CS361 Course Project

Air Quality Alert System for Guwahati

Group Name: Frequentists
Harsh Katara, 210101045
Himanshi Gautam, 210101051
Dhruv Patel, 210101075
Ridhiman Kaur Dhindsa, 210101088

Introduction and Motivation

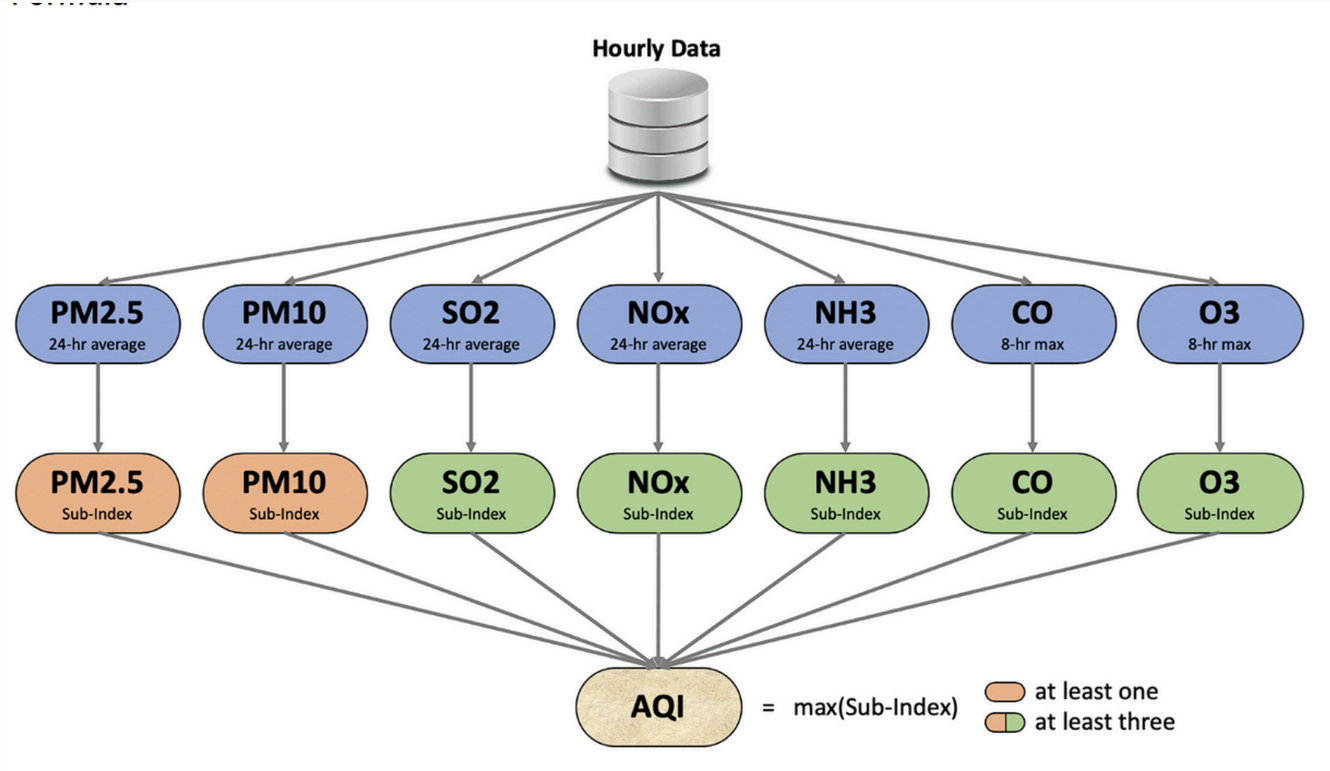
- Urban air pollution forecasting is crucial for mitigating its detrimental effects. Our project focuses on using Machine Learning methods to develop an alert system for the Air Quality Index (AQI),which is a critical metric for assessing pollution levels and environmental health.
- Motivation: Predicting AQI would enable us to provide timely information that enables authorities to take proactive measures and safeguards public health.
- 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.
- This study employs feature engineering and various regression techniques to predict the Air Quality Index (AQI) for key pollutants including PM2.5, PM10, CO, NO2, SO2, and O3. Various ML models are implemented both from scratch as well as utilizing library functions, and their results are compared.
- Evaluation metrics including Mean Square Error and a self-defined metric, Downside Error are employed to assess model accuracy.

Dataset Acquisition and Preparation

We have acquired data from the Central Pollution Control Board (CPCB) pertaining to pollutant concentrations for 4 stations in the city of Guwahati namely LGBI Airport, Pan Bazaar, IIT Guwahati, Railway Colony. We deal with specifically the pollutants PM2.5, PM10, SO2, NOx, NH3, CO and O3. Data was provided with a granularity of 15 min timestamps. We preprocessed the data and calculated AQI for each timestamp according to the standard AQI calculation method provided by CPCB. To complete preparing the Dataset for our prediction problem generate the target column by shifting the AQI by 15 days for each station,

AQI Calculation Method

The method we have used for calculating AQI for our dataset is is visualized below



Problem Statement

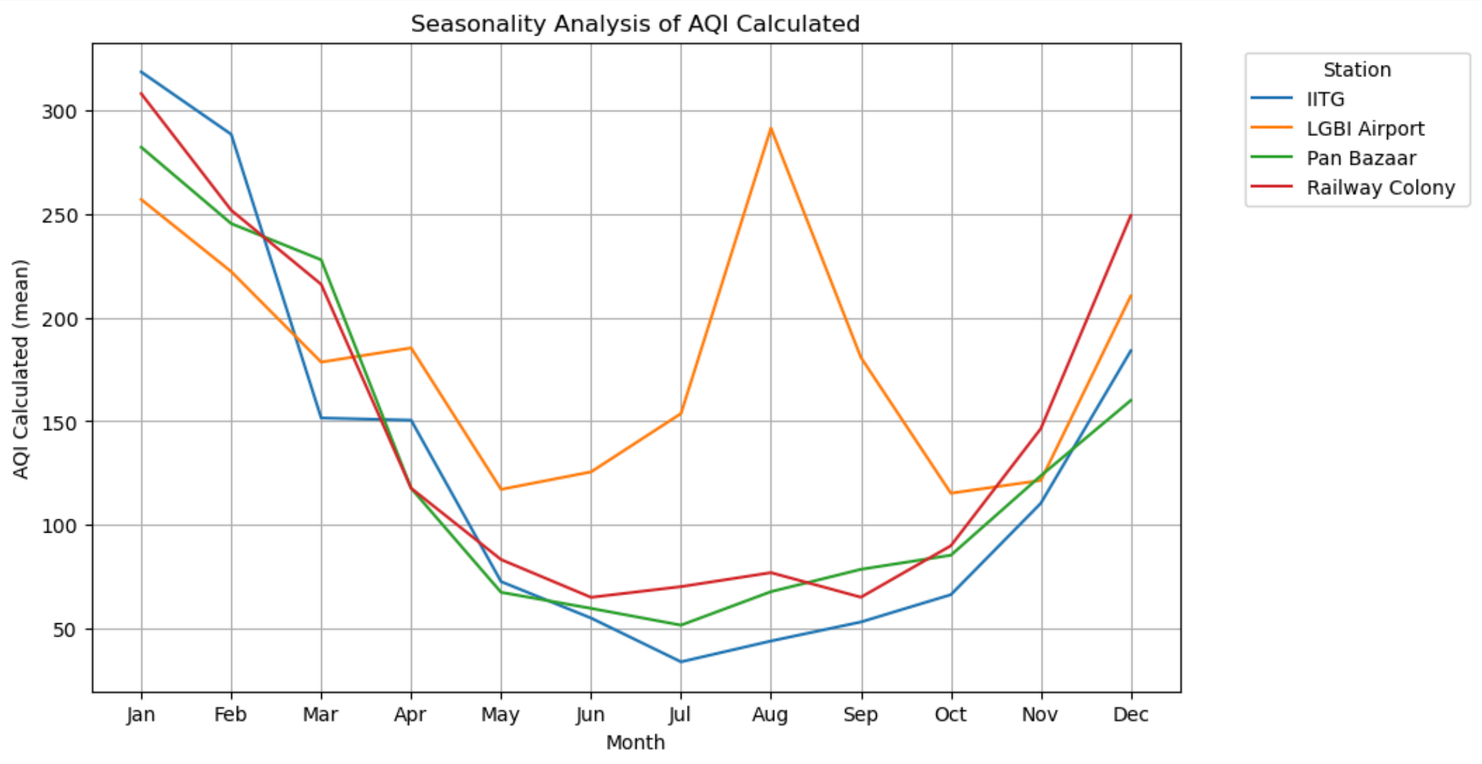
The rapidly deteriorating air quality in Guwahati poses significant challenges to public health and environmental sustainability. With the city being reported as the 2nd most polluted in 2023, there is an urgent need for an effective Air Quality Alert System that can provide timely and accurate forecasts of the Air Quality Index (AQI). We have taken on the problem to generate alerts for 15 days prediction of AQI. The task is meticulous as it involves analyzing various features and models used for prediction. We have also came up with a new metric for evaluating models known as down_side_error.

Exploratory Data Analysis (EDA)

We analyzed our air quality dataset using statistical methods and visualizations in the following ways:

- Seasonality Analysis
- Correlation Analysis of Air Quality Indicators
- Scatter Plot Analysis
- Weekly Analysis

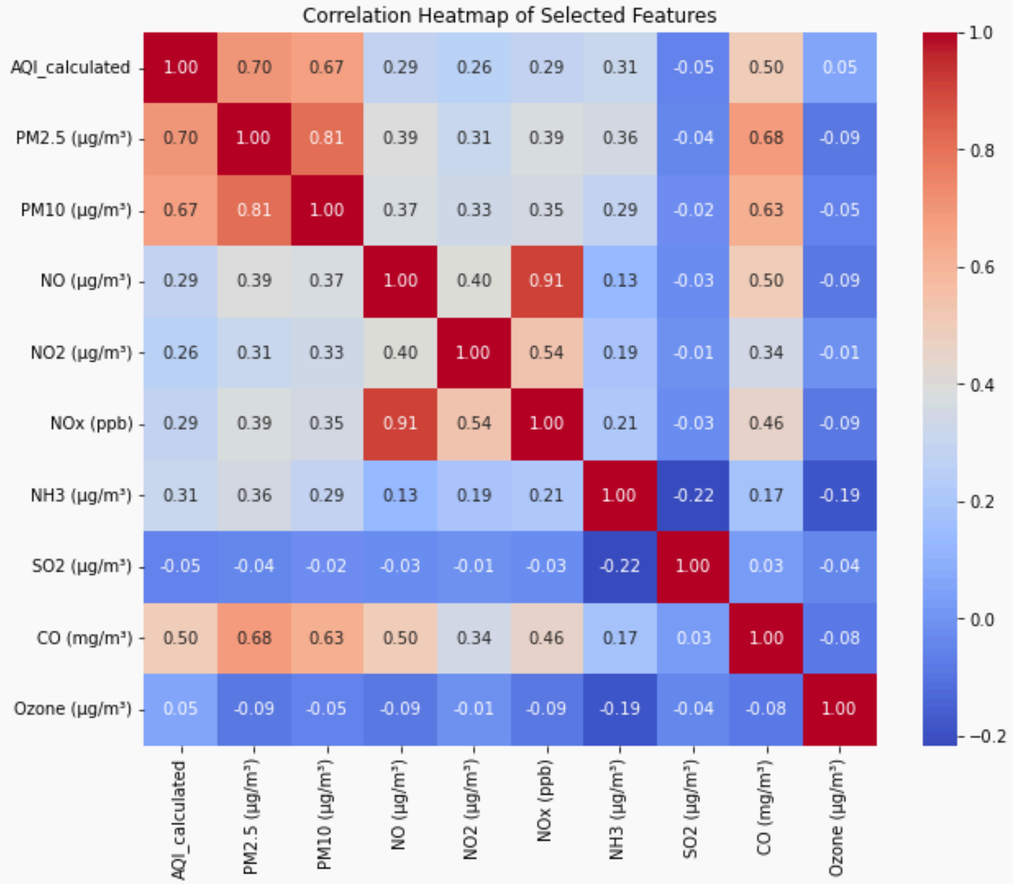
Seasonality Analysis



Observations::

- There seems to be a peak in pollution during winter season in Dec-Jan.
- There is a peak in August at Airport in the graph because not enough data is available for that month

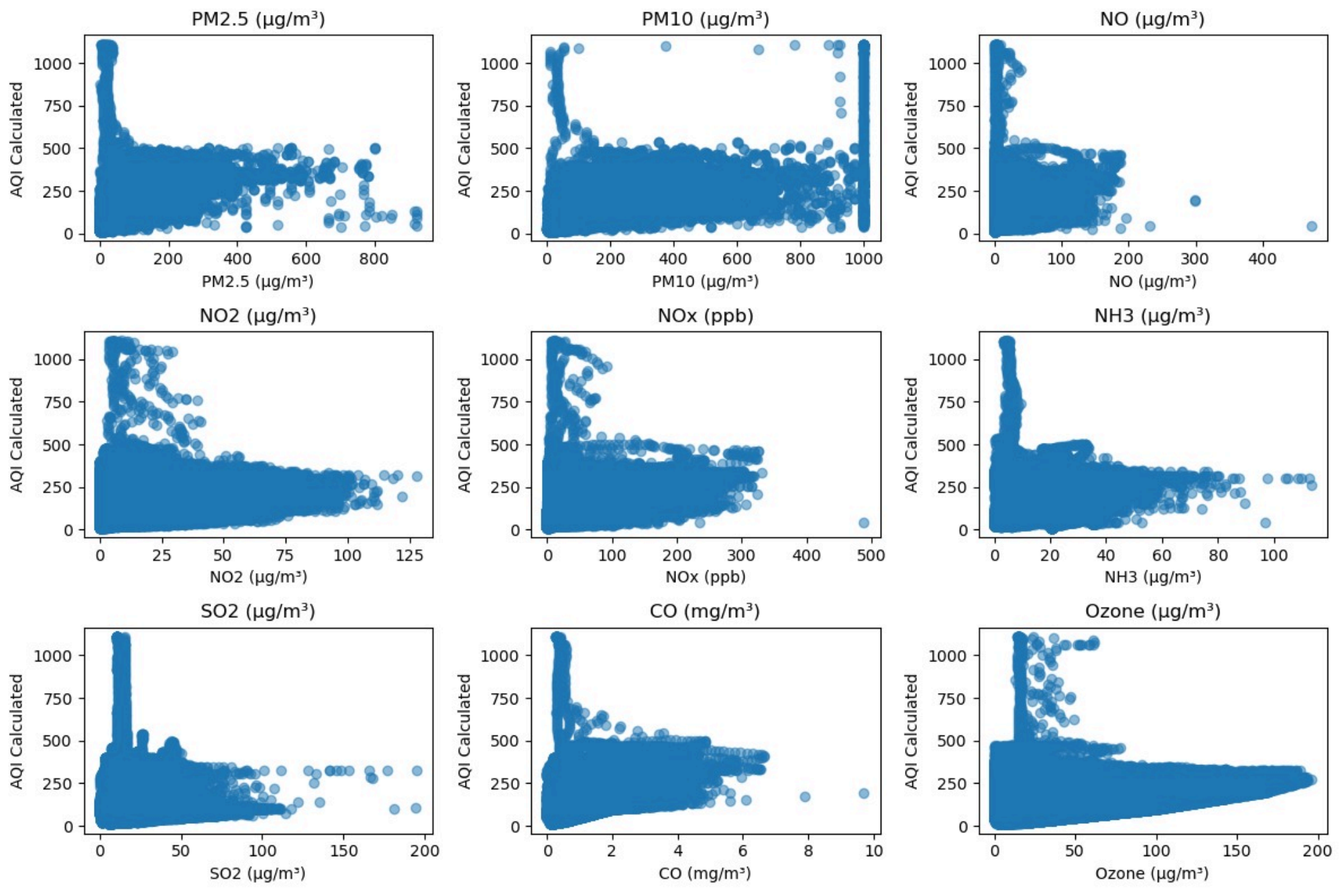
Correlation Analysis for Air Quality Indicators



Observations::

- **Positive Correlation:** Strong positive correlations are observed between PM2.5 and PM10.
- **Negative Correlation:** AQI shows a negative correlation with SO2 concentrations.
- **Weak Correlation:** Some pollutants exhibit weak correlations, such as NH3 with other pollutants.

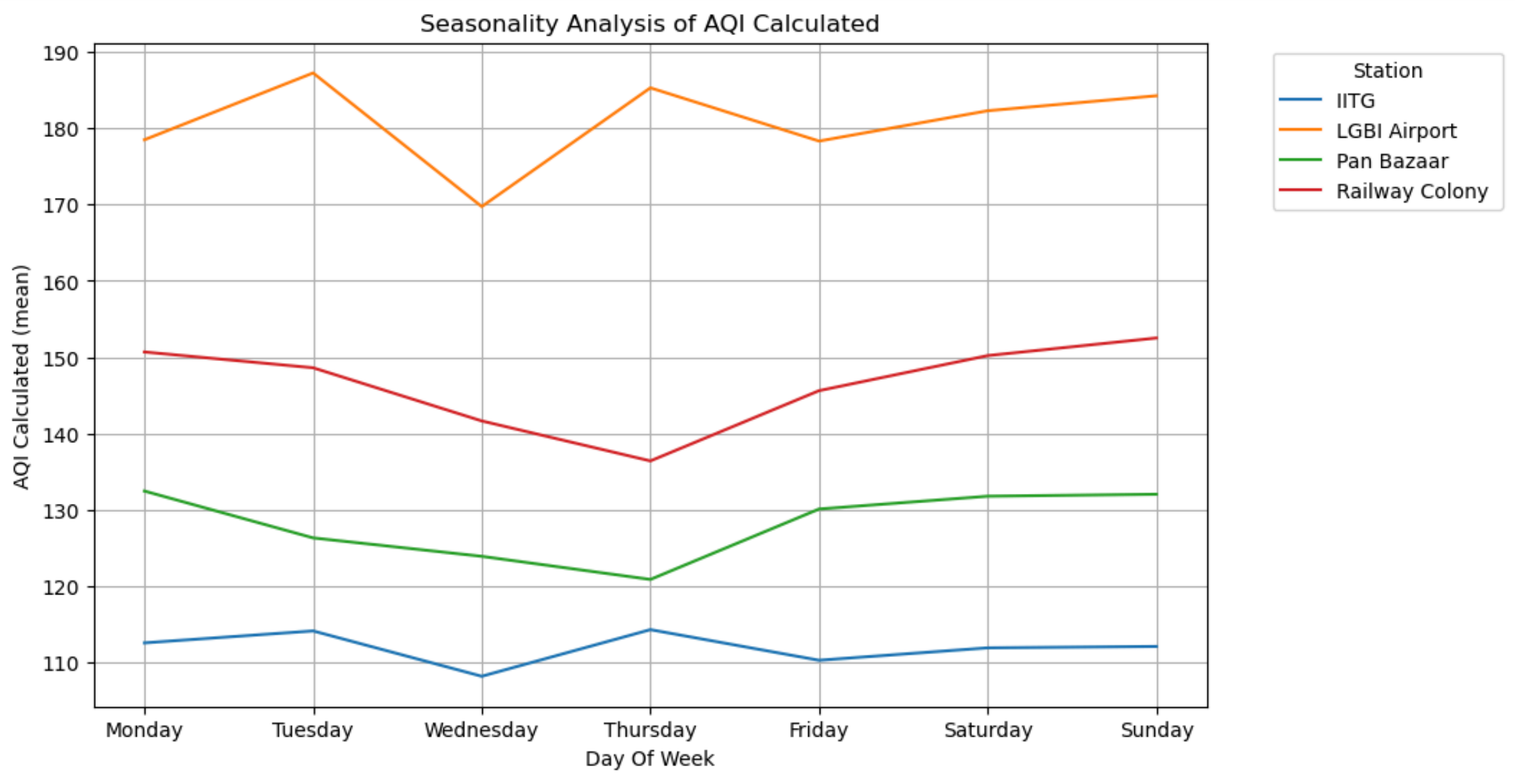
Scatter Plot Analysis



Observations:

- 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.

Weekly Analysis



Observations:

- At Airport,Railway colony and Pan Bazaar pollution peaks in weekends whereas in IITG pollution doesn’t.

Feature Engineering

After analyzing the variation of AQI with various attributes of dataset we extract additional features from our dataset to improve performance of our models.

We notice AQI varies significantly upon :

- Station
- Month
- Day of the Week
- Year

We added Onehot Encodings for stations and extracted the rest features from timestamp to be given as an input to our models. We tested both on non feature engineered dataset and feature engineered dataset and noticed significant improvements in performance.

PCA

We explored the utility of Principal Component Analysis (PCA) for dimensionality reduction, but observed a degradation in results, indicating its limited effectiveness in this context. As loss in explainable variance increases from the original dataset our models incur a noticeable decrease in performance. Hence we did not use it for feature selection in our dataset.

Methods used

Linear Regression

A statistical method to model the relationship between a dependent variable and one or more independent variables, aiming to minimize the difference between observed and predicted values using a straight line.

Lasso Regression

A linear regression technique that adds a penalty to the absolute values of the coefficients. It encourages sparsity by shrinking some coefficients to zero, effectively performing feature selection and improving model interpretability.

Ridge Regression

A linear regression technique that adds a penalty to the squared magnitudes of the coefficients. It shrinks coefficients and mitigates multicollinearity, improving model generalization and stability at the cost of some bias.

Decision Tree

A supervised learning algorithm that organizes data into a tree-like structure of decisions and their possible outcomes. It splits the data based on features to make sequential decisions, leading to a final prediction.

Random Forest

An ensemble learning method for classification and regression tasks. It constructs multiple decision trees and reduces overfitting, increasing accuracy by introducing randomness in tree construction and feature selection.

Gradient Boosting

An ensemble learning technique that builds a strong predictive model by combining multiple decision trees in a sequential manner, minimizing errors by fitting each new tree to residuals of previous ones.

Result without Feature Engineering

Model Name	Mean Square Error		Root Mean Square Error		Down side Error	
	Our Result	Sklearn Result	Our Reult	Sklearn Reult	Our Result	Sklearn Result
Linear Regression	6661.4984	6661.4984	81.6180	81.6180	4242.0040	4242.0040
Lasso Regression	6674.4869	6677.4766	81.6975	81.7158	4274.3934	4243.1611
Ridge Regression	6661.2710	6660.3983	81.6166	81.6112	4258.0875	4240.2167
Decision Tree	2722.1665	2899.0544	52.1744	53.8428	1473.2392	1544.4888
Random Forest	1287.0119	1277.8464	35.8749	35.7469	551.9467	547.3786
Gradient Boosting	1174.7717	1344.9871	34.2749	36.6740	709.4935	752.5551

Result with Feature Engineering

Model Name	Mean Square Error		Root Mean Square Error		Down side Error	
	Our Result	Sklearn Result	Our Reult	Sklearn Reult	Our Result	Sklearn Result
Linear Regression	6530.9969	6530.9969	80.8145	80.8145	4173.4740	4173.4740
Lasso Regression	6548.6875	6554.4219	80.9239	80.9594	4213.2328	4181.7961
Ridge Regression	6529.6065	6528.7336	80.8059	80.8005	4186.7543	4168.8758
Decision Tree	587.8481	782.5233	24.2455	27.9736	297.8543	345.6849
Random Forest	343.5356	339.0998	18.5347	18.4146	127.8588	124.0062
Gradient Boosting	498.2000	404.4660	22.3203	20.1113	318.2624	224.5465

Conclusion

After conducting extensive data preprocessing and exploratory data analysis, we used various models to predict AQI for Guwahati across four stations. Random forest emerged as the optimal model, outperforming others. While we explored Principal Component Analysis for dimensionality reduction, but we observed degradation in results so we decided to not use it.

References

- Using Machine Learning to Predict Air Quality Index in New Delhi
- Forecasting of aqi (pm2. 5) for the three most polluted cities in india during covid-19
- Forecasting of daily air quality index in delhi. Science of The Total Environment
- Dataset Source
- Pollution Report - Guwahati, The Times of India