VOL. 14, NO. 6, MARCH 2019 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

PROMULGATION AND DETECTION OF AIR POLLUTION DISEASES USING CLASSIFICATION

Rubidha Devi D[1], Venkatanathan N[2], Umamaheswari P[1] and Vanitha M[1]
[1]Department of Computer Software Engineering, Shanmugha Arts, Science, Technology and Research Academy deemed to be University, Kumbakonam, Tamil Nadu, India
[2]Department of Physics, Shanmugha Arts, Science, Technology and Research Academy deemed to be University, Thanjavur, Tamil Nadu, India
E-Mail: rubidhadevi@src.sastra.edu

ABSTRACT

During modern years, the advancement in Delhi leads to increase in pollution. Air pollution is able to cause a drastic impact on human health and environment. So this paper explored the far-reaching outcome of ambient air smog in sensitive stations of Delhi and also detects the diseases caused by the air pollutions. Based on that, the data mining algorithm "Naive Bayesian" is used. The various air pollutants like Sulphur dioxide (SO2), Nitrogen dioxide (NO2), Respirable particulate matter (RSPM), Suspended Particulate Matter (SPM) [1] and weather conditions and seasons have been noticing in this paper. By using the above data mining algorithm, this paper proposes which are used to detect the various air pollution diseases and will afford the awareness about that syndrome to people.

Keywords: Naive Bayesian, preprocessing, classification, prediction, detection.

INTRODUCTION

In our world, there is a massive quantity of data are existing but people don't know the knowledge and importance of the data. So that the intention of this paper is to cram a massive amount of data set to ascertain the useful comprehension to people by using data mining Algorithm. Delhi is the capital metropolis of India and the most settled metropolitan in the world. According to WHO (world health organization), the air eminence of Delhi city is the nastiest and the most stiffly polluted metropolis [2]. The awful air quality affects the human health and causes global warm. The human transience increased by the vulnerable air pollution diseases. Any substances that presence in the ambiance and harming the air and the living things in the air contamination [3]. Air contamination can be essentially categorized into two, one is natural sources and the other one is anthropogenic sources [1]. Nature and the human can both caused the air smog. The air pollutes by nature of tree-plant fires, volcanic eruptions, blustery weather erosion are called natural sources. And the human tricks that contaminate the air such as industries, vehicular emission, marine vessels, and airplanes are called anthropogenic sources. The element compounds that inferior the air prominence are called air pollutants.

This paper handles both the sources and collects an assortment of atmospheric pollutants such as SO2, NO2, RSPM, SPM and weather conditions and seasons. Data mining play an imperative responsibility in the medical area to advance the health care and diminish the cost. The preceding method measured the air pollutants concentration and analyzed only bare minimum air pollution diseases by using mixed effects model and some other sphere of influence [4].

The diseases that are concerned with the prior system are allergies and asthma. This system has been dealt the concentrations range of air pollutants and weather condition data in the susceptible areas of Delhi. Nowadays, air smog plays an essential role in deteriorating health conditions and death [5]. So based on the data mining algorithm naive Bayesian this paper builds the model for various air pollution diseases to grant the consciousness to the public.

RELATED WORK

Classification is the imperative and crucial tasks in data mining. About an assortment of investigating has been handled to concern data mining techniques and on dissimilar datasets to categorize air pollution. Many types of research concerning patients with diseases include chronic obstructive pulmonary diseases (COPD)[6], [7] and asthma [8]–[10] has been conceded to estimate the intense property of air smog. Using mixed-effects models the relations between the air pollutants concentrations and pulmonary utility were analyzed. The result of the study shows a BC concentration were automatically allied with decreases in FEV1 ((-27.28) [95% (CI) confidence interval -54.10, -0.46] for a 0.23 µg/m3 (IQR) interquartile range assortment enhancement). The indoor concentrations o3 was automatically allied with decreases in PEF ((-8.03 L/min) [95% (CI) confidence interval: -13.02, -3.03] for 11 ppb (IQR) interquartile range enhancement).

This paper proposes to predict and detect the air pollution major diseases using CPCB (Central Pollution Control Board) datasets to build decisions [11]. The purpose is to attain the finest precision with the minimum error rate in the statistic. These tentative results show the achievement of Naive Bayes highest accuracy (98.36%) with the minimum error rate of (0.02%).

EXPERIMENT

In order to predict the air pollution diseases the data mining algorithm "Naive Bayesian" is used to afford accuracy result and detect the various diseases with the level of impacts. The usage of data mining algorithm in

disease prediction is to trim down the test and enhance the accuracy of the rate of detection.

Experiment environment

This paper described the classifier experiment using the libraries from Weka machine learning environment [12]. Data preprocessing, feature selection and classification are done by WEKA which has a group of algorithms. WEKA implements Machine learning techniques that are useful to a collection of actual world harms. The plan provides a distinct structure to estimate and constructs the models for planners and adventurers.

Data preprocessing is a significant step in data mining. It is used to correct the incorrect data in real data. In this paper, it is done by binning it replaces the missing values by mean values in the dataset. The mean value is calculated by the formula:

Mean = all values/total number of values.

The method feature selection is used to find the meaningful input on the statistics. Generally, data contains more information in that some of them are irrelevant to the process. So this paper used Correlation-Based Feature Selection (CFS) to extract the useful feature for classification. It finds the feature which is highly correlated to the class.

The detection is also done by using WEKA and the Naive Bayesian classifier based on seasons, weather conditions and concentration limits of various air pollutants. The concentration limit is in the range of microgram per cubic meter (µg/m3). Based on the concentration limit this paper detects the various diseases with impact.

Air pollution dataset

The Ambient Air Quality (original) datasets from the Central Pollution Control Board (CPCB) is used in this study. Ambient Air Quality Dataset has 366 instances, 2 classes (Diseases: 261 No Diseases: 105) and 7 attributes.

Experimental consequences

The analysis of data consequences is reputed in this sector. To evaluate and concern the classifier, this paper using a technique 10-fold cross-validation test [12] to estimating the promulgating models that obtain a training dataset to train the model by splitting the original dataset, and a test dataset to estimate. Before the feature selection method, this paper attempt to analyze the information virtually and construct the value distribution in provisions of efficiency and competence.

Classifier effectiveness

The efficiency of the classifier in stipulations of time to construct the model is estimated in this sector. The accuracy, correctly and incorrectly classified instances are also estimated. The result is exposed in Table 1 and Figure-1.

Table-1. The Naive Bayes classifier performance.

Estimation criteria	Classifier
	Naive Bayes
Time to build a model (sec)	0.0
Incorrectly classified instances	106
Correctly classified instances	260
Accuracy (percentage (%))	98.36

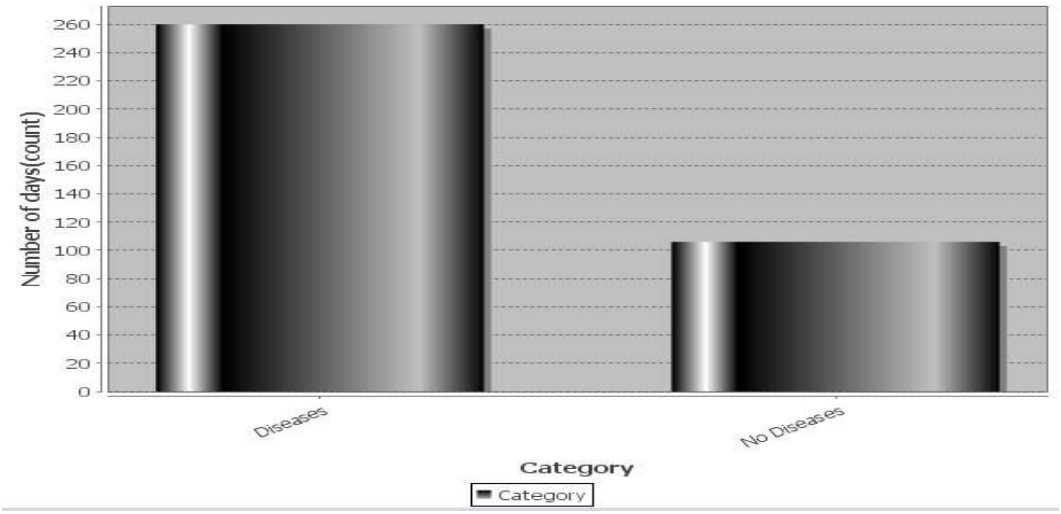

Figure-1. Correctly and incorrectly classified instances graph.

This study considered the classifier performance, simulation error for better measurement. This paper also evaluates the classifier efficiency based on the following conditions

- (RAE) - Relative Absolute Error
- (KS) - Kappa statistic

- (RRSE) - Root Relative Squared Error
- (MAE) - Mean Absolute Error
- (RMSE) - Root Mean Squared Error

RMSE and MAE are in integer values. The values of RAE, RRSE, and KS are in (%) percentage. The outcomes are exposed in Table-2.

Table-2. Training and errors.

Estimation criteria	Classifier
	Naive Bayes
RAE - Relative Absolute Error (%)	4.342
KS - Kappa Statistic	0.959
RRSE - Root Relative Squared Error (%)	28.216
MAE - Mean Absolute Error	0.017
RMSE - Root Mean Squared Error	0.127

Classifier efficiency

After building the analytical model, Table-3 has shown the accuracy of precision, False Positive rate, True Positive rate and recall values for Naïve Bayes Classifier.

Table-3. The accuracy of Naïve Bayes classifier.

True positive rate	False positive rate	Recall	Precision	F-score	Receiver operating characteristic area	Class
0.989	0.029	0.989	0.989	0.989	0.996	Diseases
0.971	0.011	0.971	0.971	0.971	0.996	No Diseases

Figure-2 has shown the better accuracy of the Naive Bayes classifier, the ROC curve of the classifier that has shown the efficiency of the classifier in Figure-3. The ROC curve graph illustrates the classifier performance. Optimal models can be simply selected from the plot. The Confusion matrices (CF) characterize a valuable approach for calculating classifier, In Table-4 every row represents the rates in a real class even as every column views promulgation.

Table-4. The confusion matrix.

	Diseases	No diseases	Class
Naive Bayes	**258**	**3**	**Diseases**
	3	102	No Diseases

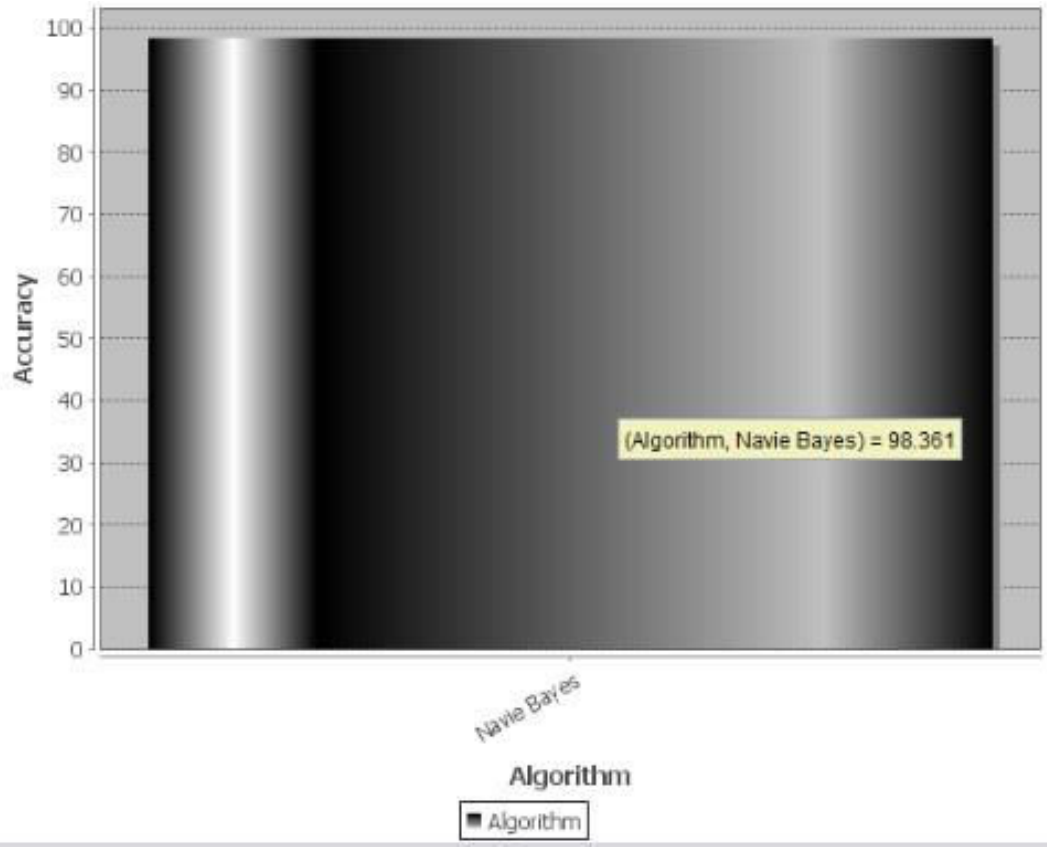

Figure-2. Accuracy of the Naive Bayesian classifier.

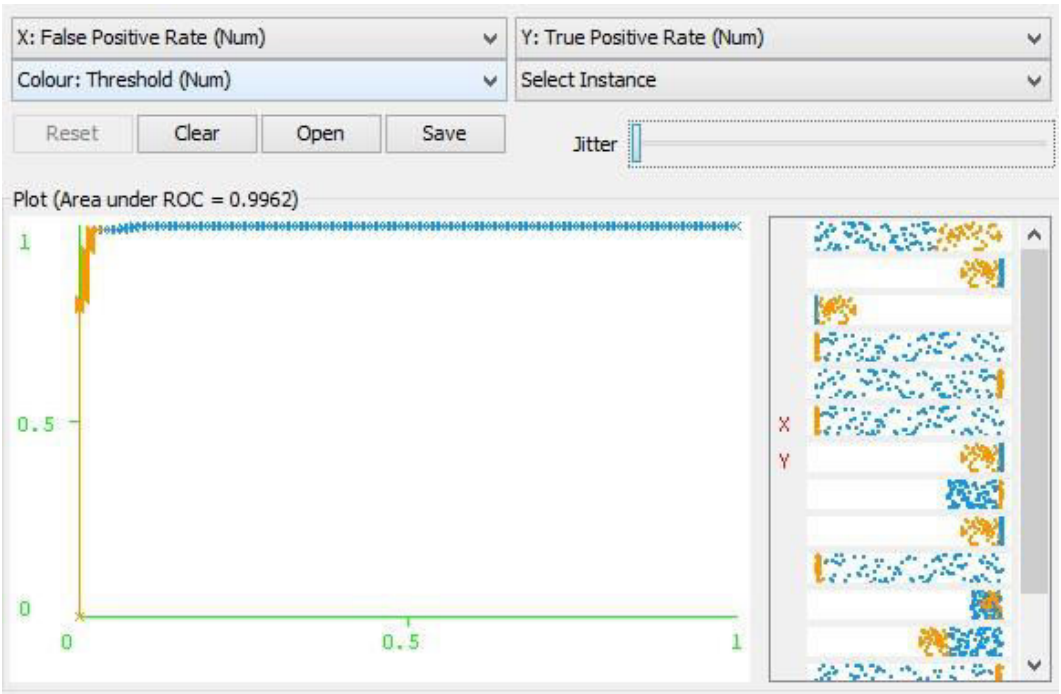

Figure-3. ROC curve.

Finally, the detection of diseases is done by Naive Bayesian classifier and WEKA based on seasons, weather conditions and concentration limits (µg/m3) of pollutants. The detection of various diseases with the impact is shown in Figure-4.

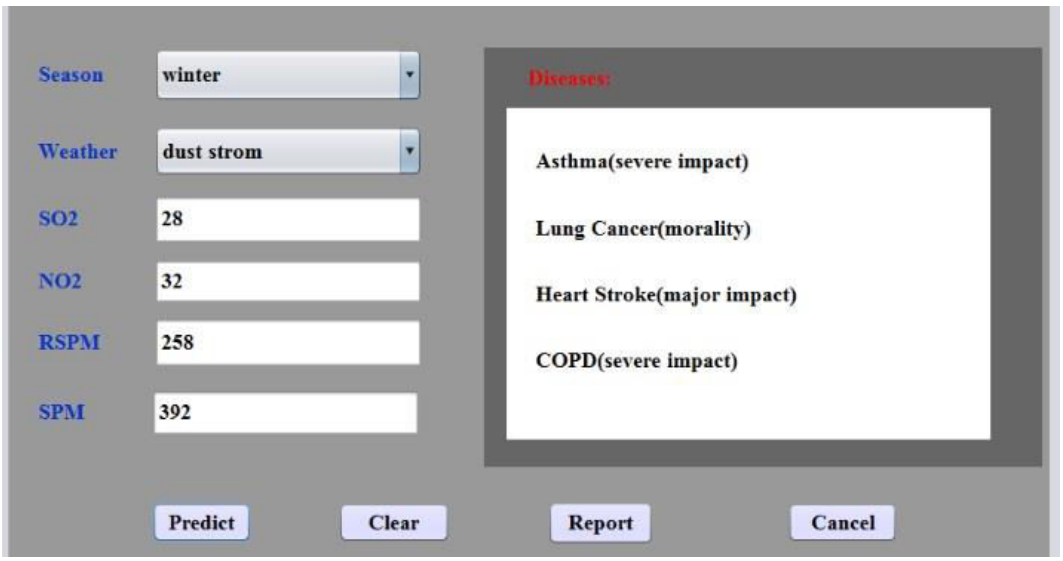

Figure-4. Detection of air pollution diseases with the impact.

CONCLUSIONS

Different methods of data mining are available to analyze the medical data. To construct the precise and valuable classifiers for therapeutic utilization is the dispute in data mining and machine learning areas. This paper propagates the "Naive Bayesian" data mining algorithms on the Ambient Air Pollution (original) dataset. The competence and efficacy of the method are enhanced in this paper in requisites of accuracy, exactness, precision, and sensitivity to find the classification accuracy. Naive Bayesian classifier produced 98.36% accuracy rate in air pollution diseases prediction with low error rate and also detects the various diseases with impacts caused by the air pollutants which can be more accurate in lateral work.

REFERENCES

[1] I. Sources, P. No, D. Air, D. Source, D. C. Measures and D. References. No Title.

[2] S. Taneja, N. Sharma, K. Oberoi, and Y. Navoria. 2017. Predicting trends in air pollution in Delhi using data mining. India Int. Conf. Inf. Process. IICIP 2016 - Proc. pp. 1-6.

[3] J. Appl. 2016. Data mining methods for prediction of air pollution. 26(2): 467-478.

[4] Y. Yoda, H. Takagi, J. Wakamatsu, T. Ito, R. Nakatsubo and Y. Horie. 2017. Acute effects of air pollutants on pulmonary function among students : a panel study in an isolated island. pp. 1-8.

[5] A. Quality, I. Value, V. Poor, V. Poor, V. Poor and P. H. Impacts. 2017. Air Quality Index on Nov 13, 2017 @ 04 : 00 PM Air Quality Index on Nov 13, 2017 @ 04 : 00 PM Ludhiana. pp. 1-3.

[6] M. B. Rice, P. L. Ljungman, E. H. Wilker, D. R. Gold, J. D. Schwartz, P. Koutrakis, G. R. Washko, G. T. O'Connor and M. A. Mittleman. 2013. Short-term exposure to air pollution and lung function in the Framingham heart study. Am. J. Respir. Crit. Care Med. 188(11): 1351-1357.

[7] I. Stewart, P. M. Webb, P. J. Schlutter and G. R. Shaw. 2006. Obesity, physical activity, and the urban environment: public health research needs. Environ. Heal. 5(2): 1-13.

[8] S. E. Sarnat, A. U. Raysoni, W.-W. Li, F. Holguin, B. a Johnson, S. Flores Luevano, J. H. Garcia, and J. a Sarnat. 2012. Air pollution and acute respiratory response in a panel of asthmatic children along the U.S.-Mexico border. Environ. Health Perspect. 120(3): 437-44.

[9] Z. Qian, H. Lin, V. M. Chinchilli, E. B. Lehman, W. F. Stewart, N. Shah, Y. Duan, T. J. Craig, W. E. Wilson, D. Liao, S. C. Lazarus and R. Bascom. 2010. Associations between air pollution and peak expiratory flow among patients with persistent asthma. J Toxicol Env. Heal. 72(1): 39-46.

[10] L. Zhu, X. Ge, Y. Chen, X. Zeng, W. Pan, X. Zhang, S. Ben, Q. Yuan, J. Xin, W. Shao, Y. Ge, D. Wu, Z. Han, Z. Zhang, H. Chu, and M. Wang. 2017. Short-term effects of ambient air pollution and childhood lower respiratory diseases. Sci. Rep. 7(1): 1-7.

[11] Ambient Air Quality Data. p. 392.

[12] H. Asri, H. Mousannif, H. Al and T. Noel. 2016. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. Procedia - Procedia Comput. Sci. 83(Fams): 1064-1069.