International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 02, 2020

ISSN: 1475-7192

DETECTION AND PREDECTION OF AIR POLLUTION USING ML MODELS

P.VEERESH. Y.NARASIMHA REDDY,.
Associate Professor & Head of CSE Dept, Associate Professor,
Department of CSE
St.Johns College of Engineering and Technology, Yemmiganur, Kurnool (Dist).

Abstract
Governments in both developed and developing countries are fully aware that air quality control is a crucial responsibility that must be completed. Conditions such as weather and traffic congestion, fossil fuel burning, and industrial features such as power plant emissions all have a substantial impact on environmental contamination and are thus considered to be environmental polluting factors. In terms of influence on air quality, particulate matter (PM 2.5) is the most significant of all the particulate matter that can be measured, and it deserves more attention than it now receives. Human health may be negatively affected when there is an excess of ozone in the air, which is conceivable when the amount of ozone is high in the atmosphere. No amount of emphasis can be placed on how vital it is to monitor its concentration in the atmosphere on a regular basis in order to effectively control it. In this study, logistic regression is used to determine if a data sample is contaminated or not polluted, based on the distribution of the data sample data. It is possible to estimate future levels of PM2.5 using authoregression, which is a statistical method that is based on previously gathered data. Being aware of the amount of PM2.5 that will be present in the air in the following years, months, or weeks allows us to work toward lowering its concentration to levels lower than those considered to be hazardous. Based on a data collection that includes daily atmospheric conditions in a certain city, this technique was developed to attempt to anticipate PM2.5 levels and identify air quality in a given place.
Keywords — Pollution detection, Pollution Prediction, Logistic Regression, Linear Regression, Autoregressio

1. INTRODUCTION

Throughout the history of our planet, air has been regarded as the most important characteristic asset for the survival and existence of all life, and it is absolutely necessary for the survival and presence of all life. Today, air is considered the most important characteristic asset for the survival and existence of all life. Aerial oxygen is required by all forms of life, including plants and animals, for their essential endurance and presence in order to live and to be present in their surrounding environment. As a result, in order to thrive, all living things require a large amount of clean, fresh air that is free of harmful gases in order to sustain their existence. An alarming

amount of pollution is being released into the atmosphere by an expanding global population, its automobiles, and commercial enterprises all over the world. Depending on the conditions, being exposed to a polluted environment can have a variety of long- and short-term repercussions for a person's health.

Given that air quality forecasting models can provide early warnings when pollution levels are anticipated to approach unsafe levels of contamination, it may be beneficial to improve the accuracy of air quality forecasting models in certain situations. Vehicle exhaust, industrial facilities, and activities with a restricted geographic reach are some of the most prominent causes of pollution fixations in metropolitan areas. A variety of factors, including the following, have the potential to negatively impact human health and the environment as a result of air pollution, including: A high concentration of pollutants in the air, such as particulate matter (10 microns and 2.5 microns), carbon monoxide (CO), and Nitrogen Oxides (NO+NO2), causes thousands of people to die prematurely every year. In recent years, the city's urban and contemporary neighbourhoods have suffered greatly from poor air quality, which has been exacerbated by the tremendous natural piling that has occurred in the area. On the next day, we officially launched our investigation into the air quality in Ghaziabad, which was the formal start of our investigation. There has been a substantial amount of public interest in the task of monitoring and improving the quality of the air we breathe in recent years. This research's overarching purpose is to identify and evaluate acceptable artificial intelligence strategies that can aid in better forecasting the fixation of air pollution, which is the primary focus of this investigation. The data collecting mechanism is provided by CPCB (Central Pollution Control Board) internet information and sensors strategically positioned across the target region. Whenever the spread of suspended particles such as PM10, PM2.5, SO2, and NO2 in polluted climatic air is acknowledged, it is referred to as having occurred.

It is possible to analyse the material for up to 5 months, utilising information mining techniques, in order to identify patterns and trends in the information (2017-2018). As a result of our research, we discovered that climatic factors had a significant impact on air contamination throughout the whole process of forecasting air contamination. We take into account a variety of meteorological elements in order to predict air pollution, such as temperature, minimum temperature, maximum temperature, wind speed, relative humidity, and a few other features in our research. For example, we consider temperature, minimum temperature, maximum temperature, wind speed, relative humidity, and a few other features. As far as projecting air quality in urban areas and contemporary sections of Ghaziabad is concerned, this will be a significant resource for government organisations who are reviewing and developing potential air pollution measures in the future. Our investigation is predicated on the idea that the degree of air pollution at a specific or specified point in a certain area is high enough to warrant investigation. When it comes to air pollution forecasting, precision, productivity, and adaptability of models are all important considerations to consider.

2. LITERATURE SURVEY

It is the technique for doing a literature review that is the most critical phase in the software development process. It is vital to examine the time factor, the economics, and the overall strength of the organisation before creating the tool. In order to design and develop the tool, it is necessary to first choose which operating system and programming language will be utilised. After these requirements have been met, the eleventh stage may begin. The programmers will require a significant amount of assistance from other sources in order to accomplish their task once they begin working on the instrument. For example, older programmers may provide guidance, as can books and websites, among other sources of information and assistance. It is vital to analyse the aspects described above before to commencing development on the proposed system in order to verify that it will perform effectively.

On this paper, Ian H. Witten cooperated with authors such as Eibe Frank, Mark A. Hall, and Christopher J. Pal, as well as other authors. When applied to a real context, data mining refers to the application of machine learning techniques and approaches to solve problems. Originally released in 2016, this book was published by Morgan Kaufmann Publishing Company. 'Data Mining: Practical Machine Learning Tools and Techniques' is a must-have resource for anyone studying, teaching, or researching data mining methods and applications. It is a must-have resource for anyone interested in learning about, implementing, and deploying data mining methods and applications, and it is a must-have resource for anyone interested in data mining methods and applications. This book discusses aspects of machine learning, intelligent systems, bioinformatics, and biomedical informatics that are relevant to undergraduate or graduate programmes in these subjects, as well as related fields, and is intended for students pursuing undergraduate or graduate degrees in these subjects or related fields.

The research "Data mining to simplify policy making in air pollution management," written by S. Li and L. Shue, and published in Expert Systems with Applications, explores the use of data mining to assist in policy making in the field of air pollution management. In this publication, S. Li and L. Shue are both co-authors (vol. 27, pgs. 331-340, 2004).Our mining method, which we cover in depth in this article, was used to extract information from a meteorological dataset.

The information for this study came from the DAV BDL public school in Bhanur, as well as the Medak weather station, which both gave information. Daily weather measurements were taken and saved in the data collection during the course of four years [2011-2015]. Clustering, association, and classification are three data mining methodologies that we are looking to put into practise in our research and development efforts. Among other things, we measured the following parameters for the aim of obtaining meteorological data: temperature, pressure, relative humidity and dew point, wind speed, precipitation amount, and wind direction, among other things. By a third parameter, one of these parameters is associated to the other parameter. Outliers were identified, data was analysed, and experimental findings were interpreted using the assistance of the Weka data mining tool and graphs created by the Excel tool, respectively.

Outliers were discovered, data was analysed, and experimental findings were interpreted These tools deliver information in the form of tables and graphs, and the information they provide is both valuable and dependable. Agriculture, air pollution, disaster management, and even weather forecasting are some of the areas that might benefit from this type of knowledge. In the coming years, we will concentrate our efforts on the development of an autonomous, efficient, and accurate weather forecast system.

3. System analysis

3.1 Existing system:

Even when applied to new issue settings, such as weather forecasting and climate change, the characteristics of air pollution that constantly fluctuate generate new challenges for these well-known study issues, which include: On the other side, people frequently create their tweets in a casual way, which is a terrific trend in the social media sector right now. When acronyms, misspellings, and special tokens are used, tweets get clogged, and methods that have been created for formal written communication become useless when applied to tweets, as seen in the following example.

The following are some of the downsides of choosing this option: • It is the most cost-effective choice that you have accessible to you at this time.Inaddition, its performance falls short of expectations.

According to the recommendations made, the following components should be included in the proposed system:In addition to being based on a specific city's air pollution data record collection, the prediction models that have been provided are also transferable and may be used to other types of reports, albeit with some alterations. Foremost, it is necessary to determine whether or not the three geolocation challenges on cities, which are as follows: forecasting home location based on pollution rate, city location, and the previously mentioned location, are applicable to the target platform before considering model adaptations and model enhancements. Then model adjustments and model upgrades can be considered, if they are appropriate. Depending on where the city is located, it may not be able to rely on forecasted weather forecasts posted on different photo and video sharing websites, as well as other information reporting platforms.

Advantages:

There are more cost-effective solutions accessible to you, therefore this is not the best choice.

Overall, it has been an outstanding performance in all areas of the business.

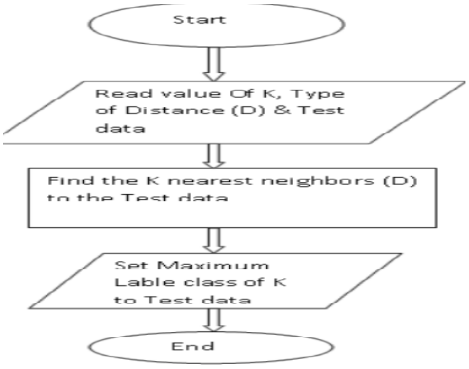
4. ALGORITHMS

4.1 K-Nearest Neighbors Algorithm:

Nonparametric supervised learning approach that makes use of training sets to classify data points into defined groups of data points based on their attributes is used to classify data points into specified groups of data points. Check out the following basic classification: the term gets

information from all educational contexts and compares it to the new example. In this stage, you will examine the training data for the K examples that are the most comparable (neighbours) to the new instance (x), and you will forecast the new instance (x) by averaging the output variables for the K cases examined in the previous step. If you look at a graphical depiction, the classification mode is represented by the class value mode (or most commonly). Figure 2 is a flow diagram of the KNN algorithm in operation, which illustrates how the algorithm works in practise.

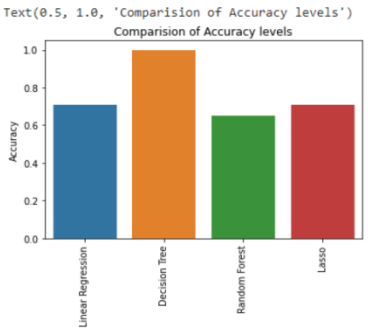
FLOWCHART



KNN Algorithm Flowchart

5. Results

	StationId	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
0	AP001	2017-11-24	71.96	115.75	1.75	20.65	12.40	12.19	0.10	10.76	109.26	0.17	5.92	0.10	NaN	NaN
1	AP001	2017-11-25	81.40	124.50	1.44	20.50	12.08	10.72	0.12	15.24	127.09	0.20	6.50	0.06	164.0	Moderate
2	AP001	2017-11-26	78.02	129.06	1.26	26.00	14.85	10.25	0.14	26.96	117.44	0.22	7.95	0.08	197.0	Moderate
3	AP001	2017-11-27	88.76	135.32	6.80	30.85	21.77	12.31	0.11	33.53	111.81	0.29	7.63	0.12	158.0	Moderate
4	AP001	2017-11-28	64.16	104.09	2.56	28.07	17.01	11.42	0.09	19.00	138.18	0.17	5.02	0.07	168.0	Moderate



Bar Graph of Air pollution

Conclusion

It goes without saying that, given the efficacy of the suggested approach in other cities, there is little doubt that it will be useful in the future development of air pollution forecasting techniques in our smart city. Timing Series Prediction may be performed via the use of Multivariate Multistep Techniques. As seen in Figure 1, the use of the Random Forest approach improves the performance of the air pollution prediction model while also decreasing its complexity. In addition, we are utilising the feature selection approach, which aids in the improvement of the accuracy of our forecasting outcomes.

Increasingly sophisticated machine learning algorithms are being developed all of the time. Future smart cities will need to be able to anticipate and evaluate air quality in real time, and this will be a requirement in the near future. The prediction of air pollution in a given or specific location is at the centre of our investigation. Forecasting air pollution requires careful consideration of a number of factors, including the accuracy, efficiency, and flexibility of the models. The results of a recent examination of air pollution levels in the Ghaziabad region have generated good results, which are provided in this project, along with a comparison of the results acquired using different approaches.

A characteristic of this system allows customers to view the expected quantity of pollution on their mobile phones through websites that they may access using their mobile devices, which is one of the features of this system. Giving citizens the opportunity to participate in the process adds an additional layer of value to the final product. As a result of the fact that everyone is equally aware of and concerned about the environment, the notion of an air pollution prediction system is advantageous to the overall well-being of the community. Furthermore, it is carried out through the application of the most up-to-date technical solutions now accessible. A number of notes are made on potential issues and needs throughout the article, as well.

Reference:

1. Witten, Ian H., Eibe Frank, Mark A. Hall, and Christopher J. Pal. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

2. Li S., and Shue L., "Data mining to aid policy making in air pollution management," Expert Systems with Applications, vol. 27, pp. 331-340, 2004.

3. Gu, Ke, JunfeiQiao, and Weisi Lin. "Recurrent Air Quality Predictor Based on Meteorology and Pollution Related Factors." IEEE Transactions on Industrial Informatics (2018).

4. García Nieto, P.J., Sánchez Lasheras, F., GarcíaGonzalo, E. et al. "Estimation of PM10 concentration from air quality data in the vicinity of the major steel works site in the metropolitan area using machine learning techniques"Stoch Environ Res Risk Assess (2018), https://link.springer.com/article/10.1007/s00477- 018-1565-6.

5. Hu, Ke, AshfaqurRahman, HariBhrugubanda, and Vijay Sivaraman. "Hazeest: Machine learning based metropolitan air pollution estimation from fixed and mobile sensors." IEEE Sensors Journal 17, no. 11 (2017): 3517-3525.

6. K. B. Shaban, A. Kadri, and E. Rezk, "Urban Air Pollution Monitoring System With Forecasting Models" IEEE Sensors Journal, vol. 16, no. 8, pp. 2598–2606, Apr. 2016.

7. Xiao Feng, Qi Li, Yajie Zhu, "Artificial Neural Network Forecasting of PM2.5 Pollution using Air Mass Trajectory based Geographic Model and Wavelet Transformation" Atmospheric Environment Journal, www.elsevier.com/locate/atmosenv, 2015.

8. M. S. Baawain and A. S. Al-Serihi, "Systematic approach for the prediction of ground-level air pollution (around an industrial port) using an artificial neural network," Aerosol and Air Quality Research, vol. 14, pp. 124–134, 2014.

9. W.-Z. Lu and D. Wang, "Learning machines: Rationale and application in groundlevel ozone prediction," Applied Soft Computing`, vol. 24, pp. 135–141, Nov. 2014.

10. A. Sotomayor-Olmedo, M. A. Aceves-Fernández, E. Gorrostieta-Hurtado, C. Pedraza-Ortega, J. M. RamosArreguín, and J. E. Vargas-Soto, "Forecast Urban Air Pollution in Mexico City by Using Support Vector Machines: A Kernel Performance Approach," International Journal of Intelligence Science, vol. 3, no. 3, pp. 126–135, Jul. 2013.

11. ShwetaTaneja, Dr. Nidhi Sharma, KettunOberoi, YashNavoria ,"Predicting Trends in Air Pollution in Delhi using Data Mining", IEEE(2016)