NYC Air Quality Data Analysis using Machine Learning

Aman Sarawgi
Computer Science and Application
Virginia Tech
Falls Church, Virginia, USA
amansarawgi@vt.edu

Dvijen Trivedi
Computer Science and Application
Virginia Tech
Falls Church, Virginia, USA
tdvijenmahesh20@vt.edu

Adinew Zeleke
Computer Science and Application
Virginia Tech
Falls Church, Virginia, USA
azeleke@vt.edu

Sara Alsalamah
Computer Science and Application
Virginia Tech
Falls Church, Virginia, USA
salsalamah@vt.edu

ABSTRACT

Given the ever-increasing pace of urbanization/urban growth, pollution and air quality have become amongst the key-central focus of governments, public policy architects, decision-makers (executives), and people almost everywhere in the world irrespective of political boundaries, countries' level of economic development, technical constraints, and the socio-economic complexities associated with pollution and air quality. Accordingly, understanding pollution levels, i.e., classification and intensities, is of paramount importance in ever-expanding global economic activities, energy consumption, high growth of urban living, and unprecedented urbanization. Accordingly, our project report details our work, and the approach taken to developing an air quality data analysis and ML classification modeling tools by using the NYC air-quality data collected across the five boroughs and 59 community districts over time.

While the air quality surveillance dataset, obtained from the NYC open-data platform, and the array of air quality indicator readings were taken 2009-2018, we focused on our general modeling approach and model development. The model can be calibrated using more recent data to improve the classification and prediction capacities. Accordingly, our analysis and model, fitted and calibrated with more recent data, is intended to provide categorical ranking and classification/prediction of air quality, and answer important threshold questions based on the national NAAQS, and NYC local-mean values derived from the dataset. Furthermore, our analysis and model provide a data-driven analytical tools and may supports decision-making relative to pollution and air-quality risk mitigation -- within a given city, district, and region, and enable tailoring of mitigation plans based on specific air quality classification. And, these tools can assist in the mobilization of resources, public-health awareness, leveraging of economic incentives, deployment of technology, fostering cooperation, and applying various policy interventions to manage pollution and improve air quality.

CCS CONCEPTS

• Computing Methodologies → Machine Learning → Cross Validation • Information systems → Data mining

KEYWORDS

Air quality, National Air Quality Standards, Logistic Regression, Decision Tree, Naïve Bayes

ACM Reference format:

Aman Sarawgi, Dvijen Trivedi, Adinew Zeleke and Sara alsalamah. 2021. NYC Air Quality Data Analysis using Machine Learning. In *Proceedings of Class Project (Urban Computing). ACM, New York, NY, USA, 8 pages.* https://doi.org/10.1145/1234567890

1 Introduction and Background

With the urbanization and intensifying economic dynamism of today's world and expanding cities (mega-cities) come several multifaceted challenges associated with air pollution and air quality which would otherwise be inconceivable a few decades ago. The mitigation or control measures deployed by the relevant governmental, public, and private entities locally, nationally, and globally also come with various trade-offs (cost-benefit) challenges, policy dilemmas, socio-economic complexities, and many unintended consequences. Similarly, the direct economic costs, social and public-health expenses, and various negative externalities associated with pollution/air quality have become intolerable with harmful effects on public health. While not included in our modeling, NYC AQI datasets provide extensive information on emergency department visits, hospitalizations, and fatalities due to asthma, cardiac, and respiratory diseases attributable to air pollution and the concentrations of pollutants, such as Ozone (O3) and PM2.5, negatively affecting public health. Furthermore, balancing various competing interests, such as the social and public-health costs of pollution vis-à-vis economic growth and improved living standards, do require multifaceted approaches and multidisciplinary problem-solving. Emerging innovations and advancements in technology, science, mathematics, engineering, computer science, computations, and big data can provide unparalleled benefits to undertake and tackle these challenges. By utilizing data and data mining techniques, one can see patterns, detect

anomalies, gain insights, make inferences, and enable evidence base and data-driven decision-making at various levels. The vast amount of data (big-data) accumulated over the past several decades in tandem with computing and algorithmic/stochastic modeling will continue to be of great benefit to understand, inform, suggest possible remedies and interventions, and test the efficacy of interventions, and ultimately assist with the mitigation and remedy pollution-related AQ problems arising from unprecedented urbanization. Accordingly, the deployment of analytical decision support tools, data-driven mathematical models, and similar empirical techniques that can support policy decisions that affect the long-term economic and social wellbeing of people is of utmost significance and necessity. And such data-driven decision tools can enable local, regional, state, and federal entities to make methodical and data-driven decisions supported by analytical models.

Accordingly, our model captures the air quality data of the City of New York and classifies (categorical ranks) the five boroughs and 59 community districts in terms of air-quality grades, such as A+, A, B, C, D, E, and F, using the US national ambient air quality standards (NAAQS) as a baseline. Additionally, our model ranks and further classifies the five boroughs and 59 community districts, using as a baseline the local-mean of air-pollutants (AQ readings) taken over time across the city. The model ranks (scores) using the "local-mean" and "NAAQS" values, and applies two unique baselines, capturing the six key AQ indicators and the relative concentration of these air pollutants in the NYC local ambient

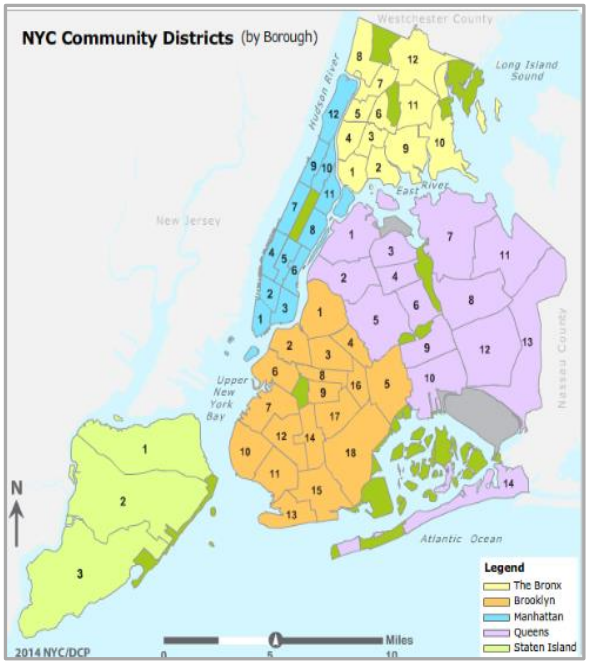


Figure 1: NYC Community Districts of each Borough

air. The six major AQ indicators included in our model are SO2(Sulfur-Dioxide), NO2(Nitrogen-Dioxide), O3(ozone), PM2.5 (Particulate-Matters), and C6H6 (Benzine), CH20 (Formaldehyde). The rankings and classifications based on NYC local mean is intended to support local-level evaluations and area-based assessments across the NYC five boroughs and 59 community districts. NYC neighborhoods, comprised of boroughs and community districts, are illustrated in the graphics above for context, clarity, and visualization of the city's neighborhoods and geography. Figure 1 shows NYC -- Manhattan, Queens, Brooklyn, Bronx, and Staten Island, and the respective 59 Community Districts.

2 Related Research

A rich body of related research exists on the analysis of air quality all over the world. With Pollution and global warming increasing at an alarming rate, it has a massive impact on the health of humans [2,7,16]. The studies range from predicting air quality index in Taiwan [5] to developing machine learning models to study the air quality in California [15].

Various models like SVM and Random Forest have been used to predict the air quality index [1,3] and even simpler models like Naive Bayes, Decision Tree and Logistic Regression were used to built to predict air quality for a region [4]. Obtaining these results is important in an urban setting as it helps the policymakers make suitable decisions to reduce further pollution of that toxin. Studying the daily fluctuation of the toxin provides crucial information on which toxin is affecting a particular region [12]. Thus, countermeasures could be taken to reduce such emissions in that area. Certain models have been built to predict the amount of toxins that could be generated based on the current emissions [8,9]. The main toxins responsible for air pollution are Fine Particulate Matters(PM2.5), Ozone (O3), Nitrogen Dioxide (NO2) , Sulfur Dioxide (SO2), Benzene (C6H6), Formaldehyde (CH20) and these models help monitor the presence of the toxins in the surrounding regions. The paper written by Zhongjie Fu1 is based on the air quality in Hangzhou which helps build model for other regions [10]. A supervised learning approach has been used to train and build a model which helps understand the behavior of the toxins in any region [13,14]

A review on the performance of machine learning algorithms was done by Huda W Ahmed and Dr Jameelah H Alamire which helps understand and analyze which model performs the best for the dataset [6]. These studies can help predict the pollution in a particular region and help in early warning [11] in regions where the air pollution is out of control and so it can help prevent loss of lives and help improve quality of life.

3 Approach

3.1 NYC Air Quality Surveillance Data

The NYC air quality surveillance dataset contains various attributes (features) shown in the table below; also, the same dataset contains public health related surveillance data which is beyond the scope of our project.

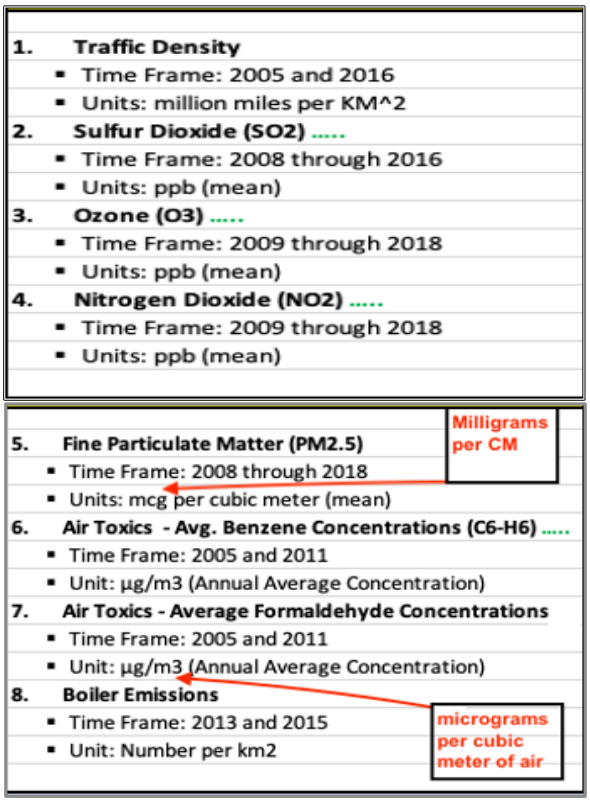


Figure 2: NYC AQ original dataset attributes.

Accordingly, we have selected six key air quality indicators (surveillance attributes) for our model. Given that national (NAAQS) and other comparable standards, such as air quality standards of the State of California, EU, and other economically developed countries, have been established for the six key air quality attributes included in our model, we were able to set our baselines(benchmark) correspondingly and determine rankings and classify. The six (6) major air quality indicators, per city-wide recorded data of pollutants and particulate matters, used in our model are listed below:

- ***Fine Particulate Matters(PM2.5)***
- ***Ozone (O3)***
- ***Nitrogen Dioxide (NO2)***
- ***Sulfur Dioxide (SO2)***
- ***Benzene (C6H6)[1]***
- ***Formaldehyde (CH20)[2]***

[1] *AAQ value of 5 ug/m^3 is used as a baseline for benzene concentration; 5 ug/m^3 is comparable and consistent with EU, WHO-minimum, and other country standards. NAAQS for benzene concentrations could not be determined based on various searches conducted.*

NOTE: Carbon Monoxide (CO) and Lead (Pb) are specific parts of the top six of the NAAQS and most US states' air quality standards, however, the NYC AQ surveillance dataset did not register these indicators. Instead, Benzene (C6H6) and Formaldehyde (CH20) are included in the NYC dataset which we have included in our modeling to account for a total of six(6) air quality indicators as listed above.

3.2 Air Quality Compartmental Ranking

Air Quality Cleanness (CL) Scoring-Grading scale is established based on the six major AQ indicators. The scale, shown in Figure 3 below, ranks (scores) each air quality indicator's readings as A+, A, B, C, D, E, and F by applying the predefined values of the US National Ambient Air Quality Standards (NAAQS) and NYC local-mean as baselines (benchmarks).

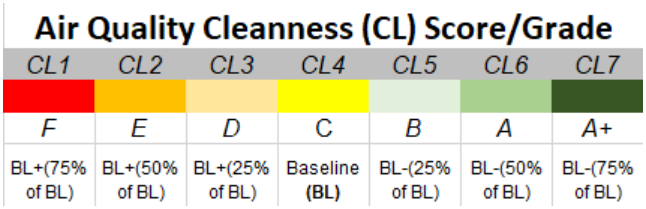


Figure 3 : Air Quality Cleanness (CL) Scoring/Grading Scale

As indicated in Figure 3 above, the grading scale evaluates air quality through benchmarking at 75%>BL(baseline), 50%>BL(Baseline), 25%>BL(Baseline), 25%<BL(Baseline), 50%<BL(Baseline), 75%<BL(Baseline) -- by using the United States national air quality standards (NAAQS), and the NYC local-mean (averages) of readings of the 6 AQ indicators taken overtime. Subsequently, formulas, "if/then" conditionals, were applied to each air quality indicator (unique reading) included in a dataset of thousands of surveillance readings. By way of the above scoring process, grades were determined based on the scale shown in Figure 4 and Figure 5 below. For example, Ozone (O3) values (readings)were scored applying the formulas shown below, which factor in the national air quality standards value of 140 ug/m3, and NYC local-mean values of 60.33ug/m3. Hence, similar formulas were applied across all Ozone (O3) readings, taken in each borough and community district over several years. The local-mean values

[2] *California AQ standard of 3 ug/m^3 is used as a baseline given NAAQS for Formaldehyde concentrations is not available based on various searches.*

for the six air quality indicators, i.e., SO2, N02, O3, CH20, PM 2.5, and C6H6, for each borough and community district, were extracted from the original AQ dataset to ascertain mean-values and to be used as baselines.

NAAQS value of 140 ug/m3 for Ozone

⇒ *IF(L2<35,"A+",IF(L2<70,"A",IF(L2<105,"B",IF(L2<140,"C",IF(L2<175,"D",IF(L2<210,"E","F"))))))*

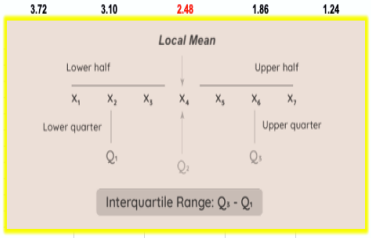
NYC Mean value of 60.33 ug/m3 for Ozone

⇒ *IF(L2<15.08,"A+",IF(L2<30.17,"A",IF(L2<45.25,"B",IF(L2<60.33,"C",IF(L2<75.41,"D",IF(L2<90.5,"E","F"))))))*

Similarly, all instances (thousands of data readings) were treated with similar formulas to obtain line-item scoring/rank of surveillance readings (indicators) taken across the various neighborhood and districts. Figures 4 and 5 show a compartmental scoring (grading) scale for each AQ Indicator.

Air Quality Grading Scale based on local-mean

	BL+(75% of BL) 75% higher	BL+(50% of BL) 50% higher	BL+(25% of BL) 25% higher	Base-line (BL) (using local-mean)	BL-(25% of BL) 25% lower	BL-(50% of BL) 50% lower	BL-(75% of BL) 75% lower
	CL1	CL2	CL3	CL4	CL5	CL6	CL7
	F	E	D	C	B	A	A+
AQ Indicators				micro-gram/m3 (ug/m3)			
1. Nitrogen Dioxide (NO2)	72.10	61.80	51.50	41.20	30.90	20.60	10.30
2. Ozone (O3)	105.58	90.50	75.41	60.33	45.25	30.17	15.08
3. Sulfur Dioxide (SO2)	11.99	10.28	8.56	6.85	5.14	3.43	1.71
4. Fine Particulate Matter (PM2.5)	17.40	14.91	12.43	9.94	7.46	4.97	2.49
5. Benzene (C6H6)	3.55	3.05	2.54	2.03	1.52	1.02	0.51
6. Formaldehyde (CH2O)	4.34	3.72	3.10	2.48	1.86	1.24	0.62



NOTE: the arithmetic mean of the NYC data set is used for the above distributions and rankings

Figure 4: Air Quality Ranking & Classification Scale based on NYC local-mean

Air Quality Grading Scale based on local-mean NAAQS *

	BL+(75% of BL) 75% higher	BL+(50% of BL) 50% higher	BL+(25% of BL) 25% higher		BL-(25% of BL) 25% lower	BL-(50% of BL) 50% lower	BL-(75% of BL) 75% lower
	X1	X2	X3	X4	X5	X6	X7
	F	E	D	C	B	A	A+
AQ Indicators				micro-gram/m3 (ug/m3)			
1. Nitrogen Dioxide (NO2)	174.37	149.46	124.55	99.64	74.73	49.82	24.91
2. Ozone (O3)	245.00	210.00	175.00	140.00	105.00	70.00	35.00
3. Sulfur Dioxide (SO2)	343.88	294.75	245.63	196.50	147.38	98.25	49.13
4. Fine Particulate Matter (PM2.5)	21.00	18.00	15.00	12.00	9.00	6.00	3.00
5. Benzene (C6H6)	8.75	7.50	6.25	5.00	3.75	2.50	1.25
6. Formaldehyde (CH2O)	5.25	4.50	3.75	3.00	2.25	1.50	0.75

Figure 5: Air Quality Ranking & Classification Scale based on NAAQS

3.3 Data Cleaning

As part of our phase I data engineering/data organization effort, the following additional attributes (features) were added, and rankings/grades were applied to each AQ surveillance indicator AQ readings initially recorded and presented in the original dataset.

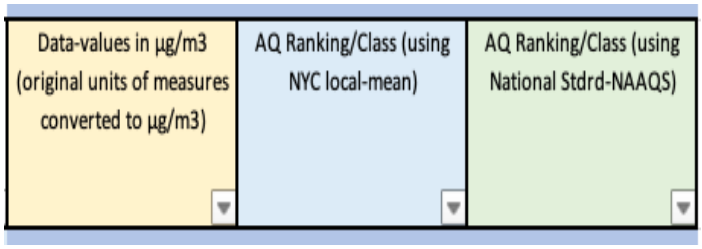


Figure 6: Three additional attributes added to the original NYC AQ surveillance dataset.

The six air quality indicators and associated surveillance readings (data points), recorded by the City of New York from 2009 to 2018, were initially presented in the original dataset in ununiform units such as PPM, PPB, and mg/m3. As part of the data organization and cleaning process, and in the interest of model output uniformity, all data-points (AQI readings) were converted to micro-gram per cubic meter (ug/m3) by applying the relevant conversion formula, following the simplified conversion formula with the application of the Avogadro's constant, from molecular theory, at a temperature of 25 degrees centigrade.

The conversion of units enabled data compatibility and standardization of the various units of measurements used included in the original dataset -- and ensured that all comparisons between the six surveillance readings are "apples to apples" .

Ozone [O_3]	1.963 mg/m^3	M = 48.00 g/mol
Sulfur dioxide [SO_2]	2.620 mg/m^3	M = 64.06 g/mol
Nitrogen Dioxide [NO_2]	1.882 mg/m^3	M = 46.01 g/mol

Figure 7: Conversion values used to obtain single standard unit across all attributes.

3.4 Exploratory Analysis

The histogram plots below show that NYC, boroughs and districts, air quality distribution is right-skewed when evaluated against the national standard – NAAQS (Fig. 8.1) and reflect a normal distribution when assessed the local-mean values of surveillance readings taken across the city. These observations and data patterns demonstrate that NYC maintains lower-emission levels and relatively cleaner ambient air, and people make good use of public transit, biking, and walking.

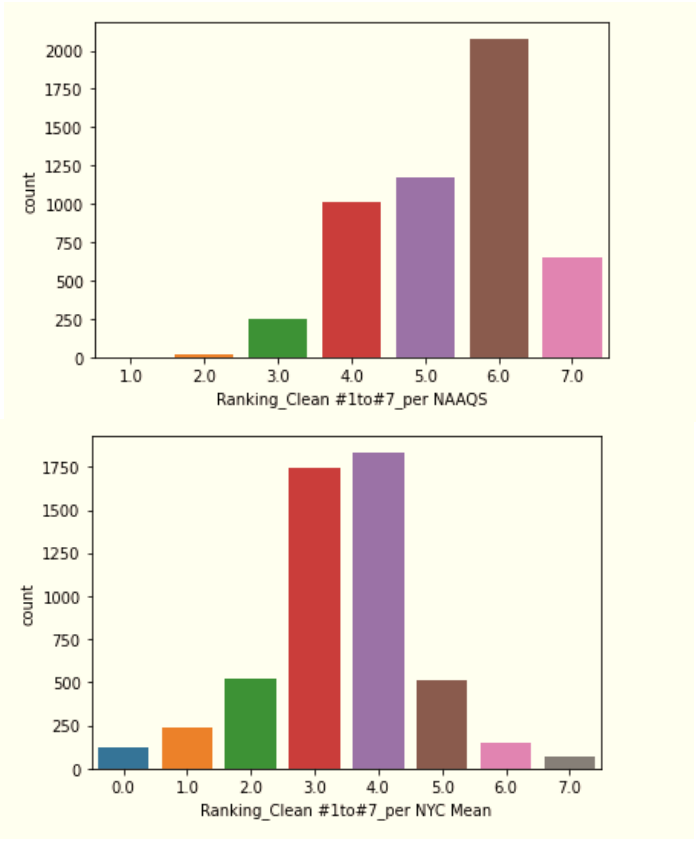


Figure 8: Ranking Distribution based on NAAQS and NYC Mean

Additionally, the scatter plots below illustrate how air quality change (vary) across the city when the readings are evaluated across the five boroughs relative to the geo-IDs assigned, i.e., geo-ID: 100, 200, 300, 400, and 500. The general trend shows that air quality improves when one travels from geo-ID 100 to 500. Note: Ozone (O3) readings (data) appear not to follow the same pattern as the five other indicators.

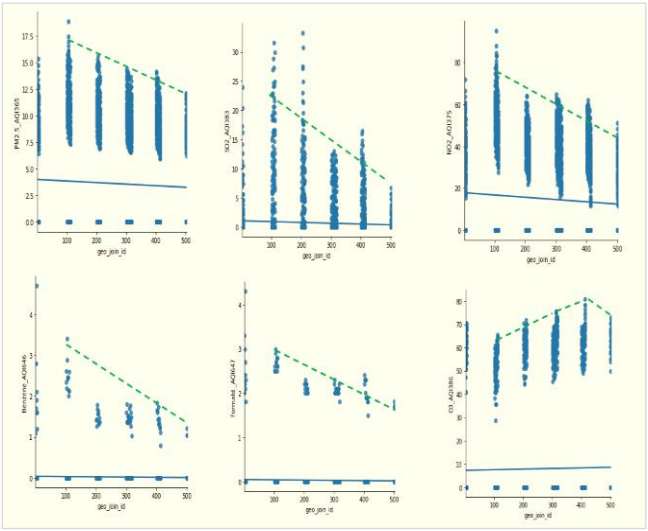


Figure 9: Scatter plots of air quality readings (y-axis) with respect to Geo IDs (x-axis)

With respect to air quality improvements over time, the data shows NYC is trending and performing very well in lowering the PM2.5, SO2, NO2, and other pollutants. The time-plots below show a descending pollution trends, and one can observe that NYC improving air quality over time.

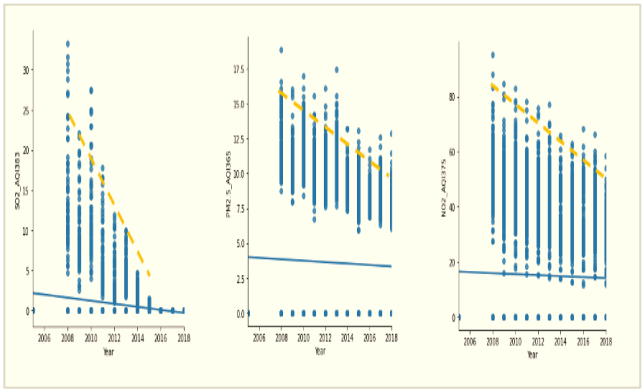


Figure 10: Scatter plots of air quality indicator readings (y-axis) with respect to time (x-axis).

Additionaly, the five boroughs and neighborhood locations (districts) were given corresponding latitudinal and longitudinal data -- which was completed as part of the data augmentation process. Including latitudinal and longitudinal coordinates enabled visualization of the genral air quality and detect potential hot spots in the city.

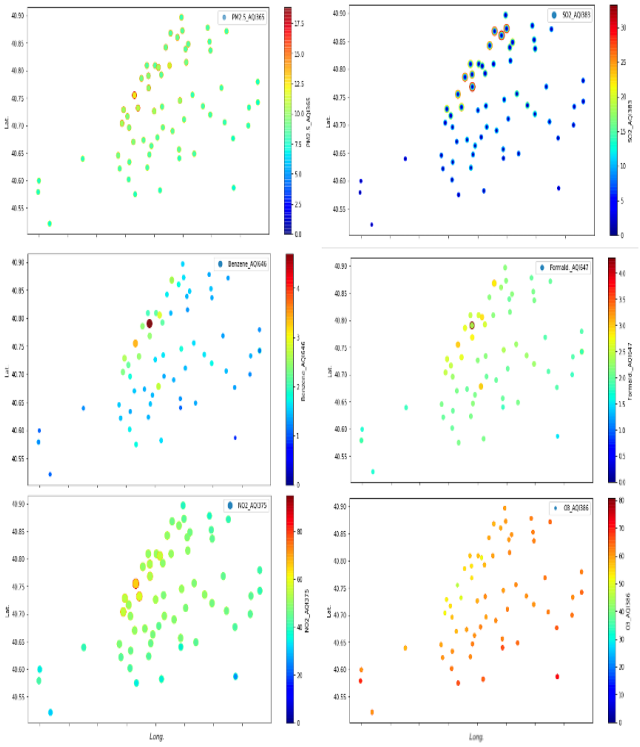


Figure 11: Air Quality Hotspots

3.5 Preproceesing and Feature Engineering

Initial data cleaning process included identifying the six pollutants viz. Fine Particulate Matters(PM2.5), Ozone (O3), Nitrogen Dioxide (NO2) , Sulfur Dioxide (SO2), Benzene (C6H6), Formaldehyde (CH20). These six pollutants (toxins) formed the features of interest. Concentrations of all six toxins were individually recorded for each of the UHFs, community districts, boroughs and for the overall city of New York in the dataset. As toxins were recorded in different units, they were converted to a common unit of ($\mu g/m^3$). Thresholds (baseline measures) for each of the toxins were used to identify recorded toxin measures using interquartile ranges into seven different categories from A^+ to F which serves as ranking of the air quality based on the recorded toxin measure for a given region. Again, such rankings were done using two different sets of thresholds or baselines one using the national standard and other using the local mean for the city of New York. Presence of data in the above format falls prey to several challenges which needed to be resolved before building our model.

	unique_id	indicator_id	name	measure	measure_info	geo_type_name	geo_join_id	geo_place_name	time_period	start_date	data_value (units vary)	data-value_µg/m3 (all converted to µg/m3)
0	315585	646	Air Toxics Concentrations- Average Benzene Con...	Annual Average Concentration	µg/m3	Borough	1	Bronx	2011	2011-01-01	1.10	1.10
1	130728	646	Air Toxics Concentrations- Average Benzene Con...	Annual Average Concentration	µg/m3	Borough	1	Bronx	2005	2005-01-01	2.80	2.80
2	315633	647	Air Toxics Concentrations- Average Formaldehyd...	Annual Average Concentration	µg/m3	Borough	1	Bronx	2011	2011-01-01	2.00	2.00
3	130776	647	Air Toxics Concentrations- Average Formaldehyd...	Annual Average Concentration	µg/m3	Borough	1	Bronx	2005	2005-01-01	3.30	3.30
4	605648	365	Fine Particulate Matter (PM2.5)	Mean	mcg per cubic meter	Borough	1	Bronx	Annual Average 2018	2018-01-01	7.25	7.25
5	547354	365	Fine Particulate Matter (PM2.5)	Mean	mcg per cubic meter	Borough	1	Bronx	Annual Average 2017	2017-01-01	7.72	7.72
6	410722	365	Fine Particulate Matter (PM2.5)	Mean	mcg per cubic meter	Borough	1	Bronx	Annual Average 2016	2015-12-31	7.75	7.75

Figure 12: Distribution of Data after Initial Cleaning

Overall air quality for a region needs to be collectively determined based on all six toxin concentrations for a given region (which could be any of the different categorizations of the NYC viz. UHFs, community districts, boroughs and overall city of New York). For the same region let's say R_1 there were multiple records in the dataset pertaining to each of the toxins recorded in different time frames. Hence it was imperative to organize the data concentrated in a single column but spread across multiple records to multiple columns, one for each toxin. Thus, a pivot kind of operation was needed to get six different columns (one for each toxin) which would capture the particular toxin concentration in a given record for a region R_1. Moreover, for the same region R_1 various toxin measures were recorded in different timeframes. It was essential to identify a time period in which all toxins were measured for all regions in the dataset. This was done to avoid the case of missing values as in most scenarios only one or two toxins were recorded for a region across different timeframes. Toxin concentrations measured across different timeframes cannot be considered to predict the overall air quality for a given region. Only considering toxin concentration recorded for all regions in a single timeframe clearly serves the problem statement. Hence after carefully analyzing the data period of 2011 was used, as all toxin concentrations for all the regions were recorded in this timeframe.

At this point for each region R_1 we had recorded toxin concentrations in ($\mu g/m^3$) as well as their respective grades using the National Standard as explained before for each of the pollutants from the 2011 timeframe. Still, we did not have a cumulative grade/ranking for air quality of the region R_1 which would be representative of air quality for the given region. Based on our study, we found that air quality must be determined by the worst pollutant in a given region. Consider the Bronx in NYC. It is ranked good for pollutants Sulphur Dioxide and Benzene (A^+ grade), Ozone and Nitrogen Dioxide (A grade) but ranks on average for Formaldehyde (B) and Fine Particulate Matter (C). Hence the overall air quality for the Bronx would still be classified as average with a grade of C. A final ranking column was generated for the dataset using the above approach for all the regions in NYC.

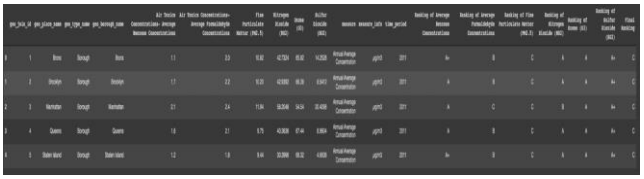


Figure 13: dataset obtained preprocessing and feature extraction

3.6 Model Building

Naive Bayes, Decision Tree and Logistic Regression Models were built to predict air quality for a region based on the recorded toxin concentrations. The above-mentioned Final Ranking column serves as the target class for air quality measure of a given region. Each of the models were trained using the actual values of each of the six toxin concentrations recorded in ($\mu g \backslash m^3$) and not on their respective rankings or grades. Those rankings were given based on the interquartile ranges from the national baselines measure just to get to the overall ranking of the air quality for the given region. This way the goal is that the model will learn actual recorded values for each of the six pollutants. The model based on its learnt parameters will then be able to predict air quality for any region in the US (even for recorded values in future).

As the models were trained on recorded values of the toxin concentrations it is important to standardize them before training the model. This is done to ensure that any large values for a given toxin concentration do not dominate and skew the results. StandardScaler from scikit-learn library was used for standardizing. Fivefold cross validation was used for each model to optimize their training.

3.7 Model Evaluation

The final dataset was split into training and testing sets in 80:20 ratio by randomly selecting regions in each set.

Various scoring metrics like accuracy, precision, recall and f1-score were used to evaluate each model's performance on the testing set. Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Recall is the ratio of correctly predicted positive observations to all observations in actual class. F1-score is the harmonic mean of precision and recall and is generally useful to evaluate performance on a skewed dataset for a combined evaluation of two competing forces in precision and recall.

Below figure shows for each model the average performance on training set using 5-fold cross validation for each of the four metrics average, precision, recall and f1-score.

'Logistic Regression Mean performance on 5 folds is:'
'\n'

	0
test_accuracy	0.999106
test_precision_weighted	0.998660
test_recall_weighted	0.999106
test_f1_weighted	0.998838

Figure 14: Performance of Logistic Regression on Training Dataset

'Decision Tree Mean performance on 5 folds is:'
'\n'

	0
test_accuracy	0.999642
test_precision_weighted	0.999553
test_recall_weighted	0.999642
test_f1_weighted	0.999583

Figure 15: Performance of Decision Tree on Training Dataset

'Naive Bayes Mean performance on 5 folds is:'
'\n'

	0
test_accuracy	0.963888
test_precision_weighted	0.998611
test_recall_weighted	0.963888
test_f1_weighted	0.980311

Figure 16: Performance of Naïve Bayes on Training Dataset

We then run all the models built on unseen Testing Dataset.

	precision	recall	f1-score	support
0	1.0	1.0	1.0	58.0
1	1.0	1.0	1.0	1338.0
2	1.0	1.0	1.0	3.0
accuracy	1.0	1.0	1.0	1.0
macro avg	1.0	1.0	1.0	1399.0
weighted avg	1.0	1.0	1.0	1399.0

Figure 17: Performance of Logistic Regression on Testing Dataset

	precision	recall	f1-score	support
0	1.0	1.0	1.0	58.0
1	1.0	1.0	1.0	1338.0
2	1.0	1.0	1.0	3.0
accuracy	1.0	1.0	1.0	1.0
macro avg	1.0	1.0	1.0	1399.0
weighted avg	1.0	1.0	1.0	1399.0

Figure 18: Performance of Logistic Regression on Testing Dataset

	precision	recall	f1-score	support
0	1.0	1.0	1.0	58.0
1	1.0	1.0	1.0	1338.0
2	1.0	1.0	1.0	3.0
accuracy	1.0	1.0	1.0	1.0
macro avg	1.0	1.0	1.0	1399.0
weighted avg	1.0	1.0	1.0	1399.0

Figure 19: Performance of Logistic Regression on Testing Dataset

4 Results and discussions

All regions in NYC have their air quality ranked among B, C or D. One may point out the fact that A^+, A, E and F rankings for any region in NYC are relatively absent from the dataset. This is neither a drawback of our approach for qualitatively identifying air quality of a region based on recorded toxin measures nor does it affect our model to be trained and used for other cities in the US. This can be explained as follows. As highlighted before, overall air quality for a region is denoted by the worst grade received by any of the six pollutants for the region. Any region would perform slightly worse for at least any one pollutant and hence the ranking A^+ and A is relatively missing. As observed in exploratory analysis air quality for NYC is right skewed compared to the national standard signifying that overall air quality for regions in NYC would be such that worse rankings of E and F would also be missing.

As long as the data for the same six pollutants is measured in the US for a given region and the national baseline thresholds

for each of the pollutants are known, the model can be trained after applying cleaning and preprocessing steps as highlighted in this paper and it will be able to predict the air quality for any other city in US.

One of the feedbacks we had received was the close to perfect performance of our models. It was highlighted that using a random split for training and testing may result in data leakage. That is the testing data is not truly unseen. What it means is that perhaps for a region in the test set, the model while training has already seen its neighbors and hence gives such good performance. We would like to highlight that this is surely not the case. We removed all regions in Bronx and formed a testing set from it. All models were again trained on regions in NYC other than Bronx and the model was later tested on regions in Bronx using the testing set. Close to perfect performance was again noted for all the models.

5 Conclusions

All the models perform similarly for predicting air quality of regions in NYC. Hence choice of model really depends on one's perspective. We would like to choose Decision Tree as our model for evaluating air quality of any city in the US using the approaches mentioned in the previous section. For any other city in general conditional independence of toxin concentrations which are used as features in the model cannot be guaranteed and hence Naïve Bayes is avoided. Logistic Regression can only solve linear problems which again cannot be guaranteed for any other city in the US. Decision Trees are universal approximators and hence will work for any other city using the approaches mentioned in this paper.

Our data cleaning, preprocessing and feature extraction process provides valuable insights and techniques into the general problem of qualitatively identifying air quality of a region or city in the US using measures of six toxins, Fine Particulate Matters(PM2.5), Ozone (O3), Nitrogen Dioxide (NO2) , Sulfur Dioxide (SO2), Benzene (C6H6), Formaldehyde (CH20).

6 ACKNOWLEDGMENTS

We would like to thank Open Data (https://opendata.cityofnewyork.us) published by New York City agencies and its partners. We have particularly used Air Quality Indicators Open data found here (NYCCAS Air Quality Indicators Open Data.xlsx). We were assisted in our understanding of the various regions in New York city by the geographical data published by the Department of City Planning (https://www1.nyc.gov/site/planning/data-maps/city-neighborhoods.page). We would also like to thank EPA, which is a United States Environmental Protection Agency, for providing National Ambient Air Quality Standards (NAAQS) data (https://www.epa.gov/criteria-air-pollutants/naaqs-table).

Our Code and Dataset is available here - https://github.com/trivedidvijen/NYC_Air_Quality_Data_Analysis_Using_Machine_Learning

7 AUTHOR CONTRIBUTIONS

Dataset was originally found by Adinew Zeleke. All members explored the data and came up with the overall strategy for the problem through multiple team meetings and discussions throughout the semester. Visualization plots were contributed by Adinew Zeleke and Sara Alsalamah. Initial cleaning of the dataset was done by Adinew Zeleke. Data Engineering and Preprocessing, Feature Engineering, Model Building and Evaluation, Result Analysis and Feedback Justification for potential Data Leakage Issue was collectively contributed by Aman Sarawgi and Dvijen Trivedi.

REFERENCES

[1] Yun-Chia Liang, Yona Maimury, Angela Hsiang-Ling Chen and Josue Rodolfo Cuevas Juarez. Machine Learning-Based Prediction of Air Quality. *Applied Sciences (21 December 2020).*

[2] World Health Organization. Air Pollution. Available online: https://www.who.int/health-topics/airpollution#tab=tab_1/ (accessed on 13 March 2020).

[3] Yu, R.; Yang, Y.; Yang, L.; Han, G.; Move, O.A. RAQ A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems. Sensors 2016, 16, 86.

[4] Veljanovska, K.; Dimoski, A. Air Quality Index Prediction Using Simple Machine Learning Algorithms. Int. J. Emerg. Trends Technol. Comput. Sci. 2018, 7, 25–30.

[5] Taiwan's Environmental Protection Administration. Taiwan Air Quality Monitoring Network. Available online: https://taqm.epa.gov.tw/taqm/en/b0201.aspx (accessed on 13 March 2020).

[6] Huda W Ahmed and Dr Jameelah H Alamire. A Review of Machine Learning Models in the Air Quality Research. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 9, Issue 3, March 2020, ISSN: 2278 – 1323.

[7] K . Karatzas, N . Katsifarakis, C . Orlowski , and A . Sarzyński, "Urban Air Quality Forecasting: A Regression and a Classification Approach", Springer, pp. 539–548, 2017.

[8] A. Masih, "Machine learning algorithms in air quality modeling", Global Journal of Environmental Science and Management (GJESM), 5(4): 515-534, autumn 2019.

[9] Y. Rybarczyk and R. Zalakeviciute, "Machine Learning Approach to Forecasting Urban Pollution", IEEE, 2016.

[10] Zhongjie Fu1, Haiping Lin, Bingqiang Huang and Jiana Yao. Research on air quality prediction method in Hangzhou based on machine learning. CISAT 2021 Journal of Physics: Conference Series 2010 (2021) 012011 IOP Publishing doi:10.1088/1742-6596/2010/1/012011.

[11] Yang, Z.S., Wang, J. (2017) A new air quality monitoring and early warning system: Air quality assessment and air pollutant concentration prediction. Environ. Res., 158: 105-117.

[12] Wu, Q.L., Lin, H.X. (2019) A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors. Sci. Total Environ., 683: 808-821.

[13] Yun-Chia Liang, Yona Maimury, Angela Hsiang-Ling Chen and Josue Rodolfo Cuevas Juarez. 2020. Machine Learning-Based Prediction of Air Quality. *Applied Sciences (21 December 2020).*

[14] Madhuri VM, Samyama Gunjal GH, Savitha Kamalapurkar. Air Pollution Prediction Using Machine Learning Supervised Learning Approach. International Journal of Scientific & Technology Research Volume 9, Issue 04, April 2020 ISSN 2277-8616.

[15] Mauro Castelli, Fabiana Martins Clemente, Aleš Popovič, Sara Silva and Leonardo Vanneschi. A Machine Learning Approach to predict Air Quality in California. Research Article, Volume 2020, Article ID 8049504. https://doi.org/10.1155/2020/8049504

[16] F. Caiazzo, A. Ashok, I. A. Waitz, S. H. L. Yim, and S. R. H. Barrett, "Air pollution and early deaths in the United States. Part I: quantifying the impact of major sectors in 2005," *Atmospheric Environment*, vol. 79, pp. 198–208, 2013.