Internship Project Report: AQI Prediction System

Submitted by: Syeda Laiba Rehman

Organization: 10Pearls

---

1. Executive Summary

Developed an Air Quality Index (AQI) Prediction System for Karachi that provides 3 days prediction with an accuracy of **96.6%** ($R^2 = 0.966$). The project integrates real-time data collection using the Open-Meteo API, machine learning modeling using **LightGBM,** and an interactive Streamlit dashboard for real-time predictions. An automated data pipeline and Hopsworks feature store ensure smooth data flow and consistent model performance between training and production.

2. Methodology & Implementation

2.1 Data Pipeline & Preprocessing

- **Data Source:** Used Open-Meteo APIs (Air Quality + Historical Weather) for Karachi coordinates
- **Period:** From September 2023 - Present (18,000+ hourly records)
- **Parameters Fetched:** PM2.5, PM10, CO2, Ozone, No2, So2, US AQI, Temperature, Humidity, Wind Speed
- **EDA:** Analyzed dataset structure, missing values, and distributions across 18,000+ hourly records, generated correlation heatmaps and temporal trend visualizations.
- **Preprocessing:** Handled outliers (IQR method), missing values (interpolation), and corrected skewness (Yeo-Johnson transformation), Data Validation **(**duplicate removal and timestamp consistency checks)

2.2 Feature Engineering & Selection:

- **Engineered Features:** Temporal features (hour, day, month), rolling averages (3H, 6H, 24H), lag features, and pollutant ratios.
- **Feature Store:** Data was versioned and managed in **Hopsworks** (offline and online feature groups), v1 (offline) for historical batch training and validation, and v2 (online) for daily dashboard predictions using the latest data.

- **Selection:** Finalized **15 key features** from an initial set of 46 using Random Forest importance and correlation analysis.

3. Model Training & Selection

Random Forest, XGBoost, and LightGBM were trained using Multi-Output Regression for multi-horizon AQI forecasting

3.1 Model Comparison Before Hyper-Parameter Tuning

Model	Train R^2	Test R^2	Train MSE	Test MSE	Train MAE	Test MAE
Random Forest	0.9936	0.9565	0.0064	0.0437	0.0438	0.1162
XG Boost	0.9663	0.9363	0.0337	0.0635	0.1328	0.1791
Light GBM	0.9551	0.9265	0.0449	0.0733	0.1556	0.1947

Before hyperparameter tuning, **Random Forest** achieved the highest accuracy **(95%)** but showed overfitting due to the large gap between training and testing errors. To address this, three-fold cross-validation and hyperparameter tuning were applied to enhance generalization and stability across models.

3.2 Hyper-Parameter Tuning Results

After tuning all models, **LightGBM** achieved the highest cross-validation R^2 score (0.9516) and was therefore **selected as the final model**. It was optimized using **GridSearchCV with three-fold cross-validation** and showed minimal overfitting, with a **train-test R^2 difference of only 0.0265**, indicating strong generalization performance.

4. SHAP Analysis & Model Interpretability

SHAP analysis was conducted to interpret the LightGBM model's predictions and validate its decision-making logic across the 3-day forecast horizon.

4.1 Key Findings by Forecast Horizon:

- **24-Hour Forecast:** Dominated by real-time pollution levels **(pm 2.5, us_aqi)**
- **48-Hour Forecast:** Balanced influence of weather patterns **(temp_roll24)** and seasonal trends (month)

- **72-Hour Forecast:** Primarily driven by temporal patterns **(month)** and meteorological forecasts

This progressive shift from pollutant-driven to pattern-based predictions aligns with atmospheric science principles, confirming the model's scientific validity and reliability for multi-day forecasting.

5. Dashboard Implementations & Automation

5.1 Streamlit Dashboard Features:

- Displays real-time AQI forecasts (24H, 48H, 72H) with color coded health indicators
- Interactive charts with SHAP-based feature importance
- Health and activity recommendations based on AQI levels

5.2 Automated Pipeline (GitHub Actions):

The automated pipeline **runs daily** to **fetch**, **clean**, **transform**, and **engineer new data**, **saving** the **features to** the **Hopsworks** online feature group so that the dashboard can access updated data for daily forecasting.

5.3 Prediction Workflow

Dashboard Request → Hopsworks Online FG (Latest 24H data) → Light GBM Model → 72H Predictions → Streamlit Visualization

6. Results & Key Learnings

This project successfully developed an automated AQI Prediction System for Karachi, achieving 96.6% R^2 accuracy. The optimized LightGBM model, selected for its superior generalization, was integrated into a complete MLOps pipeline capable of generating daily forecasts without manual intervention. The end-to-end implementation encompassed the entire machine learning lifecycle, from data ingestion and feature engineering to model validation with SHAP analysis ensuring transparent, interpretable predictions. This work demonstrates how machine learning combined with MLOps principles can create scalable, data-driven foundations for effective air quality management.

7. Repository Link:

https://github.com/rsyedalaiba/AQI_Project