Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Air Quality prediction using Multinomial Logistic Regression

Ahmad Najim Ali[1], Ghalia Nassreddine[2] ✉ and Joumana Younis[3]
[1]Jinan University, Faculty of Business, Tripoli, Lebanon
[2]Jinan University, Faculty of Business, Business Information Technology Department, Tripoli, Lebanon
[3]Conservatoire National des Arts et Métiers, France; [3]Paul-Valéry Montpellier 3 University, France
Corresponding Author: Ghalia Nassreddine, **E-mail**: ghalia.nasseredine@jinan.edu.lb

| ABSTRACT

Nowadays, Artificial Intelligence (AI) plays a primary role in different applications like medicine, science, health, and finance. In the past five decades, the development and progress of technology have allowed artificial intelligence to take an essential role in human life. Air quality classification is an excellent example of this role. The use of AI in this domain allows humans to predict whether the air is polluted or not. In effect, monitoring air quality and providing periodic and direct statistics are essential requirements to ensure good air quality for individuals in the community. For this reason, a decision-making system is built to decide whether the air is clean or not. Based on this system's decision, necessary practices and measures are taken to improve air quality and ensure air sustainability. In this paper, the multinomial logistic regression technique is used to detect the air pollution level. The proposed method is applied to a real dataset that consists of 145 responses recorded from an air quality multi-sensor device containing chemical sensors. The used device was placed in New York City, USA, from 1/1/2021 to 7/1/2021 (one week) and is freely available for air quality sensors deployed in the field. The result shows the efficacity of this method in air pollution prediction.

| KEYWORDS

Artificial Intelligence, Logistic regression, multinomial logistic regression, classification, prediction, air pollution.

| ARTICLE INFORMATION

ACCEPTED: 20 September 2022 **PUBLISHED:** 29 September 2022 **DOI:** 10.32996/jcsts.2022.4.2.9

1. Introduction

The enormous development of information technology tools gives rise to the creation of Artificial intelligence (AI) techniques. The term Artificial intelligence refers to the pretend human brain operation by machines and computer systems. Nowadays, AI is used in many recent applications such as healthcare, smart city, finance, banking, robotics, and others. It allows the creation of expert systems, voice recognition, machine vision, and natural language processing (NLP). Indeed, in the last few years, AI has had an essential role in daily human life.

Air pollution is one of the significant threats to health in the world. Almost all of the global population can be exposed to air pollution. This exposure increases the risk of heart disease, stroke, chronic obstructive pulmonary disease, cancer, and pneumonia. World Health Organization (WHO) monitors air pollution, exposure levels, and health impacts in the national, regional, and global groups from ambient and household air pollution. Such estimates are used for official reporting like the world health statistics and the Sustainable Development Goals. More than four million deaths occur yearly due to pollutant air exposure. In addition, nine out of ten people worldwide live where air quality exceeds WHO guidelines.

Recently, AI tools have been widely used in air pollution problems. AI techniques like classification allow humans to predict if the air is polluted or not. Indeed, surveying air quality and producing daily or weakly statistics are essential to ensure good air quality for individuals in the community.

Classification techniques are one of the best ways to solve the air pollution problem (Aditya et al., 2018; Govender & Sivakumar, 2020). Logistic regression is one classification technique limited to two-class classification problems (DeMaris, 1995). Some extensions have been developed recently, such as one-vs-rest, to allow logistic regression for multiclass classification problems. However, they require that the classification problem be transformed into multiple binary classification problems (Abramovich et al., 2021).

Instead, the multinomial logistic regression algorithm is an extension of the logistic regression model that requires changing the loss function to cross-entropy loss. In addition, the multinomial probability distribution is used instead of the predicted probability distribution to support multiclass classification problems (El-Habil, 2012). In various regression applications, the response variable of interest may have more than two qualitative possible outcomes or more than two nominal categories; thus, it can be represented by a multinomial variable (Bayaga, 2010).

In this article, the multinomial logistic regression algorithm will be used for air pollution. This article is composed of six sections. First, the air quality analysis will be described briefly in Section II. Some existing research on the air pollution problem will be presented in Section III. In Section IV, the proposed method will be described. The result of a real dataset will be discussed in Section V with a comparison with KNN and SVM classifiers. This paper will end with a conclusion in Section VI.

2. Air quality analysis

Air quality is a measure of how polluted or clean the air is. Therefore, monitoring air quality is essential because polluted air can harm our environmental health. Air quality is measured with the Air Quality Index (AQI). It works like a thermometer that runs from 0 to 500 degrees. However, instead of showing temperature changes, the AQI shows changes in the amount of air pollution (Rizwan et al., 2013).

The increase in air pollution causes severe health disorders. This pollution usually comes through factories, laboratories, and the frightening industrial development that we observe at present, which leads to increased air pollution. It directly affects millions of people who suffer from respiratory disorders, pneumonia, asthma, eye problems, etc. Table 1 shows the excellent and dangerous levels of air quality (Mirabelli et al., 2020):

Table 1: Air quality index

Levels of Concern	Values of Index	Description of Air Quality
Good	0 to 50	Air quality is good, and air pollution poses little or no risk.
Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, especially those unusually sensitive to air pollution.
Unhealthy for Sensitive Groups	101 to 150	People in this group may experience health effects. The general public is less likely to be affected.
Unhealthy	151 to 200	Here some may experience health effects; People with acute problems may experience more severe health effects.
Very Unhealthy	201 to 300	Health warning: here, the risks of health effects increase for everyone.
Hazardous	301 and higher	Emergency Health Warning: Everyone is likely to be affected within this range.

3. Multinomial logistic regression

We can define the Multinomial logistic regression for a response variable with three or more discrete outcomes. It is an extension of logistic regression when the response has only two outcomes (i.e., regression based on the binomial distribution). The multinomial model deals with data analysis cases where the response variable is ordinal (the order of response categories is essential) or nominal (the order of response categories is unimportant). Multinomial logistic regression uses maximum likelihood estimation to assess the probability of categorical membership, such as binary logistic regression. Also, it does not require careful study of the sample size and examination of small cases (Braga et al., 2013).

Multinomial logistic regression is often called polychromous logistic regression. It is used when the dependent variable has more than two unordered or nominal categories. In the risk analysis, using multinomial logistic regression, the response variable is dummy coded into multiple 1/0 variables, meaning that all categories have a variable except for one, so if there are M categories, there will be M-1 dummy variables. All categories have their dummy variable except one. Each category of the dummy variable has a value of 1, while the other variables have a value of 0. The single category and reference category do not need a dummy variable, so they are determined individually by other variables whose value is 0.

The analyst can take a risk using multinomial logistic regression and then estimate a separate binary logistic regression model for all these dummy variables. The critical factor to consider here is that each one informs the predictors of the effects of risk on the likelihood of success in that category compared to the reference category. We note here that each model has its regression coefficients and intercept because the predictors of risk analysis can influence a different category (Bayaga, 2010).

In the multinomial outcome setting, "the interpretation of odds ratios could be made generalizing the notation used in the binary outcome case to include the outcomes being compared as well as the values of the covariate" (Hosmer Jr et al., 2013). Considering that the outcome categorized with y=0 is a reference outcome. The odds ratio of outcome $\mathrm{Y} = \mathrm{j}$ versus outcome $\mathrm{Y} = 0$ for covariate values of $\mathrm{x} = \mathrm{a}$ versus $\mathrm{x} = \mathrm{b}$ is:

$$\mathrm{M_j(a,b)} = \frac{\mathrm{P(Y = j \,|\, x = a) \,/\, P(Y = 0 \,|\, x = a)}}{\mathrm{P(Y = j \,|\, x = b) \,/\, P(Y = 0 \,|\, x = b)}}$$

We can get the estimated odds ratio values, called M, by exponentiation of the estimated slope coefficients. When fitting a multinomial logistic regression model, the outcome has many outcomes (more than two or K), meaning we can think of the problem as fitting K-1 independent binary logistic regression models.

$$\ln = \frac{\mathrm{Pr\ (Yi = 1)}}{\mathrm{Pr\ (Yi = K)}} = \beta_1.\mathrm{X}_i$$

$$\ln = \frac{\mathrm{Pr\ (Yi = 1)}}{\mathrm{Pr\ (Yi = K)}} = \beta_2.\mathrm{X}_i$$

$$\cdots$$

$$\ln = \frac{\mathrm{Pr\ (Yi = 1)}}{\mathrm{Pr\ (Yi = K)}} = \beta_{\mathrm{k}-1}.\mathrm{X}_i$$

Exponentiation on both sides of the equations will provide probabilities:

$$\mathrm{Pr\ (Yi = 1) = Pr\ (Yi = K)\ e^{\beta 1.Xi}}$$

$$\mathrm{Pr\ (Yi = 1) = Pr\ (Yi = K)\ e^{\beta 2.Xi}}$$

$$\cdots$$

$$\mathrm{Pr\ (Yi = 1) = Pr\ (Yi = K)\ e^{\beta_{\mathit{k}-1}.Xi}}$$

3.1 Advantages of Multinomial Logistic Regression

Multinomial logistic regression is considered an attractive analysis. It has several primary advantages (Tabachnick et al., 2007):

- It has excellent strength against violations of multivariate normality and variance-covariance matrices across groups
- It is similar to linear regression but easy to interpret in stats. In addition, it has the advantage of analysis which increase its popularity
- Multinomial logistic regression does not undertake a linear relationship between dependent and independent variables
- Independent variables do not need to be an interval
- Multinomial logistic regression does not require that the independents be unbounded
- Distributed error terms are not assumed
- The widespread use of multinomial logistic regression as a problem-solving tool, especially in medicine, psychology, mathematical finance, and engineering

4. Literature review

Recently, many researchers have focused on using AI tools such as classification in air pollution prediction and many other applications. El-Habil (2012) used the Multinomial Logistic Regression model in many areas, including health, society, and education. The researcher used accurate data on physical violence against children to be identified through a survey of youth aged 10 to 14. The Palestinian Central Bureau of Statistics (PCBS) research was conducted for children in Gaza city, where a size of 66,935 had been selected. The response variable consisted of four categories. Eighteen explanatory variables were used to build the primary multinomial logistic regression model. The model was tested through statistical tests to ensure its appropriateness for the data. By randomly selecting two observations of the data used to predict the position of each observation in any classified group, it can be that the model had been tested by knowing the values of the explanatory variables used. The research concluded by using the multinomial logistic regression model. We can accurately define the relationship between the group of explanatory variables and the response variable, identify the effect of each of the variables, and predict the classification of any individual case.

AI tools are used in the text classification domain, where the system is constructed based on a typical assortment of document features (Hasnat et al., 2018). Text classification is the operation of allocating class labels to a text document. Text classification has many critical applications, especially for organization and for browsing within extensive collections of documents. Feature extraction is essential within this process because there are high dimensions of terms. Computationally, classification can be expensive here. In this study, the Document and Principle Component Analysis (PCA) technique was applied to improve the result's accuracy by using singular value decomposition (SVD). The improvement of computational efficiencies and classification accuracy is achieved by eliminating all irrelevant data to reach the objectives. Two classification algorithms are applied for text classification: Back Propagation Neural Network (BPNN) and Support Vector Machine (SVM). These algorithms were applied to Reuters 21.578 data collection. Then the performance of these two algorithms was compared by calculating the standard accuracy and calling the documents. The result showed that the average accuracy for SVM with SVD techniques is 95.6%, the error rate is 0.4%, the SVM with MI is 94.4%, and the average accuracy for BPNN with SVD techniques is 99.5%. The error rate is 0.1%, and BPNN with MI average accuracy is 99.4%.

Because Air pollution is considered one of the biggest health threats, the author focuses on predicting and classifying air pollution using machine-learning algorithms based on real-time environmental data (Tlais et al., 2019). These algorithms would help decision-makers to take action to improve critical situations. Machine-learning algorithms are evaluated with offline and real-time data collected through pollution sensors. The results revealed that artificial neural networks had the best performance and the highest accuracy among KNN, SVM, and Naïve Bayes Classifiers. At last, the aim of the study was that the system presented a good use of experience and knowledge with intelligent technologies representing the world's future.

Iskandaryan et al. (2020) covered the revision of the studies on air pollution prediction using machine learning algorithms based on sensor data in the context of intelligent cities because of the increase in machine learning techniques and their entry into all fields, especially air pollution forecasting. After a comprehensive review of the most relevant papers regarding using the most famous databases and executing the corresponding filter, the main features are extracted to link and compare them. As a result, they can conclude that:

- Instead of using simple machine learning techniques, currently, the authors apply advanced and sophisticated techniques.
- China was the leading country in terms of case studies.
- PM, with a diameter of 2.5 micrometers, was the main prediction target.
- In 41% of the publications, the authors carried out the prediction for the next day.
- 66% of the studies used data that had an hourly rate.
- 49% of the papers used open data, and since 2016 it has increased.
- For efficient air quality prediction, it is essential to consider external factors such as weather conditions, spatial characteristics, and temporal features.

Mohammad et al. (2020) studied Particular matter (PM_{10}) and forecasting to control and reduce environmental and human health damage. Studied datasets have been taken from the Kuala Lumpur meteorological station, Malaysia using a generalized linear model to build Logistic regression as a particular case of linear statistical methods. Therefore, when used with non-linear datasets, the model may reflect inaccurate results. The time stratified (TS) method in different styles is proposed to satisfy more datasets homogeneity. It includes ordering similar seasons in different years to formulate new variables more smoothly than their original ones. The results of the LR model in this study reflect outperforming for time-stratified datasets compared to the entire dataset. The study concluded that LR forecasting could depend on after-time stratifying to satisfy more accuracy with non-linear multivariate datasets in which PM_{10} is the dependent variable.

5. Multinomial logistic regression technique in air quality detection

In this section, the multinomial logistic regression will be applied to the air pollution problem. First, a small description of multinomial regression will be done. Then, the proposed algorithm will be described. Air quality assessment requires atmospheric quality to be described using qualitative indicators based on quantitative classification. The United States Clean Air Act has set specific limits for characterizing the quality of the atmospheric environment .Air quality indicators information has been identified as necessary for any operating air quality management system. Figure 1 shows the classification algorithms approach.

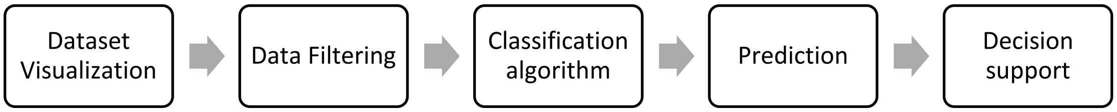


Figure 1: The classification algorithm's approach starts with dataset visualization and ends with decision support.

A. Dataset visualization: This step is usually used for more understanding of the available attributes used in the dataset.

B. Filtering and cleaning dataset

Filtering and cleaning data entails filtering, editing, correcting, and organizing the data within the dataset so that it is generally standardized and formatting the language so that computers can correctly understand and analyze it. This process is usually time-consuming, but it is necessary to obtain data with good results and statistics. If the data is poor during the model training process, the final analysis results will be unreliable, causing the organization to suffer. When data is good, it improves the organization's efficiency by saving time and money. The filtering and cleaning of data are accomplished by performing the following steps:

- Removing irrelevant data
- Remove duplicated data
- Fix structural errors
- Deal with missing data
- Filter data outliers
- Validate data

C. Classification

It can be defined as a regression model that uses a straight line to describe a relationship between variables. This model finds the most suitable data line by searching for the regression coefficient value that reduces the overall error of the model (Kwak & Clayton-Matthews, 2002). It is beneficial both from a practical and conceptual point of view.

The multinomial logistic regression algorithm allows us to predict a categorical dependent variable with more than two levels. The multinomial output can be predicted like any other regression model using one or more independent variables (Lee, Ahn, Moon, Kodell, & Chen, 2013). The multinom() function from the nnet package e in R language is used to implement multinomial logistic regression. Once the function is fed an array of features, the model performs a series of arithmetic operations to normalize the input values into a vector of values that follows a probability distribution, illustrated in Figure 4.

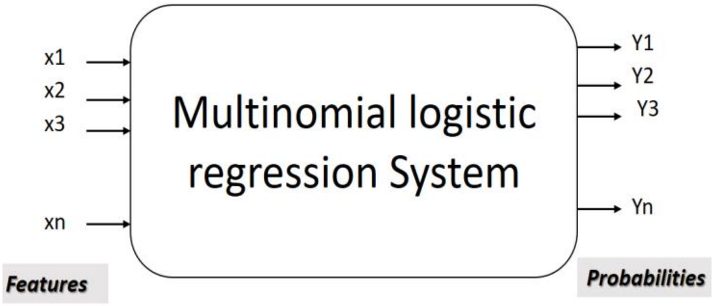


Figure 4: The Multinomial Regression function calculates probabilities for potential target classes from the given features.

The classification model classifies a data point into a particular category by predicting its probabilities of belonging to all available classes. The class with the highest probability is declared as the class of that data point. In this data set, we have the Response variable (AQI), which is a quantitative variable whose values are confined between (0-500), where the value of this indicator increases or decreases depending on the value of ($PM_{2.5}$ and PM_{10}) at most. Moreover, for the classification, the response variable here must be qualitative because we will classify the air quality index as (Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, and Hazardous) (Refer to Table 1).

Using Sturges's rule, a qualitative variable can be converted to a quantitative one. The latter means that the data range should be divided into equally spaced categories, the number of groups or classes is (1 + 3.3 log n), where n is the number of observations) (Legg, Rosin, Marshall, & Morgan, 2007). By following these steps:

- Find the range of the data set.
  Where, range = max value - min value = 78-7 = 71.
- Sturges' rule is applied to determine the number of classes.
  Classes = 1 + 3.3 (log n) where n is the number of observations.
  Classes = 1 + 3.3 (log 145) = 1+3.3*2.1 = 7.9 = 8 groups.
- We define the class width. class width = rang / number of classes = 71 / 8 = 8.87 = 9. It means that nine samples will separate eight groups/classes.
  Class 1 = 0 to 9
  Class 2 = 10 to 18
  Class 3 = 19 to 27
  Class 4 = 28 to 36
  Class 5 = 37 to 45
  Class 6 = 46 to 54
  Class 7 = 55 to 63
  Class 8 = 64 to 72

6. Result

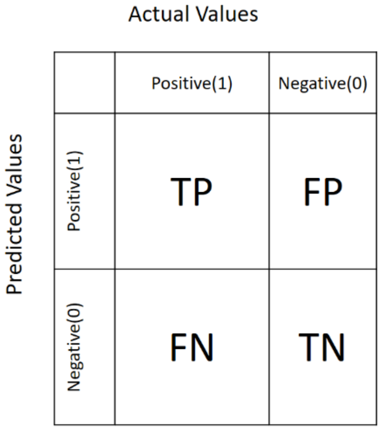
The dataset used in this study is divided into two sets:

1- Training set (80% of the available data)
2- Testing set (20% of the available data).

After applying the classifier algorithm to the dataset, the classification models' results are computed using performance measures. İn this study, a confusion matrix with an accuracy value is used in order to evaluate the performance of the Logistic regression classifier.

A. Confusion Matrix and Accuracy

A confusion matrix (CM) is a performance measurement for a machine learning classification problem where the output is classified into two or more classes. Figure 5 represents four different combinations of predicted and actual values.



		Actual Values	
		Positive(1)	Negative(0)
Predicted Values	Positive(1)	TP	FP
	Negative(0)	FN	TN

Figure 5: Confusion matrix for classification

It is beneficial for computing Recall, Specificity, and Accuracy:

- TP represents the value of True Positive (predicted positive and **true).**
- TN represents the value of **True Negative (**predicted negative and it is negative)
- FP represents the false positive (predicted positive, and it is false). It is the first type of error.
- FN represents the false negative (predicted negative, and it is false). It is the second type of error.

Recall, precision, accuracy, and f-measure values are computed using the following formulas:

$$Recall = TP/(TP + FN) \qquad (1)$$

$$Precision = Tp/(TP + FP) \qquad (2)$$

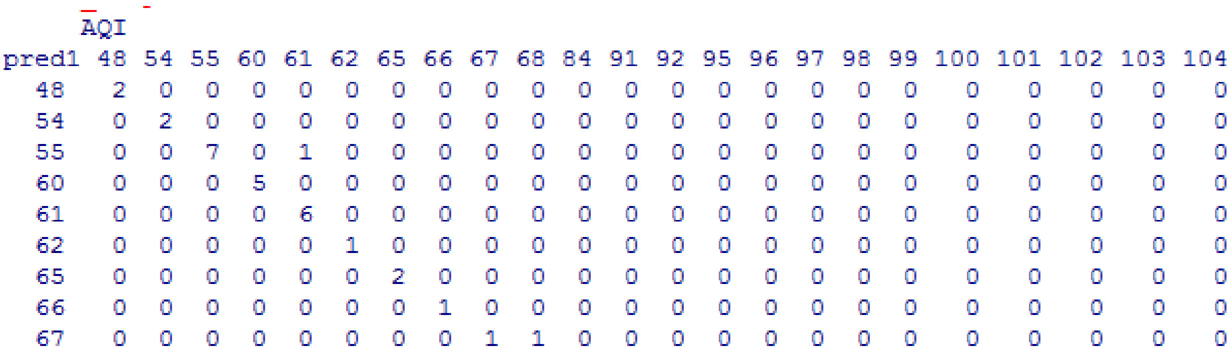
$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \qquad (3)$$

$$F - measure = 2 * \mathrm{Recall} * \mathrm{Precision}/(\mathrm{Recall} + \mathrm{Precision})\mathrm{this4})$$

İn this study, only the accuracy value will be used.

B. Results and discussion

Figure 6 shows a part of the confusion matrix of applying multinomial logistic regression to the dataset. As shown in this Figure, the classifier succeeds in almost every case in detecting the true category of air samples.



```
      AQI
pred1 48 54 55 60 61 62 65 66 67 68 84 91 92 95 96 97 98 99 100 101 102 103 104
   48  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0   0   0   0   0   0
   54  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0   0   0   0   0   0
   55  0  0  7  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0   0   0   0   0   0
   60  0  0  0  5  0  0  0  0  0  0  0  0  0  0  0  0  0  0   0   0   0   0   0
   61  0  0  0  0  6  0  0  0  0  0  0  0  0  0  0  0  0  0   0   0   0   0   0
   62  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0   0   0   0   0   0
   65  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0  0   0   0   0   0   0
   66  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0   0   0   0   0   0
   67  0  0  0  0  0  0  0  0  1  1  0  0  0  0  0  0  0  0   0   0   0   0   0
```

Figure 6: Confusion matrix for Multinomial Logistic Regression

Based on this confusion matrix, the accuracy value was computed. Table 2 shows a comparison between multinomial logistic regression, KNN, and SVM in terms of accuracy value.

	Multinomial Logistic Regression	KNN	SVM
Accuracy	0.8375	0.8279	0.8889

Table 2: Comparison between Multinomial Logistic regression, KNN, and SVM.

The accuracy is the fraction of predictions of the model got right. According to table 2, Multinomial Logistic regression is better than KNN and SVM classifiers. Therefore, this classifier is appropriate for quality prediction.

7. Conclusion

In this paper, the author proposed a new method for air quality prediction based on a multinomial logistic regression algorithm. After presenting the air quality index, the authors reviewed the concept of the multinomial logistic regression technique. Then, a review of existing works in air quality prediction was done. The classification model based on multinomial logistic regression has been described later. Indeed, detecting the air pollution level may avoid many health problems such as cancer, stroke, and pneumonia. This technique was applied later to a real dataset to detect the quality of air. After analyzing the relationship between all variables, the classifier model was applied. The classifier's efficiency was studied based on an accuracy value of around 84%. This accuracy shows that multinomial regression can be used in detecting air pollution. In addition, this classifier is compared to KNN and SVM classifiers regarding accuracy value. In future work, the model's efficiency should be tested using other metrics such as sensitivities and the ROC curve. In addition, a comparison with other models should be made. In addition, the efficiency of artificial intelligence techniques in other fields, such as water pollution, should be studied.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

ORCID iD Ghalia Nassreddine: https://orcid.org/0000-0001-9434-2914

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers.

References

[1] Abed, B. M., Shaker, K., Jalab, H. A., Shaker, H., Mansoor, A. M., Alwan, A. F., & Al-Gburi, I. S. (2016). A hybrid classification algorithm approach for breast cancer diagnosis. *Industrial Electronics and Applications Conference* (s. 269-274). IEEE.

[2] Athanasiadis, I. N., Karatzas, K. D., & Mitkas, P. A. (2006). Classification techniques for air quality forecasting. *Fifth ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence, 17th European Conference on Artificial Intelligence.* Riva del Garda, Italy.

[3] Bayaga, A. (2010). Multinomial Logistic Regression: Usage and Application in Risk Analysis. *Journal of applied quantitative methods, 5*(2).

[4] DeMaris, A. (1995). A tutorial in logistic regression. *Journal of Marriage and the Family*, 956-968.

[5] Dinalankara, W. (2015). Anti-Profiles for Anomaly Classification and Regression.

[6] El-Habil, A. M. (2012). An application of the multinomial logistic regression model. *Pakistan journal of statistics and operation research*, 271-291.

[7] Gheorghe, I. F., & Ion, B. (2011). The effects of air pollutants on vegetation and the role of vegetation in reducing atmospheric pollution. *The impact of air pollution on health, economy, environment and agricultural sources, 29*, 241-280.

[8] Harrop, O. (2018). *Air quality assessment and management: A practical guide.* CRC Press.

[9] Hirschfeld, G., & Von Brachel, R. (2014). Improving Multiple-Group confirmatory factor analysis in R–A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research, and Evaluation, 19*(1), 7.

[10] Kwak, C., & Clayton-Matthews, A. (2002). Multinomial logistic regression. *Nursing Research, 51*(6), 404-410.

[11] Lee, K., Ahn, H., Moon, H., Kodell, R. L., & Chen, J. J. (2013). Multinomial logistic regression ensembles. *Journal of Biopharmaceutical Statistics, 23*(3), 681-694.

[12] Legg, P., Rosin, P., Marshall, D., & Morgan, J. (2007). Improving accuracy and efficiency of registration by mutual information using Sturges' histogram rule.

[13] Li, X., Iervolino, E., Santagata, F., Wei, J., Yuan, C. A., Sarro, P. M., & Zhang, G. Q. (2014). Miniaturized particulate matter sensor for portable air quality monitoring devices. *SENSORS* (s. 2151-2154). IEEE.

[14] Mancl, L., & Spiekerman, C. (2013). *Advanced Regression Methods.*

[15] Niharika, V. M. (2014). A survey on air quality forecasting techniques. *International Journal of Computer Science and Information Technologies, 5*(1), 103-107.

[16] Park, H. A. (2013). An introduction to logistic regression: from basic concepts to interpretation with particular attention to the nursing domain. *Journal of Korean Academy of Nursing, 43*(2), 154-164.