Statistical Analysis of Deterioration of AirQuality in The Pandemic Era

PROF G. R. RAO[1], SHUBHAM JAIN[2], SARTHAK SONI[3], ANJALI PARMAR[4]
[1] *Associate Professor, Department of computer Engineering Bharati Vidyapeeth Deemed to beUniversity College of Engineering, Pune, India*
[2, 3, 4] *Department of computer Engineering Bharati Vidyapeeth Deemed to be University College of Engineering, Pune, India*

Abstract— Predicting air best is a complicated undertaking because of the dynamic nature, volatility, and high variability in time and area of pollutants and particulates. at the same time, being capable of version, are expecting, and screen air first-class is becoming increasingly more relevant, especially in urban areas, due to the discovered essential impact of air pollution on citizens' healt hand the environment. Managing air pollution is one of the principal environmentaldemanding situations in a smart city environment. actual-time tracking of pollutants statistics enables the metropolitans to investigate the cutting-edge site visitors state of affairs of themetropolis and take their selections as a consequence. Existing research has used specific machine learning equipment for pollutants prediction; however, comparative evaluation ofthese techniques is regularly required to have a better understanding in their processing time for more than one dataset. Our task analysis is supposed to forecast pollutant and particulate stages and to predict the air quality index (AQI).

Indexed Terms-- Air Quality, Jupyter Notebook, Deterioration,Analysis

I. INTRODUCTION

With the economic and technological development of the cities, environmental pollution problems are gradually arising, like water, noise, and air pollution. In particular, air pollution has a direct impact on human health through the exposure of the pollutants and particulates, which has now increased the interest in air pollution and the impacts among the scientific community. [1] The challenge of modern society to take air pollution abatement measures based on scientific knowledge has encouraged scientists to study the atmospheric compositionchanges, the short- and long-term pollutant effects and impacts and to simulate the air pollution scenarios all over the world. The advances achieved in this field of air pollution during the past decades are due to numerous detailed investigations, the application of large number of techniques and the acquisition of abundant monitoring data.

The substances that accumulate in the atmosphere in suchconcentration for long enough that they may harm the living organisms or produce damage to building materials are called pollutants. The World Health Organization gives us the following definition of the air pollution: "Air pollution is the contamination of indoor or outdoor environment by any chemical, physical or biological agents that modifies the natural characteristics of the atmosphere."

[1] Air pollution can also be defined as "when gases or aerosol particles emitted anthropogenically build up inthe concentrations sufficiently high to cause direct or indirect damage to plants, animals, other life forms, ecosystems, structures or the works of art"

[2] Although both definitions refer to accumulation of pollutants in the atmosphere, but the second one is arestrictive definition to anthropogenic influence on air composition.

In this respect, air quality (AQ) collocation, which is often used to express status of air pollution, can be viewed as a measure ofthe anthropogenic perturbation of natural atmospheric state. The quality of air depends on the amount of pollutants, the rateat which they are released from various sources and how quickly the pollutants are getting deposited or dispersed. Goodair

quality refers to the clean, and unpolluted air. The meteorological conditions influence significantly the number of pollutants in a region: low winds, temperature inversions and topography with mountains can trap the pollutants close tothe ground, leading to an increased number of pollutants over the region. Conversely, the presence of strong and persistent wind over an area with significant pollutant emissions butlocated in a plain can disperse the air pollutants very quickly.

Air pollution comes from many different sources like factories, electrical power or chemical plants, chimneys, landfills, oil refineries, smelters, solid waste disposal farming, home and business activities, etc. Including, all transportation activities that are using cars, buses, trucks, trains, boats and airplanes contribute to the air pollution. [7] Pollution can also be the result from wildfires, volcanic eruptions, dust storms and windblown dust. As a result, the air pollutants can have naturalor anthropogenic sources, which could come from mobile (e.g.automobiles) or stationary sources (e.g. industrial facilities), could be emitted by the local sources and may travel or be formed over long distances affecting therefore large areas. Pollutants in atmosphere can be primary pollutants (emitted directly to atmosphere) or the secondary pollutants (formed by the chemical reactions involving primary pollutants and other constituents within the atmosphere). In highly populated metropolitan areas where air pollutants are the result from a combination of stationary sources and mobile sources, we encounter these so-called air pollution hotspots.

[2] Amidst the COVID-19 pandemic, extreme steps have beentaken by the countries globally. With the reduction in major anthropogenic activities, a visible improvement in the air quality has been recorded in urban centers.

Considering the closure of industries and restrictions on the community activities, including the inter-city transportationand mobility, it is expected that these measures will havesignificant effects on the amount of air pollutants from the industries and vehicles, especially in the large cities. [3] Important environmental pollutants include particulate matters (different combinations of the solid, liquid and vapour particles), which are mainly in two types of particles with an aerodynamic diameter of < 2.5 μm (PM2.5) and < 10 μm (PM10), tropospheric ozone [O3], nitrogen dioxide [NO2], Sulphur dioxide [SO2], carbon monoxide (CO).

Nitrogen oxides are caused by combustion of fossil fuels,which is considered to be a good indicator of vehicle-related airpollution. It has been shown that more than 50% and 23% of total nitrogen oxides (NOX) in developed and developing countries are related to transportation. [4] This pollutant can also react with the volatile organic compounds and the sun's ultraviolet rays to produce the tropospheric ozone, which posesa serious health threat like airways inflammation and increasing airway hyperreactivity. In the previous studies, it was shown that the particulate pollutants, e.g., PM2.5, PM10, NO2 and ozone were the most important pollutants contributing to COVID-19 both mortality and incidence.

The analysis aims to plot and visualize the air quality before the COVID and during COVID for the states of INDIA.

This project uses machine learning classifications using ScikitLearn to predict Air quality Index.

II. TECHNOLOGY CLASSIFIERS

A. PANDAS:

Pandas is a software library written for the Python programming language for the data manipulation and analysis.

In particular, it offers the data structures and operations for manipulating numerical tables and time series. It is a free software released under the three-clause BSD license.

B. SEABORN

Seaborn is a Python data visualization library which is based on matplotlib. It also provides a high-level interface for drawing attractive and informative statistical graphics.

C. PLOTLY

The plotly Python library is one of the interactive, open- source plotting library that supports over 40

unique chart types covering a wide range of the statistical, financial, geographic, scientific, and 3-dimensional use-cases.

D. SKLEARN

Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for Python programming language. It features various classifications, regression and clustering algorithms including the support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designedto interoperate with Python numerical and scientific libraries NumPy and SciPy. Scikit-learn is NumFOCUS fiscally sponsored project.

E. RANDOM FOREST

Random forests or random decision forests is an ensemble learning method for the classification, regression and othertasks that operates by constructing a multitude of decision trees at the training time. For the classification tasks, the output of random forest is the class selected by most trees. For the regression tasks, the mean or the average prediction of the individual trees is returned. Random decision forest is correct for decision trees habit of overfitting to their training set.

F. XGBOOST

XGBoost is an implementation of the gradient boosted decision trees designed for speed and performance that is dominative competitive machine learning.

G. SUPPORT VECTOR MACHINE (SVM)

SVM algorithm creates best line or decision boundary that can segregate n-dimensional space into the classes so that we can easily put new data point in the correct category in the future. This best decision boundary is known as ahyperplane.

III. PROPOSED METHODOLOGY

A. Data Collection – The Dataset is collected from kaggle and has 29,531 rows of data. The table given below summarises various features utilised.

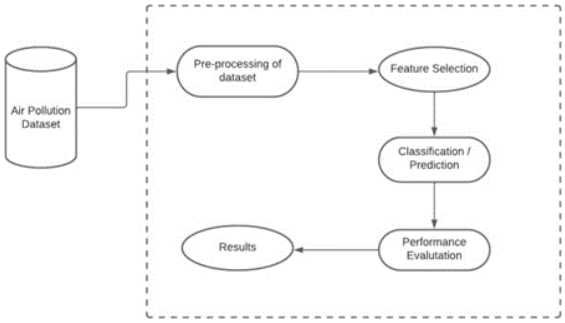
S. No.	Feature	Abbreviation
1.	City	City
2.	Date	Date
3.	PM2.5	Particulate matter 2.5 micrometer
4.	PM10	Particulate matter 10 micrometer
5.	NO	Nitrous oxide
6.	NO2	Nitrous dioxide
7.	NO-x	Any Nitrous x-oxide
8.	NH3	Ammonia
9.	CO	Carbon Monoxide
10.	SO2	Sulphur dioxide
11.	O3	Ozone
12.	Benzene	Benzene
13.	Toluene	Toluene
14.	Xylene	Xylene
15.	AQI	Air quality index
16.	AQI_Bucket	AQI categorized into buckets.

B. Data Pre-processing – The dataset contained many NULL and missing values. These were removed by adopting various strategies including filling missing values with the Mean, median and mode of existent values.

The correlation between the existing features is found by plotting a correlation matrix. The values that are closer to 1are positively correlated while those closer to –1 are negatively correlated. The Dataset is then divided into two parts that are Pre-Covid and Covid for accurate analysis and prediction of Air Quality indexBoth sections are divided into the training and test sets. A70/30 split is ensured for the training and test sets.[10]

C. Algorithms and Model Description – Several Supervised Machine learning models are used for the prediction task. The various ML algorithms which areused for the task are as follows –

Proposed Model



Logistic Regression: Logistic regression is an ML algorithm used for binary classification problems. It uses a logistic function to model a binary output. The range of the output is bounded between 0 and 1 and, it applies a non-linear Log transformation to the odds ratio. The Logistic function is described as –

$$Logistic\ function = \frac{1}{1+e^{-x}}$$

Logistic regression is a transformation of Linear regressionusing the sigmoid function. The y-axis gives the probability ofclassification while the x-axis gives value of feature. The formula for Logistic regression is –

$$F(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

The Logistic function applies a sigmoid function to restrictthe value of output to range of [0,1]

Support Vector Classifier: "Support Vector Machine" (SVM) is a supervised machine learning algorithm that can be used for both the classification or regression challenges. However, it is mostly used in the classification problems. [5] In SVM algorithm, we have to plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of particular coordinate. SVM divides the data points using the multidimension hyperplane.

The linear SVM finds the optimal hyperplane of the form f (x)
= w^T x + b where w is a dimensional coefficient vector and b isa offset. This is done by solving the subsequent optimization problem:

$$Min_{w,b,\xi_i} \frac{1}{2}w^2 + C\sum_{i=1}^{n}\xi_i$$
$$s.t.\ y_i\left(w^T x_i + b\right) \geq 1 - \xi_i, \xi_i \geq 0, \quad \forall_i \in \{1, 2, \ldots, m\}$$

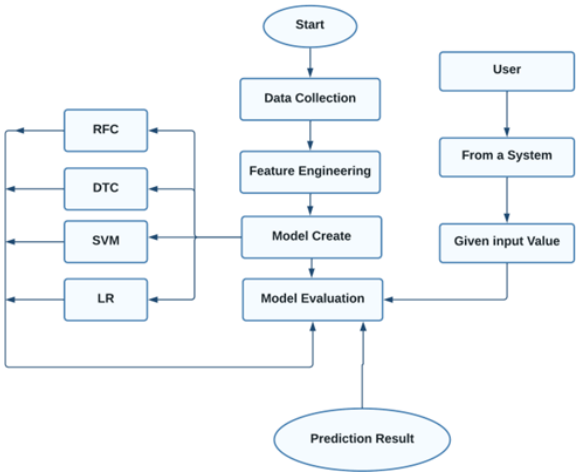
Random Forest: Random Forest is a Supervised Machine Learning Algorithm which is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for theclassification and average in case of regression. Random forestis a great choice if anyone wants to build model fast and efficiently as one of the best things about the random forest is itcan handle missing values.

Algorithm - Random Forest
Precondition: A training set S := (x1; y1),....., (xn; yn),features F, and number of trees in forest B.

1. *function* RANDOMFOREST(S , F)
2. H = ∅
3. for i in 1, ,B do
4. S(i) = A bootstrap sample from S
5. h(i) RANDOMIZEDTREELEARN(S(i), F)
6. H = H U {h(i)}
7. end for
8. return H
9. *end function*
10. *function* RANDOMIZEDTREELEARN(S , F)
11. At each node:
12. f = very small subset of F
13. Split on best feature in f
14. return The learned tree
15. *end function*

XG Boost: XGBoost is an algorithm that has been dominating applied machine learning and Kaggle competitionsfor structured or tabular data. XGBoost is an implementation of the gradient boosted decision trees designed for speed and performance. This algorithm goes by lots of different names like gradient boosting, multiple additive regression trees, stochastic gradient boosting or gradient boosting machine.

D. System Architecture



IV. RESULT AND DISCUSSION

Confusion Matrix: A confusion matrix is a way of calculatingmachine learning classification algorithms performance. Usingconfusion matrix we can calculate Accuracy, precision, recall, and F1.

		Actual Class	
		Positive (P)	Negative (N)
Predicted Class	Positive (P)	True Positive (TP)	False Positive (FP)
	Negative (N)	False Negative (FN)	True Negative (TN)

Accuracy: Accuracy = (True Positive+ True Negative)/ Total

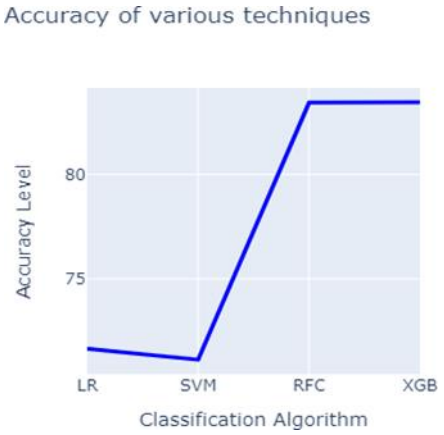


Figure 2. Accuracy of various techniques

Precision: Precision is the proportion of correctly predicted positive cases and the total predicted positive cases.

Precision = True Positive/ (True Positive + False Positive)

Recall: The recall is the proportion of correctly predicted positive cases to all the cases in the base class.

Recall = True Positive/ (True Positive + False Negative)

F1 Measure: The average of both precision and recall is F1 measure. Therefore, [9] this measure was taken both falsenegative and false positive. F1 is more useful than that of accuracy. If both false positive and false negative costs are very similar then accuracy is better but for very different casesF1 is suitable.

F1 score = (2 * Precision * Recall))/ (Precision + Recall).

Precision: Precision is the proportion of correctly predicted positive cases and total predicted positive cases.

Precision = True Positive/ (True Positive + False Positive)

Recall: The recall is the proportion of correctly predicted positive cases to all cases in the base class.

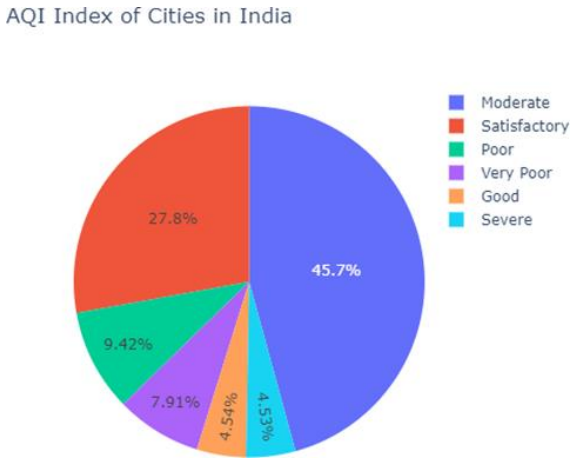
Recall = True Positive/ (True Positive + False Negative)

F1 Measure: The average of precision and recall is F1 measure.Therefore, this measure was taken both false negative and falsepositive. F1 is more useful than that of accuracy.[6] If both false positive and false negative costs are very similar then accuracy is better but for very different cases F1 is suitable.

F1 score = (2 * Precision * Recall))/ (Precision + Recall)

Techniques	Precision	Recall	F1 Score
LR	0.75	0.72	0.73
SVM	0.75	0.71	0.72
RFC	0.83	0.83	0.83
XGB	0.84	0.84	0.84

Table 1. Precision, Recall and F1 Score of different techniques



Out of 29,531 entries in dataset of 26 cities in India there are almost 4.53% (1338 days) which have severe AQI which can contains excess amount of major pollutants such as ground- level ozone, PM2.5, PM10, Carbon Monoxide, Sulphur Dioxide, Nitrogen Dioxide which can cause lung diseases such as emphysema, bronchitis and asthma and after that wheezing, chest pain, dry throat, headache or nausea and also lead to heart diseases. Also there is 7.92% have very poor AQI quality where risk of health effects is increased for everyone and9.42% have poor AQI quality in which sensitive people groups may experience more serious health effects than general public.[11][12]

V. FUTURE SCOPE

Monitoring Air pollution is an important application in many fields. For example, an IoT based model can be proposed for determining the pollution levels of its local space. Samples of pollutants can be collected for further analysis, reports can be generated based on collected samples and further predictions can be made in regards to the pollution levels of the place via the generated reports.

Further implementation could include pollution monitoring via drones and data of the same can be collected ina central database. Real time Predictions can be made and notified to individuals through a mobile application.

Further research and progress can be made by people working in the fields with the right and timely support from governmentauthorities.

CONCLUSION

The following conclusions are drawn from the above study :Major cause of generation of pollutants was common in all the monitored cities i.e. vehicles and automobiles. In past, several initiatives have been taken by Indian government to reduce air pollution across the country but still a lot of efforts need to be done to reduce it further. It is better to prevent the air pollution rather than allowing it to increase and well-known proverb 'charity begins at home' suits here. We are in need to understand the causes, effects and measures to reduce the pollution then only we can decrease pollution levels to ensure a better environment for future generations. More efforts are required for making the air quality monitoring a national issue by creating more awareness and following certain laws and rules to decrease the level of pollutants in air.

One needs to use public automobiles often and try to avoid theuse of excess personal automobiles as it will help to not only to decrease the excess vehicle congestion on roads but also help in decreasing pollution. Energy efficient commodities must be utilised to wherever possible as they tend to save energy along with having low negative effects. Existing policies and strategies that are need to be strengthened more to ensure more positive results. People should prefer public transport rather than using the personal vehicles. More vehicle evaluation on regular intervals needs to be done based on the air pollutants emissions and proper certification should be given to vehicles. Awareness needs to be created among the people related to health effects of air pollution so that people start taking it seriously and breathe a safer and pure air.

ACKNOWLEDGMENT

This paper and also the research behind it would not havebeen possible without the exceptional support of our mentor, Prof G. R. Rao. Her enthusiasm, knowledge and exactingattention to the details have been an inspiration and kept our work on track from our first encounter with the topic to the final draft of this paper. Prof G. R. Rao has shared the invaluable information from the book that she has been gathering for almost twenty years. We are also grateful for the insightful comments offered by anonymous peer reviewers at the Books & Texts. The generosity and expertise of one andall have improved the study in innumerable ways and saved us from many errors, those that inevitably remains are entirelyour own responsibility.

Studying about this topic has proved extremely costly and we are most thankful for the BHARATI VIDYAPEETH (DEEMED TO BE UNIVERSITY) Fellowship that has Provided the financial support for the larger projects fromwhich this paper grew. These things allowed us to continueour research with the book much longer than we could haveexpected. The final design of the complicated transcriptiontables in Appendices are the creative and technical work of theteam members, and the language and formatting of the paperhave benefited enormously. Finally, it is with true pleasure that we acknowledge the contribution of this amazing team,who has given up many Friday evenings and Sundayafternoons to read every version of this paper and theresponses it has generated with a combination of compassion and criticism that only he could master.

REFERENCES

[1] Air Pollution Monitoring: A Case Study from Romania (https://www.intechopen.com/chapters/52269)

[2] Air quality index variation before and after the onset of COVID-19 pandemic: a comprehensive study on 87 capital, industrial and polluted cities of the world (https://enveurope.springeropen.com/articles/10.1186/s12 302-021-00575-y)

[3] Huixiang Liu, Qing Li, Dongbing Yu, Yu Gu, "Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms", Applied Sciences, ISSN 2076-3417; CODEN: ASPCC7, 2019, 9, 4069; doi:10.3390/app9194069.

[4] Pooja Bhalgat, Sejal Pitale, Sachin Bhoite, "Air Quality Prediction using Machine Learning Algorithms", International Journal of Computer Applications Technology and Research Volume 8–Issue 09, 367-370, 2019, ISSN:-2319–8656.

[5] Ziyue Guan and Richard O. Sinnot, "Prediction of Air Pollution through Machine Learning on the cloud", IEEE/ACM5th International Conference on Big Data Computing Applications and Technologies (BDCAT), 978-1-5386-5502-3/18/$31.00 ©2018 IEEE DOI 10.1109/BDCAT.2018.00015.

[6] Heidar Malek, Armin Sorooshian, Gholamreza Goudarzi, Zeynab Baboli, Yaser Tahmasebi Birgani, Mojtaba Rahmati, "Air pollution prediction by using an artificial neural network model", Clean Technologies andEnvironmental Policy, (2019) 21:1341–1352.

[7] Air pollution. 05 March 2013 Available from http://www.cseindia.org/node/2094.

[8] Rizwan SA, Baridalyne Nongkynrih,Sanjeev Kumar Gupta. Air pollution in Delhi: Its magnitude and effects on health. Indian J Community Medicine. 2013.

[9] Air quality trends. National ambient air quality monitoring programme. CPCB.

[10] Koo Ping Shung, https://towardsdatascience.com/accuracy-precisionrecall-or-f1-331fb37c5cb9

[11] Air Quality Index (AQI) Basics https://www.airnow.gov/aqi/aqi-basics/

[12] Health Effects https://www.sparetheair.com/health.cfm#:~:text=Aggravated%20respiratory%20disease%20such%20as,Reduced%20resistance%20to%20infections