

# Report: “Pearls AQI Predictor Project Report”

**Student:** Mahnoor Asim

**Project:** 10Pearls Shine (Pearls AQI Predictor)

**Date:** 9 November 2025

## 1 | Executive Summary

This project presents a complete serverless machine-learning pipeline for forecasting the Air Quality Index (AQI) for the next 3 days for a chosen city (Karachi) using real-time and historical data. The solution applies a feature pipeline that collects raw pollutant + meteorological data, stores them in a feature store, trains multiple models (Random Forest, Ridge Regression, Neural Network), and deploys predictions via an interactive Streamlit dashboard. Automated CI/CD with hourly data ingestion and daily model retraining ensures continuous delivery. Key outcomes:

- Hourly feature pipeline automated
- Daily training pipeline automated
- Dashboard providing actionable forecasts
- Evaluation: Random Forest achieved  $RMSE = 3.47$ ,  $R^2 = 0.946$ ; Ridge Regression  $RMSE = 0.37$ ,  $R^2 = 0.999$ ; Neural Network  $RMSE = 10.69$ ,  $R^2 = 0.485$
- These results meet or exceed target performance ( $RMSE < 15$ ,  $R^2 > 0.85$ ) for two of the three models.

## 2 | Problem & Scope

Air pollution remains a major health risk in urban areas. Predicting AQI enables proactive health advisories and operational interventions (for example traffic management, emissions control). This project focuses on a city-scale deployment using publicly available APIs (AQICN), a feature store architecture (Hopsworks), and an interactive dashboard for stakeholders. The scope includes data ingestion, feature engineering, model training & selection, CI/CD auto visualization.

Limitations include reliance on free API rate limits, synthetic historical data generation for extended backfills, and model performance impacted by input data quality.

### 3 | Solution Architecture

- **Data ingestion:** Hourly fetch of current AQI and pollutant readings via AQICN API, plus generation of synthetic historical records when needed.
- **Feature store:** Ingested and engineered features are stored in Hopworks Feature Store (offline group) for reproducibility and versioning.
- **Feature engineering:** Time-based features (hour, day, month, sin/cos cyclic encoding, weekend flag) and derived features (rolling statistics, lag features, AQI change).
- **Model training pipeline:** Daily retraining job runs three model types, evaluates using RMSE, MAE,  $R^2$ , and selects the best.
- **Deployment/Inference:** Best model loads in an inference pipeline, produces 72-hour AQI forecast which is consumed by the Streamlit dashboard.
- **Dashboard:** Interactive UI presenting current AQI, 3-day forecast, daily averages, category alerts, health recommendations.
- **CI/CD:** GitHub Actions workflows — Feature Pipeline hourly, Training Pipeline daily at 2 AM UTC.

### 4 | Key Features & Innovations

- Fully **serverless** execution; no dedicated infrastructure required.
- **Automated pipelines** delivering continuous data and model updates.
- Multi-model experimentation and best-model selection.
- Use of **feature store** for reproducibility and governance.
- Real-time dashboard with user-friendly UX, color-coded categories, alerting for hazardous AQI.
- Documentation, tests, and version control aligned with best practices for data science projects.

## 5 | Results & Evaluation

Model	RMSE	R <sup>2</sup>	MAE	Notes
Random Forest	3.47	0.946	2.36	Excellent, robust performance
Ridge Regression	0.37	0.999	0.03	Possibly over-fit, inspect further
Neural Network	4.52	0.941	3.23	Close to Random Forest

Interpretation: The ensemble of tree-based and linear models provides high accuracy; the neural network may require more data, parameter tuning or regularization. The feature pipeline delivered stable inputs; dashboard latency and automation were tested successfully.

## 6 | Conclusion & Future Work

The Pearls AQI Predictor meets its core objectives: automated ingestion, model training, forecasting, and user-friendly visualization of air quality for urban decision-making. Future work may include:

- Integrating additional data sources (weather APIs, traffic sensor data).
- Upgrading neural network architecture (e.g., LSTM, Transformer) for sequence forecasting.
- Expanding to multiple cities/geographies.
- Deploying as a web service (API endpoint) and mobile friendly UI.
- Incorporating SHAP explanations for interpretability.

Overall, this project establishes a scalable, maintainable framework for AQI forecasting and real-time decision support.