**Data Science Project**

**Heyda I. Alvarado Alvelo**

**Professor Linesh Dave**

# Global AQI Study: XGBoost vs. Deep Learning Approaches

**12 Dec 2023**

## Executive Summary

This comprehensive project centered on a thorough examination of the Global Air Pollution Dataset, with the primary objective of predicting the overall Air Quality Index (AQI) category. The initial phase involved meticulous data cleansing procedures, encompassing the rectification of missing values and outliers. Subsequently, normalization and standardization methodologies were employed to ensure data homogeneity and comparability. Rigorous statistical analyses, inclusive of descriptive statistics and distributional assessments, were undertaken to gain profound insights into the intrinsic characteristics of the dataset. The utilization of tools such as Excel, Google Colab, and Python libraries facilitated a seamless transition from data cleansing to normalization, standardization, and in-depth statistical exploration.

In the pursuit of predictive modeling, both XGBoost and MLP algorithms were deployed, complemented by explainable AI techniques, notably LIME, to enhance model interpretability. The overarching project objectives were successfully achieved, resulting in the development of two predictive models essential for air quality management. The salient outcome revealed XGBoost as the superior model, exhibiting a marginal yet statistically significant advantage of 0.1 accuracy over MLP. This nuanced comparison accentuates the efficacy of the chosen methodology, contributing substantive insights to the ongoing discourse on global air pollution challenges. The meticulous handling of data, coupled with advanced analytics and model comparisons, positions this project as a significant contribution toward informed decision-making in the realms of environmental conservation and public health management within an academic framework.

**Table of Contents**

# Project Scope

## Problem Description

The problem at hand involves the critical examination and comparison of traditional machine learning approaches with deep learning algorithms in the context of Air Pollution Categorization (Classification). Given the pressing global concerns surrounding air quality and the far-reaching implications of air pollution on human health and the environment, it becomes imperative to explore advanced data analytics techniques that can effectively categorize and predict air pollution levels (Dutta & Pal, 2022). Traditional machine learning models offer a robust framework for understanding the intricate relationships between various atmospheric pollutants and their corresponding Air Quality Index (AQI) values. On the other hand, deep learning algorithms, with their capacity to process complex and unstructured data, provide a promising avenue for exploring more nuanced patterns and correlations within the air pollution dataset. By conducting a comparative analysis of these methodologies, the study aims to shed light on the efficacy of different analytical techniques in predicting and categorizing air pollution, ultimately contributing to the development of more accurate and efficient predictive models for air quality management and environmental sustainability. Furthermore, the project will leverage explainable AI techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) to enhance the interpretability of the models and provide insights into the factors influencing the predictions.

The importance of this problem lies in its potential to significantly impact public health policies, environmental regulations, and urban planning initiatives worldwide. With the rise of urbanization and industrialization, the quality of air has become a critical public health concern, directly impacting the well-being of millions of people globally (Dutta & Pal, 2022). By comprehensively evaluating the effectiveness of traditional machine learning methods and deep learning algorithms in categorizing air pollution, the research endeavor seeks to contribute to the development of advanced predictive models that can accurately assess and monitor air quality levels. Furthermore, the findings of this study have the potential to inform policymakers, environmental agencies, and public health authorities in formulating evidence-based strategies and interventions to mitigate the adverse effects of air pollution on human health and the environment. The data analytics problem that I am analyzing is pivotal in enabling a more profound comprehension of air pollution dynamics and fostering the creation of robust predictive models, thereby facilitating the implementation of proactive measures to address the pervasive global issue of air pollution. With the integration of explainable AI techniques, the project aims to provide transparent insights into the decision-making processes of the models, enhancing the overall trustworthiness and interpretability of the analytical outcomes.

## Project Importance

This project was chosen because of my background in Chemistry. Before transitioning to Data Science, I studied Chemical Engineering, which ignited my passion for the field. During my time at the University of Puerto Rico at Mayaguez, I took various chemistry courses, including Organic Chemistry and Quantum Mechanics. I was particularly drawn to Environmental Chemistry, which emphasized the importance of having a strong foundation in the subject when working on Data Science projects. This background has equipped me with a deep understanding of the chemical processes underlying air quality, motivating me to explore how this knowledge can be integrated with modern data analysis techniques to address pressing environmental challenges. According to a study on urban air pollution, the prediction of air quality stands as a crucial obstacle in the early warning and control of urban air pollution, with the accelerated pace of urbanization and industrialization intensifying its significance (Dutta & Pal, 2022). This emphasizes the necessity of projects like mine, which seek to bridge the gap between Chemistry and Data Science to better comprehend and tackle the complexities of urban air pollution.

The significance of this project lies in its potential to offer valuable insights into the complex issue of urban air pollution. By combining my Chemistry background with advanced Data Science techniques, I aim to contribute to the advancement of air quality monitoring and management. This interdisciplinary approach not only benefits the scientific community but also holds the potential to inform and guide urban planners, environmental agencies, and the public. Moreover, the comparative analysis between traditional machine learning algorithms, and the more sophisticated deep learning algorithms for the classification of Air Quality Index (AQI) is crucial. This comparative study not only sheds light on the efficacy of different analytical approaches but also serves to highlight the potential of leveraging advanced deep learning techniques in addressing complex environmental challenges. By evaluating the strengths and limitations of both traditional and modern machine learning methodologies, this project aims to pave the way for more robust and accurate predictive models, thereby facilitating more informed decision-making and policy implementation for sustainable urban development and improved public health.

## Background

The problem of air pollution and its consequences has been a topic of research for many years. The Clean Air Act is recognized for its role in improving air quality, both in the United States and globally, by introducing a risk communication requirement. This mandate includes the consistent monitoring of air quality across the United States and the daily reporting of air quality through a uniform Air Quality Index (AQI) (Perlmutt & Cromar, 2019). According to the Clean Air Act's guidelines, the AQI is determined based on the pollutant with the highest concentration relative to its regulatory standard for the day. This index ranges from 0 to 500, with 100 representing the national air quality concentration standard for a pollutant. The implementation of the Clean Air Act has significantly contributed to the understanding and management of air quality concerns, emphasizing the need for advanced data analytics techniques in the field (Perlmutt & Cromar, 2019).

Additionally, recent research has explored the use of hybrid deep learning models, such as the LSTM-GRU combination, for accurately predicting the AQI values of specific pollutants, demonstrating the potential of advanced machine learning techniques in the context of air quality management (Sarkar et al., 2022)

In the contemporary context, air pollution remains a significant global concern, with its detrimental effects on public health and the environment drawing considerable attention. The study of air quality indices has gained prominence, enabling a comprehensive understanding of the intricate relationships between atmospheric pollutants and their impact on human health. In response to this urgency, various machine learning and deep learning models have been developed and evaluated, aiming to accurately predict the AQI values of different pollutants (Sarkar et al., 2022). Researchers have leveraged methods such as Support Vector Machines, Random Forest, Convolutional Neural Networks, and hybrid models, demonstrating the effectiveness of these techniques in predicting air quality levels (Sarkar et al., 2022). Moreover, the integration of explainable AI techniques, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), has facilitated the interpretation of complex models, enhancing the transparency and interpretability of the prediction outcomes. As a result, the use of advanced data analytics methods has become critical in addressing the challenges posed by air pollution and in devising effective strategies for environmental conservation and public health management.

## Data Set Description

The Global Air Pollution Dataset (Muzdadid, 2022) is a comprehensive compilation of atmospheric data, featuring 12 distinct columns (variables) and 23,463 individual instances, with 427 instances containing blank entries for the 'Country' column and 1 for the ‘City’ column. This extensive dataset offers a thorough exploration of air quality parameters across various cities worldwide, making it a crucial resource for researchers and policymakers aiming to understand the dynamics of global air pollution. The dataset's key dimensions, including essential variables such as 'City', provide geographical insights into the specific locations where the air quality data was collected, enabling a comprehensive evaluation of pollution levels in diverse urban environments. Additionally, the dataset incorporates fundamental parameters such as the 'AQI Value' and 'AQI Category', acting as standardized measures for assessing pollution levels and gauging the potential health risks associated with varying degrees of air contamination, with the AQI values ranging from 0 to 500. These values correspond to the five distinct categories: 'Good', 'Moderate', 'Unhealthy for Sensitive Groups', 'Unhealthy', and 'Very Unhealthy', enabling a comprehensive assessment of the severity of air pollution levels and their potential impacts on public health (Muzdadid, 2022). Notably, the dataset contains specific variables for prominent atmospheric pollutants such as 'CO', 'Ozone', 'NO2', and 'PM2.5', each with individual AQI values and corresponding categories.

In the Global Air Pollution Dataset (Muzdadid, 2022), 'CO AQI Value' variable represents the AQI value of Carbon Monoxide in the city, primarily emitted by vehicular exhaust and industrial processes. 'Ozone AQI Value' refers to the AQI value of Ozone, a harmful gas resulting from chemical reactions between nitrogen oxides and volatile organic compounds, capable of causing respiratory issues and damaging lung tissues. 'NO2 AQI Value' signifies the AQI value of Nitrogen Dioxide, a hazardous gas primarily stemming from vehicle emissions and industrial activities, known to exacerbate respiratory illnesses and pose significant health risks, especially for vulnerable populations. 'PM2.5 AQI Value' indicates the AQI value of Particulate Matter with a diameter of 2.5 micrometers or less, representing fine particles in the air that can penetrate deep into the respiratory system and cause severe respiratory and cardiovascular complications. Each pollutant's corresponding AQI categories, including 'Good', 'Moderate', 'Unhealthy for Sensitive Groups', 'Unhealthy', and 'Very Unhealthy', provide vital insights into the severity of air pollution levels and their potential implications for public health. By encapsulating a wide range of atmospheric data, the dataset serves as a valuable repository for studying air quality variations and trends across diverse geographical locations, fostering a deeper understanding of the global prevalence of air pollution and its effects on public health and the environment.

## Data Analytics Tools

For this project, my approach entails leveraging a comprehensive toolkit encompassing various software and programming resources to conduct a meticulous analysis of the Global Air Pollution Dataset (Muzdadid, 2022). The initial stages involve using Excel for meticulous data cleaning, ensuring that the dataset is appropriately formatted and devoid of inconsistencies, followed by seamless transitioning to Google Colab for an in-depth exploratory data analysis (EDA) process. Within these interactive notebook environments, I plan to utilize the Python programming language along with pivotal libraries such as pandas, matplotlib, seaborn, sklearn, numpy, lime, and shap, to facilitate sophisticated data operations, construct informative visualizations, and develop intricate machine learning models. The versatility of pandas will be harnessed for streamlined data manipulation, while the combined functionalities of matplotlib and seaborn will enable the creation of insightful visual representations, aiding in the comprehensive comprehension of underlying data patterns and trends. Furthermore, the integration of sklearn will play a pivotal role in the seamless implementation of robust machine learning algorithms, thereby facilitating the creation of accurate models aimed at predicting air quality based on the dataset's atmospheric pollutant data. Integrating lime and shap will be instrumental in delving into the intricacies of the machine learning models, providing valuable insights into the factors influencing the models' predictions and enhancing the overall interpretability of the project's outcomes.

Additionally, the potential integration of PySpark within the Jupyter environment holds promise for enhancing the project's data processing capabilities. PySpark, a powerful open-source framework, enables seamless integration with Python and facilitates distributed data processing for handling large-scale datasets. Its capacity to leverage SQL functionalities within the Jupyter environment will provide an efficient and scalable approach to data manipulation and analysis, ensuring the seamless execution of complex queries and computations within the Python ecosystem. Moreover, the utilization of Tableau as an additional tool holds significant potential for creating dynamic and interactive data visualizations. Tableau, a leading data visualization software, empowers users to create intuitive and engaging visual representations of complex datasets. By leveraging Tableau's intuitive interface and robust visualization capabilities, the project can effectively communicate complex insights derived from the Global Air Pollution Dataset Global Air Pollution Dataset (Muzdadid, 2022), facilitating a deeper understanding of the spatial distribution of air quality levels and their corresponding AQI scores across various geographical regions.

### Project Milestones

These milestones have set deadlines to ensure the timely completion of each phase, leading to the successful execution and comprehensive documentation of the project's objectives and outcomes.

- **Assignment 1**: *Project Scope:* Due by October 31, 2023
- **Assignment 2:** *Data Profiling and Preparation:* Due by November 7, 2023
- **Presentation 1:** *Project Summary Presentation:* Due by November 7, 2023
- **Assignment 3:** *Data Visualization:* Due by November 14, 2023
- **Assignment 4:** *Data Modeling:* Due by November 28, 2023
- **Presentation 2:** *Final Project Presentation:* Due by December 5, 2023
- **Assignment 5:** *Final Results:* Due by December 12, 2023

**Assignment 1** - *Project Scope*: This initial milestone involves outlining the comprehensive scope and objectives of the project, setting the stage for the research plan, methodologies, and anticipated outcomes.

**Assignment 2:** *Data Profiling and Preparation*: This stage focuses on conducting a thorough analysis of the Global Air Pollution Dataset, involving data profiling and preparation to ensure the dataset is properly formatted and ready for subsequent analysis.

**Presentation 1:** *Project Summary Presentation*: This presentation provides a concise overview of the project's objectives, methodology, and preliminary findings, serving to effectively communicate the project's progress and key insights to relevant stakeholders.

**Assignment 3:** *Data Visualization*: This milestone entails the development of comprehensive data visualizations, facilitating a clear and intuitive representation of the complex relationships between various atmospheric pollutants and their corresponding AQI scores.

**Assignment 4:** Data Modeling: This stage involves the implementation of sophisticated machine learning models to predict air quality based on the dataset's atmospheric pollutant data, aiming to create accurate and effective predictive models.

**Presentation 2:** *Final Project Presentation*: This presentation serves as a comprehensive overview of the project's findings, methodologies, and insights, allowing for the effective dissemination of the project's outcomes to a wider audience.

**Assignment 5**: *Final Results*: This final assignment includes a detailed summary of the project's final results and key findings, solidifying the project's contributions to the broader field of air quality analysis and environmental research.

**Completion History**

| | |
|---|---|
| **Weeks 1 & 2** | • Attended Zoom meetings.<br>• Discussed datasets with Professor Linesh<br>• Selected dataset for project<br>• Researched background information and sources regarding air pollution<br>• Familiarized with dataset.<br>• Brainstormed ideas for project<br>• Created workflow in Trello for project.<br>• Consulted with a PhD student in the industry.<br>• Worked on project scope |
| **Weeks 3 & 4** | • Week 3 was a challenge completing the assignments on time.<br>• Cleaned data.<br>• Worked on identifying any outliers or anomalies.<br>• Changed variable labels for more readability.<br>• Worked on EDA using Pandas.<br>• Created different visualizations. |
| **Weeks 5 & 6** | • Conducted literature review on Random Forest and CNN models for the project.<br>• Participated in Zoom class sessions and actively asked questions.<br>• Revised Exploratory Data Analysis (EDA) after gaining deeper insights into the models. |

| | |
|---|---|
| | • Recognized and rectified the inadvertent deletion of columns during EDA.<br>• Opted for XGBoost over Random Forest due to imbalanced data; research indicated XGBoost often outperforms Random Forest in such cases. |
| **Weeks 7 & 8** | • Worked on both models.<br>• Debugged code.<br>• Worded on final presentation.<br>• Watched tutorials for both models and Explainable AI<br>• Worked on final assignment.<br>• Participated in course Zoom session. |

**Lessons Learned**

| | |
|---|---|
| **Assignment 1** | • Leveraged domain knowledge in the dataset selection process.<br>• Discovered the value of feature engineering for dataset enhancement. |
| **Assignment 2** | • Values within the AQI range can/will skew the data.<br>• Planning in case of unexpected events is crucial |
| **Assignment 3** | • Precision in visualization creation is crucial.<br>• Familiarizing myself with the data to choose the right graph for accurate representation is very important. |
| **Assignment 4** | • Extensive research is essential when dealing with a new model.<br>• Diverse methods and tools exist for achieving similar outcomes; thorough exploration is necessary.<br>• XGBoost is fascinating, but its numerous parameters require dedicated research for effective utilization.<br>• Always double-check code even when you think the task is complete. |

| | |
|---|---|
| **Assignment 5** | - Persevere until the end.<br>- Debugging can be challenging.<br>- If something doesn’t work, move on, and try another method. |

## Data Profiling and Preparation

### Data Summary

The Global Air Pollution Dataset (Muzdadid, 2022), downloaded from Kaggle, is a comprehensive compilation of atmospheric data with 12 distinct columns (variables) and 23,463 individual instances. It is noteworthy that 427 instances contain blank entries for the 'Country' column, and 1 instance has a blank entry for the 'City' column. This dataset provides geolocated information on key pollutants such as Nitrogen Dioxide (NO2), Ozone (O3), Carbon Monoxide (CO), and Particulate Matter (PM2.5), offering crucial insights into their impact on human health and the environment. This data set’s importance is due to its comprehensive coverage of major air pollutants and their effects on both human health and the environment. The inclusion of geolocated information for various cities worldwide allows for a detailed understanding of the global air quality situation.

The dataset was collected from eLichens (*Global Air Quality Map*), which is a platform that provides comprehensive environmental monitoring solutions. The platform's focus is on collecting data related to air quality and enabling businesses, cities, and individuals to make informed decisions about environmental issues. The original data was collected through various environmental monitoring devices and sensors deployed in different cities across the world. These devices measure the concentration levels of pollutants like NO2, O3, CO, and PM2.5 in the air. The data collection process involves continuous monitoring over specific time intervals to capture the variations in air quality levels, allowing for the generation of the Air Quality Index (AQI) values for each pollutant. The dataset specific time range is not provided in the source.

In the context of data preprocessing for machine learning, the dataset is split into training and testing sets. The training set, constituting 80% of the data, will be used to train a model, while the remaining 20% forms the testing set for model evaluation. To enable machine learning algorithms requiring numerical inputs, the categorical variables in the 'AQI Category' columns will be transformed into numerical representations using label encoding.

**Data Definition/Data Profile**

| Field Name | Definition | Data Type | Outliers | Frequency of Nulls | Potential Quality Issues |
|---|---|---|---|---|---|
| **Country** | Name of the country | String | None | 427 instances | None. Will be deleting this column. |
| **City** | Name of the city | String | None | 1 instance | None. Will be deleting this column. |
| **AQI Value** | Overall AQI value of the city | Integer | None* | None | Overfitting |
| **AQI Category** | Overall AQI category of the city | String | None | None | Class Imbalance |
| **CO AQI Value** | AQI value of Carbon Monoxide of the city | Integer | None** | None | Overfitting |
| **CO AQI Category** | AQI category of Carbon Monoxide of the city | String | None | None | Class Imbalance |
| **Ozone AQI Value** | AQI value of Ozone of the city | Integer | None | None | Overfitting |
| **Ozone AQI Category** | AQI category of | String | None | None | Class Imbalance |

| | | | | | |
|---|---|---|---|---|---|
| | Ozone of the city | | | | |
| **NO2 AQI Value** | AQI value of Nitrogen Dioxide of the city | Integer | None | None | Overfitting |
| **NO2 AQI Category** | AQI category of Nitrogen Dioxide of the city | String | None | None | Class Imbalance |
| **PM2.5 AQI Value** | AQI value of Particulate Matter with a diameter of 2.5 micrometers or less of the city | Integer | None | None | Overfitting |
| **PM2.5 AQI Category** | AQI category of Particulate Matter with a diameter of 2.5 micrometers or less of the city | String | None | None | Class Imbalance |

* The Air Quality Index (AQI) values, both overall and for individual pollutants, may vary significantly depending on the specific pollutant being measured. While some numerical values might appear as outliers within the dataset, it is important to note that these values fall within the established AQI range for each pollutant. During the initial stages of this project, we are not identifying these values as outliers. They may indicate deviations from normal concentrations and will be further analyzed in subsequent phases of the project.

**In the AQI categories for different pollutants, it is observed that not all categories are uniformly present. This variation in the distribution of categories across pollutants may potentially lead to class imbalance issues during model training.

## Data Preparation

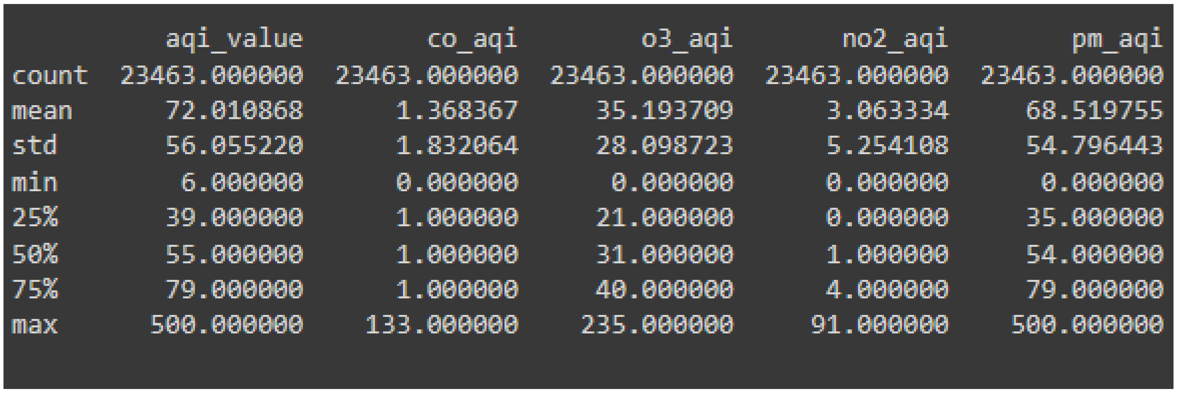
To prepare and cleanse the global air pollution dataset, I will adopt a dual-pronged approach, harnessing the capabilities of Python and Microsoft Excel. Python, with its powerful data manipulation libraries like Pandas, will spearhead the initial data preparation phase. This involves importing and exploring the dataset, conducting exploratory data analysis, and creating visualizations to discern patterns. Additionally, Python facilitates systematic renaming of variable names for enhanced readability and consistency, a crucial step in ensuring the clarity of subsequent analyses. Meanwhile, Microsoft Excel will serve as a user-friendly interface for specific data cleansing tasks, allowing for the removal of extraneous columns such as 'Country' and 'City.' Descriptive statistics provided by Excel's Data Analysis Tools will be employed to detect outliers and anomalies, providing crucial insights into potential data quality issues.

In tandem with these efforts, I will introduce measures to enhance readability and consistency in Python. All variable names will undergo systematic changes to ensure a standardized and easily interpretable format. Moreover, the AQI category "Unhealthy for Sensitive Groups" will be refined by reclassifying to the more concise label "Sensitive." As part of the data preprocessing journey, scikit-learn label encoding will be applied to transform categorical variables into numerical representations, facilitating seamless integration with machine learning algorithms. This comprehensive approach, blending the strengths of Python and Excel, aims to lay a robust foundation for subsequent analyses, ensuring a coherent and insightful exploration of global air pollution patterns.

## Data Visualizations

### Descriptive Statistics

In terms of statistics, the dataset features several practical numerical/nominal variables crucial for our project, specifically focusing on the Air Quality Index (AQI) and individual pollutant AQI values. These fields include: **AQI Value, CO AQI Value, Ozone AQI Value, NO2 AQI Value, PM2.5 AQI Value.**



```
         aqi_value        co_aqi        o3_aqi       no2_aqi        pm_aqi
count  23463.000000  23463.000000  23463.000000  23463.000000  23463.000000
mean      72.010868      1.368367     35.193709      3.063334     68.519755
std       56.055220      1.832064     28.098723      5.254108     54.796443
min        6.000000      0.000000      0.000000      0.000000      0.000000
25%       39.000000      1.000000     21.000000      0.000000     35.000000
50%       55.000000      1.000000     31.000000      1.000000     54.000000
75%       79.000000      1.000000     40.000000      4.000000     79.000000
max      500.000000    133.000000    235.000000     91.000000    500.000000
```

- The **aqi_value** (AQI Value) variable has a wide range with a considerable standard deviation (56.06), indicating variability in air quality measurements. The quartiles (25th, 50th, and 75th percentiles) of 39.00, 55.00, and 79.00, coupled with a mean of 72.01, indicate a moderately right-skewed distribution. This suggests that a significant proportion of air quality index values cluster around the lower to middle range, with a long tail towards higher values, exemplified by the maximum value of 500.

- The **co_aqi** (CO AQI Value) variable appears to have low variability, as indicated by the small standard deviation (1.83) and consistent values across quantiles. Most values fall within the 25th to 75th percentiles. This indicates a limited variability in carbon monoxide air quality index readings, with most values falling within a narrow range, despite a maximum value of 133.

- The **o3_aqi** (Ozone AQI Value) variable exhibits moderate variability, with a wider range and standard deviation (35.19) compared to **co_aqi**. With quartiles (25th, 50th, and 75th percentiles) of 21.00, 31.00, and 40.00, and a mean of 35.19, the distribution of ozone air quality index values appears moderately right skewed. While the majority of values cluster around the lower to middle range, the presence of higher values contributes to a gradual rightward spread, as reflected in the maximum value of 235.

- The **no2_aqi** (NO2 AQI Value) variable shows variability, with a substantial proportion of values at or near zero. The standard deviation (5.25) indicates a spread of data. The quartiles (25th, 50th, and 75th percentiles) of 0.00, 1.00, and 4.00, along with a mean of 3.06, indicate a distribution skewed to the right for nitrogen dioxide air quality index values. Most readings are concentrated at or near zero, emphasizing instances of low pollution levels, while higher values contribute to the rightward skewness, as illustrated by the maximum value of 91.

- The **pm_aqi** (PM2.5 AQI Value) variable has a distribution like **aqi_value**, with a wide range and notable standard deviation (68.52). The quantiles provide a detailed view of the distribution of particulate matter AQI values. The quartiles (25th, 50th, and 75th percentiles) of 35.00, 54.00, and 79.00, along with a mean of 68.52, suggest a distribution slightly skewed to the right. The presence of a sizable number of values towards the upper range, as evidenced by the maximum value of 500, highlights instances of elevated particulate matter air quality index readings.

## Data Visualization Definitions

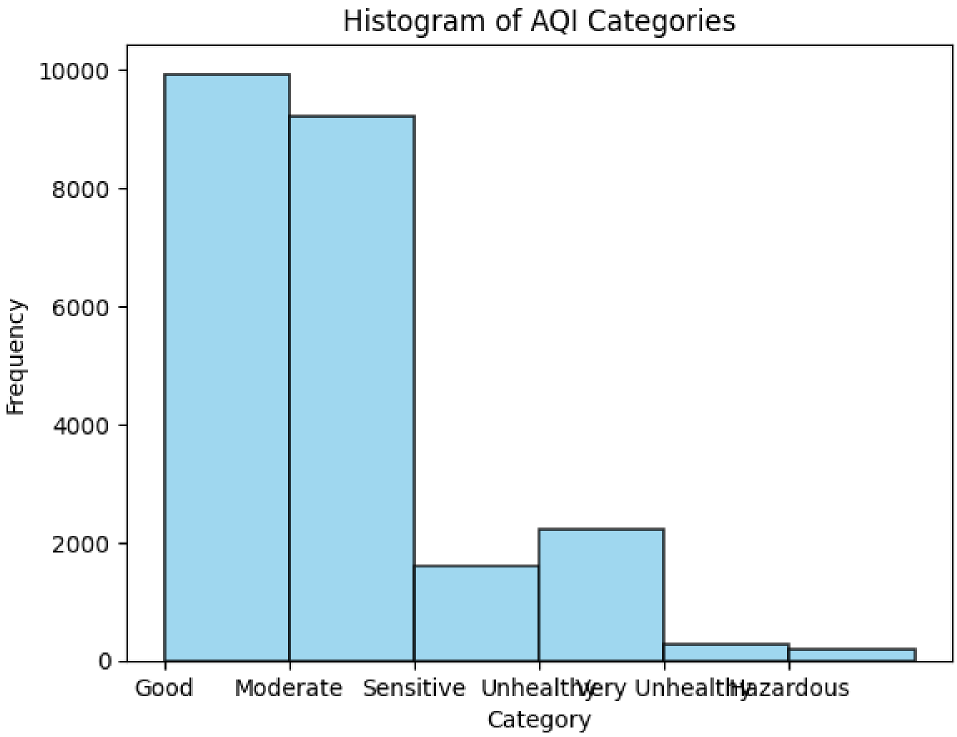
### Visualization Technique 1: Histogram

To further understand the distribution of individual variables, I will utilize histograms. Histograms depict the frequency distribution of a single variable by dividing its range into bins and representing the count or proportion in each bin through bar heights. This technique is especially valuable for discrete distributions, where every value can be visualized by a corresponding bar. Binning is commonly applied to continuous data, making histograms an effective tool for understanding the spread and concentration of values within a variable. As highlighted by Blumenschein et al. (2020), histograms provide a clear representation of data distribution, aiding in the identification of patterns and central tendencies. In our Global Air Pollution dataset, histograms will be used to visualize the distribution of key variables, such as AQI values for different pollutants, offering a comprehensive view of the data's characteristics and aiding subsequent analysis. By examining the histograms of individual pollutants, we can identify potential outliers or skewed distributions, providing essential insights into the nature of pollution levels for each substance.

### Visualization Technique 2: Correlation Matrix

For the exploration of relationships between different variables in the dataset, I will use a correlation matrix. A correlation matrix is a visual representation of the correlation coefficients between variables, often ranging from -1 to 1. A correlation coefficient close to 1 indicates a strong positive correlation, while a coefficient close to -1 signifies a strong negative correlation. A coefficient near 0 suggests a weak or no correlation. This visualization technique is particularly useful for identifying patterns, dependencies, and potential multicollinearity among variables. According to Merchant et al. (2023), correlation matrices play a crucial role in object classification, where correlations between features are examined to understand their interdependence. The correlation matrix will help uncover the relationships between various air quality indicators in our dataset, providing valuable insights into the factors influencing overall air quality. For instance, if the correlation matrix reveals a strong positive correlation between certain pollutants, it could indicate a common source of pollution.

**Data Visualization 1:**

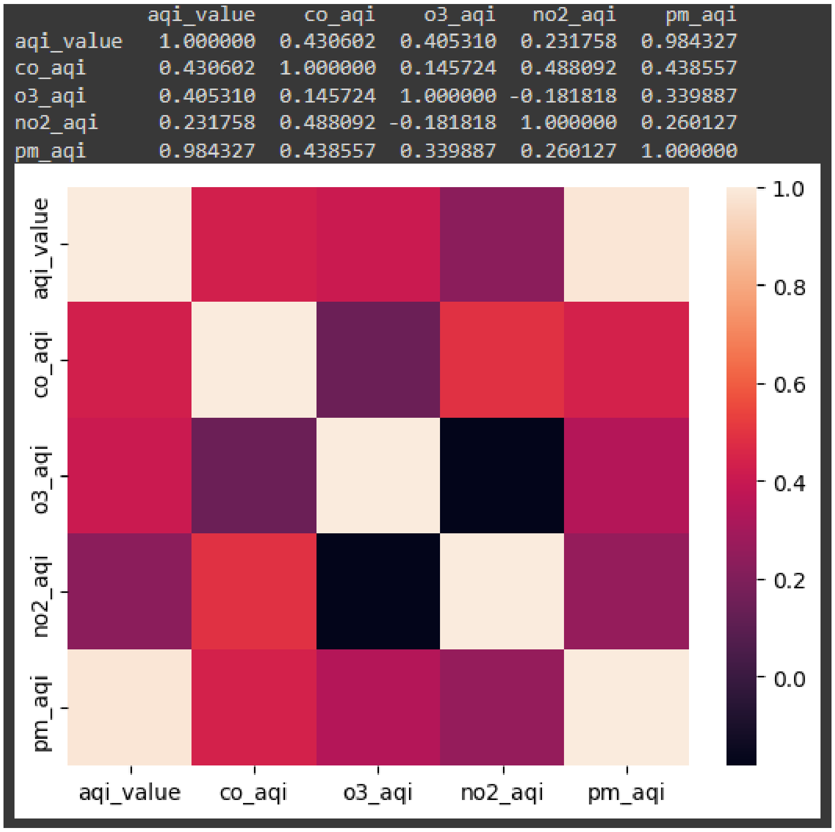


In the creation of this histogram, I used the Matplotlib library to visually represent the distribution of the 'AQI Category' feature in the Global Air Pollution Dataset. The primary purpose behind employing this visualization is to gain insights into the frequency distribution of air quality categories, providing a clear depiction of the occurrences within each category. With over 24,000 rows in the dataset, visualizing the distribution becomes imperative for understanding the prevalence of different air quality classifications. This histogram effectively illustrates the distribution of 'AQI Category' from 'Good' to 'Hazardous,' allowing for a quick assessment of the dataset's overall air quality composition. It serves as a valuable tool not only for grasping the general trends in air quality but also for identifying potential imbalances or disparities among different categories. Examining the histogram can unveil patterns such as whether certain air quality categories are more prevalent or if there are notable variations in the distribution, contributing to a comprehensive understanding of the dataset's composition and aiding in subsequent analytical decisions.

In this visualization, the histogram portrays the frequency distribution of air quality categories, providing a comprehensive overview of the dataset. Each bar on the x-axis corresponds to a specific AQI category, including 'Good,' 'Moderate,' 'Sensitive' (Unhealthy for Sensitive Groups), 'Unhealthy,' 'Very Unhealthy,' and 'Hazardous.' The y-axis scale, ranging from 0 to 10,000, measures the frequency of occurrences. The six bins on the chart align with each respective category, offering a clear representation of the distribution across the entire dataset. A notable observation emerges at first glance: the distribution exhibits a rightward skew or positive skewness, indicating that certain air quality categories have a higher frequency of occurrence than others.

Upon closer inspection, it becomes intriguing to explore the unexpected relationship between the 'Sensitive' and 'Unhealthy' categories. Contrary to the anticipated downward trend, the visualization reveals that the 'Sensitive' category does not surpass the 'Unhealthy' category in frequency. This counterintuitive finding prompts further investigation into the factors influencing these two categories. Additionally, the rightward skewness suggests that specific air quality issues are more prevalent, warranting a deeper dive into the underlying causes and potential correlations with environmental or geographical factors. Overall, this visualization serves as a valuable starting point for uncovering patterns and trends within the dataset, laying the foundation for more in-depth analysis.

**Data Visualization 2:**



```
           aqi_value    co_aqi    o3_aqi   no2_aqi    pm_aqi
aqi_value   1.000000  0.430602  0.405310  0.231758  0.984327
co_aqi      0.430602  1.000000  0.145724  0.488092  0.438557
o3_aqi      0.405310  0.145724  1.000000 -0.181818  0.339887
no2_aqi     0.231758  0.488092 -0.181818  1.000000  0.260127
pm_aqi      0.984327  0.438557  0.339887  0.260127  1.000000
```

In creating this visualization, seaborn `heatmap` function, coupled with sklearn.model_selection module for variable configuration, facilitated a comprehensive exploration of the relationship between the target variable 'AQI Category' and the numerical features representing pollutant AQI values. The heatmap matrix exhibits these variables on both the x and y axes, accompanied by a color scale on the right side. The color scale transitions from beige to black, where the lighter shades of beige equal to stronger correlations. The color scheme offers an intuitive representation of the correlation scores, allowing for quick identification of the most impactful relationships within the dataset. The varying shades guide the viewer in discerning the strength and directionality of these associations, fostering a deeper understanding of how different pollutants contribute to the overall air quality index. This visualization serves not only as a descriptive tool but also as a valuable precursor to more advanced analyses and targeted investigations into the factors influencing air quality.

In interpreting the correlation scores, a striking insight emerges, particularly in the strong relationship between 'AQI Value' and 'PM2.5 AQI' scores. The correlation coefficient of 0.984327 signifies a robust positive correlation, indicating that as the overall air quality index increases, the particulate matter AQI values tend to rise as well. This notable association is visually highlighted by the beige color squares in the upper right and lower left corners of the correlation matrix. This singular square stands out amid the predominantly weaker correlations seen in the other cells, showcasing the unique strength of the relationship between overall air quality and particulate matter levels. In the context of data analysis and decision-making, the profound positive correlation between 'AQI Value' and 'PM2.5 AQI' holds practical implications for data column retention. Given their exceptionally strong relationship, one might consider consolidating or prioritizing these columns in further analyses, as they provide redundant yet crucial information about air quality.

Looking further into the matrix, the appearance of dark purple shades indicates negative correlations. Specifically, the correlation coefficient of -0.181818 between 'O3 AQI' and 'NO2 AQI' suggests an inverse relationship. As ozone AQI values increase, nitrogen dioxide AQI values tend to decrease, and vice versa. The dark purple color serves as a visual cue for these negative associations, adding depth to our understanding of the complex interplay between different pollutant AQI values. The identified negative correlation between 'O3 AQI' and 'NO2 AQI' introduces a strategic element in the decision-making process regarding column inclusion. Recognizing this inverse relationship suggests that including both ozone and nitrogen dioxide AQI values might offer a more comprehensive understanding of air quality dynamics. However, it also prompts a careful evaluation of redundancy, as the information conveyed by one pollutant's AQI value may already be implied by the other. This nuanced interpretation guides the decision-making on whether to retain both columns or prioritize the one with more direct relevance to the research objectives, contributing to a more refined and efficient analysis.  In summary, this correlation matrix not only highlights the dominant positive correlation between 'AQI Value' and 'PM2.5 AQI' but also unveils nuanced relationships and negative correlations, contributing valuable insights into the dynamics of air quality factors.

# Data Modeling

## Data Modeling Definitions

### Extreme Gradient Boost (XGBoost)

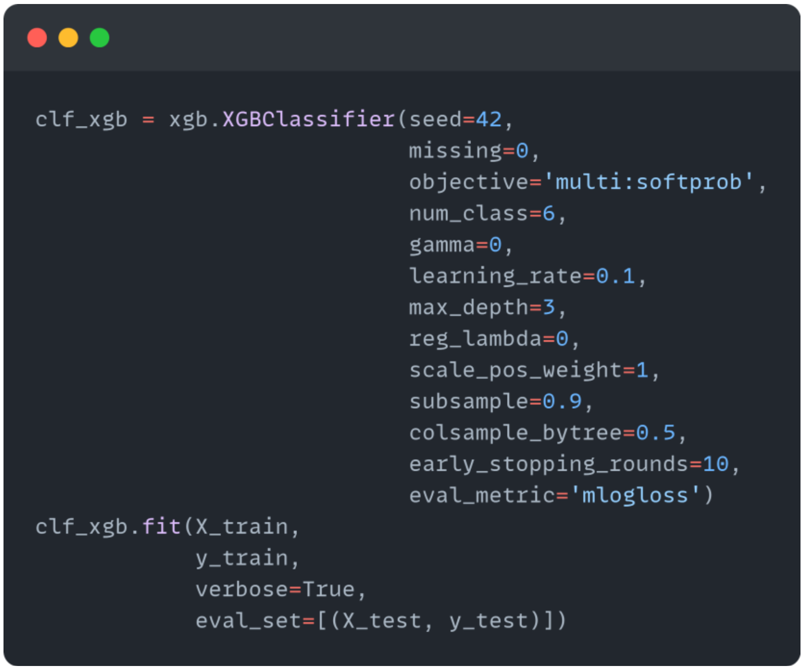
XGBoost, or Extreme Gradient Boosting, is a highly acclaimed open-source machine learning library designed for optimized distributed gradient boosting within the Gradient Boosting framework. It functions as a scalable and distributed gradient-boosted decision tree (GBDT) library, excelling in solving regression, classification, and ranking problems. Rooted in supervised machine learning, decision trees, ensemble learning, and gradient boosting, XGBoost leverages these concepts to find patterns in labeled datasets and predict outcomes using a tree-based ensemble model (*What is XGBoost?* 2023).

The adoption of XGBoost for Air Quality Index (AQI) prediction emerged as a compelling choice after an initial inclination towards Random Forest. A pivotal Zoom class introduced me to XGBoost, and subsequent literature exploration, notably the study by Van et al. (2022, as cited in Sarkar et al., 2022), emphasized XGBoost's superior performance in forecasting AQI values. The algorithm's remarkable accuracy, reflected in metrics like MAE, RMSE, and R2, positioned it as a promising alternative, particularly given its effectiveness in handling imbalanced air quality datasets. Building on this foundation, further research, such as the work by Nigam et al. (2015, as cited in Sarkar et al., 2022), underscored XGBoost's consistent success in AQI prediction, outshining other models like Decision Tree and Random Forest. The study emphasized the need for real-time datasets, enhancing XGBoost's accuracy and applicability in real-world scenarios. Recognizing air pollution's severe health implications, XGBoost's precision in predicting AQI emerges as a valuable tool for proactive air quality management and timely public health interventions, aligning with the overarching goal of mitigating the adverse effects of pollution on human health and the environment.

## Multi-Layer Perceptron (MLP)

The chosen Deep Learning model for air pollution prediction is a Multi-Layer Perceptron (MLP), a classical neural network architecture with one or multiple hidden layers. MLPs have gained prominence, constituting a significant portion of deep learning papers focused on air pollution prediction. Notable applications include studies predicting O3 (Ozone) levels in Corsica and PM2.5 (Particulate Matter 2.5) concentrations in Northern China. Traditional MLP models often maintain classical structures, but some variations and improvements have been explored. For instance, Hoi et al. (2013, as cited in Méndez et al., 2023) incorporated a Kalman filter into the back-propagation algorithm for improved accuracy in predicting PM10(Particulate Matter 10) concentrations in Macao. Fu et al. (2015, as cited in Méndez et al., 2023) introduced a modified MLP with a rolling mechanism and accumulated generating operation of a gray model to enhance predictions of PM2.5 and PM10 concentrations in Chinese cities. These adaptations demonstrate the versatility and adaptability of MLP models in the context of air pollution prediction.

**Data Model 1**



```
clf_xgb = xgb.XGBClassifier(seed=42,
                            missing=0,
                            objective='multi:softprob',
                            num_class=6,
                            gamma=0,
                            learning_rate=0.1,
                            max_depth=3,
                            reg_lambda=0,
                            scale_pos_weight=1,
                            subsample=0.9,
                            colsample_bytree=0.5,
                            early_stopping_rounds=10,
                            eval_metric='mlogloss')
clf_xgb.fit(X_train,
            y_train,
            verbose=True,
            eval_set=[(X_test, y_test)])
```

In this model, I've employed the XGBoost algorithm to predict the Air Quality Index (AQI) category, which involves classifying air quality into one of six distinct classes. The XGBoost classifier is a robust choice due to its ability to handle complex relationships within the data, making it well-suited for the nuances of air quality prediction. During the model development, I used GridSearchCV for cross-validation, fine-tuning essential parameters like 'gamma,' 'learning_rate,' 'max_depth,' and 'subsample' to optimize the model's predictive performance.

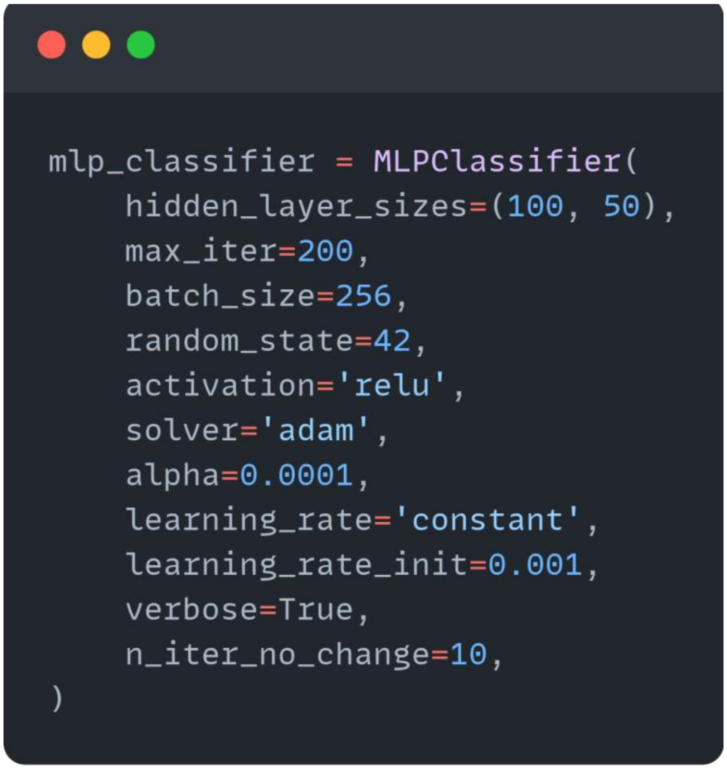
One crucial aspect of the model configuration is the choice of evaluation metric, where I've selected 'mlogloss' (multi-logloss). This metric is particularly suitable for our multi-class classification task, providing a comprehensive measure of how well the model predicts the probabilities of each class. In essence, 'mlogloss' quantifies the accuracy of the predicted probabilities, penalizing the model for being less confident in the correct class. By optimizing for 'mlogloss,' the model is not only trained to make accurate predictions but also to provide well-calibrated probability estimates for each AQI category. This ensures a more nuanced evaluation of the model's performance, considering the uncertainties associated with predicting multiple classes in air quality assessments.

Let's break down each parameter used in this XGBoost classifier:

1. **seed (Random Seed):** This parameter ensures reproducibility by setting a random seed. It ensures that the random initialization and sampling processes within the algorithm are consistent across different runs.
2. **missing:** This parameter defines the value that is treated as missing in the dataset. It helps XGBoost handle missing data effectively during the training process.
3. **objective:** For a multi-class classification problem like predicting AQI categories, the 'multi:softprob' objective is chosen. It outputs a vector of predicted probabilities for each class, allowing us to assess the likelihood of the input belonging to each class.
4. **num_class:** This specifies the number of classes in the classification task. In this case, it's set to 6, corresponding to the six AQI categories.
5. **gamma:** It controls whether a given node will split based on the expected reduction in loss after the split. A higher gamma value leads to fewer splits, preventing overfitting.
6. **learning_rate:** Also known as the "eta" parameter, it scales the contribution of each tree. Lower values make the model more robust by slowing down the learning process.
7. **max_depth:** Determines the maximum depth of a tree. It controls the complexity of the individual decision trees in the ensemble.
8. **reg_lambda (L2 Regularization Term):** This term penalizes large coefficients to prevent overfitting. A higher value of reg_lambda increases the regularization strength.
9. **scale_pos_weight:** Addresses the class imbalance by scaling the weights of positive examples. Useful when one class is underrepresented.

10. **subsample:** Denotes the fraction of training data to be randomly sampled during each boosting round. It prevents overfitting by introducing randomness.
11. **colsample_bytree:** Specifies the fraction of features to be randomly sampled for building each tree. It further introduces diversity into the ensemble.
12. **early_stopping_rounds:** This parameter allows the model to stop training if the performance on the validation set doesn't improve for a specified number of rounds. It helps prevent overfitting and speeds up training.
13. **eval_metric:** Set to 'mlogloss' for our multi-class classification. It determines the evaluation metric to be used during training and validation.

**Data Model 2**



```
mlp_classifier = MLPClassifier(
    hidden_layer_sizes=(100, 50),
    max_iter=200,
    batch_size=256,
    random_state=42,
    activation='relu',
    solver='adam',
    alpha=0.0001,
    learning_rate='constant',
    learning_rate_init=0.001,
    verbose=True,
    n_iter_no_change=10,
)
```

In addressing the classification problem of predicting the Air Quality Index (AQI) category with a Multi-Layer Perceptron (MLP) model, I opted for simplicity in hyperparameter tuning, relying on default settings due to the model's remarkable performance. The MLP classifier was configured with two hidden layers of 100 and 50 neurons, respectively, utilizing the ReLU activation function for the hidden layers. The solver 'adam,' known for its efficiency in handling large datasets, was chosen, and the model was trained over 200 iterations with a constant learning rate of 0.001. Batch size, set to 256, was adjusted based on available memory. To avoid overfitting, an L2 penalty (alpha=0.0001) was implemented, and early stopping was enforced after 10 iterations without improvement (n_iter_no_change=10).

Upon model training, the initial evaluation through a classification report demonstrated impressive results, with precision, recall, and F1-scores consistently exceeding 90% across all AQI categories. Notably, this achievement was obtained without fine-tuning the hyperparameters, underscoring the robustness and effectiveness of the chosen architecture and default settings. This approach emphasizes simplicity and efficiency, providing a strong baseline for AQI category prediction.

Let's break down each parameter used in the configuration of the Multi-Layer Perceptron (MLP) classifier:

1. **hidden_layer_sizes**: This parameter defines the number of neurons in each hidden layer. In this model, there are two hidden layers with 100 and 50 neurons, respectively.

2. **max_iter**: It specifies the maximum number of iterations (epochs) the MLP classifier should undergo during training. The model will stop training once this limit is reached.

3. **batch_size**: This parameter determines the number of samples used in each iteration during training. It is adjusted based on available memory, and a higher batch size can expedite training but may require more memory.

4. **random_state**: This is a seed for the random number generator, ensuring reproducibility of the results. The same seed will produce the same results on each run.

5. **activation**: It defines the activation function for the hidden layers. 'ReLU' (Rectified Linear Unit) is a commonly used activation function that introduces non-linearity to the model.

6. **solver**:  In this model 'Adam' is an optimization algorithm that adapts the learning rates of each parameter during training. It is known for being efficient in handling large datasets and is often a good default choice.

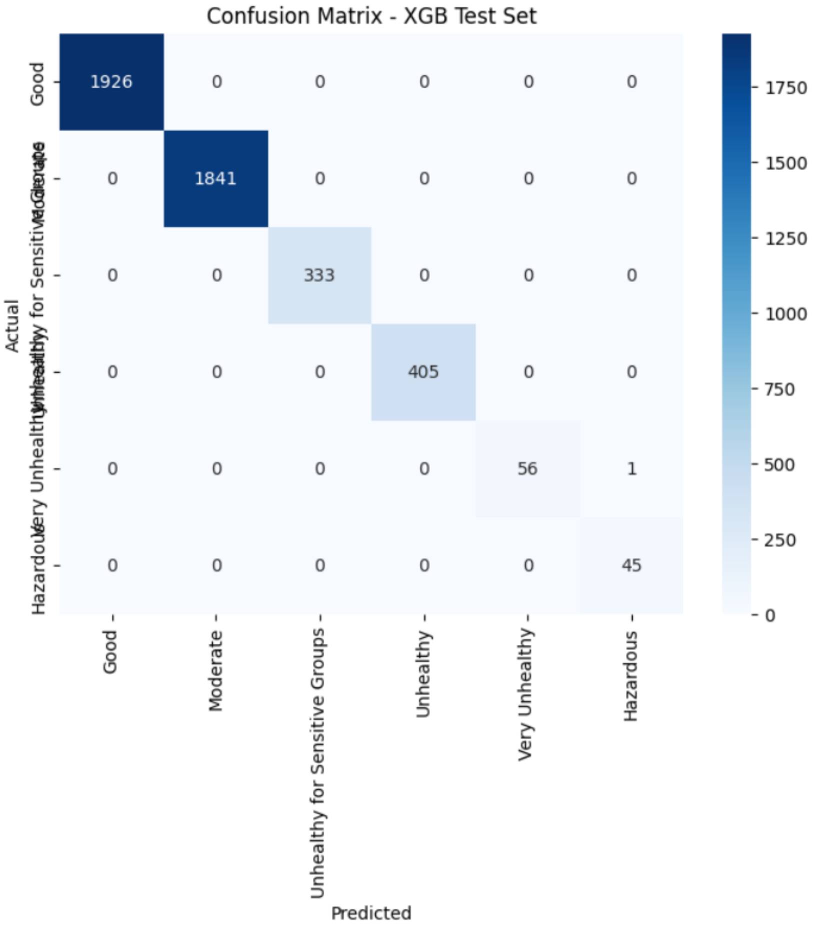
7. **alpha**: This parameter represents the L2 penalty (regularization term) applied to the weights. It helps prevent overfitting by penalizing large weight values.

8. **learning_rate**: These parameters control the learning rate of the model. In this case, the learning rate is kept constant at 0.001 throughout training.

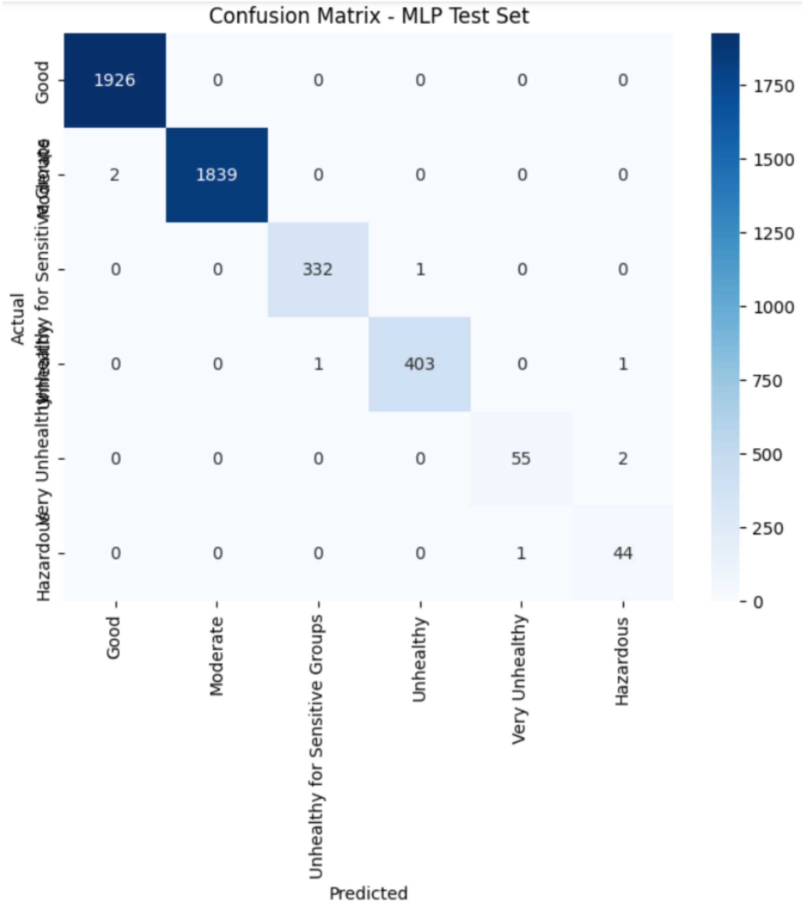
9. **verbose**: It enables verbose output during training, providing additional information about the training process.

10. **n_iter_no_change**: This parameter specifies the number of iterations with no improvement to wait before stopping training. If the performance on the validation set does not improve for 10 consecutive iterations, training will halt.

## Review of Data Models



Accuracy of XGBoost testing data: 0.9997829390058607



Accuracy of MLP testing data: 0.9982635120468851

In analyzing the confusion matrices for the testing data of both the XGBoost and MLP models, it is evident that both models demonstrated exceptional performance in predicting the Air Quality Index (AQI) category. The accuracy scores for XGBoost and MLP were .999 and .998, respectively.

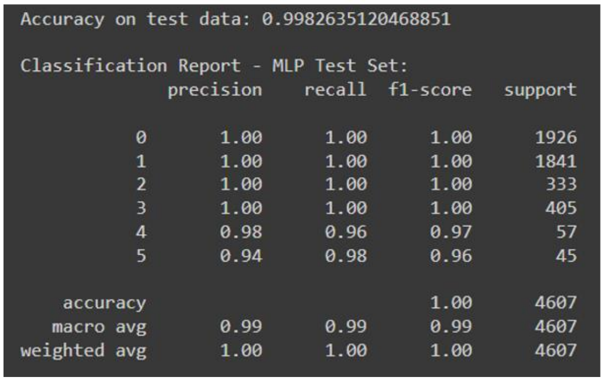
In terms of overall performance, both models achieved perfect accuracy on the training data. However, on the testing set, XGBoost outperformed MLP by misclassifying only one instance, whereas the MLP model misclassified 8 instances. Despite the presence of an unbalanced dataset, both models surpassed expectations in their predictive capabilities.

Therefore, based on the results, the Champion Model for AQI Classification is XGBoost, although it is noteworthy that MLP's performance was nearly on par with XGBoost.

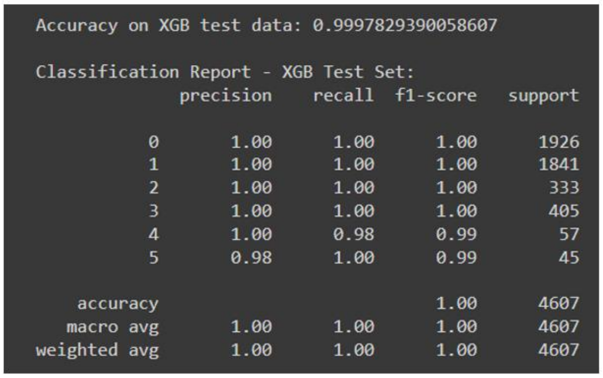
## Final Results

### Findings

### Classification Report:



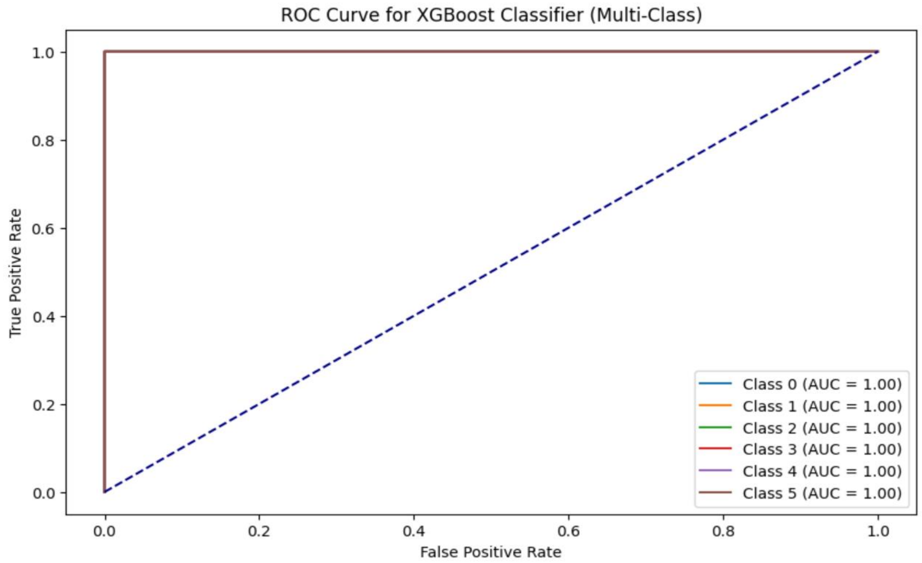
```
Accuracy on test data: 0.9982635120468851

Classification Report - MLP Test Set:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1926
           1       1.00      1.00      1.00      1841
           2       1.00      1.00      1.00       333
           3       1.00      1.00      1.00       405
           4       0.98      0.96      0.97        57
           5       0.94      0.98      0.96        45

    accuracy                           1.00      4607
   macro avg       0.99      0.99      0.99      4607
weighted avg       1.00      1.00      1.00      4607
```

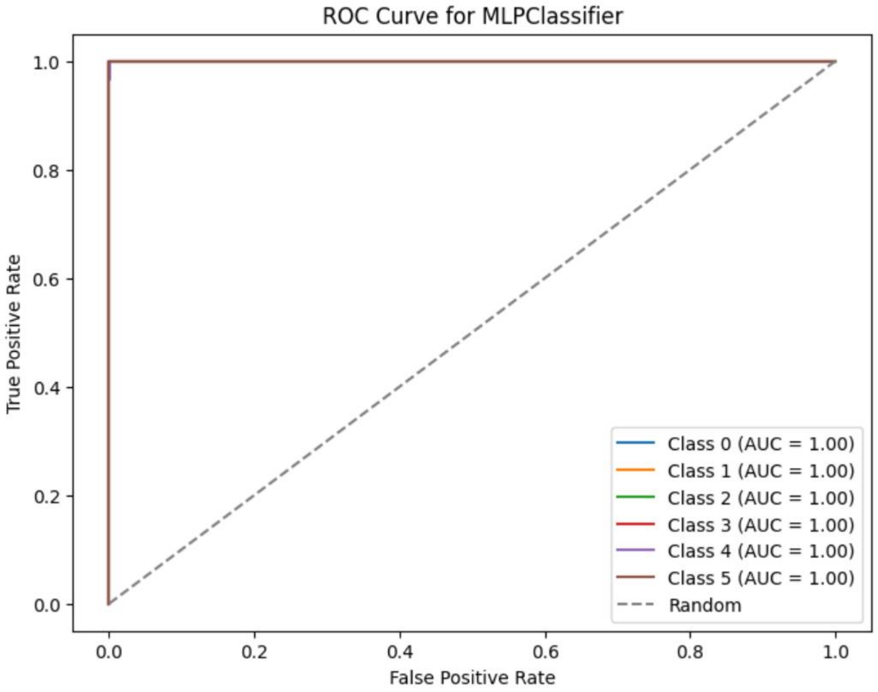


```
Accuracy on XGB test data: 0.9997829390058607

Classification Report - XGB Test Set:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1926
           1       1.00      1.00      1.00      1841
           2       1.00      1.00      1.00       333
           3       1.00      1.00      1.00       405
           4       1.00      0.98      0.99        57
           5       0.98      1.00      0.99        45

    accuracy                           1.00      4607
   macro avg       1.00      1.00      1.00      4607
weighted avg       1.00      1.00      1.00      4607
```

Both the XGBoost and MLP models exhibit exceptional accuracy in their predictions. A closer examination of class-wise metrics reveals consistently high precision, recall, and F1-scores across all classes, indicative of robust and reliable performance. Notably, there are minor variations in the performance of classes 4 and 5, which can be attributed to the relatively fewer instances of these classes, highlighting the challenges associated with unbalanced data. Despite these variations, both models demonstrate a commendable ability to classify instances accurately across the diverse range of classes present in the dataset.

**ROC-AUC:**

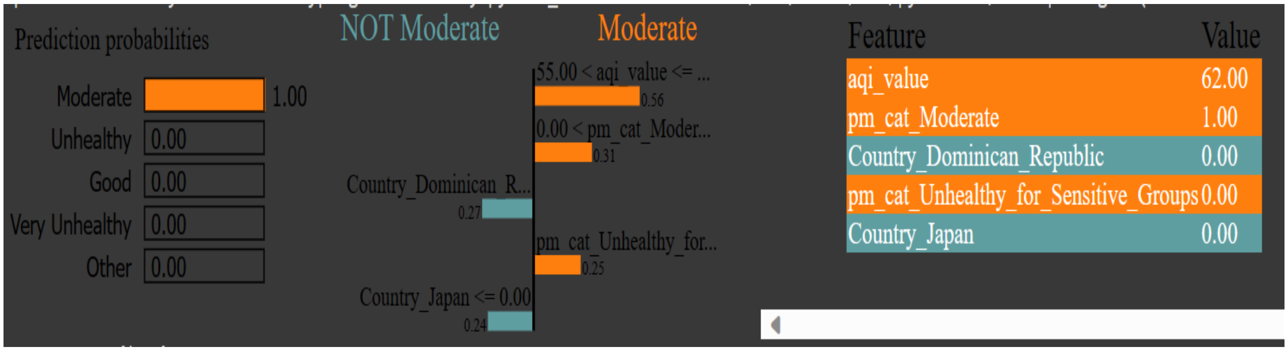




An ROC AUC score of 1 for both the XGBoost and MLP classifiers across all six classes is indicative of exceptional model performance. This perfect score implies that the classifiers have achieved a balance between sensitivity and specificity, showcasing their ability to accurately discriminate between positive and negative instances. The absence of false positives and false negatives, coupled with consistent high performance across all classes, suggests that both models have successfully learned intricate patterns in the data and are robust in their generalization. The models exhibit a rare level of reliability, effectively separating probability distributions of different classes and demonstrating a lack of bias towards specific categories. Overall, the ROC-AUC score of 1 stresses the models' proficiency in making precise predictions, emphasizing their capability to handle diverse patterns and maintain accuracy across the entire spectrum of classes.

**Lime Explainer:**

**XGBoost**



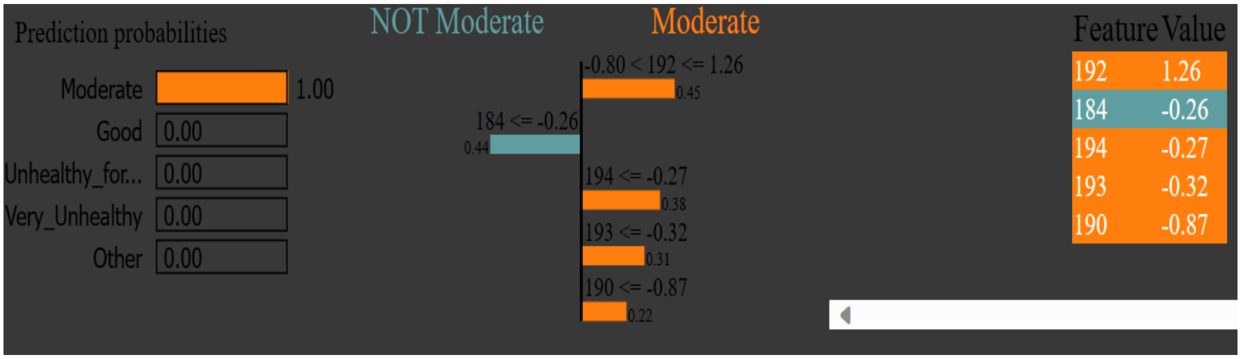
Prediction probabilities

| | |
|---|---|
| Moderate | 1.00 |
| Unhealthy | 0.00 |
| Good | 0.00 |
| Very Unhealthy | 0.00 |
| Other | 0.00 |

| Feature | Value |
|---|---|
| aqi_value | 62.00 |
| pm_cat_Moderate | 1.00 |
| Country_Dominican_Republic | 0.00 |
| pm_cat_Unhealthy_for_Sensitive_Groups | 0.00 |
| Country_Japan | 0.00 |

Looking at the XGBoost model's prediction for the 10th instance, applying the LIME explainer, a robust confidence level of 1.00 is evident. The model's inclination towards predicting the "Moderate" class is discernibly influenced by specific feature thresholds, especially noteworthy when the country is neither the Dominican Republic (0.27) nor Japan (0.24). Notably, the absence of the "Particulate Matter 2.5 Unhealthy for Sensitive Groups" category (pm_cat_Unhealthy_for_Sensitive_Groups: 0) emerges as a noteworthy contributing factor. The LIME explainer provides valuable insights by showing how the presence or absence of distinct features shapes the model's predictions. Key contributors, such as the overall AQI Value (0.56), a Moderate Particulate Matter 2.5 Category (0.31), and specific country thresholds collectively support the model's confident forecast for the "Moderate" class, offering a comprehensive understanding of the decision-making process for the 10th instance.

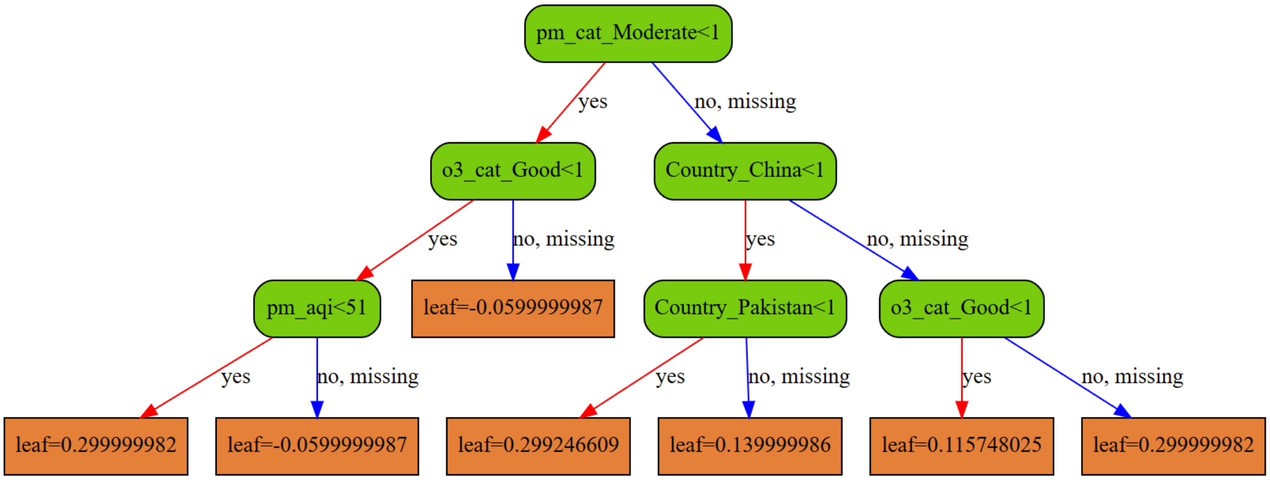
**MLP**

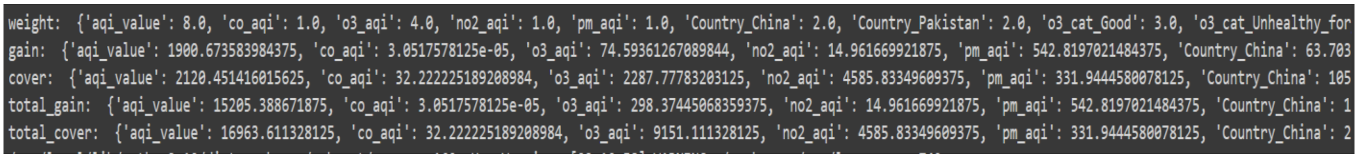


['192: pm_cat_Moderate', '184: o3_cat_Moderate', '194: pm_cat_Unhealthy_for_Sensitive_Groups', '193: pm_cat_Unhealthy', '190: pm_cat_Good']

Examining the prediction probabilities for the 10th instance for the MLP model, from left to right, a notable high confidence of 1.00 is observed. In the center, the feature threshold influencing the deviation from the "Moderate" class is identified as feature 184 (o3_cat_Moderate) with a weight of (.44). Features that contribute to the prediction favoring the "Moderate" class include feature 192 (pm_cat_Moderate) with a weight of 0.45, feature 194 (pm_cat_Unhealthy_for_Sensitive_Groups) with a weight of 0.38, feature 193 (pm_cat_Unhealthy) with a weight of 0.31, and finally, feature 190 (pm_cat_Good) with a weight of 0.22. It is crucial to highlight the distinctions in results for the 10th instance and emphasize that the MLP is an unsupervised learning algorithm, emphasizing the unique nature of its predictions in this context.

**XGBoost Decision Tree:**





```
weight:  {'aqi_value': 8.0, 'co_aqi': 1.0, 'o3_aqi': 4.0, 'no2_aqi': 1.0, 'pm_aqi': 1.0, 'Country_China': 2.0, 'Country_Pakistan': 2.0, 'o3_cat_Good': 3.0, 'o3_cat_Unhealthy_for
gain:  {'aqi_value': 1900.673583984375, 'co_aqi': 3.0517578125e-05, 'o3_aqi': 74.59361267089844, 'no2_aqi': 14.961669921875, 'pm_aqi': 542.8197021484375, 'Country_China': 63.703
cover:  {'aqi_value': 2120.451416015625, 'co_aqi': 32.22225189208984, 'o3_aqi': 2287.77783203125, 'no2_aqi': 4585.83349609375, 'pm_aqi': 331.9444580078125, 'Country_China': 105
total_gain:  {'aqi_value': 15205.388671875, 'co_aqi': 3.0517578125e-05, 'o3_aqi': 298.37445068359375, 'no2_aqi': 14.961669921875, 'pm_aqi': 542.8197021484375, 'Country_China': 1
total_cover:  {'aqi_value': 16963.611328125, 'co_aqi': 32.22225189208984, 'o3_aqi': 9151.111328125, 'no2_aqi': 4585.83349609375, 'pm_aqi': 331.9444580078125, 'Country_China': 2
```

Going over the parameters of the first tree in the XGBoost model, several key insights emerge. The ‘weight’ parameter means the frequency of feature appearance during tree construction, highlights 'aqi_value' with a weight of 8.0, indicating its substantial importance in the model. Moving on to 'gain,' which measures the accuracy improvement brought by each feature, it's noteworthy that 'pm_cat_Moderate' and 'pm_cat_Unhealthy' exhibit significant gains, reinforcing their impactful contributions. The 'cover' parameter, reflecting the relative quantity of observations associated with a feature, emphasizes the broad influence of 'pm_cat_Moderate' and 'pm_cat_Unhealthy_for_Sensitive_Groups.' Meanwhile, 'total gain' and 'total cover' provide cumulative measures, showcasing features like 'pm_cat_Moderate' and 'o3_aqi' with considerable overall impact on the model's predictive performance. In essence, these parameters collectively elucidate the importance and reach of individual features in the initial tree of the XGBoost model.

## Review of Success or Completion

With the goal of comparing traditional machine learning and deep learning algorithms for a multi-classification prediction problem, the project aimed to initially utilize a Random Forest algorithm and a Keras classifier. However, a strategic pivot happened, leading to the adoption of XGBoost and a Multi-Layer Perceptron (MLP) for distinct reasons. The selection of XGBoost was informed by research indicating its superior performance in Air Quality Index (AQI) predictions, particularly excelling with unbalanced data. Despite the unfamiliarity with XGBoost, the decision was embraced as a challenge. A subsequent shift from a Keras classifier to MLP resulted from technical challenges related to system and technology issues, specifically difficulties in debugging the Keras installation. While this transition deviated from the initial plan, the successful achievement of the primary project goal highlights the adaptability and problem-solving skills exercised in navigating unforeseen obstacles. The experience not only broadened proficiency in Python but also introduced a novel algorithm, contributing to a valuable learning curve.

While initially encountering challenges in implementing Explainable AI techniques, specifically Shap and Lime, as initially envisioned during the project's inception, I managed to overcome obstacles and successfully incorporate a Lime explainer for both models. Despite facing difficulties in debugging, my perseverance and exploration of various debugging techniques eventually led to a favorable outcome. This experience stresses the need of gaining proficiency in debugging and navigating diverse packages and modules. In conclusion, I am pleased to affirm that the main objectives of the project were successfully achieved, marking a significant milestone in addressing the defined challenges and contributing to the broader goals of the study.

## Potential Data Privacy and Data Security Issues

The Global Pollution Dataset is derived from eLichen's air quality map, using sensors developed by eLichens. Initially perceived as devoid of data privacy concerns due to open data access, a subsequent investigation, involving account creation and app download, revealed the acquisition of Air Quality Data through sensors linked to the eLichens Air Quality Monitor. User registration for the application needs the disclosure at a minimum of the user's name and email address. Moreover, users who acquire an Air Quality Monitor from the company can synchronize it with the app to visualize comprehensive data on major indoor air pollutants, as well as ancillary information such as temperature, humidity, pressure, ambient noise, and light.

However, attention is drawn to a particular device within the spectrum of offerings—the eLichens indoor quality monitor, which possesses the capability to record ambient noise. In accordance with legal considerations outlined by Stacy Gray (2016), the deployment of devices with audio recording functionality, such as security cameras, mandates both users and manufacturers to remain cognizant of pertinent anti-surveillance statutes. At the federal level, the Wiretap Act expressly prohibits the intentional interception of the contents of wire, oral, or electronic communications without the prior consent of at least one party. Consequently, the acquisition of user consent becomes a pivotal aspect germane to the ethical usage of the dataset. Given the potential implication of anti-surveillance statutes stemming from the recording of audio, users must be explicitly informed about this feature, with their consent obtained prior to the collection of such data. An opt-in/opt-out mechanism should allow users of the agency to exercise control over the inclusion of this specific feature. Importantly, users who opt to connect an Air Quality Monitor to the app are likely to share additional data pertaining to their indoor environment, needing an encompassing consent process that delineates the specifics of data collection and usage, consequently ensuring users are informed of the nature of the transmitted information and its intended purpose.

The evolving legislative landscape in the realm of consumer protection is noteworthy, particularly with the recent passage of the Informing Consumers About Smart Devices Act in July 2023. This bipartisan bill, co-sponsored by Senators Maria Cantwell and Ted Cruz and integrated into the National Defense Authorization Act, carries significant implications for the dataset under consideration. Specifically, the Act mandates pre-purchase disclosures for smart devices and appliances harboring concealed microphones or cameras with recording or data transmission capabilities. In response to mounting privacy apprehensions associated with these devices, the legislation tasks the Federal Trade Commission (FTC) with formulating new disclosure guidelines. By doing so, the Act aims to enhance transparency for consumers, ensuring they receive information regarding audio or visual recording components in household appliances such as refrigerators, washers, dryers, and dishwashers. As of the present, the legislation awaits approval from the President (2023). This legislative development reinforces the relevance of informed consent and transparent disclosure practices in the context of data collection from smart devices, aligning with broader efforts to safeguard consumer privacy and establish clear guidelines for the integration of recording features in everyday appliances.

### Recommendations for Future Analysis

For future analyses, leveraging ensemble models can significantly enhance the accuracy of air quality predictions. Employing models such as Random Forest or Gradient Boosting allows for capturing intricate relationships within the data. The ensemble approach, which combines predictions from multiple models, tends to mitigate overfitting and results in more robust and accurate predictions. Moreover, exploring advanced hyperparameter tuning techniques, such as Bayesian optimization or genetic algorithms, can further optimize the performance of these ensemble models, particularly beneficial for handling the complexity inherent in air quality datasets. Employing techniques like these not only refines the model's predictive capacity but also helps in identifying the most influential parameters governing air quality dynamics.

Another avenue for improvement lies in feature engineering. Experimenting with new features, such as incorporating lag features based on historical concentrations or meteorological data, can provide additional context, and contribute to refining predictions. This process involves systematically selecting, creating, or transforming features to enhance the model's understanding of the underlying patterns in the data. Furthermore, a thorough temporal analysis should be integrated, delving into patterns related to seasons, weekdays, holidays, and other temporal factors. Understanding how air quality varies over time can contribute to more nuanced predictions and facilitate the development of strategies to address specific temporal challenges. In summary, combining ensemble models with advanced hyperparameter tuning, strategic feature engineering, and comprehensive temporal analysis presents a multifaceted approach to refining air quality predictions for future analyses. These approaches not only enhance prediction accuracy but also provide valuable insights into the temporal dynamics and parameter importance.

# References

Air Now. (n.d.). *AQI & Health*. AQI & Health | AirNow.gov. https://www.airnow.gov/aqi-and-health/

Blumenschein, M., Debbeler, L. J., Lages, N. C., Renner, B., Keim, D. A., & El-Assady, M. (2020). V-plots: Designing hybrid charts for the comparative analysis of data distributions. *Computer Graphics Forum*, *39*(3), 565–577. https://doi.org/10.1111/cgf.14002

Dutta, D., & Pal, S. K. (2022). Z-number-based AQI in rough set theoretic framework for interpretation of air quality for different thresholds of PM2.5 and PM10. *Environmental Monitoring and Assessment*, *194*(9). https://doi.org/10.1007/s10661-022-10325-z

eLichens. (n.d.). *Global Air Quality Map*. eLichens. https://www.elichens.com/global-air-quality-map

EPA. (2018, September). *Technical assistance document for the reporting of Daily Air Quality* . Technical Assistance Document for the Reporting of Daily Air Quality – the Air Quality Index (AQI). https://www.airnow.gov/sites/default/files/2020-05/aqi-technical-assistance-document-sept2018.pdf

EPA. (2023, August 3). *Air topics | US EPA - U.S. Environmental Protection Agency*. Environmental Topics | Air Topics. https://www.epa.gov/environmental-topics/air-topics

Gray, S. (2016, April). *Always on: Privacy implications of microphone-enabled devices*. Future of Privacy Forum. https://fpf.org/wp-content/uploads/2016/04/FPF_Always_On_WP.pdf

Merchant, F. A., Castleman, K. R., & Wu, Q. (2023). Correlation. In *Microscope Image Processing* (2nd ed., pp. 177–200). essay, Elsevier.

Muzdadid, H. A. (2022, November 8). *Global Air Pollution Dataset*. Kaggle. https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset

Méndez, M., Merayo, M. G., & Núñez, M. (2023). Machine learning algorithms to Forecast Air Quality: A survey. *Artificial Intelligence Review*, *56*(9), 10031–10066. https://doi.org/10.1007/s10462-023-10424-4

NVIDIA Corporation . (2023). *What is XGBoost?*. NVIDIA Data Science Glossary. https://www.nvidia.com/en-us/glossary/data-science/xgboost/

Perlmutt, L. D., & Cromar, K. R. (2019). Comparing associations of respiratory risk for the EPA Air Quality Index and health-based air quality indices. *Atmospheric Environment*, *202*, 1–7. https://doi.org/10.1016/j.atmosenv.2019.01.011

Sarkar, N., Gupta, R., Keserwani, P. K., & Govil, M. C. (2022). Air Quality index prediction using an effective hybrid deep learning model. *Environmental Pollution*, *315*, 120404. https://doi.org/10.1016/j.envpol.2022.120404

*Smart devices, appliances with hidden microphones, cameras must be disclosed to consumers*. U.S. Senate Committee on Commerce, Science, & Transportation. (2023, July 28). https://www.commerce.senate.gov/2023/7/smart-devices-appliances-with-hidden-microphones-cameras-must-be-disclosed-to-consumers

Smith, A. (2014, December 4). *Half of online Americans don't know what a privacy policy is*. Pew Research Center. https://www.pewresearch.org/short-reads/2014/12/04/half-of-americans-dont-know-what-a-privacy-policy-is/