**Data Science Project**

**Fallon Wynn**

**Global Air Pollution**

**February 25, 2025**

**Executive Summary**

**Project Scope**

The dissemination of air pollutants is an ongoing global issue that has detrimental impacts on public health and the environment. Using the Global Air Pollution Dataset from Kaggle, I plan to analyze the impact of global air pollution on public health. The data analytics tools that I plan to use for this project, include Excel, Python in Google Colab, and Tableau. The analysis of historical data for global air pollution will identify the trends in global air pollution to predict future air pollution levels. **The insights provided can be used by government agencies to make important decisions about regulations, policies, and legislature to further decrease air pollution in problematic geographic locations.** It is critical that air pollution trends are evaluated at a local and global level, in order to implement plans for mitigation that can decrease the number of premature deaths.

* Air pollution was the second largest risk factor of deaths in 2021.
* In 2021, air pollution was the cause of 8.1 million deaths worldwide (1 in 8 deaths).
* 709,000 of the total deaths from air pollution were children under the age of five.
* Pollutants such as, particulate matter, carbon monoxide, ozone, nitrogen dioxide and sulfur dioxide are of major public health concern.
* Particulate Matter (PM2.5) has been associated with the most harmful health impacts and is present in harmful levels around the world.

**Data Profiling and Preparation**

The Global Air Pollution Dataset consists of U.S. Air Quality Index (AQI) values for different pollutants for many cities world-wide. The AQI includes 6 ranges of index values, which corresponds to six color-coded health risk categories. Smaller AQI values are indicative of lower levels of air pollution, while larger AQI values are indicative of higher levels of air pollution. Real-time data was collected from elichens.com by the Kaggle author Hasib Al Muzdadid. The temporal coverage start date was 11/2/2022, but the data set was last updated on Kaggle in 2023.

* Dataset consists of 23,463 rows of data and 12 columns.
* There are 175 unique countries and 23,462 cities.
* There were some data quality issues, which involved missing data and one questionable outlier, which was addressed before data analysis.
* Different tools and techniques were utilized to process, prepare, and clean the data.
* The data was divided into an 80\20 split before model training. 80% of the data was used for model training and the other 20% was used for model testing and validation.

**Data Visualizations**

There are different visualization techniques that can be utilized for exploratory data analysis. To aide in a thorough analysis of the “Global Air Pollution Dataset”, I will be using visualization techniques, such as box plots, count plots, bar plots, histograms, scatter plots, Seaborn’s (SNS) pair plot, heat maps, and tree maps. A definition for each visualization type was provided. However, only a few of the created visualizations were thoroughly discussed, which include a box plot, a correlation heat map, and a tree map. The descriptive statistics, overall data distribution, the correlation of variables, and the proposed selected features for model training will be discussed.

**Data Modeling**

A multiple linear regression model and a classification model will be used to analyze the “Global Air Pollution Dataset” to predict or forecast future air pollution levels to determine its’ future effect on public health in different geographical regions. To provide a comprehensive analysis for the “Global Air Pollution Dataset”, a multiple linear regression model was created, trained, tested, and analyzed and 4 classification models (Gradient Boosting decision tree model, a Random Forest decision tree model, and two Support Vector Machine (SVM) models) were created, trained, tested, and analyzed to find the best performing model.

* The multiple linear regression model had an overall great performance.
* The SVM Model #1 was determined to be the most reliable and most accurate classification model with the least amount of overfitting.
* The multiple linear regression model can be used for quantitative forecasting of PM2.5 AQI Values and to study the relationship of the specific pollutants in a particular geographic region.
* The SVM Model #1 can be used for a qualitative forecasting of PM2.5 AQI Categories in specific geographic regions.
* Utilizing both models can capture both the linear and nonlinear trends in global air pollution.

**Final Results**

* The dataset contains air pollution data for only 89.7% of existing countries.
* There is insufficient or missing air pollution data for some countries, most likely due to the lack of high-quality air pollution monitoring systems or air pollution sensors and/or faulty low-cost air pollution sensors, and due to some governments not making their data public.
* 104 countries (includes 4088 cities) have PM2.5 AQI Values greater than 100, which indicates that there is poor air quality in 59.4% of all counties in the dataset.
* Asian and African countries have the worst average PM2.5 AQI Values.
* The top ten countries with the highest average PM2.5 AQI Values include, Republic of Korea, Bahrain, Mauritania, Pakistan, Aruba, Kuwait, United Arab Emirates, Senegal, India, and Saudi Arabia.
* The top ten countries with the lowest average PM2.5 AQI Values include, Andorra, Uruguay, Finland, Sweden, Papua New Guinea, Norway, Iceland, Maldives, Palau, and Solomon Islands.
* The USA has the average PM2.5 AQI Value in the moderate air quality range.
* The top ten countries with the most predictions represented in the data include, the USA, India, China, Brazil, Germany, Somalia, Finland, Turkey, Japan, and the Neverlands.
* The evaluation and visualization of PM2.5 AQI Value predictions by country for the multiple linear regression model showed that the model made very accurate predictions that were close to the actual PM2.5 AQI Values.
* The evaluation and visualization of PM2.5 AQI Category predictions by country for the SVM Model #1 showed that the model made relatively good predictions with minimal error, but can use more finer hyperparameter tuning and optimization to improve the model’s performance.

**Recommendations for Future Analysis**

Although the air pollution data gap needs to be filled to create the most accurate and productive models, both models can still be useful to help determine the possible future impacts of PM2.5 air pollution in different countries and regions in order to implement policy changes that can mitigate PM2.5 air pollution. Adding additional dimensions to the analysis for global air pollution could provide more useful insights to be used as a guidance by policy makers. Recommendations include:

* Collecting more data, which leads to better performing models.
* Incorporating additional features to further enhance models, such as emissions sources, hourly analysis to pinpoint timeframes when pollution levels peak, and population weighted annual average concentration for each country by calculating the PM2.5 per Capita to provide the level of exposure per person enabling a more equal comparison between countries.
* Use model ensembling, which is a better approach to model selection and has been proven to be better than any single model since it corrects for errors of each individual model.

**Table of Contents**

[Project Scope 9](#_Toc191279879)

[Problem Description 9](#_Toc191279880)

[Project Importance 10](#_Toc191279881)

[Background 11](#_Toc191279882)

[Data Set Description 14](#_Toc191279883)

[Data Analytics Tools 15](#_Toc191279884)

[Project Milestones 15](#_Toc191279885)

[Completion History 16](#_Toc191279886)

[Lessons Learned 16](#_Toc191279887)

[Data Profiling and Preparation 17](#_Toc191279888)

[Data Summary 17](#_Toc191279889)

[Data Definition/Data Profile 19](#_Toc191279890)

[Data Preparation 22](#_Toc191279891)

[Data Visualizations 25](#_Toc191279892)

[Descriptive Statistics 25](#_Toc191279893)

[Data Visualization Definitions 26](#_Toc191279894)

[Data Visualization 1 28](#_Toc191279895)

[Data Visualization 2 31](#_Toc191279896)

[Data Visualization 3 34](#_Toc191279897)

[Data Modeling 35](#_Toc191279898)

[Data Modeling Definitions 35](#_Toc191279899)

[Data Model 1: Linear Regression Model 40](#_Toc191279900)

[Data Model 2: Decision Tree Model # 1 43](#_Toc191279901)

[Data Model 3: Decision Tree Model #2 46](#_Toc191279902)

[Data Model 4: SVM Model #1 48](#_Toc191279903)

[Data Model 5: SVM Model #2 50](#_Toc191279904)

[Review of Data Models 53](#_Toc191279905)

[Final Results 54](#_Toc191279906)

[Findings 54](#_Toc191279907)

[Review of Success or Completion 62](#_Toc191279908)

[Potential Data Privacy and Data Security Issues 64](#_Toc191279909)

[Recommendations for Future Analysis 65](#_Toc191279910)

[Appendix 68](#_Toc191279911)

[References 95](#_Toc191279912)

# Project Scope

# Problem Description

The dissemination of air pollutants is an ongoing global issue. Global air pollution has detrimental impacts on public health and the environment (Health Effects Institute, 2024). There are studies that linked air pollution to triggering and worsening asthma and other respiratory diseases, increased risks of cardiovascular diseases and cancer, and the cause of death. There is evidence that even low levels of air pollution cause health issues (Hunter, 2020). This project will focus on the impact of global air pollution on public health. It will discuss which countries have the highest and lowest air pollution and how do other countries compare to the United States of America.

Decreasing air pollutants can increase human life expectancy. In the past, multiple air quality legislatures were implemented due to the rising health toll associated with air pollution (Hunter, 2020). As the concentration of air pollutants differs geographically, so do the sources or origin of the pollutants. According to Amann et al. (2020), “In the past, changes in air pollutant emissions have been driven by economic development and policy choices” (p. 5). Analyzing historical data for global air pollution can identify the trends in global air pollution, enabling the prediction or forecasting of future air pollution levels (Amann et al., 2020). **The data analytics problem that I am analyzing is based on historical global air pollution and its’ future effect on public health. The insights provided can be used by government agencies to make important decisions about regulations, policies, and legislature to further decrease air pollution in problematic geographic locations.**

# Project Importance

This project on global air pollution was selected because it is an ongoing issue that affects everyone worldwide. According to the Health Effects Institute (2024), air pollution was the second largest risk factor of deaths in 2021 (p. 3). For instance, in 2021 air pollution was the cause of 8.1 million deaths worldwide, amounting to 1 in 8 deaths. Furthermore, 709,000 of the total deaths from air pollution were children under the age of five. In addition, air pollution was responsible for causing 28% of deaths from ischemic heart disease, 30% of deaths from lower respiratory infections, and 48% of deaths from chronic obstructive pulmonary disease (Health Effects Institute, 2024, p. 3). Although all countries worldwide are being negatively impacted by air pollution, it appears that countries in South Asia and Africa have the highest health impacts (Health Effects Institute, 2024). Therefore, it is essential that air pollution trends are evaluated at a local and global level, in order to implement plans for mitigation that can decrease the number of premature deaths.

Although short-term exposure to air pollution has a negative impact on health, long-term exposure has the biggest impact and is more likely to cause health issues and premature deaths (Health Effects Institute, 2024). Although there are already air quality legislatures, policies, and guidelines in place, there is evidence that air pollution is still a rising issue. According to the Health Effects Institute (2024), air pollution caused more deaths in 2021 than any other year, dating back to 1990 (p. 14). Children, the elderly, and people with preexisting health conditions are more susceptible to the effects of air pollution. As air pollutants are inhaled, they can be deposited into the lungs, altering lung defenses, and some “enter directly into the bloodstream and deeper tissues, including the heart, brain and other organs (Health Effects Institute, 2024, p. 20). Nonetheless, as restrictions are placed on activities or emissions to reduce air pollution in some countries, other countries lack the appropriate and effective air quality restrictions (Health Effects Institute, 2024). Therefore, it is essential to mankind that locally and globally, the most efficient and effective policy decisions are implemented that can decrease emissions of pollution that can significantly improve future air quality and population exposure (Amann et al., 2020).

# Background

Air pollution is the release of harmful chemicals or physical or biological particles in the indoor and outdoor atmosphere, which can cause harm to the health of humans, animals, plants, and the environment in general (National Geographic Society, n.d.; World Health Organization, 2025). There are different sources of air pollution, and air pollutants can be in the form of gases, solid particles or liquid droplets (National Geographic Society, n.d). The origin of air pollution is derived both from natural sources and “anthropogenic sources” (human-made sources). Pollutants emitted from wildfire smoke, volcano ash, and windblown sand or dust are examples of natural sources. On the other hand, the burning of fossil fuels that are associated with vehicles, airplanes, power plants, and factories, and cigarette/e-cigarette smoke are examples of anthropogenic sources (National Geographic Society, n.d, para. 3). Asian and African countries’ important source of exposure to air pollution is due to continuing the household use of solid fuels for cooking, which was already discontinued in other countries (Health Effects Institute, 2024). Nonetheless, there is a need to improve air quality and reduce greenhouse gas emissions, as rising temperatures are making air pollution and its health effects worse (Health Effects Institute, 2024, p. 29).

According to the World Health Organization (WHO) (2025), 99% of the global population is exposed to air that consists of high levels of pollutants and exceeds WHO guideline limits (para. 3). The WHO guidelines (see table 1 in the appendix) are air quality guidelines (AQG) that should be used to reduce human exposure to air pollution to minimize adverse health effects, and when the AQG are exceeded, it presents a major public health risk. Furthermore, the purpose of the WHO guidelines is to serve as a guidance for government air quality management and mitigation interventions by adapting legislation and policies that can reduce air pollution (WHO, 2021). Pollutants such as, particulate matter, carbon monoxide, ozone, nitrogen dioxide and sulfur dioxide are of major public health concern (WHO, 2025, para. 2). However, particulate matter (PM2.5) has been associated with the most harmful health impacts and is present in harmful levels around the world (Amann et al. 2020; Health Effects Institute, 2024). Nonetheless, all pollutants contribute to the overall air quality.

***Particular Matter (PM2.5)***

Particular matter, which is particle pollution, is a mixture of macroscopic and microscopic solid particles and liquid droplets found in the air, such as dust, dirt, soot, or smoke that can be 10 micrometers and smaller (PM10) or 2.5 micrometers and smaller (PM2.5). Particular matter particles can be made up of hundreds of different chemicals and they directly originate from construction sites, unpaved roads, fields, smokestacks or fires (EPA, 2024). PM2.5 is associated with the most harmful health impacts because these fine particles are easily inhaled and travels deep into the lungs and the blood stream, and it is a carcinogen for lung cancer with longtime exposure (EPA, 2024; WHO 2025). PM2.5 is the main pollutant that is used as an indicator for assessing the health effects of air pollution exposure (WHO, 2025).

***Carbon Monoxide (CO)***

Carbon monoxide (CO) is a colorless and odorless gas that is released when something is burned, such as wood, petrol, coal, natural gas, kerosene, or from the emissions of motor vehicles. Motor vehicles or machinery that burn fossil fuels is the source of outdoor CO, while space heater, chimneys, furnaces, and gas stoves is the source of indoor CO. CO can be deadly when exposed to high levels indoor since it causes a lack of oxygen in the body, making it difficult to breath and causing damage to tissues and cells (EPA, 2024; WHO 2025).

***Ozone (O3)***

There are two types of ozone’s; The good naturally occurring one located in the upper atmosphere is called stratospheric ozone. The ground level ozone (O3), which is called the tropospheric ozone, is formed by chemical reactions between oxides of nitrogen (Nox) and volatile organic compounds (VOC) that are emitted by cars, power plants, industrial boilers, refineries, chemical plants, and other sources chemically react in the presence of sunlight (EPA, 2024). Excessive exposure to O3 can cause respiratory problems (WHO, 2025).

***Nitrogen Dioxide (NO2)***

Nitrogen Dioxide (NO2) is an indicator for the larger groups of nitrogen oxides, which are highly reactive gases known as oxides of nitrogen or nitrogen oxides (NOx). NO2 is emitted into the air from the burning of fuel from heating, motor vehicles, and power plants. Exposure to NO2 causes respiratory issues (EPA, 2024, WHO, 2025).

***Sulfur Dioxide (SO2)***

Sulfur Dioxide (SO2) is a colorless, water-soluble gas that is emitted from fossil

fuel combustion from domestic heating, industries, and power generation. SO2 exposure

worsens asthma symptoms requiring emergent medical treatment (WHO, 2025).

# Data Set Description

The proposed data set is the “Global Air Pollution Dataset” from the website Kaggle. The data set consists of 23,463 rows of data and 12 columns, and contains both numerical and categorical data. The author of the data set is Hasib Al Muzdadid. The temporal coverage start date for the data set was 11/2/2022. The data set was last updated in 2023 (Muzdadid, 2022). The data set consists of U.S. Air Quality Index (AQI) values for different pollutants for many cities world-wide (AirNow, n.d.; Muzdadid, 2022).

The AQI values are for the pollutants, particulate matter (PM2.5), carbon monoxide (CO), ozone (O3), nitrogen dioxide (NO2), and the overall AQI (Muzdadid, 2022). The EPA uses the AQI tool for communication about the outdoor air quality and health risks. It is a tool for public health that alerts the public about the current air quality and the associated health risk. The AQI includes 6 ranges of index values, which corresponds to six color-coded health risk categories. Smaller AQI values are indicative of lower levels of air pollution, while larger AQI values are indicative of higher levels of air pollution. AQI values 100 and below are considered to be satisfactory, while AQI values above 100 are considered to be unhealthy, and hazardous above 300 (see table 2 in the appendix) (AirNow, n.d.).

# Data Analytics Tools

The data analytics tools that I plan to use for this project, include Excel, Python in Google Colab, and Tableau. The “Global Air Pollution Dataset” was downloaded as an Excel comma separated value (CSV) file. Excel is a spreadsheet-based software that can be used to store, organize, and manipulate tabular data. The benefit of using a CSV file is it is compatible with many other data analytics software (French, 2020; Wikipedia, 2024). Excel will be used for part of the data exploration process. The Excel dataset file will then be uploaded into Google Colab and Tableau for further data exploration.

Google Colab is a free cloud-based environment that functions as an as-a-service version of a Jupyter Notebook, which enables the manipulation of data by writing and executing Python code (Burke, 2023). Python in Google Colab will be used to complete more data exploration, data preprocessing and cleaning, and data analysis. It will be used to determine if there is missing data, duplicates, or outliers. Python will also be used to obtain information about the data such as, data types, data shape, data distribution, data visualization, descriptive statistics, machine learning, and predictive modeling. Tableau will also be used for data exploration, data preprocessing, and data visualization. Tableau is software that can be used to easily explore and manipulate data from a large range of data sources and creating data visualizations that can be shared. Tableau is able to create charts, graphs, and interactive visualization and dashboards to analyze data (GeeksforGeeks, 2024).

# Project Milestones

The major milestones that will be performed for this project includes the following:

* Selecting the project idea
* Data review and acquisition
* Defining the Project Scope
* Data profiling and preparation through exploratory data analysis
* Creating Data visualizations
* Data modeling and model testing and validation
* Discussion of final results

# Completion History

**Assignments 1 - 5**

|  |  |
| --- | --- |
| **Weeks 1 & 2** | Selected project idea, reviewed and acquired data set, researched data set, and defined project scope. |
| **Weeks 3 & 4** | Begin exploratory data analysis, summarized data, created tables outlining data profiles and data definitions for each variable, and described the processes and tools used to prepare and clean the data. Described descriptive statistics in detail, described visualization techniques used, and described visualizations in detail and insights from visualizations. |
| **Weeks 5 & 6** | Created, trained, tested, and analyzed a multiple linear regression model and 4 classification models. Identified and defined the modeling techniques used. Presented and compared the results of all 5 models and identified the best 2 performing models. |
| **Weeks 7 & 8** | Discussed the project findings, reviewed and discussed project success or completion and lessons learned, discussed potential data privacy and data security issues, provided recommendations for future analysis, and summarized the key points of the project in the executive summary. |

# Lessons Learned

**Assignments 1 - 5**

|  |  |
| --- | --- |
| **Assignment 1** | I learned that it is essential to define the project scope to ensure that the outcome of the project is successful. The problem needs to be defined and the appropriate data must be acquired and of quality. It is important to do the necessary research to gain a thorough understanding of the problem and the data, and how the data can be analyzed to provide meaningful insights. The project scope clearly outlines the importance of the problem and what processes, tools, and resources are required to complete the project. |
| **Assignment 2** | I learned more about the data set and the method for how it was collected. I learned through exploratory data analysis that the dataset has some missing data, which will need to be addressed to ensure the analysis results will not be skewed. I also learned that there is a class imbalance between the distribution of health risk classes; This will also need to be addressed. |
| **Assignment 3** | I learned that a thorough analysis of the descriptive statistics, in conjunction with visualization techniques can uncover deeper insights about the data. I learned that there is a disparity in the amount of data points between countries, which could lead to bias in the classification model. I learned what would be the best input features and target variable to use for the machine learning models. |
| **Assignment 4** | I learned that all models are not equal and all models have different challenges and limitations. I also learned that using a technique to balance classes for global pollution data can lead to a model making unrealistic predictions. A high accuracy does not necessarily mean that the model is making realistic or meaningful predictions. |
| **Assignment 5** | I learned that due to data limitations for the Global Air Pollution Dataset, the models created for this project may not fully capture the true aspects of PM2.5 air pollution for some countries or regions and with the limited amount of data available, any comparisons made between countries may be inaccurate. This project revealed that there is insufficient air pollution data available for some countries that is not just an issue with this dataset, but is an issue pertaining to all global air pollution data due to the lack of high-quality air pollution monitoring systems or air pollution sensors and/or faulty low-cost air pollution sensors, and also due to some governments not making their data public. To provide more usable insights for policy makers, it may be necessary to incorporate hourly analysis, emission sources, and population data. To provide better performing models, it may be best to use model ensembling. |

# Data Profiling and Preparation

### Data Summary

The “Global Air Pollution Dataset” CSV file was obtained from the website Kaggle. The data set consists of U.S. Air Quality Index (AQI) values for different pollutants for many cities world-wide (AirNow, n.d.; Muzdadid, 2022). The data set is important for the evaluation and impact of global air pollution on public health. The data was collected from elichens.com, which is the website of the company eLichens that manufactures different pollutant sensors for air quality monitoring and also provides real-time global air quality data. eLichens obtains air quality data through the “real-time measurements from eLos Air Quality Stations and Local Air Quality Agencies” (eLichens, 2024, Key Features). The author or collaborator of Kaggle’s “Global Air Pollution Dataset” is Hasib Al Muzdadid. Muzdadid collected the data using web scrapping and hand engineered preprocessing (Muzdadid, 2022).

The temporal coverage start date for the “Global Air Pollution Dataset” was 11/2/2022, but the data set was last updated on Kaggle in 2023 (Muzdadid, 2022). The data set consists of 12 columns and 23,463 distinct rows of data. The data includes 5 continuous (numerical) fields and 7 categorical fields (see Table 1.1 and 2 below for details). There are some data quality issues, which involve missing data and one questionable outlier. There are 427 rows of data that are missing the name of the country and there is 1 row of data that is missing the name of the city. Although there is data missing from the Country and City columns, there is not any data missing from the other 10 columns.

On the other hand, there is a class imbalance between the distribution of health risk classes, as the AQI Category, CO AQI Category, Ozone AQI Category, NO2 AQI Category, and the PM2.5 AQI Category do not all have data categorized for all six health risk categories, only the AQI Category and the PM2.5 AQI Category have data representation across all six health risk categories (see Table 3 below for details). The Ozone AQI Category has data representation across all health risk categories, except the “Hazardous” class. The CO AQI Category has data representation across only the “Good”, “Moderate”, and “Unhealthy for Sensitive Groups” classes. The NO2 AQI Category has data representation across only the “Good” and “Moderate” classes. Nevertheless, data quality issues will need to be addressed during the data preparation and preprocessing steps. Once the data quality issues have been addressed during data preparation and preprocessing, the data will be divided into an 80\20 split before model training. 80% of the data will be used for model training and the other 20% will be used for model testing and validation.

### Data Definition/Data Profile

**Table 1.1**

**Examination of Continuous (Numerical) Fields:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name of field** | **Brief description** | **Minimum value** | **Maximum value** | **Average** | **Standard**  **deviation** | **Potential Quality**  **Issue?** | **Nulls** | **Outliers** |
| AQI Value | The overall level of air pollution. | 6.0 | 500.0 | 72.01 | 56.06 | No | 0 | 0 |
| CO AQI Value | The overall level of Carbon Monoxide pollution in the air. | 0.0 | 133.0 | 1.37 | 1.83 | Yes | 0 | 1 |
| Ozone AQI Value | The overall level of Ozone pollution in the air. | 0.0 | 235.0 | 35.19 | 28.10 | No | 0 | 0 |
| NO2 AQI Value | The overall level of Nitrogen Dioxide pollution in the air. | 0.0 | 91.0 | 3.06 | 5.25 | No | 0 | 0 |
| PM2.5 AQI Value | The overall pollution level of Particulate Matter with a diameter of 2.5 micrometers or less in the air. | 0.0 | 500.0 | 68.52 | 54.80 | No | 0 | 0 |

**Table 2**

**Examination of Categorical Fields:**

| **Name of field** | **Brief description** | **Number**  **of categories** | **Potential Quality Issue?** | **Nulls** | **Outliers** |
| --- | --- | --- | --- | --- | --- |
| Country | The name of the country. | **Total:** 23036  **Unique:** 175 | Yes | 427 | 0 |
| City | The name of the city. | 23462 | Yes | 1 | 0 |
| AQI Category | **Good**= 0-50; satisfactory air quality, which posed little to no risk.  **Moderate**= 51-100; acceptable air quality that may pose some risk to very sensitive groups.  **Unhealthy for Sensitive groups**= 101-150; Impacts the health of only sensitive groups.  **Unhealthy**= 151-200; May impact the health of the general public, but may severely impact the health of sensitive groups.  **Very Unhealthy**= 201-300; Health Alert: Increased risk for health impact for everyone.  **Hazardous**= 301 and higher; Health warning of emergency condition: everyone more likely to have health impact. | 6 | No | 0 |  |
| CO AQI Category | **Good**= 0-50; satisfactory air quality, which posed little to no risk.  **Moderate**= 51-100; acceptable air quality that may pose some risk to very sensitive groups.  **Unhealthy for Sensitive groups**= 101-150; Impacts the health of only sensitive groups. | 3 | No | 0 | 0 |
| Ozone AQI Category | **Good**= 0-50; satisfactory air quality, which posed little to no risk.  **Moderate**= 51-100; acceptable air quality that may pose some risk to very sensitive groups.  **Unhealthy for Sensitive groups**= 101-150; Impacts the health of only sensitive groups.  **Unhealthy**= 151-200; May impact the health of the general public, but may severely impact the health of sensitive groups.  **Very Unhealthy**= 201-300; Health Alert: Increased risk for health impact for everyone. | 5 | No | 0 | 0 |
| NO2 AQI Category | **Good**= 0-50; satisfactory air quality, which posed little to no risk.  **Moderate**= 51-100; acceptable air quality that may pose some risk to very sensitive groups. | 2 | No | 0 | 0 |
| PM2.5 AQI Category | **Good**= 0-50; satisfactory air quality, which posed little to no risk.  **Moderate**= 51-100; acceptable air quality that may pose some risk to very sensitive groups.  **Unhealthy for Sensitive groups**= 101-150; Impacts the health of only sensitive groups.  **Unhealthy**= 151-200; May impact the health of the general public, but may severely impact the health of sensitive groups.  **Very Unhealthy**= 201-300; Health Alert: Increased risk for health impact for everyone.  **Hazardous**= 301 and higher; Health warning of emergency condition: everyone more likely to have health impact. | 6 | No | 0 | 0 |

**Table 3**

**Counts of Air Quality Health Risk Categories:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Name of field** | **Good** | **Moderate** | **Unhealthy for Sensitive Groups** | **Unhealthy** | **Very Unhealthy** | **Hazardous** |
| AQI Category | 9688 | 9087 | 1568 | 2215 | 286 | 191 |
| CO AQI Category | 23032 | 2 | 1 | 0 | 0 | 0 |
| Ozone AQI Category | 20672 | 1419 | 488 | 404 | 52 | 0 |
| NO2 AQI Category | 23020 | 15 | 0 | 0 | 0 | 0 |
| PM2.5 AQI Category | 9950 | 8939 | 1601 | 2118 | 255 | 172 |

### Data Preparation

There will be different tools and techniques utilized to process, prepare, and clean the data. The “Global Air Pollution Dataset” CSV file will be uploaded into Google Colab, where a Jupyter Notebook will be used to execute Python code to process, prepare, and clean the data. Using Python’s Pandas library, the data set will be imported and read as a Pandas data frame, which is a spreadsheet like format, which enables the easy manipulation of data. Python’s Numpy library will be used to help manage numerical calculations. To fully explore the contents and integrity of the dataset, Python code will be written to view the shape of the data, the first five rows of the data, the last five rows of the data, and five random rows of the data. Python code will also be written to view information pertaining to the data types of each column, along with the non-null counts (count of valid entries), to determine if there are null or missing data, and to obtain the sum of any missing values grouped by column. Furthermore, Python code will also be written to calculate and display the descriptive statistics of the dataset.

As shown in Table 2 above, there are 427 rows of data that are missing the name of the country and 1 row of data that is missing the name of the city. Python code will be written to drop all rows of data that contain missing data and then the shape of the data will be rechecked. After dropping all rows that contained missing data, there were now 12 columns and 23,035 distinct rows of data to be used for analysis. Once again, Python code was written to calculate and display the descriptive statistics of the dataset after dropping all rows that contained missing data, but the descriptive statistics did not change much from the initial results (see Table 1.2 below). Python code was written to determine the number of unique countries since there are duplicate country entries. To determine if the data set contains duplicate rows of data, Python code was written to check for duplicate cities, which there were none. Now I can further explore the data by creating visualizations using Python code.

Visualizations will also be created using Tableau. First, the “Global Air Pollution Dataset” CSV file will be uploaded into Tableau. There is no special data preparation required to analyze the data in Tableau. There is also no extra step involved in cleaning the data before creating visualizations. When creating visualizations, the rows of data that have missing data will be excluded by setting filters. The created visualizations will be discussed in the next section.

*.*

**Table 1.2**

**Examination of Continuous (Numerical) Fields after data cleaning:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name of field** | **Brief description** | **Minimum value** | **Maximum value** | **Average** | **Standard**  **deviation** | **Potential Quality**  **Issue?** | **Nulls** | **Outliers** |
| AQI Value | The overall level of air pollution. | 6.0 | 500.0 | 72.34 | 56.36 | No | 0 | 0 |
| CO AQI Value | The overall level of Carbon Monoxide pollution in the air. | 0.0 | 133.0 | 1.38 | 1.84 | Yes | 0 | 1 |
| Ozone AQI Value | The overall level of Ozone pollution in the air. | 0.0 | 235.0 | 35.23 | 28.24 | No | 0 | 0 |
| NO2 AQI Value | The overall level of Nitrogen Dioxide pollution in the air. | 0.0 | 91.0 | 3.08 | 5.28 | No | 0 | 0 |
| PM2.5 AQI Value | The overall pollution level of Particulate Matter with a diameter of 2.5 micrometers or less in the air. | 0.0 | 500.0 | 68.88 | 55.06 | No | 0 | 0 |

# Data Visualizations

### Descriptive Statistics

The descriptive statistics for the “Global Air Pollution Dataset” were recalculated after data cleaning (see Table 1.2 above). Overall, the descriptive statistics show the variability of the distribution of data for the AQI Value, the CO AQI Value, the Ozone AQI Value, the NO2 AQI Value, and the PM2.5 AQI Value. The AQI Values range from 6.0 to 500, with the average value of 72.34 (Moderate Health Risk Category Range), and a standard deviation of 56.36, which indicates a high range of variability in the values. The CO AQI Values range from 0.0 to 133, with the average value of 1.38 (Good Health Risk Category Range), and a standard deviation of 1.84, which indicates there is no significant variability in the values and that the CO levels are consistently low.

The Ozone AQI Values range from 0.0 to 235, with the average value of 35.23 (Good Health Risk Category Range), and a standard deviation of 28.24, which indicates a moderate range of variability in the values. The NO2 AQI Values range from 0.0 to 91, with the average value of 3.08 (Good Health Risk Category Range), and a standard deviation of 5.28, which indicates a slight variability in the values, but overall NO2 levels are usually low. The PM2.5 AQI Values range from 0.0 to 500, with the average value of 68.88 (Moderate Health Risk Category Range), and a standard deviation of 55.06, which indicates a high range of variability in the values. According to the descriptive statistics, the Good Health Risk Category has the highest frequency than any other category for all pollutants. The descriptive statistics also show that the United States is the top country with the highest frequency count (2,872) within the dataset.

### Data Visualization Definitions

There are different visualization techniques that can be utilized for exploratory data analysis. To aide in a thorough analysis of the “Global Air Pollution Dataset”, I will be using visualization techniques, such as box plots, count plots, bar plots, histograms, scatter plots, Seaborn’s (SNS) pair plot, heat maps, and tree maps. Displaying data as a box plot helps to quickly capture the distribution of data and aides in identifying outliers, which can skew and negatively impact the data analysis (Reliance Jio, 2024). A box plot displays a five number summary for the data including, the minimum value, lower quartile, median value, upper quartile, and the maximum value (see Figure 1 in the appendix). Any data points below the minimum value line or above the maximum value line are considered to be outliers since the minimum and maximums represents the lower and upper boundaries of standard deviation (Agrawal, 2024; GeeksforGeeks, 2024, Reliance Jio, 2024). Count plots are used to show the frequency of specific elements in a category in the form of a bar graph (Agrawal, 2024).

Bar plots are used to display the relationship between categorical and numerical variables in the form of a bar. Bar plots make it easy to compare different categories or elements of data within the same category (Singla, 2024). Histograms are used to display the numerical distribution of values of one variable in columns that are binned for specific ranges (Agrawal, 2024). Histograms can also detect outliers and abnormal data distribution (Singla, 2024). Scatter plots are used to show the relationship between two numerical variables (Agrawal, 2024). Scatter plots can determine whether 2 variables have positive correlation, negative correlation, or no correlation. A SNS pair plot (see Figure 5 in the appendix) is a comprehensive visualization of multiple variables, which consists of a matrix of scatter plots and histograms and shows the relationship between each pair of variables (Abluwalia, 2024). A SNS pair plot can be made using Seaborn, which is a library that can be used to make statistical visualizations using Python code (Seaborn, n.d.).

Correlation heat maps are used to determine the numerical strength of correlation between variables before selecting input features for a model. It is best practice to select input features that have strong correlations with the dependent or target variable. However, by all means avoid selecting input features that strongly correlate with one another because this could negatively affect model performance. Correlation values range from -1 to 1; -1 represents no correlation, while 1 represents perfect correlation (Spencer, 2022). In addition, the intensity of the correlation is also represented by a color gradient. Not only will correlation heatmaps be used, but also a regular heatmap that represents the count of data elements in a tabular format and a geographical format.

Moreover, a tree map is a visualization that helps to illustrate a part-of-whole relationship between different categories in a hierarchal format. A tree map, which is composed of nested rectangles, is organized by category and by size. The rectangles are organized with the largest rectangle being placed in the top left corner and progresses down to the smallest in the bottom right. Furthermore, the largest part of the whole is represented by the largest rectangle and the smallest part of the whole is represented by the smallest rectangle. A tree map color scheme can be representative of categories or measures (Rowe, n.d.; Tableau, 2025). Nevertheless, although I have used multiple different visualization techniques, I will only be discussing a few below.

### Data Visualization 1

This visualization, which was created using Python code, shows a box plot of the numerical data for AQI Value, CO AQI Value, Ozone AQI Value, and PM2.5 AQI Value. The x-axis shows the continuous variable names. The y-axis shows the numerical range of the data, which is 0 to 500. Each box plot for each variable is represented by a colored rectangle, which the bottom line represents the lower quartile (Q1), the middle line represents the median value (Q2), and the upper line represents the upper quartile (Q3) (GeeksforGeeks, 2024). The bottom line of the whisker, which is the line that is perpendicular to the line connected to the Q1 line, is the minimum value. The top line of the whisker, which is the line that is perpendicular to the line connected to the Q3 line, is the maximum value.

Looking at the box plot for the AQI Value, it shows that 25% of the data is less than 39, 50% of the data is less than 55, and 75% of the data is less than 80. There are not any data points below the minimum value, but there are many data points above the maximum value, extending all the way up to 500. The box plot for the PM2.5 AQI Value is very similar to the AQI Value’s boxplot. Looking at the box plot for the PM2.5 AQI Value, it shows that 25% of the data is less than 35, 50% of the data is less than 54, and 75% of the data is less than 79. There also are not any data points below the minimum value, but there are many data points above the maximum value, extending all the way up to 500. Looking at the box plot for the Ozone AQI Value, it shows that 25% of the data is less than 21, 50% of the data is less than 31, and 75% of the data is less than 40. There are not any data points below the minimum value, but there are some data points above the maximum value, extending all the way up to 235. There is also a data point that is slightly higher and segregated from the other data points, which after comparing the actual value to the other data points, it is not an outlier.

Looking at the box plot for the CO AQI Value, since the values are much smaller, we cannot see the complete box plot. However, we can see that the majority of the data is less than 1. There are data points above the maximum value and a data point (CO AQI= 133) that is significantly higher and segregated from the other data points, which appears to be an outlier. Looking at the box plot for the NO2 AQI Value, since the values also are much smaller, we cannot see the complete box plot. However, we can see that the majority of the data is less than 4. There are data points above the maximum value and a data point that is slightly higher and segregated from the other data points, which after comparing the actual value to the other data points, it is not an outlier.

In my opinion, outliers on a box plot are data points that are significantly higher and segregated from the other data points as seen for the CO AQI Value. However, due to the nature of air pollution data, there could have been a major pollution event that contributed to the higher value that appears to be an outlier, such as a wild fire (AQI, 2024; Singh, 2022). The specific data point that appeared to be an outlier for the CO AQI Value was for the City Durango (assumed to be in Colorado) in the USA. Durango also had an AQI and PM2.5 AQI value of 500; therefore, the CO AQI Value is most likely not an outlier.

The box plots in general, visually displays the descriptive statistics. Overall, the distribution of data is different between the variables, except for the AQI Value and PM2.5 AQI Value which have very similar data distributions. The CO AQI and the NO2 AQI values are on a much smaller scale and do not contribute much to the overall air pollution, while the Ozone AQI Value does make a contribution to the overall air pollution. Nonetheless, including the CO AQI Value and the NO2 AQI value could lead to bias in the model or inaccurate predictions. Although the box plot visualization provided important insights into the distribution of the data and the identification and investigation of possible outliers, it is crucial to create and analyze a multitude of visualizations to get a complete overview of the data. For instance, by creating and analyzing histograms for each variable, I could see a better breakdown and frequency of the data ranges. I could see that there were more zero values for the CO AQI Value and the NO2 AQI Value (see Figures 2 and 3 in the appendix).

### Data Visualization 2

This visualization, which was created using Python code, shows a correlation heatmap of the selected variables, Country, CO AQI Value, Ozone AQI Value, AQI Value, PM2.5 AQI Value, and PM2.5 AQI Category. As you can see, the correlation heatmap displays the numerical correlation values between every variable in a tabular format. In addition to the numerical correlation values, the magnitude of correlation is also represented by a purple-white-green color scale. Dark purple to white represent zero correlation to approximately 0.50 correlation. Light green to dark green represents correlation from approximately 0.55 to 1.0. The legend on the right of the correlation heat map shows mapping of the color gradient to the numerical scale.

The correlation heat map for the selected data has all positive correlations that range between 0 and 1. The diagonal ones represent the correlation of each variable to itself, which explains why there is perfect correlation. Looking at the correlation heat map, we can see that the variable “Country” has an extremely weak correlation with all the other variables. The CO AQI Value has a weak correlation with the Ozone AQI Value. The CO AQI Value has an almost moderate amount of correlation with the AQI Value and the PM2.5 AQI Value. The CO AQI Value has fair amount of correlation with the PM2.5 AQI Category.

The Ozone AQI Value has an almost moderate amount of correlation with the AQI Value. The Ozone AQI Value has a fair amount of correlation with the PM2.5 AQI Value. The Ozone AQI Value has a low correlation with the PM2.5 AQI Category. The AQI Value has an extremely strong correlation with the PM2.5 AQI Value. The AQI Value has a more than moderate correlation with the PM2.5 AQI Category. The PM2.5 AQI Value has a more than moderate correlation with the PM2.5 AQI Category.

Based on my research for global air pollution, it was determined that the PM2.5 AQI Value would be the best dependent variable to use for a linear regression model analysis, and the PM2.5 AQI Category would be the best target variable for a machine learning classification model, since PM2.5 is associated with the most harmful health impacts and is the main pollutant that is used as an indicator for assessing the health effects of air pollution exposure. The NO2 AQI Value was omitted from this correlation heat map because this correlation heatmap is the second of two correlation heat maps that I created. The first correlation heat map (see Figure 4 in the appendix) was initially made for feature selection for the linear regression model and only included the CO AQI Value, the Ozone AQI Value, the NO2 AQI Value, the AQI Value, and the PM2.5 AQI Value. The first correlation heat map showed that the NO2 AQI Value had the weakest correlation (0.26) with the PM2.5 AQI Value compared to the other pollutants; therefore, it was omitted from the second correlation heat map that was created for feature selection for the classification model.

From the above correlation heat map, it was determined that the classification model will be trained using the input variables, Country, CO AQI Value, Ozone AQI Value, and AQI Value, with the target variable PM2.5 AQI Category. Although “Country” is weakly correlated with all other variables, it is necessary to include it so the model will be able to make classification predictions by country. Since the AQI Value and the PM2.5 AQI Value almost perfectly correlates with one another, it was best to select the AQI Value and omit the PM2.5 AQI Value, especially since the PM2.5 AQI Category will be the target variable.

### Data Visualization 3

This visualization, which was created using Tableau, shows a snap shot of the available data by country using a tree map. Each country is represented by a single rectangle. The size of the rectangles correlates with the count of data points for that specific country. The magnitude of the count is also represented by a blue color scale. The legend on the right of the tree map shows mapping of the color gradient to the numerical scale. The larger the data count, the larger the rectangle and the darker the blue color. On the other hand, the smaller the data count, the smaller the rectangle and the lighter the blue color. Since there are 175 unique countries, all countries could not be represented on this tree map. However, the tree map overall shows that the data count is not equal for all countries and varies and differs significantly, which could lead to bias in the machine learning models, especially the classification model.

### Data Modeling

### Data Modeling Definitions

A linear regression model and a classification model will be used to analyze the “Global Air Pollution Dataset” to predict or forecast future air pollution levels to determine its’ future effect on public health in different geographical regions. Linear regression models are the most basic and popular techniques used for understanding relationships between variables. A linear regression model involves the supervised learning (the use of labeled data to train a model) of a model, in which one continuous variable can be predicted based on the linear relationship between other input continuous features. The variable to be predicted is referred to as the dependent variable or the target variable (Y). The input variables or features are referred to as independent variables or the predictors (X). The line of best fit, which represents the linear relationship between the target variable and predictors, is used to make the predictions. A linear relationship is formed in a linear regression model, based on the assumption that changes in the predictors directly correspond to the proportional changes in the target variable. Furthermore, the line of best fit selected is the line that has the smallest amount of error (Mean Square error) or residuals, which is the difference between the independent data points and the predicted data points (Joy, 2025). The intercept for the model is the value at which the line of best fit crosses the y-axis. The coefficients for the model represent the slope (the steepness) of the line of best fit.

Furthermore, there are two types of linear regression models, simple and multiple. Only one predictor is used in simple linear regression and multiple predictors are used in multiple linear regression. A multiple linear regression model will be utilized for this project. To evaluate the performance of the linear regression model, the Mean Absolute Error (MAE), the Mean Squared Error (MSE), the Root Mean Squared Error (RMSE), and the R-squared (R²)will be calculated and evaluated**.** The MAE is the average absolute (positive) distance between the predicted and actual values, and a lower MAE is better. The MAE is less sensitive to outliers and does not penalize large errors. The MSE is the average squared difference between the predicted and actual values, and a lower MSE is better. The MSE is sensitive to outliers and heavily penalizes large errors. The RMSE is the average distance between a statistical model's predicted values and the actual values, and is the least possible error of the best fit line from the data points; the lower the RMSE, the better. The RMSE is also very sensitive to outliers and heavily penalizes large errors. The RMSE has the same units as the target variable, making it more interpretable. The R² is the variance of the dependent variable explained by the independent variables, and is the measure of accuracy of the model, as far as how well the model fits the data. The R² is evaluated from 0 to 1; the higher the R², the better (Joy, 2025).

In addition to the linear regression model, a classification model will also be used for this project. In order to find the optimal performing classification model, several types of classification models were created, trained, tested, and compared. The classification types modeled include, a Gradient Boosting decision tree model, a Random Forest decision tree model, and two Support Vector Machine (SVM) models. A decision tree model is a machine learning model that hierarchically iterates through data specific questions and partitions data in a logical manner that leads to different possible outcomes or solutions. The starting point of the decision tree is called the root node and it is where the main question is posed. The question is based on the features of the dataset. Based on a series of yes or no questions, the data is divided into smaller subsets based on specific attributes, which is represented by branches from the root node connecting to decision nodes, which have branches connecting to leaf nodes (see Figures 6 and 7 in the appendix). The leaf node shows the final outcome or decision. As more questions are asked, smaller subsets are created until all necessary questions have been asked and answered, leading to the final outcome or decision (GeeksforGeeks, 2025; machinelearningplus, 2025).

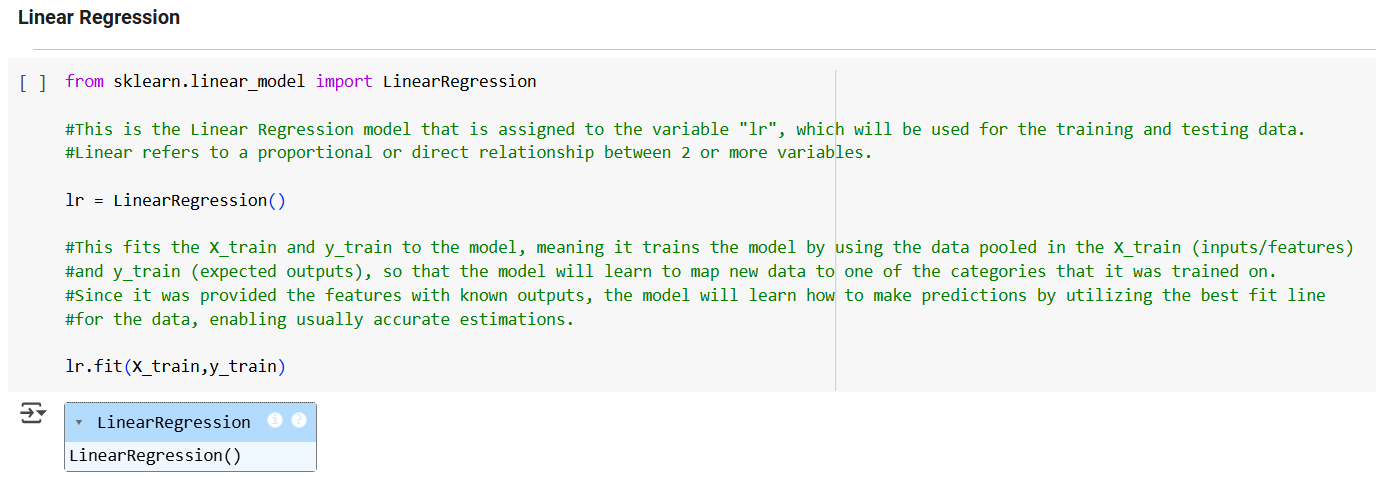
Nonetheless, overfitting is a major problem for basic decision tree models. Overfitting of data occurs when a model performs exceptionally well on the training data, but performs poorly to unseen data, which indicates that the model did not learn the trends in the data, but just simply memorized the data. A technique called “Gradient Boosting” can be utilized to help reduce overfitting. A Gradient Boosting decision tree model is a machine learning model that can be used for both classification and regression. A Gradient Boosting decision tree model uses ensemble learning, which combines many weak and slow learning models to form one strong learning model. Each weak model is trained sequentially and each model tries to minimize the error from the previous model. As each new model is trained, the predictions from the new model are added to the ensemble and this process repeats until a stopping criterion is reached. The result of each step is aggregated by the final model, leading to a strong learning decision tree model (GeeksforGeeks, 2025; machinelearningplus, 2025).

Another technique that can be utilized to help reduce the problem of overfitting of a decision tree model is called “Bagging”. A Random Forest decision tree model uses the ensemble learning application of bagging. A Random Forest decision tree model is a machine learning model that can be used for both classification and regression. A Random Forest decision tree model trains many different decision trees in parallel. Although training occurs in parallel, each tree is trained on a random subset of the same data and a random subset of features. The classification of the data is determined through a majority vote of predictions from the results from all trees, in which the final classification prediction is the category that was predicted by most trees (GeeksforGeeks, 2025; machinelearningplus, 2025).

A Support Vector Machine (SVM) model uses supervised machine learning for classification and regression, but is mainly used for classification. SVM models perform classification by finding the best boundary or hyperplane that separates data into different classes. Support vectors are the closest data points to the boundary (GeeksforGeeks, 2023). The boundary is chosen to maximize the margin, “which is the distance between the boundary and the closest data points from each class” (GeeksforGeeks, 2023, para. 4). By utilizing the technique called the “kernel trick”, SVM models can be used for both linear and non-linear classification. The data becomes linearly separable because the kernel trick maps the input data into a higher-dimensional space. Since SVM models only use a subset of data to make prediction, they are very efficient and are less prone to overfitting (GeeksforGeeks, 2023).

Different metrics will be used to evaluate the performance of the classification models. Accuracy, precision, recall, F1-score, and a confusion matrix will be used to evaluate the performance of the classification models. The accuracy is the percentage of correct predictions made by the model. The precision is the percentage of true positives predicted by the model, taking into account the true positive and false positive predictions. The recall is the percentage of true positives predicted by the model, taking into account the true positive and false negative predictions. The recall evaluates the model’s sensitivity or ability to detect positive samples. The F1-score is the balance between the precision and recall, and ranges between 0 and 1; a higher F1-score is indicative of better model performance. A confusion matrix displays the number of correct and incorrect predictions made by the model compared with the actual classifications in a tabular format. It clearly shows the true positive, false positive, true negative, and false negative predictions for each class or category (Goyal, 2021). The evaluation of the metrics obtained from the training data and testing data predictions must be compared to evaluate overfitting. A model accuracy of 100% or near perfect is indicative of overfitting (Kaplan, 2024). If the accuracy score for the training data is higher than the accuracy score for the testing data, the model has a degree of overfitting. A lower precision, recall, and F1-score for the testing data, in comparison to the training data, is also indicative of overfitting (btd, 2023).

### Data Model 1: Linear Regression Model



A correlation heatmap (see Data Visualization 2 above) was used to help complete feature selection for all classification models for analysis of the ‘Global Air Pollution Dataset”. After analyzing the correlation heatmap, three features and one target variable were selected to be used in the analysis. The features selected to be used as input variables were the CO AQI Value, the Ozone AQI Value, and the AQI Value. The PM2.5 AQI Value would be the best target variable to be used.

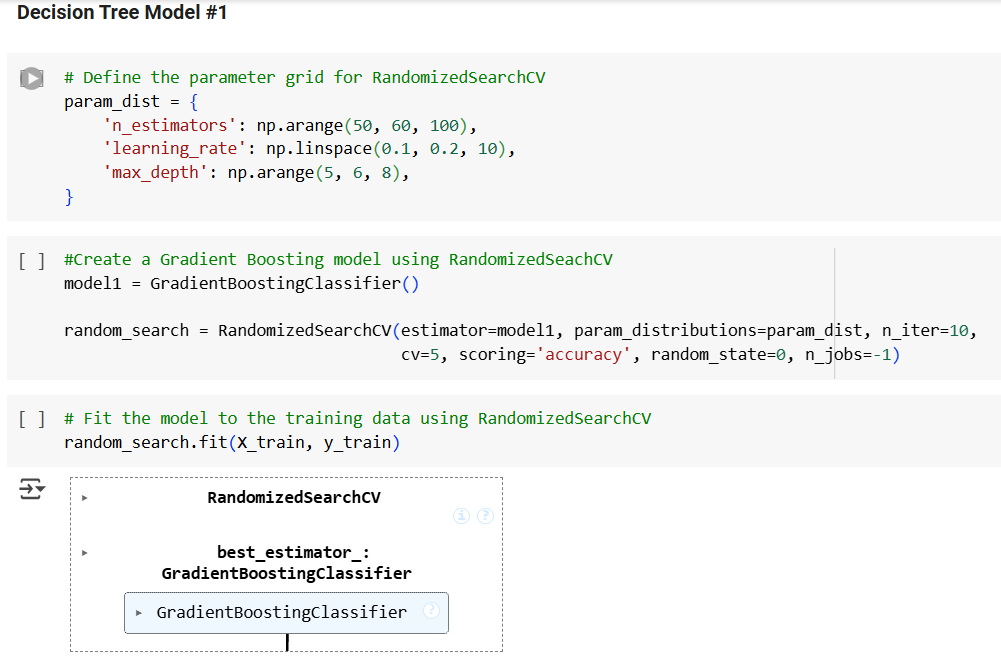
For model preparation, the CO AQI Value, the Ozone AQI Value, and the AQI Value columns were assigned to the variable “X”, and the PM2.5 AQI Value column was assigned to the variable “y” (see above). Then the data was split into train and test data for further model preparation. The train\_test\_split function was used to split the data into X and y, creating variables "X\_train" and "y\_train" for the training data pool, and creating variables "X\_test" and "y\_test" for the testing data pool; 80% of data was allocated to the training pool and the other 20% of the data was allocated to the testing pool. The random\_state was set to 0, which tells Sklearn not to use random selection so that we will always get the same results.

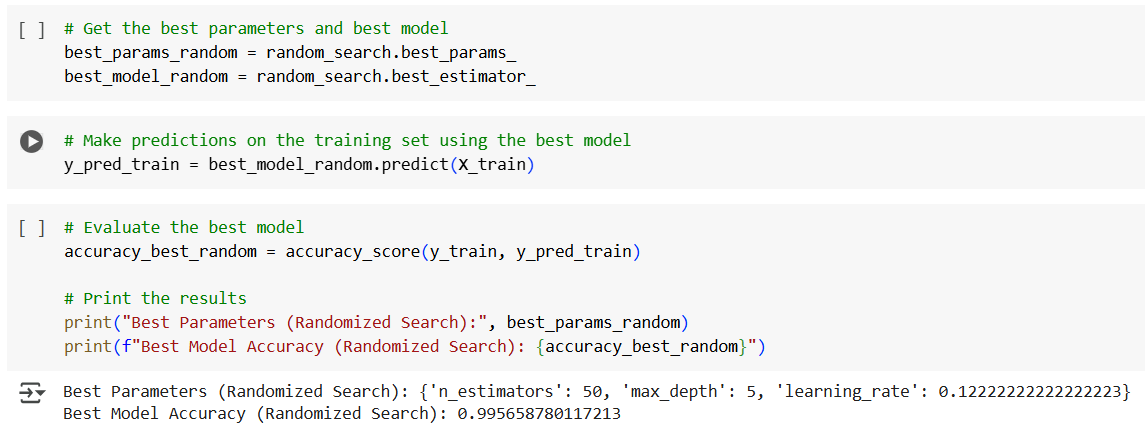
After splitting the data, the X\_train had 18,428 rows of data and 3 columns, the y\_train had 18,428 rows of data and 1 column, the X\_test had 4,607 rows of data and 3 columns, and the y\_test had 4,607 rows of data and 1 column. The “LinearRegression()” model was imported from the sklearn.linear\_model library, which uses “Ordinary least squares Linear Regression” (Sci-kit learn, n.d.). The “LinearRegression()” model was assigned to the variable "lr". Then the X\_train and y\_train data was fitted to the model, meaning the model was trained using the data pooled in the X\_train (inputs/features) and y\_train (expected outputs), so that the model would learn to map new data to one of the categories that it was trained on. Since the model was provided the features with known outputs, the model learned how to make predictions by utilizing the best fit line for the data.

Next, the trained model was used to predict the PM2.5 AQI Value for the test values by using X\_test data inputs/features. The predicted values were plotted against the actual values of the target (y\_test) for comparison to evaluate the model’s performance (see Figure 8 in the appendix). The plot shows that the predicted PM2.5 AQI Values and the actual PM2.5 AQI Values of the target, overall have a strong positive correlation indicating that the model’s performance is good. However, when the PM2.5 AQI Values increase above 300, the correlation is weaker indicating that the model’s predictions for PM2.5 AQI Values above 300 are less accurate than PM2.5 AQI Values below 300.

The intercept of the model is 1.82. The model’s coefficients for the CO AQI Value, the Ozone AQI Value, and the AQI Value are respectively, 0.39, -0.13, and 0.98, which shows that the AQI Value has the strongest positive impact on the predicted PM2.5 AQI Values. The CO AQI Value has a weaker positive impact and the Ozone AQI Value has a very weak negative impact on the predicted PM2.5 AQI Values. The MAE was 4.98, the MSE was 83.90, the RMSE was 9.16, and the R² was 0.98, which shows that the linear regression model has an overall great performance.

### Data Model 2: Decision Tree Model # 1





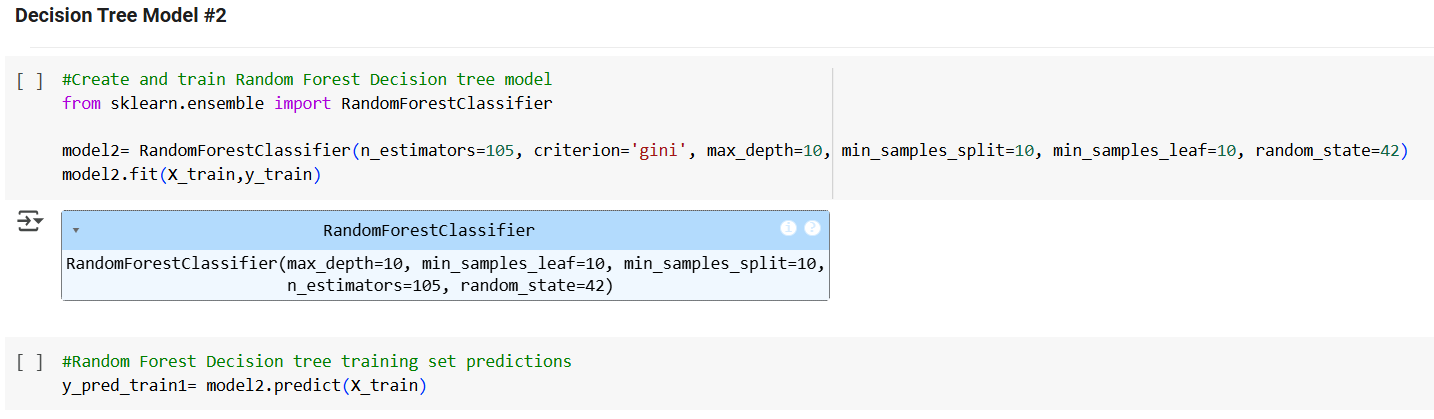
The GradientBoostingClassifier decision tree model was trained using the input features, Country, CO AQI Value, Ozone AQI Value, and AQI Value, with the target variable PM2.5 AQI Category. The PM2.5 AQI Category classes are imbalanced (see Figure 10 in the appendix), which could affect the model’s performance. Since the Country and PM2.5 AQI Category columns consisted of categorical data, the columns were encoded into numerical data using the LabelEncoder function from the sklearn.preprocessing library. The encoding mapped 0 to Good, 1 to Moderate, 2 to Unhealthy for Sensitive Groups, 3 to Unhealthy, 4 to Very Unhealthy, and 5 to Hazardous. Once this preprocessing step was completed, the input features Country, CO AQI Value, Ozone AQI Value, and AQI Value columns were assigned to the variable “X”, and the PM2.5 AQI Category column was assigned to the variable “y”. Then the data was split into train and test data for further model preparation. The train\_test\_split function was used to split the data into X and y, creating variables "X\_train" and "y\_train" for the training data pool, and creating variables "X\_test" and "y\_test" for the testing data pool. 80% of the data was allocated to the training data and 20% of the data was allocated to the testing pool, with the random\_state set to 0. After splitting the data, the X\_train had 18,428 rows of data and 4 columns, the y\_train had 18,428 rows of data and 1 column, the X\_test had 4,607 rows of data and 4 columns, and the y\_test had 4,607 rows of data and 1 column.

The GradientBoostingClassifier model was created using RandomizedSeachCV from the sklearn library. The parameter grid defined to be used by the RandomizedSearchCV include, n\_estimators= 50, 60, 100, learning\_rate= 0.1, 0.2, 10, and max\_depth= 5, 6, 8. n\_estimators designates the number of trees used. learning\_rate controls the step-size taken by each new tree, which shrinks the contribution of each tree according to each specific number. max\_depth is the number of layers the tree will go down before it stops (GeeksforGeeks, 2024; Scikit learn, n.d.). The training data was fitted to the GradientBoostingClassifier by using RandomizedSearchCV for 5-fold cross validation to find the best combinations of parameters and models. The RandomizedSearchCV randomly samples different combinations of hyperparameters from the parameter grid iterating 10 times through each n\_estimator, learning rate, and max-depth. Since the model runs all computations in parallel at the same time, this particular model has slow performance when fitting the data to the model.

The RandomizedSearchCV determined that the best model with the highest accuracy consisted of n\_estimators = 50, max\_depth= 5, and learning\_rate= 0.12 as hyperparameters with the overall accuracy of 100%. The confusion matrix and classification reports for both the training and testing data were compared (see Figure 11 in the appendix). Looking at the confusion matrix for the training data, the model classified 99.7% (rounded by classification report (CR) as 100%) of “0” predictions correctly, correctly classifying 7,944 as “0” and incorrectly classifying 22 as “2”. The model classified 100% of “1” predictions correctly. The model classified 99.6% (rounded by CR as 100%) of “2” predictions correctly, correctly classifying 7,103 as “2” and incorrectly classifying 18 as “0” and 12 as “4”. The model classified 99% of “3” predictions correctly, correctly classifying 1,729 as “3” and incorrectly classifying 12 as “4”. The model classified 99% of “4” predictions correctly, correctly classifying 1,240 as “4” and incorrectly classifying 11 as “2” and 5 as “3”. The model classified 100% of “5” predictions correctly. The overall accuracy, precision, recall, and F1-score were 100%.

Looking at the confusion matrix for the testing data, the model classified 99% of “0” predictions correctly, correctly classifying 1,976 as “0” and incorrectly classifying 14 as “2”. The model classified 95% of “1” predictions correctly, correctly classifying 37 as “1” and incorrectly classifying 2 as “5”. The model classified 99% of “2” predictions correctly, correctly classifying 1,776 as “2” and incorrectly classifying 12 as “0”, 1 as “3”, and 5 as “4”. The model classified 98% of “3” predictions correctly, correctly classifying 377 as “3” and incorrectly classifying 7 as “4” and 1 as “5”. The model classified 94% of “4” predictions correctly, correctly classifying 325 as “4” and incorrectly classifying 13 as “2” and 6 as “3”. The model classified 93% of “5” predictions correctly, correctly classifying 51 as “5” and incorrectly classifying 4 as “1”. The overall accuracy was 99%. The overall precision, recall, and F1-score were 96%. The overall evaluation for both the training data and testing data shows that the model is very overfitted, especially since the training data had 100% scores for all metrics.

### Data Model 3: Decision Tree Model #2

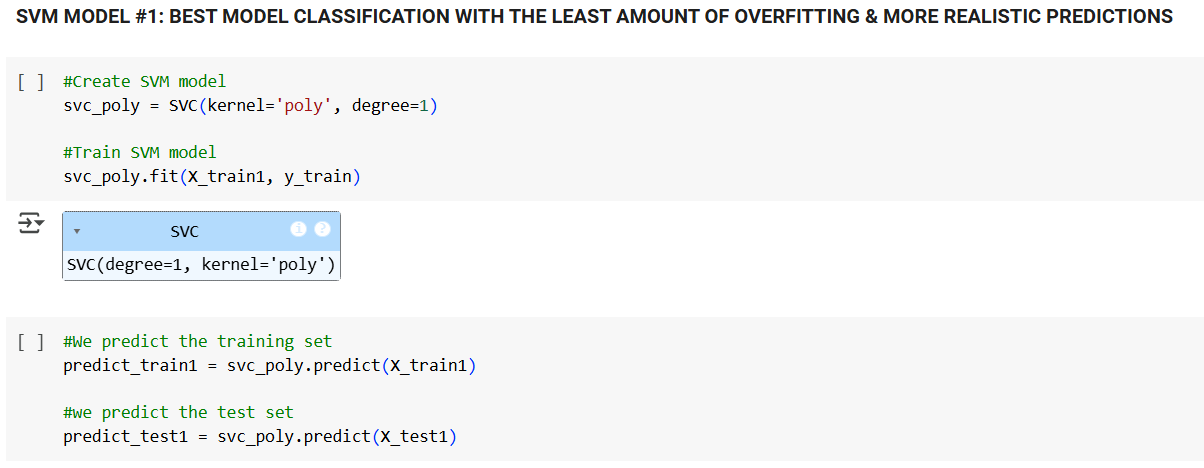


For the RandomForestClassifier decision tree model, the same input features and target variable were used as for the GradientBoostingClassifier model and the same preprocessed split data was used. The training data was fitted to the RandomForestClassifier model by using the criterion “gini”, which splits the nodes systematically to yield the least amount of impurity based on the probability distributions of the class target attribute’s values (GeeksforGeeks, 2024). n\_estimators=105 is the number of trees used. max\_depth=10 which means the tree will only go down 10 layers before it stops. min\_samples\_split= 10 means that an internal node will not be split unless it contains the minimum of 10 samples. min\_samples\_leaf=10 means that each leaf node requires a minimum of 10 samples remaining on each of the left and right branches after a split. random\_state=42 controls the random selection of the training data (Skikit-learn, n.d.).

The confusion matrix and classification reports for both the training and testing data were compared (see Figure 12 in the appendix). Looking at the confusion matrix for the training data, the model classified 99% of “0” predictions correctly, correctly classifying 7,921 as “0” and incorrectly classifying 41 as”2”. The model classified 98% of “1” predictions correctly, correctly classifying 127 as “1” and incorrectly classifying 2 as “5”. The model classified 99% of “2” predictions correctly, correctly classifying 7,047 as “2” and incorrectly classifying 40 as “0” and 11 as “4”. The model classified 98% of “3” predictions correctly, correctly classifying 1,721 as “3” and incorrectly classifying 42 as “4” and 2 as “5”. The model classified 95% of “4” predictions correctly, correctly classifying 1,211 as “4” and incorrectly classifying 1 as “0”, 48 as “2”, and 11 as “3”. The model classified 97% of “5” predictions correctly, correctly classifying 197 as “5” and incorrectly classifying 4 as “1” and 2 as “3”. The overall accuracy was 99%. The overall precision, recall, and F1-score were 98%.

Looking at the confusion matrix for the testing data, the model classified 99% of “0” predictions correctly, correctly classifying 1,965 as “0” and incorrectly classifying 16 as “2”. The model classified 95% if “1” predictions correctly, correctly classifying 36 as “1” and incorrectly classifying 2 as “5”. The model classified 99% of “2” predictions correctly, correctly classifying 1,770 as “2” and incorrectly classifying 23 as “0” and 2 as “4”. The model classified 98% of “3” predictions correctly, correctly classifying 378 as “3” and incorrectly classifying 1 as “2” and 5 as “4”. The model classified 94% of “4” predictions correctly, correctly classifying 330 as “4” and incorrectly classifying 16 as “2” and 5 as “3”. The model classified 90% of “5” predictions correctly, correctly classifying 52 as “5” and incorrectly classifying 5 as “1” and 1 as “3”. The overall accuracy was 98%. The overall precision, recall, and F1-score were 96%. Since the training data has a close to perfect accuracy and the accuracy and the other metrics for the testing data is slightly lower than the accuracy and other metrics for the training data, the RandomForestClassifier model is overfitted, but not as much as the GradientBoostingClassifier model.

### Data Model 4: SVM Model #1

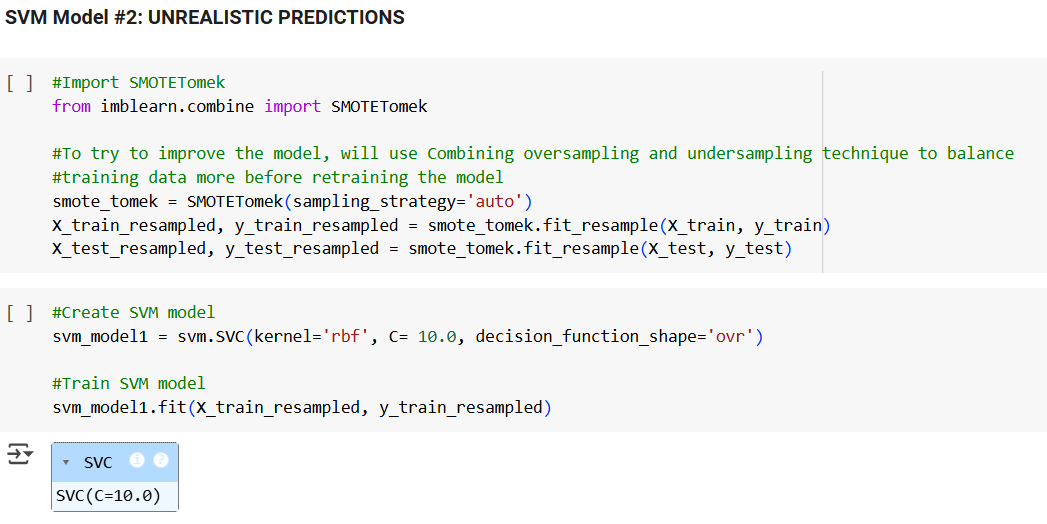
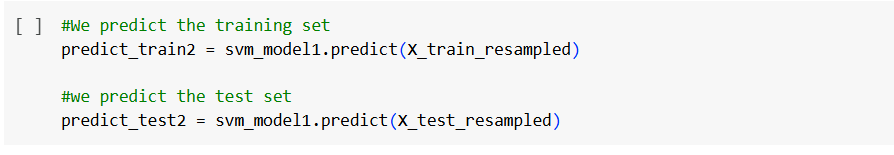


For the SVM Model #1, the same input features and target variable were used as for the RandomForestClassifier and GradientBoostingClassifier models. The same preprocessed split data was also used. However, the StandardScaler() function from the sklearn.preprocessing library was used to normalize the X\_train and X\_test data before model training, which means that the data was transformed to the same standard scale. A kernel trick technique was used for this model. The training data was fitted to the SVM Model #1 using the C-Support Vector Classification (SVC) function utilizing the kernel=’poly’ and degree=1 (SciKitlearn, n.d.). The poly (polynomial) kernel is used to separate data that has a curved shape (Sampaio, 2023). Using “degree=1” tells the model to use a 1-degree polynomial to find the hyperplane to split the data. Degree is only used with the poly kernel (Fraj, 2018).

The confusion matrix and classification reports for both the training and testing data were compared (see Figure 13 in the appendix). Looking at the confusion matrix for the training data, the model classified 98% of “0” predictions correctly, correctly classifying 7,606 as “0” and incorrectly classifying 130 as”2”. The model classified 98% of “1” predictions correctly, correctly classifying 123 as “1” and incorrectly classifying 3 as “5”. The model classified 94% of “2” predictions correctly, correctly classifying 6,923 as “2” and incorrectly classifying 355 as “0” and 94 as “4”. The model classified 89% of “3” predictions correctly, correctly classifying 1,710 as “3” and incorrectly classifying 15 as “2”, 165 as “4”, and 34 as ‘3”. The model classified 92% of “4” predictions correctly, correctly classifying 1,005 as “4” and incorrectly classifying 1 as “0”, 68 as “2”, and 23 as “3”. The model classified 95% of the “5” predictions correctly, correctly classifying 164 as “5” and incorrectly classifying 8 as “1” and 1 as “3”. The overall accuracy was 95%. The overall precision was 94%. The overall recall was 91%. The overall F1-score was 92%.

Looking at the confusion matrix for the testing data, the model classified 98% of “0” predictions correctly, correctly classifying 1,8912 as “0” and incorrectly classifying 35 as”2”. The model classified 95% of “1” predictions correctly, correctly classifying 35 as “1” and incorrectly classifying 2 as “5”. The model classified 93% of “2” predictions correctly, correctly classifying 1,741 as “2” and incorrectly classifying 96 as “0” and 37 as “4”. The model classified 89% of “3” predictions correctly, correctly classifying 379 as “3” and incorrectly classifying 4 as “2”, 36 as “4”, and 5 as “5”. The model classified 90% of “4” predictions correctly, correctly classifying 264 as “4” and incorrectly classifying 23 as “2” and 5 as “3”. The model classified 89% of “5’ predictions correctly, correctly classifying 47 as “5” and incorrectly classifying 6 as “1”. The overall accuracy was 95%. The overall precision was 92%. The overall recall was 90%. The overall F1-score was 91%. Since the overall accuracy is exactly the same (95%) for both the training data and testing data, the data seems to fit the SVM Model #1 perfectly. Although the other metrics for the testing data is slightly lower than the other metrics for the training data, in my opinion the model is not overfitting or may have very low insignificant amount of overfitting.

### Data Model 5: SVM Model #2

For the SVM Model #2, the same input features and target variable were used as for the RandomForestClassifier, GradientBoostingClassifier, and SVM #1 models. The normalized X\_train and X\_test data was used. To try to design a better performing SVM model, a Combining oversampling and undersampling technique was used to better balance the training and testing data before training the model. This was done by importing “SMOTETomek” from imblearn.combine, which automatically uses a interpolating sampling strategy that resamples the training data set, transforming it into a more balanced dataset, while fitting the “X\_train” and “y\_train” data to the model. The resampled data was stored in the new variables “X\_train\_resampled” and “y\_train\_resampled” (Imbalanced-learn, n.d.). Then the resampled training data was fitted to SVM Model #2 using the SVC function utilizing the kernel=’rbf’, C=10.0, and decision\_function\_shape=’ovr’.

The rbf kernel is the “Radial Basis Function” kernel, which is used to separate data that has circular shapes. The “C” hyperparameter is a regularization parameter that tells the model how much to avoid misclassifying training samples by designating a specific value that represents the margin, and the larger the value of “C”, the smaller the margin of error; Therefore, “C=10.0” is a large value, which designates a smaller margin of error (Sampaio, 2023). The “decision\_function\_shape = ovr”, is the One-vs-rest method that splits multiclass datasets into multiple binary datasets by modeling each class against all the other classes and then predictions are made using the model that is most confident (Brownlee, 2021; StackOverflow, n.d.). The decision\_function\_shape “ovr” has the shape (n\_samples, n\_classes) (SciKit-learn, n.d.).

The confusion matrix and classification reports for both the training and testing data were compared (see Figure 14 in the appendix). Looking at the confusion matrix for the training data, the model classified 99% of “0” predictions correctly, correctly classifying 7,532 as “0” and incorrectly classifying 90 as”2”. The model classified 99% of “1” prediction correctly, correctly classifying 7,957 as “1” and incorrectly classifying 80 as “5”. The model classified 95% of “2” predictions correctly, correctly classifying 7,669 as “2” and incorrectly classifying 396 as “0” and 12 as “4”. The model classified 96% of “3” predictions correctly, correctly classifying 7,800 as “3” and incorrectly classifying 1 as “2” and 287 as “4”. The model classified 96% of “4” predictions correctly, correctly classifying 7,656 as “4” and incorrectly classifying 1 as “0”, 165 as “2”, and 141 as “3”. The model classified 99.7% (rounded by CR as 100%) of “5” predictions correctly, correctly classifying 7,880 as “5” and incorrectly classifying 3 as “1” and 18 as “3”. The overall accuracy was 97%. The overall precision was 98%. The overall recall and F1-score were 97%.

Looking at the confusion matrix for the testing data, the model classified 99% of “0” predictions correctly, correctly classifying 1,880 as “0” and incorrectly classifying 23 as”2”. The model classified 98% of “1” predictions correctly, correctly classifying 1,775 as “1” and incorrectly classifying 42 as “5”. The model classified 95% of “2” predictions correctly, correctly classifying 1,891 as “2” and incorrectly classifying 89 as “0” and 8 as “4”. The model classified 97% of “3” predictions correctly, correctly classifying 1,924 as “3” and incorrectly classifying 1 as “2” and 61 as “4”. The model classified 95% of “4” predictions correctly, correctly classifying 1,914 as “4” and incorrectly classifying 48 as “2” and 50 as “3”. The model classified 90% of “5” predictions correctly, correctly classifying 1,941 as “5” and incorrectly classifying 208 as “1” and 11 as “3”. The overall accuracy was 95%. The overall precision was 96%. The overall recall and F1-score were 95%. When comparing the metrics for the training and testing data, the SVM Model #2 is overfitted less than the RandomForestClassifier decision tree model, but more than the SVM Model #1. In addition, using the combination of oversampling and undersampling did not help improve model predictions, it actually led to the model making close proportions of predictions for all 6 classes, which is very unrealistic for this dataset.

### Review of Data Models

To provide a comprehensive analysis for the “Global Air Pollution Dataset”, a multiple linear regression model was created, trained, tested, and analyzed and 4 classification models were created, trained, tested, and analyzed. The multiple linear regression model had an overall great performance, although when the PM2.5 AQI Values increases above 300, the correlation is weaker indicating that the model’s predictions for PM2.5 AQI Values above 300 are less accurate than PM2.5 AQI Values below 300, confirming that non-linear data is challenging for linear regression models. On the other hand, after comparing the 4 classification models’ performance (see Tables 3.1 and 3.2 in the appendix), it was determined that the most reliable and most accurate model with the least amount of overfitting was the SVM Model #1.

The overall accuracy for SVM Model #1 was 95%, staying consistent between the training and testing data. The precision of 92%, recall of 90%, and F1-score of 91% for the testing data showed a slight drop from the training data, but the scores display the metrics for an overall good performing model. However, the SVM Model #1 is best at accurately classifying the “Good” PM2.5 AQI Category, is good at accurately classifying the “Moderate” and “Unhealthy for Sensitive Groups” PM2.5 AQI Categories, but can use some improvement in more accurately classifying the “Unhealthy”, “Very Unhealthy”, and “Hazardous” PM2.5 AQI Categories. Nonetheless, the multiple linear regression model can be used for quantitative forecasting of PM2.5 AQI Values and to study the relationship of the specific pollutants in a particular geographic region, while the SVM Model #1 can be used for a qualitative forecasting of PM2.5 AQI Categories in specific geographic regions. Utilizing both models can capture both the linear and nonlinear trends in global air pollution.

### Final Results

### Findings

Exploratory data analysis of the Global Air Pollution Dataset was utilized to find insights about the distribution and characteristics of air pollution among different countries. There are 175 unique countries and 23,462 unique cities represented in the dataset. However, there are actually 195 recognized unique countries in the world, meaning the dataset contains air pollution data for only 89.7% of existing countries (Worldometer, n.d.). 104 countries (includes 4088 cities) have PM2.5 AQI Values greater than 100, which indicates that there is poor air quality in 59.4% of all counties in the dataset. 19 (includes 256 cities) of the 104 countries with PM2.5 AQI Values greater than 100 have PM2.5 AQI Values that are very unhealthy (201-300), representing 10.9% of all countries in the dataset (see Figure 15 in the appendix). 10 (includes 171 cities) of the 104 countries with PM2.5 AQI Values greater than 100 have PM2.5 AQI Values that are hazardous (>300), representing 5.7% of all countries in the dataset (see Figure 16 in the appendix).

The average PM2.5 AQI Value was calculated for each country. Tableau was used to create a global heatmap showing the average PM2.5 AQI Values for the different countries (see Figure 17 in the appendix). Based on the heatmap, Asian and African countries have the worst average PM2.5 AQI Values, which confirms the literature stated under the “Project Importance” section. The top ten countries with the highest average PM2.5 AQI Values, the top ten countries with the lowest average PM2.5 AQI Values, and the USA were plotted for comparison (see Figure 18.1 in the appendix). The top ten countries with the highest average PM2.5 AQI Values include, Republic of Korea, Bahrain, Mauritania, Pakistan, Aruba, Kuwait, United Arab Emirates, Senegal, India, and Saudi Arabia. The top ten countries with the lowest average PM2.5 AQI Values include, Andorra, Uruguay, Finland, Sweden, Papua New Guinea, Norway, Iceland, Maldives, Palau, and Solomon Islands.

Looking at the bar graph, we can see how the USA compares to the countries with the highest and lowest average PM2.5 AQI Values. The USA has the average PM2.5 AQI Value in the moderate air quality range. To help visual the distribution of data, Tableau was used to create a count plot that shows the count of PM2.5 AQI Categories for the top 10 countries with the highest and lowest average PM2.5 AQI Values, and the USA (see Figure 18.2 in the appendix). As seen from the count plot, the average PM2.5 AQI Values are biased for some countries since the amount of data varies and there is an unequal amount of data available for each country. 6 out of 21 countries have only one PM2.5 AQI Value. 5 out of 21 countries have only three PM2.5 AQI Values. The other 10 countries have more PM2.5 AQI Values. Furthermore, India, the USA, and Pakistan have the most PM2.5 AQI Values. Nonetheless, India has significantly more PM2.5 AQI Values than any other country and even has double the amount of PM2.5 AQI Values than the USA.   
 Nevertheless, there could be disparaties in the available data among different countries due to the differences in the size and the population of the country, and the number of cities within a country. For instance, India has a population of 1,450,935,791 and 2,488 cities, while the USA has a population of 345,426,571 and 2,872 cities. Senegal has a population of 18,501,984 and only 33 cities. Palau has a population of 17,695 and only 1 city (see Tables 4 and 5 in the appendix) (Worldometer, 2025). Another issue with the dataset is that all cities are not included for each country. For instance, according to the World Population Review (2024), India, the USA, and Senegal actually have more cities than what is included in the data set. Therefore, the dataset is not completely representative for the true amount of PM2.5 air pollution for each country. With that said, with the limited amount of data available, any comparisons made between countries may be inaccurate (World Health Organization, 2024).

Moreover, knowing that there are data limitations for the Global Air Pollution Dataset, the models created for this project may not fully capture the true aspects of PM2.5 air pollution for some countries or regions. However, as discussed in the Data Modeling section, to capture both the linear and nonlinear trends in global air pollution, the multiple linear regression model and the SVM Model #1 can be used. To analyze the predictions made from the multiple linear regression model, a dataframe was created linking each of the 4,067 PM2.5 AQI Value predictions to each country. Then the average PM2.5 AQI Value predictions were calculated for each country and categorized by PM2.5 AQI Categories. Next, the linear regression model average predicted PM2.5 AQI Values by Country was plotted on a graph to visualize the predictions (see Figure 9 in the appendix). Looking at the predictions, India’s average PM2.5 AQI Value was predicted to be Unhealthy for Sensitive Groups (148.9), which is close to what the average PM2.5 AQI Value was for the original data (149.5).

USA’s average PM2.5 AQI Value was predicted to be Moderate (57.9), which is close to what the average PM2.5 AQI Value was for the original data (58.0). Senegal’s average PM2.5 AQI Value was predicted to be Unhealthy for Sensitive Groups (150.1), which is close to what the average PM2.5 AQI Value was for the original data (152.4), but due to the difference of 2.3, the predicted PM2.5 AQI Category is Unhealthy for Sensitive Groups, while the actual PM2.5 AQI Category is Unhealthy. Palau’s average PM2.5 AQI Value was predicted to be Good (15.4), which is a higher value compared to what the average PM2.5 AQI Value was for the original data (7.0), but is still within the same PM2.5 AQI Category. Looking at another country Aruba that has only one city included in the data, Aruba’s average PM2.5 AQI Value was predicted to be Unhealthy (159.4), which is close to what the average PM2.5 AQI Value was for the original data (163.0). Evaluating the average PM2.5 AQI Value predictions from only the top ten countries with the highest average PM2.5 AQI Values, the top ten countries with the lowest average PM2.5 AQI Values, and the USA previously discussed above, overall, the linear regression model does a great job at making predictions.

All except three average PM2.5 AQI Value predictions matched the original PM2.5 AQI Category for each country. Three countries’ actual vs. predicted average PM2.5 AQI Values were close, but only differed by one PM2.5 AQI Category. Seven countries did not have any predictions (see Table 6 in the appendix). There were not any average predictions in the Very Healthy or Hazardous PM2.5 AQI Categories. However, policy makers can further analyze the predictions specifically by city if they would want to create new policies to target PM2.5 air pollution mitigation in higher polluted cities, which according to the single PM2.5 AQI predictions, include cities within the Democratic Republic of the Congo, India, Mexico, South Africa, Nigeria, Chile, Pakistan, Indonesia, Brazil, China, and Uzbekistan; these were countries had cities with PM2.5 AQI Value predictions >200.9 that were either in the Very Unhealthy or Hazardous PM2.5 AQI Categories.

To analyze the country specific predictions made from the SVM Model #1, which was the best classification model, a dataframe was created to show the predicted PM2.5 AQI Categories for each country. However, there were PM2.5 AQI Categories predicted for only 43 countries, but a total of 4,067 predictions were made overall (see Figure 19 in the appendix). There was a disparity in the number of PM2.5 AQI Category predictions for each of the 43 countries, as shown by Figure 19, which is expected since there is limited data for some countries. To narrow down the analysis, the following PM2.5 AQI Category predictions of countries from the top ten countries with the highest average PM2.5 AQI Values, the top ten countries with the lowest average PM2.5 AQI Values, and the USA, previously discussed above, were plotted on a count plot. The count plot was filtered to show PM2.5 AQI Category predictions for the Republic of Korea, India, Mauritania, Pakistan, Kuwait, Norway, Papua New Guinea, Finland, Sweden, Solomon Islands, and the USA, but since there were not any predictions made for all countries, the count plot only showed the PM2.5 AQI Category predictions for the USA, India, Finland, and Pakistan (see Figure 18.3 in the appendix). The same information was also visualized in a heatmap (see Figure 18.4 in the appendix).

For the USA and Finland, the SVM Model #1 was able to make predictions for all PM2.5 AQI Categories. For Finland, the model made predictions for all PM2.5 AQI Categories, except for the Moderate category. The model only made predictions for the Good, Unhealthy for Sensitive Groups, and Unhealthy PM2.5 AQI Categories for Pakistan. Due to the disparity in the data, we can see that the USA and India had the highest number of predictions. In addition, the USA had the highest number of predictions for each PM2.5 AQI Category. However, since these particular countries and the previously mentioned other countries from the top ten countries with the highest average PM2.5 AQI Values and the top ten countries with the lowest average PM2.5 AQI Values cannot be accurately compared due to data disparities, the top ten countries with the most predictions represented in the data were plotted on a count plot (see Figure 20.1 in the appendix).

The top ten countries with the most predictions represented in the data include, the USA, India, China, Brazil, Germany, Somalia, Finland, Turkey, Japan, and the Neverlands. This data was also visualized using a heatmap (see Figure 20.2 in the appendix). When comparing these countries, China had the highest number of Good and Unhealthy for Sensitive Groups PM2.5 AQI Category predictions. The USA had the highest number of Moderate, Unhealthy, Very Unhealthy, and Hazardous PM2.5 AQI Category predictions. Turkey had the lowest number of Unhealthy for Sensitive Groups PM2.5 AQI Category predictions. The Neverlands and Turkey both had the lowest number of Unhealthy PM2.5 AQI Category predictions. The Neverlands was the only country that did not have any predictions for the Very Unhealthy PM2.5 AQI Category. The Neverlands, Somalia, and Turkey were the only three countries that did not have any predictions for the Hazardous PM2.5 AQI Category. Based on the Data Modeling section above, it was already determined that the SVM Model #1 incorrectly classified 5% of the predictions incorrectly and the distribution of class predictions was illustrated by the confusion matrix, but the amount of error is minimal (see Figure 13 in the appendix).

When comparing the analysis of PM2.5 AQI predictions for both the multiple linear regression model and the SVM Model #1, both models can be very useful to assist government officials to enact new policy changes that can help mitigate the future health impact of PM2.5 air pollution. The multiple linear regression model would enable the government to complete a high-level city review of PM2.5 air pollution so they could target the cities with the worst PM2.5 air pollution. On the other hand, the SVM Model #1 can be used to get a low-level country review of PM2.5 AQI Categories within specific countries. If the government wanted to get a quick snap shot of a country as a whole, they should use the SVM Model #1. Although the ideal goal would be to decrease PM2.5 air pollution levels down to the “Good” category, this would be infeasible. The more realistic goal may be to enact policies that could possibly decrease PM2.5 air pollution levels to at least the “Unhealthy for Sensitive Groups” category, but this could also be a far reach. However, the government could use the models to identify the countries or cities with the least PM2.5 air pollution levels to determine what policies and factors present in those countries or cities can be enacted in the countries or cities with the PM2.5 air pollution levels greater than 150.9.

The main goal for completing this analysis was to aide in policy changes, as this analysis can **be used by government agencies to make important decisions about regulations, policies, and legislature to further decrease air pollution in problematic geographic locations. However, an issue lies within the lack of representation of cities and countries within the data due to the lack of available air pollution data.** which most likely led to bias in the linear regression and SVM Model #1. As previously noted, there are even some countries that only have 1 data point within the entire data set. The lack of air pollution data may be due to the lack of high-quality air pollution monitoring systems or air pollution sensors and/or faulty low-cost air pollution sensors (Berrisford et al., 2024).

A study done by Berrisford et al. (2024) showed that the lack of real-world air pollution data for a specific country or region leads to poor and inaccurate predictions. Therefore, PM2.5 AQI predictions made by the designated linear regression models and SVM Model #1 can be inaccurate or underrepresented for a specific country or region, leading to the model having biases against specific countries or regions, which was demonstrated in this project. Consequentially, this can lead to the lack or neglect for the implementation of new policy changes that could lower PM2.5 air pollution in specific countries or regions. Therefore, it is crucial to have sufficient and good quality data for all countries or regions. On the other hand, the discovery of the lack of data in specific regions could possibly prompt the government to disseminate and install air pollution monitoring systems or air pollution sensors in the countries or regions that lack air pollution data.

### Review of Success or Completion

**Iterating through each step of the data science life cycle has enabled me to reflect on the challenges and limitations associated with the project.** The overall goal of this project was to use the Global Air Pollution Dataset to analyze historical data to identify the trends in global air pollution, in order to predict or forecast the future PM2.5 air pollution levels to determine the future effect on public health in different countries or geographical regions. **The insights provided can then be used by government agencies to make important decisions about regulations, policies, and legislature to further decrease air pollution in problematic geographic locations. The project idea was selected, the problem was defined, and the necessary research was done to gain a thorough understanding of the problem and the data.** The project scope clearly outlined the importance of the problem and what processes, tools, and resources were required to complete the project. During exploratory data analysis, the creation of different visualizations, tables, and the review of descriptive statistics revealed that the dataset had some missing data, a possible outlier, and a class imbalance between the distribution of health risk classes. In addition, while further exploring the general distribution of available air pollution data by country, it was discovered that there were disparities in the amount of air pollution data among the different countries. These data quality issues presented a challenge and the risk of biases for the machine learning models’ ability to make accurate and realistic predictions for countries which lack sufficient data.

To address the missing data, the data entries with missing values were excluded from the analysis and the one possible outlier was determined to most likely be a true value and was included in the analysis. After data preprocessing, one multiple linear regression model and 4 classification models were created, trained, tested, and analyzed. Through careful feature selection, I was able to create an overall good performing multiple linear regression model, and with data normalization, a relatively good performing SVM classification model. Although I thought that using a technique to balance the health risk classes for global pollution data would make the SVM model perform better, it actually led to the model making unrealistic predictions. The resulting outcome of balancing the health risk classes revealed that health risk class imbalance is not an issue for global air pollution data and leads to more ideal and realistic predictions. However, since the dataset is not completely representative for the true amount of PM2.5 air pollution for each country, any comparisons made between countries may be inaccurate and the models may not be capturing the comprehensive levels of PM2.5 air pollution.

Nonetheless, although there is some bias in the multiple linear regression model, the model made really great and highly accurate PM2.5 AQI Value predictions. Although the SVM Model #1 had more bias, it made relatively good PM2.5 AQI classification predictions, but most definitely need more finer hyperparameter tuning and optimization to improve the model’s performance. Nevertheless, the air pollution data gap needs to be filled to create the most accurate and productive models. At this point of time, both models can still be useful to help determine the possible future impacts of PM2.5 air pollution in different countries and regions in order implement policy changes that can mitigate PM2.5 air pollution; However, stakeholders must be well informed and aware of the possible model limitations.

### Potential Data Privacy and Data Security Issues

Since the Global Air Pollution Dataset contains data that is not comprised of sensitive or personal health information, there is minimal data privacy and security issues associated with it. The data was collected from elichens.com who collected the real-time air pollution data themselves using their own developed outdoor eLos Air Quality Stations. In addition to measuring PM2.5, CO, O3, NO2, and SO2, the eLos Air Quality Stations also monitor temperature, humidity, noise, and pressure, but nothing directly connected to any personal information (eLichens, 2025). eLichens also uses air quality data collected by Local Air Quality Agencies (eLichens, 2024). However, all the data is made available for public access and use. Nonetheless, there are general issues relating to global air pollution data.

This project revealed that there is insufficient air pollution data available for some countries. This is not just an issue with this dataset, this is an issue pertaining to all global air pollution data. It is crucial to have access to reliable air pollution data for all countries in order for policy makers to understand and take corrective action to improve air quality (Sawant et al., 2022). As there is insufficient or missing air pollution data for some countries, due to the lack of high-quality air pollution monitoring systems or air pollution sensors and/or faulty low-cost air pollution sensors, there is also a lack of data due to some governments not making their data public. Only 53% of countries share their air quality data publicly. 26% of countries that share their air pollution data do not provide the geographical location coordinates of the monitoring station, lacking full transparency (Sawant et al., 2022).

The lack of full transparency and the lack of sharing air pollution data by some countries may be due to the difference in national and international laws and regulations pertaining to the release of government data due to privacy and security issues (ODC, n.d.). A report written by Hasenkopf et al. (2023), disclosed that available open-source air pollution data, such as provided by the open-air quality data platform OpenAQ, combined air quality data from government and other sources across the world. OpenAQ is known to be the largest open-source air quality data platform in the world (Sawant et al., 2022).Some local actors in countries where the government does not publicly share or does not monitor air pollution, self-report air quality data they have collected from their own initiatives to OpenAQ. These local actors, which may consist of individuals from community organizations, government agencies, non-profit groups, medical fields, media entities, the private sector, and research and academic institutions, may want to remain anonymous (Hasenkopf et al., 2023); Therefore, OpenAQ must protect their privacy.

### Recommendations for Future Analysis

As the availability of global air pollution data was a challenge and limitation faced in the analysis of the Global Air Pollution Dataset, to create more accurate and productive machine learning models in the future, it is crucial to collect more data. It may be best to incorporate additional collected data from OpenAQ since OpenAQ collects non-aggregated ground-level ambient air quality measurements from hundreds of different sources and is known to be the largest air pollution data platform (OpenAQ, 2025). However, the OpenAQ database allows programmatic and queryable access through API, which would require additional resources, such as a big data platform, such as AWS and more complex preprocessing of the data since the data is provided in physical units instead of AQI units (OpenAQ, n.d.; OpenAQ, 2025). There are also other websites that contain tabular air quality data that was collected and extracted from OpenAQ, such as Kaggle and Opendatasoft (OpenAQ & Dane, n.d.; Opendatasoft, n.d.). In addition to requiring more resources, the collection of data from multiple sources can lengthen the data preprocessing and data cleaning timeframe. Of course, more data can lead to a better performing model, but also including additional features could further enhance models.

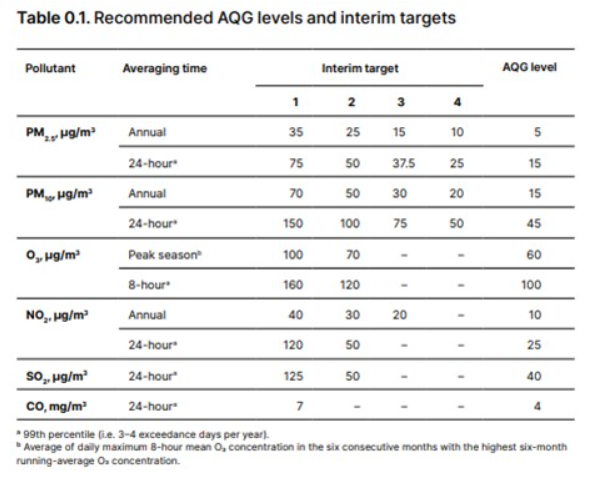
For instance, in order for the government to formulate better policies that can more efficiently mitigate air pollution, data pertaining to emission sources could be added to the analysis and training of models. I could also try to incorporate hourly analysis to pinpoint timeframes when pollution levels peak (Berrisford et al., 2024). Since some countries have smaller or larger populations, I could also incorporate a population-weighted annual average concentration for each country by calculating the PM2.5 per Capita, which would provide the level of exposure per person enabling a more equal comparison between countries. I would need to calculate the total PM2.5 emissions per country and would need to also include population data and feature engineer a new column containing the calculated PM2.5 per Capita by dividing the total PM2.5 emissions by the population size (Microsoft Bing, 2025; State of Global Air, 2025). Adding the additional above dimensions to the analysis for global air pollution will provide more useful insights to be used as a guidance by policy makers. However, it will make the analysis much more complex and time consuming.

Moreover, there might be a better approach to model selection for future analysis. Instead of testing different models separately, I could train a library of different models with different parameter or features, which is termed “model ensembling”. The use of ensemble selection will automate the selection of the best ensemble within the model library. This process is a type of ensemble learning used for model building that combines multiple predictive models to produce a new model that is more accurate than all individual models. Model ensembling was proven to be better than any single model since it corrects for errors of each individual model, which was demonstrated in a Kaggle competition described by Al-Taie et al. (2017). Nonetheless, optimal data preprocessing and feature engineering, and selecting the most significant features will optimize model training, leading to better performing models.

# Appendix

**Table 1**

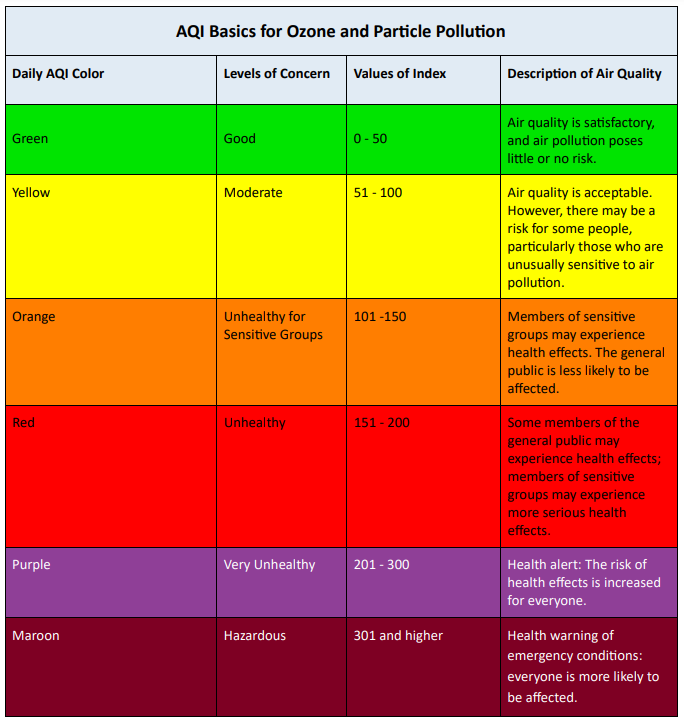
*WHO AQG Guidelines table*



(WHO, 2024)

**Table 2**

*EPA AQI Table for O3 and PM5.2*



(AirNow. n.d.)

**Table 3.1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Overall Precision** | **Overall Recall** | **Overall F1-Score** |
| **Decision Tree Model #1**  **(Using Training Data)** | 1.00 | 1.00 | 1.00 | 1.00 |
| **Decision Tree Model #1**  **(Using Testing Data)** | 0.99 | 0.96 | 0.96 | 0.96 |
| **Decision Tree Model #2**  **(Using Training Data)** | 0.99 | 0.98 | 0.98 | 0.98 |
| **Decision Tree Model #2**  **(Using Testing Data)** | 0.98 | 0.96 | 0.96 | 0.96 |
| **SVM Model #1**  **(Using Training Data)** | 0.95 | 0.94 | 0.91 | 0.92 |
| **SVM Model #1**  **(Using Testing Data)** | 0.95 | 0.92 | 0.90 | 0.91 |
| **SVM Model #2**  **(Using Training Data)** | 0.97 | 0.98 | 0.97 | 0.97 |
| **SVM Model #2**  **(Using Testing Data)** | 0.95 | 0.96 | 0.95 | 0.95 |

*Comparison of Classification Model Performance Metrics (Best model highlighted)*

**Table 3.2**

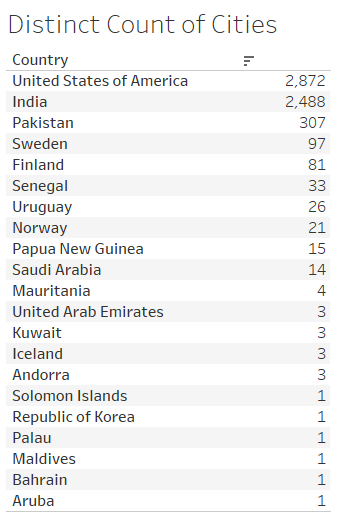
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Percentage of Correct “Good” Classifications**  **(Precision)** | **Percentage of Correct “Moderate” Classifications**  **(Precision)** | **Percentage of Correct “Unhealthy for Sensitive Groups” Classifications**  **(Precision)** | **Percentage of Correct “Unhealthy” Classifications**  **(Precision)** | **Percentage of Correct “Very Unhealthy” Classifications**  **(Precision)** | **Percentage of Correct “Hazardous” Classifications**  **(Precision)** |
| **Decision Tree Model #1** | 99% | 95% | 99% | 98% | 94% | 93% |
| **Decision Tree Model #2** | 99% | 95% | 99% | 98% | 94% | 90% |
| **SVM**  **Model #1** | 98% | 95% | 93% | 89% | 90% | 89% |
| **SVM**  **Model #2** | 99% | 98% | 95% | 97% | 95% | 90% |

*Comparison of Classification Model Performance on Testing Data*

**Table 4**

*Count of Cities for the Top 10 Countries with the Highest and Lowest Average PM2.5*

*AQI, & the USA*



|  |  |
| --- | --- |
| **Country** | **Population** |
| United States of America | 345,426,571 |
| India | 1,450,935,791 |
| Pakistan | 251,269,164 |
| Sweden | 10,606,999 |
| Finland | 5,617,310 |
| Senegal | 18,501,984 |
| Uruguay | 3,386,588 |
| Norway | 5,576,660 |
| Papua New Guinea | 10,576,502 |
| Saudi Arabia | 33,962,757 |
| Mauritania | 5,169,395 |
| United Arab Emirates | 11,027,129 |
| Kuwait | 4,934,507 |
| Iceland | 393,396 |
| Andorra | 81,938 |
| Solomon Islands | 819,198 |
| Republic of Korea | 51,717,590 |
| Palau | 17,695 |
| Maldives | 527,799 |
| Bahrain | 1,607,049 |
| Aruba | 108,066 |

**Table 5**

*Population for the Top 10 Countries with the Highest and Lowest Average PM2.5*

*AQI, & the USA*

(Worldometer, 2025)

**Table 6**

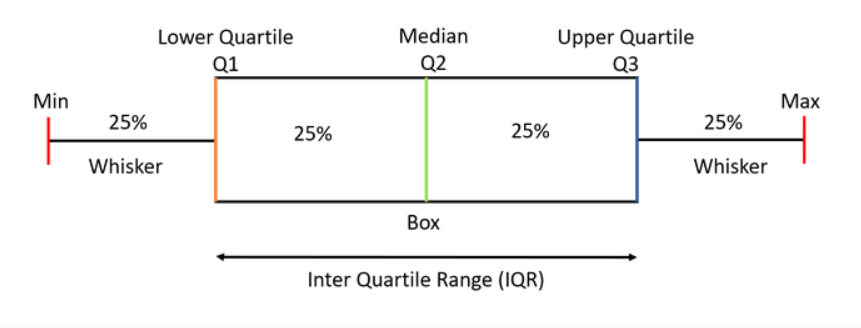
*Multiple Linear Regression Model’s Actual vs Predicted Average PM2.5 AQI Values for*

*the Top 10 Countries with the Highest and Lowest Average PM2.5 AQI, & the USA*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Country** | **Actual Average PM2.5 AQI Value** | **Actual PM2.5 AQI Category** | **Predicted Average PM2.5 AQI Value** | **Predicted PM2.5 AQI Category** |
| United States of America | 58.0 | Moderate | 57.9 | Moderate |
| India | 149.5 | Unhealthy for Sensitive Groups | 148.9 | Unhealthy for Sensitive Groups |
| Pakistan | 173.1 | Unhealthy | 173.6 | Unhealthy |
| Sweden | 21.0 | Good | 32.6 | Good |
| Finland | 21.6 | Good | 38.5 | Good |
| Senegal | 152.4 | Unhealthy | 150.1 | Unhealthy for Sensitive Groups |
| Uruguay | 21.7 | Good | 26.6 | Good |
| Norway | 18.6 | Good | 39.7 | Good |
| Papua New Guinea | 20.5 | Good | 19.6 | Good |
| Saudi Arabia | 149.3 | Unhealthy for Sensitive Groups | 166.2 | Unhealthy |
| Mauritania | 179.0 | Unhealthy | 164.5 | Unhealthy |
| United Arab Emirates | 152.7 | Unhealthy | 146.1 | Unhealthy for Sensitive Groups |
| Kuwait | 162.0 | Unhealthy | No prediction | No prediction |
| Iceland | 18.3 | Good | No prediction | No prediction |
| Andorra | 22.0 | Good | No prediction | No prediction |
| Solomon Islands | 6.0 | Good | No prediction | No prediction |
| Republic of Korea | 415.0 | Hazardous | No prediction | No prediction |
| Palau | 7.0 | Good | 15.4 | Good |
| Maldives | 15.0 | Good | No prediction | No prediction |
| Bahrain | 188.0 | Unhealthy | No prediction | No prediction |
| Aruba | 163.0 | Unhealthy | 159.4 | Unhealthy |

**Figure 1**

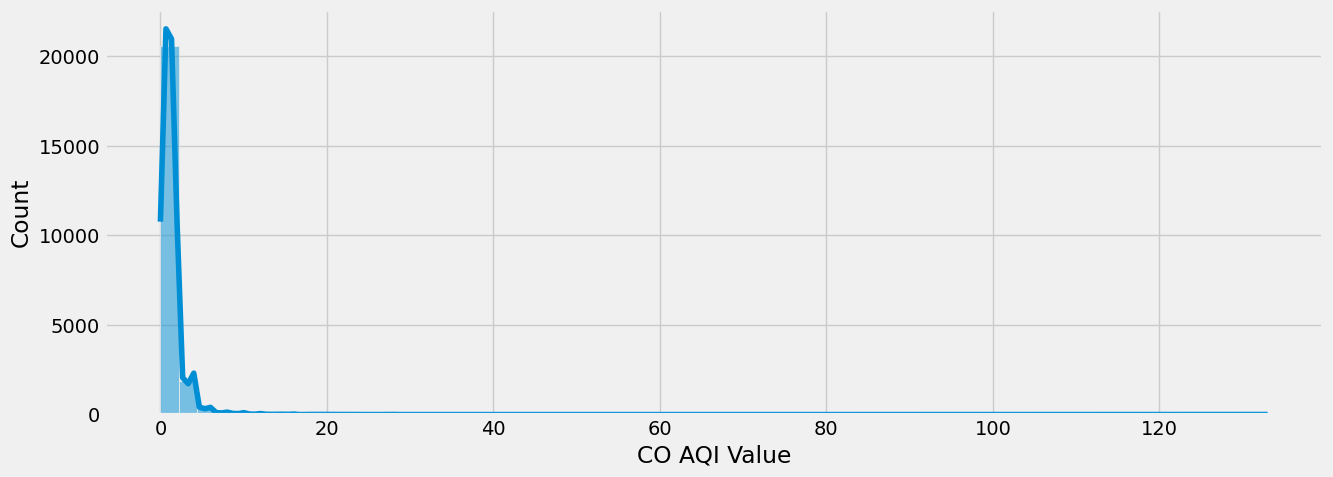
*Diagram of Box Plot*



(GeeksforGeeks, 2024)

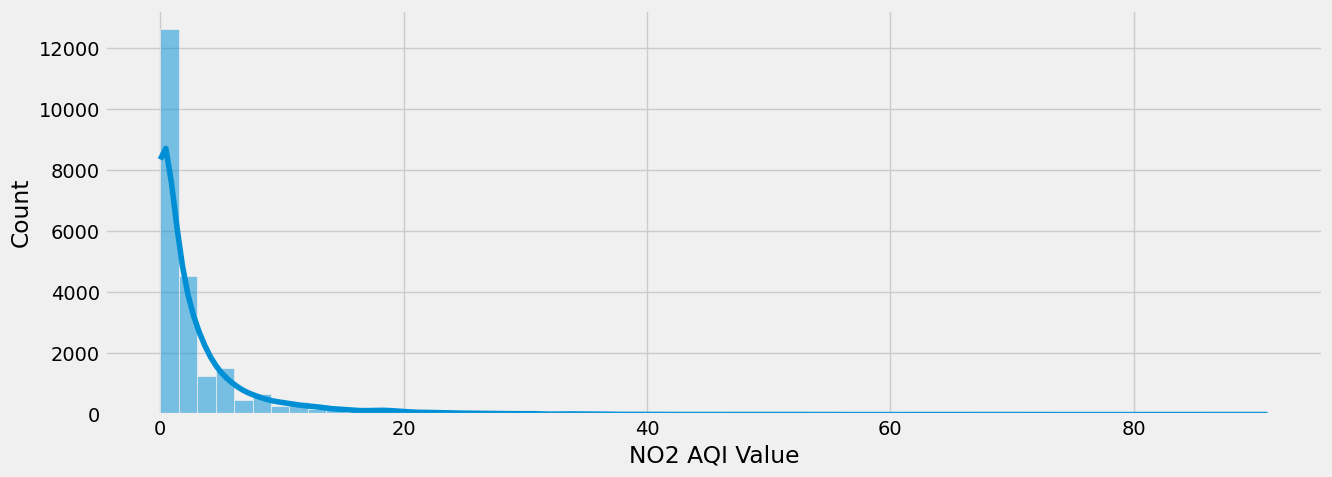
**Figure 2**

*Histogram showing the distribution of data for the CO AQI Value*



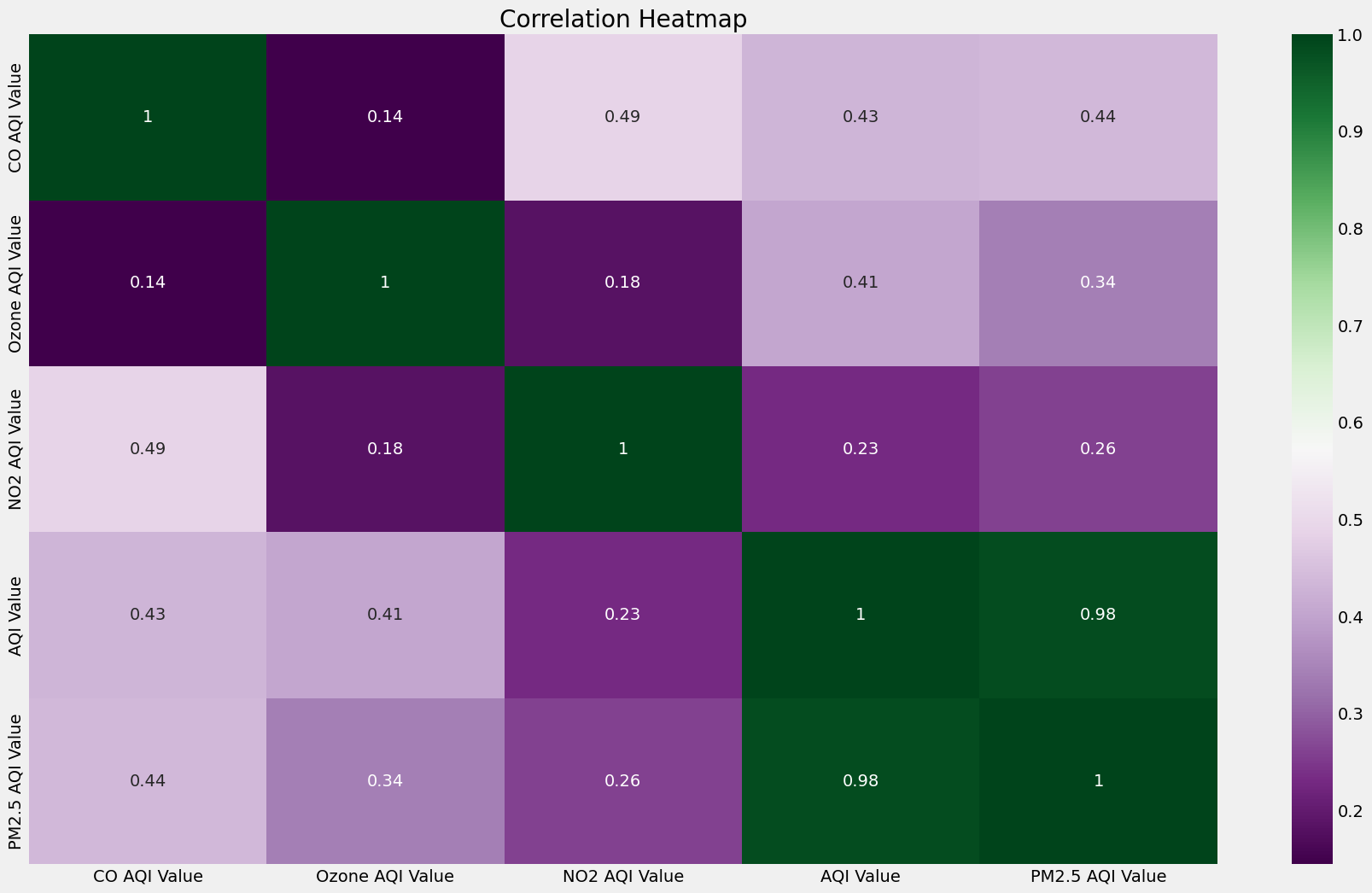
**Figure 3**

*Histogram showing the distribution of data for the NO2 AQI Value*



**Figure 4**

*First Correlation Heat Map Linear Regression Model Feature Selection*



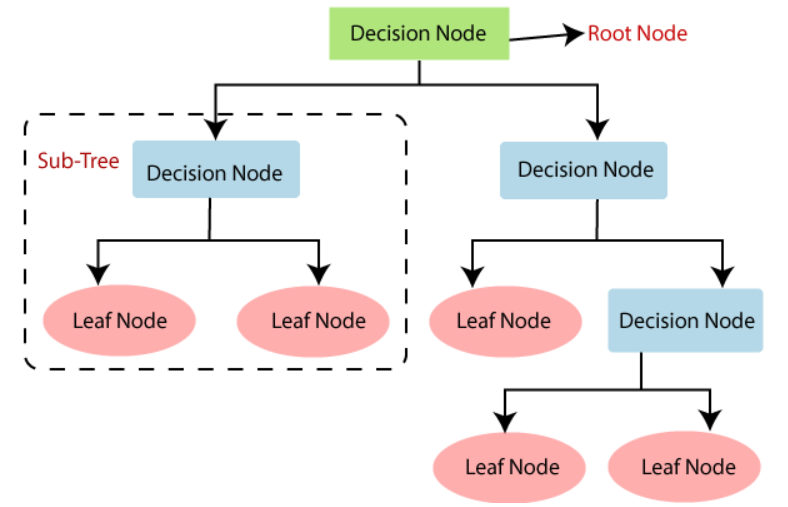
**Figure 5**

*SNS Pairplot showing the distribution of data*



**Figure 6**

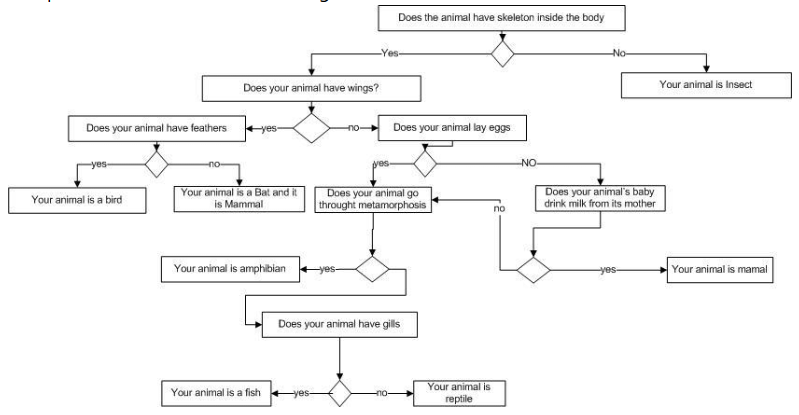
*Decision tree model structure*



(machinelearningplus, 2025)

**Figure 7**

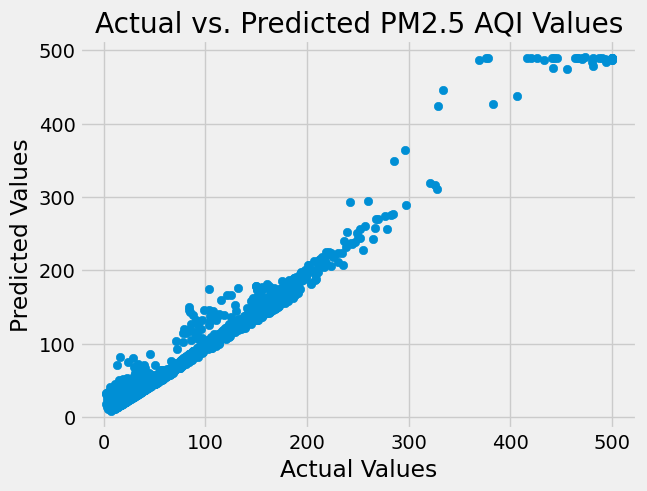
*Example workflow of a decision tree model*



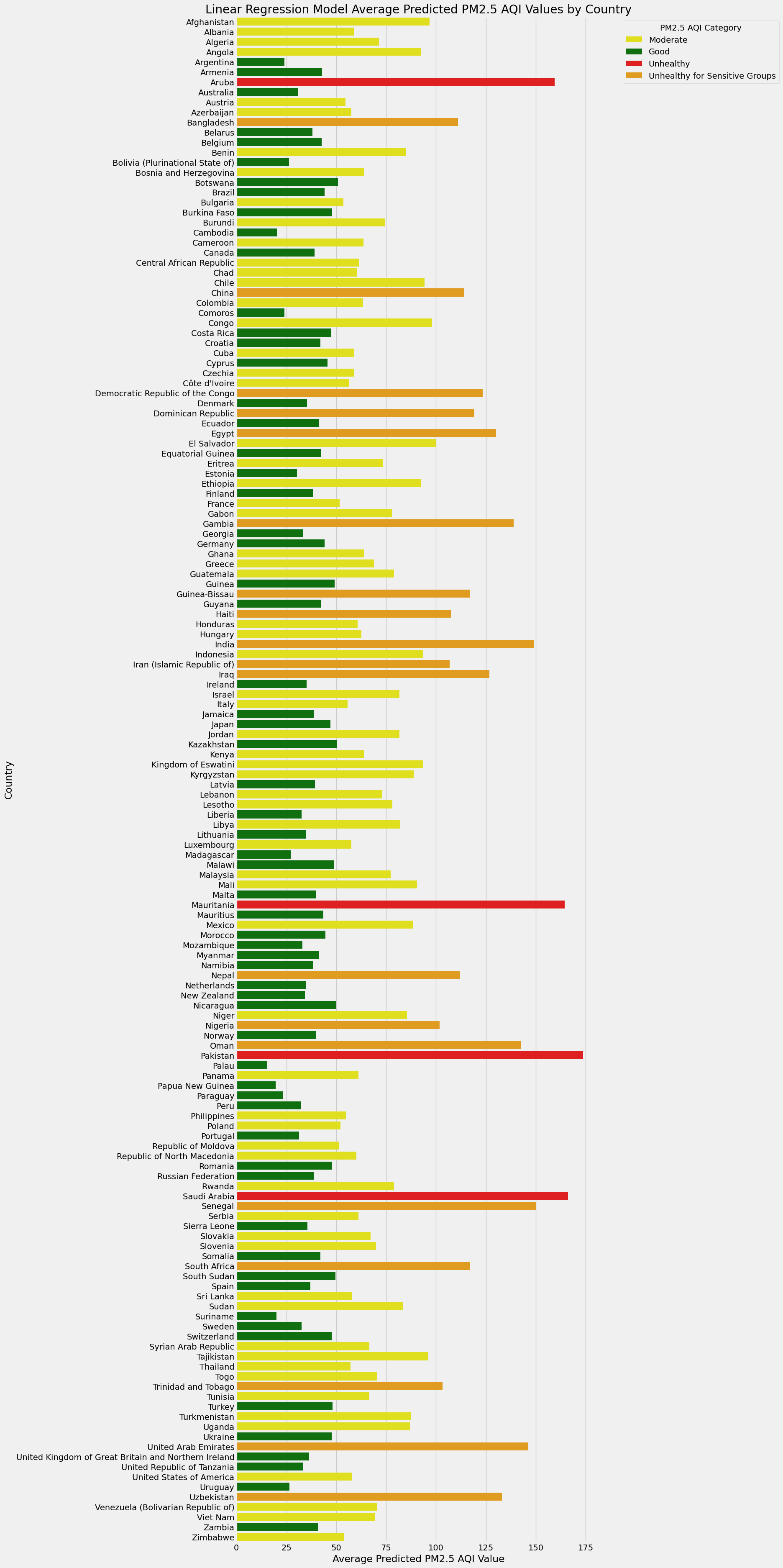
(machinelearningplus, 2025)

**Figure 8**

*Linear Regression Model comparison of actual vs predicted PM2.5 AQI Values*

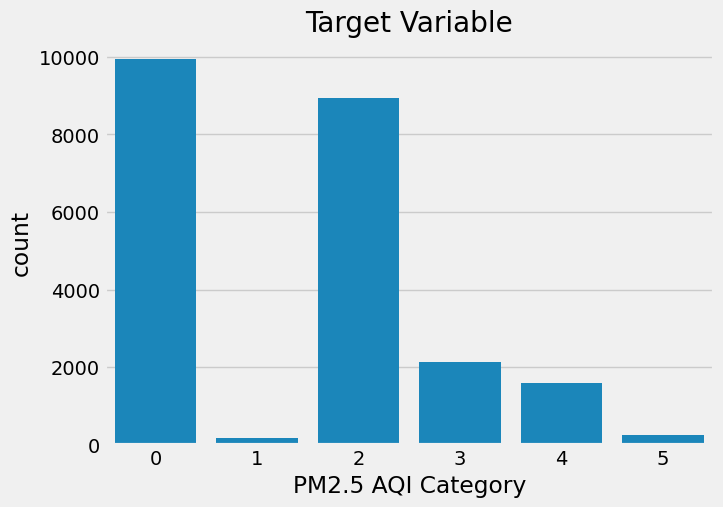


**Figure 9**

*Linear Regression Model Average PM2.5 Predictions by Country*

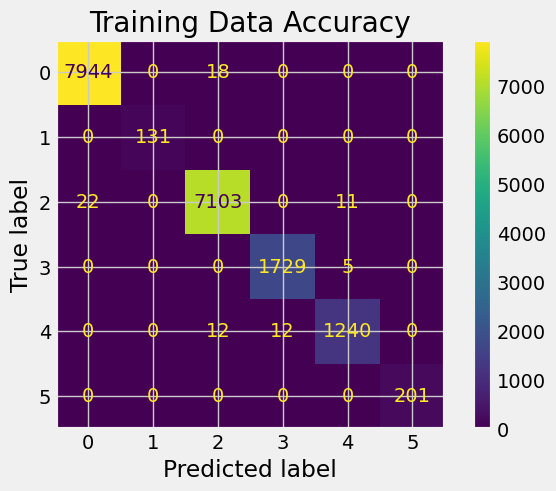
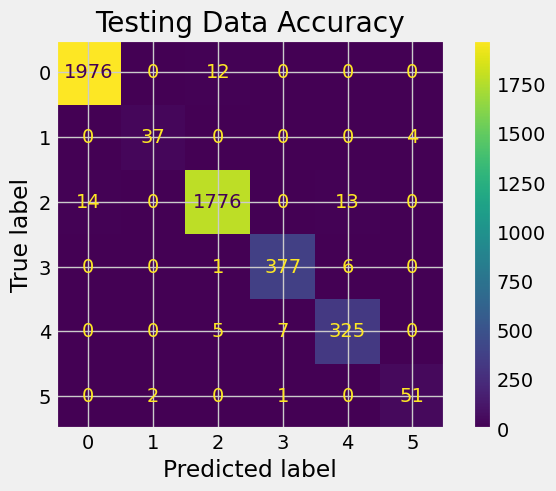
**Figure 10**

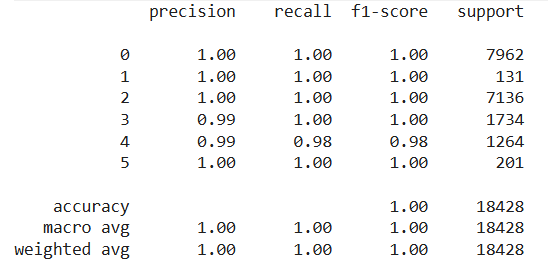
*The count of PM2.5 AQI Categories*



**Figure 11**

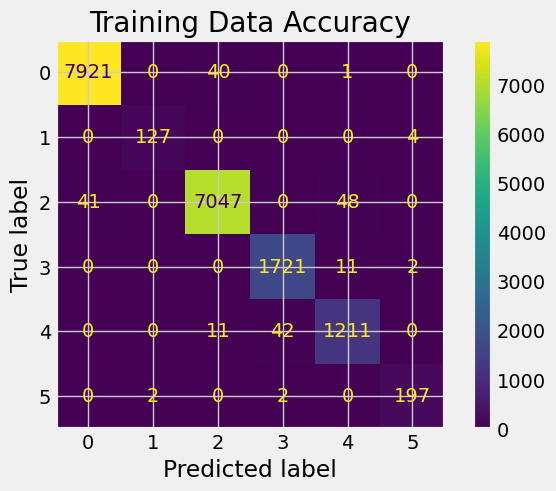
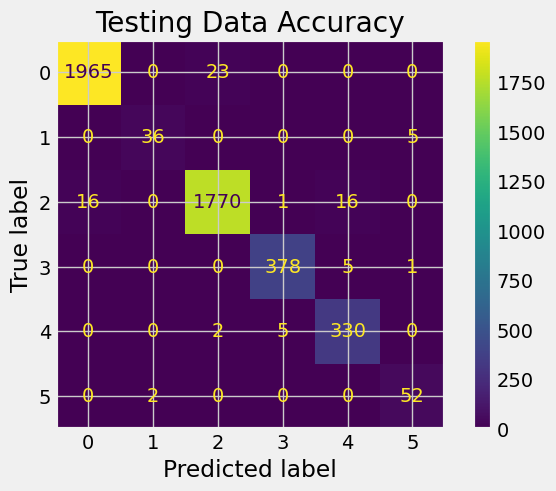
*Decision Tree Model #1 Confusion Matrixes and Classification Reports*

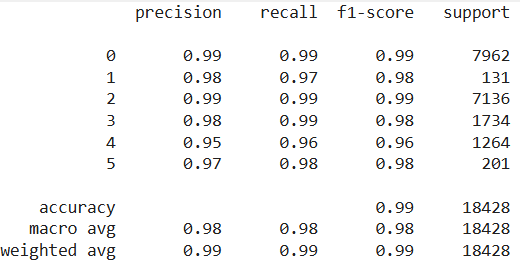
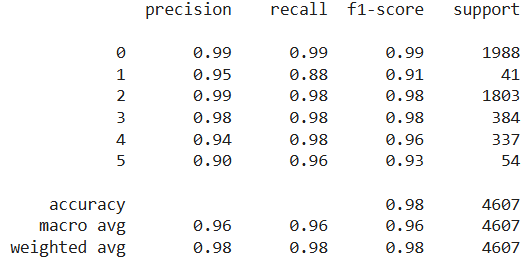
 

**Figure 12**

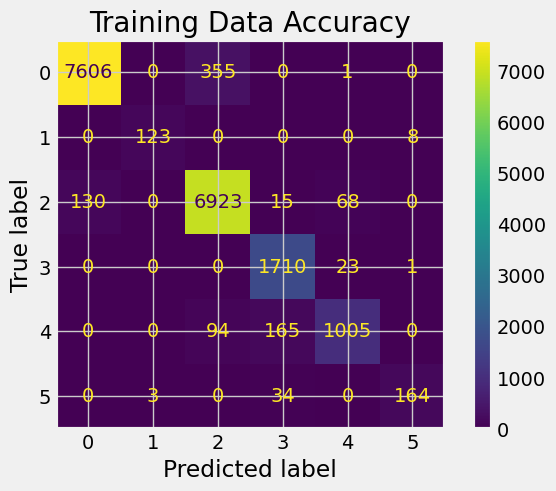
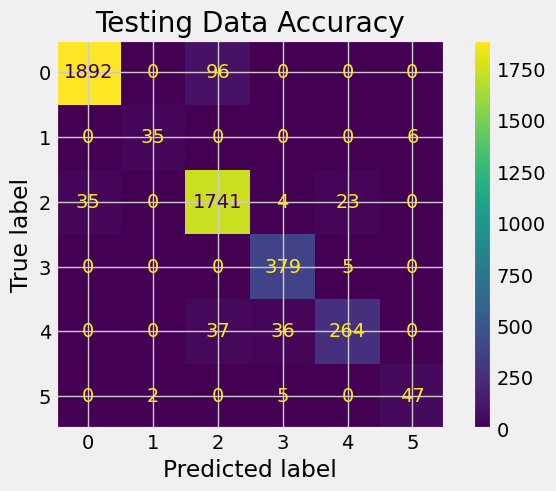
*Decision Tree Model #2 Confusion Matrixes and Classification Reports*

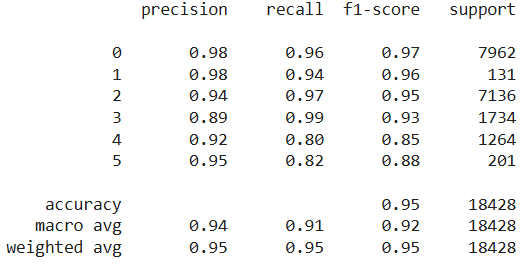
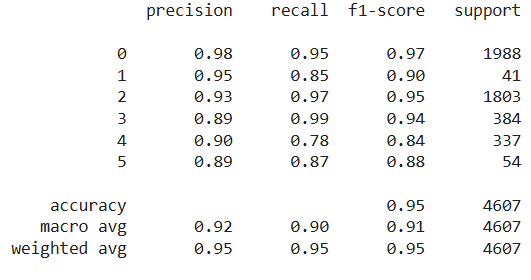
 

**Figure 13**

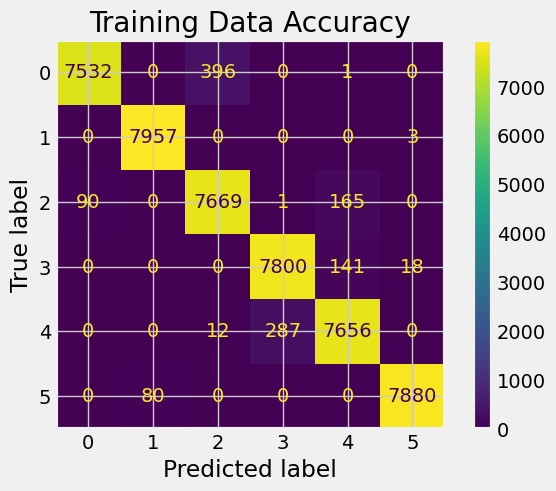
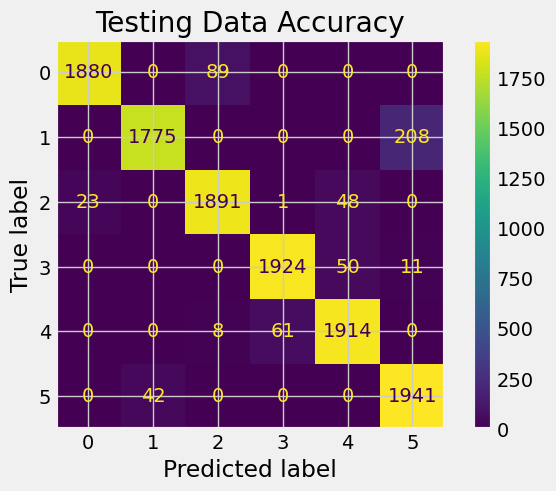
*SVM Model #1 Confusion Matrixes and Classification Reports*

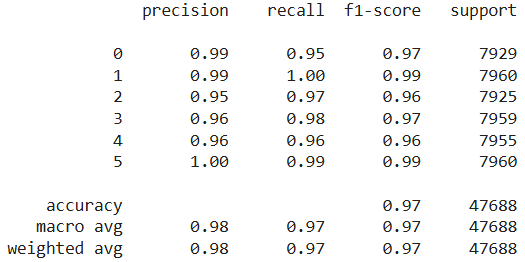
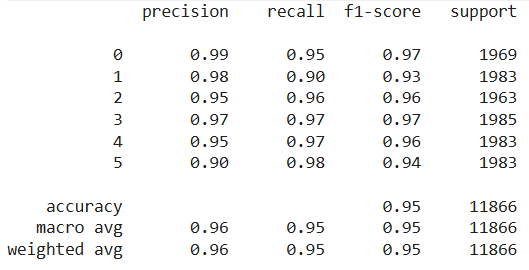
 

**Figure 14**

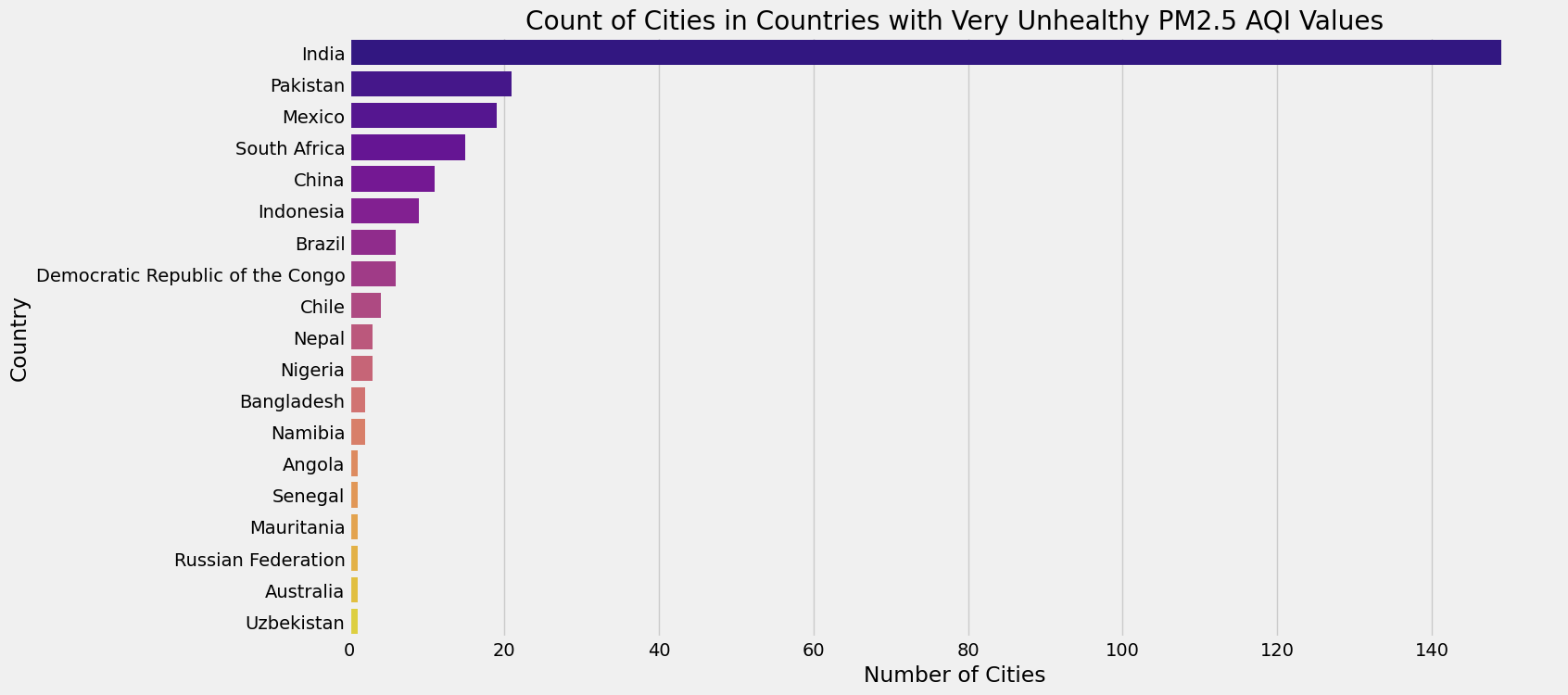
*SVM Model #2 Confusion Matrixes and Classification Reports*

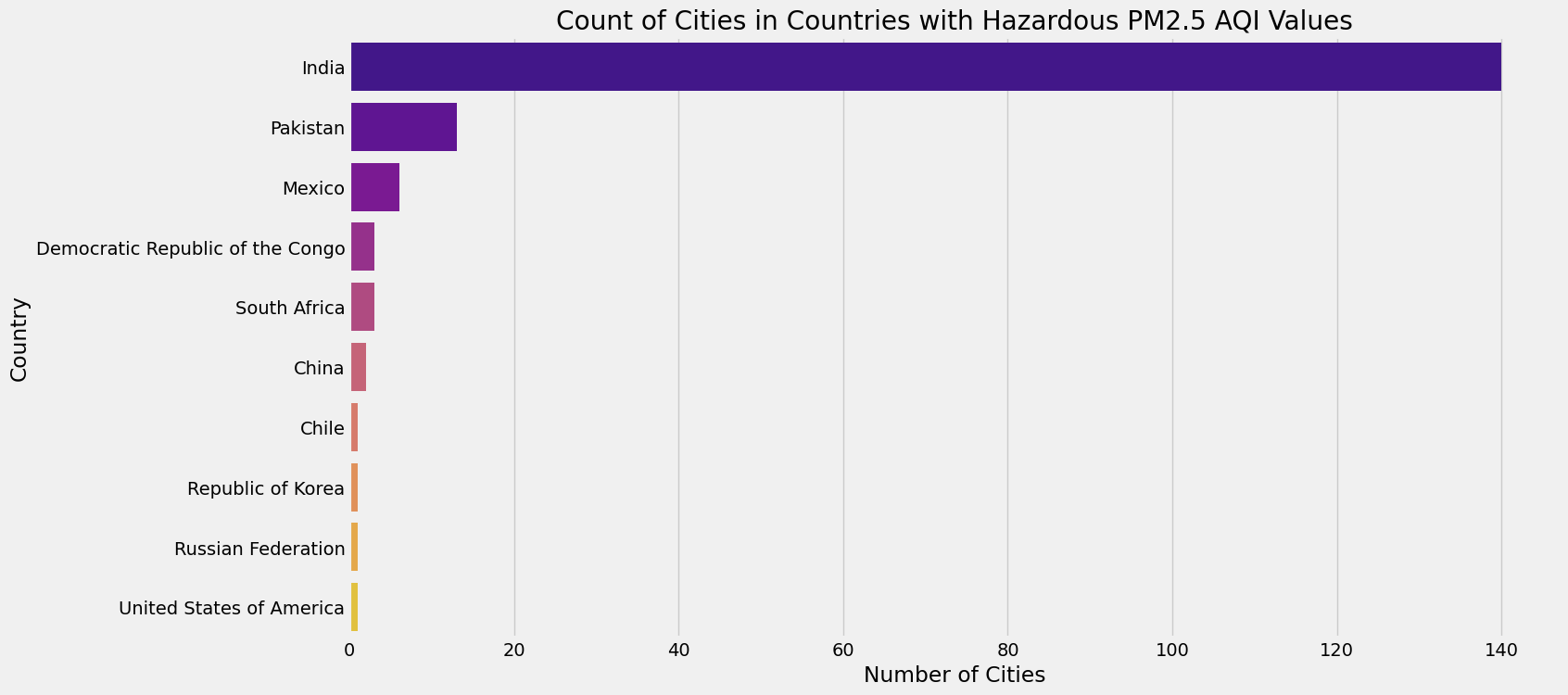
**Figure 15**

*Counts of Cities in Countries with Very Unhealthy PM2.5 Values*

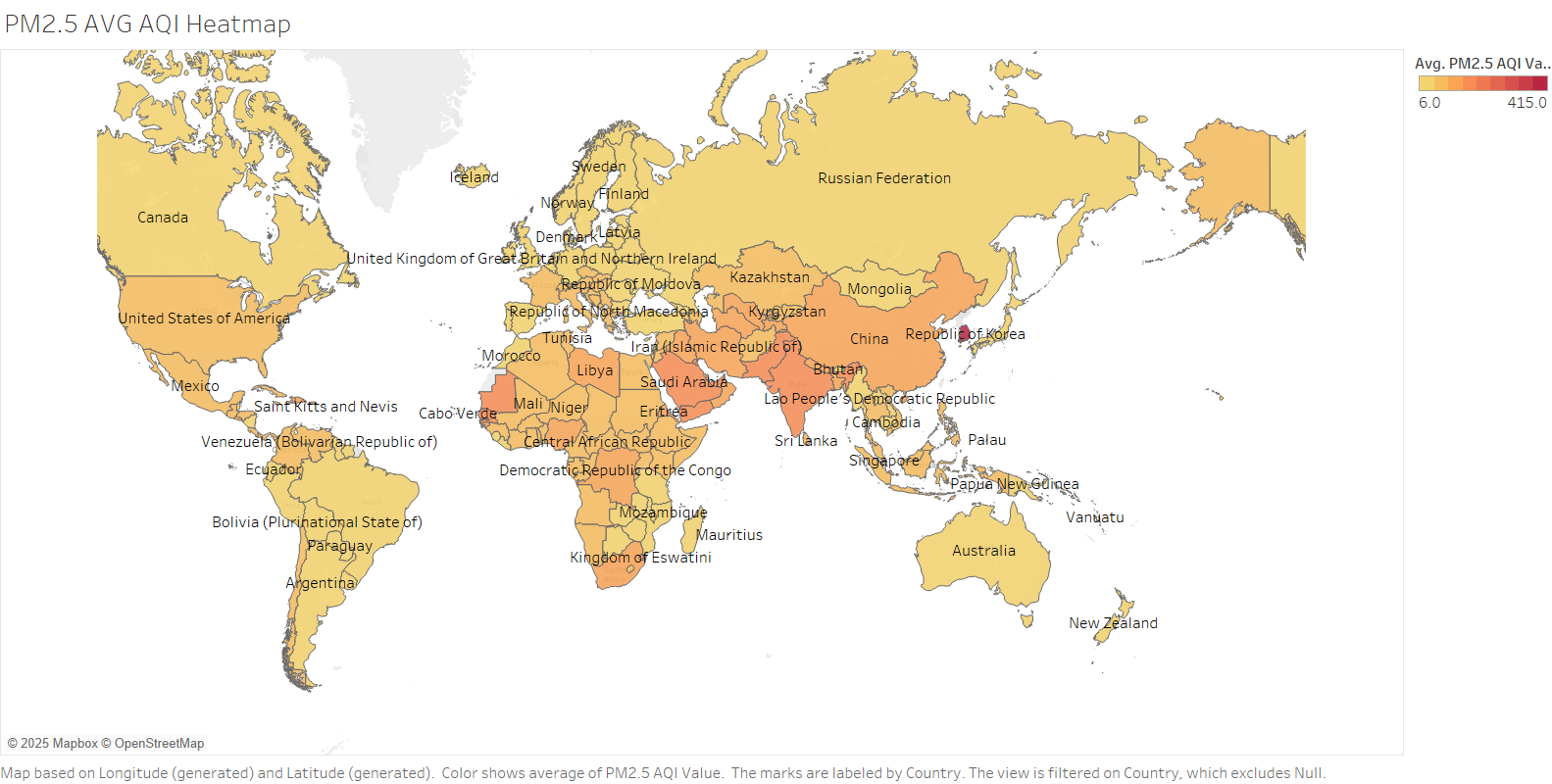


**Figure 16**

*Count of Cities in Countries with Hazardous PM2.5 AQI Values*

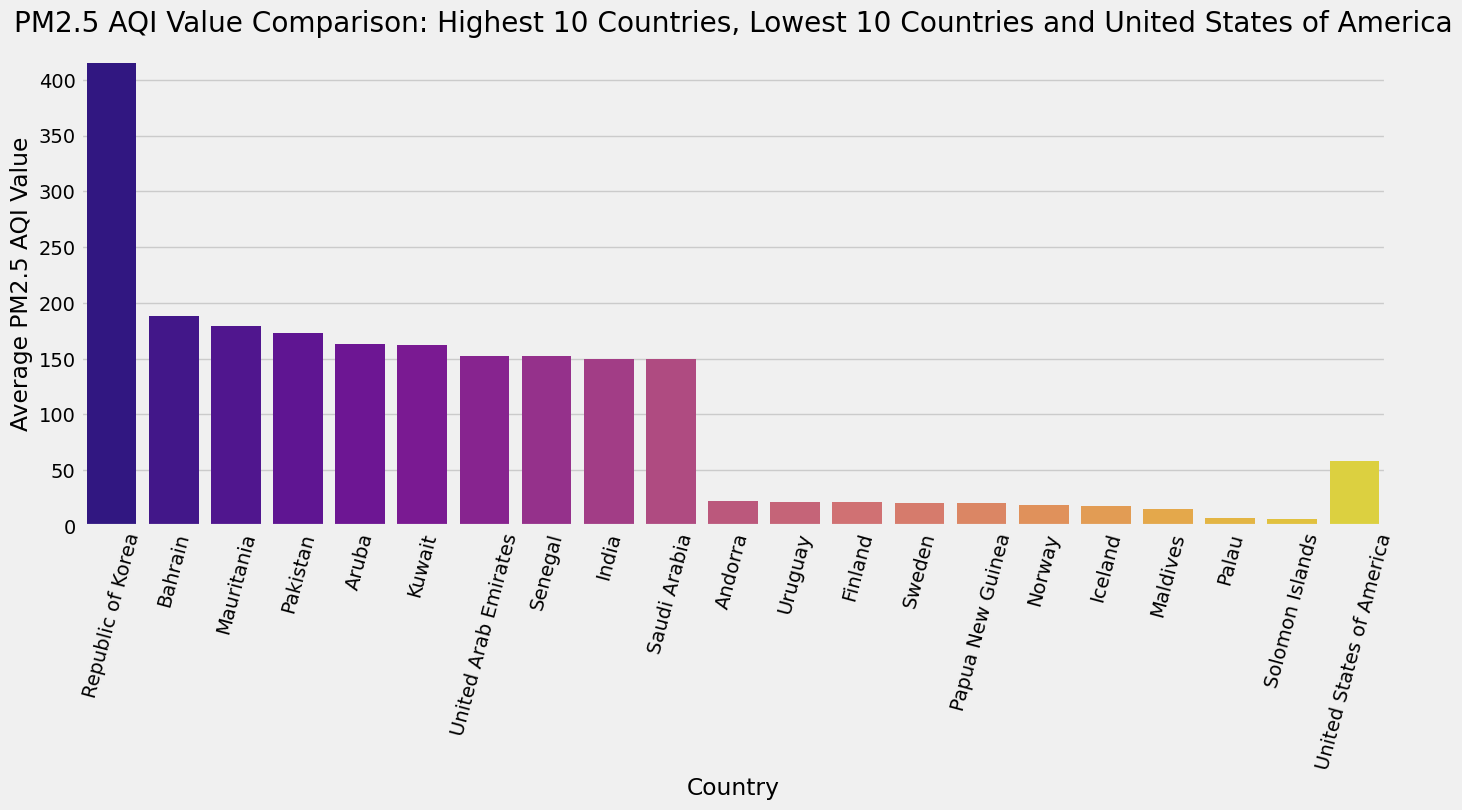


**Figure 17**

*Tableau Global Heatmap Showing the Average PM2.5 AQI Values for different countries*

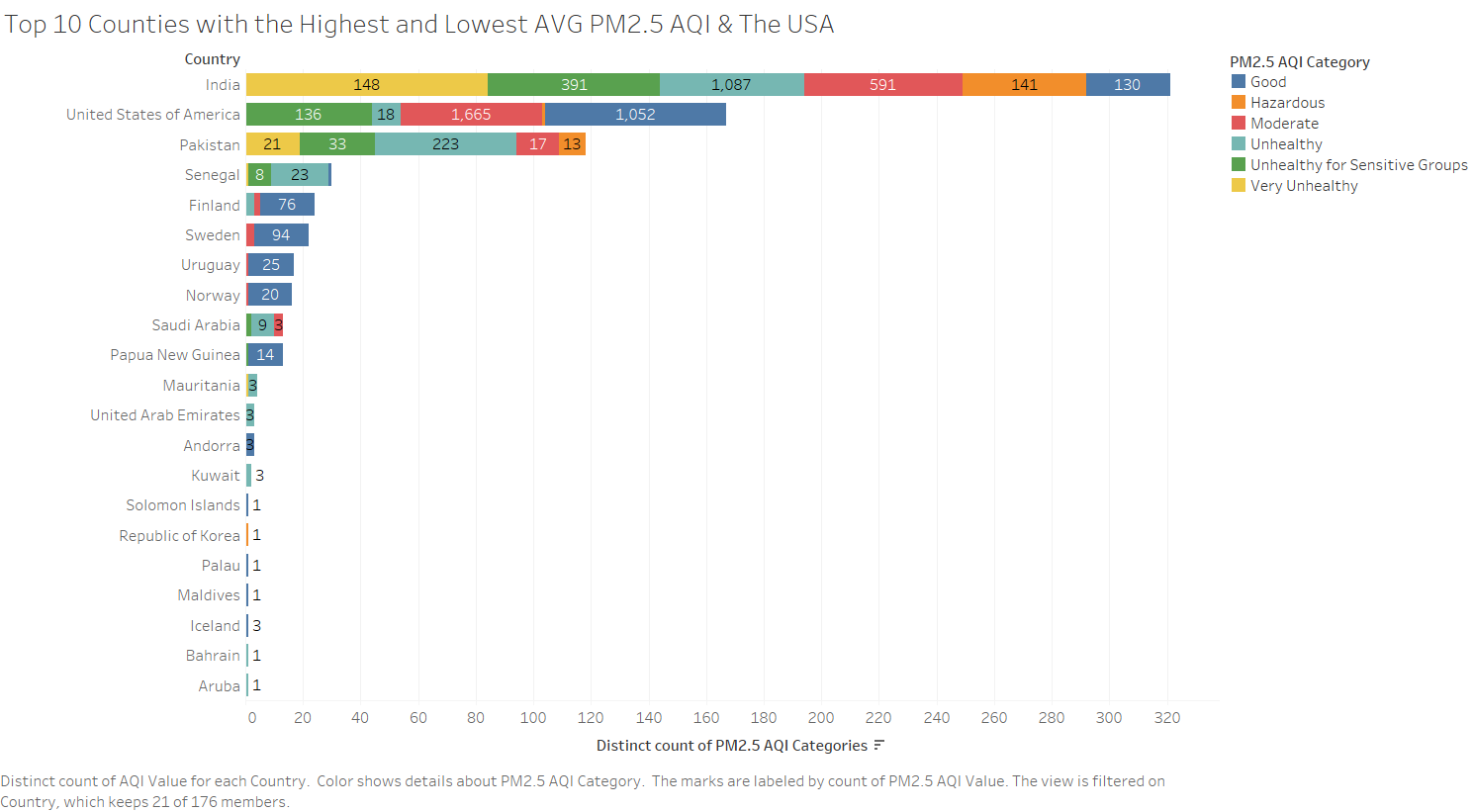
**Figure 18.1**

*Average PM2.5 AQI Values in the original data set from the Highest 10 Countries, Lowest 10 Countries, & the USA*



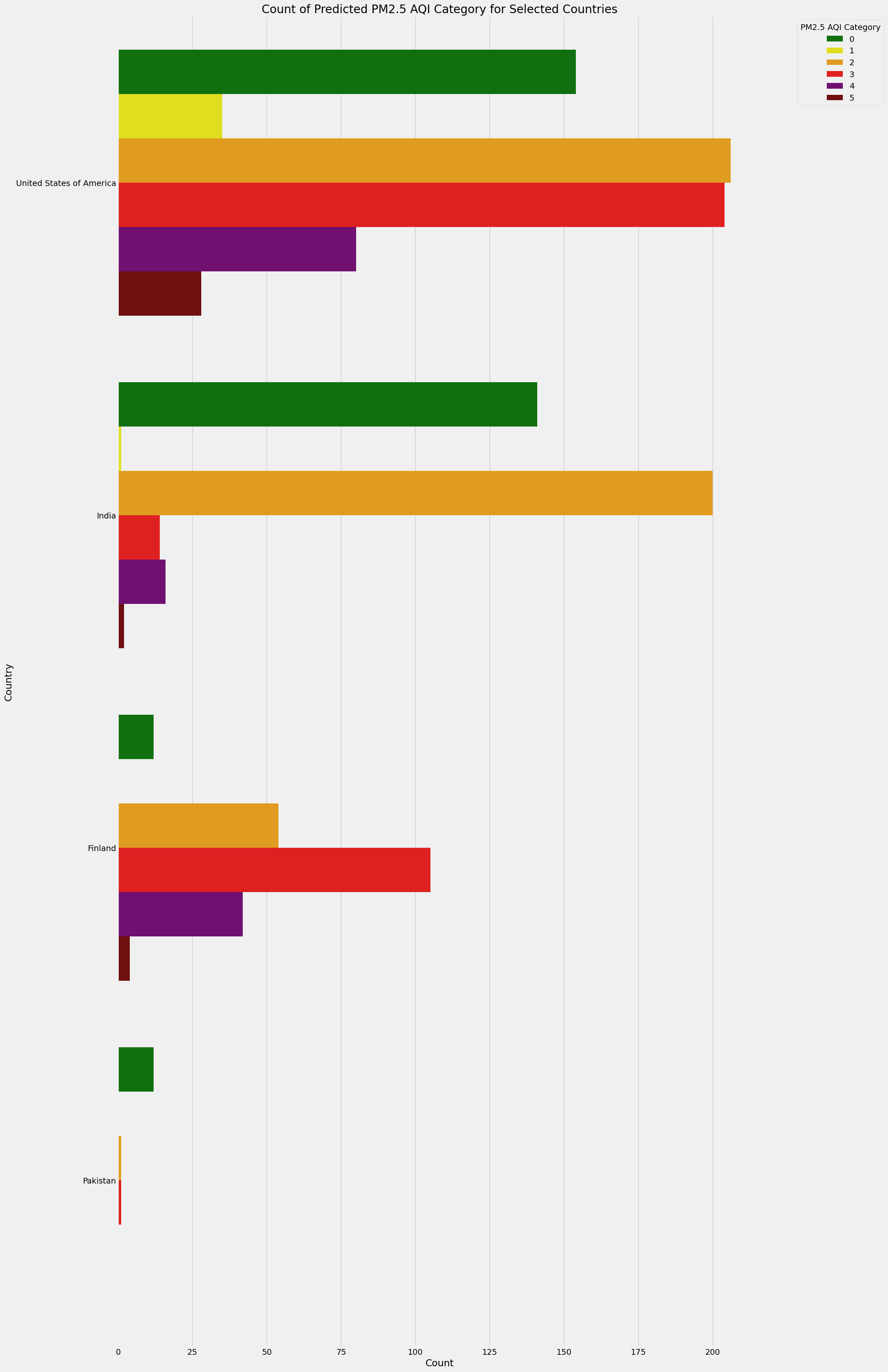
**Figure 18.2**

*Count of PM2.5 AQI Categories for the Top 10 Countries with the Highest and Lowest Average PM2.5 AQI, & the USA*



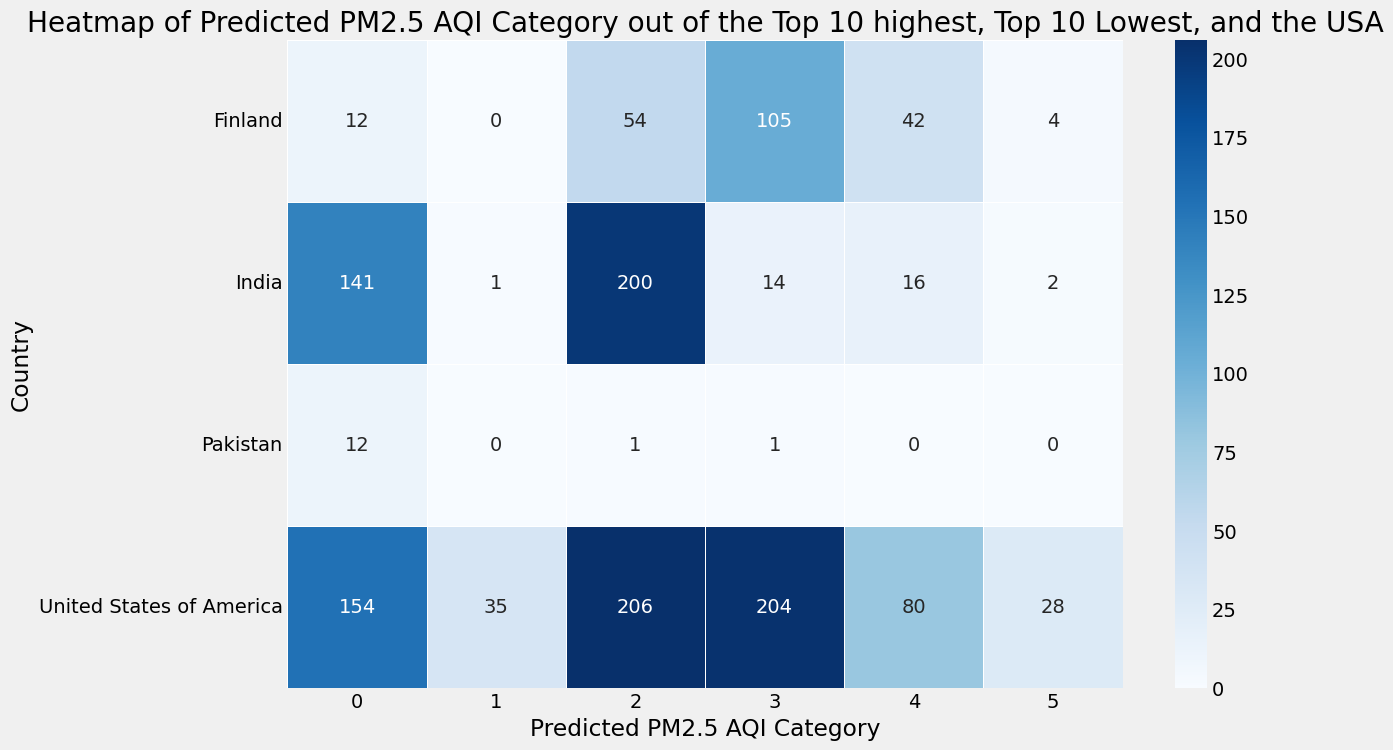
**Figure 18.3**

*SVM Model # 1 Count of Predicted PM2.5 AQI Category for USA, India, Finland, & Pakistan*



**Figure 18.4**

*SVM Model # 1 Heatmap of Predicted PM2.5 AQI Category for USA, India, Finland, & Pakistan*

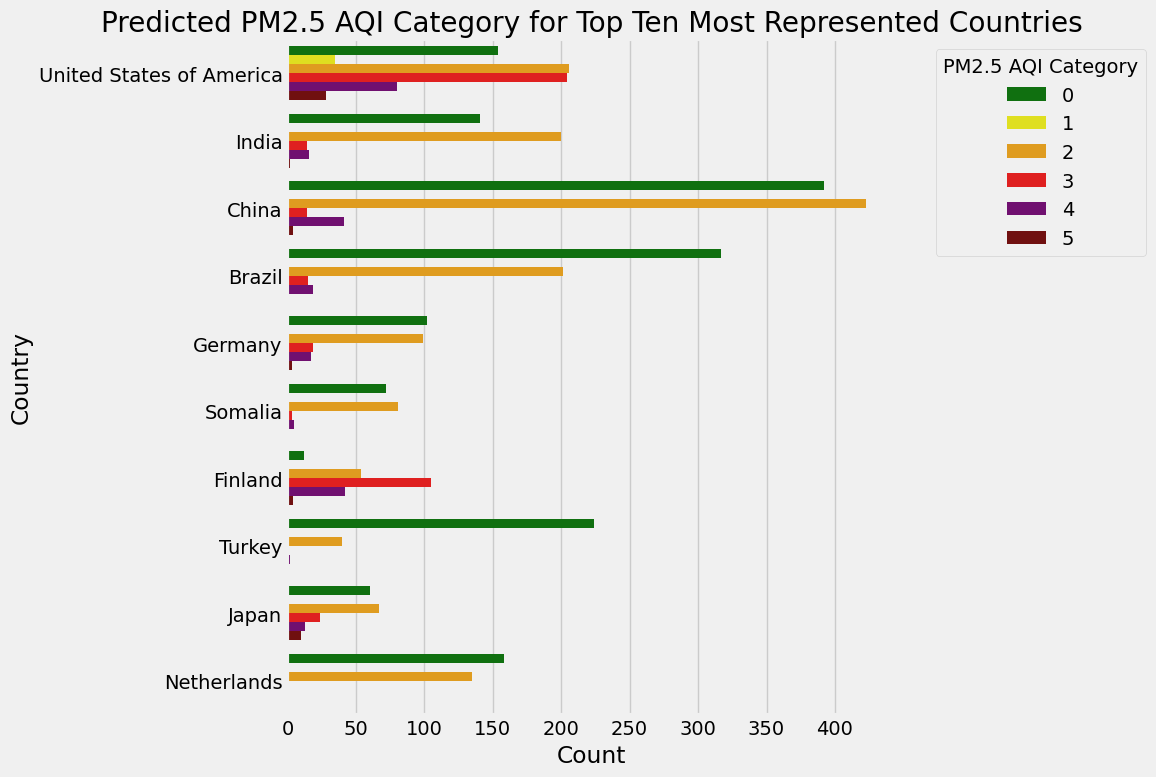


**Figure 19**

*SVM Model #1 Predicted PM2.5 AQI Category for All Countries*

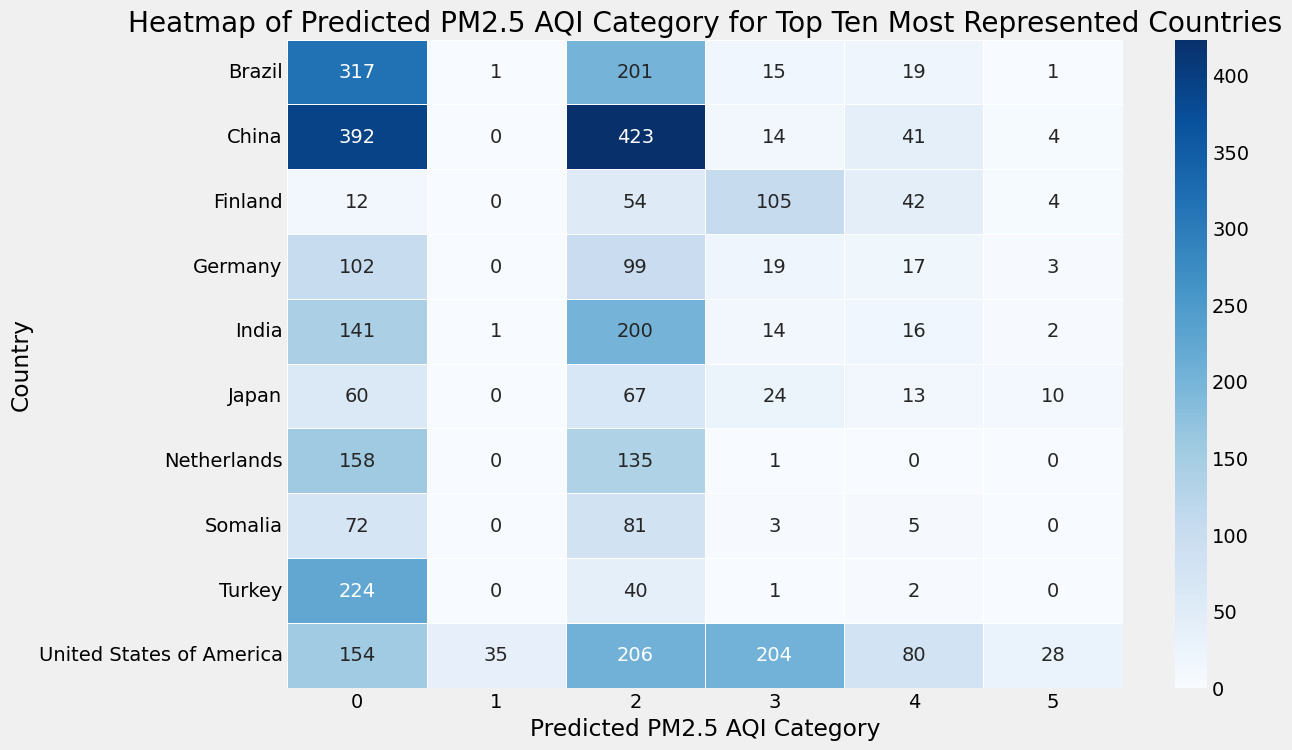
**Figure 20.1**

*SVM Model # 1 Predicted PM2.5 AQI Category for Top Ten Most Represented Countries in the Test Data*



**Figure 20.2**

*SVM Model # 1 Heatmap of Predicted PM2.5 AQI Category for Top Ten Most Represented Countries in the Test Data*



# References

Abluwalia, H. (2024, March 19). Pair plots in machine learning. <https://www.analyticsvidhya.com/blog/2024/02/pair-plots-in-machine-learning/>

Agrawal, R. (2024, December 16). Exploratory data analysis using data visualization techniques. <https://www.analyticsvidhya.com/blog/2021/06/exploratory-data-analysis-using-data-visualization-techniques/>

AirNow. (n.d.). Air quality index (AQI) basics. Retrieved January 8, 2025, from <https://www.airnow.gov/aqi/aqi-basics/>

AirNow. (n.d.). Using the air quality index. Retrieved January 8, 2025, from <https://www.airnow.gov/aqi/aqi-basics/using-air-quality-index/>

Al-Taie, M, Salim, N. & Obasa, A. (2017). Successful data science projects: Lessons learned from kaggle competition. *Kurdistan Journal of Applied Research*, *2*(3), 40–49. <https://doi-org.ezproxy.umgc.edu/10.24017/science.2017.3.18>

Amann, M., Kiesewetter, G., Schöpp, W., Klimont, Z., Winiwarter, W., Cofala, J., Rafaj, P., Höglund-Isaksson, L., Gomez-Sabriana, A., Heyes, C., Purohit, P., Borken-Kleefeld, J., Wagner, F., Sander, R., Fagerli, H., Nyiri, A., Cozzi, L., & Pavarini, C. (2020). Reducing global air pollution: the scope for further policy interventions. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical, and Engineering Sciences*, *378*(2183), 1–27. <https://doi.org/10.1098/rsta.2019.0331>

Antonenko, D. (2022, July 31). Big data privacy and security challenges: What you need to know. <https://www.businesstechweekly.com/operational-efficiency/data-management/big-data-privacy-and-security-challenges/>

AQI. (2024, June 28). How wildfires affect air quality: Recent global case examples. <https://www.aqi.in/blog/wildfires-affect-air-quality-recent-case/>

Atiqah, S. N. (2024). Linear regression of global air pollution [Notebook]. Kaggle. <https://www.kaggle.com/code/serinuratiqah/linear-regression-of-global-air-pollution>

Berrisford, L. J., Barbosa, H., & Menezes, R. (2024). A data-driven supervised machine learning approach to estimating global ambient air pollution concentrations with associated prediction intervals. <https://arxiv.org/pdf/2402.10248>

Brownlee, J. (2021, April 27). One-vs-rest and one-vs-one for multi-class classification. Retrieved from <https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/>

btd. (2023, November 10). Model fitness check: 11 Evaluation metrics to identify overfitting and underfitting in ML models. Medium. <https://baotramduong.medium.com/machine-learning-overfitting-vs-underfitting-96fe8b41192b>

Burke, J. (2023, September 25). Why and how to use Google colab. TechTarget. <https://www.techtarget.com/searchEnterpriseAI/tutorial/Why-and-how-to-use-Google-Colab>

Comma-separated values. (2024, December 21). In *Wikipedia.* <https://en.wikipedia.org/wiki/Comma-separated_values>

eLichens. (2024). Global air quality map. <https://www.elichens.com/global-air-quality-map>

eLichens. (2024). Our products. <https://www.elichens.com/our-products>

eLichens. (2025). eLichens elos | Outdoor air quality station. <https://www.elichens.com/elos-outdoor-air-quality-station>

EPA. (2024, June 20). Particulate matter (PM) basics. <https://www.epa.gov/pm-pollution/particulate-matter-pm-basics>

EPA. (2024, June 21). Basic information about carbon monoxide (CO) outdoor air pollution. <https://www.epa.gov/co-pollution/basic-information-about-carbon-monoxide-co-outdoor-air-pollution#What%20is%20CO>

EPA. (2024, July 16). Basic information about NO2. <https://www.epa.gov/no2-pollution/basic-information-about-no2#What%20is%20NO2>

Fraj, M. B. (2018, January 18). In depth: Parameter tuning for SVC. *Medium.* Retrieved from <https://medium.com/all-things-ai/in-depth-parameter-tuning-for-svc-758215394769>

French, T. (2020, June 10). What is Microsoft excel and what does it do? Lifewire. <https://www.lifewire.com/what-is-microsoft-excel-3573533>

GeeksforGeeks. (2023, May 07). Support vector machine in machine learning. <https://www.geeksforgeeks.org/support-vector-machine-in-machine-learning/>

GeeksforGeeks. (2024, February 29). How to tune hyperparameters in gradient boosting algorithm. Retrieved from <https://www.geeksforgeeks.org/how-to-tune-hyperparameters-in-gradient-boosting-algorithm/>

GeeksforGeeks. (2024, March 12). Box plot. <https://www.geeksforgeeks.org/box-plot/#what-is-a-box-plot>

GeeksforGeeks. (2024, April 16). How to tune a decision tree in hyperparameter tuning. Retrieved from <https://www.geeksforgeeks.org/how-to-tune-a-decision-tree-in-hyperparameter-tuning/>

GeeksforGeeks. (2024, July 4). Tableau tutorial. <https://www.geeksforgeeks.org/tableau-tutorial/>

GeeksforGeeks. (2025, January 16). Decision tree. <https://www.geeksforgeeks.org/decision-tree/>

GeeksforGeeks. (2025, January 16). Random forest algorithm in machine learning. <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>

Goyal, S. (2021, July 20). Evaluation metrics for classification models. Medium. <https://medium.com/analytics-vidhya/evaluation-metrics-for-classification-models-e2f0d8009d69>

Hasenkopf, C., Sharma, N., Kazi, F., Mukerjee, P., & Greenstone, M. (2023). The case for closing global air quality data gaps with local actors: A golden opportunity for the philanthropic community. *Energy Policy Institute at the University of Chicago (EPIC)*. <https://epic.uchicago.edu/wp-content/uploads/2023/12/The-Case-for-Closing-Global-Air-Quality-Data-Gaps-with-Local-Actors.pdf>

Health Effects Institute. (2024). *State of global air 2024. Special Report.* <https://www.stateofglobalair.org/resources/report/state-global-air-report-2024>

Hunter, P. (2020). The health toll of air pollution: Despite global efforts to clean up the air, outdoor and indoor air pollution still have a drastic negative effect on public health. *EMBO Reports*, *21*(8), 1-4. <https://doi.org/10.15252/embr.202051183>

Imbalanced-learn. (n.d.). Smotetomek. Retrieved February 9, 2025, from <https://imbalanced-learn.org/stable/references/generated/imblearn.combine.SMOTETomek.html>

Joy, A. P. (2025, February 5). A beginner’s guide to linear regression: Understanding the fundamentals. <https://www.skillcamper.com/blog/a-beginners-guide-to-linear-regression-understanding-the-fundamentals#What-is-Linear-Regression>

Kaplan, S. (2024, February 22). Can machine learning models give an accuracy of 100??[The truth]. <https://enjoymachinelearning.com/blog/can-machine-learning-models-give-an-accuracy-of-100/>

machinelearningplus. (2025). An introduction to gradient boosting decision trees. <https://www.machinelearningplus.com/machine-learning/an-introduction-to-gradient-boosting-decision-trees/>

Microsoft Bing. (2025). Copilot (January 11 version) [Large Language Model]. <https://copilot.microsoft.com/chats/ayaeR4f2F6CAXmMaWGYyo>

Microsoft Bing. (2025). Copilot (February 6 version) [Large Language Model]. <https://copilot.microsoft.com/chats/wNLBCongu7CuWEHFYWDp7>

Microsoft Bing. (2025). Copilot (February 9 version) [Large Language Model]. <https://copilot.microsoft.com/chats/P8UWADwLhjBGaV89k7fX8>

Microsoft Bing. (2025). Copilot (February 21 version) [Large Language Model]. <https://copilot.microsoft.com/chats/eWKXAKQGwNzDS6tAqmCtT>

Muzdadid, H. A. (2022). *Global air pollution dataset* [Data set]. Kaggle. <https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset/data>

National Geographic Society. (n.d.). Air pollution. In *National Geographic Encyclopedia*. Retrieved January 12, 2025, from <https://education.nationalgeographic.org/resource/air-pollution/>

ODC. (n.d.). ODC principles. Retrieved February 20, 2025, from <https://opendatacharter.org/principles/>

OpenAQ. (n.d). OpenAQ API docs. Retrieved February 20, 2025, from <https://docs.openaq.org/>

OpenAQ, & Dane, S. (n.d.). *OpenAQ Global air pollution measurements* [Data set]. Kaggle. Retrieved February 20, 2025, from <https://www.kaggle.com/datasets/open-aq/openaq>

OpenAQ. (2025). Fighting air inequality through open data. <https://openaq.org/>

OpenAQ. (2025). Platform overview. <https://openaq.org/developers/platform-overview/>

Opendatasoft. (n.d.). *World air quality- OpenAQ* [Dataset]. Retrieved February 20, 2025, from <https://data.opendatasoft.com/explore/dataset/openaq%40public/table/?disjunctive.measurements_parameter&disjunctive.location&disjunctive.city&sort=measurements_lastupdated>

Raden, N., & Klein, A. (2020, March 9). *Risks and ethical issues with predictive analytics and artificial intelligence* [PowerPoint slides]. CAS ERM Symposium. <https://www.casact.org/sites/default/files/2021-08/CS18-Risks_and_Ethical_Issues_with_Predictive_Analytics.pdf>

Reliance Jio. (2024, February 7). Importance of Boxplot in Exploratory Data Analysis (EDA). Medium. <https://medium.com/@venugopal.adep/importance-of-boxplot-in-exploratory-data-analysis-eda-5119944850a3>

Rowe, S. (n.d.) What is a treemap? Retrieved January 30, 2025, from <https://www.storytellingwithdata.com/blog/what-is-a-treemap>

Sampaio, C. (2023, July 2). Implementing SVM and kernal SVM with python’s scikit-learn. *StackAbuse*. Retrieved from <https://stackabuse.com/implementing-svm-and-kernel-svm-with-pythons-scikit-learn/>

Sampaio, C. (2023, July 2). Understanding SVM hyperparameters. *StackAbuse*. Retrieved from <https://stackabuse.com/understanding-svm-hyperparameters/>

Sawant, V., Hagerbaumer, C., Rosales, C. M. F., Isied, M., & Biggs, R. (2022). Open air quality data: The global landscape*. OpenAQ.* <https://documents.openaq.org/reports/Open+Air+Quality+Data+Global+Landscape+2022.pdf>

Sci-kit learn. (n.d.). LinearRegression. Retrieved February 8, 2025, from <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>

Scikit learn. (n.d.). sklearn.ensemble.GradientBoostingClassifier. Retrieved February 9, 2025, from <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html>

Scikit-learn. (n.d.). sklearn.ensemble.RandomForestClassifier. Retrieved February 9, 2025, from <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

SciKitlearn. (n.d.). SVC. Retrieved February 9, 2025, from <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC>

Seaborn. (n.d.). An introduction to seaborn. Retrieved January 27, 2025, from <https://seaborn.pydata.org/tutorial/introduction.html>

Seaborn. (n.d.). Choosing color palettes. Retrieved January 26, 2025, from <https://seaborn.pydata.org/tutorial/color_palettes.html>

Singh, K. (2022, May 25). More than 100 acres now burning near Durango. *Nexstar Media Group, Inc.* <https://www.fox21news.com/top-stories/more-than-100-acres-now-burning-near-durango/>

Singla, S. (2024, January 23). Different plot used in exploratory data analysis (EDA). <https://www.comet.com/site/blog/different-plots-used-in-exploratory-data-analysis-eda/>

Spencer, A. (2022, April 9). Minute EDA: Correlation heatmap. Medium. <https://medium.com/5-minute-eda/5-minute-eda-correlation-heatmap-b57bbb7bae14>

Stack Overflow. (n.d.). Which decision function shape for sklearn.svm.SVC when using onevsrestclassifier? Retrieved February 9, 2025, from <https://stackoverflow.com/questions/43505451/which-decision-function-shape-for-sklearn-svm-svc-when-using-onevsrestclassifier>

State of Global Air. (2025). How we estimate exposure. <https://www.stateofglobalair.org/data/estimate-exposure>

Stodden, V. (2020). The data science life cycle: a disciplined approach to advancing data science as a science. Communications of the ACM, 63(7), 58–66. <https://doi-org.ezproxy.umgc.edu/10.1145/3360646>

Tableau. (2025). What is a tree map? Understanding and using tree maps. <https://www.tableau.com/chart/what-is-treemap>

Using the air quality index. (n.d.) AirNow. Retrieved January 8, 2025, from <https://www.airnow.gov/aqi/aqi-basics/using-air-quality-index/>

World Health Organization. (2021, September 22). *WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Executive summary*. <https://iris.who.int/bitstream/handle/10665/345334/9789240034433-eng.pdf>

World Health Organization. (2024, January 31). Air quality database updates 2024 (Version Sixth). <https://www.who.int/data/gho/data/themes/air-pollution/who-air-quality-database>

World Health Organization. (2024, October 24). Ambient (outdoor) air pollution*.* <https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health>

World Health Organization. (2025). Air pollution. <https://www.who.int/health-topics/air-pollution#tab=tab_1>

Worldometer. (n.d.). Countries in the world. Retrieved January 24, 2025, from <https://www.worldometers.info/geography/how-many-countries-are-there-in-the-world/>

Worldometer (2025). Countries in the world by population (2025). <https://www.worldometers.info/world-population/population-by-country/>

World Population Review. (2024). How many cities are in the United States? 2024. <https://worldpopulationreview.com/us-city-rankings/how-many-cities-are-in-the-us>

World Population Review. (2024). India cities by population 2024. <https://worldpopulationreview.com/cities/india>

World Population Review. (2024). Senegal cities by population 2024. <https://worldpopulationreview.com/cities/senegal>