International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

Foresight of Health Risk Based on Air Pollutants' Air Quality Index Values

Aiswarya Johney, Namitha S J, Leena Vishnu Namboothiri

Abstract: ***This proposed work is mainly focused on the drastic air pollution data in various metropolitan cities. Rapidly growing industrialization both in automobiles and other public sectors massively increases the intensity of air pollution which drastically affects the nature and human rights for decades. The most destructive part of air pollution is that it may badly cause severe immunity problems for both the flora and fauna as well the human life. It might be life-threatening if the air pollutants cross its limit. Different software/tools are used for the prediction of air contamination and it is a formidable one. In this paper, we aim to find an accurate algorithm for implementing a system by utilizing Weka Tool for the prediction of health risks. Algorithms used are Decision tree J48 and Multiclass Classifier. Prediction of health risks is done based on different AQI values of air pollutants such as*** NO_2***,*** O_3***, CO and,*** SO_2***.***

Keywords : Air Quality Index; Data Mining; J48; Multiclass Classifier, Weka Tool.

I. INTRODUCTION

Air pollution is the contamination of air caused by the endurance of toxic substances which cause damage to the well-being of humans in the environment due to the effects of harmful gases, mainly CO2, NO2 and, SO2. Air Pollution is the major cause of fitness concerns such as nasal issues, lung cancer, bronchitis, epidermis infections, and heart problems. Air pollution is caused due to pollutants and particulate matters present in the air which are dangerous to take in. Due to the development of huge industries and enlarged use of public and private vehicles, there is an upsurge in the contamination which leads to a drastic variation in the climate which in turn leads to global warming. So by considering, estimation of air pollution levels is important to decide the necessary action plans to decrease air pollution. Thus, we can predict the escalating levels of air pollution which are merely injurious to sensitive people just by analyzing the air. The quality of atmospheric air is evaluated for the health problems relating to air quality levels.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
Aiswarya Johney*, P G Student, Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidyapeetham, India. Email: aiswaryajohney09@gmail.com
Namitha S J, P G Student, Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidyapeetham, India. Email:namitha7700@gmail.com
Leena Vishnu Namboothiri, Assistant Professor, Department of Computer Science and IT, Amrita School of Arts and Sciences,Kochi ,Amrita Vishwa Vidyapeetham, India . Email:vleena@gmail.com

Air Quality Index (AQI) is a mathematical evaluation that is applied by the government agencies to make aware of the public about the contamination level of air. An increase in air quality index causes severe threats to human health which signifies increased air pollution.

The AQI value detects if the air in the surroundings is clear or contaminated and mentions the health risks associated with that. Developed countries have their inbuilt air quality indices and concern to the public health protection, they are related to the National air quality standards. The AQI values are assorted into certain ranges authorized with a color code and with a standardized public health warning. The criteria to categorize the Air Quality Index level are Good, Moderate, Unhealthy for sensitive groups, Unhealthy and Very unhealthy.

AQI values range from 0 to 300.If the AQI shows values less than 50, it suggests that there is no hazard for public health and if the AQI value is more than 200, then it shows an unhealthy representation. An Environmental Protection Agency (EPA) for health concern which is developed under the United States based on air contamination. The quality of air is decided by AQI values in six different levels and it shows a hazardous air quality to those over 300 AQI values and good quality for the below 50.

II. LITERATURE REVIEW

In 2015, a system proposed by Ruhul Amin Dicken [1] acquires the dependency of the pollutants to the admittance of victims in the medical facilities and thus analyze the cause behind the rapid rise of disease rates in Bangladesh. The method of clustering different air pollutants in different seasons of Bangladesh is derived by K-Means clustering and the method is used for the classification of patients according to different admission rates in the CART method. Data mining is applied to prognosticate the air contamination [3] in which two procedures of feature selection such as a linear method of step-wise fit and genetic algorithm is proposed. PM10, Sulphur Dioxide, Nitrogen Dioxide, and Ozone are the pollutants that help to take part in the prediction of selected features. This paper explains the mathematical features of forecasting issues in air pollution and concentrates mainly on the data mining classification used for constructing the most authentic prediction model.

In Tamil Nadu, ambient air quality prediction is analyzed by the ANN (Artificial Neural Network) model in data mining techniques [4].For the Government policymakers in planning the upcoming standard policies of air pollution, the obtained pattern can serve as an important reference. Data mining is used as a method for the analysis of health risk on an air quality level basis [5].

For the prediction of health concerns, we use Decision Tree J48 and Naïve Bayes algorithms. Paper [9] tells about the air quality increase in the urban region compared to other regions and the random forest algorithm gives more accuracy for foreseeing the air quality in urban regions compared to others.

The data is collected from urban areas such as road information, metrology data, point of interest and real-time traffic status. Paper [10] gives an overall view about air pollution and how it becomes hazardous to the environment as well as to human life. Random forest algorithm is used to prove it by selecting some of the most important features and predictors.

This paper [11] mainly focused on PM2.5 air pollutant which develops a prediction system that helps us to predict the air quality of the next two days by using data-driven models. Random forest, gradient boosting, and linear regression are the algorithms used for this. The air quality is predicted based on Ipython implementation. Air data analysis for predicting health risks [2], here the dataset which consists the AQI values of different air pollutants such as CO, NO2, O3, and SO2 are analyzed by using two algorithms they are Random forest and Multiclassifier. The result exhibits that the Multiclass Classifier is more reliable than the other. The air quality index is predicted using the KNN (K-Nearest Neighbor) technique. The dataset consists of AQI values of SO2, NO2, CO, and O3. Here, results obtained as a relative squared error is 87.10% and mean squared error is 0.669%.

III. METHODOLOGY

In this proposed System, there are two different algorithms. By comparing these algorithms we identify the best algorithm that shows the highest accuracy and with the help of this algorithm we can predict health concern of any pollutants.

A. Multiclass Classifier

A Multiclass Classifier is used to perform classification tasks with two or more classes. This classification assumes that a particular sample is allocated to a single label.

B. Decision Tree J48

A tree-like structure which includes Root Node, Branches, and Leaf Nodes. The internal node does the test condition on attributes, resultant condition of the internal node is branch and the class label is shown by the leaf nodes.

Decision Tree J48 is an implementation of the ID3 algorithm. The J48 classifier is a C4.5 algorithm that is mainly used for classification. By applying this algorithm to our dataset will be able to predict an aimed variable of a new dataset.

C. Dataset Preparation

Our dataset is taken from kaggle which consists of 28 fields. They are,

StateCode,CountyCode, SiteNum, Address,State,Country ,City,Date Local,NO2 Units,NO2 Mean,NO2 1st Max Value,NO2 1st Max Hour,NO2 AQI,O3 Units,O3 Mean,O3 1st Max Value,O3 1st Max Hour,O3 AQI,SO2 Units,SO2 Mean,SO2 1st Max Value,SO2 1st Max Hour,SO2 AQI,CO Units,CO Mean, CO 1st Max Value,CO 1st Max Hour and,CO AQI.

Actually this much fields are not needed for developing our model .So we clean the dataset by removing all the unwanted fields and after that now our dataset contains of 4 fields .They are, NO2 AQI, O3 AQI, SO2 AQI and, CO AQI.

In addition to this four fields for predicting the health risk based on pollutants AQI value we need to find the Maximum AQI value among this 4 pollutant, that should be taken as a new field that is, AQI MAX (1) .

$$\text{AQI MAX} = \max(NO_{2AQI}, SO_{2AQI}, O_{3AQI}, CO_{AQI}) \quad (1)$$

And next we need to assign a class label to the maximum AQI reading value of each day (AQI MAX) based on below mentioned criteria.

301 – 500	Hazardous
201 – 300	Very Unhealthy
151 – 200	Unhealthy
101 – 150	Unhealthy for Sensitive Groups
51 – 100	Moderate
0 – 50	Good

And this class label is taken as next new field i.e., LABEL.

So our final dataset now contains of NO_2 AQI, O_3 AQI, SO_2 AQI, CO AQI, AQI MAX, and LABEL.

By using this dataset we can develop our model.

IV. PROPOSED METHOD

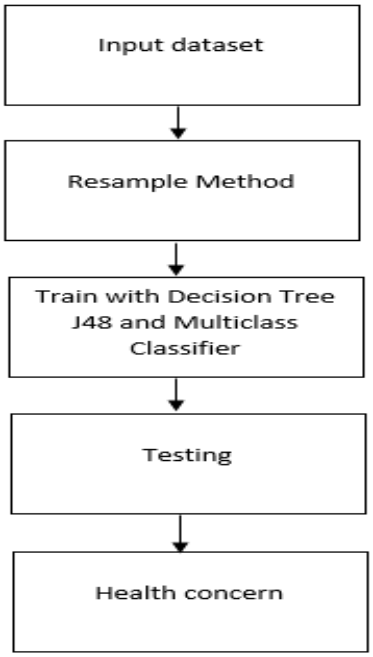

Fig.1. Proposed Model

Fig.1. shows overall steps through which we had gone to find out the algorithm that shows highest prediction accuracy.

First step is used to input dataset. Then we need to divide our dataset into two parts that are test set and train set, this is done by using the resample method. Here, we had taken 70% of data in our dataset for training and remaining 30% is for testing.

Then train the dataset with J48 and Multiclass Classifier algorithms and then save and load the train model. After that re-evaluate the model with the supplied test set.

V. RESULT

The total number of data taken for this study is 15500.From this 10850 data is taken to train and 4650 data is taken to test.

A. Multiclass Classifier

The accuracy obtained for Multiclass Classifier is 96.7527%.But for 151 records it was unable to classify correctly with this classifier which leads to an error rate of 3.2473%.

a	b	c	d	e	classified as
3144	4	0	0	0	a = good
54	1170	7	0	2	b = moderate
0	69	184	0	2	c = unhealthy_sensitive group
0	3	10	0	0	d = unhealthy
0	0	0	0	1	e = very unhealthy

Fig. 2.Confusion Matrix of Multiclass Classifier

B. Decision Tree J48

The J48 classifier has given the accuracy of 99.828%.The error rate for the same is 0.172%.

a	b	c	d	e	classified as
3147	1	0	0	0	a = good
0	1233	0	0	0	b = moderate
0	3	248	4	0	c = unhealthy_sensitive group
0	0	0	13	0	d = unhealthy
0	0	0	0	1	e = very unhealthy

Fig. 3.Confusion Matrix of Decision Tree J48

C. Performance Analyzing

After training the trained dataset, our test set is tested with this trained one to check whether the class label allocated is right or not. Accuracy of classification algorithms is determined by calculating the percentage of instances allocated to the right class label. The accuracy of a classifier can be attained from (2).

$$\text{Accuracy} = \frac{\text{Total number of correctly classified instances}}{\text{Total number of instances}} \quad (2)$$

The classifier Error Rate can be obtained from (3).

$$\text{Error Rate} = \frac{\text{Total number of incorrectly classified instances}}{\text{Total number of instances}} \quad (3)$$

Fig.4 ,below calculations and Table- I shows the Accuracy and Error Rate of Multiclass Classifier and Decision Tree J48 algorithms .

Fig.4.Comparison of Accuracy and Error Rrate

For Multiclass Classifier:

Accuracy = 4499/4650 =96.7527 %

Error Rate = 151/4650 =3.2473 %

For Decision Tree J48:

Accuracy = 4642/4650 = 99.828 %

Error Rate = 8/4650 =0.172 %

Table- I: Accuracy and Error Rate

Algorithm	Accuracy	Error Rate
Multiclass Classifier	96.7527%	3.2473%
J48	99.828%	0.172%

By using True positive [TP] rate, False positive [FP] rate, Precision and Recall the dataset is being assessed for this two algorithms. The most suitable value for Precision, Recall and TP rate is always 1 and the FP rate is 0.The value of precision is 1 which states that every data retrieved is relevant, Recall states that all relevant data is retrieved, the TP rate states that the instances are accurately grouped as the class that is already given. FP rate states that the instances are grouped inaccurately.

From Table- II we can conclude that for J48 algorithm all the class except the unhealthy_sensitive group have TP rate 1 which means almost all the data is correctly classified except few of them in unhealthy_sensitive group but in case of Multiclass classifier very unhealthy is only classified correctly. FP rate is 0 and Precision is 1 for all the class label except moderate and unhealthy in J48 but in case of Multiclass classifier except the unhealthy others have FP rate, and no class labels have good precision. Recall is good for all the class labels except the unhealthy_sensitive group in J48 but in multiclass classifier only very unhealthy have good recall i.e. 1. The outcome of precision, recall.TP rate and FP rate is shown in Fig. 5.

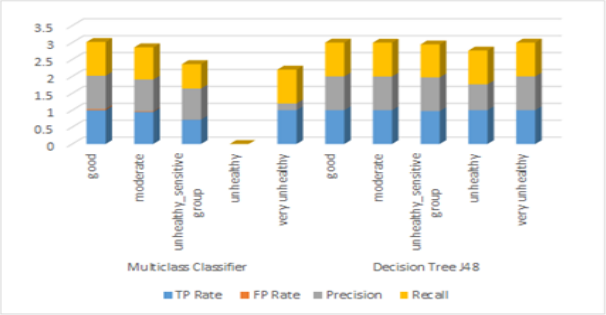

Fig.5.Performance comparison of Decision Tree J48 and Multiclass Classifier

Table- II: Performance Measures of Decision Tree J48 and Multiclass Classifier

Algorithm	Class	TP Rate	FP Rate	Precision	Recall
Mutliclass Classifier	good	0.999	0.036	0.983	0.999
	moderate	0.949	0.022	0.939	0.949
	unhealthy_sensitive group	0.722	0.004	0.915	0.722
	unhealthy	0.000	0.000	0.000	0.000
	very unhealthy	1.000	0.001	0.200	1.000
J48	good	1.000	0.000	1.000	1.000
	moderate	1.000	0.001	0.997	1.000
	unhealthy_sensitive group	0.973	0.000	1.000	0.973
	unhealthy	1.000	0.001	0.765	1.000
	very unhealthy	1.000	0.000	1.000	1.000

From the above-mentioned results and calculations, it is clear that the Decision Tree J48 algorithm gives more prediction accuracy than the Multiclass Classifier.

VI. CONCLUSION

To lead a good quality life, it is extremely essential to analyze the air quality. It plays a very prominent role in developing a smart city and to come up with environmental policies. Using the Multiclass Classifier and Decision Tree J48 algorithm techniques, the data gathered is being analyzed and compare these algorithms to know which will be the best accuracy. The accuracy for the J48 decision tree algorithm is 99.828% and for Multiclass Classifier, the accuracy is 96.7527%. So, we can conclude that the Decision Tree J48 algorithm gives a better health prediction accuracy than the Multiclass Classifier.

REFERENCES

1. Ruhul Amin Dicken, Fazle Rubby, S.A.M, Naz, and Arefin Khaled, 'Analysis and Classification of respiratory health risk with respect to air pollution levels', 2015 IEEE/ACIS.
2. Ranjana Gore and Deepa S. Deshpande ,'Air data analysis for predicting health risks', International Journal of Computer Science and Network(IJCSN),Volume 7,Issue 1, January 2018.
3. Krzysztof Siwek A and Stanislaw Osowski. 'Data Mining method for prediction of Air Pollution', International journal of Applied Mathematics Computer Science, 2016, Volume 26.
4. Dr. V.khanaa and Christy,' Data Mining in the prediction of impacts of ambient air quality data analysis in Urban and Industrial area', IJRITCC, Volume 4 issue 2.
5. Ranjana Gore and Deepa S. Deshpande, 'An Approach for classification of health risks based on air quality levels', International conference on intelligent systems and information management(ICISIM),2017.
6. Jiawei Han and Micheline Kamber,Jian Pei, 'Data Mining: Concepts and Techniques 3rd edition',2011.
7. Elia Georgiana Dragomir,'Air quality index prediction using K-Nearest Neighbor technique', 2010.
8. Ioannis N. Athanasiadis, Kostas D.Karatzas and Pericles A.Mitkas, 'Classification techniques for air quality forecasting', 2007.
9. Ruiyun Yu, Yu Yang, Leyou Yang, Guangjie Han and Oguti Ann Move,'RAQ –A Random Forest approach for predicting air quality in urban sensing systems', 2016.
10. Suketha,Pooja N S and Vanishree B S,'Air Pollution estimation using Data Mining approach',2018,Volume 4,Issue 4.
11. Dr Sandhya P,'Ensemble learning on forecasting fine grained pollutant levels in air using Random Forest ,Naïve Bayes,Decision Tree algorithms',International journal of civil engineering and technology(IJCIET),2018,Volume 9,Issue 7.

AUTHORS PROFILE

Aiswarya Johney ,P G Student, Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi, Amrsita Vishwa Vidyapeetham, India.
Email: aiswaryajohney09@gmail.com

Namitha S J , P G Student, Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidyapeetham, India.
Email:namitha7700@gmail.com

Leena Vishnu Namboothiri,,Assistant Professor, Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi ,Amrita Vishwa Vidyapeetham, India .
Email:vleena@gmail.com