CS361: Air Quality Index Forecasting

Harsh Katara Himanshi Gautam Dhruv Patel Ridhiman Kaur Dhindsa

Abstract

Forecasting urban air pollution is vital to combat its adverse impacts. Various machine learning methods will be utilized to predict air quality, with this study employing a variety of regression techniques. The aim is to forecast the Air Quality Index for major pollutants such as PM2.5, PM10, CO, NO2, SO2, and O3. Evaluation criteria such as Mean Square Error, Mean Absolute Error, and R-squared error are to be used in determining the accuracy given by the various algorithms.

1. Introduction

This project endeavors to forecast Air Quality Index (AQI), a pivotal metric in assessing environmental health and pollution levels. Through meticulous preprocessing, data cleaning, and exploratory data analysis (EDA), we aim to uncover underlying patterns and insights within the air quality dataset. This serves as a foundational step towards developing accurate forecasting models for AQI, contributing to proactive measures in pollution control and public health management.

1.1. Motivation

Central to our efforts is a shared concern – the pressing challenge of air pollution and its profound impacts on public health and the environment. Through our exploration of Air Quality Index (AQI) forecasting, we are motivated by the imperative to equip stakeholders with actionable insights for effective pollution mitigation strategies. By meticulously analyzing data through preprocessing, cleaning, and exploratory techniques, our objective is to elucidate the intricate dynamics of air quality. Ultimately, we aspire to provide decision-makers with evidence-based strategies and interventions to combat pollution and facilitate the creation of cleaner, healthier environments for all.

Recent Reports: Guwahati was reported to be the 2nd most polluted city in the year of 2023 by IQAir, a Swiss Air Quality technology company with average PM2.5 concentration of 105.04 micrograms per cubic meter, which was 21 times the WHO guidelines.
Guwahati with an estimated population of 13 lakh was far cleaner in 2022 when its average annual PM2.5 concentration 51 (μg/m³). The report says, "Causing an estimated one in every nine deaths worldwide, air pollution is the greatest environmental threat to human health. According to the World Health Organization (WHO), air pollution is responsible for an estimated seven million premature deaths worldwide every year.

2. Methods

2.1. Data Preparation

The data is acquired from the Central Pollution Control Board (CPCB) pertaining to pollutant concentrations. The primary objective of this analysis was the meticulous calculation of the Air Quality Index (AQI) based on the acquired data. Methodologically, the data underwent rigorous processing to ensure accuracy and consistency, followed by AQI calculation utilizing prescribed formulae and breakpoints provided by regulatory authorities. The resultant AQI values were categorized into distinct air quality levels, ranging from 'Good' to 'Severe,' aligning with established standards.

- The AQI calculation uses 7 measures: PM2.5, PM10, SO2, NOx, NH3, CO and O3.
- For PM2.5, PM10, SO2, NOx and NH3 the average value in last 24-hrs is used with the condition of having at least 16 values.
- For CO and O3 the maximum value in last 8-hrs is used.
- Each measure is converted into a Sub-Index based on pre-defined groups.
- Sometimes measures are not available due to lack of measuring or lack of required data points.
- Final AQI is the maximum Sub-Index with the condition that at least one of PM2.5 and PM10 should be available and at least three out of the seven should be available.

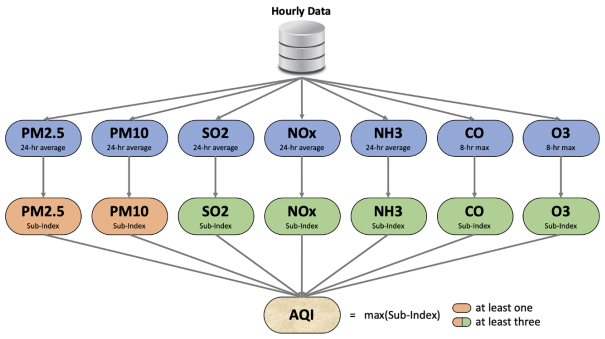

Figure 1. How AQI is calculated

- There is no theoretical upper value of AQI but it is rare to find values over 1000.
- The AQI values are divided into buckets.
- The pre-defined buckets of AQI are as follows:

Good (0–50)	Minimal Impact	Poor (201–300)	Breathing discomfort to people on prolonged exposure
Satisfactory (51–100)	Minor breathing discomfort to sensitive people	Very Poor (301–400)	Respiratory illness to the people on prolonged exposure
Moderate (101–200)	Breathing discomfort to the people with lung, heart disease, children and older adults	Severe (>401)	Respiratory effects even on healthy people

Figure 2. AQI Buckets

2.2. Data Analysis

This section presents a comprehensive analysis of the air quality dataset using statistical methods and visualizations. Various analyses, including seasonality, correlation, weekly patterns, scatter plots, and trend assessments, are conducted using pandas, matplotlib, and seaborn. These analyses provide valuable insights into air quality dynamics and trends.

2.2.1. Seasonal Analysis

We analysed the calculated AQI dataset for four different stations namely: IIT Guwahati,Pan bazaar, Guwahati airport and railway colony for each month seasonally.
Observation:

- There seems to be a peak in pollution during winter season in Dec-Jan. The main reason for this could be that cold air is denser and moves slower than warm air. This density means that cold air traps the pollution but also doesn't whisk it away.
- There is a peak in August at Airport in the graph but this is not a valid observation as it is generated because not enough data is available for that month.

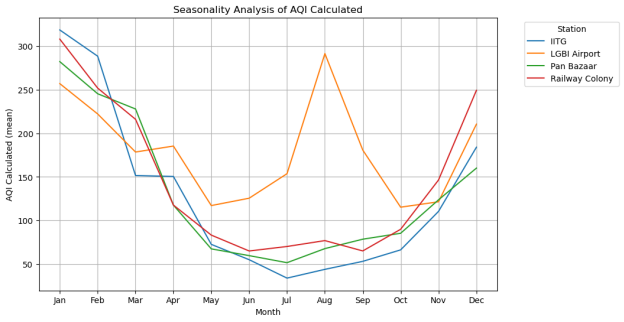

Figure 3. Seasonal AQI

2.2.2. Correlation Analysis of Air Quality Indicators

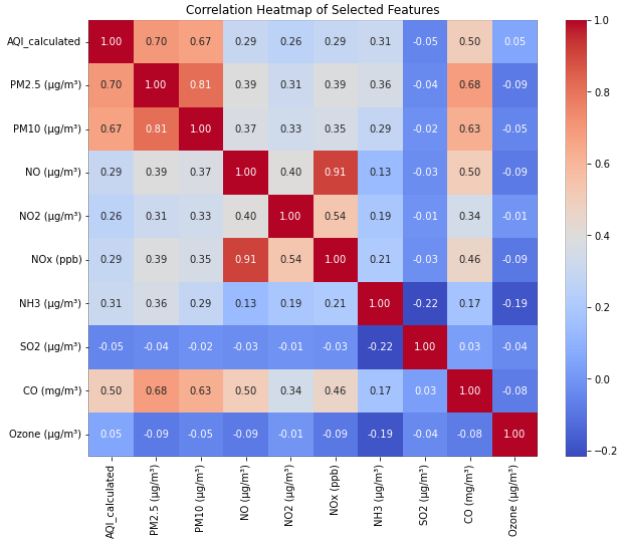

Figure 4. Weekly Analysis

The correlation heatmap (Figure 4) illustrates pairwise relationships between air quality indicators. Key observations include:

- **Positive Correlation:** Strong positive correlations are observed between PM2.5 and PM10, as well as between NO, NO2, and NOx, and between C0 and PM2.5 indicating shared emission sources.
- **Negative Correlation:** AQI shows a negative correlation with SO2 concentrations. It could suggest that factors such as particulate matter or ozone have a stronger influence on overall air quality, or that localized dispersion patterns play a role in lowering AQI values despite elevated SO2 levels.

- **Weak Correlation:** Some pollutants exhibit weak correlations, such as NH3 with other pollutants, suggesting independent sources or atmospheric behaviors.

2.2.3. Scatter plot Analysis

We have created scatter plots for concentration of each pollutant v/s AQI and analysed their relationship.
Observation:

- PM10 has a limit of 1000, sensor cannot measure more than 1000.This is evident from the scatter plot of PM10.

- PM10 and PM2.5 are the major Pollutants.

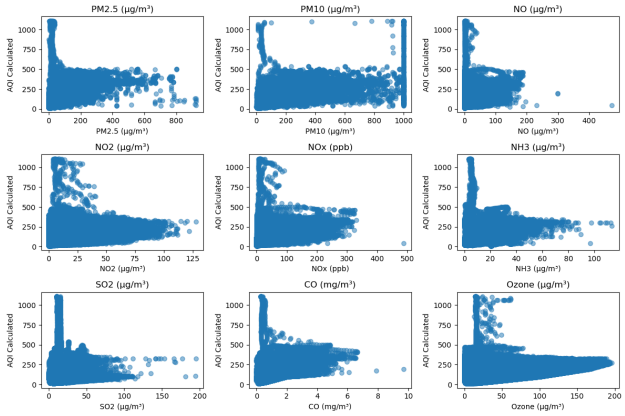

Figure 5. Scatter plots

2.2.4. Weekly Analysis

We analysed the calculated AQI dataset for four different stations namely: IIT Guwahati,Pan bazaar, Guwahati airport and railway colony for day of the week and compared them.
Observation

- At Airport,Railway colony and Pan Bazaar pollution peaks in weekends whereas in IITG pollution doesn't the reason behind this could be that in weekend the former three station are visited more frequently by people and are more crowed. Like due to more frequency of flights the pollution at airport increases. Frequency of vehicles on road increases which leads to more traffic which in turn increases AQI in pan Bazaar.

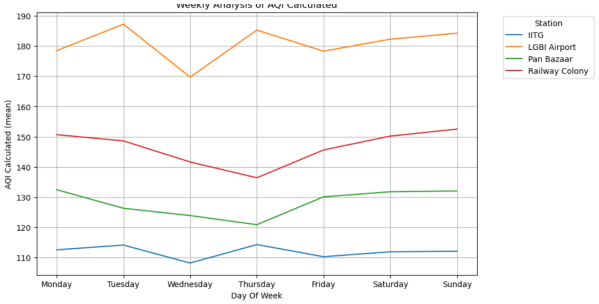

Figure 6. Weekly Analysis

2.2.5. Data availability and Trend Analysis

In this section, we evaluate data availability and trends in Air Quality Index (AQI) values across monitoring stations over time. We scrutinize data distribution to ensure reliability and pinpoint any gaps or irregularities.

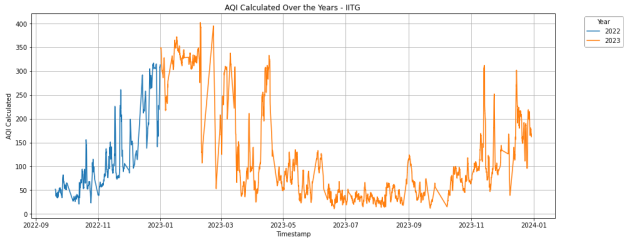

Figure 7.

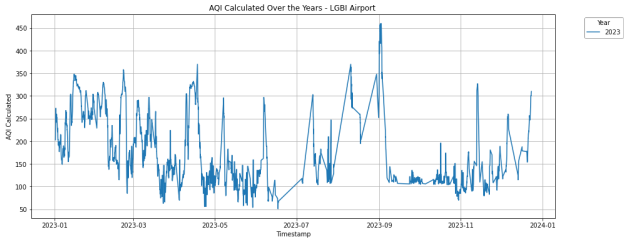

Figure 8.

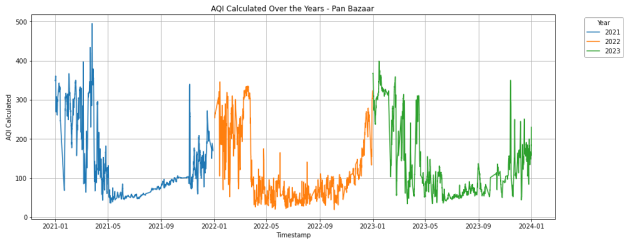

Figure 9.

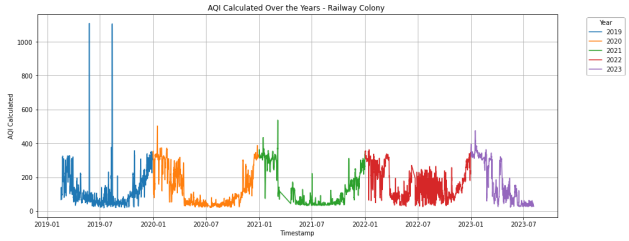

Figure 10.

3. Progress

The milestones achieved in this report include data preparation and exploratory data analysis. The data preparation process included searching for appropriate air pollution based datasets in the Indian context, with special emphasis on Guwahati. It also comprised of data cleaning operations which removed redundant columns not required for AQI calculation, such as concentrations of subsidiary pollutants- Benzene, Toluene, Ozone etc, as well as atmospheric parameters- Temperature, Relative Humidity, Barometric Pressure etc. Following this, AQI calculation was performed over 15-minute intervals and NaN values were dealt with appropriately.

The process of exploratory data analysis revealed that AQI calculations were available for four locations in Guwahati, namely- LGBI airport, IITG, Pan Bazaar and Railway station. Various line graphs and scatter plots were used to make observations about trends in the data from seasonal, location-based, correlational and temporal perspectives.

The proposed direction of the project post-midterm is to explore various statistical, ensemble and kernel methods in order to predict AQI in Guwahati. This is particularly relevant as the city was recently declared the world's second most polluted city, primarily due to high PM2.5 concentrations.

4. Conclusion

Our exploratory data analysis (EDA) of the Guwahati AQI dataset across four stations has unveiled crucial insights into regional air quality dynamics. Distinct patterns emerged, with Airport, Railway Colony, and Pan Bazaar experiencing higher pollution levels on weekends, likely due to increased human activity. Scatter plots highlighted measurement limitations, particularly with PM10 concentrations near saturation, while the correlation heatmap underscored significant relationships among pollutants.

Notably, a peak in pollution during winter months and an anomaly in August at the Airport station were observed, indicating seasonal variability and data quality concerns. These findings underscore the complexity of air quality dynamics and highlight the importance of considering external factors and data integrity in analysis.

In summary, our EDA lays a solid foundation for further research and policy interventions aimed at improving air quality and public health in the Guwahati region.

5. Resources and Links

- Dataset Source
- Pollution Report- Guwahati, The Times of India

References

Bhattacharya, S. and Shahnawaz, S. Using machine learning to predict air quality index in new delhi. Technical report, Jadavpur University, 2021. URL https://arxiv.org/abs/2112.05753.

Kaur J., S. S. and S., P. K. Forecasting of aqi (pm2. 5) for the three most polluted cities in india during covid-19 by hybrid daubechies discrete wavelet decomposition and autoregressive (db-dwd-arima) model. Technical report, Environmental Science and Pollution Research, 2023. URL https://doi.org/10.1007/s11356-023-29501-w.

Kumar, A. and Goyal, P. Forecasting of daily air quality index in delhi. Technical report, Science of The Total Environment. URL https://www.sciencedirect.com/science/article/pii/S0048969711009661.