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

Fetched 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.

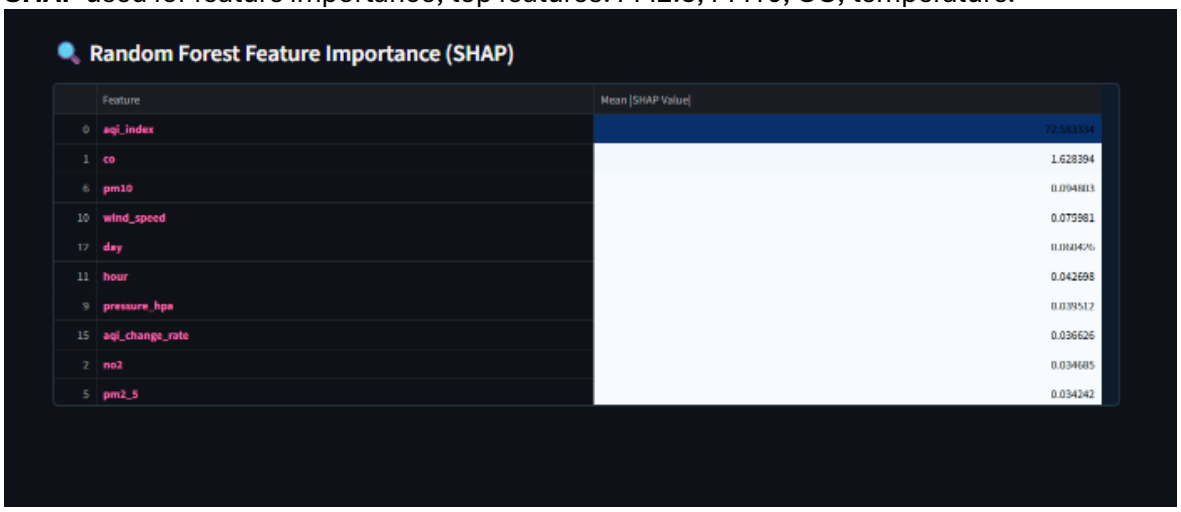
Model Training Pipeline

Models trained: **Linear Regression, Random Forest, Gradient Boosting**.
Time-based train-test split (80%-20%).
Evaluation metrics: **MAE, RMSE, R^2**.

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.



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.