CS4048 Data Science
Final Report AQI Value Prediction

Muhammad Hassan Muzaffar
Mohammad Waleed Ikram
Syed Ahmer Zaidi

I. Introduction

Air pollution is one of the most pressing environmental and public health challenges affecting urban areas worldwide. The Air Quality Index (AQI) serves as a standardized indicator to communicate air pollution levels and their potential health effects to the public. High levels of pollutants such as PM2.5, PM10, O_3 (Ozone), NO_2 (Nitrogen Dioxide), CO (Carbon Monoxide) and SO_2 (Sulfur Dioxide) are known to contribute significantly to poor air quality, posing health risks such as respiratory illnesses and cardiovascular diseases. Monitoring and predicting AQI trends are essential for informing the public, guiding policymaking, and implementing timely interventions to reduce pollution exposure.

A. Motivation

To address the growing need for effective air quality management, this project focuses on developing a predictive system for the Air Quality Index (AQI) using data collected from the IQAir Websdite [1]. The dataset includes key pollutants such as PM2.5, PM10, O_3 (Ozone), NO_2 (Nitrogen Dioxide), CO (Carbon Monoxide), and SO_2 (Sulfur Dioxide), along with meteorological factors like temperature, humidity, and wind speed. By leveraging these features, the project aims to build an advanced machine learning system capable of forecasting AQI for major global cities. Accurate AQI prediction is essential for enabling timely health advisories, formulating pollution control policies, and supporting climate action strategies. This project utilizes modern Machine Learning and Deep Learning algorithms such as Random Forest, Gradient Boosting, Linear Regression and Dense neural networks to improve prediction accuracy. By exploring the relationships between pollutants and AQI, this project seeks to provide valuable insights that can guide decision-making for both policymakers and the general public.

B. Data set Description

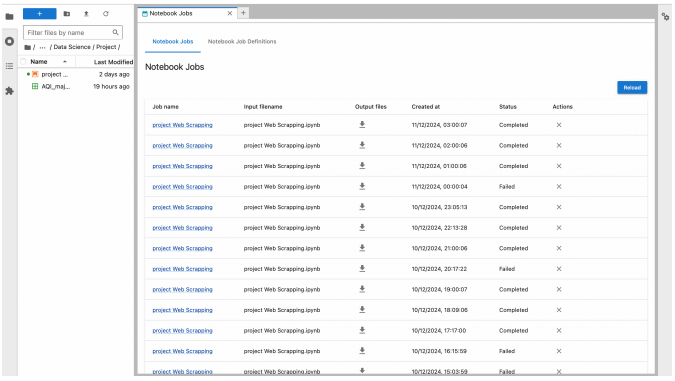
The dataset used for this project is scrapped from the IQAir Website [1], which provides comprehensive air quality data for major global cities. The dataset includes both pollutant concentrations and meteorological features that influence air quality. Below is a description of the key features included in the dataset:

a) *City:* The name of the city for which air quality data is recorded. This feature allows the model to differentiate between different urban environments, as pollution sources and weather conditions vary from city to city.
b) *Date and Time:* Timestamps for each record, used to capture temporal patterns in pollution levels. Time-based features are crucial for capturing daily, weekly, and seasonal trends in AQI.
c) *PM2.5:* The concentration of fine particulate matter with a diameter of 2.5 micrometers or smaller. PM2.5 is one of the most harmful pollutants due to its ability to penetrate deep into the lungs and bloodstream.
d) *PM10:* The concentration of particulate matter with a diameter of 10 micrometers or smaller. Although less harmful than PM2.5, PM10 still poses health risks, especially for individuals with respiratory conditions.
e) *O_3 (Ozone):* Ground-level ozone, which forms when pollutants react with sunlight. Ozone is a major component of smog and can cause respiratory irritation.
f) *NO_2 (Nitrogen Dioxide):* This pollutant is primarily released from vehicle emissions and industrial processes. It contributes to the formation of ground-level ozone and particulate matter.
g) *SO_2 (Sulfur Dioxide):* Released from the burning of fossil fuels, SO_2 can cause respiratory issues and contribute to the formation of fine particulate matter.
h) *CO (Carbon Monoxide):* A colorless, odorless gas that is produced by incomplete combustion of fossil fuels. Exposure to high CO levels can be hazardous to human health.
i) *Humidity:* Represents the amount of moisture in the air, which can influence the concentration and dispersion of certain pollutants.
j) *Temperature:* Temperature influences the formation of pollutants like ozone, as warmer temperatures accelerate photochemical reactions.

k) *Wind Speed:* Wind plays a crucial role in dispersing pollutants and clearing them from urban areas. Higher wind speeds are often associated with improved air quality.

These features were selected due to their strong influence on AQI prediction. By including weather-related factors alongside pollutant concentrations, the system can capture the complex interactions that drive air quality changes.

1) Data Collection: The data for this project was collected from the IQAir website [1] through an automated web scraping process. To extract data efficiently, we crated a .ipynb file which utilized techniques like Selenium and BeautifulSoup, along with ChromeDriver to enable browser automation. This approach allowed for real-time scraping of AQI and pollutant data from the website. The process was automated using Jupyter Scheduler within JupyterLab, ensuring that the scraping task was executed every hour from December 5 to December 11. The collected data includes information on key pollutants such as PM2.5, PM10, O_3, NO_2, SO_2, and CO, as well as weather-related factors like temperature, humidity, and wind speed. To ensure data integrity and continuity, log files were maintained for each scraping session as evidence of successful data collection.We have scrapped the website for 102 times. This method of data collection provided a rich and dynamic dataset that enabled the development of an accurate and robust AQI prediction system.



2) Data Preparation:

a) *Data Loading:* The dataset, comprising 4822 rows and 13 columns, was loaded using the panda's read_csv () function. Methods like info (), head (), tail (), and describe () provided initial insights into data types and summary statistics.

b) *Data Exploration:* The dataset includes key pollutants such as PM2.5, PM10, O_3 (Ozone), NO_2 (Nitrogen Dioxide), CO (Carbon Monoxide), and SO_2 (Sulfur Dioxide), along with meteorological factors like temperature, humidity, and wind speed. Visual inspection and summary statistics helped in understanding the distributions and ranges of these features.

i) df.info ()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4822 entries, 0 to 4821
Data columns (total 13 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   City       4822 non-null   object
 1   Date       4822 non-null   object
 2   Time       4822 non-null   object
 3   Wind       4822 non-null   object
 4   Humidity   4822 non-null   object
 5   Weather    4822 non-null   object
 6   PM2.5      4822 non-null   float64
 7   PM10       3356 non-null   float64
 8   O₃         3230 non-null   float64
 9   NO₂        3269 non-null   float64
 10  SO₂        2576 non-null   float64
 11  CO         2467 non-null   object
 12  AQI        4822 non-null   int64
dtypes: float64(5), int64(1), object(7)
memory usage: 489.9+ KB
```

ii) df.describe ()

	PM2.5	PM10	O_3	NO_2	SO_2	AQI
count	4822.000000	3356.000000	3230.000000	3269.000000	2576.000000	4822.000000
mean	29.461012	39.139303	35.712291	33.419027	7.932376	80.710701
std	35.875902	49.819774	33.630769	27.771785	14.242155	54.667613
min	0.000000	0.000000	0.000000	0.000000	0.000000	6.000000
25%	7.700000	11.500000	11.000000	15.500000	2.000000	44.000000
50%	17.000000	22.000000	28.500000	26.000000	4.300000	66.000000
75%	35.000000	45.025000	49.375000	45.000000	8.200000	99.000000
max	303.700000	529.500000	223.400000	249.700000	162.000000	457.000000

iii) df.head ()

	City	Date	Time	Wind	Humidity	Weather	PM2.5	PM10	O_3	NO_2	SO_2	CO	AQI
0	Dhaka	05 December 2024	07:00	0 km/h	88%	17°	268.0	NaN	10.0	NaN	NaN	NaN	386
1	Kolkata	05 December 2024	06:30	0 km/h	82%	18°	138.8	278.3	8.5	36.5	9.2	1,600.0	214
2	Lahore	05 December 2024	06:00	0 km/h	76%	10°	132.0	NaN	NaN	NaN	NaN	NaN	207
3	Ulaanbaatar	05 December 2024	09:00	3.6 km/h	84%	-22°	119.7	205.5	0.1	53.9	89.5	3,250.0	196
4	Karachi	05 December 2024	06:00	11.1 km/h	67%	14°	105.5	NaN	NaN	NaN	NaN	NaN	186

c) *Data Cleaning:* To ensure the accuracy and reliability of the dataset, several data cleaning techniques were applied. Missing values were identified and addressed by replacing them with zeros. This approach was chosen because, in air quality datasets, a missing pollutant reading often indicates that the pollutant concentration was negligible or not recorded due to equipment constraints. Duplicate records were also removed to avoid redundancy, ensuring that only unique records were retained for analysis and model training. Additionally, outliers were detected and handled using a combination of Z-score and Interquartile Range (IQR) methods. Z-score was used to identify values that were more than three standard deviations from the mean, while the IQR method flagged data points lying outside the range defined by 1.5 times the interquartile range (IQR) from the first and third quartiles. These cleaning methods improved the quality of the dataset and enhanced the robustness of the AQI prediction model.

d) *Data Transformation:* To prepare the dataset for analysis and modeling, several data transformation steps were performed. The Date and Time columns were merged into a single datetime column to facilitate time-based analysis. Pollutant features like CO were converted to numeric format to ensure consistency, while Wind, Humidity, and Weather values were converted from string representations to numerical values. Additionally, a new variable called

AQI_category was created to classify AQI levels into categories such as "Good," "Moderate," and "Unhealthy." The City feature was encoded into numerical labels to enable machine learning models to process it effectively. Finally, standardization was applied to ensure that features with different scales had a mean of zero and a standard deviation of one, improving the convergence of machine learning algorithms.

i) df.info ()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4822 entries, 0 to 4821
Data columns (total 12 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   City      4822 non-null   object
 1   Wind      4822 non-null   float64
 2   Humidity  4822 non-null   float64
 3   Weather   4822 non-null   int64
 4   PM2.5     4822 non-null   float64
 5   PM10      3356 non-null   float64
 6   O₃        3230 non-null   float64
 7   NO₂       3269 non-null   float64
 8   SO₂       2576 non-null   float64
 9   CO        2174 non-null   float64
 10  AQI       4822 non-null   int64
 11  datetime  4822 non-null   datetime64[ns]
dtypes: datetime64[ns](1), float64(8), int64(2), object(1)
memory usage: 452.2+ KB
```

ii) df.describe ()

	Wind	Humidity	Weather	PM2.5	PM10	O_3	NO_2	SO_2	CO	AQI
count	4822.000000	4822.000000	4822.000000	4822.000000	3356.000000	3230.000000	3269.000000	2576.000000	2174.000000	4822.000000
mean	11.573725	0.716869	11.197428	29.461012	39.139303	35.712291	33.419027	7.932376	399.193974	80.710701
min	0.000000	0.050000	-27.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	6.000000
25%	5.500000	0.590000	3.000000	7.700000	11.500000	11.000000	15.500000	2.000000	229.000000	44.000000
50%	9.300000	0.770000	10.000000	17.000000	22.000000	28.500000	26.000000	4.300000	379.500000	66.000000
75%	16.700000	0.870000	20.000000	35.000000	45.025000	49.375000	45.000000	8.200000	572.500000	99.000000
max	63.000000	1.000000	34.000000	303.700000	529.500000	223.400000	249.700000	162.000000	996.800000	457.000000
std	8.740096	0.193969	10.537948	35.875902	49.819774	33.630769	27.771785	14.242155	252.584719	54.667613

iii) df.head ()

	City	Wind	Humidity	Weather	PM2.5	PM10	O_3	NO_2	SO_2	CO	AQI	datetime
	Chengdu	10.8	0.62	9	25.0	43.0	37.0	30.0	4.0	400.0	81	2024-12-06 23:00:00
	Seattle	7.4	0.89	3	7.0	10.0	9.4	27.1	0.3	266.1	38	2024-12-06 08:00:00
	Dhaka	0.0	0.72	18	116.9	0.0	14.0	0.0	0.0	0.0	194	2024-12-07 00:00:00
	Osaka	11.1	0.70	5	7.0	9.0	0.0	18.8	10.5	343.5	39	2024-12-07 04:00:00
	Karachi	7.4	0.59	17	105.0	0.0	0.0	0.0	0.0	0.0	186	2024-12-07 00:00:00

C. *Related Works*

The prediction of AQI is a critical task that draws on methods from regression analysis, machine learning, and deep learning. In this project, we utilize models such as Linear Regression, Random Forest, Gradient Boosting, and Dense Neural Networks to predict AQI. Traditional regression methods like Linear Regression are widely used due to their simplicity and interpretability, while tree-based models like Random Forest and Gradient Boosting are known for their ability to handle non-linear relationships between pollutants and AQI [2], [3]. Recent studies have also demonstrated the effectiveness of deep learning models, such as Dense Neural Networks (DNNs), which can capture complex relationships in large datasets [4].

Feature engineering is another key component of AQI prediction. Features like PM2.5, PM10, O_3, NO_2, SO_2, CO, humidity, wind, and temperature are essential in capturing the dynamics of air pollution. Encoding city names using label encoding enables models to consider geographic differences in pollution patterns [5]. Outliers are identified and addressed using Z-score and IQR methods, which remove anomalies that can distort model predictions [6].

The inclusion of machine learning models like Random Forest, Gradient Boosting, and Dense Neural Networks enhances predictive accuracy. These models use ensemble techniques and deep neural architectures to generalize better to unseen data, enabling more reliable AQI prediction.

D. Regression Evaluation Metrices

To evaluate the performance of the AQI prediction models, we used a set of commonly used regression evaluation metrics, namely R-squared (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics provide a comprehensive understanding of the accuracy and error of the prediction models.

R-squared (R^2): This metric measures how well the predicted values fit the actual data. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (1)$$

Mean Absolute Error (MAE): This metric calculates the average absolute difference between the predicted and actual values. It is defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (2)$$

Mean Squared Error (MSE): MSE calculates the average squared difference between the predicted and actual values. It is given by:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (3)$$

Root Mean Squared Error (RMSE): RMSE is the square root of the MSE and provides an error measure in the same units as the AQI. It is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (4)$$

E. AQI Prediction

The task of AQI prediction is to forecast the Air Quality Index (AQI) based on historical pollutant concentrations and meteorological features. The prediction task is treated as a regression problem where the model learns to estimate the AQI value at a specific time and location based on input features like PM2.5, PM10, O_3, NO_2, SO_2, CO, temperature, humidity, and wind speed.

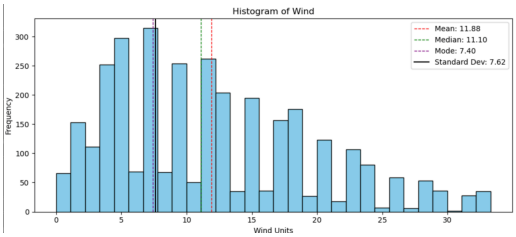
To evaluate the performance of AQI prediction models, we rely on a set of standard regression metrics, namely R-squared (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) and the formulas are Eq. (1), Eq. (2), Eq. (3), Eq. (4).

II. Data Analysis

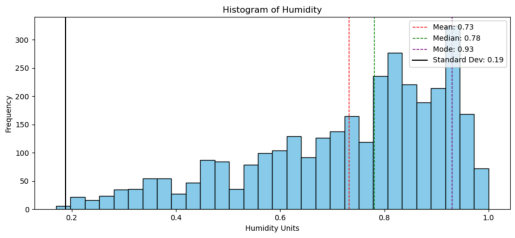
It plays a pivotal role in understanding the structure, patterns, and relationships within the dataset.In this project, univariate analysis was performed to visualize the distribution of individual features, such as PM2.5, PM10, CO, NO_2, SO_2, O_3, temperature, humidity, and wind speed.Histograms was used to detect skewness.Bivariate analysis was conducted to examine relationships between AQI and individual pollutants, revealing key contributors to poor air quality.Correlation analysis was also performed to understand the strength of relationships among features, with a heatmap highlighting significant correlations.By identifying strong correlations, essential features for model development were prioritized. Additionally, pair plots were created to visualize interactions among multiple features simultaneously, enabling the detection of complex relationships.This thorough analysis informed the feature selection process for predictive models.

A. *Univariate Analysis:*

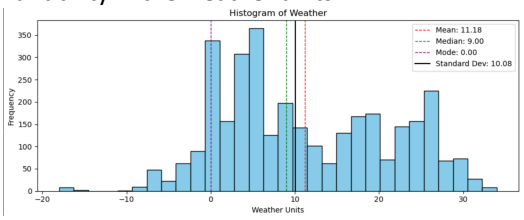
a. Wind: The distribution of wind values, highlighting the frequencies of various wind speeds. The mean (11.88), median (11.10), and mode (7.40) are marked, showing slight skewness in the data, with most values clustering below 15. The standard deviation (7.62) indicates moderate variability in wind speeds across the dataset.



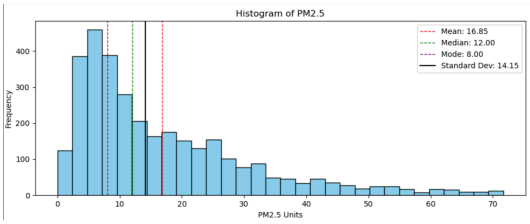
b. Humidity: The distribution of humidity levels, with the x-axis representing humidity units and the y-axis showing their frequency. The mean humidity is 0.73, the median is 0.78, and the mode is 0.93, suggesting a slightly right-skewed distribution. The standard deviation of 0.19 indicates relatively low variability in humidity values.



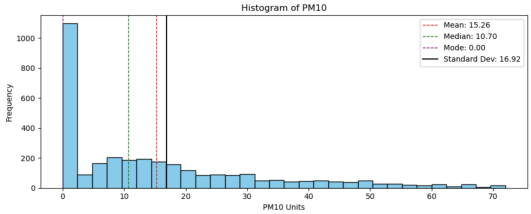
c. Weather: The distribution of weather-related units, with the x-axis indicating the values and the y-axis their frequency. The mean is 11.18, the median is 9.00, and the mode is 0.00, indicating a positively skewed distribution. A standard deviation of 10.08 suggests considerable variability in the weather units.



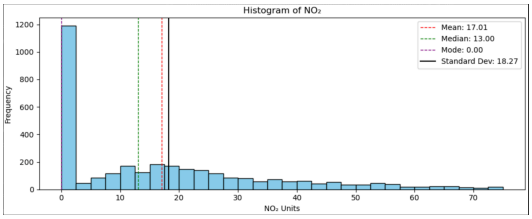
d. PM2.5: The distribution of PM2.5 levels in the dataset. The mean PM2.5 concentration is 16.85, the median is 12.00, and the mode is 8.00, indicating a right-skewed distribution. The standard deviation of 14.15 reflects significant variability in PM2.5 levels, with most values concentrated below 20.



e. PM10: This histogram depicts the distribution of PM10 levels in the dataset. The mean is 15.26, the median is 10.70, and the mode is 0.00, indicating a right-skewed distribution with many zero values. The standard deviation of 16.92 suggests a high variation in PM10 concentrations.

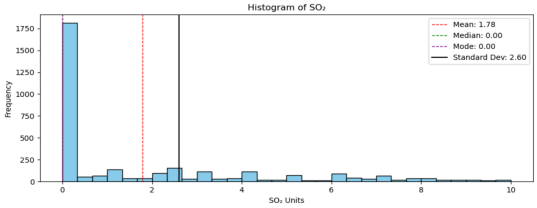


f. NO2: The distribution of NO_2 levels in the dataset. The mean concentration is 17.01, the median is 13.00, and the mode is 0.00, indicating a right-skewed distribution with a significant number of zero values. The standard deviation of 18.27 suggests considerable variability in NO_2 levels.

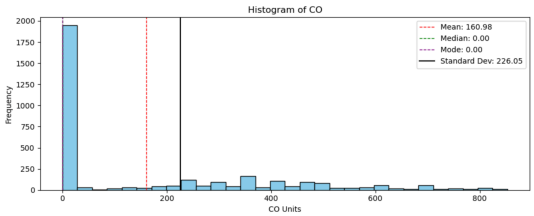


g. So2: This histogram represents the distribution of SO_2 levels. The mean value is 1.78, while both the median and mode are 0.00, showing a highly

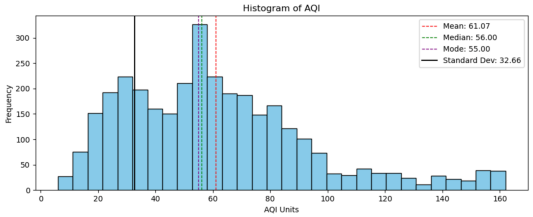
right-skewed distribution dominated by zero values. The standard deviation of 2.60 indicates moderate variability, with a significant portion of data clustered near zero.



h. CO: This histogram shows the distribution of CO levels in the dataset. The mean is 160.98, while both the median and mode are 0.00, indicating a highly skewed distribution with a significant number of zero values. The standard deviation of 226.05 suggests a very high variability in CO concentrations, with most data clustered near zero and a few extreme values.



i. AQI: This histogram illustrates the distribution of AQI (Air Quality Index) values. The mean AQI is 61.07, the median is 56.00, and the mode is 55.00, suggesting a relatively symmetrical distribution around the median. The standard deviation of 32.66 indicates moderate variability in AQI levels across the dataset.



B. *Bivariate and Multivariate Analysis:*

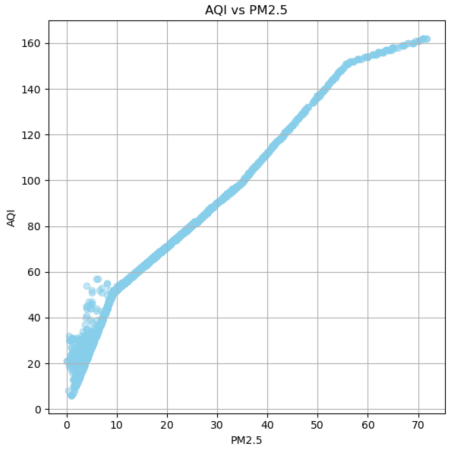
a. AQI vs PM2.5:

i. There is a strong positive correlation, indicating that as PM2.5 levels increase, AQI also rises.

ii. The points form a clear upward trend, suggesting a near-linear relationship between the two variables.

iii. At lower PM2.5 levels, AQI values are concentrated, while at higher PM2.5 levels, the spread of AQI narrows.

iv. This pattern highlights PM2.5 as a significant contributor to the overall AQI.



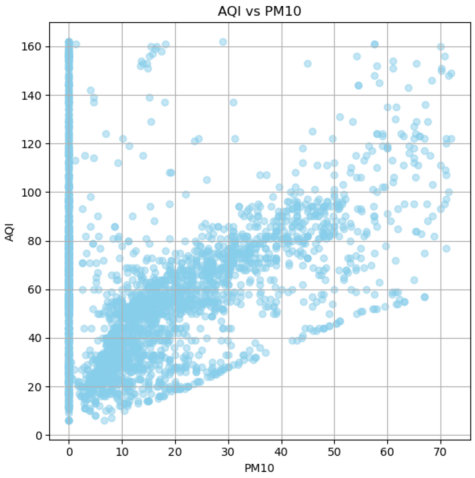
b. AQI vs PM10:

i. There is a visible positive correlation, though less linear compared to AQI vs. PM2.5.

ii. A significant number of points are clustered at lower PM10 values, with AQI ranging widely.

iii. Higher PM10 levels show a general trend of increased AQI, but with greater variability.

iv. The spread indicates that while PM10 contributes to AQI, its impact may interact with other factors.



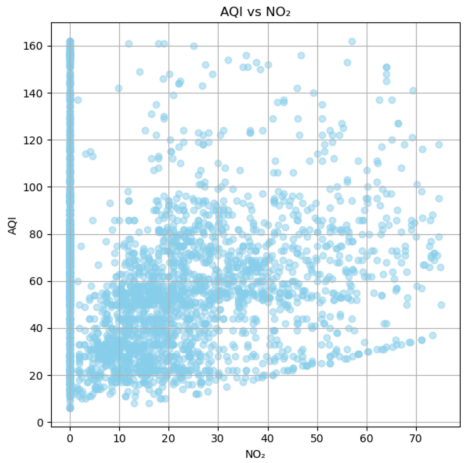
c. AQI vs NO2:

i. There is a weak positive correlation, with AQI generally increasing as NO_2 levels rise.

ii. A large concentration of points is observed at low NO_2 values, with AQI spread across a wide range.

iii. The variability in AQI increases with higher NO_2 levels, indicating other contributing factors to AQI.

iv. The trend is less pronounced compared to PM2.5 or PM10 correlations with AQI.



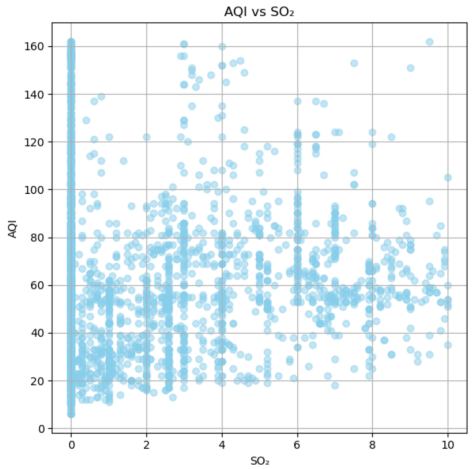
d. *AQI vs SO2:*

i. here is no clear or strong correlation, as AQI values are widely scattered across all levels of SO_2.

ii. A significant cluster of points is observed at lower SO_2 levels, with AQI ranging broadly.

iii. Higher SO_2 values show more variability, but no consistent trend with AQI.

iv. This suggests that SO_2 may have a minimal or less direct impact on AQI compared to other pollutants.



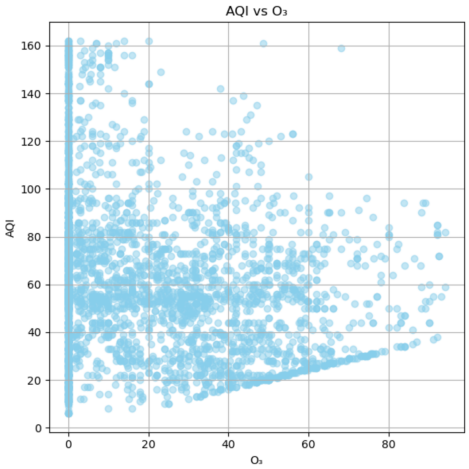
e. *AQI vs O3:*

i. There is a weak positive correlation, with AQI generally increasing as O_3 levels rise, but the relationship is not strong.

ii. A dense cluster of points is observed at lower O_3 levels, with AQI values spread widely.

iii. As O_3 levels increase, the spread in AQI narrows slightly, indicating a more consistent impact at higher concentrations.

iv. This suggests that while O_3 contributes to AQI, its influence may vary depending on other environmental factors.



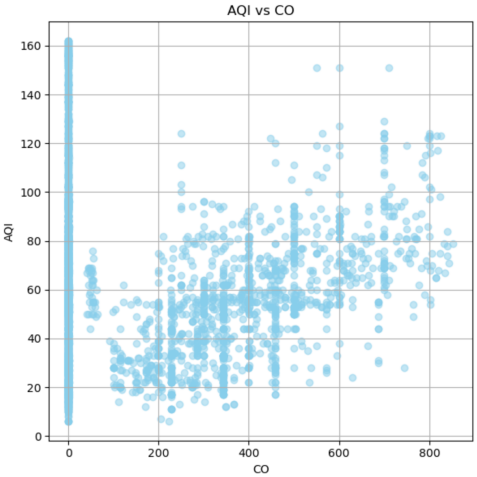
f. *AQI vs CO:*

i. There is a weak positive correlation, with AQI increasing slightly as CO levels rise.

ii. A large cluster of points is concentrated at lower CO values, showing a wide spread in AQI.

iii. At higher CO levels, the AQI values remain scattered, indicating variability and potential influence from other factors.

iv. This suggests that CO contributes to AQI but does not solely determine its values.

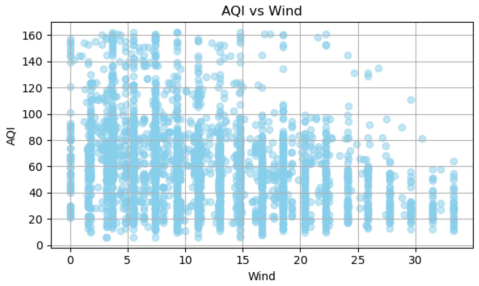


g. *AQI vs Wind:*

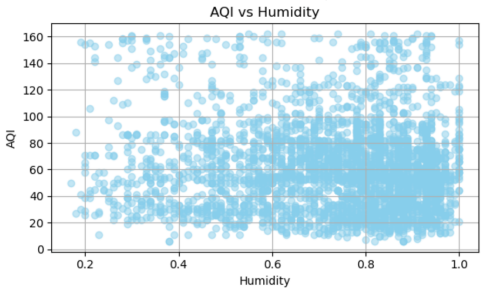
i. There is no clear or strong correlation observed between wind speed and AQI values.

ii. A significant number of AQI values are spread across all ranges of wind speeds, indicating weak dependence.

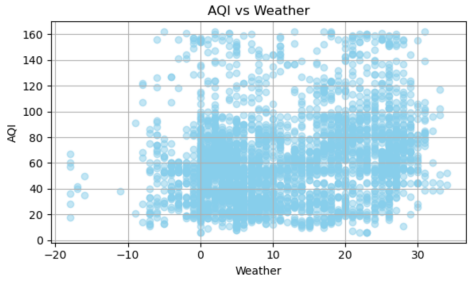
iii. The variability in AQI at different wind speeds suggests other factors may have a stronger influence on AQI levels.
iv. Wind appears to have a minimal direct impact on AQI based on this visualization.



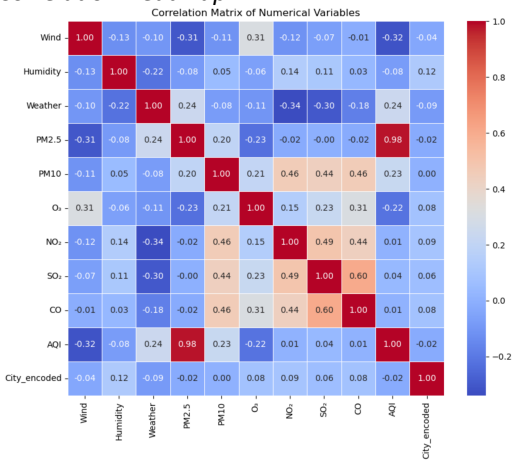
h. *AQI vs Humidity:*
   i. There is no strong or clear correlation, as AQI values are spread across all ranges of humidity.
   ii. A dense cluster of AQI values is observed between 40 and 80 across various humidity levels.
   iii. High variability in AQI at all humidity levels suggests that humidity alone does not significantly influence AQI.
   iv. This indicates that humidity may have an indirect or minimal impact on AQI.



i. *AQI vs Waether:*
   i. There is no clear correlation, as AQI values are dispersed widely across all weather conditions.
   ii. The data points are densely clustered around certain weather values, with no noticeable trend.
   iii. This indicates that the "Weather" variable has minimal or no direct impact on AQI in this dataset.
   iv. Variability in AQI at similar weather values suggests other influencing factors are more significant.



j. *Correlation HeatMap:*



k. *Feature Analysis:*
   i. Top Features (High Correlation with AQI):
      1. **PM2.5 (Correlation: 0.98)**: This feature has the highest correlation with AQI, indicating it is the most significant factor influencing air quality.
      2. **PM10 (Correlation: 0.23)**: Moderately correlated with AQI, suggesting it is also a significant contributor to air quality.
      3. **O_3 (Correlation: -0.22)**: Though negatively correlated, it has a noticeable relationship with AQI.

   ii. Low Features (Low or Insignificant Correlation with AQI):
      1. **Wind (Correlation: -0.32)**: Weak negative correlation, implying minimal influence on AQI.
      2. **Humidity (Correlation: -0.08)**: Negligible impact on AQI, suggesting it is not a strong predictor.
      3. **Weather (Correlation: 0.24)**: Shows a small positive correlation but is relatively

insignificant compared to other factors.

4. **SO_2 (Correlation: 0.01)**: Virtually no correlation with AQI.
5. **CO (Correlation: 0.01)**: Similarly, no significant impact on AQI.

III. Regression Models for AQI Prediction

A. *Feature Selection:*

Feature selection is a crucial step in machine learning that aims to identify the most relevant features for model training. In this project, we employed the Forward Selection method to determine the optimal set of features. The process began by selecting PM2.5, as it is highly correlated with the target variable, AQI. Subsequently, other features such as PM10, CO, NO_2, SO_2, O_3, humidity, wind, and temperature were added one by one, and the R^2 score and Mean Squared Error (MSE) were recorded for each iteration. It was observed that while the R^2 score remained constant, the MSE was minimized when all the features were included.

Given that the calculation of AQI inherently relies on the values of each pollutant and meteorological data, it was concluded that using all features would lead to the most accurate prediction. By including all available features, we aim to capture the complex interactions among pollutants and environmental factors, thereby improving the predictive performance of the model.

B. *Model Selection:*

Linear Regression was chosen as a baseline model because it is simple to implement and interpret. However, due to the presence of non-linear relationships in the dataset, Random Forest was also considered, as it is well-suited for handling non-linear interactions.Gradient Boosting was then used as it performs well on small datasets and is often superior to Random Forest in terms of accuracy and efficiency. For deep learning, a Simple Fully Connected Neural Network was implemented to explore the dataset's potential for complex pattern recognition.

C. *Model Training and Evaluation:*

a. *Linear Regression:* Linear Regression is a simple yet effective approach to predicting AQI. It establishes a linear relationship between the input features and the target variable. The steps involved in the implementation are:

i. Data Splitting: The data was split into training and testing sets (80% training, 20% testing) using train_test_split.

ii. Model Training: A Linear Regression model was trained using the training set.

iii. Prediction: The trained model predicted AQI values for the test set.

iv. Evaluation: The model's performance was evaluated using R2, MSE, MAE, and RMSE.

v. Results:

```
R2 score with all features:  0.9682995546668377
MSE with all features:  32.8869467053774
MAE: 4.57724770195141
RMSE: 5.7347141781763975
```

b. *Random Forrest Regressor:* Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions to achieve better generalization. The steps involved in the implementation are:

i. Data Splitting: Similar to Linear Regression, the data was split into training (80%) and testing (20%) sets.

ii. Model Training: A Random Forest Regressor with 100 estimators was trained on the training data.

iii. Prediction: The trained model was used to predict AQI values for the test set.

iv. Evaluation: R2, MSE, MAE, and RMSE were computed for performance evaluation.

v. Cross-Validation: Cross-validation is often used to detect overfitting by providing an estimate of model performance on different subsets of the data. If the model performs well across all folds, it suggests that the model is generalizing well and is less likely to be overfitting.

In our case, the cross-validation results show an R^2 score close to 1 in all folds, with an average R^2 score of 0.9994. This indicates that the model is performing extremely well on the data and is likely not overfitting, as the R^2 score is very high for all folds.

vi. Results:

```
R2 score:  0.9996563818569673
MSE:  0.3564792682926829
MAE: 0.22600609756097564
RMSE: 0.5970588482659669
```

```
Cross-Validation R2 Scores: [0.99949574 0.99969227 0.99961037 0.99955373 0.998462  ]
Average R2 Score: 0.9993628204718072
```

c. *Gradient Boosting Regressor:* Gradient Boosting is a powerful ensemble technique that builds multiple weak learners sequentially, each correcting the errors of its predecessor. The steps in its implementation are:

i. Data Splitting: The data was split into 80% training and 20% testing sets.

ii. Model Training: A Gradient Boosting Regressor with 100 estimators, a learning rate of 0.1, and a maximum tree depth of 3 was trained.

iii. Prediction: Predictions for the test set were generated using the trained model.

iv. Evaluation: The model was evaluated using R^2, MSE, MAE, and RMSE to measure the accuracy and magnitude of prediction errors.

v. Results:

```
Gradient Boosting R2 score:  0.9993276333793947
Gradient Boosting MSE:  0.6975323212633866
MAE: 0.39147880755242426
RMSE: 0.8351840044345836
```

d. *Dense Neural Network (DNN):* The Dense Neural Network (DNN) is a deep learning model capable of capturing complex patterns in the data. It uses multiple layers of interconnected neurons. The steps for implementation are as follows:

i. Data Splitting: The data was split into training (80%) and testing (20%) sets.

ii. Feature Scaling: The features were scaled using StandardScaler to ensure all features have a similar range. This prevents dominance of features with larger magnitudes.

iii. Model Architecture:

1. Input Layer: The input layer has the same number of neurons as the features in the dataset.
2. Hidden Layers: Two hidden layers were implemented with 128 neurons and 64 neurons, each using a ReLU activation function.
3. Output Layer: A single neuron with a linear activation function was used to predict AQI.

iv. Model Compilation: The model was compiled using the Adam optimizer and mean squared error (MSE) loss function.

v. Training: The model was trained for 100 epochs with a batch size of 10.

vi. Prediction and Evaluation: After training, the model predicted AQI values for the test set, and the results were evaluated using R2, MSE, MAE, and RMSE.

vii. Results:

```
Neural Network R2 score:  0.9974197001905498
Neural Network MSE:  2.6768766629442515
MAE: 1.0202211548642415
RMSE: 1.6361163353943544
```

D. *Model Comparison:*

The performance of the four models—Linear Regression, Random Forest, Gradient Boosting, and Neural Network—was compared based on key evaluation metrics, including R^2, MSE, MAE, and RMSE. The results are presented in the following table:

Model	R2 Score	MSE	MAE	RMSE
Linear Regression	0.9682	32.8869	4.5772	5.7347
Random Forrest	0.9996	0.3564	0.2260	0.5970
Gradient Bossting	0.9993	0.6975	0.3914	0.8351
Dense Neural Network	0.9974	2.6768	1.0202	1.6361

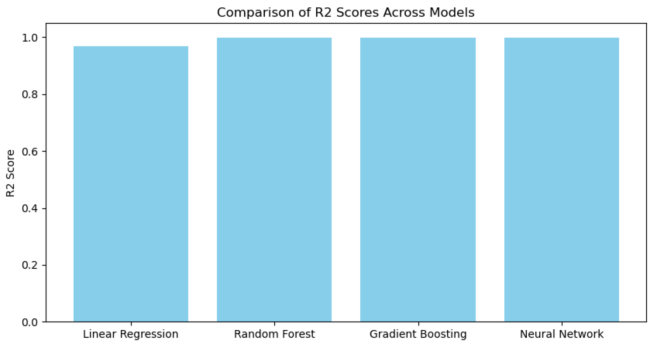
TABLE I

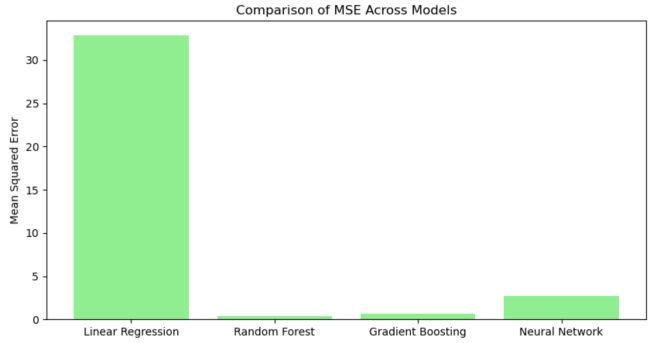
TABLE SHOWING r2 score, MSE, MAE, RMSE OF DIFFERENT MODELS

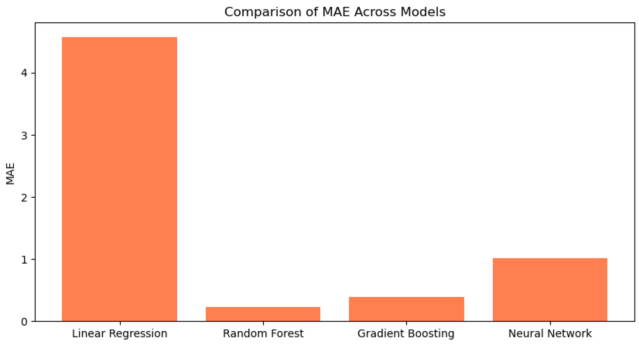
Based on the analysis, the **Random Forest Regressor** emerged as the best model for predicting AQI from the dataset.

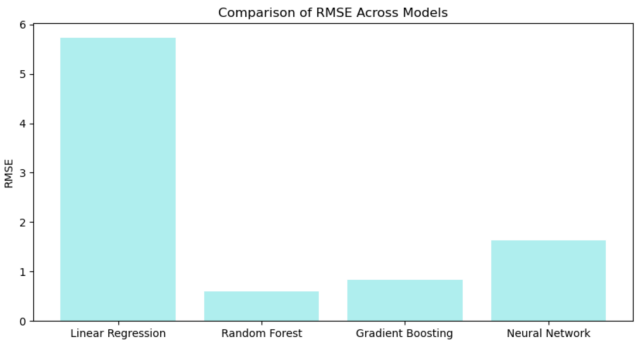
Reason for Choosing Random Forest

- **High Accuracy**: Random Forest achieved the highest **R^2 score**, close to 0.999, indicating its ability to explain nearly all the variance in AQI.
- **Low Error**: It exhibited the lowest **MSE, MAE, and RMSE**, which demonstrates its superior prediction accuracy compared to other models.









IV. Conclusion and Future Direction

This project successfully demonstrated the use of meteorological and pollutant data for AQI prediction. Key findings revealed that features like PM2.5, PM10, O_3, NO_2, SO_2, CO, humidity, wind, and temperature play a critical role in determining AQI. While all models exhibited strong predictive performance, the Random Forest Regressor emerged as the most effective model, achieving the highest R^2 score and the lowest MSE, MAE, and RMSE. This indicates its superior capacity to explain variance in AQI and minimize prediction errors.

Other models, including Gradient Boosting, Linear Regression, and Neural Networks, also demonstrated reasonable predictive power. However, Random Forest outperformed them in all key performance metrics. The study met its primary objectives of building an accurate AQI prediction system and identifying key pollutants and weather factors influencing AQI.

While the project achieved its goals, several areas for improvement remain. Future work can focus on incorporating **time-series models** such as **LSTM (Long Short-Term Memory)** and **ARIMA (Auto-Regressive Integrated Moving Average)** to better capture temporal dependencies and improve prediction accuracy. Expanding the dataset by collecting AQI data over a longer period and across more cities can enhance the model's generalization capability. Additionally, integrating **time-related features** like **hour, day of the week, and seasonality** could help capture periodic patterns in AQI changes. Lastly, with larger datasets, **Neural Networks** can be fine-tuned to achieve higher accuracy and handle more complex interactions among features. These improvements would enhance the robustness, accuracy, and interpretability of AQI prediction models, making them more effective in supporting real-time air quality monitoring and decision-making.

V. Individual Contribution

Muhammad Hassan Muzaffar and Syed Ahmer Zaidi took the lead in web scraping by utilizing tools like Selenium and BeautifulSoup to extract data from IQAir. Muhammad Hassan Muzaffar handled the essential tasks of data preprocessing, including cleaning, encoding, and scaling the dataset to ensure it was suitable for model training. Mohammad Waleed Ikram was responsible for data visualization and feature selection, playing a key role in identifying the most critical features for AQI prediction. He also went a step further by creating an interactive dashboard using Tableau, which provides a user-friendly interface for visualizing AQI trends and model predictions. The tasks of model training, implementation, and evaluation were collectively handled by all three team members: Muhammad Hassan Muzaffar, Mohammad Waleed Ikram, and Syed Ahmer Zaidi, ensuring a collaborative approach to optimizing model performance. Additionally, the project's final report was co-authored by Syed Ahmer Zaidi and Muhammad Hassan Muzaffar, ensuring clear and comprehensive documentation of the project's objectives, methods, and outcomes.

References

[1] IQAir. (2024). Air Quality of major cities. Available at: https://www.iqair.com/world-air-quality-ranking

[2] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[3] Friedman, J. H. (2002). Stochastic gradient boosting. Computational statistics & data analysis, 38(4), 367-378.

[4] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

[5] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research.

[6] Aggarwal, C. C. (2016). Outlier analysis. Springer.