

10 PEARLS AQI PREDICTION PROJECT REPORT

By Rezan Sohail

Tools Used: Python, Streamlit, Hopsworks, GitHub Actions, Google Colab, VS Code

Project Overview

The goal of this project is to predict the Air Quality Index (AQI) in Karachi for the next 3 days using a fully serverless, automated pipeline. The system fetches real-time weather and pollutant data, computes derived features, trains multiple ML models, and provides an interactive dashboard for monitoring AQI trends.

Data and Feature Pipeline

Fetches hourly data from **Open-Meteo APIs** for pollutants (PM2.5, PM10, CO, NO2, SO2, O3) and weather (temperature, humidity, pressure, wind).

Merged datasets and computed derived features:

AQI index (mean of pollutants)

AQI change rate

Time features: hour, day, month, day of week

Stored features in **Hopsworks Feature Store** (karachi_aqi_pollution_history).

Historical backfill performed to cover past 3 months.

Model Training Pipeline

Models trained: **Linear Regression, Random Forest, Gradient Boosting**.

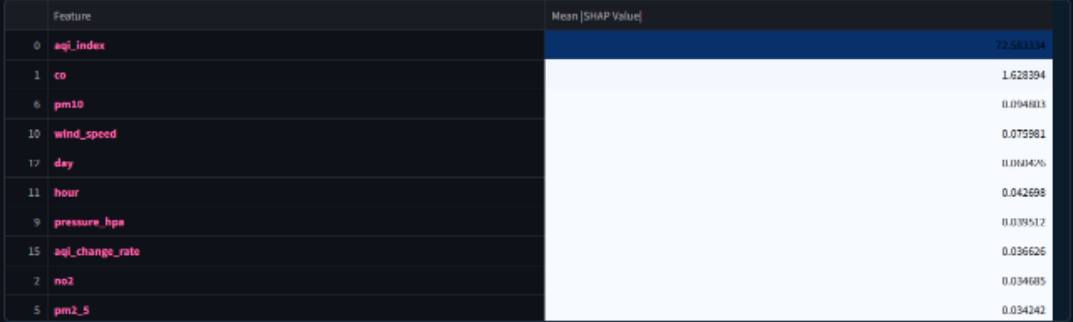
Time-based train-test split (80%-20%).

Evaluation metrics: **MAE, RMSE, R²**.

Model	MAE	RMSE	R ²
Random Forest	15.2	22.4	0.92
Gradient Boosting	17.1	24.5	0.89
Linear Regression	22.3	30.1	0.82

Random Forest selected as best: handles non-linear patterns, outliers, and feature interactions.

SHAP used for feature importance; top features: PM2.5, PM10, CO, temperature.



The image shows a screenshot of a web application titled "Random Forest Feature Importance (SHAP)". It displays a table with two columns: "Feature" and "Mean |SHAP Value|". The table lists 15 features, with "aqi_index" having the highest importance value of 72.581394. Other features include "co", "pm10", "wind_speed", "day", "hour", "pressure_hpa", "aqi_change_rate", "no2", and "pm2_5".

	Feature	Mean SHAP Value
0	aqi_index	72.581394
1	co	1.628394
6	pm10	0.07048013
10	wind_speed	0.075981
17	day	0.00004076
11	hour	0.042698
9	pressure_hpa	0.039512
15	aqi_change_rate	0.036626
7	no2	0.034685
5	pm2_5	0.034242

Web Dashboard

Built with **Streamlit**.

Features:

- Hourly AQI forecast for 3 days.
- Color-coded AQI alerts (Hazardous → Excellent).

Interactive plots: actual vs predicted AQI.

Summary metrics (MAE, RMSE, R^2).

SHAP feature importance table.

Fetches **real-time AQI** from Hopsworks.

Automation & CI/CD

GitHub Actions automate:

Feature script runs **hourly**.

Training script runs **daily**.

Ensures pipeline is **scalable and fully automated**.

Conclusion

Successfully implemented **end-to-end AQI prediction system**.

Random Forest provides most accurate forecasts.

Dashboard shows **real-time alerts and trends**.

Future improvements: deep learning models, multi-city predictions, cloud deployment.