# Air Quality Index (AQI) Prediction System

**Author: Taha Tahir**

**Organization: 10Pearls**

**Date: 09 November 2025**

# 1. Introduction

The Air Quality Index (AQI) is a standardized indicator that quantifies the concentration of major air pollutants such as PM2.5, PM10, $O_3$, $NO_2$, $SO_2$, and CO. This project aims to build a fully automated and production-ready AQI prediction system capable of:

- Collecting real-time data from World Air Quality Index (WAQI) API,
- Cleaning and processing the data through a robust 10-step pipeline,
- Engineering multiple predictive features,
- Training and comparing multiple ML models, and
- Deploying predictions through a web-based dashboard.

# 2. Objectives

The main objectives of the system are:

1. Automate hourly data collection from the WAQI API.
2. Develop a reliable data preprocessing and cleaning pipeline.
3. Perform advanced feature engineering for time-series forecasting.
4. Train and evaluate multiple machine learning models for AQI prediction.
5. Deploy the best model for real-time inference using a Streamlit dashboard.
6. Integrate MLOps practices with Hopsworks Feature Store and GitHub Actions.

# 3. System Architecture

The overall architecture integrates data collection, model training, feature storage, and deployment in a cohesive MLOps pipeline.

## 3.1 Data Flow

WAQI API ---> Data Fetcher ---> Data Cleaning ---> Feature Engineering
---> Hopsworks Feature Store ---> Model Training ---> Predictions
---> Streamlit Dashboard

## 3.2 Components

| Component | Description |
|---|---|
| **Data Fetcher** | Collects hourly air quality and weather data using WAQI API. |
| **Data Cleaning** | Applies a 10-step cleaning pipeline for high-quality input data. |

| Component | Description |
|---|---|
| **Feature Engineering** | Generates multiple lag, rolling, and domain-specific features. |
| **Model Training** | Compares XGBoost, Random Forest, LSTM, and CNN models. |
| **Feature Store** | Manages features and ensures training-serving consistency. |
| **Streamlit Dashboard** | Provides a real-time, interactive visualization of predictions. |
| **GitHub Actions** | Automates hourly data updates and predictions. |

# 4. Methodology

## 4.1 Data Collection

Data was fetched every hour via the World Air Quality Index (WAQI) API.
Collected pollutants and weather features include:

- PM2.5, PM10, $O_3$, $NO_2$, $SO_2$, CO
- Temperature, Humidity, Pressure, Wind Speed

## 4.2 Data Cleaning Pipeline

A comprehensive 10-step cleaning pipeline was implemented:

1. Duplicate removal
2. AQI range validation (0–500)
3. Pollutant-level verification
4. Outlier detection (IQR + Z-score)
5. Domain-based value capping
6. Temporal consistency checks
7. Feature correlation analysis
8. Low variance feature removal
9. Rolling median smoothing
10. Final validation and integrity check

**Outcome:** 85–92% data retention with >95% quality assurance.

# 5. Machine Learning Models

Four models were trained and evaluated for comparative performance:

| Rank | Model | RMSE | MAE | $R^2$ | Performance |
|---|---|---|---|---|---|
| 1 | XGBoost | 9.44 | 5.55 | 0.947 | Excellent |
| 2 | Random Forest | 9.59 | 5.70 | 0.945 | Excellent |
| 3 | LSTM | 16.34 | 10.90 | 0.842 | Good |
| 4 | CNN 1D | 45.42 | 38.53 | -0.22 | Poor |

XGBoost achieved 94.7% variance explanation and was chosen for deployment.

# 6. Model Details

## 6.1 XGBoost Configuration

```
XGBRegressor(
   n_estimators=100,
   learning_rate=0.1,
   max_depth=6,
   subsample=0.8,
   colsample_bytree=0.8,
   random_state=42
)
```

## 6.2 Cross-Validation

- **Method:** TimeSeriesSplit (5-fold)
- **Metrics:** $R^2$, RMSE, MAE
- **Result:** Mean $R^2 = 0.931 \pm 0.023$, Mean RMSE = $7.51 \pm 2.92$

# 7. MLOps & Deployment

## 7.1 Hopsworks Feature Store

- Centralized repository for engineered features
- Version-controlled and time-travel capable
- Ensures feature consistency between training and production

## 7.2 GitHub Actions Automation

- **Frequency:** Every hour
- **Tasks Automated:**
    1. Environment setup
    2. Data collection and cleaning
    3. Feature generation and upload
    4. Model inference
    5. JSON update for dashboard visualization

### 7.3 Streamlit Dashboard

- Displays current AQI with color-coded health categories
- Provides 3-day forecast with trend visualization
- Showcases model analytics and feature importance

## 8. Results and Discussion

The XGBoost model demonstrated superior performance with:

- RMSE: 9.44
- MAE: 5.55
- $R^2$: 0.947

This means the model can predict AQI within ±9 points of actual values on average, representing strong predictive capability. The feature analysis confirmed the dominance of PM2.5-related lag and rolling mean features, consistent with domain understanding that PM2.5 is a primary driver of air quality fluctuations.

## 9. Conclusion

This project successfully demonstrates a production-grade AQI prediction system integrating machine learning, MLOps, and data automation.
By achieving 94.7% prediction accuracy, the system provides reliable air quality forecasts and forms a strong foundation for scalable environmental monitoring platforms.
With further improvements and broader data integration, this approach can support smart city air quality management and public health decision-making.

StreamLit Deployed Link: https://airqualityindexprediction.streamlit.app/

# 10. References

1. EPA Air Quality Index Guide: https://www.airnow.gov/aqi/aqi-basics/
2. WAQI API Documentation: https://aqicn.org/api/
3. Hopsworks Documentation: https://docs.hopsworks.ai/
4. Scikit-learn Documentation – *Time Series Cross-Validation.*
5. YouTube videos provided in Discord server.
6. ChatGPT for learning new techniques and to cater low R2 scores.