Air Quality Index (AQI) Prediction Project

Prepared By: Manav Birjani

Company: 10 Pearls Shine

This project report explains the design, implementation, and outcomes of the AQI Prediction System built using machine learning, data pipelines, and CI/CD automation. It highlights objectives, methodologies, results, and future scope in detail

1. Introduction

The AQI Prediction Project is designed to provide real-time and forecasted air quality information for Karachi. This system integrates machine learning with CI/CD pipelines to automate data fetching, model training, and predictions. It was built as part of a professional project under the name '10 Pearls Shine'.

2. Objectives

- Automate the process of collecting AQI data from WAQI API. - Preprocess and engineer features for model input. - Train and deploy machine learning models for AQI prediction. - Provide real-time predictions and 3-day forecasts. - Enable CI/CD automation using GitHub Actions. - Ensure explainability using SHAP. - Create alerts for AQI threshold breaches. - Maintain proper documentation and reporting.

3. Data Sources

- WAQI API for real-time AQI data. - Raw AQI datasets stored in CSV files such as `raw_aqi_data_karachi.csv`, `daily_predictions.csv`, and `features_store.csv`. - Features include PM2.5, PM10, O3, CO, SO2, NO2, and AQI change rate.

4. Methodology

The project follows these steps: 1. **Data Collection**: Fetch real-time data via API. 2. **Feature Engineering**: Extract pollutants, compute AQI change rate, and store features. 3. **Model Training**: Train Random Forest and Gradient Boosting models. 4. **Model Registry**: Store and version models. 5. **Prediction Pipeline**: Generate real-time and forecast AQI. 6. **Explainability**: Use SHAP for model interpretation. 7. **Deployment**: Automate via CI/CD pipelines on GitHub Actions. 8. **Visualization**: Streamlit dashboard for interactive analysis.

5. Results & Findings

- The Random Forest model performed best with high accuracy in AQI prediction. - CI/CD workflows ensure predictions every 3 hours and model retraining daily. - Forecasts help predict AQI trends for the next 3 days. - SHAP plots provide feature importance for transparency. - Alerts highlight poor air quality for proactive action.

6. Challenges

- Handling missing or inconsistent API data. - Ensuring compatibility of workflows with Windows runner. - GitHub Actions authentication issues for CI/CD automation. - Balancing accuracy with computational efficiency. - Managing multiple datasets and maintaining consistency.

7. Conclusion

The AQI Prediction Project successfully automates air quality monitoring for Karachi. It integrates machine learning, data engineering, and DevOps to provide reliable, real-time, and forecasted AQI insights. The system demonstrates how AI can be used for environmental monitoring and public health awareness.