Air Quality Bot Index – Project Report

1. Introduction

Air pollution poses one of the most severe threats to human health and the environment. Monitoring and predicting the Air Quality Index (AQI) helps authorities and citizens take timely precautions against harmful pollutants. The Air Quality Bot Index is an AI-driven forecasting system that automatically collects weather and pollutant data, processes features, trains predictive models, and displays AQI trends on a modern Streamlit-based dashboard. This project integrates machine learning, cloud automation, and feature engineering to forecast the next 3 days of AQI for any given city using real-time environmental data. The goal is to provide a self-updating, explainable, and visually informative system for AQI analysis and prediction.

2. System Overview

The project consists of four major modules:

1. Feature Pipeline – Collects AQI and weather data, performs feature engineering, and saves the processed dataset.
2. Training Pipeline – Trains and evaluates a Random Forest model, logs results to MLflow, and stores artifacts in Hopsworks.
3. CI/CD Automation – Uses GitHub Actions to trigger the feature and training pipelines automatically.
4. Streamlit Dashboard – Displays live AQI, historical trends, 3-day forecasts, and model performance metrics.

3. Methodology

A. Feature Pipeline

The feature extraction pipeline fetches 360 days of historical air pollution and weather data using the OpenWeather API. It processes data to remove missing values and derives time-based features such as hour, day, month, weekday, and AQI change rate. The merged dataset is stored in both CSV format and the Hopsworks Feature Store for persistent storage and versioning.

B. Training Pipeline

The training pipeline loads processed data from the Feature Store, selects relevant features, and trains a Random Forest Regressor. It computes evaluation metrics (RMSE, MAE, R^2) and uses Prophet for forecasting the next three days of AQI. Models and metrics are logged using MLflow and stored locally as well as in Hopsworks.

4. Key Features

- Fetches raw AQI and weather data using OpenWeather APIs.
- Computes time-based and derived features like AQI Change Rate.
- Stores processed features in Hopsworks Feature Store.
- Trains Random Forest and Prophet models for AQI forecasting.
- Logs metrics (RMSE, MAE, R^2) and artifacts using MLflow.

- Automates pipelines with GitHub Actions.
- Provides interactive Streamlit dashboard with AQI visualization.

5. Results

Model performance metrics:
RMSE: 0.032
MAE: 0.002
R^2: 0.999

6. Limitations and Future Enhancements

Despite the strong performance and automation, a few limitations remain:

- Currently limited to Random Forest and Prophet models – future versions can add Ridge Regression, LSTM, or Transformer-based models.
- The pipeline writes to Hopsworks but does not read directly from it – future enhancement will include two-way integration.
- Only basic feature importance is provided – SHAP/LIME can be added for explainability.
- No real-time alerts for hazardous AQI levels – alert notifications can be integrated.
- Currently supports one city at a time – future updates can include multi-city predictions.

7. Conclusion

The Air Quality Bot Index successfully demonstrates an automated AI system for AQI prediction. It integrates data engineering, model training, and visualization into a cohesive pipeline. The project leverages Hopsworks for feature storage, MLflow for tracking, and Streamlit for deployment. With future improvements such as explainability tools, deep learning models, and multi-city expansion, the system can evolve into a scalable environmental intelligence platform.