10PEARLS SHINE INTERNSHIP PROGRAM

AIR QUALITY PREDICTION APP

DEVELOPED BY: ALIYA AKHTAR

(DATA SCIENCE INTERN)

PROBLEM STATEMENT

Air pollution has emerged as one of the most pressing environmental and public health issues in urban areas around the world. Karachi, being one of the largest and most densely populated cities in Pakistan, faces severe air quality challenges due to rapid industrialization, increased vehicular traffic, construction activities, and meteorological factors. High levels of air pollutants such as particulate matter (PM2.5, PM10), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and ozone (O_3) have a direct impact on human health, contributing to respiratory diseases, cardiovascular problems, and reduced life expectancy.

Despite the significant health and environmental risks, timely and accurate information about air quality remains limited. Existing monitoring systems provide real-time measurements, but they often lack predictive capabilities, which prevents residents, policymakers, and urban planners from taking proactive measures to reduce exposure. Forecasting air quality can enable better decision-making, such as issuing health advisories, planning outdoor activities, or implementing temporary restrictions on pollution sources.

However, predicting air quality is a complex task due to its dependence on multiple dynamic factors, including weather conditions, traffic patterns, industrial emissions, and seasonal variations. The lack of an automated, data-driven system for predicting air quality in Karachi leaves a critical gap in environmental management and public health preparedness.

SOLUTION

To address the challenge of limited predictive insight into air quality in Karachi, a comprehensive **Air Quality Prediction Application** has been developed. This application provides both **short-term forecasts** and **historical trend analysis**, enabling users to make informed, data-driven decisions regarding air quality exposure.

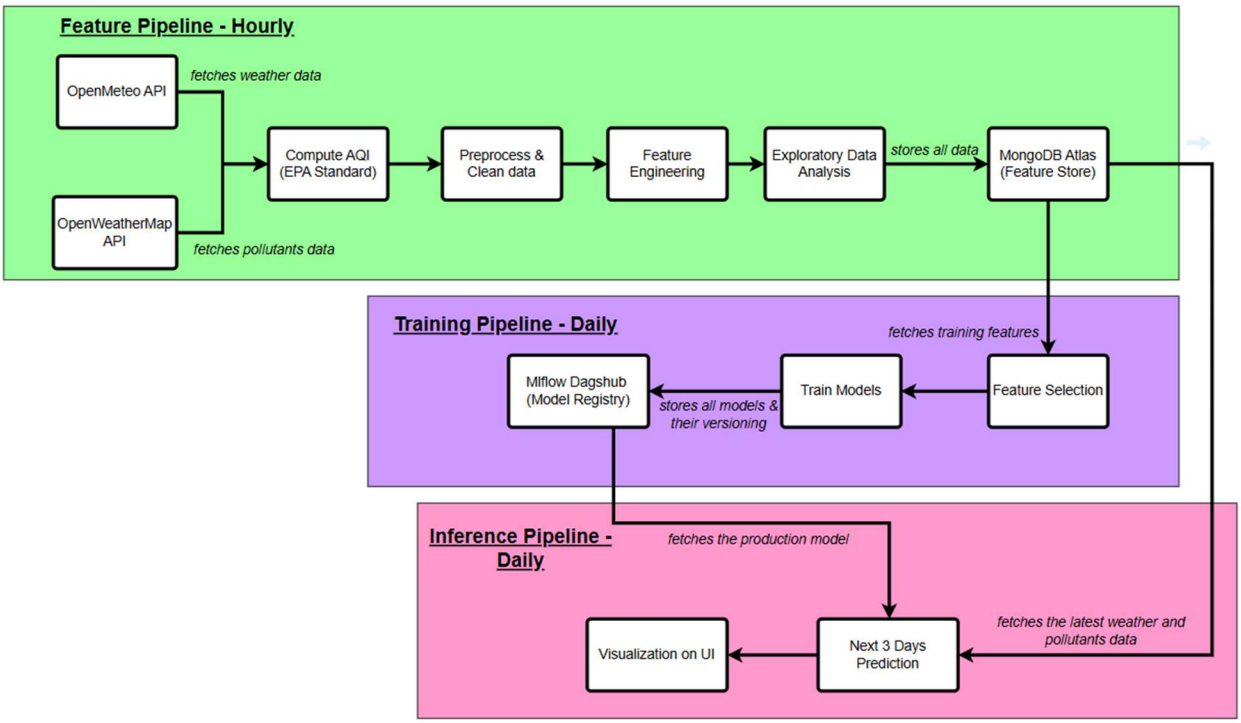
The key features of the solution are:

1. **Three-Day AQI Forecasting**
    - The application predicts the Air Quality Index (AQI) for the next three days using historical and real-time environmental data.
    - These forecasts allow citizens, health authorities, and policymakers to anticipate periods of high pollution and take proactive measures to minimize exposure.
2. **Historical Data Visualization**
    - The application provides clear visualizations of AQI trends over the past seven days.
    - Users can easily identify patterns, fluctuations, and recurring pollution spikes, supporting better understanding of the city's air quality dynamics.
3. **Actionable Insights**
    - By combining forecasted and historical data, the application offers actionable insights, such as high-risk periods for outdoor activities or recommendations for vulnerable populations.
    - The user-friendly interface ensures that insights are accessible to both technical and non-technical audiences.
4. **Data-Driven Decision Support**
    - The system empowers urban planners, environmental agencies, and the general public with predictive analytics, bridging the gap between raw air quality data and actionable knowledge.

This solution leverages machine learning models trained on historical AQI data, meteorological parameters, and other relevant environmental indicators. The integration of predictive analytics with intuitive visualizations provides a comprehensive tool to **monitor, understand, and forecast air quality in Karachi**, helping to mitigate the health risks associated with air pollution.

METHODOLOGY

The development of the Air Quality Prediction Application follows a **structured, pipeline-driven approach** to ensure data reliability, accurate modeling, and seamless deployment. The methodology is divided into **three main pipelines**: hourly feature fetching, daily model training, and daily inference.



1. Data Collection and Preparation

- **Historical Data:** Initially, three months of historical weather data were collected from **Open-Meteo**, and pollutant concentration data were fetched from **OpenWeather**.
- **AQI Computation:** Using the United States Environmental Protection Agency (EPA) standards, the Air Quality Index (AQI) values were calculated based on the pollutant data.
- **Preprocessing & Feature Engineering:** The raw datasets underwent cleaning, normalization, and feature engineering to create meaningful input features for the predictive models. The processed data was then stored in a **feature store** for downstream tasks.

2. Exploratory Data Analysis (EDA)

Before building predictive models, a detailed Exploratory Data Analysis (EDA) was performed on the collected dataset to understand trends, relationships, and data quality issues. The key objectives of EDA were to identify patterns in air quality, guide feature engineering, and detect anomalies in the data.

Key Insights from EDA:

- **Temporal Trends:** Analysis of AQI over the historical period revealed daily and weekly fluctuations, indicating higher pollution levels during certain hours of the day and weekdays, likely due to traffic and industrial activity.

- **Correlation Analysis:** Relationships between weather parameters (temperature, humidity, wind speed) and pollutant concentrations were analyzed to identify features with predictive power for AQI.
- **Pollutant Behavior:** Individual pollutant levels (PM2.5, PM10, NO_2, SO_2, O_3) were visualized to understand their contribution to overall AQI and seasonal variations.
- **Missing Data & Outliers:** Data completeness and quality checks were performed. Missing values were imputed, and outliers were identified and treated to prevent model bias.
- **Visualization:** Graphs and heatmaps were used to depict trends, correlations, and pollutant distributions. These visualizations helped in feature selection and understanding the dynamics of air quality in Karachi.

By conducting EDA, the dataset was better understood, which informed preprocessing steps, feature engineering decisions, and ultimately improved model performance. The visual insights also provide an intuitive understanding of air pollution trends, which are later reflected in the application's UI visualizations.

3. Hourly Feature Pipeline

- This pipeline runs **every hour** to fetch real-time weather and pollutant data from the APIs.
- The newly fetched data is preprocessed and stored in the **feature store**, ensuring that the models always have access to the most recent information for accurate predictions.

4. Daily Model Training Pipeline

- **Model Selection:** Four machine learning models were evaluated — **LightGBM, Random Forest, XGBoost, and Ridge Regression**.
- **Training Process:** Each model retrieves training data from the feature store, and multiple versions are trained and tracked in a **model registry** using **MLflow** integrated with **DagsHub**.
- **Model Evaluation:** The performance of each model is evaluated based on the **Root Mean Squared Error (RMSE)** metric.
- **Production Model:** The model with the lowest RMSE is selected and promoted to production for inference.

5. Daily Inference Pipeline

- **Model Retrieval:** The best-performing model is automatically retrieved from the model registry.
- **Prediction:** Using the latest features from the feature store, the model generates AQI predictions for the next **three days**.

- **Visualization:** The forecasted AQI values are then displayed on the application's UI alongside visualizations of the past **seven days** to provide context and trend analysis.

6. Data and Model Management

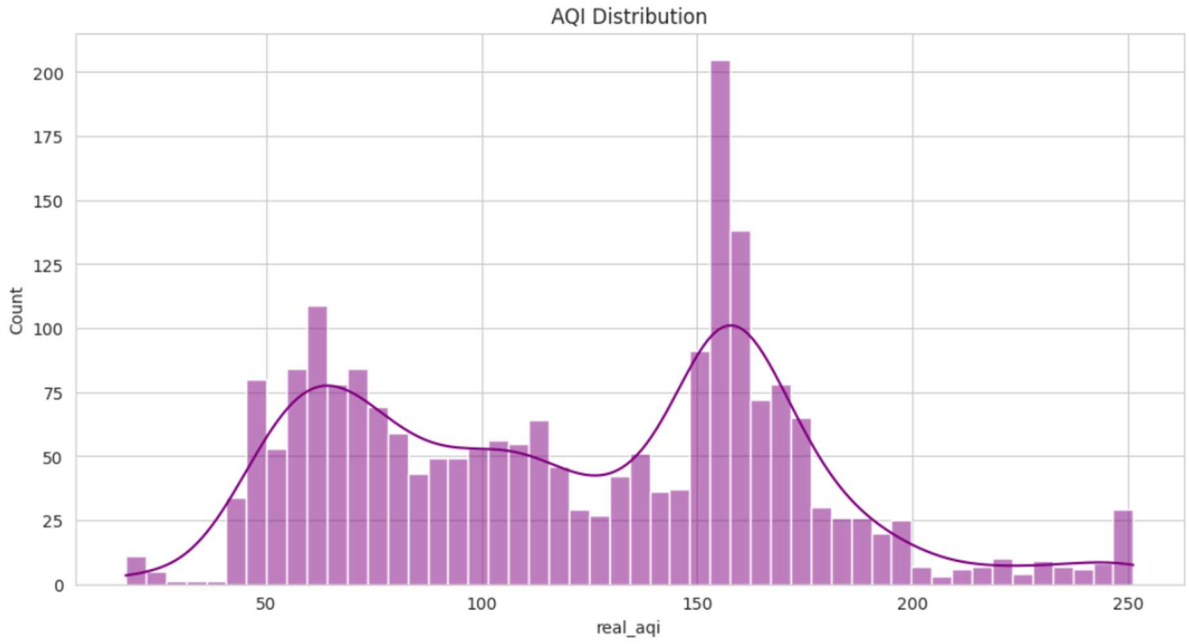
- All feature and model versions are systematically tracked, ensuring reproducibility and version control.
- The pipeline architecture allows for continuous data ingestion, model retraining, and real-time predictions, creating a robust and scalable AQI prediction system.

This methodology ensures that the application is **data-driven, automated, and continuously updated**, providing accurate and timely air quality forecasts to users.

EXPLORATORY DATA ANALYSIS

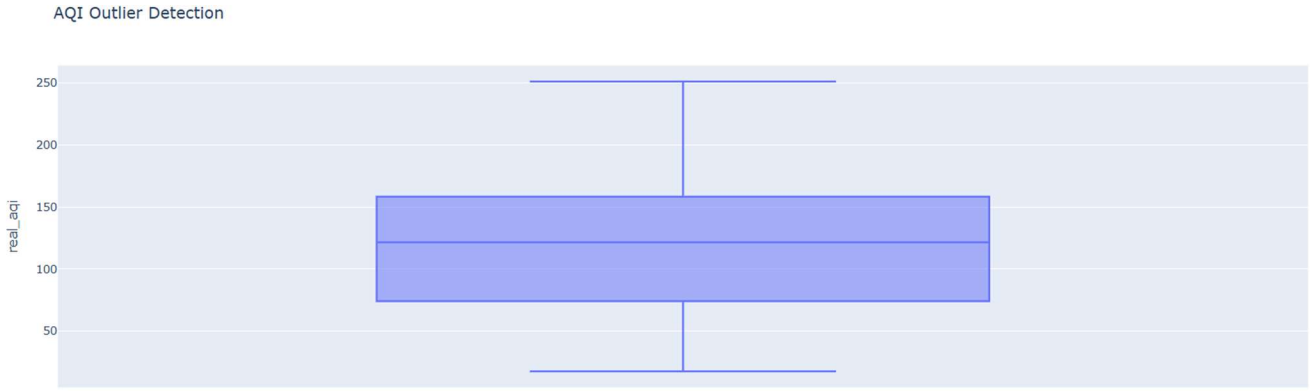
AQI Distribution and Outlier Analysis

The histogram with the density curve shows how the Air Quality Index (AQI) values are distributed across the dataset. It helps us understand the frequency of different AQI levels and reveals the overall pattern of air quality. The plot suggests that AQI values are spread across a wide range, with noticeable concentrations around moderate and higher pollution levels, indicating periods of varying air quality conditions.



The box plot is used to detect potential outliers and summarize the spread of AQI values. It shows the median AQI, the interquartile range (middle 50% of values), and the overall variability. Any points far from the box would indicate unusually high or low AQI readings. From the plot, we can observe that there are no significant outliers in the AQI data, as all observations fall within the

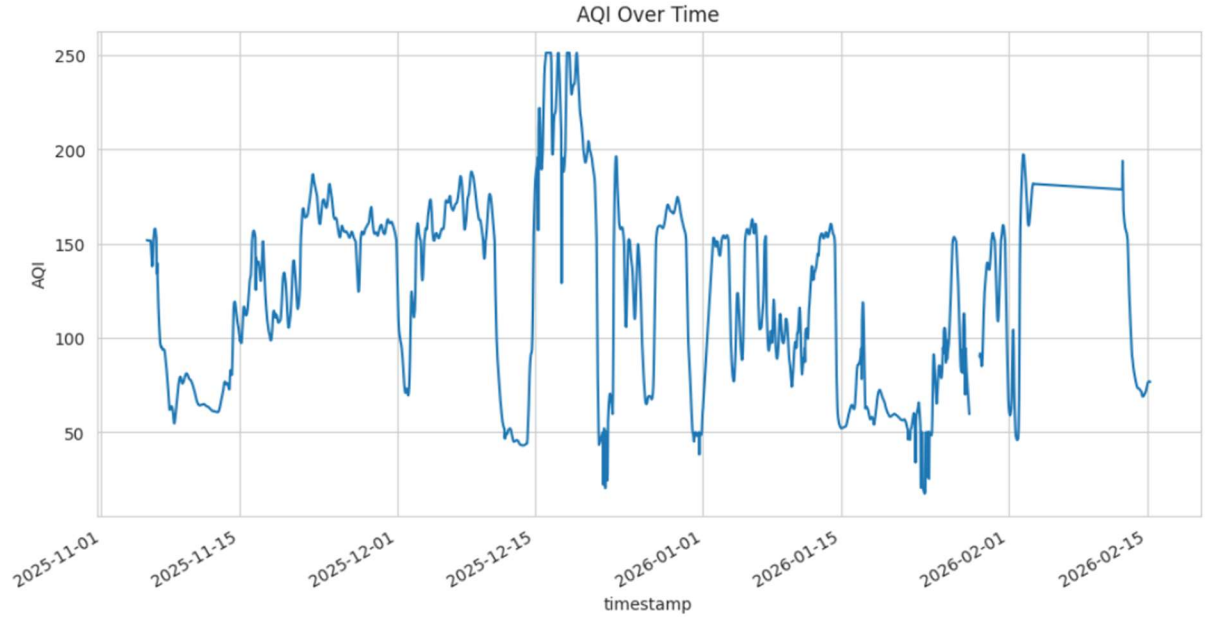
whisker range, suggesting that extreme values are within expected variability rather than being anomalous.



AQI Trend Over Time

The line plot illustrates how the Air Quality Index (AQI) changes over the observed period. It helps visualize trends, fluctuations, and sudden spikes or drops in air pollution levels. By examining this plot, we can identify periods of poor air quality (high AQI values) and relatively cleaner periods (low AQI values), as well as understand how variable the air quality is over time.

The graph shows noticeable variability, indicating that AQI does not remain constant and can change significantly due to factors such as weather conditions, and pollutants. Peaks in the plot represent episodes of higher pollution, while dips indicate improved air quality.

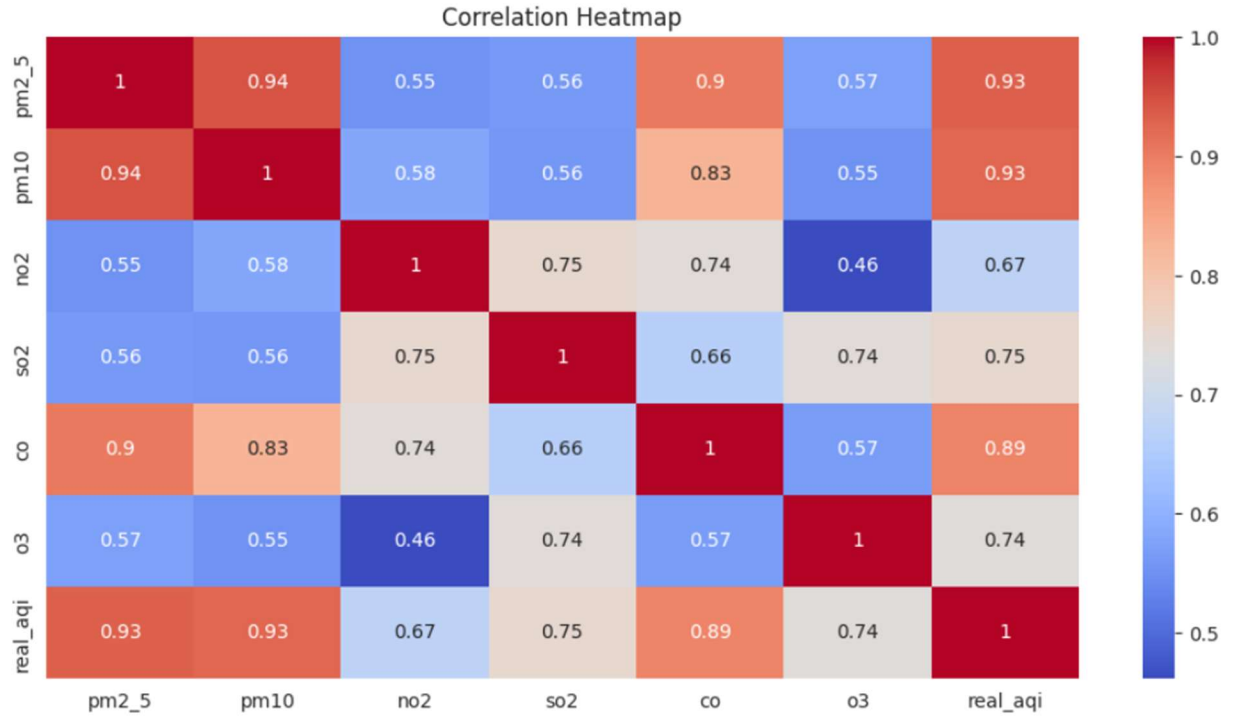


Correlation Analysis of Pollutants and AQI

The correlation heatmap illustrates the strength and direction of relationships between different air pollutants (such as PM2.5, PM10, NO_2, SO_2, CO, and O_3) and the Air Quality Index (AQI). Each cell in the heatmap represents a correlation coefficient ranging from −1 to 1, where values closer to 1 indicate a strong positive relationship, values near 0 indicate little or no relationship, and negative values indicate an inverse relationship.

From the heatmap, particulate matter pollutants (PM2.5 and PM10) show a strong positive correlation with AQI, suggesting that increases in these pollutants are closely associated with worsening air quality. Carbon monoxide (CO) also exhibits a strong relationship with AQI, indicating its significant contribution to pollution levels. Other gases such as NO_2, SO_2, and O_3 show moderate correlations, implying they also influence air quality but to a lesser extent.

Additionally, strong correlations among certain pollutants suggest that they may originate from similar sources, such as traffic emissions or industrial activities. This analysis helps identify the most influential variables and supports feature selection for building predictive models.

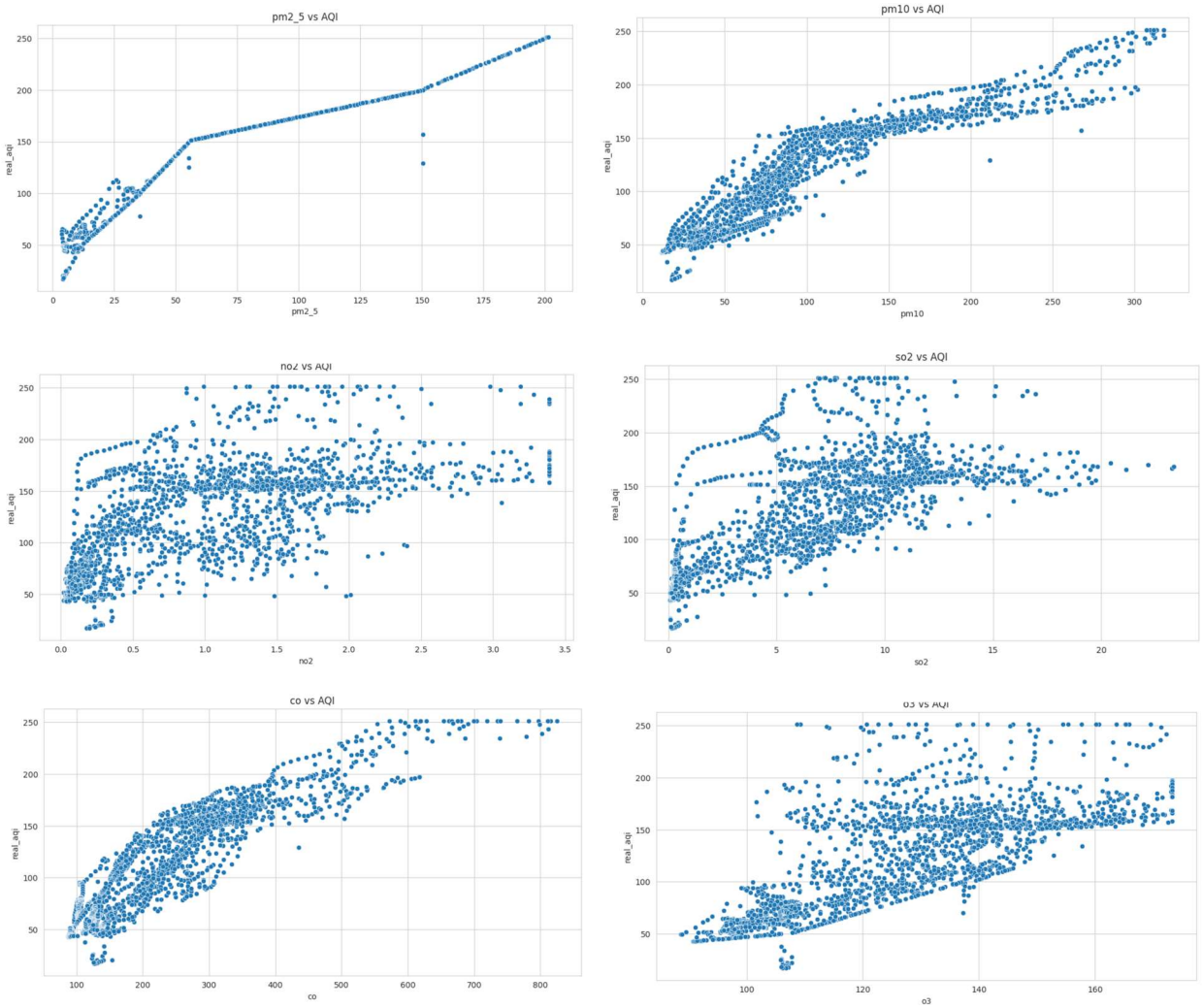


Analysis of Pollutant-AQI Relationships

The scatter plots reveal how different air pollutants correlate with the Air Quality Index (AQI). The analysis shows that **particulate matter (PM10 and PM2.5) exhibits the strongest linear relationships with AQI**, indicating these are primary drivers of air quality deterioration. PM2.5 demonstrates an especially clear linear trend, suggesting it's the most reliable predictor of overall air quality.

In contrast, **gaseous pollutants show weaker or non-linear patterns**. SO2 displays significant scatter with no clear trend, while NO2 shows a relationship that plateaus at higher concentrations. CO maintains a moderate positive correlation, and O3 exhibits a scattered distribution, reflecting its complex formation process through photochemical reactions.

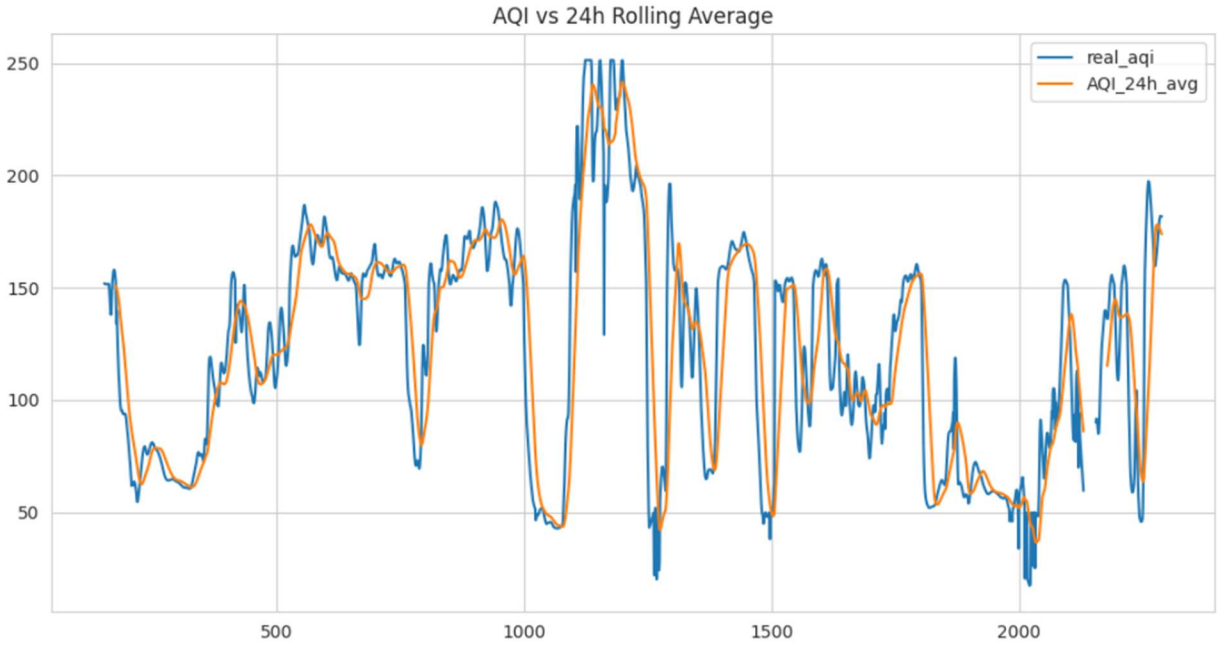
These patterns indicate that **particulate pollution is the dominant factor** affecting AQI in the dataset, while gaseous pollutants contribute less consistently to overall air quality measurements. This insight is crucial for developing targeted air quality improvement strategies focused on reducing particulate emissions.



AQI Temporal Trend Analysis

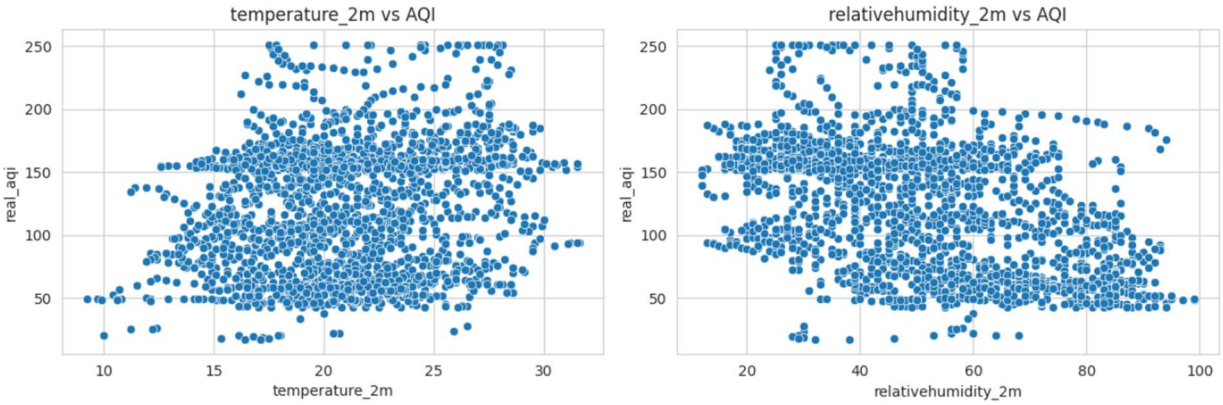
The time series plot displays actual AQI measurements alongside a 24-hour rolling average. The raw AQI (blue) exhibits significant day-to-day variability with sharp peaks and valleys, reflecting the dynamic nature of air pollution. The rolling average (orange) smooths these fluctuations, revealing the underlying trend and making sustained pollution episodes more apparent. Notable observations include a severe pollution spike around the 1000-day mark reaching AQI levels above

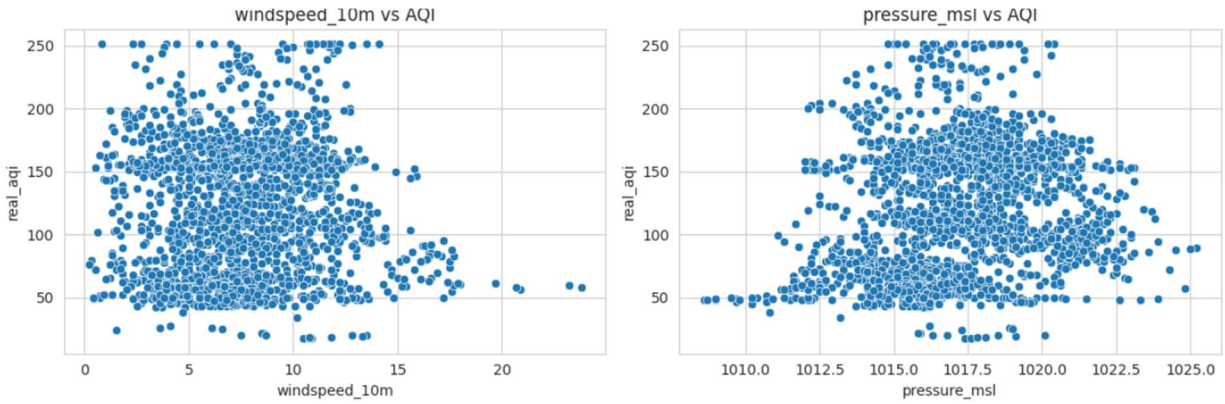
250, and a general declining trend toward the end of the observation period, suggesting improving air quality conditions.



Weather Variables and AQI Analysis

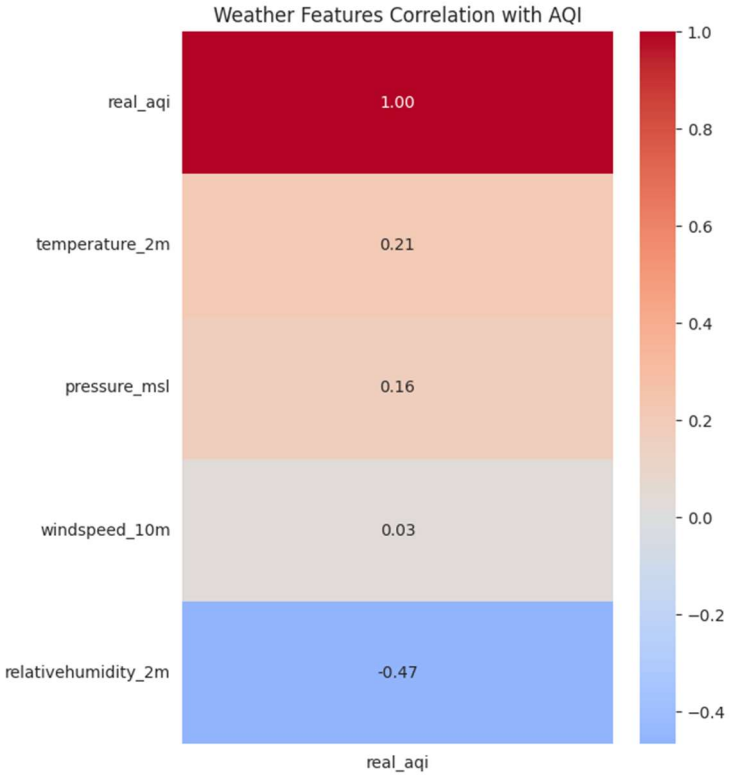
The scatter plots examine the relationship between meteorological parameters and air quality. Unlike pollutant concentrations, weather variables show weaker correlations with AQI. **Relative humidity exhibits a slight negative trend**, suggesting that higher moisture levels may aid in settling particulate matter. **Wind speed shows a notable pattern** where lower wind speeds (below 5 m/s) are associated with higher AQI values, indicating that calm conditions allow pollutants to accumulate. Temperature and atmospheric pressure display scattered distributions with no clear linear relationships, suggesting their influence on AQI is either indirect or mediated by other factors. These findings highlight that while weather conditions can influence pollutant dispersion and accumulation, they are secondary factors compared to direct emission sources.





Weather-AQI Correlation Heatmap

The correlation matrix quantifies the relationships between meteorological variables and AQI. **Relative humidity shows the strongest correlation at -0.47**, indicating a moderate negative relationship where increased moisture is associated with lower pollution levels, likely due to wet deposition of particulates. Temperature exhibits a weak positive correlation (0.21), while atmospheric pressure shows minimal association (0.16). Wind speed demonstrates virtually no linear correlation (0.03), though this may mask non-linear effects observed in the scatter plots. These quantitative results confirm that among weather variables, humidity has the most significant direct relationship with air quality, while other meteorological factors play more subtle or indirect roles.



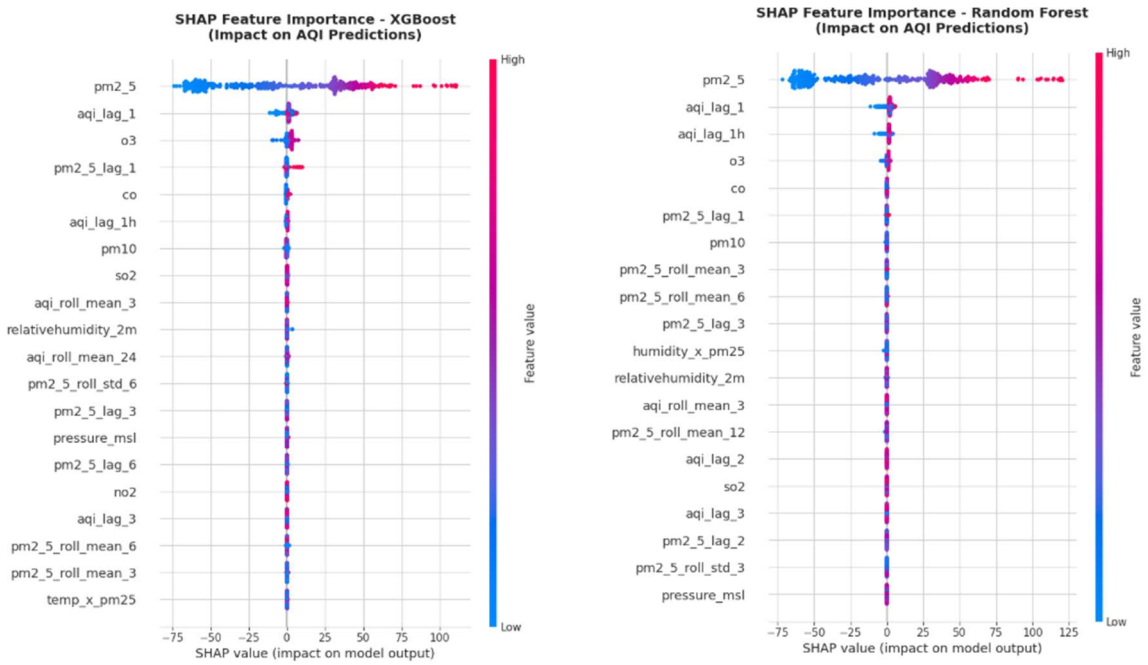
FEATURE IMPORTANCE

We conducted SHAP (SHapley Additive exPlanations) analysis to quantify the contribution of each feature to AQI predictions using Random Forest and XGBoost models on 500 test samples.

Key Findings

1. PM2.5 Dominates Predictions

PM2.5 exhibits the highest SHAP value (~40), significantly outweighing all other features. The dependence plot reveals a strong positive relationship: as PM2.5 increases, its impact on AQI predictions increases almost linearly, confirming particulate matter as the primary driver of air quality.



2. Temporal Persistence is Critical

Recent AQI history features (aqi_lag_1, aqi_lag_1h) show strong importance, indicating air quality exhibits temporal autocorrelation. Current AQI is heavily influenced by recent past values, validating the use of lag features in forecasting.

3. Other Pollutants Contribute Moderately

O3, CO, PM10, and SO2 demonstrate moderate SHAP values (2-5), serving as secondary predictors. Their dependence plots show non-linear relationships, particularly for O3 which exhibits varying impact across different concentration ranges.

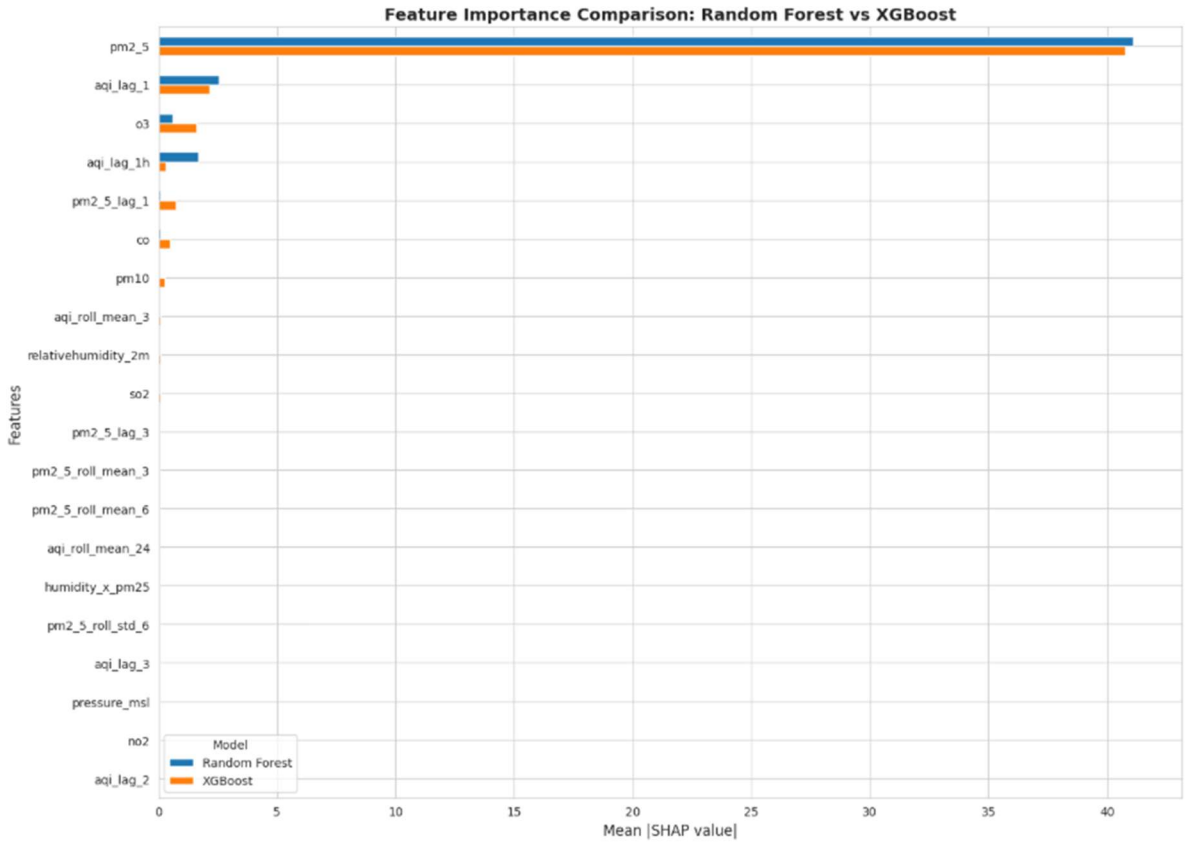
4. Feature Category Analysis

By category, **Pollutants** account for the highest total importance (sum = 42.6), followed by **Lag Features** (sum = 3.98) and **Rolling Means** (sum = 0.25). Weather features show minimal impact

(sum = 0.11), while temporal features (hour, day of week, month) contribute negligibly (sum = 0.03).

5. Model Consistency

Feature importance rankings are highly consistent between Random Forest and XGBoost models. PM2.5, aqi_lag_1, and O3 rank in the top 5 for both models, indicating robust feature importance independent of algorithm choice.



Feature Selection Recommendation

Based on SHAP analysis, we identified:

- **15 high-importance features** (top 30%) to retain for modeling
- **11 low-importance features** (bottom 20%) recommended for removal, including temporal cyclical features (hour_cos, dow_sin, dow_cos, month) and redundant long-lag features (pm2_5_lag_12, aqi_lag_24h, pm2_5_lag_72)

This reduction enables **20% dimensionality reduction** while preserving predictive power, improving model efficiency and reducing overfitting risk.

TOOLS AND TECHNOLOGIES

CHALLENGES FACED

During the development of the Air Quality Prediction Application, several technical challenges were encountered, which required careful troubleshooting and strategic decisions to ensure the reliability and performance of the system.

1. Feature Store Implementation

Initially, **Hopsworks** was used as the feature store to manage and serve features for model training and inference. However, during implementation, recurring **versioning errors** occurred when retrieving features, causing inconsistencies in the data pipeline. Despite multiple attempts to resolve these issues, the errors persisted due to compatibility and versioning conflicts. To overcome this challenge, the feature store was migrated to **MongoDB**, which provided a **stable, flexible, and scalable solution** for storing both feature data and model metadata. MongoDB allowed for easier data management, seamless integration with ML pipelines, and reliable retrieval of features for training and inference.

2. Automation & Scheduling Issues

Another significant challenge was related to the automation of the pipelines. The **scheduled workflows** for hourly data ingestion and daily model training were not triggering at the expected

times. Extensive debugging was required to identify potential causes, including configuration issues, API rate limits, and scheduling conflicts. After spending significant time troubleshooting, the workflows eventually started functioning correctly. Upon investigation, it was observed that **GitHub Actions workflows sometimes experience activation delays**, especially when newly configured or recently updated, due to internal propagation and initialization processes. This delay was outside the application's control but highlighted the importance of monitoring and validating automated workflows after deployment.

Reflection

These challenges underscored the importance of **choosing robust infrastructure components** and being aware of platform-specific behavior when implementing automated ML pipelines. Overcoming these obstacles improved the stability, reliability, and maintainability of the system.