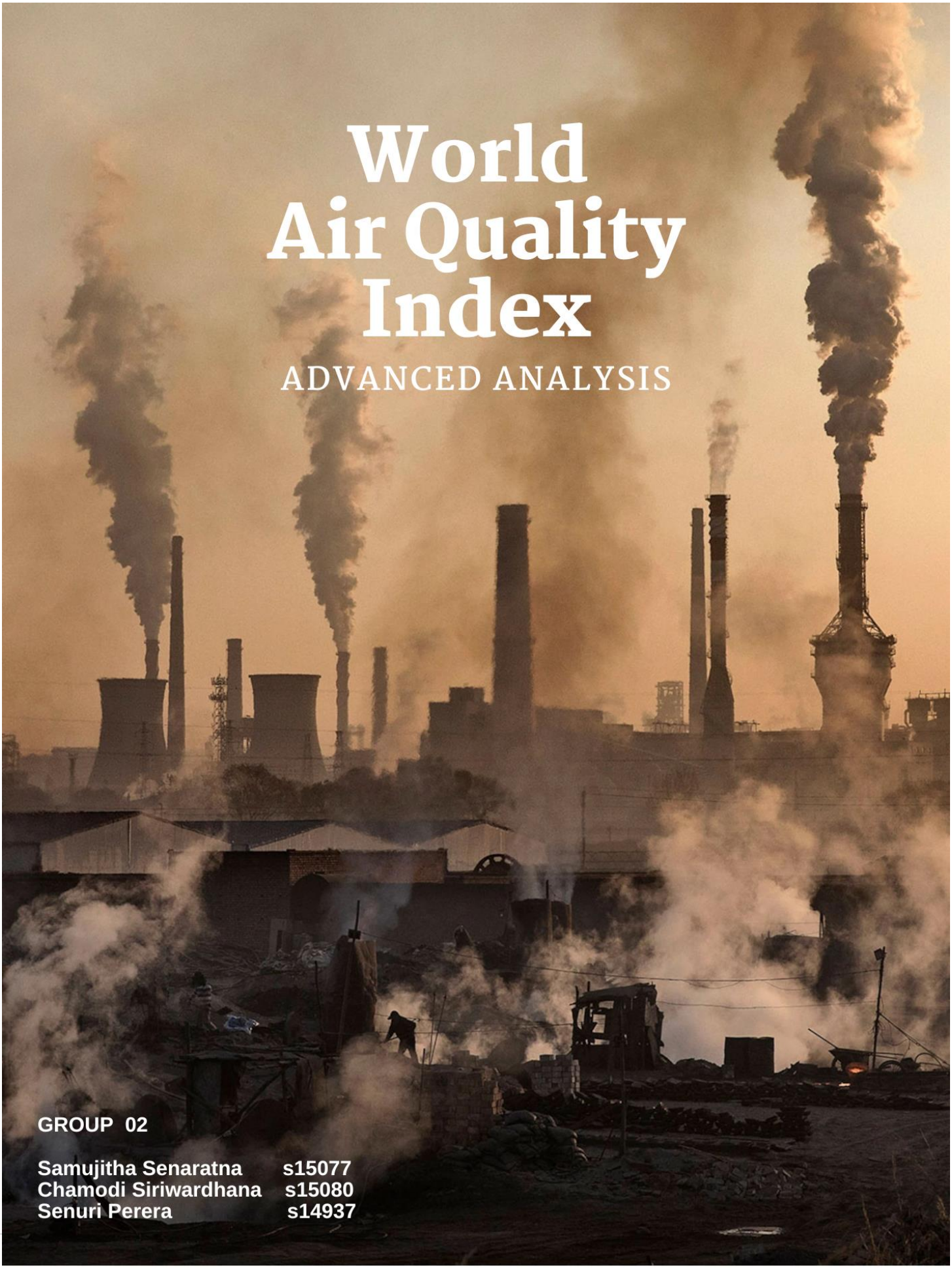


World Air Quality Index

ADVANCED ANALYSIS

GROUP 02

Samujitha Senaratna s15077
Chamodi Siriwardhana s15080
Senuri Perera s14937

ABSTRACT

The prediction of the AQI category holds immense importance and offers numerous advantages in the environmental and public health domains. By accurately predicting the AQI category, we gain valuable insights into the air quality and potential health risks associated with varying pollutant levels. This knowledge empowers policymakers, environmental agencies, and the public to take timely and targeted actions to mitigate pollution and protect public health. This study focuses on predicting the AQI category through a detailed descriptive analysis. The aim is to comprehensively evaluate the factors influencing the AQI category. The analysis involves statistical data representation techniques, numerical summaries, and comparisons between predictor variables and the response variable to establish significance and relationships. To uncover potential patterns or clusters within the data, partial least squares regression (PLSR) is implemented, and Loadings plots are used to identify correlations between predictors and the response variable. VIP plots are utilized to identify less important variables, enhancing model simplicity and preventing overfitting. The findings from this analysis offer valuable insights for more advanced analysis and model building in predicting AQI category accurately.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

Air quality is a fundamental prerequisite for sustaining life on Earth, with direct and far-reaching implications for human health, environmental preservation, and overall well-being. However, the global challenge of air pollution has emerged as a pressing concern, affecting countless individuals worldwide. As such, recognizing the profound significance of world air quality is very important to ensuring the protection of human health, safeguarding the environment, and fostering sustainable development.

AQI Score Range	AQI Category
0-50	Good
51-100	Moderate
101-150	Unhealthy
151-200	Unhealthy
201-300	Very Unhealthy
301+	Hazardous

The Air Quality Index (AQI) serves as a crucial tool in assessing and communicating the quality of the air we breathe. It provides a standardized scale for evaluating the concentration of various air pollutants and their potential health impacts. It is important to note that the AQI scale may vary from one country to another due to the inclusion of different air pollutants in different regions. For instance, India incorporates measurements of ammonia levels in addition to other pollutants. To make these readings more accessible, the AQI has a scoring system that runs from 0 to 500, using data collected from air monitoring stations in cities around the world. Scores below 50 are considered good, with very little impact to human health. The higher the score gets, the worse the air quality is.

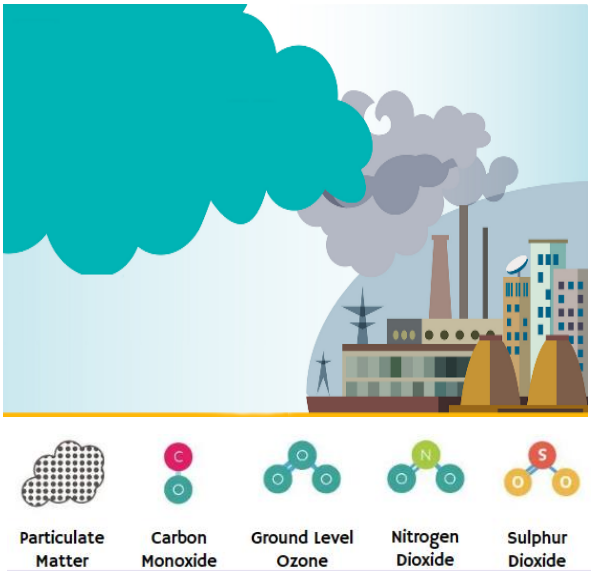
In the United States, the AQI is calculated based on the concentrations of five major air pollutants that are regulated under the Clean Air Act. These pollutants are:

1. Ground-level ozone
2. Carbon monoxide
3. Sulfur dioxide
4. Particle pollution (Particulate Matter)
5. Nitrogen dioxide

The primary goal of this advanced analysis is to create an effective predictive model for determining the Air Quality Index (AQI) category in different regions worldwide since it is very crucial due to the substantial health risks posed by air pollution, including respiratory and cardiovascular diseases. Also areas with high pollution levels experience increased premature mortality rates as well.

2. PROBLEM STATEMENT

Air pollution poses a significant threat to public health, with pollutants like PM2.5, Ozone, CO, NO2 and SO2 causing respiratory and cardiovascular diseases. Vulnerable populations suffer the most, and premature mortality rates escalate in regions with high pollution levels. Therefore, assessing and understanding the level of air pollution in different regions around the world, and identifying the significant relationships between the air pollutants with the AQI category is utmost important. Thus, the objective of this analysis is to develop the optimal predictive model that predicts the AQI category of a location given its important features.



3. DATA PRE-PROCESSING

Since there was a limited number of observations within the categories "Unhealthy for Sensitive Groups," "Very Unhealthy," "Unhealthy," and "Hazardous", these categories were lumped together and were named as “Unhealthy”.

There were 302 missing values in the ‘Country’ variable. It was imputed using the city name from ‘City’ variable and the Geopy library in Python with the help of GeoNames web database.

A new variable named ‘Continent’ was created by categorizing each country to its respective Continent.

Web-scraping was used to obtain the SO2 AQI Value, which was not available in the original dataset. AccuWeather, which is a reputable source of weather and air quality information, was utilized for the web-scraping task. Selenium, a popular tool in web automation, was utilized to interact with the web page dynamically. XPath, on the other hand, is a query language used to navigate through the structure of an XML or HTML document. Note that the dataset that we originally obtained, has not mentioned any date of data collection. Thus, by the process of including the SO2 data, we generalize the dataset to the present.

4. DESCRIPTION OF THE DATASET

The 'World Air Quality Index by City and Coordinates' dataset was acquired from the Kaggle website. It contains 16695 records under 14 variables, where the response variable is 'AQI Category' (categorical).

Variable	**Type**	**Description**
Country	Character	Name of the Country
City	Character	Name of the City
CO AQI Value	Integer	The AQI value of Carbon Monoxide
CO AQI Category	Factor	The AQI category of Carbon Monoxide
Ozone AQI Value	Integer	The AQI value of Ozone
Ozone AQI Category	Factor	The AQI category of Ozone
NO2 AQI Value	Integer	The AQI value of Nitrogen Dioxide
NO2 AQI Category	Factor	The AQI category of Nitrogen Dioxide
PM2.5 AQI Value	Integer	Fine particulate matter less than 2.5 micrometers in diameter value
PM2.5 AQI Category	Factor	Fine particulate matter less than 2.5 micrometers in diameter category
lat	Float	Latitude value of the city
lng	Float	Longitude value of the city
AQI Value	Integer	Overall air quality index value
AQI Category	Factor	Overall air quality index category with respect to the AQI score range.

Table 1: About the dataset

5. IMPORTANT RESULTS OF THE DESCRIPTIVE ANALYSIS

- Class 'Unhealthy' constitutes a relatively lower percentage, accounting for only 11.6% of the total observations.
- There exists moderate multicollinearity between several explanatory variables AQI value.
- From the Spearman's Correlation we have found that PM2.5 AQI value is highly correlated with AQI category.
- Also from the Partial Least Squares Analysis, PM2.5 shows a massive importance to the response AQI category.
- Asia and Africa continents have the worst air quality out of all the continents in the World. The variable importance plot guarantees this result by showing higher importance than the other continents.
- Ozone also shows a moderate importance to our classification.
- From the partial least squares analysis, it was seen that moderately seperable clusters were identified. So, can consider some linear classification algorithms.

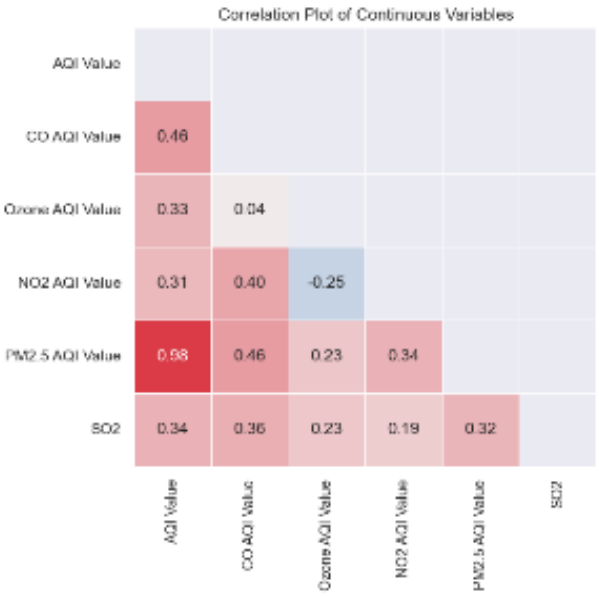


Figure 1: Correlation Plot of Continous Varaibles

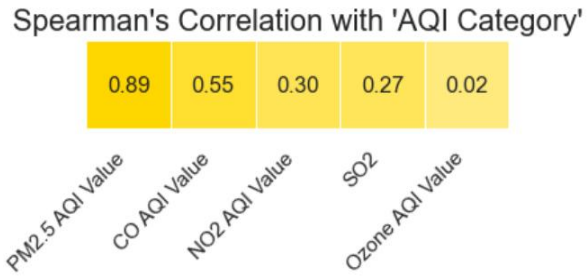


Figure 2: Spearman's Correlation

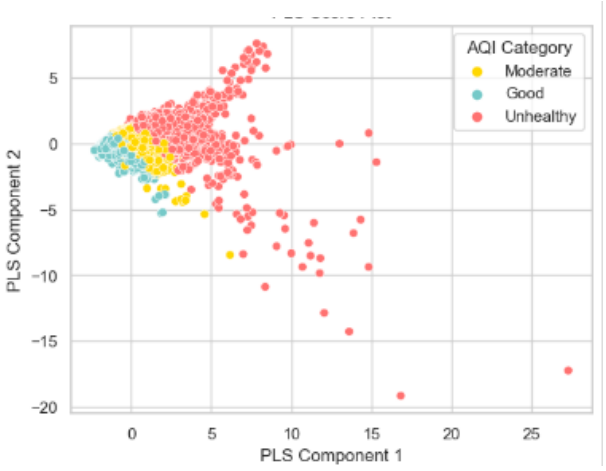


Figure 3: PLS Score Plot

6. IMPORTANT RESULTS OF THE ADVANCED ANALYSIS

Proportional odds model

The proportional odds model is a class of generalized linear models used for modeling the dependence of an ordinal response on discrete or continuous covariates. In this analysis, the primary aim is to predict an ordinal response variable—the AQI (Air Quality Index) category, comprising three ordered outcomes: "good," "moderate," and "unhealthy."

To achieve this, Proportional Odds Logistic Regression is employed as the initial modeling approach, utilizing AQI category as the response variable and considering the explanatory variables CO.AQI.Value, Ozone.AQI.Value, NO2.AQI.Value, PM2.5.AQI.Value, SO2, and Continent.

It is essential to bear in mind that this methodology relies on the assumption of proportional odds, signifying that the influence of these predictors remains consistent across different AQI categories. Additionally, the analysis must ensure the absence of multicollinearity among the explanatory variables to maintain the model's robustness and reliability. The summary of the fitted model is as follows.

	Value	p_value	odds ratio
CO . AQI . Value	0.29080494	3.111860e- 04	1.337504 +00
Ozone . AQI . Value	0.11647021	0	1.123524e +00
NO2.AQI . Value	0.12150697	0	1.129197e + 00
PM2.5.AQI.Value	0.42508834	0	1.529726e +00
so2	-0.04009425	3.706211e - 03	9.606989e - 01
ContinentAsia	-0.85499574	3.118521e - 05	4.252850e - 01
ContinentEurope	-1.67242973	0	1.877902e - 01
ContinentNorthAme	-1.56084959	4.884981e - 15	2.099576e - 01
ContinentOceania	-0.93960868	1.450243e - 01	3.907807e - 01
ContinentSouthAme	0.38344083	8.354357e - 02	1.467325e + 00
Good | Moderate	23.69574020	0	1.954021e + 10
Moderate |Unhealth	45.85272894	0	008.195721e + 19

Interpreting the odds ratio

- CO.AQI.Value :

  For every one unit increase in CO.AQI value, the odds of being in a higher category are associated with a 1.337 times increase (that is from 33.7 %) holding constant all other variables.

- Continent South America :

  For those belonging to South America, the odds of being higher AQI Category increase 2.85 times that of countries that do not belong to South America, holding constant all other variables.

Below present the key performance metrics on both training and testing datasets.

	Accuracy	Precision	Recall	F1-score
Train	0.9548	0.9528	0.9401	0.9464
Test	0.9518	0.9421	0.946	0.944

The model demonstrates strong predictive power without overfitting, aligning well with the goal of accurate predictions for unseen data.

Testing the proportional odds assumption

The analysis of the Proportional Odds Logistic Regression model involved testing the proportional odds assumption. That is that the relationship between predictors and the ordered categories of the response variable is assumed to be constant across all category transitions, implying consistent odds ratios.
This was done by creating binary variables for various category transitions within the response variable and fitting separate binary logistic regression models for each transition. By comparing the coefficient estimates from these models, it was assessed whether the coefficients were similar across different transitions.

	moderate	unhealthy	Difference
CO . AQI . Value	0.40059827	0.36335838	-0.037239885
Ozone.AQI . Value	0.12699389	0.12965888	0.002664994
NO2.AQI . Value	0.15339525	0.05266513	-0.100730126
PM2.5.AQI.Value	0.42480380	0.46016549	0.035361683
S02	-0.02069605	-0.12427668	-0.103580628
ContinentAsia	-0.82773784	-1.56611819	-0.738380354
ContinentEurope	-1.89898633	-2.11403066	-0.215044329
ContinentNorth America	-1.92877841	0.08535841	2.014136820
ContinentOceania	-1.29280836	6.38898421	7.681792568
ContinentSouth America	0.38307624	0.86530395	0.482227710

The above findings indicate that the coefficient differences were relatively small. This suggests that the model adheres to the proportional odds assumption, indicating consistent predictor variable effects on ordinal categories across transitions. Therefore, the model is suitable for predictive purposes, meeting the essential assumption for reliable ordinal regression analysis.

Random Forest Classifier

Random Forest Classifier is a powerful and widely used machine learning algorithm that falls under the ensemble learning category. The Random Forest algorithm is known for its robustness, versatility, and excellent performance across a wide range of datasets and applications.

A Random Forest model, employing default parameters, was trained and evaluated. The model's performance was assessed on both the test set and the training set, yielding the following results.

Performance on the Training Set			
Accuracy	Precision	Recall	F1-score
1.00	1.00	1.00	1.00

Performance on the Test Set			
Accuracy	Precision	Recall	F1-score
1.00	1.00	1.00	1.00

Table 2: Performance of Random Forest Classifier

The results obtained from the Random Forest Classifier are exceptional, reflecting a high level of accuracy and precision in classifying the Air Quality Index (AQI) into three categories: "Good," "Moderate," and "Unhealthy." For all three categories, the precision, recall, and F1-score are exceptionally high, each equal to 1.00. This indicates that the model achieved perfect classification accuracy for all classes.

Therefore, among all the fitted models, it is evident that the Random Forest Classifier stands out as the top-performing model.

Model Interpretation

Interpretability is the degree to which a human can understand the cause of a decision. The higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made. The need for interpretability arises from an incompleteness in problem formalization, which means that for certain problems or tasks it is not enough to get the prediction (the what). The model must also explain how it came to the prediction (the why), because a correct prediction only partially solves your original problem. It allows stakeholders, policymakers, and environmental experts to make informed decisions based on the model's output. For instance, if the model predicts an "Unhealthy" AQI category, interpretability helps us comprehend which specific factors, such as PM2.5 or CO levels, contributed most to that classification. The following reasons drive the demand for interpretability and explanations for our Random Forest model.

Global Model-Agnostic Methods

Global methods describe the average behavior of a machine learning model. Global methods are often expressed as expected values based on the distribution of the data.

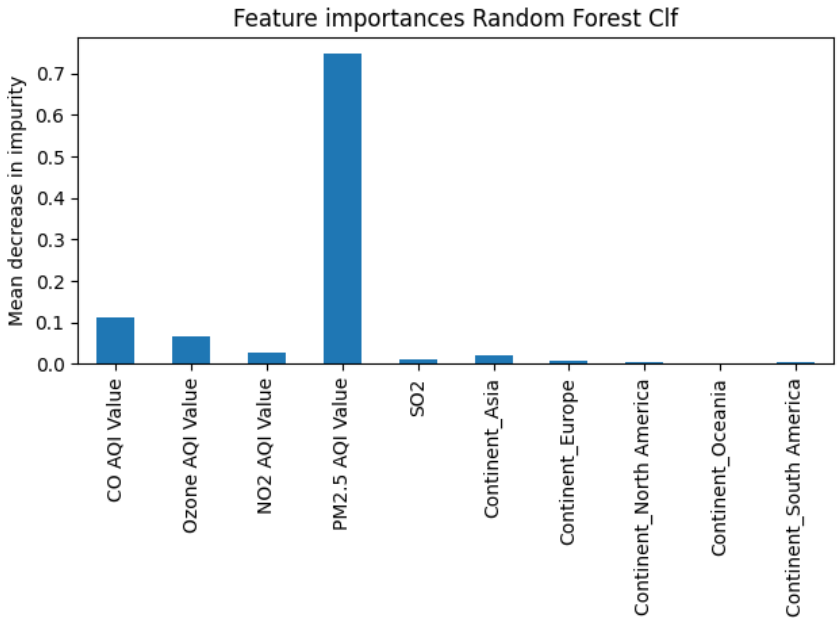
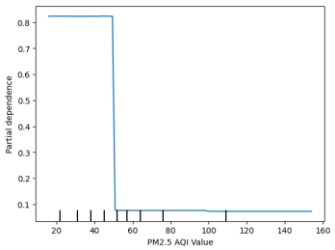
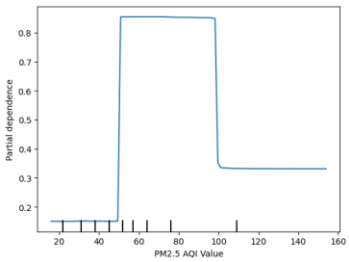
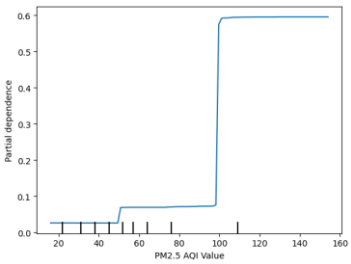


Figure 4: Feature Importance of RF

An essential aspect of our predictive model's interpretability is the examination of feature importance values. These values shed light on the relative significance of each feature in determining the Air Quality Index (AQI) categories, including "Good," "Moderate," and "Unhealthy."

The most influential feature in our model, PM2.5 AQI values, dominates the decision-making process. This underscores the paramount importance of PM2.5 particles in assessing air quality. CO, Ozone, and NO2, are the other most critical factors influencing AQI category predictions. Additionally, geographic location, represented by continent features, does have some influence on the model's classifications, although to a lesser degree. Understanding feature importance enhances our comprehension of the key determinants affecting air quality classification and informs our model's interpretability.

The partial dependence plot (short PDP or PD plot) shows the marginal effect one or two features have on the predicted outcome of a machine learning model. Let us see the PDPs for the 3 most important features of our Random Forest model.

Feature	**“Good” AQI Category**	**“Moderate” AQI Category**	**“Unhealthy” AQI Category**
PM 2.5 AQI Value			

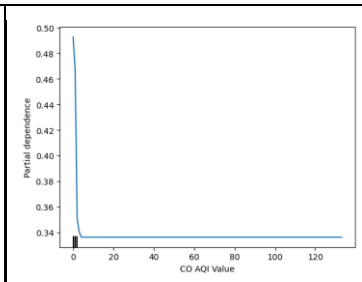
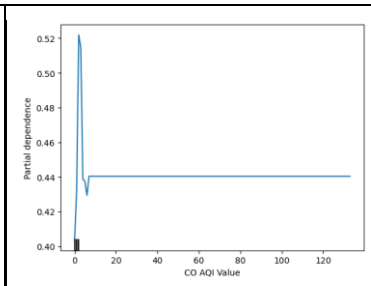
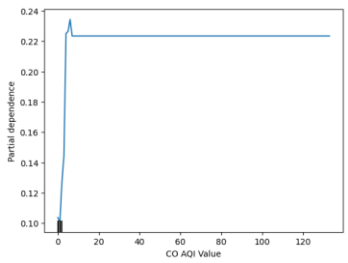
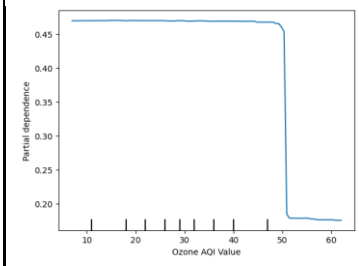
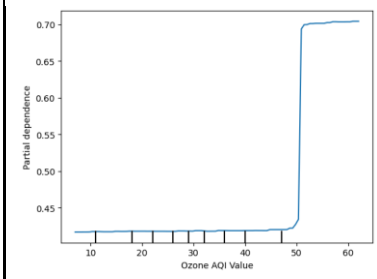
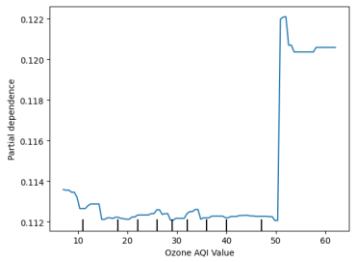
CO AQI Value			
Ozone AQI Value			

Table 3: PDPs for model interpretation

When considering the first row of PDPs in the above table, it can be seen that when the PM 2.5 AQI Value is low the there is high probability that the AQI Category is “Good”. Then, for medium ranges of PM 2.5 AQI Value, there is high probability that the AQI Category id “Moderate” and for higher values, there is high probability to be “Unhealthy” AQI Category. Reports from the World Health Organization (WHO) confirms this observation, where it states that "there is a clear relationship between PM2.5 levels and air quality, with higher PM2.5 levels associated with worse air quality."

When considering the PDPs for the CO AQI Value, there cannot be seen an obviously understanding result as seen in the above scenario of PM 2.5 AQI Value. Here, the reason can be due to the finding from the descriptive analysis that, some observations have very low Ozone AQI values they have very high AQI Values which reflects the Unhealthy Air category.

When considering the PDPs for the Ozone AQI Value, it is visible that the probability of “Moderate” and “Unhealthy” AQI Category is high for high Ozone AQI Values.

Local Model-Agnostic Methods

Local interpretation methods explain individual predictions.

SHAP (SHapley Additive exPlanations) is a method to explain individual predictions. **TreeSHAP**, a variant of SHAP for tree-based machine learning models such as decision trees, random forests and gradient boosted trees. Using Force Plots, we can visualize feature attributions such as Shapley values as "forces". Each feature value is a force that either increases or decreases the prediction. The prediction starts from the baseline. The baseline for Shapley values is the average of all predictions.

Now, let us consider an example (city) where the CO AQI Value is 2, Ozone AQI Value is 25, NO2 AQI Value is 8, PM 2.5 AQI Value is 150 and the Continent is Asia. The random forest model predicts the outcome as "Unhealthy".

Let us see the force plots for each category of the response (AQI Category).

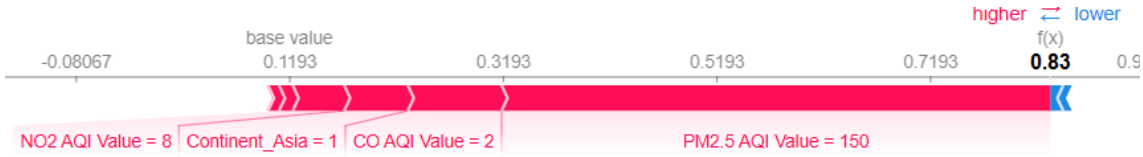
AQI Category	**Force Plot**
Good	
Moderate	
Unhealthy	

Table 4: Force Plots - SHap

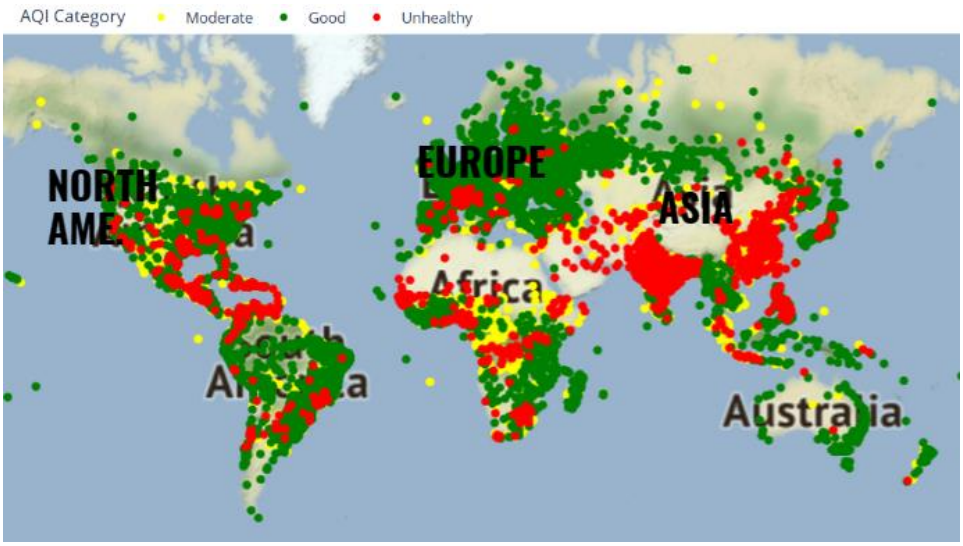
Consider the "Unhealthy" AQI Category. The baseline – the average predicted probability – is 0.1193. This city with the above-mentioned predictor values, has a very high predicted risk (0.83) of being in the "Unhealthy" category. From the baseline of 0.1193, the risk has been increased due to the fact that the PM 2.5 AQI Value is 150, which has a huge impact on the prediction (shown by the length of the red strip). Also,

other factors such as the city belonging to Asian Continent, and the CO AQI Value being 2 has had a considerable role to play in the prediction being “Unhealthy”.

Now, considering the “Moderate” and “Good” AQI Category, the baseline for both the categories is just above 40%. However, this city has a very low predicted probability of being in the “Good” or “Moderate” AQI Category because of the decreasing effects of PM 2.5 AQI Value equal to 150 and being in the Asian Continent. These decreasing effects are signified by the blue strips of the force plot.

7. ISSUES ENCOUNTERED AND PROPORSED SOLUTIONS

One major issue that was encountered is that the observations of the dataset were not representative of the whole world. Majority of the observations were from The USA and India, which could have lead to misleading results in the descriptive analysis as well as in the advanced analysis.



The proposed solution would be to include observations from other cities, which are not mentioned in the dataset, so that the sample becomes representative. This can be done through web-scrapping.

8. REFERENCES

4.1. partial dependence and individual conditional expectation plots (no date) *scikit*. Available at: https://scikit-learn.org/stable/modules/partial_dependence.html (Accessed: 16 September 2023).

Ališauskas, B. (2023) *Web scraping with selenium and python tutorial + example project*, *ScrapFly Blog*. Available at: https://scrapfly.io/blog/web-scraping-with-selenium-and-python/ (Accessed: 16 September 2023).

Molnar, C. (2023a) *Interpretable machine learning*, *9.6 SHAP (SHapley Additive exPlanations)*. Available at: https://christophm.github.io/interpretable-ml-book/shap.html#treeshap (Accessed: 16 September 2023).

Molnar, C. (2023b) *Interpretable machine learning*, *9.1 Individual Conditional Expectation (ICE)*. Available at: https://christophm.github.io/interpretable-ml-book/ice.html#ice (Accessed: 16 September 2023).

Appendix:

StatlearningGroup2/AQI_Prediction_Advanced_Analysis (github.com)