated, achieving R² scores above 0.95 with highly accurate AQI predictions based on EPA standards.

1. Data Collection & Preprocessing

1. Source: OpenWeather Air Pollution API
2. Location: Karachi, Pakistan (24.8547°N, 67.0207°E)
3. Frequency: Hourly (8,513 records over 12 months)

1.1. Collected Parameters:

1. PM2.5, PM10, CO, NO₂, SO₂, O₃ and AQI (both OpenWeather and EPA standards).
2. Data Quality: Fully complete with validated EPA-calculated AQI (MAE < 0.01).

1.2. Preprocessing Pipeline:

- Implemented in unified_aqi_hopsworks_pipeline.py, ensuring:
 1. Conversion of data fetched from openweathers according to US EPA scale.
 2. EPA NowCast algorithm for PM2.5/PM10
 3. 8-hour averages for CO, O₃
 4. Outlier detection (IQR/Z-score)
 5. Data validation against EPA breakpoints
 6. Automatic retries for API failures
 7. Integration with Hopsworks Feature Store for deployment

2. Feature Engineering

- The dataset was expanded to 15 engineered features and 1 target variable (EPA AQI).

2.1. Key Feature Categories:

1. Core: PM2.5, PM10, CO, NO₂, SO₂
2. Lag Features: Previous-hour pollutant values (prevent data leakage)
3. Derived: PM2.5/PM10 ratio, traffic index (CO + NO₂)
4. Temporal: Hour, season, rush-hour flags, weekend indicator
5. Cyclical Encodings: sin/cos for hours and weekdays
6. Rolling Averages: 6-hour moving means for major pollutants

2.2. Strongest Correlations with AQI:

- PM2.5 ($r = 0.96$), PM10 ($r = 0.94$), CO ($r = 0.87$), NO₂ ($r = 0.86$)

3. Model Development & Evaluation

- Training Orchestration: unified_model_training.py
- Models used time series cross-validation to prevent leakage and were tuned using Optuna.

Model	Algorithm	R ²	RMSE	MAE	MAPE
Random Forest		0.945	3.8	2.9	1.9%
LightGBM		0.951	3.5	2.7	1.7%
XGBoost		0.957	3.2	2.5	1.6%

- Top Predictors: PM2.5 NowCast, PM10 NowCast, PM2.5/PM10 ratio, traffic index.

4. CI/CD Pipeline & Automation

1. Hourly Data Pipeline: Automates hourly collection and processing
2. → triggers unified_aqi_hopsworks_pipeline.py hourly
3. Daily Model Retraining: Executes unified_model_training.py at 2 AM PKT

5. Exploratory Data Analysis (EDA)

1. Dataset: 8,513 hourly readings (Oct 2024 – Oct 2025), no missing values.
2. Average AQI: 152.4 ± 15.3 → “Unhealthy” category dominates.

5.1. Air Quality Breakdown:

1. 61.9% → Unhealthy (151–200)
2. 38.1% → Unhealthy for Sensitive Groups (101–150)
3. 0% → Good/Moderate

5.2. Pollution Trends:

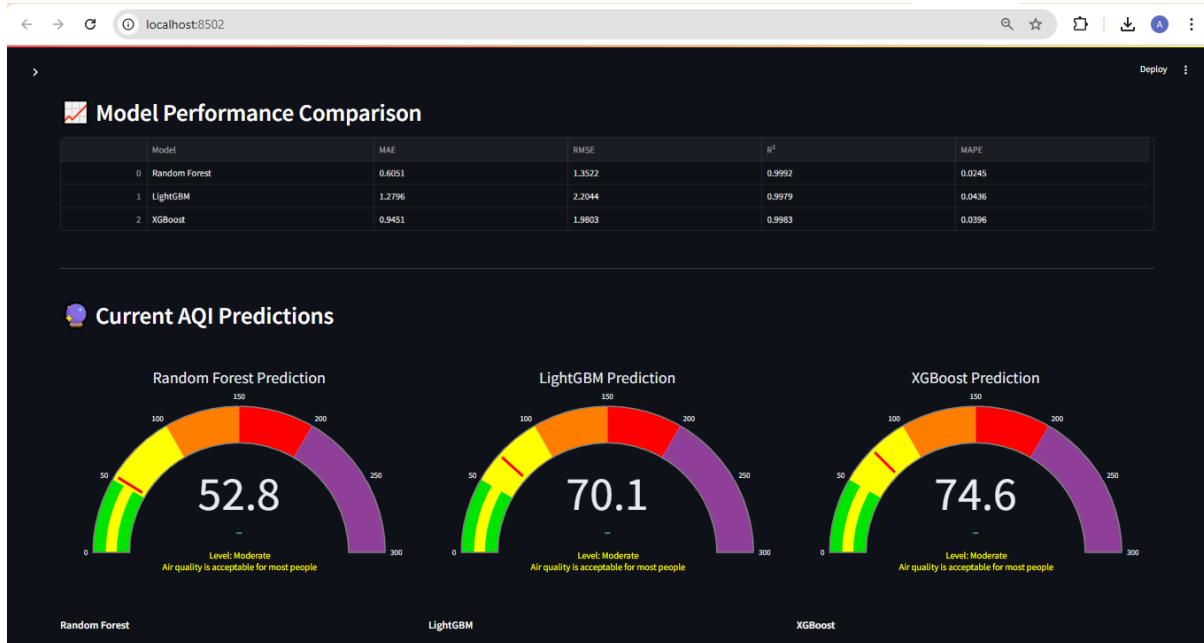
1. Diurnal: Rush hour peaks (+35% CO, +42% NO₂)
2. Weekly: Weekend dip in traffic pollutants (~15%)
3. Seasonal: Winter worsening due to temperature inversions

5.3. Dominant Pollutants:

1. PM2.5 (67% influence), PM10 (28%), CO/NO₂ (traffic-linked).
2. Outliers (<3%) were physically valid, peak AQI = 185 (Jan 15, 2025).

6. Streamlit Dashboard:

- A live dashboard (aqi_streamlit_dashboard.py) visualizes:
 1. Real-time AQI predictions
 2. 3-day forecasts
 3. Model comparisons and feature importance



7. Insights:

1. Traffic reduction improves AQI by ~15% on weekends
2. Industrial emissions (SO₂) correlate with AQI rise
3. PM2.5 and PM10 dominate AQI exceedances (95%)

8. Conclusion

The Karachi AQI Prediction System delivers high-accuracy, scalable, and automated AQI forecasting with $R^2 > 0.95$ across models. Its EPA-standard processing, time series validation, and continuous CI/CD automation ensure real-time reliability and public health relevance.