Air Quality Index Project Report

1. Project Objective

The Air Quality Index (AQI) Forecasting Project develops an end-to-end machine learning system that predicts air quality levels for London, UK, across multiple time horizons (1h, 6h, 12h, 24h, 48h, 72h). The system provides real-time forecasts through an automated ML pipeline with data collection, feature engineering, model training, and web dashboard deployment.

Key Goal: Forecast AQI up to 72 hours ahead, provide actionable insights via interactive dashboard, implement automated pipeline updates, and monitor hazardous air quality conditions.

2. System Architecture & Implementation

2.1 Data Pipeline Architecture

The system follows a modular, production-oriented architecture with six layers: (1) Data Collection from Open Meteo API, (2) Feature Engineering for ML-ready features, (3) Hopsworks Feature Store for versioned storage, (4) Multi-output regression models for simultaneous predictions, (5) Hopsworks Model Registry for version control, and (6) Streamlit dashboard for visualization.

Data Source: Open Meteo Air Quality API for London, UK (51.5074°N, 0.1278°W) providing air quality metrics (PM2.5, PM10, Ozone, NO_2, SO_2, CO) and weather data (temperature, humidity, pressure, wind, precipitation, cloud cover).

2.2 Implementation Logic

1. Data fetching,
2. EDA snapshots for quality assessment
3. Feature engineering
4. Hopsworks storage with versioning
5. Model training with 80/20 chronological split
6. Evaluation using R^2 and MAE metrics
7. Model persistence to registry
8. Alert system monitoring
9. Streamlit dashboard deployment

3. Feature Engineering Strategy

3.1 Feature Categories (124 Total Features)

Time-Based (12 features): Categorical encodings (hour, day_of_week, month, is_weekend, is_rush_hour, season) and cyclical transformations (sin/cos) capture temporal patterns including diurnal cycles and seasonal trends.

Lag Features (22 features): 1h and 6h lags for 11 variables (PM2.5, PM10, Ozone, NO_2, SO_2, CO, Temperature, Humidity, Wind Speed/Direction, AQI) capture short-term dependencies.

Rolling Windows (54 features): 12h, 24h, and 48h windows compute mean and standard deviation for 9 variables (PM2.5, PM10, Ozone, NO_2, Temperature, Wind Speed, Humidity, Pressure, AQI) to capture medium-term trends.
Derived Features (5 features): temp_humidity_interaction, wind_chill, pm_ratio (PM2.5/PM10), pm_combined, and aqi_change_1h provide domain-specific insights.

Base Features (18 features): 8 pollutant concentrations + 7 weather variables + 3 computed values (AQI, timestamp, target).

3.2 Optimization Strategy

Lag windows reduced from [1,2,3,6] to [1,6] preventing overfitting (33% reduction). Rolling windows optimized from [3,6,12,24] to [12,24,48] for better long-horizon signals. Wind direction added to lag features for pollutant dispersion modeling. Humidity and pressure added to rolling features as weather stability indicators.

Feature Store: Hopsworks london_air_quality_6h (Version 5) with automatic schema compatibility checks and daily incremental updates.

4. Model Selection & Training

4.1 Model Architecture

Random Forest Regressor: Configuration (n_estimators=50, max_depth=5, min_samples_split=10, min_samples_leaf=5, max_features='sqrt') handles non-linear relationships, provides robustness to outliers, and delivers feature importance rankings.

Ridge Regression: Configuration (alpha=20.0) provides linear baseline with strong L2 regularization, fast training/inference, and better generalization for longer horizons.

Multi-Output Implementation: Uses sklearn.multioutput.MultiOutputRegressor to predict all 6 horizons simultaneously, enabling shared representation learning, reduced variance through joint optimization, and 83% training time reduction.

4.2 Training Process

Data Split: Chronological 80/20 split yielding ~5,800 training and ~1,450 test samples from 1 year of hourly data.

Workflow: (1) Load features from Hopsworks, (2) Create multi-horizon targets by shifting AQI forward [1,6,12,24,48,72] hours, (3) Prepare feature matrix excluding targets, (4) Train models using MultiOutputRegressor, (5) Calculate horizon-specific metrics, (6) Save models and metrics locally and to registry.

Performance Optimization: Prediction clipping to [0,500], robust metrics handling NaN/inf values, and train/test gap monitoring for early overfitting detection.

5. CI/CD Pipeline

5.1 GitHub Actions Automation

Configuration: .github/workflows/retrain.yml executes hourly (cron: '0 * * * *' UTC) on ubuntu-latest with Python 3.12. Manual triggers supported via workflow_dispatch. Environment variables (HOPSWORKS_API_KEY, HOPSWORKS_PROJECT_NAME) stored as GitHub Secrets.

5.2 Hourly Operations

The workflow fetches 24 hours of new data, generates EDA snapshots, engineers and appends features to Hopsworks, retrains models with updated dataset, saves new versions to Model Registry, runs alert system, and executes SHAP interpretability analysis. Execution time: ~1-2 minutes. Status: 110+ successful runs.

Model Versioning: Models saved as multi_output_{algorithm}_model.pkl with metrics in multi_output_{algorithm}_h{horizon}_metrics.json. Hopsworks tracks version history with descriptions.

6. Web Dashboard & Alert System

6.1 Dashboard Components

Technology: Streamlit framework providing six key sections: Current AQI status with color-coded gauge, Alert panel for hazardous levels, Model registry metrics with comparison tables, Historical AQI trends visualization, multi-horizon forecasts for all time windows, and Side-by-side model comparison (Random Forest vs Ridge).

Features: Interactive date range selector, real-time Hopsworks data loading, automated model inference, and visual performance metrics (R^2 scores, MAE, bar charts).

6.2 Alert System

Implementation: EPA AQI threshold monitoring (Low: 0-50, Moderate: 51-100, High: 101-150, Critical: >200) for current and predicted values across all horizons. Alert types include severity levels (Low, Moderate, High, Critical) with 1-hour deduplication window and 24-hour JSON-based storage.

7. Conclusion

The Air Quality Index Forecasting Project successfully implements a production-ready MLOps pipeline for multi-horizon AQI prediction in London. The system demonstrates effective feature engineering (124 features), complementary model selection (Random Forest and Ridge Regression), and fully automated CI/CD with hourly GitHub Actions execution. The Streamlit dashboard provides real-time predictions and health alerts, while Hopsworks ensures robust version control for features and models. The multi-output regression approach achieves 83% training efficiency improvement while maintaining prediction accuracy across all six-time horizons (1h to 72h).