

AQI Prediction System - Technical Report

1. Project Overview

A comprehensive machine learning platform that forecasts Air Quality Index values using environmental data. The system provides accurate 3-day AQI forecasts through an automated pipeline integrating data collection, ML modeling, API services, and interactive visualization. It serves both technical users and the general public with reliable air quality predictions.

2. System Architecture

2.1 Data Flow Pipeline

The system follows a modular architecture with four key components working in sequence:

Data Flow: OpenWeather API → Data Processing → ML Models → REST API → Dashboard

Component Integration:

- **Backend Service:** FastAPI application handling data processing and model inference
- **Frontend Interface:** Streamlit dashboard for real-time visualization
- **Feature Store:** Hopsworks integration for incremental feature management
- **Automation:** GitHub Actions for scheduled retraining and data updates

This design ensures seamless data flow while allowing independent component scaling and maintenance.

3. Data Pipeline & Feature Engineering

3.1 Data Collection & Processing

Data Acquisition:

- 365 days of historical air quality data collected in 30-day batches
- Real-time hourly polling for current conditions
- Pollutant monitoring: PM2.5, PM10, NO2, O3, SO2, CO
- Automated AQI calculation using US EPA standards

Data Quality:

- Timestamp standardization and chronological sorting
- Missing value handling and duplicate removal
- Validation checks for data consistency

3.2 Feature Engineering Strategy

Temporal Features:

- Time components: hour, day, month, day of week
- Cyclical encoding using sine/cosine transformations
- Seasonal indicators and peak period detection components, i.e., is_weekend, is_peak_hour

Statistical Features:

- 24-hour lag features for key pollutants

- Rolling statistics and moving averages
- Rate of change calculations and pollutant interactions

Optimization:

- 22 optimal features selected from correlation analysis
- PM2.5 identified as the strongest predictor (0.954 correlation)
- Seasonal pattern incorporation based on monthly variations

4. Machine Learning Core

4.1 Model Architecture

Algorithm Portfolio:

- Ridge Regression: ($R^2 = 0.8264$, MAE = 13.0012, RMSE = 17.9829)
- Random Forest: ($R^2 = 0.7087$, MAE = 18.9470, RMSE = 23.2958)
- XGBoost: ($R^2 = 0.6682$, MAE = 19.4325, RMSE = 24.8657)

Training Methodology:

- Time-series cross-validation, preventing data leakage
- Temporal train-test split (85-15 ratio)
- Automated model selection based on R^2 thresholds

4.2 Prediction & Validation

Forecasting Engine:

- Recursive multi-day prediction
- Temporal feature adjustment for future points
- Realistic AQI boundary enforcement
- Confidence scoring decreases with horizon

Validation Framework:

- Multi-metric evaluation (RMSE, MAE, R^2)
- Overfitting detection through gap analysis
- Prediction quality and consistency checks

5. API & Dashboard Services

5.1 Backend API

Service Design:

- RESTful endpoints with automatic model loading
- In-memory caching for rapid responses
- Comprehensive error handling

Key Endpoints:

- Current AQI with real-time conditions
- 3-day forecasts with confidence scores
- Historical data for analysis

- Model performance monitoring

5.2 Interactive Dashboard

Visualization:

- Real-time AQI display with color coding
- Interactive forecast charts
- Historical trend analysis
- Pollutant correlation views

User Experience:

- Configurable prediction horizons
- Data export functionality
- Responsive design
- Alert systems for critical conditions

6. Automation & Performance

6.1 Feature Store & Automation

Data Management:

- Hopworks integration for feature versioning
- Hourly incremental updates
- Automated merge and deduplication

Training Pipeline:

- Daily model retraining
- Performance comparison and versioning
- Rollback capabilities

6.2 System Performance

Key Insights:

- Average AQI: 118.77 (Moderate to Unhealthy for Sensitive Groups)
- Seasonal patterns: High in winter, low in monsoon
- PM2.5 is the strongest predictor with a 0.954 correlation

Best Model Results:

- R^2 scores > 0.8 on validation
- MAE < 15 AQI points
- Minimal overfitting (gap < 0.15)
- Stable cross-seasonal performance

7. Conclusion

The AQI Prediction System provides a comprehensive ML pipeline, spanning from data collection to deployment. By combining strong feature engineering with reliable modeling and automated workflows, it ensures accurate, stable predictions. Interactive visualizations make the platform user-friendly and suitable for ongoing air quality monitoring.