International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT2283118

Predictive Disease Data Analysis of Air Pollution Using Supervised Learning

Manikanta Sirigineedi [*1], Padma Bellapukonda [2], R N V Jagan Mohan [3]

[*1]Department of Information Technology, Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India
[2]Department of Information Technology, Shri Vishnu Engineering College for Women, Bhimavaram, Andhra Pradesh, India
[3]Department of Computer Science and Engineering, SRKR Engineering College, Bhimavaram, Andhra Pradesh, India

Article Info
Publication Issue :
Volume 8, Issue 4
July-August-2022

Page Number : 105-110

Article History
Accepted: 05 July 2022
Published: 16 July 2022

ABSTRACT

Air pollution is a combination of natural and manmade substances in the air we breathe. It is classified into two major categories, i.e. outdoor air pollution and indoor air pollution. Outdoor air pollution involves exposures that take place outside the built environment where as, indoor air pollution involves exposure to particulates, carbon oxides, and other pollutants carried by indoor air or dust. In this paper, we would like to propose that air pollution relates to increased cardiovascular and breathing related problems data rate, prediction with supervised machine learning. The study is largest of its benevolent to investigate the short-term impacts of air pollution is conducted completed a 30-years epoch. This study analyzes the experiments data on air pollution and humanity in India and other regions. The experimental result is on Risk of Cardiovascular Illness in several patients data classification is used.

Keywords: Air pollution, supervised machine learning, carbon dioxide, cardiovascular, breathing, asthma, Radon Gas.

I. INTRODUCTION

Machine learning is a benevolent that provides computers with the ability to learn without existence openly involuntary. It is a well-known method of Predictive data analysis and focuses on the development of computer programs such that they can teach themselves to grow and change when exposed to new data. The process of machine learning is similar to that of courses of actions in data mining. These systems are seeking through the data to look for patterns. However, instead of extracting data for human comprehensions the circumstance in data mining application like medical applications and machine learning uses that data to improve the program's own understanding. Machine learning programs detect patterns in data and adjust program actions accordingly. Machine Learning is closely related to Computational Statistics in which helps us in making predictions or decisions. Exposure to toxic air pollutants linked to increased cardiovascular and respiratory death rates, according to a new study by researchers. The study to analyze data on air pollution and mortality in cities across countries and regions,

and found increases in total deaths linked with exposure to inhalable particles (PM10) and fine micro particles (PM2.5) emitted from fires or formed through atmospheric chemical transformation. There is no threshold for the association between particulate matter (PM) and mortality, even low levels of air pollution can increase the risk of death. Radon gas pollution also leads to lung cancers, which is very much prevalent in Indian context. People who live in surrounding areas of Industrial Estates are much prone to Radon gas impact. This is leading to lung cancers and Brain cancers in some of the cities. Given the extensive evidence on their health impacts, PM10 and PM2.5 are regulated through the World Health Organization (WHO) Air Quality Index Guidelines and standards, however more attention to the sudden increase in air pollution. The results are comparable to previous findings in other multi-city and multi-country studies, and suggest that the levels of particulate matter below the current air quality guidelines and standards are still hazardous to public health. Asthma is an increasing cause of concern in both adults and children. It has increased it is prevalence in the last five decades for so many times. Nearly 20% of the Indian population is suffering from Asthmatic attacks particularly in the winter seasons. Particularly, the PM2.5 (particulate matter) is proved most dangerous for asthma patients and even ordinary people too. Over the past 30 years, researchers have dig up a wide array of health effects in which it is believed to be associated with air pollution exposure. Among those are respiratory diseases including asthma and changes in lung functioning, cardiovascular diseases, adverse pregnancy outcomes (such as preterm birth), and sometimes death. While climate change is a global process, it has very local impacts that can profoundly affect some communities, not the least of which is air pollution. Increasing temperatures are directly linked to poor air quality, which in turn can affect the heart and exacerbate cardiovascular disease. This paper might be include a rise in pollen, due to increased plant growth, or a rise in molds, due to severe storms, both of which can worsen allergies and other lung diseases, such as asthma. Outdoor air pollution exposures can be reduced by checking one's Air Quality Index (AQI), avoiding heavy traffic when possible, and avoiding passive tobacco smoke.

To appreciate them respond of this paper positioned on seven sections: The general introduction Section-1 deals with the Air Pollution Trajectory Clustering Network Analysis. In section-2 is the Predictive Air Pollution Data Analysis Using Supervised Learning. In section-3is about Linear Regression. The section-4 deals with Foreseeable result. The conclusion is in section-5. The last section-6 is reference.

II. Air Pollution Trajectory Clustering Network Analysis

The efficient way to calculate the inbuilt of data and unknown schemes is clustering. With the development of GPS devices, maximum-to-maximum number of objects, which are travelling on air, can also be recorded. Now the concentration on moving object identification is improved. The air pollution contains different oxides, which can be easily identified this technology.

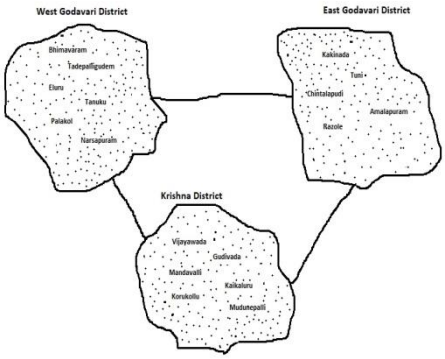

Figure-A: Air Pollution Trajectory Clustering Network

In this technology first of all the moving particles can be identifies first. Later the similarities and non-similarities of these particles are determined by the trajectories. Finally, the result, which is given by this trajectory, is correct or not will be check. The

pollutant particles in the air will be identified by these trajectories and will form a cluster for easy identification. With observing the above network diagram we can easily understand the concept of trajectory clustering network. Here west Godavari, Krishna and East Godavari are considering as cluster each one.

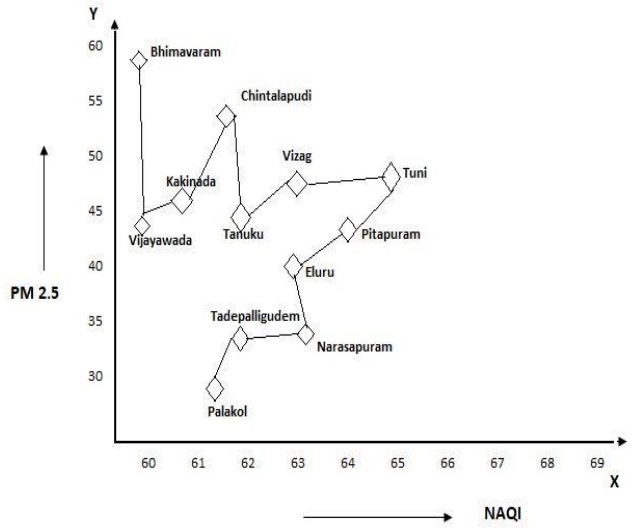

Figure-B: Graph on Air Pollution Trajectory Clustering Network for Rural Areas

While considering the above graph we can easily observe that all the values of PM2.5 and NAQI are at peak level for some areas. In addition, remaining areas compete with the each other in those values.

III.PREDICTIVE AIR POLLUTION DATA ANALYSIS USING SUPERVISED LEARNING

Supervised machine learning is the exploration for algorithms, which produces general hypotheses from externally supplied instances and then makes predictions about future instances. The basics of analyst features are used to build a brief model of the circulation of class labels, as it is the main aim of supervised learning. Then, the output classifier is used to allocate class labels to the testing instances where the values of the interpreter features are known, but the value of the class label is unknown. A Predictive analysis is usually pave the way by a systematic review, as this allows identification and critical appraisal of all the relevant evidence (thereby limiting the risk of bias in summary estimates).

The Steps are as follows:

1. Predictive Source Analysis of Air Pollution allied to Risk of Cardiovascular Disease.
2. Find the PatientHistoryof Air Pollution allied to Risk of Cardiovascular Disease.
3. Selection of studies ('incorporation criteria') Based on quality criteria, e.g. the requirement of randomization and blinding in a clinical trial Selection of specific studies on a well-specified subject, e.g. the treatment of cardiovascular diseases and lung cancer.
4. Decide whether unpublished studies are included to avoid publication bias (file drawer problem) decide which dependent variables or summary measures are allowed. For instance, aggregate the data.
5. Selection of a prediction analysis, e.g. fixed effect or random effects diagnosis analysis.
6. Examine sources of between-study heterogeneity, e.g. using subgroup analysis or regression.

IV. LINEAR REGRESSION

Regression refers to a machine learning technique where as “LINEAR” refers to a straight line. This means that, when we draw a graph between the variables present in the given problem, if the points are situated as a cream around a straight line, then those variables are said to be have “Linear regression”. The main purpose of linear regression is to predict and forecast values based on cardio disease analysis of Air Pollution information. There are two types of variables i.e., targeting variables and explanatory variables influence the regression and they are the major elements in achieving Linear Regression. Using Linear Association, we can identify the impact of Target variables on Explanatory Variables and on their change.

The mathematical Expression for regression formula is y=ax+e

There are other formulas also necessary to calculate slope of Regression line and for intercept point of regression.

This formula isthe slope of the regression line is given by: a= (N∑xy-(∑x)(∑y))/(N∑x2 –(∑x)2)

This formula is the intercept point of regression is given by: e= (∑y-b (∑x))/N

Where x and y are the variables that form a dataset and N is the total number of values.

A. FORESEEABLE RESULT ON AIR POLLUTION:

As per the information is most polluted cities among the countries are very high. According to PM2.5 and NAQI (National Air Quality Index), the pollution severity is as follows and based upon these values only we can decide the concentration of pollution. In the below table consider the values of NAQI and PM 2.5 for some of the rural areas.

The experiment result is air pollution data analysis of West Godavari, East Godavari and Krishna District. It is a linear relation between the NAQI and PM2.5.

S. No	NAQI	PM 2.5	Rural Areas of West Godavari
1.	68	59	Bhimavaram
2.	62	56	Tadepalligidem
3.	66	59.5	Eluru
4.	63	48	Tanuku
5.	61	45	Palakol
6.	60	42	Narasapuram
Total	380	309.5	

In this regard we have to find out regressive line and estimate the severity of NAQI value is 425.

As per the above table the n=6, because the count of the rural areas taken from west Godavari is 6.

Here, ∑x=380,∑y=309.5, $\sum x^2$ =144400 and ∑x*∑y =117610

So,

$$a_0=\frac{(309.5*144400-380*117610)}{(6*144400-144400)}$$

$$a_0=\frac{(44691800-44691800)}{(866400-144400)}$$

$$a_0=0$$

And

$$a_1=\frac{(6*117610-380*309.5)}{(6*144400-144400)}$$

$$a_1=\frac{(705660-117610)}{(866400-144400)}$$

$$a_1=\frac{588050}{722000}$$

$$a_1=0.814$$

The regression line is,

$$a_0+a_1=0+0.814$$

$$a_0+a_1=0.814$$

So at the critical value 425 of NAQI

That is, 0.814*425 = 345.95

≈ 346

The severe of the West Godavari NAQI WHO's severe value is 425 that is equal to 346.

Now regression line for East Godavari is calculating here

S. No	NAQI	PM 2.5	Rural Areas of East Godavari
1.	67	58	Kakinada
2.	64	50	Tuni
3.	62	48	Amalapuram
4.	61	42	Razole
5.	62	46	Chintalapudi
6.	63	41	Gannavaram
Total	379	244	

In this regard we have to find out regressive line and estimate the severity of NAQI value is 425.

As per the above table the n=6, because the count of the rural areas taken from East Godavari is 6.

Here, ∑x=379,∑y=244.5, $\sum x^2$ =143641 and ∑x*∑y=92476

So,

$$a0 = \frac{(244*143641 - 379*92476)}{(6*143641 - 143641)}$$

$$a0 = \frac{(35048404 - 35048404)}{(861846 - 143641)}$$

a0=0

And

$$a1 = \frac{(6*92476 - 379*244)}{(6*143641 - 143641)}$$

$$a1 = \frac{(554856 - 92476)}{(861846 - 143641)}$$

$$a1 = \frac{462380}{718205}$$

a1 = 0.643

The regression line is,

a0+a1 = 0+0.643

a0+a1 = 0.643

So at the critical value 425 of NAQI

That is, 0.643*425 = 273.275

≈ 273

The severe of the East Godavari NAQI WHO's severe value is 425 that is equal to 273.

Now regression line for Krishna District is calculating here,

S. No	NAQI	PM 2.5	Rural Areas of Krishna District
1.	69	59	Vijayawada
2.	66	55	Gudiwada
3.	64	49.5	Mandavalli
4.	63	52	Kakinada
5.	61	51	Korukollu
6.	65	54	Mudinepalli
Total	388	320.5	

In this regard we have to find out regressive line and estimate the severity of NAQI value is 425.

As per the above table the n=6, because the count of the rural areas taken from Krishna District is 6.

Here, $\sum x=388, \sum y=320.5, \sum x^2 =150544$ and $\sum x^* \sum y=124354$

So,

$$a0 = \frac{(320.5*150544 - 388*124354)}{(6*150544 - 150544)}$$

$$a0 = \frac{(48249352 - 48249352)}{(903264 - 150544)}$$

a0=0

And

$$a1 = \frac{(6*124354 - 388*320.5)}{(6*150544 - 150544)}$$

$$a1 = \frac{(746124 - 124354}{(903264 - 150544)}$$

$$a1 = \frac{621770}{752720}$$

a1 = 0.826

The regression line is,

a0+a1 = 0+0.826

a0+a1 = 0.826

So at the critical value 425 of NAQI

That is, 0.826*425 = 351.05

≈ 351

The severe of NAQI WHO's value is 425 and for the West Godavari, East Godavari and Krishna Districts severe values are 346, 273 and 351.

When compared among three districts, the Krishna District air pollution linear regression severe value is high. In general also we know that when compared with these three districts Krishna district's air pollution level is high. So the derived values are true.

IV.CONCLUSION

As shown in the above we have already seen that with the linear regression formula and PM, NAQI values we can able to calculate the actual level of pollution in each area. With knowing the values like x, y and n, we can easily calculate the severe level of air pollution for the particular area. So with this we can know the actual severity level of the air pollution. Then we can alert local authorities to make necessary arrangements to reduce the depth of the pollution problem. Because many are died with this pollution

problem and so many in sick also. Mostly the weak, such as children, sick elderly are at risk write now. In this regard the predictive disease data analysis of air pollution method is so useful to reduce air pollution in danger areas with alerting local authorities.

V. REFERENCES

[1] Erik Melén: Air pollution and IgE sensitization in 4 European birth cohorts—the MeDALL project, Journal of Allergy and Clinical Immunology, Vol 147, Pages: 713-722 , 2021.

[2] Melén E: Air pollution exposure and allergic sensitization during childhood and adolescence in four European birth cohorts - the MeDALL project, Environmental Epidemiology,2019.

[3] Wolf K: Low-level air pollution and incidence of acute coronary events, Environmental Epidemiology, Vol 3, Pages:443, 2019.

[4] Liu S: Low-level air pollution and incidence of asthma among adults, Environmental Epidemiology, Vol 3, Pages:246,2019.

[5] Stafoggia M: Low-level air pollution and natural cause mortality in Europe, Environmental Epidemiology, Vol 3, Pages:380-381,2019.

[6] Joachim Heinrich:Traffic-Related Air Pollution Exposure and Asthma, Hayfever, and Allergic Sensitisation in Birth Cohorts: A Systematic Review and MetaAnalysis, Geoinformatics & Geostatistics: An Overview,vol 4, 2016.

[7] Zorana Jovanovic Andersen: Long-term Exposure to Ambient Air Pollution and Incidence of Brain Tumor in 12 European Cohorts: the European Study of Cohorts for Air Pollution Effects (ESCAPE), ISEE Conference Abstracts,Vol 2016, 2016.

[8] Patrizia Schifano: Heat, Air Pollution and Preterm Birth: Which Weeks of Gestation Are Susceptible? The Rome and Barcelona Birth Cohorts, ISEE Conference Abstracts, Vol 2014, Pages:2534,2014.

[9] M. Srikanth, R. N. V. Jagan Mohan, M Chandra Naik: Blockchain based Crop Farming Application Using Peer-to-Peer, xidian journal, Volume 16, Pages, 168 – 175, 2022.

[10] M. Srikanth, R. N. V. Jagan Mohan: Query Response Time in Blockchain Using Big Query Optimization, Apple Academy Press and CRC Press,2022.

[11] M. Srikanth, R. N. V. Jagan Mohan: Stop spread corona based on voice, face and emotional recognition using machine learning, query optimization and Block chain Technology, Solid State Technology, Vol. 63 No. 6, 2020.

[12] M. Srikanth, R. N. V. Jagan Mohan: Machine Learning for Query Processing System and Query Response Time using Hadoop, IJMTST,2020.

[13] Shrawan Kumar, S Rajeswari, M Srikanth, T Raghunadha Reddy: A New Approach for Authorship Verification Using Information Retrieval Features, Springer, Pages 23-29, 2019.

[14] Nagendra Panini Challa,Suma Bharathi T,Padma B,Manikanta Sirigineedi,JS Shyam Mohan,GP Siva Kumar:Recent Trends, Challenges and Applications of Cyber Physical Systems and Internet of Things,Information Classification,2022.

[15] M. Srikanth: An Enhanced and Naive Clustering Algorithm for Text Classification Based on Weight, International Journal & Magazine of Engineering, Technology, Management and Research, Vol 1, Pages 7, 2014.

[16] M. Srikanth, R. N. V. Jagan Mohan: Block-level based Query Data Access Service Availability for Query Process System, IEEE, 2020.

[17] M. Srikanth, Padma Bellapukonda, Manikanta Sirigineedi: Protecting tribal peoples nearby patient care centres use a hybrid techniques based on a distribution network, International Journal of Health Sciences, 2022.

Cite this article as :

Manikanta Sirigineedi, Padma Bellapukonda, R N V Jagan Mohan, "Predictive Disease Data Analysis of Air Pollution Using Supervised Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 4, pp. 105-110, July-August 2022. Available at doi : https://doi.org/10.32628/CSEIT2283118

Journal URL : https://ijsrcseit.com/CSEIT2283118