End-to-End MLOps Project Report

Project: Pearls AQI Predictor

Author: Uzair A. Jokhio

Date: November 9, 2025

1. Executive Summary

This project built a fully automated **MLOps pipeline** to predict the **Air Quality Index (AQI)** for **Karachi** over the next **3 days**.
The system automatically collects air pollution and weather data every hour, cleans it, and retrains five machine learning models daily using **GitHub Actions**.

All trained models and performance results are saved in the **Hopsworks Model Registry**. A **Streamlit dashboard** connects to this registry, downloads the latest models, and shows live forecasts using data from **OpenWeatherMap**.
Users can compare all models and see which gives the most accurate AQI predictions.

The **Gradient Boosting model** performed best, achieving an **R^2 of 0.9941** and a **Mean Absolute Error (MAE)** of **0.77 AQI points**.

---

2. Exploratory Data Analysis (EDA) & Feature Selection

EDA was used to understand how pollutants and weather affect AQI. The process was so useful that it was added as a page in the Streamlit app.

Key Findings:

- The AQI from OpenWeather was too simple (scale 1–5), so we calculated our own **continuous AQI** using U.S. EPA standards.
- **Main pollutants:** *PM2.5* and *PM10* had the strongest effect on AQI.
- **Time patterns:**
  - NO_2 and *CO* rose during traffic hours (6–9 AM, 5–8 PM).
  - O_3 peaked in the afternoon (1–3 PM).
  - Therefore, **hour_of_day** was a key feature.
- **Correlation results:**
  - AQI increased with *PM2.5* and *PM10*.
  - AQI decreased with higher *wind_speed*.

Final Features Used for Modeling:

1. PM2.5
2. PM10
3. O_3
4. Temperature
5. Hour of the day
6. Day of the month

3. Model Evaluation & Selection

Models Used

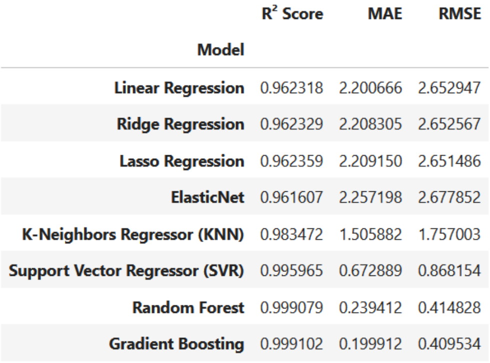
The pipeline's **model_training.yml** workflow runs daily, using the latest clean data to train five regression models:

1. **Ridge Regression** – Linear model
2. **KNeighborsRegressor (KNN)** – Distance-based
3. **Support Vector Regressor (SVR)** – Non-linear
4. **RandomForestRegressor** – Ensemble
5. **GradientBoostingRegressor** – Ensemble

Evaluation Process

All models followed the same testing process:

- **Data split:** 80% training, 20% testing
- **Scaling:** Used one shared *StandardScaler* for all models
- **Metrics:**
    - **R^2 (R-squared):** Shows how much variance is explained (1.0 = perfect).
    - **MAE (Mean Absolute Error):** Shows how many AQI points predictions were off by.



	R^2 Score	MAE	RMSE
Model			
Linear Regression	0.962318	2.200666	2.652947
Ridge Regression	0.962329	2.208305	2.652567
Lasso Regression	0.962359	2.209150	2.651486
ElasticNet	0.961607	2.257198	2.677852
K-Neighbors Regressor (KNN)	0.983472	1.505882	1.757003
Support Vector Regressor (SVR)	0.995965	0.672889	0.868154
Random Forest	0.999079	0.239412	0.414828
Gradient Boosting	0.999102	0.199912	0.409534

Best Model
Results were stored in the **Hopsworks Model Registry** and shown in the Streamlit app.
The **Gradient Boosting model** performed best, with the **highest R^2** and **lowest MAE**.
It was chosen as the final model, but users can still compare all five in the dashboard.

4. Final Evaluation & Limitations

Final Evaluation
The project met all goals. It built a fully automated pipeline that collects data, creates features, and retrains five models daily. All models are stored in the Hopsworks registry, and the Streamlit app displays their forecasts and performance in a clear, interactive dashboard.

Dashboard Features:

- **3-Day Forecast:** Shows predicted average and peak AQI for each day.
- **Model Comparison:** 72-hour forecast line chart for all models.
- **Performance Tab:** R^2, MAE, and feature importance for top models.
- **EDA Page:** Interactive analysis of historical data.

Limitations:

1. **Dependence on OpenWeather Data:** Model accuracy relies on the quality of OpenWeather's pollutant forecasts.
2. **Demo Model (Data Leakage):** The model predicts AQI using pollutant data from the same hour, giving an unrealistically high R^2. True forecasting would use future targets (T+3).
3. **Single-City Model:** Trained only on Karachi data may not perform well for other cities