Sains Malaysiana 52(10)(2023): 2971-2983
http://doi.org/10.17576/jsm-2023-5210-18

Classifying Severity of Unhealthy Air Pollution Events in Malaysia: A Decision Tree Model

(Mengelaskan Keparahan Kejadian Pencemaran Udara Tidak Sihat di Malaysia: Hasil Model Pokok Keputusan)

NURULKAMAL MASSERAN[1,*], RAZIK RIDZUAN MOHD TAJUDDIN[1] & MOHD TALIB LATIF[2,3]

[1]Department of Mathematical Sciences, Faculty of Science and Technology
Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
[2]Department of Earth Sciences and Environment, Faculty of Science and Technology
Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
[3]Department of Environmental Health, Faculty of Public Health, Universitas Airlangga, Surabaya, East Java 60115, Indonesia

Received: 16 June 2023/Accepted: 2 October 2023

ABSTRACT

The application of data mining technique in dealing with real problems is popular and ubiquitous in various knowledge domains. This study proposes the concept of severity measures correspond to the characteristics of duration and intensity size for evaluating unhealthy air pollution events. In parallel with that, the present study also proposes a decision tree as a predictive model to deal with a binary classification corresponding to extreme and non-extreme unhealthy air pollution events, which is established based on threshold of the power-law behavior. In a similar vein, other characteristics, such as duration and intensity size, were also determined as important related features. A case study was conducted using the air pollution index data of Klang, Malaysia, from January 1st, 1997 to August 31st, 2020. The results found that the decision tree model can provide a high degree of precision and generalization with 100% accuracy in classifying a class for extreme and non-extreme events for the air pollution severity in the Klang area. In addition, a duration size is the most influential feature that leads to the occurrence of an extreme air pollution event. Thus, this study also suggests that authorities should exercise some vigilance precautions with respect to pollution incidents with a consecutive duration exceeding 11 hours.

Keywords: Air pollution classification; data mining; extreme air pollution; predictive model

ABSTRAK

Pengaplikasian teknik perlombongan data dalam menangani masalah dunia sebenar adalah popular dalam pelbagai domain pengetahuan. Kajian ini mengusulkan konsep ukuran keparahan sepadan dengan ciri tempoh masa dan saiz keamatan untuk menilai kejadian pencemaran udara yang tidak sihat. Selari dengan itu, kajian ini juga mengusulkan kaedah pokok keputusan sebagai model ramalan bagi kes pengelasan binari terhadap kejadian pencemaran udara tidak sihat yang melampau dan tidak melampau yang boleh dikenal pasti berdasarkan nilai ambang tingkah laku hukum-kuasa. Di samping itu, ciri lain iaitu tempoh masa dan saiz keamatan, juga dikenal pasti sebagai ciri berkaitan yang penting bagi suatu kes pencemaran udara. Dalam kajian ini, kajian kes telah dijalankan menggunakan data indeks pencemaran udara di Klang, Malaysia, dari 1 Januari 1997 hingga 31 Ogos 2020. Hasil kajian mendapati model pokok hasil dapat memberikan tahap ketepatan dan pengitlakan yang tinggi dengan ketepatan 100% dalam mengelaskan bagi kejadian pencemaran melampau dan tidak melampau merujuk kepada keparahan suatu pencemaran udara di kawasan Klang. Selain itu, saiz tempoh masa dikenal pasti sebagai adalah ciri berpengaruh yang membawa kepada berlakunya kejadian pencemaran udara yang melampau. Oleh itu, kajian ini juga mencadangkan bahawa pihak berkuasa harus melaksanakan beberapa langkah berjaga-jaga jika kejadian pencemaran udara didapati berlaku dalam tempoh berturut-turut melebihi 11 jam.

Kata kunci: Model peramal; pencemaran udara melampau; pengelasan pencemaran udara; perlombongan data

Introduction

Air pollution has always been a main concern particularly for developed and developing countries as the process of urbanization increases (Al-Kindi et al. 2020; Masseran, 2017; Ouyang et al. 2019). The air pollution problem always exerts damaging effects closely related to human, environmental health, and socioeconomic impacts. The health effect has been reported by many researchers, particularly in terms of the effects corresponding to the respiratory and cardiovascular systems (Maji, Ghosh & Ahmed 2018; Thongtip et al. 2022; Wang, Feng & Chen 2019), morbidity and mortality (Brønnum-Hansena et al. 2018; Sanyal et al. 2018), lung cancer (Hvidtfeldt et al. 2021; Wang et al. 2019), and other diseases (Chau & Wang 2020; Schraufnagel et al. 2019; Zhao et al. 2019). In terms of the environmental effect, air pollution is reported to potentially be a cause of a reduction in agricultural crops (Zhao, Zheng & Wu 2018; Zhao et al. 2021), affect forest sustainability (Agathokleous, Feng & Saitanis 2022), increase plant susceptibility to diseases (Agathokleous & Saitanis 2020), and influence pests and other environmental stresses (Emberson 2020; Masui et al. 2021). Meanwhile, the spillover effect of air pollution has a high risk to influence the socioeconomic situation (Lanzi, Dellink & Chateau 2018), such as decrease life satisfaction and increase anxiety and mental disorders (Lu 2020). Thus, a timely investigation on the risk for air pollution events in any country is a must to mitigate and plan for severe effects from the occurrences of such events.

Air pollution can be classified as extreme or non-extreme depending on the magnitude or severity of an event. As described by Masseran (2021a), in this scenario, an air pollution event that exceeds the threshold of the power-law behavior is extremely unhealthy corresponding to a high level of severity. Thus, a precautionary measure needs to be taken to prevent the occurrence of this event. Accordingly, this study expands the classification analysis of extreme air pollution events by attempting to develop a machine learning model that can accurately predict the features that lead to the occurrences of such events. Thus, a better judgment can be made to mitigate the risk of extreme air pollution events. In the literature, many available techniques can predict an air pollution event. Among the most popular ones are neural network and deep learning model (Bakar et al. 2022; Bekesiene, Meidute-Kavaliauskiene & Vasiliauskiene 2021; Cabaneros, Calautit & Hughes 2019; Haldorai & Ramu 2021; Kow et al. 2022). However, such models are solely designed to achieve a high accuracy of estimation corresponding to a large assemblage's value of parameters. Thus, a neural network model and a deep learning model are difficult to be interpreted and hence is sometimes referred as a black-box model.

On the contrary, classification techniques, such as decision trees, provide a comprehensive criterion in terms of interpretability (Rokach & Maimon 2015). In fact, a decision tree model provides more advantages compared to other classification techniques. For example, because this technique is a nonparametric approach, there is no rigor statistical or mathematical assumption that needs to be fulfilled before applying to any dataset (Rokach & Maimon 2009). In parallel with that, a decision tree is a versatile approach that can be used to deal with the data mining task involving air pollution analysis, such as regression (Ndong et al. 2021), classification (Sarkhosh et al. 2021), feature selection (Zhang et al. 2020), and clustering (Zalakeviciute et al. 2020). In addition, a decision tree is a self-explanatory technique that produces results that are transparent and easy to be interpreted (Lantz 2019; Rizvi, Rienties & Khoja 2019). A decision tree is also a very flexible technique that can be easily used in a variety of data types, including numeric, nominal, and textual values (Malik et al. 2019; Rokach & Maimon 2015). It is also very flexible to deal with missing values, outliers, or errors in a dataset (Feldman & Gross 2005; Hodge & Austin 2004).

The application of decision tree methods in dealing with air pollution data has been presented in several studies around the world. For instance, Tileubai et al. (2023) used a decision tree technique to provide a classification model for defining high and low rates of mortality in Ulaanbaatar, Mongolia, based on 11 attributes that representing air pollution and temperature. They found that the accuracy of decision tree model for their cases is between 60% and 70% along with a sensitivity and the specificity values ranging from 0.50 to 0.75. They conclude that a decision tree able to produce a satisfying results with a simple model development. In fact, a decision tree is very flexible model that can easily be extended to form a new variant model for the purpose of increasing its accuracy, specificity and sensitivity. Among the popular extended version decision tree model are bagging (Breiman 1996), random forest (Breiman 2001), gradient boosting (Friedman 2001), and adaptive boosting (Schapire & Freund 2013). Examples of application of these models in air pollution data can be referred to Mustakim et al. (2023), Putra and Sitanggang (2020), and Shaziayani et al. (2022). However, these modified models should not be used indiscriminately.

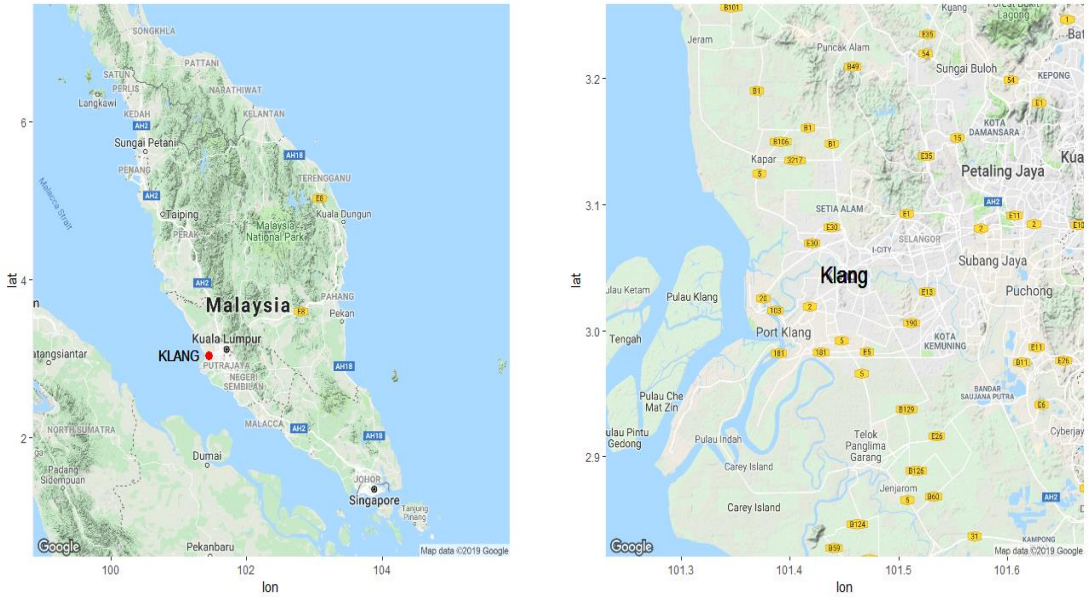
This main reason is because, although these models are able to provide better accuracy, however, the structure of the resulting decision tree is more complex which is generally difficult to interpret against the data.

Along with all of the advantages of decision tree technique, this study try to look at a different perspective in classifying air pollution data by attempting to describe the extreme and non-extreme of air pollution events in terms of their characteristic defined as a duration, intensity and severity size.

STUDY AREA AND DATA

This study analyzed the air pollution index data of Klang, Malaysia. As illustrated in Figure 1, Klang is located at a latitude of 101°26′44.023″ E and longitude of 3°2′41.701″ N, and it is one of the largest cities with a land area of approximately 573 km2 (Google, 2019). The main economic activities in Klang are import and export which operate in Port Klang. Its import and export activities encompass a wide range of products across various industries which mainly include; i) electronics and electrical equipment, ii) petroleum and chemical products, iii) machinery and equipment, iv) automobiles and automotive parts, v) textiles and apparel, vi) consumer goods, vii) plastics and rubber products, and vii) agricultural products. Apart from that, Klang is also an active area for important industrial and economic interests in Malaysia. Klang was recognized as the 13th busiest transshipment port and the 16th busiest container port in the world (Gin 2009). Despite its importance, Klang has an elevated risk of exposure to air pollution (Masseran & Safari 2020). Thus, it is extremely important to investigate the behaviors of the air pollution index (API) in Klang for the purpose of planning and alleviating the risks of extreme air pollution events. The data used in this study were obtained from the Department of Environment Malaysia from January 1st, 1997 to August 31st, 2020.

In general, the Department of Environment Malaysia measures the API values to provide intelligible information on the status of the air quality to the public. Five main sub-pollution indices, namely, nitrogen dioxide (NO_2), sulfur dioxide (SO_2), ozone (O_3), suspended particulate matter of less than 10 microns (PM_{10}), and carbon monoxide (CO), are integrated to represent the API values at a particular time (Department of Environment 1997). The process of determining the API is illustrated in Figure 2 (Masseran 2022b).



(a) (b)

FIGURE 1. (a) Map of Peninsular Malaysia (Klang identified by the red dot); (b) map of Klang

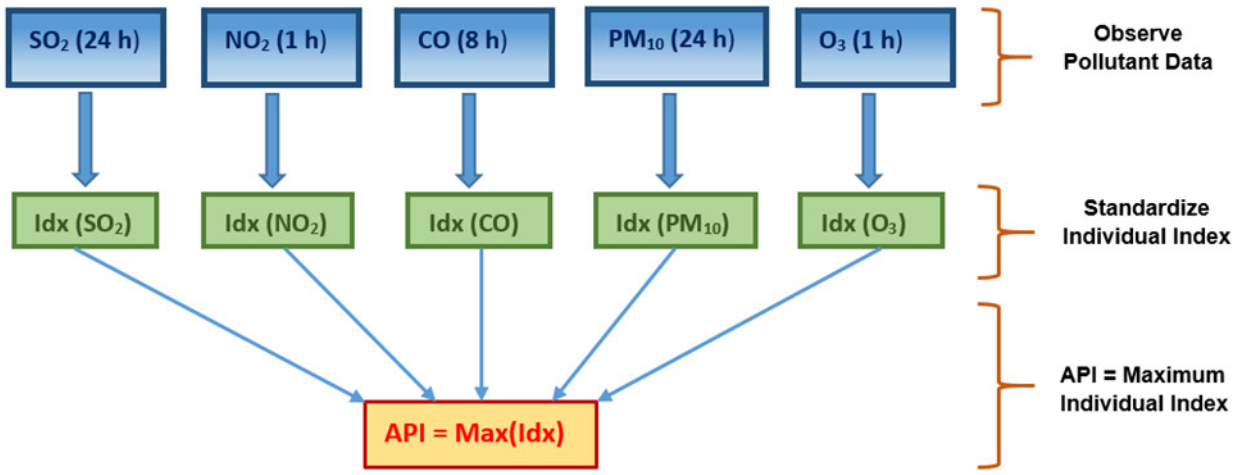


FIGURE 2. Process of determining the API value

AIR POLLUTION CHARACTERISTICS

Unhealthy air pollution events are identified with API values greater than 100 (Department of Environment 1997). For some particular unhealthy air pollution events, information about their duration size can be derived based on the consecutive periods of having API values higher than 100. In particular, the duration of unhealthy air pollution events can be represented as follows

$$D_i = \sum_{j=1}^{N} I_i\left(API_j\right), \qquad \text{for} \quad i = 1,2,3,\ldots,n, \tag{1}$$

where D_i is a random variable for air pollution duration, and $i=1,2,3,\ldots,n$ represents the i-th air pollution event with n as the total number of air pollution events occurring throughout the observed period. In addition, API_j for $j=1,2,3,\ldots,N,$ is an observed time series data with N as the total number of observations (Masseran 2021a). In addition, the characteristic of air pollution severity can be derived from each particular i-th unhealthy air pollution event. Let the indicator function represent data points with an unhealthy state ($API>100$) as follows.

$$I_i\left(API_j\right) = \begin{cases} 1, & if \quad API_j > 100 \\ 0, & if \quad API_j \le 100 \end{cases} \tag{2}$$

Then, the severity of unhealthy air pollution events can be determined from a cumulative of API values greater than 100 corresponding to their duration D_i. This severity measure can be represented as follows,

$$S_i = \sum_{j=1}^{D_i} API_j, \qquad for \qquad \forall D_i, \tag{3}$$

where S_i is a random variable representing the severity of air pollution events. Meanwhile, maximum API value within each particular air pollution event can be determined from the information about the intensity (I_i). Figure 3 illustrates the relationship between the characteristics of the duration, intensity, and severity size corresponding to unhealthy air pollution events.

These three characteristics of air pollution event could provide an important attributes that should be utilized as indicators for analyzing the risk for occurrence of extreme air pollution events. In parallel with that, the higher the value of severity, the more serious the air pollution event. For instance, a prolonged duration or high level of intensity and severity indicates the occurrences of extreme pollution events. The occurrence of this scenario will negatively affect the public health, disrupt the economic activities, and deteriorate the environmental ecosystems (Masseran 2021b). As reported by Masseran (2022a), air pollution events with severity levels greater than a threshold of 1221 exhibit a power-law behavior. Air pollution events corresponding to a severity level that obeys a power-law mechanism have a high risk to provide a disastrous effect on the air quality. Thus, in this study, any unhealthy air pollution event can be categorized as an extreme event if their severity level exceeds the threshold of 1221. Meanwhile, a non-extreme event is determined by a severity measure below the threshold of 1221.

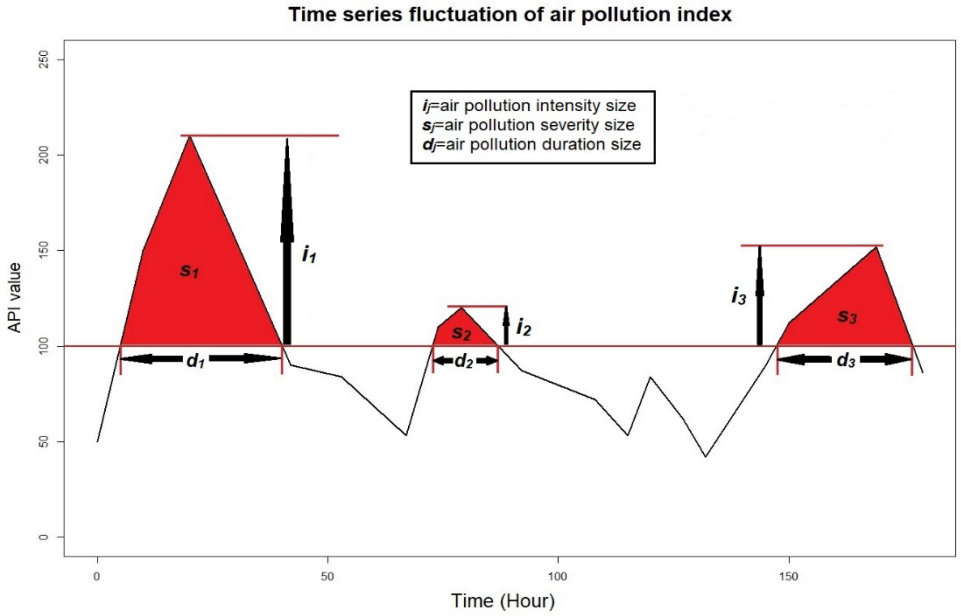


FIGURE 3. Air pollution characteristics based on their duration, intensity, and severity

DECISION TREE CLASSIFICATION

To investigate a suitable predictive model for discriminating the class of extreme air pollution events, this study proposes the application of a decision tree model which is also known as a classification tree. This technique is commonly referred as a classifier of the instance space based on the concept of recursive partition (Cohen, Rokach & Maimon 2007; Rokach & Maimon 2005). As described by Lantz (2019), recursive partition corresponds to the concept, where the dataset will be split into several subsets up to the smallest subset. The construction of a decision tree commonly uses the classification and regression tree algorithm (Breiman 1984). This algorithm provides a splitting process that will only stop if all the subsets have sufficiently homogenous samples or the prescribed stopping criterion conditions are reached. Although the concept of classifiers in a decision tree is quite simple, in most scenarios, this technique is quite popular because it can provide good results in terms of the prediction, classification, and description of the relationship between independent and dependent variables in a dataset (Chang & Wang 2006; Delen, Kuzey & Uyar 2013; Kumar, Mishra & Choudhary 2022; Rizvi, Rienties & Khoja 2019). In fact, the decision tree technique is often used to extract meaningful features and patterns in large datasets for discrimination and predictive modeling (Kamiran, Calders & Pechenizkiy 2013; McCarthy et al. 2019; Myles et al. 2004).

A decision tree is formed from a combination of nested nodes. Here, based on graph theory, a decision tree provides a hierarchical structure in terms of a directed graph starting from a root note (node without any incoming edges). Then, the root node will grow out by splitting into several factions of nodes until it reaches their terminal nodes, known as single-class subspaces, which is referred as a leaf (Rokach & Maimon 2009). As a directed graph, all other nodes except the root node in the decision tree model will have an incoming edge. The nodes between the root and terminal nodes are referred as decision nodes. The process of dividing a node into two or more subspaces are carried out in here, where two or more branches (child nodes) may grow from each decision node (parent node). The splitting process involves certain discrete functions of the input attribute values (Rokach & Maimon 2015). Among the most popular discrete functions used in decision tree classifiers are the information gain and Gini index (Raileanu & Stoffel 2004). The information gain uses the concept of entropy, which can be described as follows,

$$IG = I_f\left(Parent\right) - \sum_k p_k I_f\left(Child_k\right), \tag{4}$$

where $I_f(.)$ is an information function given as

$$I_f(t) = -\sum_j \left(\frac{N_j(t)}{N(t)}\right) \log_2 \left(\frac{N_j(t)}{N(t)}\right), \quad (5)$$

where $N(t)$ is the number of samples in a node t, and $N_j(t)$ is the number of samples belonging to class j that are found to be available in a node t (Myles et al. 2004). Information gain described in Equations (4) and (5) measure the effectiveness of splitting a dataset based on a particular attribute. In which it determines which feature should be chosen as the decision node to create a split that best separates the data into different classes. At each node, the splitting process occurs in essence to maximize the information gain between a parent node and its child node. Meanwhile, the Gini index can be described as

$$Gini = I_m(Parent) - \sum_k p_k I_m(Child_k), \quad (6)$$

where $I_m(.)$ is an impurity function given as

$$I_m(t) = 1 - \sum_j \left\| p(j) \frac{N_j(t)}{N_j} \right\|^2, \quad (7)$$

where N_j is the number of samples belonging to class j (Myles et al. 2004). Equations (6) and (7) measure the impurity of distribution class in decision trees model during data classification. The value of Gini index range from 0 to 1. A node with Gini index equal to 0 or 1 indicates greatest purity of the classification among various classes. On the other hand, a node with Gini index of 0.5 indicates a lowest purity corresponding to equal distribution of elements across different classes.

The process of splitting in decision tree model will continue until all the leaves in a decision tree mode have a homogeneous class or some stopping criteria, such as the maximum depth, are reached. Then, based on the constructed decision tree model, a certain class can be predicted based on the majority representation in that particular class. Predicted probabilities can also be determined based on the proportion of each class within the subgroup (Boehmke & Greenwell 2020).

DATA MINING APPROACH

Data mining has become popular owing to its ability to produce an accurate prediction about certain phenomena. Most techniques in data mining use the concept of inductive learning, in which the original observed data will be divided into two non-overlapping random partitioned datasets known as training and test data samples. In general, 70% of original data will be randomly selected for the training set, whereas the remaining 30% of original data are allocated as a test dataset. The selection of training and test data sets with a ratio of 70:30 was made by random sampling without replacement. In training data, the selection of 70% of the original observation data will be used for the fitting and construction of the decision tree classification model. Meanwhile, the remaining 30% test data is an observational data that is not used for model building (Boehmke & Greenwell 2020; Tan et al. 2019).

A model will be constructed by generalizing it from the training dataset. In our cases, the training dataset will train the decision tree (fitted model) for constructing a suitable classifier. Then, based on the inductive approach, the trained model should be applicable for other unobserved datasets. This assumption will be assessed by evaluating the accuracy of the trained model using the test data sample (Aggarwal 2015; Maimon & Rokach 2009). If the decision tree classification model identified from the training data shows good results, by induction, this model should also give good results on the test data. However, if the opposite is happening, it means that the model is having overfitting problem that needs to be fixed (James et al. 2013; Tan et al. 2019). Likewise, the test dataset will be used for forecasting evaluations that can be assessed using the confusion matrix. For the two problem classes, Table 1 represents their confusion matrix.

TABLE 1. Confusion matrix of the two data classes

	True class		
Predicted class	Extreme event	Non-extreme event	Total
Extreme event	*A (True Event)*	*B (False Event)*	*A+B*
Non-extreme event	*C (False Non-Event)*	*D (True Non-Event)*	*C+D*
Total	*A+C*	*B+D*	*N=A+B+C+D*

Based on the trained model, true-positive and true-negative results indicate the number of data that have been correctly classified in their class. Meanwhile, false-positive and false-negative results indicate the number of elements that are incorrectly classified, also known as errors. In a similar vein, the information provided in the confusion matrix can be used to determine other performance measures for a trained model, which are known as sensitivity, specificity, accuracy, and misclassification rate (Rokach & Maimon 2015). The sensitivity measure is also known as recall, which is described as

$$Sensitivity = \frac{A}{A+C}, \tag{8}$$

This measure determines how well a trained model can recognize positive samples. The specificity measure provides information about how well the trained model can recognize negative samples, which is described as,

$$Specificity = \frac{D}{B+D}, \tag{9}$$

The accuracy of the trained model can be obtained as

$$Accuracy = \left(\frac{A+D}{N}\right) \tag{10}$$

A misclassification rate can be determined as

$$\begin{aligned} Misclassification\ rate &= 1 - Accuracy \\ &= \frac{B+C}{N}, \end{aligned} \tag{11}$$

In general, the training set and test set errors should be low to provide a good estimation of the generalization error fitted model.

Results and Discussions

As mentioned earlier, this study classified air pollution events into two categories. The first category represents the extreme class (1), which corresponds to severity levels greater than the threshold of the power-law behavior, which is 1221. The second category represents the non-extreme class (0), which is determined based on a severity level below the threshold of 1221. Altogether with this binary category, the properties of unhealthy air pollution events can also be obtained from the observed hourly API, as described in Section 'Air Pollution Characteristics'. Then, based on the obtained dataset, 70% of data that represent the severity and the features of air pollution in Klang were randomly selected to be a training dataset to construct a decision tree model. The remaining 30% of the data were allocated as a test dataset to test the prediction accuracy for the fitted decision tree model. Based on the procedures presented in Section 'Data Mining Approach', Figure 4 and Table 2 show the results of the fitted decision tree model for the training data using the concept of the partitioned feature space. This trained decision tree model is found to perfectly classify the class of extreme and non-extreme events of air pollution in Klang. Based on Figure 4, only the duration feature for unhealthy air pollution events is found to be the main factor for the occurrence of extreme air pollution events. In particular, unhealthy air pollution events with a duration size greater than 11 h have a substantial risk of leading to implications. The complexity parameter (cp) plot in Figure 5 provides an agreement with the decision tree plot. The cp plot shows the values of the complexity parameter on the x-axis, while y-axis show a performance metric of decision tree model. The goal is to find the value of the complexity parameter that minimizes the performance metric based on relative error. Based on Figure 5, the cp plot suggests that the size of the decision tree, which is equal to 2,

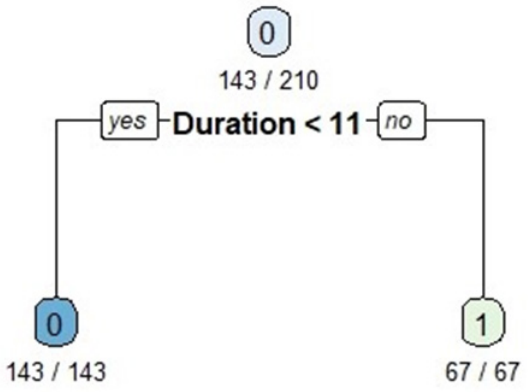


FIGURE 4. Decision tree model of air pollution severity for training data

is an optimal tree size with minimum of relative error. Hence, the fitted decision tree with a simple structure as presented in Figure 4 is sufficient and optimal to represent the relationship of unhealthy air pollution events between their class (extreme and non-extreme event) and its features (duration size and intensity).

Table 2 provides information on the confusion matrix for training data. A similar conclusion can be derived in this matrix information. That is, the computed values for sensitivity, specificity, and accuracy are all equal to 1, whereas the misclassification rate is equal to 0. This result indicates a high precision performance for the fitted decision tree model to our training data. In a similar vein, the ROC curve plot in Figure 6 summarizes the trade-off between the true-positive rate and false-positive rate for a decision tree model. In particular, Figure 6 provides a graphical performance of a decision tree as a classification model to our data in terms of its ability to distinguish between positive and negative instances. Based on Figure 6, it is found that the area under the curve for the ROC plot is equal to 100%, which implies that the decision tree model is a perfect classifier to deal with the data of air pollution events in this study. Nonetheless, it is still very important to evaluate this fitted model into a test dataset to justify that this model is not producing an overfitting problem with a poor generalization performance.

Table 3 shows the results of the class prediction on the test dataset using the fitted decision tree obtained from the training data. The decision tree model can perfectly predict the class of extreme and non-extreme pollution events in our test dataset. Based on the confusion matrix, the computed values for sensitivity, specificity, and accuracy are all equal to 1, whereas the misclassification rate is equal to 0.

In fact, these results provide an agreement with the whole distributional class of extreme and non-extreme air pollution event, as shown in Figure 7. The discriminant line determines based on duration size > 11 h clearly can clearly separate well the two classes of air pollution events. Thus, in overall, we can conclude that the classifier based on the decision tree model produces remarkably high precision accuracy of the prediction corresponding to a good generalization performance.

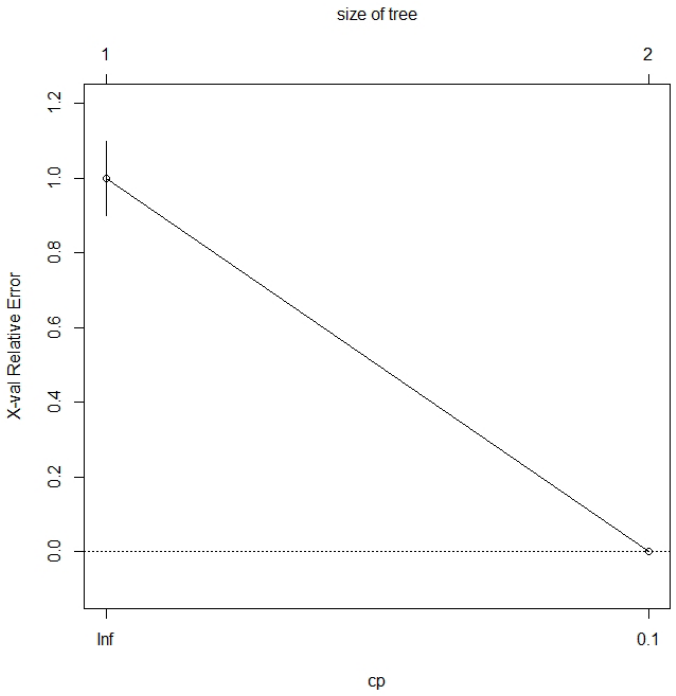


FIGURE 5. Parameter of the cp value for the fitted decision tree model

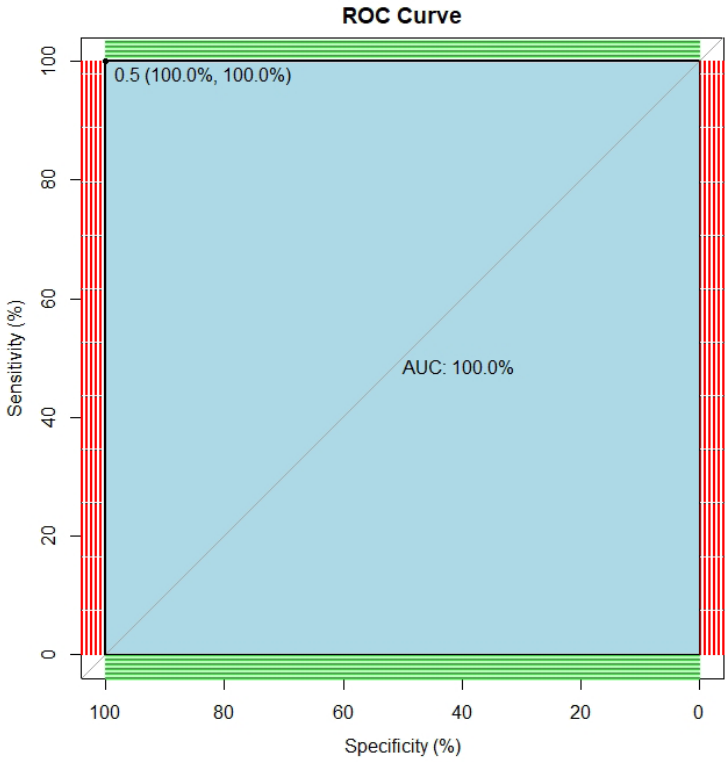


FIGURE 6. ROC plot for the fitted decision tree model

TABLE 2. Confusion matrix for the training data

Predicted class event	True class event	
	Non-extreme severity air pollution	Extreme severity air pollution
Non-extreme severity air pollution	143	0
Extreme severity air pollution	0	67

TABLE 3. Confusion matrix for the test data

Predicted class event	True class event	
	Non-extreme severity air pollution	Extreme severity air pollution
Non-extreme severity air pollution	64	0
Extreme severity air pollution	0	27

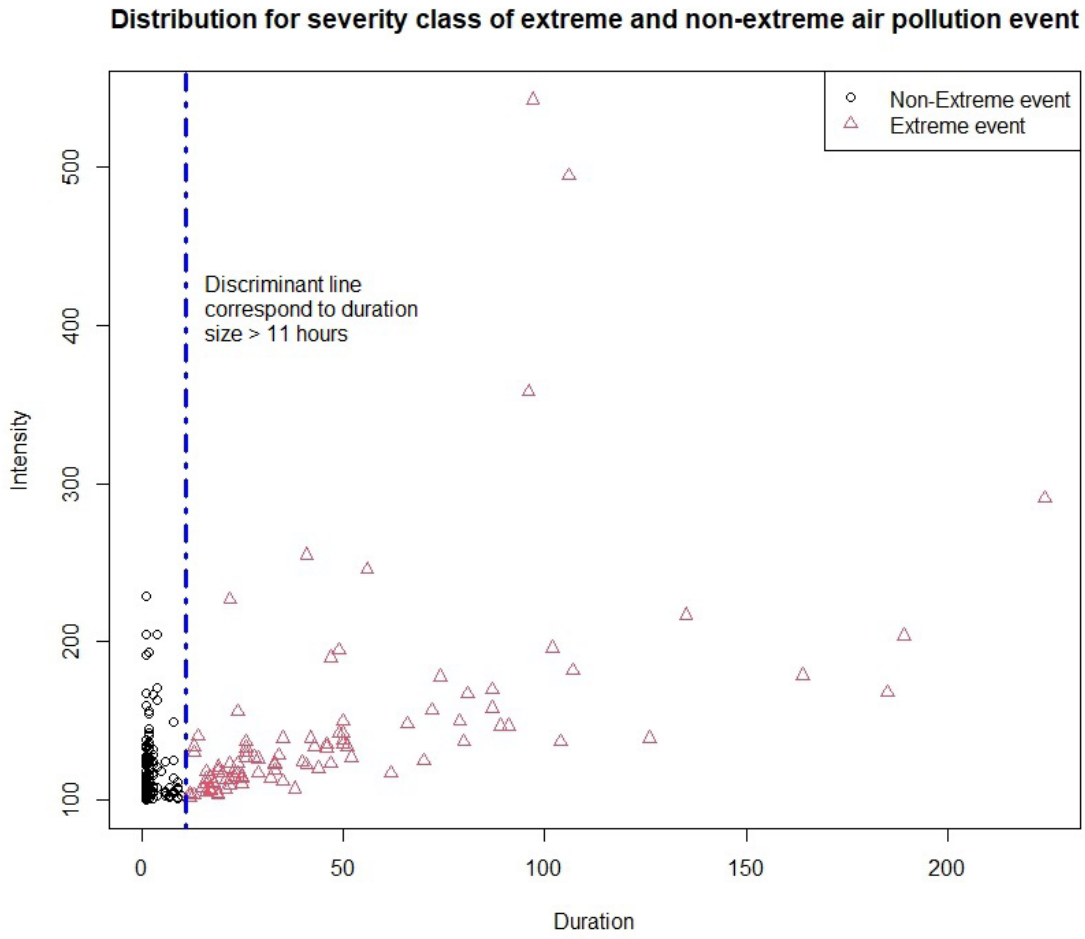


FIGURE 7. Scatter plot for distribution of severity class of extreme and non-extreme air pollution event

CONCLUSIONS

This study proposes a concept to determine a characteristics of unhealthy air pollution event based on a measure of duration, intensity and severity size. Based on these three air pollution characteristics, a class for extreme events is determined using severity level corresponds to a threshold of the power-law behaviors. This scenario leads to a problem of the binary classification class that can be solved using data mining and machine learning. Thus, this study proposes the application of decision trees as a potential machine learning model for classifying extreme and non-extreme air pollution events based on their features described by the characteristics of duration and intensity size of the past occurrence of unhealthy air pollution events. A case study was conducted using the data from Klang, Malaysia. Then, an inductive approach based on a data mining framework was used to train a decision tree model that can represent well the training and testing datasets with a high precision degree. The obtained results show that decision trees can predict well a class for extreme and non-extreme events for air pollution severity with a high degree of precision and generalization in classifying a class for extreme and non-extreme events for air pollution severity class. The results also show that a duration size greater than 11 h is the most important feature that leads to the occurrence of extreme air pollution events in Klang. Thus, this study suggests that authorities should exercise some vigilance precautions with respect to pollution incidents with a consecutive duration exceeding 11 h. Overall, this study concludes that a decision tree is a good machine learning model

with transparent results and provides easy interpretation for the application on air pollution classification. Further research is recommended to evaluate the topological features of air pollution event. Evaluating topological features is an interesting alternative approach that can be used to investigate the air pollution data for the purpose of extracting important features hidden in the data before identifying a classification model.

ACKNOWLEDGEMENTS

The author acknowledges the Dana Impak Perdana 2.0, grant number DIP-2022-002, funded by the Universiti Kebangsaan Malaysia. The authors are also indebted to the Malaysian Department of Environment for providing air pollution data.

REFERENCES

Agathokleous, E. & Saitanis, C.J. 2020. Plant susceptibility to ozone: A tower of Babel? *Sci. Total Environ.* 703: 134962.

Agathokleous, E., Feng, Z. & Saitanis, C.J. 2022. *Effects of Ozone on Forests*. In *Handbook of Air Quality and Climate Change,* edited by Akimoto, H. & Tanimoto, H. Singapore: Springer.

Aggarwal, C. 2015. *Data Mining*. Cham: Springer.

Al-Kindi, S.G., Brook, R.D., Biswal, S. & Rajagopalan, S. 2020. Environmental determinants of cardiovascular disease: Lessons learned from air pollution. *Nat. Rev. Cardiol.* 17: 656-672.

Bakar, M.A.A., Ariff, N.M., Bakar, S.A., Chi, G.P. & Rajendran, R. 2022. Air quality forecasting using temporal convolutional network (TCN) deep learning method. *Sains Malaysiana* 51(11): 3785-3793.

Bekesiene, S., Meidute-Kavaliauskiene, I. & Vasiliauskiene, V. 2021. Accurate prediction of concentration changes in ozone as an air pollutant by multiple linear regression and artificial neural networks. *Mathematics* 9(4): 356.

Boehmke, B. & Greenwell, B. 2020. *Hands-on Machine Learning with R*. Boca Raton: Chapman & Hall/CRC.

Breiman, L. 2001. Random Forests. *Mach. Learn.* 45: 5-32.

Breiman, L. 1996. Bagging predictors. *Mach. Learn.* 24: 123-140.

Breiman, L. 1984. *Classification and Regression Tree*. Boca Raton: Chapman & Hall/CRC.

Brønnum-Hansena, H., Bender, A.M., Andersen, Z.J., Sørensen, J., Bønløkke, J.H., Boshuizen, H., Becker, T., Diderichsen, F. & Loft, S. 2018. Assessment of impact of traffic-related air pollution on morbidity and mortality in Copenhagen Municipality and the health gain of reduced exposure. *Environ. Int.* 121(Part 1): 973-980.

Cabaneros, S.M., Calautit, J.K. & Hughes, B.R. 2019. A review of artificial neural network models for ambient air pollution prediction. *Environ. Model. Softw.* 119: 285-304.

Chang, L-Y. & Wang, H-W. 2006. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accid. Anal. Prev.* 38(5): 1019-1027.

Chau, T.T. & Wang, K.Y. 2020. An association between air pollution and daily most frequently visits of eighteen outpatient diseases in an industrial city. *Sci. Rep.* 10: 2321.

Cohen, S., Rokach, L. & Maimon, O. 2007. Decision-tree instance-space decomposition with grouped gain-ratio. *Inf. Sci.* 177(17): 3592-3612.

Delen, D., Kuzey, C. & Uyar, A. 2013. Measuring firm performance using financial ratios: A decision tree approach. *Expert Syst. Appl.* 40(10): 3970-3983.

Department of Environment. 1997. *A Guide to Air Pollutant Index in Malaysia (API)*. Kuala Lumpur: Ministry of Science, Technology and the Environment. https://aqicn.org/images/aqi-scales/malaysia-api-guide.pdf

Emberson, L. 2020. Effects of ozone on agriculture, forests and grasslands. *Philos. Trans. Royal Soc. A.* 378(2183): 20190327.

Feldman, D. & Gross, S. 2005. Mortgage default: Classification trees analysis. *J. Real Estate Finan. Econ.* 30: 369-396.

Friedman, J.H. 2001. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 29(5): 1189-1232.

Gin, O.K. 2009. *Historical Dictionary of Malaysia*. Lanham: Scarecrow Press.

Haldorai, A. & Ramu, A. 2021. Canonical correlation analysis based hyper basis feedforward neural network classification for urban sustainability. *Neural Process. Lett.* 53: 2385-2401.

Hodge, V. & Austin, J. 2004. A survey of outlier detection methodologies. *Artif. Intell. Rev.* 22: 85-126.

Hvidtfeldt, U.A., Severi, G., Andersen, Z.J., Atkinson, R., Bauwelinck, M., Bellander, T., Boutron-Ruault, M-C., Brandt, J., Brunekreef, B., Cesaroni, G., Chen, J., Concin, H., Forastiere, F., van Gils, C.H., Gulliver, J., Hertel, O., Hoek, G., Hoffmann, B., de Hoogh, K., Janssen, N., Jöckel, K.H., Jørgensen, J.T., Katsouyanni, K., Ketzel, M., Klompmaker, J.O., Krog, N.H., Lang, A., Leander, K., Liu, S., Ljungman, P.L.S., Magnusson, P.K.E., Mehta, A.J., Nagel, G., Oftedal, B., Pershagen, G., Peter, R.S., Peters, A., Renzi, M., Rizzuto, D., Rodopoulou, S., Samoli, E., Schwarze, P.E., Sigsgaard, T., Simonsen, M.K., Stafoggia, M., Strak, M., Vienneau, D., Weinmayr, G., Wolf, K., Raaschou-Nielsen, O. & Fecht, D. 2021. Long-term low-level ambient air pollution exposure and risk of lung cancer - A pooled analysis of 7 European cohorts. *Environ. Int.* 146: 106249.

James, G., Witten, D., Hastie, T. & Tibshirani, R. 2013. *An Introduction to Statistical Learning with Application in R*. New York: Springer.

Kamiran, F., Calders, T. & Pechenizkiy, M. 2013. *Techniques for Discrimination-Free Predictive Models*. In *Discrimination and Privacy in the Information Society. Studies in Applied Philosophy, Epistemology and Rational Ethics*, vol 3, edited by Custers, B., Calders, T., Schermer, B. & Zarsky, T. Berlin: Springer.

Kow, P-Y., Chang, L-C., Lin, C-Y., Chou, C.C-K. & Chang, F-J. 2022. Deep neural networks for spatiotemporal $PM_{2.5}$ forecasts based on atmospheric chemical transport model output and monitoring data. *Environ. Pollut.* 306: 119348.

Kumar, S., Mishra, A.K. & Choudhary, B.S. 2022. Prediction of back break in blasting using random decision trees. *Eng. Comput.* 38: 1185-1191.

Lantz, B. 2019. *Machine Learning with R: Expert Techniques for Predictive Modeling*. 3rd ed. Birmingham: Packt Publishing.

Lanzi, E., Dellink, R. & Chateau, J. 2018. The sectoral and regional economic consequences of outdoor air pollution to 2060. *Energy Econ*. 71: 89-113.

Lu, J.G. 2020. Air pollution: A systematic review of its psychological, economic, and social effects. *Curr. Opin. Psychol.* 32: 52-65.

Maimon, O. & Rokach, L. 2009. Introduction to knowledge discovery and data mining. In *Data Mining and Knowledge Discovery Handbook*, edited by Maimon, O. & Rokach, L. Boston: Springer.

Maji, S., Ghosh, S. & Ahmed, S. 2018. Association of air quality with respiratory and cardiovascular morbidity rate in Delhi, India. *Int. J. Environ. Health Res.* 28(5): 471-490.

Malik, S., Kanwal, N., Asghar, M.N., Sadiq, M.A.A., Karamat, I. & Fleury, M. 2019. Data driven approach for eye disease classification with machine learning. *Appl. Sci.* 9: 2789.

Masseran, N. 2022a. Power-law behaviors of the severity of unhealthy air pollution events. *Nat. Hazards* 112: 1749-1766.

Masseran, N. 2022b. Multifractal characteristics on multiple pollution variables in Malaysia. *Bull. Malaysian Math. Sci. Soc.* 45: 325-344.

Masseran, N. 2021a. Power-law behaviors of the duration size of unhealthy air pollution events. *Stoch. Environ. Res. Risk Asses.* 35: 1499-1508.

Masseran, N. 2021b. Modeling the characteristics of unhealthy air pollution events: A copula approach. *Int. J. Environ. Res. Public Health* 18(16): 8751.

Masseran, N. 2017. Modeling fluctuation of PM_{10} data with existence of volatility effect. *Environ. Eng. Sci* 34(11): 816-827.

Masseran, N. & Safari, M.A.M. 2020. Risk assessment of extreme air pollution based on partial duration series: IDF approach. *Stoch. Environ. Res. Risk Asses.* 34: 545-559.

Masui, N., Agathokleous, E., Mochizuki, T., Tani, A., Matsuura, H. & Koike, T. 2021. Ozone disrupts the communication between plants and insects in urban and suburban areas: An updated insight on plant volatiles. *J. For. Res.* 32: 1337-1349.

McCarthy, R.V., McCarthy, M.M., Ceccucci, W. & Halawi, L. 2019. *Applying Predictive Analytics*. Cham: Springer.

Mustakim, N.A., Ul-Saufie, A.Z., Shaziayani, W.N., Mohamad Noor, N. & Mutalib, S. 2023. Prediction of daily air pollutants concentration and air pollutant index using machine learning approach. *Pertanika J. Sci. & Technol.* 31(1): 123-135.

Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A. & Brown, S.D. 2004. An introduction to decision tree modeling. *J. Chemom.* 18(6): 275-285.

Ndong, G.O., Villerd, J., Cousin, I. & Therond, O. 2021. Using a multivariate regression tree to analyze trade-offs between ecosystem services: Application to the main cropping area in France. *Sci. Total Environ.* 764: 142815.

Ouyang, X., Shao, Q., Zhu, X., He, Q., Xiang, C. & Wei, G. 2019. Environmental regulation, economic growth and air pollution: Panel threshold analysis for OECD countries. *Sci. Total Environ.* 657: 234-241.

Putra, F.M. & Sitanggang, I.S. 2020. Classification model of air quality in Jakarta using decision tree algorithm based on air pollutant standard index. *IOP Conf. Ser.: Earth Environ. Sci.* 528: 012053.

Raileanu, L.E. & Stoffel, K. 2004. Theoretical comparison between the Gini Index and information gain criteria. *Ann. Math. Artif. Intell.* 41: 77-93.

Rizvi, S., Rienties, B. & Khoja, S.A. 2019. The role of demographics in online learning; A decision tree based approach. *Comput. Educ.* 137: 32-47.

Rokach, L. & Maimon, O. 2015. *Data Mining with Decision Trees: Theory and Applications*. 2nd ed. Singapore: World Scientific Publishing.

Rokach, L. & Maimon, O. 2009. Classification trees. In *Data Mining and Knowledge Discovery Handbook*, edited by Maimon, O. & Rokach, L. Boston: Springer.

Rokach, L. & Maimon, O. 2005. Decision trees. In *Data Mining and Knowledge Discovery Handbook*, edited by Maimon, O. & Rokach, L. Boston: Springer.

Rokach, L. & Maimon, O. 2005. Top-down induction of decision trees classifiers - A survey. *IEEE Trans. Syst. Man. Cybern. B Cybern.* 35(4): 476-487.

Sanyal, S., Rochereau, T., Maesano, C.N., Com-Ruelle, L. & Annesi-Maesano, I. 2018. Long-term effect of outdoor air pollution on mortality and morbidity: A 12-year follow-up study for metropolitan France. *Int. J. Environ. Res. Public Health* 15(11): 2487.

Sarkhosh, M., Najafpoor, A.A., Alidadi, H., Shamsara, J., Amiri, H., Andrea, T. & Kariminejad, F. 2021. Indoor air quality associations with sick building syndrome: An application of decision tree technology. *Build. Environ.* 188: 107446.

Schapire, R.E. & Freund, Y. 2013. Boosting: Foundations and Algorithms. *Kybernetes* 42(1): 164-166.

Schraufnagel, D.E., Balmes, J.R., Cowl, C.T., Matteis, S.D., Jung, S-H., Mortimer, K., Perez-Padilla, R., Rice, M.B., Riojas-Rodriguez, H., Sood, A., Thurston, G.D., To, T., Vanker, A. & Wuebbles, D.J. 2019. Air pollution and noncommunicable diseases: A review by the Forum of International Respiratory Societies' Environmental Committee, Part 2: Air pollution and organ systems. *CHEST* 155(2): 417-426.

Shaziayani, W.N., Ul-Saufie, A.Z., Mutalib, S., Mohamad Noor, N. & Zainordin, N.S. 2022. Classification prediction of PM_{10} concentration using a tree-based machine learning approach. *Atmosphere* 13: 538.

Tan, P-G., Steinbach, M., Karpatne, A. & Kumar, V. 2019. *Introduction to Data Mining*. 2 ed. Boston: Pearson Education.

Tileubai, A., Tsend, J., Oyunbileg, B-E., Luvsantseren, P., Luvsan-Ish, A., Chilhaasuren, B., Puntsagdash, J., Chuluunbaatar, G. & Tsagaan, B. 2023. Study of decision tree algorithms: Effects of air pollution on under five mortality in Ulaanbaatar. *BMJ Health Care Inform*. 30: e100678.

Thongtip, S., Srivichai, P., Chaitiang, N. & Tantrakarnapa, K. 2022. The influence of air pollution on disease and related health problems in Northern Thailand. *Sains Malaysiana* 51(7): 1993-2002.

Wang, C., Feng, L. & Chen, K. 2019. The impact of ambient particulate matter on hospital outpatient visits for respiratory and circulatory system disease in an urban Chinese population. *Sci. Total Environ.* 666: 672-679.

Wang, N., Mengersen, K., Tong, S., Kimlin, M., Zhou, M., Wang, L., Yin, P., Xua, Z., Cheng, J., Zhang, Y. & Hu, W. 2019. Short-term association between ambient air pollution and lung cancer mortality. *Environ. Res.* 179(Part A): 108748.

Zalakeviciute, R., Bastidas, M., Buenaño, A. & Rybarczyk, Y.A. 2020. Traffic-based method to predict and map urban air quality. *Appl. Sci.* 10: 2035.

Zhang, Y., Zhang, R., Ma, Q., Wang, Y., Wang, Q., Huang, Z. & Huang, L. 2020. A feature selection and multi-model fusion-based approach of predicting air quality. *ISA Trans.* 100: 210-220.

Zhao, C-N., Xu, Z., Wu, G-C., Mao, Y-M., Liu, L-N., Wu, Q., Dan, Y-L., Tao, S-S., Zhang, Q., Sam, N.B., Fan, Y-G., Zou, Y-F., Ye, D-Q. & Pan, H-F. 2019. Emerging role of air pollution in autoimmune diseases. *Autoimmun. Rev.* 18(6): 607-614.

Zhao, H., Zheng, Y. & Wu, X. 2018. Assessment of yield and economic losses for wheat and rice due to ground-level O_3 exposure in the Yangtze River Delta, China. *Atmos. Environ*. 191: 241-248.

Zhao, H., Zhang, Y., Qi, Q. & Zhang, H. 2021. Evaluating the impacts of ground-level O_3 on crops in China. *Curr. Pollution Rep.* 7: 565-578.

*Corresponding author; email: kamalmsn@ukm.edu.my