Journal of Air Pollution and Health (Summer 2022); 7(3): 287-298

Original Article

Available online at http://japh.tums.ac.ir

Air quality modeling for effective environmental management in Uttarakhand, India: A comparison of logistic regression and naive bayes

Alka Pant[1,], Sanjay Sharma[1], Ramesh Chandra Joshi[2]*

[1] Department of Computer Applications, School of Computer Applications and Information Technology, Shri Guru Ram Rai University, Dehradun, Uttarakhand, India
[2] Department of Computer Science and Engineering, School of Engineering, Graphic Era (Deemed to be University), Dehradun, Uttarakhand, India

ARTICLE INFORMATION

Article Chronology:
Received 12 June 2022
Revised 12 August 2022
Accepted 28 August 2022
Published 29 September 2022

Keywords:
Pollutants; Air quality; Logistic regression; Naive bayes; Environmental management

CORRESPONDING AUTHOR:

alkapant392@gmail.com
Tel: +91 8958128955
Fax: +91-135-2721762

ABSTRACT

Introduction: Air pollution increases the load of hospitalization cases, especially for those who have respiratory problems. For effective environmental management, this study aims to compare the performance of two classification algorithms in machine learning (logistic regression and naive bayes) and to evaluate the selection of the best algorithm for predicting the air quality class.

Materials and methods: Pollutants data (PM_{10}, SO_2, NO_2) have been collected from the Haldwani, Kashipur and Rudrapur regions in Uttarakhand (India). In part I of the study, the Air Quality Index (AQI) is calculated and assigned a class accordingly. In part II, the performance of algorithms is compared, and the air quality class is predicted through the best algorithm. In part III, accuracy is calculated after comparing the predicted class with the actual class. Then, it is compared with the accuracy of our selected algorithm.

Results: The study finds a positive correlation between PM_{10} and SO_2 pollutants. The result shows that the highest accuracy is achieved through logistic regression to predict the air quality class. Further, logistic regression has achieved the same accuracy i.e., 98.70% after comparing predicted values with the actual values.

Conclusion: Logistic regression is the best algorithm to predict the air quality class in the regions of Uttarakhand, where pollutants are being measured in the Government's hospital. The research also indicates that asthma patients in the Kashipur and Rudrapur regions may experience more health effects due to moderately polluted air quality; however, the situation is improving during the monsoon season.

Introduction

A famous saying is "Prevention is better than cure," but natural phenomena and human activities produce pollutants discharged directly from identifiable sources. In developing countries

Please cite this article as: Pant A, Sharma S, Joshi RCh . Air quality modeling for effective environmental management in Uttarakhand, India: A comparison of logistic regression and naive bayes. Journal of Air Pollution and Health. 2022; 7(3): 287-298.

like India, the initiatives are operational. Every industry has a different range of products being manufactured, resulting in different kinds of waste and generating a combination of harmful gases like Nitrogen dioxide (NO_2), Sulfur dioxide (SO_2), and Carbon monoxide (CO) with varying contents and percentages. Pesticides, insecticides, and fertilizers are used by farmers who directly polluted the air.

Indoor and outdoor air pollution is the two main categories of air pollution. The air pollution concentrations are measured at various locations, so air pollution monitoring stations play an essential role. The state and or local agencies use an Air Quality Index (AQI) to disclose to the general population how safe or sterile the environment is predictable. AQI is used to evaluate ambient air quality. The higher the AQI value, the greater the level of air pollution and the greater the health concern. When air concentrations exceed the ambient air's self-cleaning properties and adversely impact human health, the air is said to be polluted [1]. Particulate matter, nitrogen dioxide, sulfur dioxide, and ozone are the most prevalent and dangerous pollutants. Particles of a diameter of fewer than 10 μm, or 100 times smaller than a millimeter, are referred to as PM_{10} (particulate matter). The Particulate Matter (PM_{10}) is generated from construction work and indoor activities (cooking, fireplaces, and cigarette smoking) whereas NO_2 is emitted by road traffic or indoor combustion sources [2]. The particulate matter is small in size that can quickly enter into lungs through the nose and throat which directly affects and creates diseases [3]. On the other side, the SO_2 is shaped by petroleum product ignition at power plants [4]. It has additionally been found to fuel cardiovascular illnesses [5]. It is the leading cause of respiratory and heart diseases. The SO_2 also disturbs the nose, throat, and lungs and may deteriorate existing respiratory diseases [6]. Compared to similar pollutants in food, even a tiny concentration of contaminants in the air eventually becomes more harmful to human health [7]. The sources of pollution are pollutant emissions and the weather is what regulates how air pollutants move through the atmosphere and diffuse into other parts of the ecosystem [8]. The air becomes polluted by the black smoke generated by brick production, and chemical-textile industries make a lot of garbage, causing air pollution. The increase of vehicles and industries in the recent up-going trend has caused great concern for the people in various fields and climatic changes [9].

To protect the environment, the role of scientific researchers has also become more important as they can use powerful machine-learning techniques. The machine learning algorithms like logistic regression, support vector machine, naive bayes, ANN, Decision Tree, etc., have been used to predict the pollutants and Air Quality Index (AQI) by researchers for effective environmental management. Previous studies show that the random forest regression algorithm is best to predict the $PM_{2.5}$ [10] and the M5P algorithm is best for NO_2 and SO_2 prediction, whereas, many researchers go for the decision tree classifier for AQI prediction in different regions [11, 12].

Nowadays, air pollution is also seen as a widespread problem in India. Due to high traffic volumes, air pollution is a significant challenge in India's capital cities and tourist places [13]. In India, the connection between humans and the environment has deteriorated, placing ecosystem stability in danger. According to the Centre for Research on Energy and Clean Air (C.R.E.A.), India continues to occupy the top emitter position for the fifth consecutive year. However, a significant decline of six percent in 2019 as compared to 2018 [14].

Artificial intelligence and machine learning have been phenomenal technological advancements to forecast air quality [15]. Machine-learning is a function of artificial intelligence, which is based on the idea that machines can learn from data, identify patterns, and draw conclusions with minimal human interference. The ability to adapt to machine learning algorithms was also one of the reasons that we prefer this to predict the quality of air [16]. In past, various comparative studies of machine learning algorithms have been conducted to predict air quality, but in India, there are very fewer studies that have been carried out. Nowadays, India is also facing a severe air pollution problem. For effective environmental management, it is important to understand the pollutant's status in different states of this country (India). Based on the weekly dataset of air pollutants, this study aims to compare the performance of two classification algorithms in machine learning i.e., Logistic Regression and Naive Bayes, and to evaluate the selection of the best algorithm for predicting the air quality class in the state of Uttarakhand (India).

Materials and methods

With the expansion of industries and increased automobiles in different locations, India's natural clean air is no longer available. India is a developing country and its capital (Delhi) is also facing severe air pollution. Moreover, it is the most polluted capital city in the world for the fourth consecutive year (World Air Quality Report 2021). Due to technological advancement in India, burning coal increased SO_2 emissions [17]. For this, it is essential to understand and be able to predict changes in the levels of poisonous substances in a particular region [18]. Various nations have their measures of air quality and according to the concentration range of air pollutants, the National Ambient Air Quality Standards (NAAQS) has the following categories of Air Quality Index (AQI)-good (0-50), satisfactory (51-100), moderately polluted (101-200), poor (201-300), very poor (301-400), and severe (401-500) [19]. The past data, the experience, and the trends help us predict future trends and events.

India has a very heavy population and is divided into many states and territories. We have selected our state i.e., 'Uttarakhand' for the study purpose. This state is also known as the 'Hill State' of India. To know the risk of sensitive people (like asthma patients) on hospital admission [20-22] and to understand the pollutant's status, we have covered three regions (namely Haldwani, Kashipur, and Rudrapur) of this state. For these regions, the air pollutants are being measured in Government hospitals. We have collected the weekly dataset of air pollutants (Particulate Matter - PM_{10}, Sulfur dioxide - SO_2, and Nitrogen dioxide - NO_2 [23] for the year 2020 from the Uttarakhand Pollution Control Board (U.P.C.B.).

To set a prediction model for air quality [24], appropriate visualization of the data is required. So we use python programming language and the libraries of matplotlib, pandas, and seaborn during the data processing and visualization stage. The computation of the AQI, accuracy, and prediction has been made through the Jupyter notebook to measure the efficiency of our work. The following parameters are used for the Performance Assessment of Air Quality: Accuracy, Precision, and Recall [25, 26].

Accuracy: Accuracy is the ratio of the number of observations predicted correctly to the total number of observations.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (1)$$

Precision (P): Precision is the ratio of the number of positive observations predicted correctly to the total number of the predicted positive observations.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (2)$$

Recall (R): Recall is the attribute of the model that sums up the model's ability to predict all the positive samples.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

Here, TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative.

To draw some conclusions with the help of machine learning, we have divided our work into three parts. In part I, we calculated the AQI with the help of the sub-index of all the air pollutants and assigned them their class as per the National Ambient Air Quality Standards (NAAQS). The mean value of pollutants fills the missing values to know the status of pollutants and analyze it for machine-learning model purposes. In part II, the performance of Logistic Regression and naive bayes [27] have been compared based on precision, recall, and accuracy [28], and the predicted values of the air quality class have been turned out from the best algorithm. For this, all the weekly pollutants had been divided into two data sets: train and test. Here, the training dataset contains 70% of the data, and the test dataset includes 30% of the data. In part III, accuracy has been predicted after comparing the predicted values of the air quality with the actual class [29]. At last, this accuracy is compared with the accuracy of our selected algorithm for effective environmental management.

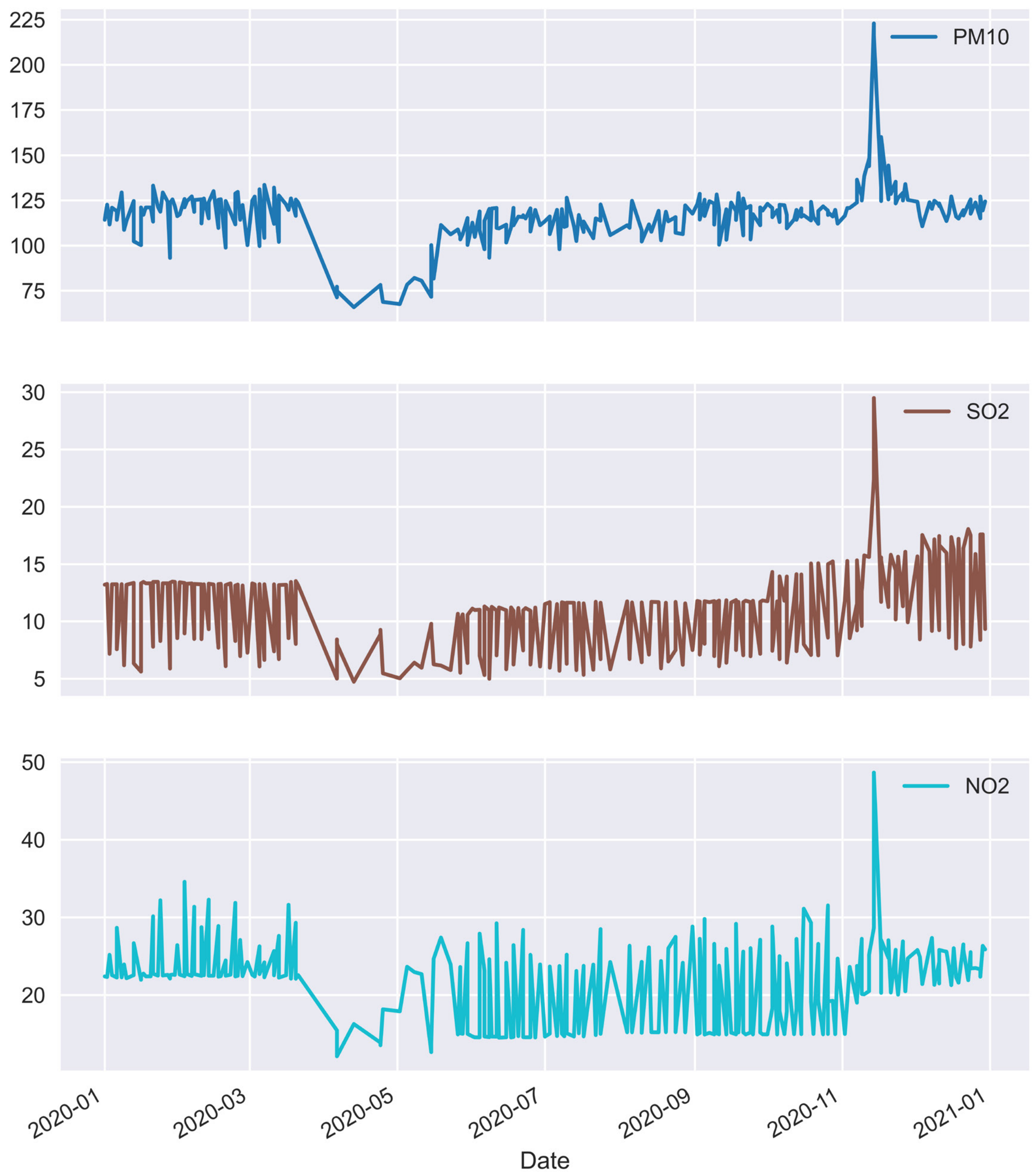

Fig. 1. Graphical representation of air pollutants

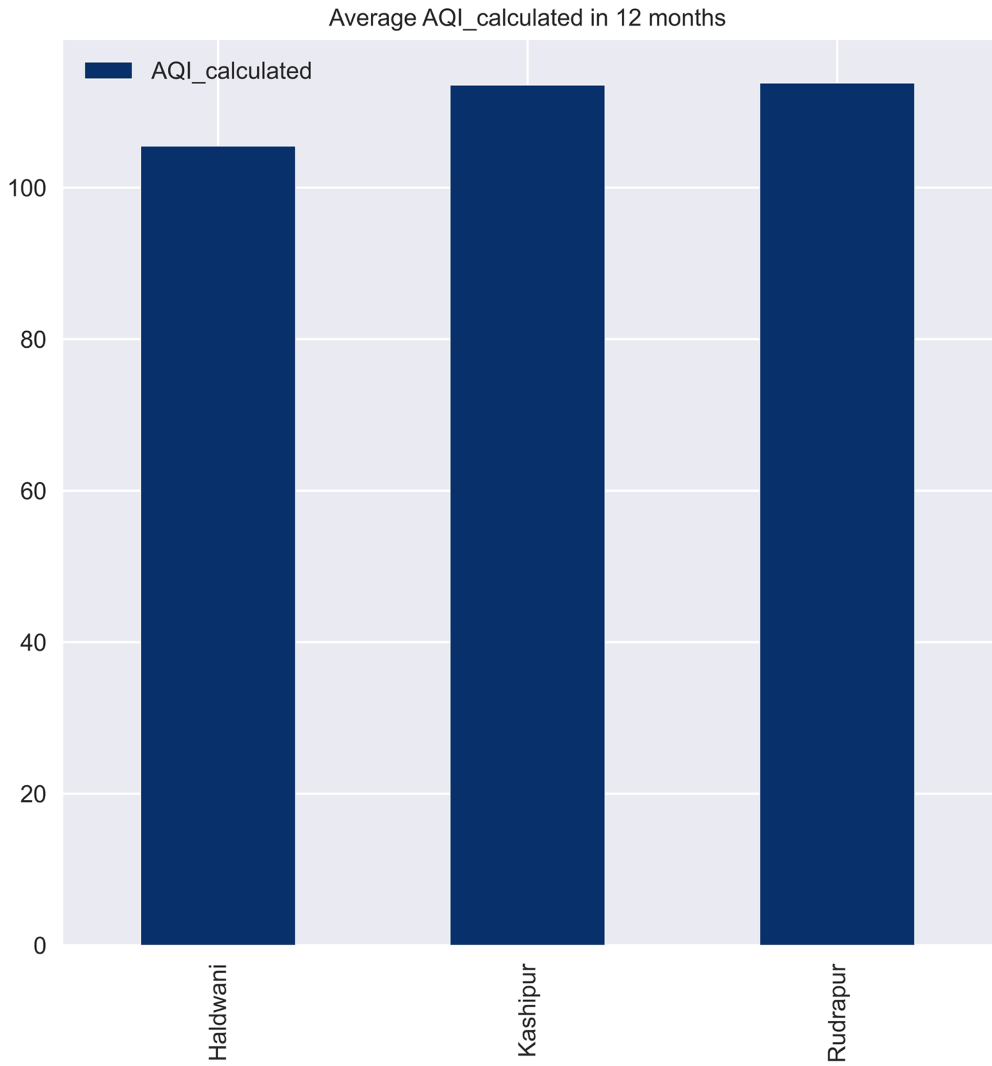

Fig. 2. Air Quality Index (AQI)

Results and discussion

For effective environmental management in Uttarakhand (India), it is important to know the risk for sensitive people (like asthma patients). Focusing on this, we have used two classification algorithms in machine learning (Logistic regression and Naïve Bayes) for the Haldwani, Kashipur and Rudrapur region(s) of this state. In our study, the SO_2 pollutant is maximum for the Rudrapur region (Fig. 3) as this city is known for its industrial area which supports that environmental pollution is also affected by energy consumption and industrial structure [30]. The study finds a positive correlation of 0.73 between PM_{10} and SO_2 pollutants [31], showing that if PM_{10} increases, SO_2 increases and vice-versa (Fig. 4) for all the regions of this state. The study also finds a sharp rise in air pollutants in winters [32, 33]. It may be due to the festival seasons; however, the situation improves during the monsoon season (Fig. 1).

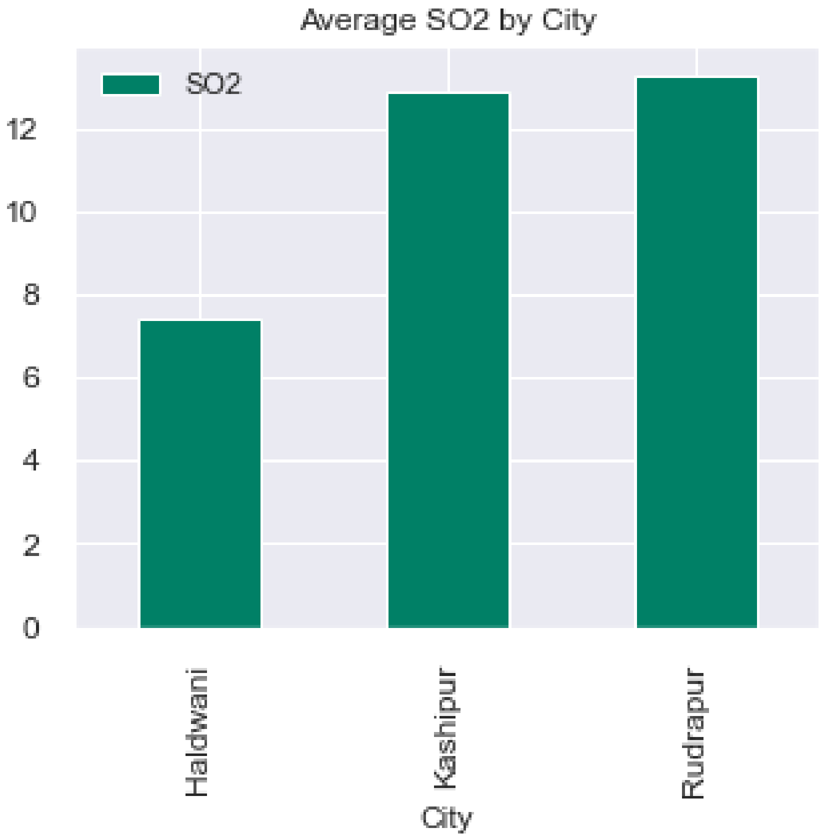

Fig. 3. Pollutant's status (SO_2 and NO_2)

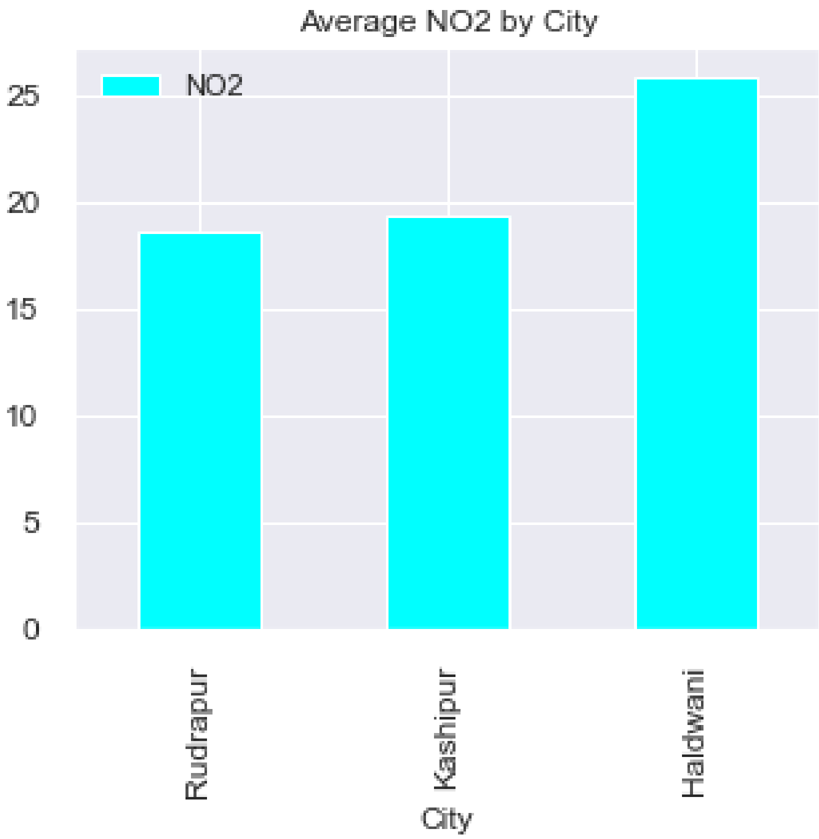

Fig. 4. Heatmap of air pollutants
(continuous variables)

The Machine learning algorithms estimate the new output values using the historical data as input. So, we used the two classification algorithms (Logistic Regression and Naive Bayes) to evaluate the predicted output for the categorical data, i.e., the air quality class where Good: 0. Satisfactory: 1, Moderate: 2, Poor: 3, and so on. Another reason to choose both the classification algorithms is the study on urban air pollution based in Tehran city. Here, the Naive Bayes algorithm is more reliable to handle the imbalanced dataset whereas the Logistic Regression classifier shows a better result in predicting the class just because of the zero precision values [34]. So, we used the precision, recall, f1-score, and confusion matrix measures to evaluate the performance of machine learning algorithms [35]. We find that the results of the employed algorithms are not the same, and the precision-recall is high in the case of Logistic Regression, with the highest accuracy rate of 98.70% (Table 1).

This study corroborates the findings of a previous study [34], where it was found that the Logistic Regression is best for the prediction of air quality. To confirm the selection of our algorithm, we calculated the predicted values of the air quality class through Logistic Regression. We compared these predicted values (class) with the actual values, and it was found that the same accuracy has been achieved i.e., 98.70%, for all the regions of Uttarakhand where the pollutants are being measured in the Government hospitals (Table 2 and Fig. 5). The architecture diagram in Fig. 6 also shows a glimpse of a proposed model for the selection of a classification algorithm in machine learning for predicting air quality. This paper provides a new solution for air quality prediction under a large amount of pollutants data and plays an essential role in preventing and protecting the environment of a developing country like India.

Table 1. Performance Measures

Algorithms and Accuracy	Precision	Recall	F1 score	Confusion matrix $\begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix}$	AQI Class
Logistic Regression (98.70%)	0.80	1.00	0.89	$\begin{bmatrix} 4 & 0 \\ 1 & 59 \end{bmatrix}$	1 as satisfactory.
	1.00	0.98	0.99		2 as moderate.
Naive Bayes (92.20%)	0.44	1.00	0.62	$\begin{bmatrix} 4 & 0 \\ 5 & 55 \end{bmatrix}$	1 as satisfactory.
	1.00	0.92	0.96		2 as moderate.

Table 2. Accuracy of actual vs predicted class
(98.70% through Logistic Regression)

Date	A_d	P_d	Date	A_d	P_d
11/9/2020	2	2	3/3/2020	2	2
1/24/2020	2	2	7/14/2020	2	2
12/21/2020	2	2	2/11/2020	2	2
4/6/2020	1	1	2/13/2020	2	2
10/13/2020	2	2	1/20/2020	2	2
6/2/2020	2	2	2/17/2020	2	2
6/18/2020	2	2	7/28/2020	2	2
10/22/2020	2	2	9/7/2020	2	2
11/12/2020	2	2	9/9/2020	2	2
8/24/2020	2	2	11/14/2020	2	2
10/9/2020	2	2	1/29/2020	2	2
10/19/2020	2	2	12/11/2020	2	2
5/19/2020	2	2	11/23/2020	2	2
2/24/2020	2	2	11/14/2020	2	2
10/22/2020	2	2	12/29/2020	2	2
12/17/2020	2	2	10/29/2020	2	2
6/25/2020	2	2	8/10/2020	2	1
12/24/2020	2	2	10/9/2020	2	2
11/7/2020	2	2	3/5/2020	2	2
3/11/2020	2	2	9/10/2020	2	2
2/20/2020	2	2	2/22/2020	2	2
9/11/2020	2	2	10/19/2020	2	2
11/21/2020	2	2	12/21/2020	2	2
9/3/2020	2	2	1/28/2020	2	2
5/23/2020	2	2	2/13/2020	2	2
12/28/2020	2	2	2/6/2020	2	2
1/30/2020	2	2	12/9/2020	2	2
10/15/2020	2	2	6/20/2020	2	2
9/5/2020	2	2	5/16/2020	1	1
7/6/2020	2	2	11/4/2020	2	2
4/6/2020	1	1	9/28/2020	2	2
7/17/2020	2	2	11/24/2020	2	2
5/27/2020	2	2	8/6/2020	2	2
5/15/2020	1	1	2/24/2020	2	2
12/15/2020	2	2	2/25/2020	2	2
9/24/2020	2	2	3/7/2020	2	2
2/17/2020	2	2	12/24/2020	2	2
1/21/2020	2	2	5/11/2020	1	1

A_d *means Actual Data*
P_d *means Predicted Data*

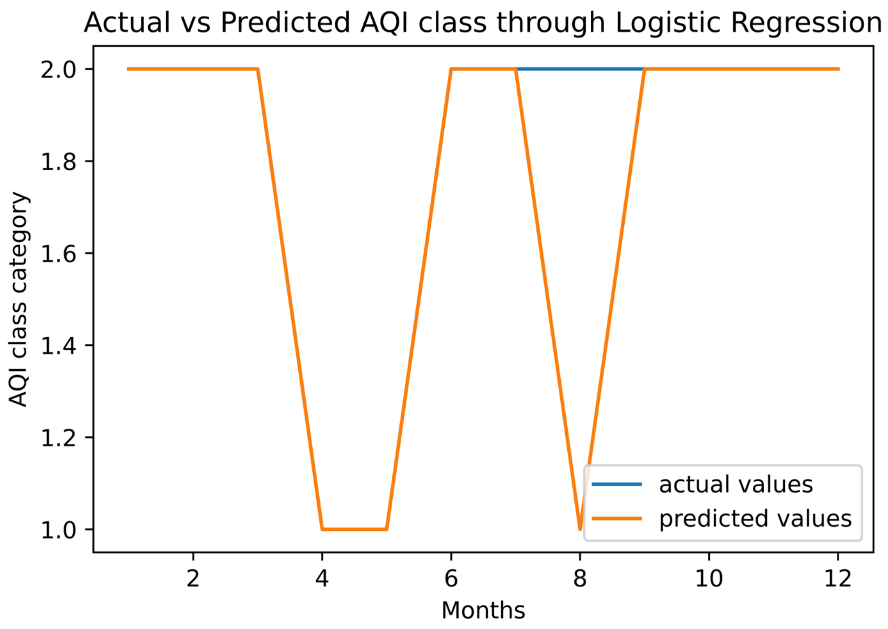

Fig. 5. Comparison of actual with predicted class

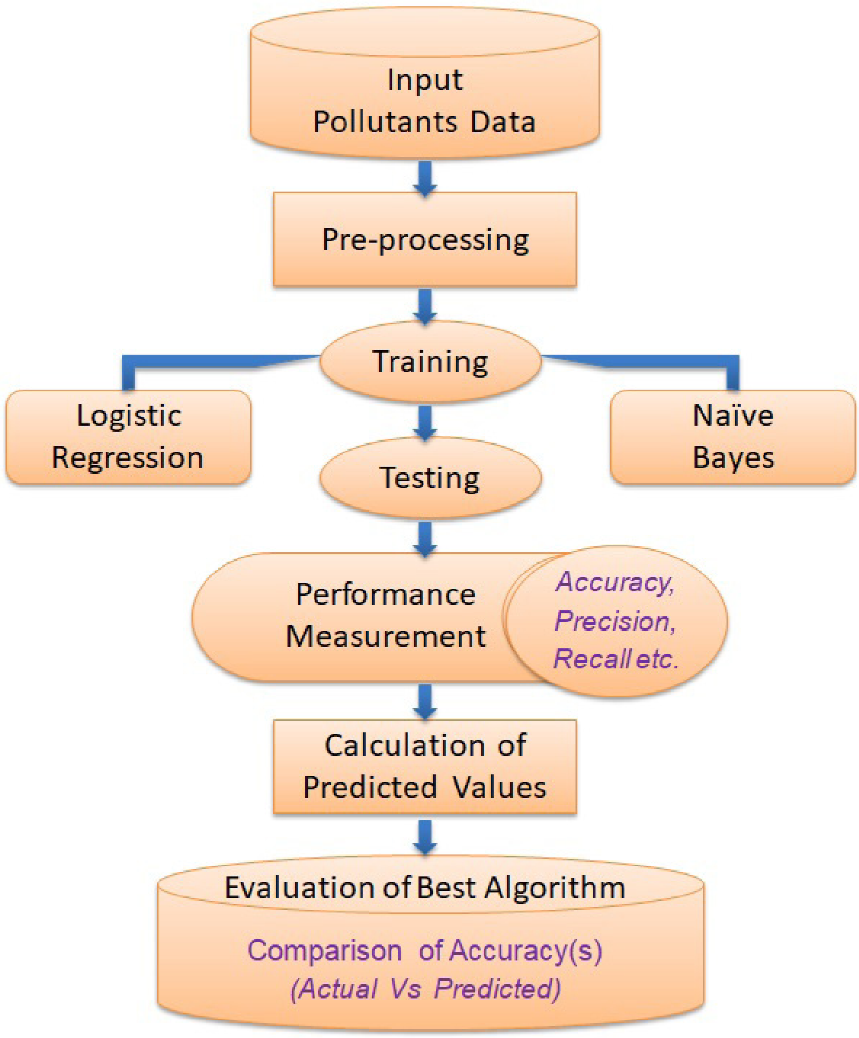

Fig. 6. Prediction model of air quality class using classification algorithms

Conclusion

The air pollution crises have occurred from time to time in India, just as they did in the early years of development in developed countries. Hence, it is essential to predict the pollution events in India, like pollution from stubble burning continuously. To timely prevent air pollution, we have established two classification algorithms in machine learning. The research indicates that the members of sensitive groups (like asthma patients) of the Kashipur and Rudrapur regions may experience more health effects due to the moderately polluted air quality. The result shows that the Logistic Regression is more accurate, with an accuracy rate of 98.70% than the Naive Bayes for predicting the air quality on the weekly dataset of air pollutants of Haldwani, Kashipur, and Rudrapur regions of Uttarakhand (India). Based on the existing dataset of air pollutants, the results confirm the effective environmental management, as the Logistic Regression has achieved the same accuracy after comparing the predicted values with the actual ones of air quality class. Hence, Logistic Regression is the best algorithm as compared to Naive Bayes to predict the air quality class for all the regions of Uttarakhand (India) where the air pollutants are measured in the government hospitals. In near future, the time series forecasting methods can also be used for each pollutant concentration to make a better solution to the air pollution problem for the above regions.

Financial supports

In this research, no grants were received by the authors.

Competing interests

The authors have declared that no competing interests exist.

Acknowledgements

The authors would like to acknowledge the support of Ms. Niharika Dimri, Information Officer in Uttarakhand Pollution Control Board (Uttarakhand, India) for providing the pollutants data for different regions of the state, and Dr. Nagendra Kumar Singh, Faculty (NIELIT - Lucknow, Government of India) for guiding the machine learning algorithms.

Ethical considerations

The authors declare that ethical issues (including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed.

References

1. Zhang K, Batterman S. Air pollution and health risks due to vehicle traffic. Science of the Total Environment. 2013;450-451:307-316.

2. Kopnina H. Vehicular air pollution and asthma: implications for education for health and environmental sustainability. Local Environment - The International Journal of Justice & Sustainability. 2017;22(1):38-48.

3. Delavar MR, Gholami A, Shiran GR, Rashidi Y, Nakhaeizadeh GR, Fedra K, et al. A novel method for improving air pollution prediction based on machine learning approaches: a case study applied to the capital city of Tehran. ISPRS International Journal of Geo-Information. 2019 Feb 23;8(2):99.

4. P. Predd. A market for clean air. IEEE Spectrum. 2005;42(6):60.

5. Meyers DG, Neuberger JS, He J. Cardiovascular effect of bans on smoking in public places. A systematic review and meta-analysis. Journal of American College of Cardiology. 2009;54(14):1249-1255.

6. Kim D, Cho S, Tamil L, Song DJ, Seo S. Predicting asthma attacks: Effects of indoor PM concentrations on peak expiratory flow rates of asthmatic children. IEEE Access. 2020;8:8791-8797.

7. Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. Environmental and health impacts of air pollution: A review. Frontiers in Public Health. 2020;8(14).

8. Bai Y, Li Y, Wang X, Xie J, Li C. Air pollutants con¬centrations forecasting using back propagation neural network based on wavelet decomposition with meteo¬rological conditions. Atmospheric Pollution Research. 2016;7(3): 557-66.

9. Douglas MJ, Watkins SJ, Gorman DR, Higgins M. Erratum: Are cars the new tobacco? Journal of Public Health (Bangkok). 2011;33(3):472.

10. Ameer S, Shah MA, Khan A, Song H, Maple C, Asghar MN. Comparative analysis of machine learning techniques for predicting air quality in smart cities. IEEE Access. 2017;20.

11. Jamal A, Nabizadeh Nodehi R. Predicting air quality index based on meteorological data: A comparison of regression analysis, artificial neural networks and decision tree. Journal of Air Pollution & Health. 2017;2(1).

12. Mahesh Babu K, Rene Beulah J. Air quality prediction based on supervised machine learning methods. International Journal of Innovative Technology & Exploring Engineering. 2019;8(9 Special Issue 4):206-212.

13. A. Pant, S. Sharma, M. Bansal, M. Narang. Comparative analysis of supervised machine learning techniques for AQI prediction. IEEE International Conference on Advanced Computing Technologies and Applications (ICACTA). March 2022:1-4.

14. India world's largest emitter of sulfur dioxide, emissions. Greenpeace India. 2019: October.

15. Shaban KB, Kadri A, Rezk E. Urban air pollution monitoring system with forecasting models. 2016;16(8):2598-2606.

16. Mahalingam U, Elangovan K, Dobhal H, Valliappa C. A machine learning model for air quality prediction for smart cities. International Conference on Wireless Communications Signal Processing & Networking. 2019;452-457.

17. Xu C, Zhao W, Zhang M, Cheng B. Pollution haven or halo? The role of the energy transition in the impact of FDI on SO_2 emissions. The Science of the Total Environment. April 2021;763.

18. M. Volkodaeva, A. Kiselev. On development of system for environmental monitoring of atmospheric air quality. Journal of Mining Institute. 2017;227:589-596.

19. Irani T, Amiri H, and Deyhim H. Evaluating visibility range on air pollution using NARX neural network simulation. Journal of Environmental Treatment Techniques. 2021;9(2):540-547.

20. Chciałowski A, Agata D, Badyda A, Piotr D. Ambient air pollution and risk of admission due to asthma in the three largest urban agglomerations in Poland : A Time-Stratified, Case-Crossover Study. International Journal of Environmental Research and Public Health. 2022;19(10):5988.

21. Gharehchahi E, Mahvi AH, Amini H, Nabizadeh R, Akhlaghi AA. Health impact assessment of air pollution in Shiraz, Iran: a two-part study. Journal of Environmental Health Sciences and Engineering. 2013:1-8.

22. Nadali, A., Leili, M., Karami, M. et al. The short-term association between air pollution and asthma hospitalization: a time-series analysis. Air Quality, Atmosphere and Health. 2022;15:1153-1167.

23. Li, S., Song, S., Fei, X. Spatial characteristics of air pollution in the main city area of Chengdu, China. 19th International Conference on Geoinformatics, Shanghai, China. 2011:1-4.

24. Moradi, H., Talaiekhozani, A., Kamyab, H. et al. Development of equations to predict the concentration of air pollutants indicators in Yazd City, Iran. Journal of Inorganic and

Organometallic Polymers and Materials. July, 2022.

25. Senthivel S, Chidambaranathan M. Machine learning approaches used for air quality forecast : Revue d'Intelligence Artificielle. 2022;36(1):73-78.

26. S. Yarragunta, M. A. Nabi, J. P., R. S. Prediction of air pollutants using supervised machine learning. 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021:1633-1640.

27. R. W. Gore, D. S. Deshpande. An approach for classification of health risks based on air quality levels. 1st International Conference on Intelligent Systems and Information Management (ICISIM). 2017:58-61.

28. Kumar, K., Pande, B.P. Air pollution prediction with machine learning: a case study of Indian cities. International Journal of Environmental Science & Technology. 2022.

29. Doreswamy, Harishkumar K S1, Yogesh KM, Ibrahim Gad. Forecasting air pollution particulate matter ($PM_{2.5}$) using machine learning models. Third International Conference on Computing & Network Communications. Procedia Computer Science. 2020;171:2057-2066.

30. Yingying Li and Dongxiao Niu. Research on the relationship between China's economic growth and SO_2 emission. IEEE International Conference on Test and Measurement. 2009:88-90.

31. L. Song. Impact analysis of air pollutants on the air quality index in Jinan Winter. IEEE International Conference on Computational Science and Engineering (CSE) & IEEE International Conference on Embedded and Ubiquitous Computing (EUC), 2017: 471-474.

32. Chauhan, Avnish, Mayank Pawar, Rajeev Kumar, P. C. Joshi. Ambient air quality status in Uttarakhand (India): A Case Study of Haridwar and Dehradun using Air Quality Index. 2010;6(9):565-574.

33. Madonsela BS, Maphanga T, Silas Chidi B, Shale K, Zungu V. Assessment of air pollution in the informal settlements of the Western Cape, South Africa. JAPH. 2022;7(1):1-14.

34. Marjan Asgari, M., Farnaghi, M. Ghaemi, Z. Predictive mapping of urban air pollution using apache spark on a hadoop cluster. ICCBDC International Conference on Cloud & Big Data Computing, London: United Kingdom. September, 2017:89–93.

35. S. Yarragunta, M. A. Nabi, J. P., R. S. Prediction of air pollutants using supervised machine learning. 5th International Conference on Intelligent Computing annd Control Systems (ICICCS), 2021:1633-1640.