The University of Texas at Austin

Air Quality Analysis
Project Report

CE 395 R 5-Data Mining
Prof. Carlos Caldas
Spring 2018

TEAM MEMBERS:
Shreshtha Shukla (ss83452)
Sonakshi Garg (sg46745)

Introduction

Rapid industrialization has given rise to land, air and water resources sites being contaminated with toxic waste and pollutants, harming the ecosystem and leading to serious health risks among humans, flora, and fauna. Various rules and regulations were put into effect to curb the effects and regulate the levels of pollutants, but it has been a difficult endeavor. Analyzing the historical data and trying to find out the various factors affecting the levels of pollution in various regions may help us in regulating it. The proposed solutions may help in determining emission limits and reduce the pollution by taking necessary action and implementing precautionary measures.

Problem Definition

With a rise in pollution in the metropolitan cities, it has become a priority to identify the highly polluted areas and take precautionary measures. To understand the reason and identify the affected regions, it is imperative to classify the highly polluted areas, analyze pollution levels, and find correlations between the factors affecting these areas and the rate at which the pollution may increase in the future. A detailed analysis would help us understand the measures that need to be taken to manage the pollution levels around the country.

Dataset

The Environmental Protection Agency (EPA) creates air quality trends using measurements from monitors located across the country. All this data comes from EPA's Air Quality System (AQS). Data collection agencies report their data to the EPA via this system and it calculates several types of aggregate (summary) data for EPA internal use.

Field Descriptions:

1. State Code: The FIPS code of the state in which the monitor resides.
2. County Code: The FIPS code of the county in which the monitor resides.
3. Site Num: A unique number within the county identifying the site.
4. Parameter Code: The AQS code corresponding to the parameter measured by the monitor.
5. POC: This is the "Parameter Occurrence Code" used to distinguish different instruments that measure the same parameter at the same site.
6. Latitude: The monitoring site's angular distance north of the equator measured in decimal degrees.
7. Longitude: The monitoring site's angular distance east of the prime meridian measured in decimal degrees.
8. Datum: The Datum associated with the Latitude and Longitude measures.
9. Parameter Name: The name or description assigned in AQS to the parameter measured by the monitor. Parameters may be pollutants or non-pollutants.
10. Sample Duration: The length of time that air passes through the monitoring device before it is analyzed (measured). So, it represents an averaging period in the atmosphere (for example, a 24-hour sample duration draws ambient air over a collection filter for 24 straight hours). For continuous monitors, it can represent an averaging time of many samples (for example, a 1-

hour value may be the average of four one-minute samples collected during each quarter of the hour).

11. Pollutant Standard: A description of the ambient air quality standard rules used to aggregate statistics. (See description at beginning of the document.)
12. Metric Used: The base metric used in the calculation of the aggregate statistics presented in the remainder of the row. For example, if this is Daily Maximum, then the value in the Mean column is the mean of the daily maximums.
13. Method Name: A short description of the processes, equipment, and protocols used in gathering and measuring the sample.
14. Year: The year the annual summary data represents.
15. Units of Measure: The unit of measure for the parameter. QAD always returns data in the standard units for the parameter.
16. Event Type: Indicates whether data measured during exceptional events are included in the summary. A wildfire is an example of an exceptional event; it is something that affects air quality, but the local agency has no control over. No Events means no events occurred. Events Included means events occurred and the data from them is included in the summary. Events Excluded means that events occurred but data from them is excluded from the summary. Concurred Events Excluded means that events occurred but only EPA concurred exclusions are removed from the summary. If an event occurred for the parameter in question, the data will have multiple records for each monitor.
17. Observation Count: The number of observations (samples) taken during the year.
18. Observation Percent: The percent representing the number of observations taken with respect to the number scheduled to be taken during the year. This is only calculated for monitors where measurements are required (e.g., only certain parameters).
19. Completeness Indicator: An indication of whether the regulatory data completeness criteria for valid summary data have been met by the monitor for the year. Y means yes, N means no or that there are no regulatory completeness criteria for the parameter.
20. Valid Day Count: The number of days during the year where the daily monitoring criteria were met if the calculation of the summaries is based on valid days.
21. Required Day Count: The number of days during the year which the monitor was scheduled to take samples if measurements are required.
22. Exceptional Data Count: The number of data points in the annual dataset affected by exceptional air quality events (things outside the norm that affect air quality).
23. Null Data Count: The count of scheduled samples when no data was collected and the reason for no data was reported.
24. Primary Exceedance Count: The number of samples during the year that exceeded the primary air quality standard.
25. Secondary Exceedance Count: The number of samples during the year that exceeded the secondary air quality standard.
26. Certification Indicator: An indication whether the completeness and accuracy of the information on the annual summary record have been certified by the submitter. Certified means the submitter has certified the data (due May 01 the year after collection). Certification not required means that the parameter does not require certification, or the deadline has not yet passed. Uncertified (past due) means that certification is required but is overdue. Requested but not yet concurred means the submitter has completed the process, but EPA has not yet acted to certify the data. Requested but denied means the submitter has completed the process,

but EPA has denied the request for the cause. Was certified but data changed means the data was certified but data was replaced, and the process has not been repeated.

27. Num Obs Below MDL: The number of samples reported during the year that were below the method detection limit (MDL) for the monitoring instrument. Sometimes these values are replaced by 1/2 the MDL in summary calculations.
28. Arithmetic Mean: The average (arithmetic mean) value for the year.
29. Arithmetic Standard Dev: The standard deviation about the mean of the values for the year.
30. 1st Max Value: The highest value for the year.
31. 1st Max DateTime: The date and time (on a 24-hour clock) when the highest value for the year (the previous field) was taken.
32. 2nd Max Value: The second highest value for the year.
33. 2nd Max DateTime: The date and time (on a 24-hour clock) when the second highest value for the year (the previous field) was taken.
34. 3rd Max Value: The third highest value for the year.
35. 3rd Max DateTime: The date and time (on a 24-hour clock) when the third highest value for the year (the previous field) was taken.
36. 4th Max Value: The fourth highest value for the year.
37. 4th Max DateTime: The date and time (on a 24-hour clock) when the fourth highest value for the year (the previous field) was taken.
38. 1st Max Non-Overlapping Value: For 8-hour CO averages, the highest value of the year.
39. 1st NO Max DateTime: The date and time (on a 24-hour clock) when the first maximum non-overlapping value for the year (the previous field) was taken.
40. 2nd Max Non-Overlapping Value: For 8-hour CO averages, the second highest value of the year that does not share any hours with the 8-hour period of the first max non-overlapping value.
41. 2nd NO Max DateTime: The date and time (on a 24-hour clock) when the second maximum non-overlapping value for the year (the previous field) was taken.
42. 99th Percentile: The value from this monitor for which 99 percent of the rest of the measured values for the year are equal to or less than.
43. 98th Percentile: The value from this monitor for which 98 percent of the rest of the measured values for the year are equal to or less than.
44. 95th Percentile: The value from this monitor for which 95 percent of the rest of the measured values for the year are equal to or less than.
45. 90th Percentile: The value from this monitor for which 90 percent of the rest of the measured values for the year are equal to or less than.
46. 75th Percentile: The value from this monitor for which 75 percent of the rest of the measured values for the year are equal to or less than.
47. 50th Percentile: The value from this monitor for which 50 percent of the rest of the measured values for the year are equal to or less than (i.e., the median).
48. 10th Percentile: The value from this monitor for which 10 percent of the rest of the measured values for the year is equal to or less than.
49. Local Site Name: The name of the site (if any) given by the State, local, or tribal air pollution control agency that operates it.
50. Address: The approximate street address of the monitoring site.
51. State Name: The name of the state where the monitoring site is located.
52. County Name: The name of the county where the monitoring site is located.

53. City Name: The name of the city where the monitoring site is located. This represents the legal incorporated boundaries of cities and not urban areas.
54. CBSA Name: The name of the core based statistical area (metropolitan area) where the monitoring site is located.
55. Date of Last Change: The date the last time any numeric values in this record were updated in the AQS data system.

Data Exploration

We used Tableau for data exploration. Since our dataset had many geographical attributes, we used the maps available in Tableau to understand the initial distribution of our data in the United States of America.

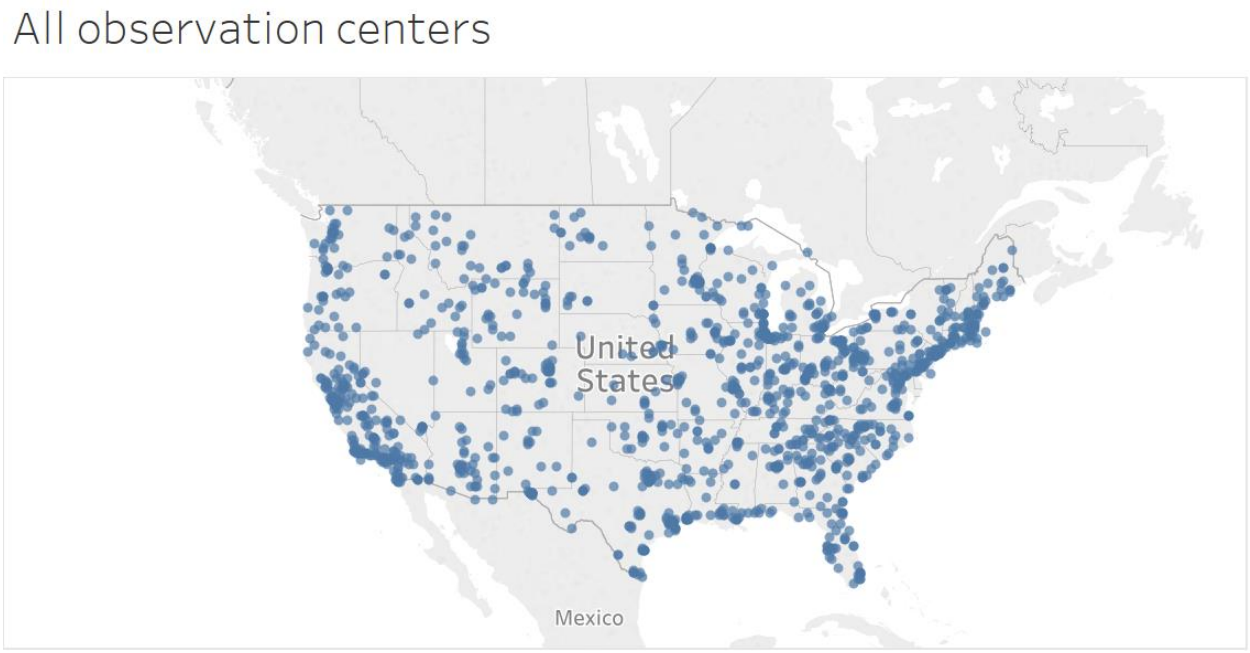


Figure 1*: Using the Latitude and Longitude attributes we were able to visualize the location of observation centers recording air pollution levels in different cities, across the country.*

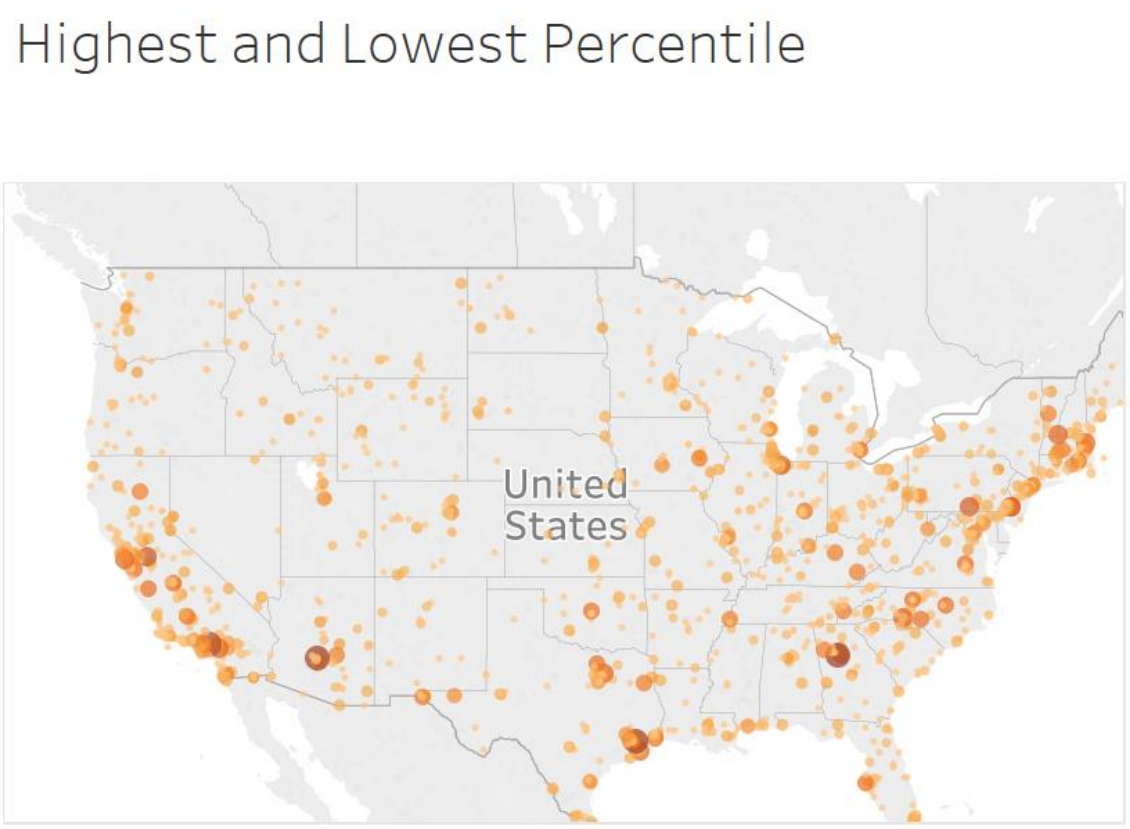


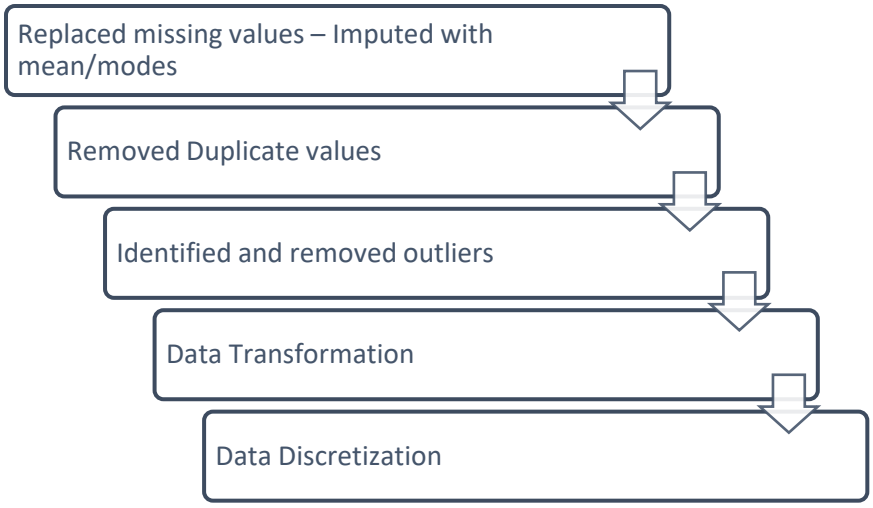
Figure 2: *Highest and Lowest percentile of pollution levels across the country. Size – 99th percentile and Shade – 10th percentile*

Attribute Categorization: To find better inferences from our data, we segregated available attributes before performing feature selection based on domain knowledge.

- Location – State_code, county_code, site_num, state_name
- Time and Resolution – year, sample_duration, metric_used
- Measurement Method – parameter_code, parameter_name, units_of_meausre
- Yearly Readings Coverage – observation_count, observation_percent
- Readings Statistics – arithmetic_mean, ninety_nine_percentile, ten_percentile
- Validation – completeness_indicator, event_type

Preprocessing

Data pre-processing was one of the most time-consuming steps in our Data Mining pipeline. We used the below procedure to finally arrive at 2557 instances from our dataset.



- Dimensionality – Along with having 1.5 million data points, one of our biggest challenges were 55 attributes in our dataset. To be able to find any concrete conclusions from our data, reducing the dimension to be considered while modeling was our focus. We narrowed the attributes by referring to multiple articles on air pollution and understanding what parameters are of utmost importance.
- Complex and Heterogeneous Data – Our dataset was very difficult to understand since it heavily focused on domain-specific attributes. We had both categorical and nominal data. Most of the value data were categorical data, which posed a problem with missing values. The nominal data was associated with features like the metric system used, percentiles, etc. which led to confusion regarding the choice of our most important features.
- Missing values – We used mean and mode to impute missing values for nominal and categorical data, respectively. But as our categorical data included cities, counties, and states, replacing them with the mode would produce incorrect results. We had to impute the values and correct the data to work with it.
- Duplicate data – We used "RemoveDuplicates" in WEKA to identify and remove duplicate values. This helped us remove the redundant records that would not add any value to our models.
- Outliers – Outliers were initially giving us skewed results and hence we used "InterQuartileRange" in WEKA to identify outliers and remove them. We also removed the noise in our data.

Objectives

- Isolating pollutants which heavily impact pollution levels. The parameter field in the dataset provides us with this information. The parameter can be both pollutant and non-pollutant. This isolation can help us define if any preventive measures can be taken to address pollution levels in various regions.
- Understanding associations between attributes by using different association and clustering methods. The analysis would focus on:
  Area wise clustering – We aim to aggregate high risk and most polluted areas
  Pollutant wise clustering – We aim to explore if a specific type of pollutants is causing an increase in pollution levels, which can be further used to infer if any industrial presence in those regions are the contributing factors

Methods

We decided to perform three data mining tasks using four different algorithms.

1) Classification – Decision Tree and Naïve Bayes

2) Clustering – K-Means

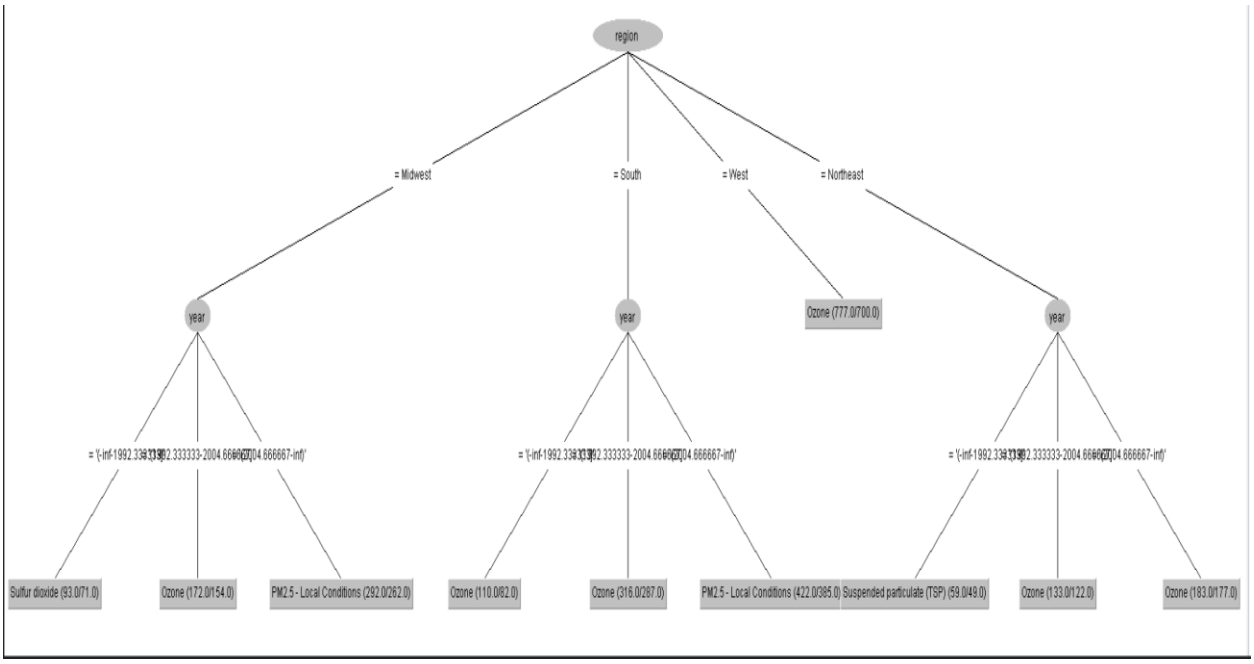
3) Association – Apriori Algorithm

1. Classification

a. DECISION TREE:

We tried multiple permutations and combinations for decision trees to be able to arrive at one from which we can deduce conclusive results. We observed that if we used all our features available, the final tree was not interpretable. Enhancing interpretability guided our feature selection for a decision tree. We also used the "pruning" feature in Weka to create a clearer decision tree.

The region was selected as the root node and data was further split on years. We used entropy as the splitting criteria for our decision tree. It was evident that the west was dominated by "Ozone" pollutant the most over the years. The years for other regions played an important role in determining the dominant pollutant. Years were clubbed in a four-year bucket to finally arrive at the pollutant leaf nodes. Another interesting observation was that Ozone was dominant for most regions in most years, but we believe this can also be due to our data being biased as we had the most records for Ozone even after random subset selection of our data.



b. NAÏVE BAYES:

Next classification algorithm that we considered was Naïve Bayes. Since it classifies with an assumption of independence between features we had some very interesting results on our dataset.

We split our data in an 80-20 spit and trained our model on 80% of the data available. With the classification results on our dataset reaching an accuracy of 99%, we believe that our model is overfitting our data.

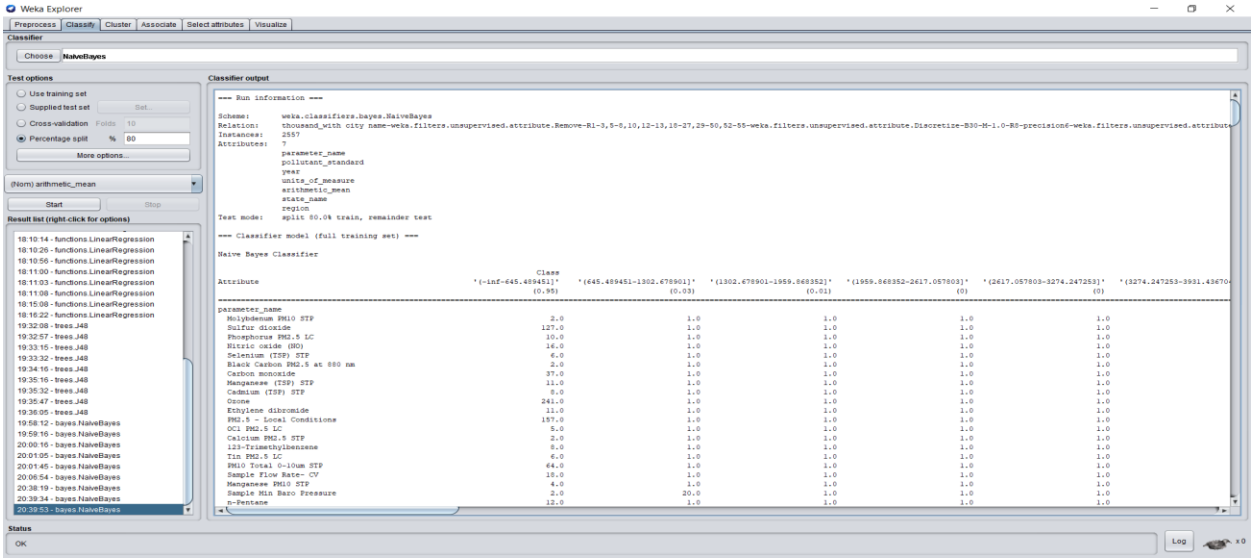


Figure 3: *Naïve Bayes classifier with 80-20 split*

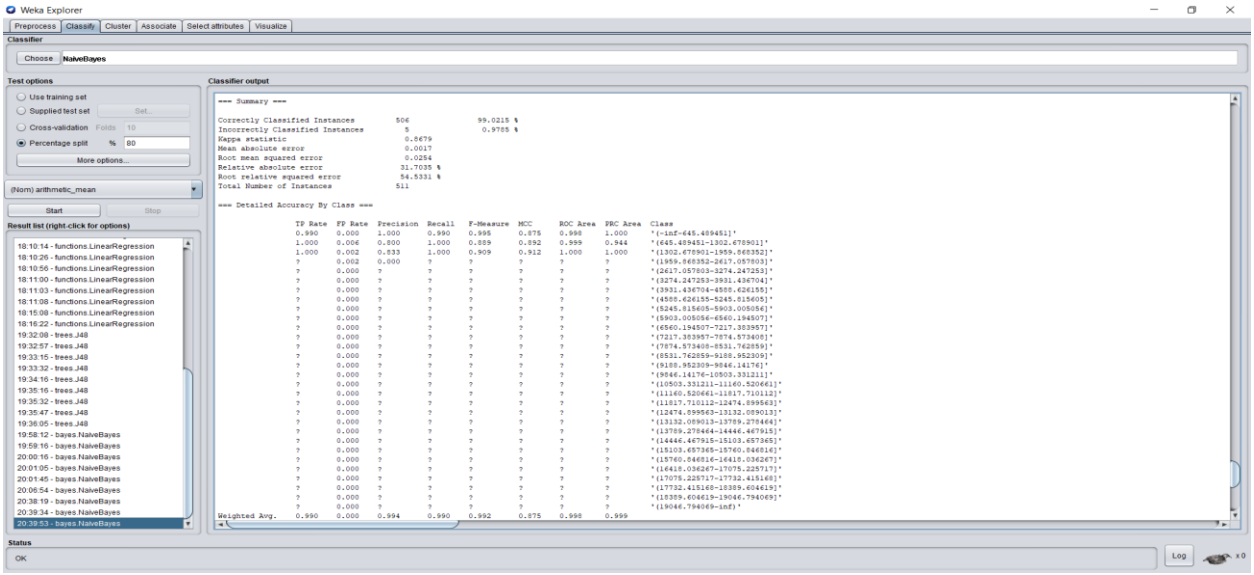


Figure 4: *Naïve Bayes classifier – Detailed Accuracy by class*

COST/BENEFIT ANALYSIS:

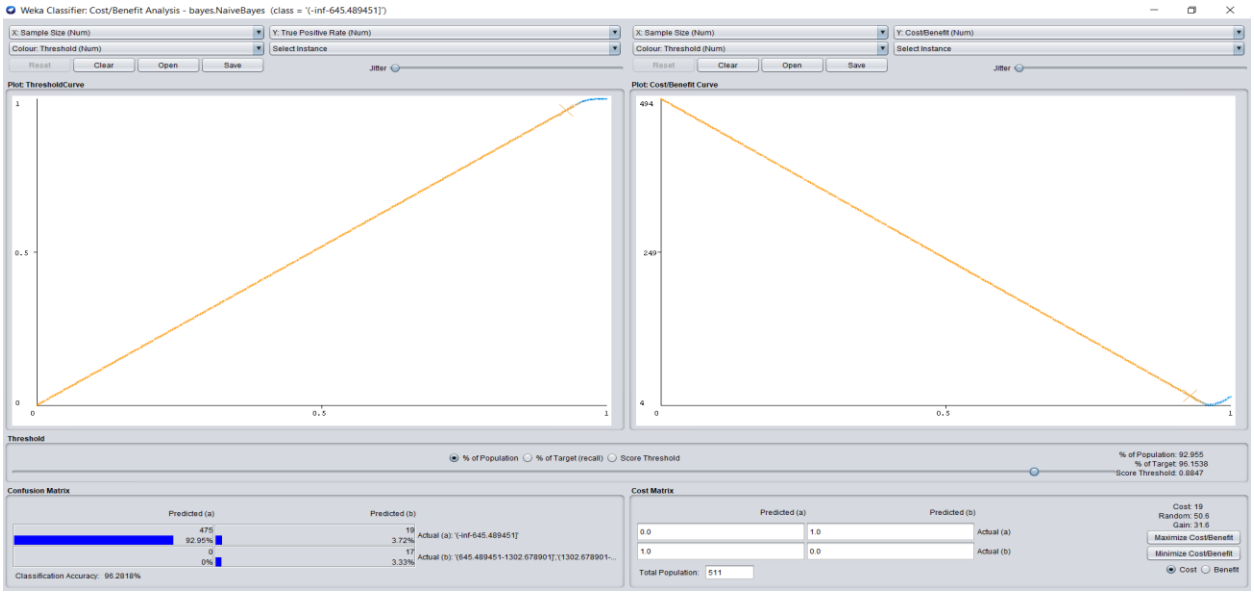


Figure 5: *Threshold Curve value = 0.8 and Cost Benefit Curve value = 96.67%*

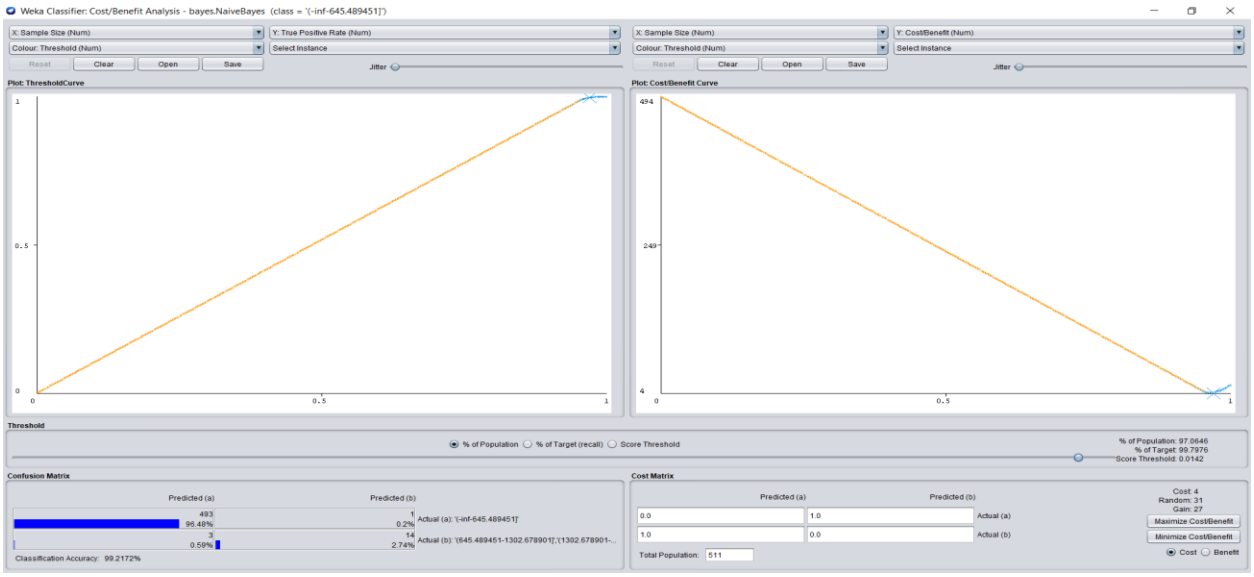


Figure 6: *Threshold Curve value = 0.014 and Cost Benefit Curve value = 99.2%*

2. Clustering

a. Simple K-Means

We used Simple K Means as our clustering algorithm. The model was successfully able to classify 51% of the instances. The clusters were formed on the average value of pollutants. We tried multiple values of clusters to see how the inferences change and if more insights can be discovered. But our best recognizable patterns were visible in 3 clusters: Cluster 0 (Blue) has the minimum average values and cluster 1 (Red) has the highest average value for pollutants.

When we visualized our results, we could see that parameters with a highest average value over the years were dominant in the South region more than Midwest, west or the North East. The Midwest follows closely after.

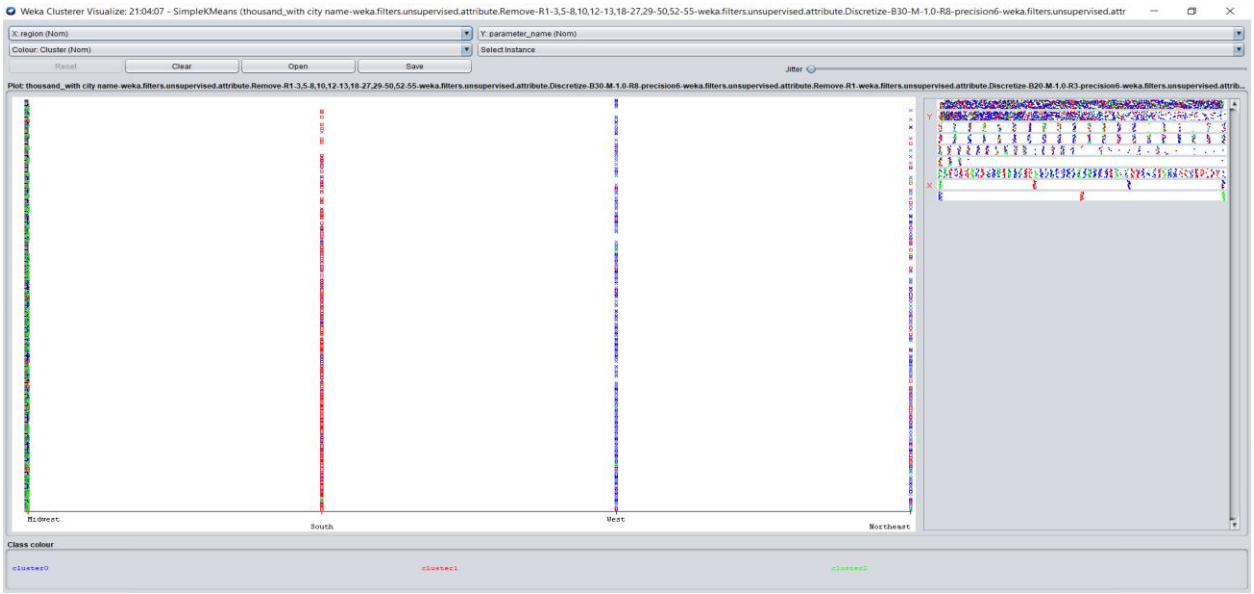


Figure 7: *Simple K Means clustering*

3. Association Rules

We discovered Association rules using “Apriori” algorithm. The rules echoed some of the findings we had from clustering, where the regions were associated with the mean values of pollutants over time. This is visible from rows 3,9, 11-20.

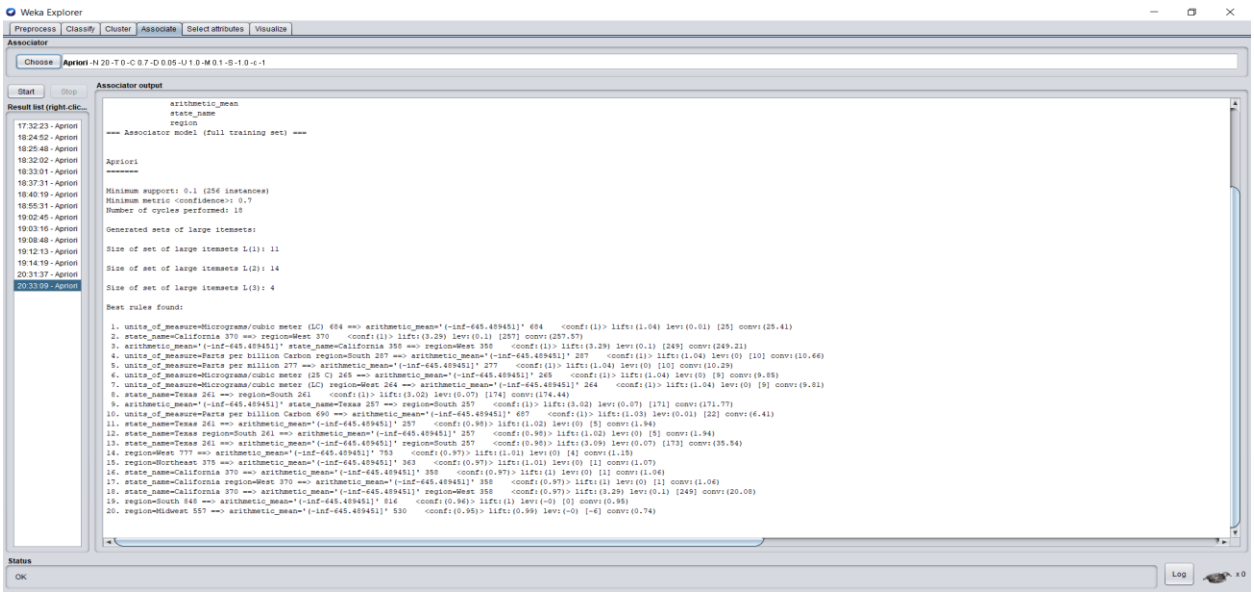


Figure 8: *Association rules using “Apriori” algorithm.*

Findings & Analysis

Based on the analysis of the results we obtained from all the data mining tasks, we can observe that the best results were obtained from the decision tree. The data was clearly classified according to the region, years and the highest amount of pollutants in the area. As an additional benefit, the decision tree was more interpretable than some of the other results we obtained and allowed us to isolate specific pollutants according to regions.

We can observe that the Region is the root node in our decision tree. This is a clear indicator that the region-wise division of the data was very helpful in narrowing down the scope of the classification and thus, it gave us more accurate results. We decided this was important to note as it implies that our model is considering the correct feature to calculate the possibility and produce the results.

The second-best results were obtained from the Naive Bayes classification where we were able to predict the approximate mean value of the pollutants in the various regions based on the various attributes in our dataset. Analysis of the association rule and the k-means produced similar results, which allowed us to conclude that the models were working accurately. Given the amount of data

and the dimension, the results can be made more precise. This will be possible by transforming our data and focusing on a region or set of pollutants.

Lastly, the features that did not show much importance on the classification were fairly expected from our analysis above. The pollutant standard, the metric used, Primary Exceedance Count, etc. all had little significance on our model. We concluded that if we were to obtain features that had more correlation to each other and had some definitive impact on the amount of pollutants in the air, we would have been able to come up with better results.

Challenges

The biggest challenge we faced while working with our dataset was its quality. Our dataset consisted of approximately 1 million records and had inconsistencies and missing data. This led us to spend a lot of time on data cleaning and transformation. Also, due to computational reasons, we had to choose only 2557 records for our analysis. The data was also not chronologically arranged. This made it very difficult for us to choose the records randomly in hopes of obtaining results in a broader range.

We wanted to do a detailed analysis to identify the highest pollutants and their amounts in certain areas. Based on the region where these pollutants were concentrated, suggestions could be provided as to what type of policies or measures could be implemented in the future to prevent these pollutant levels from increasing. For instance, if there is a concentration of a pollutant which is caused by certain types of industries, the government could enforce laws to reduce the waste produced by these industries or maybe put a restriction on the number of industries in that region. Since the data we had was inconsistent, our idea to incorporate the extra information was not fruitful and we could not reach substantial conclusions.

Further, most of our useful data was categorical data, which limited the implementation of models to only qualitative models. Because of the large size of the dataset and limited computational power, running the models was time consuming task. Some of the models were not suitable for Weka and this limited our scope of exploration.

Conclusion

The analysis of the results obtained from the data mining tasks revealed that the largest concentration of pollutants can be found in the South. Ozone and PM2 5 seemed to be the pollutants that are widely present in the atmosphere. This leads us to believe that if we remove these two pollutants from our dataset and explore the data some more, we may be able to find some new patterns and come up with better conclusions. The restriction in terms of computational power was also a limitation for us and maybe the use of some better resources would help us utilize the dataset to its fullest and get results that a small number of data points were unable to provide.