Hybrid Method Air Quality Classification Analysis Model

Musli Yanto, Syafri Arlis, Jufriadif Na'am, Yuhandri Yuhandri, Deri Marse Putra

Universitas Putra Indonesia YPTK, Padang, Indonesia

Abstract – **This paper aims to present a discussion of air quality based on the classification analysis developed by hybrid method. Problems that occur, where the longer the air quality is getting worse and can cause serious problems for human life. The method in performing the used analysis consisted of K-Means clustering, Multiple Linear Regression (MRL), Artificial Neural Network (ANN), and Decision Tree Algorithm C.45. The results of MRL measurement show that correlation relationship is the same as the variable with an output of 70.7%. Then the results of determination with ANN showed an MSE value of 0.0018197 and output accuracy of 99.99%**.

Keywords **– analysis model, classification, hybrid method, knowledge based, quality air.**

1. Introduction

Air pollution often experiences a high enough increase to have an impact on environmental pollution and human health [1]. Broadly speaking, the perceived health impacts are respiratory problems [2]. Related to this, the negative effects that will be felt in the future are damage to the lungs, heart, and other organs [3]. For the results to be generated, and analysis process is needed in determining the status of air conditions that will occur next in the short and long term, so that adverse impacts can be anticipated early on.

DOI: 10.18421/TEM112-41
https://doi.org/10.18421/TEM112-41

Corresponding author: Syafri Arlis,
Universitas Putra Indonesia YPTK Padang, Indonesia.
Email: syafri_arlis@upiyptk.ac.id

Received: 02 February 2022.
Revised: 05 May 2022.
Accepted: 11 May 2022.
Published: 27 May 2022.

In the classification of air quality, previous analytical models are generated, such as the model developed to perform a multivariable air pollution prediction [4]. Another discussion also states that the model using the MultiLayer Perceptron (MLP) gives good accuracy results in determining air quality [5]. Other studies also explain the air quality by adopting mathematical calculations of variables that affect air pollution [6]. The analytical model of the Multiple Linear Regression (MRL) and Artificial Neural Network (ANN) methods can be used to predict air quality [7]. MRL and ANN are performance models that are quite good at making predictions and providing accurate results [8].

This paper discusses the air quality classification analysis model. The update contained in this study presents the results of grouping air quality status based on output for the evaluation of knowledge-based system analysis. The evaluation results are taken into consideration in the control and monitoring process. The hybrid method is used to enhance the analysis model that already exists. The stages of analysis in the proposed model start from (1) the preprocessing analysis stage with K-Means dataset clustering to produce a classification pattern, (II) Multiple Linear Regression (MRL) is used to measure the correlation of variables in the resulting pattern, (III) Artificial Neural Network (ANN) learning process. (IV) Classification using the C.45 decision tree method to provide an overview of the classification results on air quality status in the form of a decision tree.

The clustering process can group objects originating from facts or events to find stored information [9], [10]. Discussion in the case of clusters is also able to group factors that affect air quality [11]. Multiple Linear Regression (MRL) is a method to test the correlation between variables and outputs. This correlation test process an overview of the results measurement variables [12]. In previous studies, MRL between two or more variables and was able to measure the relationship between the predictor variables used [13]. Previous studies have explained that MRL can analyze factors that affect air quality with a fairly minimal error rate [14].

The ANN is a supervised learning method used in solving problems with optimal solutions [15]. It can do learning to provide a form of knowledge-based evaluation [16]. In other cases, the ANN is used to identify based on network variables and models [17]. In the same case, the ANN has shown significant and superior results in determining air quality [18]. C.45 is an algorithm in the Decision Tree data mining method used in the making form of patterns that contain information and knowledge [19]. This algorithm is a process used to convert facts into information [20]. The C.45 algorithm is used to find new knowledge [21].

From the explanation above, this paper presents a good concept and analysis model from the previous model. This study also presents another update, namely a new analytical model in the classification of air quality status. The working representation of this model is seen based on the output of the data mining clustering process forming a classification pattern. After that, the MRL will be used to test the accuracy of the pattern in the given correlation relationship. The hybrid method in the classification analysis model provides precise and accurate results of air quality status. The purpose of this paper is to present the concept of a more structured classification analysis model on the status of air quality and to be taken into consideration in decision making. So that the developed model will answer all the doubts of the previous model.

2. Research Method

The classification analysis model that is carried out in the process of determining air quality with this hybrid method provides a structured form of analysis in determining air quality. The following analysis model can be seen in Figure 1.

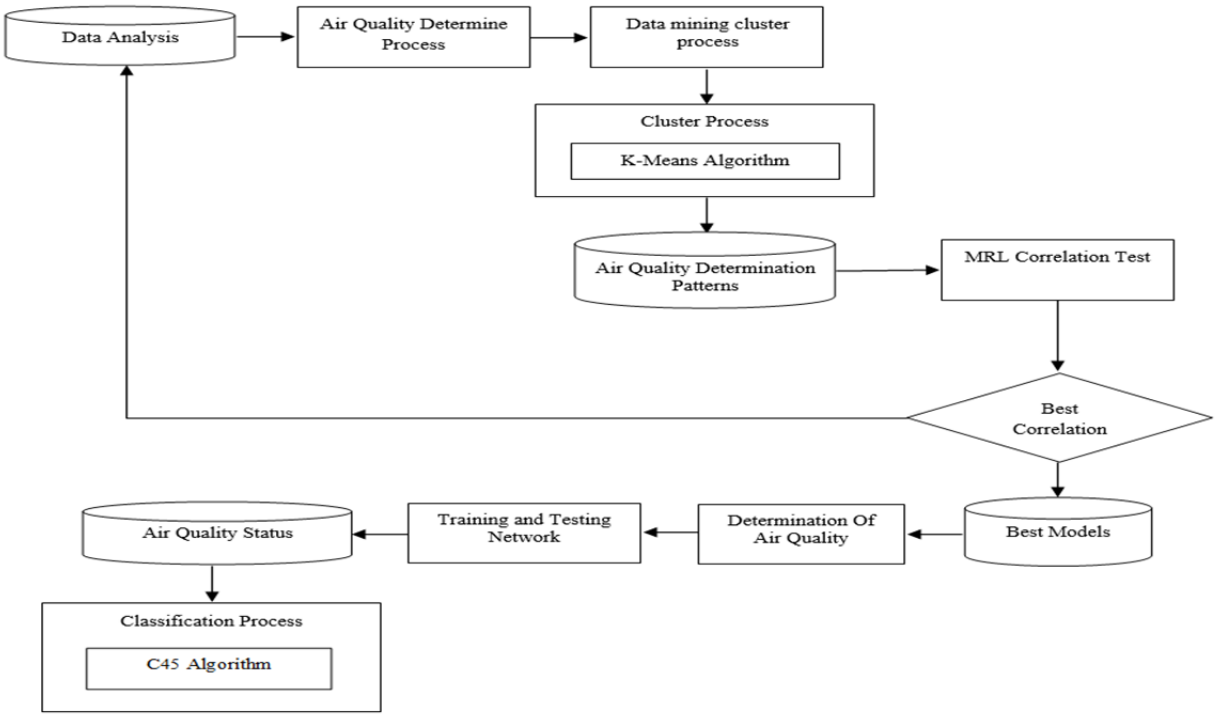

Figure 1. Air quality qnalysis model

Figure 1. above describes the analysis model that will be carried out in determining air quality. The following steps are carried out:

1. The data analysis stage is the stage in determining the variables that affect air quality.
2. Stages of clusters with data mining methods. The preprocessing analysis stage. The clustering process will use the K-means algorithm.
3. Stages of variable correlation test using MRL. This stage aims to test the relationship level of the previously generated rule pattern variables. The output obtained provides the measurement results of the variables used in determining the status of air quality.
4. Stages of ANN learning. This stage will carry out the training and testing process on the previously formed rule pattern. The initial process starts from the formation of the network architecture to arrive at the final process by classifying air quality status. The output results given will be measured by the percentage level of accuracy, sensitivity, and the resulting error value.
5. The final stage in the classification analysis model uses the decision tree algorithm C.45 method. The results given from this process provide an overview of the air quality status rules in the form of a decision tree.

K-Means cluster

A cluster process is an approach that is widely used in data mining [22]. The purpose of the K-Means method is to use grouping by mining data to generate information [23]. The K-means is used to analyze air quality status data. Cluster analysis begins by determining the number of clusters (C1, C2, C3, and C4). After the number of clusters is determined, the process of calculating the distance from the cluster can be continued by using Formula 1 [24]:

$$D_e = \sqrt{(M_{ix,} - C_{ix})^2 + (M_{iy} - C_{iy})^2} \quad (1)$$

Where:
D_e = Euclidean Distance
M_{ix} , M_{iy} = Data Object Coordinates
C_{ix} , C_{iy} = Centroid Coordinate Center

Multiple Regression Linear (MRL)

The Multiple Regression Linear (MRL) method is used to measure the level of relationship between variables [25]. The MRL in principle performs a knowledge-based evaluation based on the correlation between variables [26]. The equations in the MRL can be seen in Formulas 2 & 3 [27]:

$$E\,(Y|X=x_1, X2=x_2,..X_m\, x_m)=\beta_0+\beta_1x_1+....+\beta_m x_m) \quad (2)$$

$$Y=\beta_0+\beta_1x_1+....+\beta_m x_m+\varepsilon, E\varepsilon=0,\ D\varepsilon=DY=\alpha^2 \quad (3)$$

Arificial Neural Network (ANN)

The air quality classification process uses ANN, the approach used is backpropagation. The approach is in the form of a series of feedforward processes or is called reverse inference [28]. Learning is applied to mathematical calculation models and acceptable logic to produce decisions [29]. The network architecture pattern that is built consists of an input layer, a hidden layer, and an output layer [30]. This process aims to produce an optimal network by learning variations of the number of hidden layers [31].

Method Decision Tree

The Decision Tree method is a concept that performs the classification process with the results of the decision tree as the output [32]. The output of this method forms a series of trees and is widely used in describing information [33]. A decision tree has nodes. Each node represents a decision based on the attributes or variables of the dataset. In each of them, there will be a label and a value from the range of attribute values [34].

3. Results and Discussions

Analysis Cluster K-means

In the analysis air quality status classification, the variables used include: (X1) Ozone (O3), (X2) Carbon Monoxide (CO), (X3) Air Particles (PM10), (X4) Nitrogen Dioxide (NO2), (X5) Sulfur Dioxide (SO2), and (X6) Air Pollution Standard Index (ISPU). This variable is found from data on air content conditions in the Padang City area, West Sumatra Province, Indonesia, which occurs for 1 year. The results of the sample K-means cluster analysis process can be seen in Table 1.

Table 1. Air quality satus classification pattern

X1	X2	X3	X4	X5	X6	Y
58.89	22.284	166.5	14.1	0	54	Medium
78.96	23.947	148.6	17.2	0	64.48	Medium
82.46	22.617	112.7	17.7	0	66.23	Medium
55.63	23.501	96.4	14	0	52.83	Medium
38.4	22.428	115.5	13.5	0	38.4	Medium
35.69	24.171	81.1	12.8	0	35.69	Medium
72.38	25.639	137.1	18.2	0	61.23	Medium
77.46	20.705	279.4	17.4	0	63.73	Not Healthy
70.25	20.473	254	14.9	0	60.1	Not Healthy
57.12	22.962	264.1	11.1	0.037	53.56	Not Healthy
103.46	18.437	228.2	13.4	0.085	76.73	Not Healthy
159.21	24.355	242.7	24.7	0.03	105	Not Healthy
275.21	22.266	247.4	34.8	0	225.21	Dangerous
73.75	0.5693	221.3	19.8	0.013	61.88	Dangerous
41.04	0.6157	147.6	9.3	0.04	41.04	Medium
41.92	0.5986	133.3	9.5	0.084	41.92	Medium
54.04	0.5719	116.6	13.1	0.082	52.05	Medium
60.25	0.5241	106.7	11.3	0.069	55.1	Medium
46.83	0.5521	122.5	6	0.099	46.83	Medium
33.58	0.6013	246	9.7	0.102	33.58	Dangerous
46.2	0.573	271.2	9	0.082	46.2	Dangerous
51.2	0.576	221.7	10.5	0.065	50.6	Dangerous
59.4	0.516	551.5	10.4	0.079	54.7	Healthy
94.7	0.531	671.8	17.2	0.07	72.4	Healthy
100.9	0.558	414.5	18.6	0.028	75.5	Healthy
44.7	0.54	245.2	7.7	0.054	44.7	Dangerous
9.6	0.544	368.2	9.1	0.04	9.6	Dangerous
34.7	0.551	376.6	11.3	0.062	34.7	Dangerous
55.7	0.52	254.8	18.8	0	52.9	Dangerous
22.8	0.516	281.9	9.3	0.075	22.8	Dangerous
34	0.545	289.6	8.6	0.052	34	Dangerous
59.4	0.766	317.7	14.1	0.054	54.7	Dangerous
63.3	0.82	314.7	19	0	56.7	Dangerous
45.2	0.844	329.2	9.1	0.041	45.2	Dangerous
39.9	0.761	227.9	10.4	0.053	39.9	Dangerous
11.7	0.856	236.2	14.1	0	11.7	Dangerous
21.2	0.933	260.5	10.2	0.041	21.2	Dangerous
20.5	0.91	226.4	10	0.081	20.5	Dangerous
22.3	0.837	180.1	10.5	0.054	22.3	Not Healthy
16.5	0.838	302.67	10.2	0.037	16.5	Dangerous

Table 1. above, shows that the clustering process forms a classification pattern of air quality. The visualization of the results of the clusterization of air quality status can be seen in Figure 2.

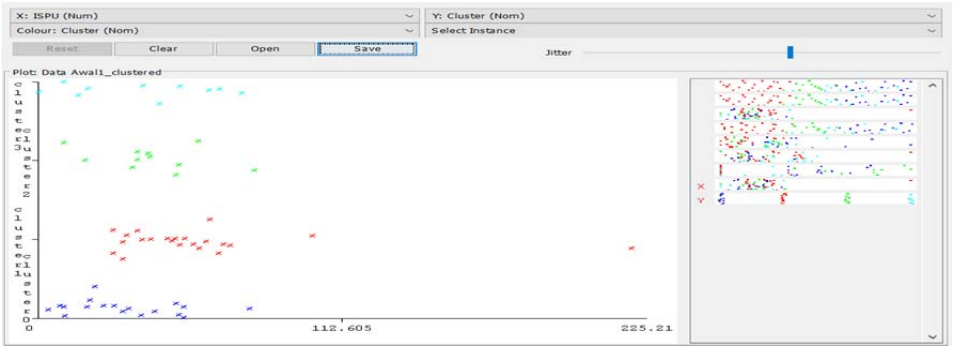

Figure 2. Visualization of air quality status clustering

Figure 2. explains that the cluster results consist of C1 = Heathy by 16%, C2 = Not Healthy by 18%, C3 = Medium by 30% and C4 = Dangerous by 36%. From the results of this cluster analysis, it can be concluded that based on the average of the dataset used, the air quality is in Dangerous status. To be able to see the relationship between variables and the output of air quality status, the analysis process is continued in the analysis process using MRL.

Analysis MRL

In this MRL analysis, the process will start from testing the level of the relationship between variables in determining air quality. The process of regression can be used in measuring the relationship between predictor variables [35]. In its implementation, the MRL can model the independent variable (X) and the dependent variable (Y) [36]. This process can be used as a parameter to see the relationship based on the factors that affect the results [37]. The MRL process for the Coefficient of Determination Test that has been carried out can be seen in Table 2.

Table 2. Results of the Coefficient of Determination

Model Summary				
Model	R	R Square	Adjusted R	Std. Error
1	.707[a]	.500	.440	.40044
a. Predictors: (Constant), ISPU, NO, CO, O3, SO2, PM10				

Table 2. above shows that the results of the measurement of the relationship between the variable and the dependent result produce a result of 70.7%. These results are sufficient to prove that the variables used can affect air quality. To re-test the analysis produced on the results of the determination test contained in Table.2, the correlation test was carried out to see the relationship between the variables and the air quality status. The results of the correlation test can be seen in Table 3.

Table 3. Variable correlation test results

Correlations								
Control Variables			PM10	SO2	CO	O3	NO	ISPU
Status Quality Air	PM10	Correlation	1,000	0,654	-0,179	0,562	-0,251	0,815
		Significance (2-tailed)		0,000	0,172	0,000	0,053	0,000
		df	0	58	58	58	58	58
	SO2	Correlation	0,654	1,000	-0,272	0,302	-0,532	0,540
		Significance (2-tailed)	0,000		0,036	0,019	0,000	0,000
		df	58	0	58	58	58	58
	CO	Correlation	-0,179	-0,272	1,000	-0,035	0,066	-0,159
		Significance (2-tailed)	0,172	0,036		0,793	0,619	0,003
		df	58	58	0	58	58	58
	O3	Correlation	0,562	0,302	-0,035	1,000	-0,435	0,402
		Significance (2-tailed)	0,000	0,019	0,793		0,001	0,005
		df	58	58	58	0	58	58
	NO	Correlation	-0,251	-0,532	0,066	-0,435	1,000	-0,196
		Significance (2-tailed)	0,053	0,000	0,619	0,001		0,002
		df	58	58	58	58	0	58
	ISPU	Correlation	0,815	0,540	-0,159	0,402	-0,196	1,000
		Significance (2-tailed)	0,000	0,000	0,003	0,005	0,002	
		df	58	58	58	58	58	0

Table 3. explains that the relationship between each variable has shown good results with air quality status. This can be seen from the significant value which states less than 0.005. PM10 variable obtained a significant value of 0.000 < 0.005, SO2 variable of 0.000 < 0.005, CO variable of 0.003 < 0.005, NO variable of 0.002 < 0.005 and ISPU variable of 0.004 < 0.005. This is in line with research conducted [38], which states that NO, CO, and SO2 are related to air pollution. Later in the same study it is also explained that PM10 and SO2 have a good relationship in influencing air quality [39]. In another paper it is also stated that the element of NO compounds can also affect air quality [40]. In the research that has been done, it is explained that CO has a significant level of coefficient on the level of air quality [41]. In this case, the MRL analysis can describe the relationship between variables and gives good enough results to test the pattern generated in the previous analysis process so that it can be used in the process of determining air quality. In the analysis that is also carried out in conducting the F-test in Table 4.

Table 4. F Test results

ANOVA[a]						
	Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.017	6	1.336	8.333	.000[b]
	Residual	8.018	50	.160		
	Total	16.035	56			

a. Dependent Variable: Quality Air
b. Predictors: (Constant), ISPU, NO, CO, O3, SO2, PM10

Table 4. illustrates that the F-test results give a significant result of 0.000 which is lower than 5%. This indicates that the variables that have been used together can be used in determining air quality. After the MRL analysis process is carried out, this identification pattern can be continued in the learning process in classifying air quality.

ANN machine learning

The resulting ANN learning process gives good results in analyzing the classification of air quality status. The results are given with an MSE value of 0.000333 and an accuracy rate of 99.99%. In addition, the performance value is 0.0009997 and the MAPE value is 0.000544048. The process of analyzing the status air quality that has been carried out can be seen in Figure 3.

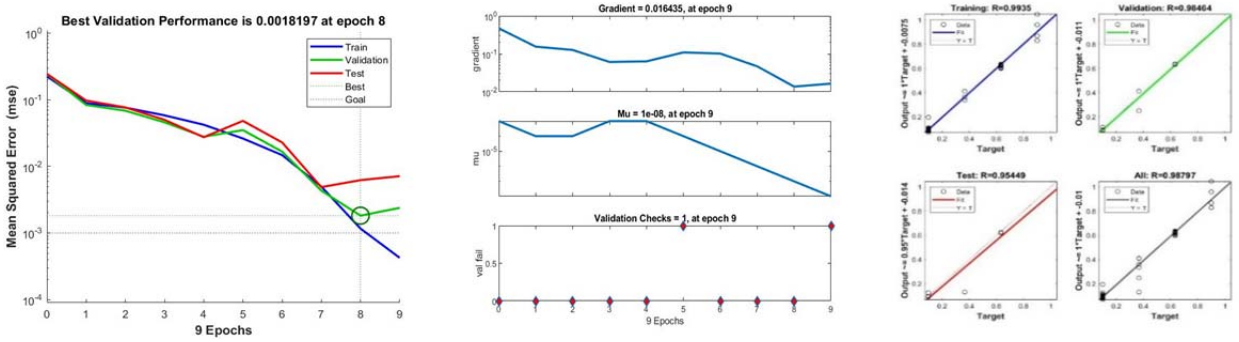

Figure 3. ANN Learning Outcomes

In Figure 3. It can be seen that the ANN learning output graph presents quite good results. The testing process for the classification pattern gives quite good results based on the MSE graph which has a value of 0.0018197. While the resulting validity value is 98.464%. The classification process is continuing in the decision tree analysis process to see the pattern that will be described in a decision tree in the case of air quality.

Classification of Decision Tree

The pattern obtained will be re-analyzed by comparing the results of the cluster carried out using the C.45 approach. This process is able to describe the relationship of the grouping to the data used. The results obtained are able to provide the same results with the aim of classifying the data [42]. In this case, the C.45 method also gives results in the form of a decision tree from the formed pattern [43]. In the process carried out using the C.45 approach, the decision tree is found in Figure 4.

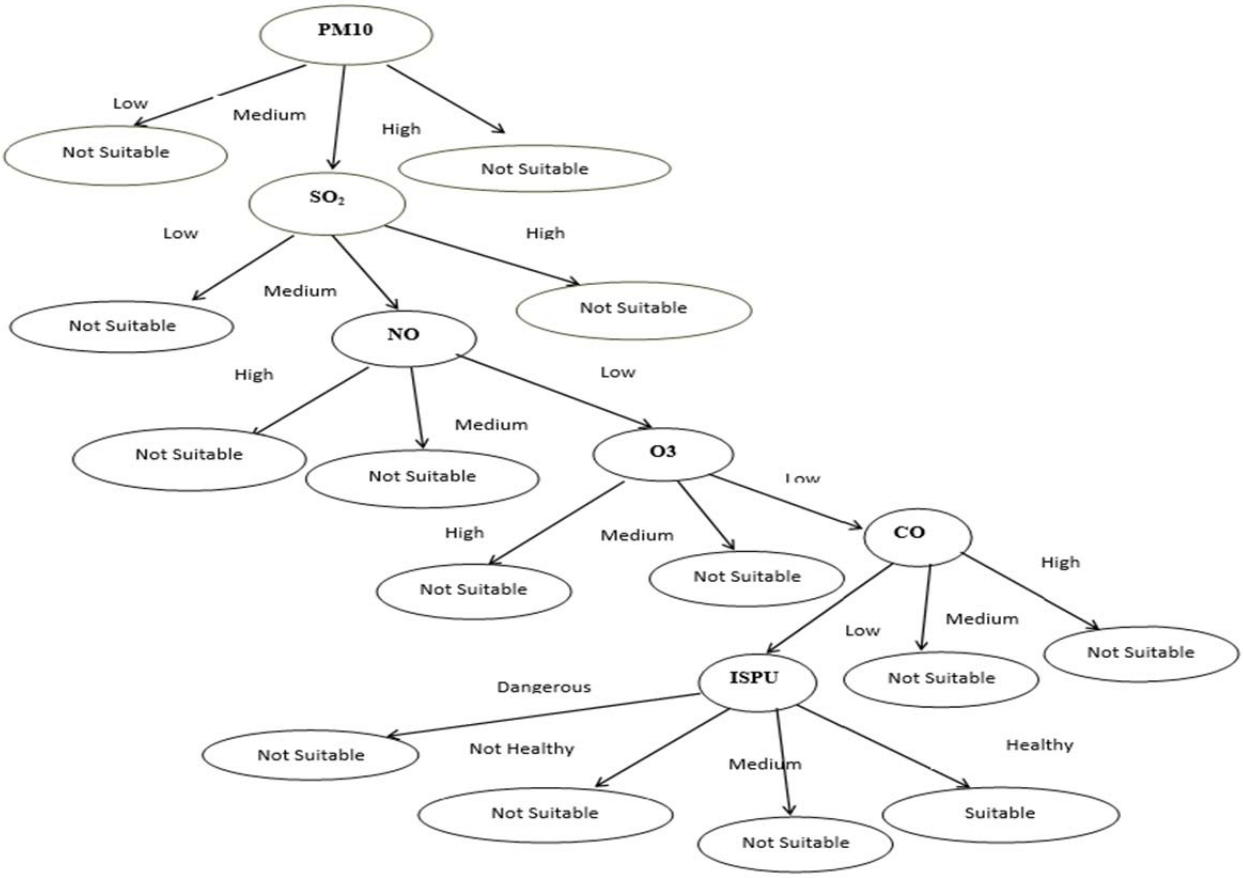

Figure 4. Decision Tree results

Figure 4. explains that the decision tree generated by the decision tree algorithm C.45 can describe the pattern of air quality status. In the picture, it can be seen that the notes and knots used are derived from the patterns obtained previously. This form of decision tree description can be used for controlling and monitoring the development of air quality status.

4. Conclusion

The analysis that has been carried out using the hybrid method approach provides a structured and systematic analysis model in the process of determining air quality. The proposed method can optimize the classification analysis process previously seen from the presented process. The given results also have a fairly good level of accuracy so that they can provide precise and accurate output. The application of the data mining clustering process can group data to form a classification pattern. The MRL analysis developed plays an important role in measuring the accuracy of the variables as well as the correlation with the output which is quite significant. ANN learning on the classification analysis model can be applied to train and test networks using backpropagation algorithms. The results given are quite precise in providing output in determining air quality. In this case, based on the output obtained, it is explained that the research objective is to produce a much better form of the classification analysis process. The results given are used as consideration for decision making.

References

[1]. Adigun, O., & Kosko, B. (2020). Bidirectional Backpropagation. *Ieee Transactions on Systems, Man, and Cybernetics: Systems*, *50*(5). https://doi.org/10.1109/Tsmc.2019.2916096

[2]. Albergaria, J. T., Martins, F. G., Alvim-Ferraz, M. D. C. M., & Delerue-Matos, C. (2014). Multiple linear regression and artificial neural networks to predict time and efficiency of soil vapor extraction. *Water, Air, & Soil Pollution*, *225*(8), 1-9. https://doi.org/10.1007/S11270-014-2058-Y

[3]. Bernstein, J. A., & Levy, M. L. (Eds.). (2014). *Clinical Asthma: Theory and Practice*. CRC Press. https://doi.org/10.1201/b16468

[4]. Alimissis, A., Philippopoulos, K., Tzanis, C. G., & Deligiorgi, D. (2018). Spatial estimation of urban air pollution with the use of artificial neural network models. *Atmospheric environment*, *191*, 205-213. https://doi.org/10.1016/J.Atmosenv.2018.07.058

[5]. Allahyari, E. (2019). Predicting elderly depression: An artificial neural network model. *Iran. J. Psychiatry Behav. Sci*, *13*, 398497. https://doi.org/10.5812/Ijpbs.98497

[6]. Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20-28. https://doi.org/10.38094/jastt20165.

[7]. Chen, B., Zhu, G., Ji, M., Yu, Y., Zhao, J., & Liu, W. (2020). Air Quality Prediction Based on Kohonen Clustering and ReliefF Feature Selection. *Cmc-Computers Materials & Continua*, *64*(2), 1039-1049. https://doi.org/10.32604/Cmc.2020.010583

[8]. Chu, B., Ma, Q., Liu, J., Ma, J., Zhang, P., Chen, T., ... & He, H. (2020). Air Pollutant Correlations in China: Secondary Air Pollutant Responses to NO x and SO2 Control. *Environmental Science & Technology Letters*, *7*(10), 695-700. https://doi.org/10.1021/Acs.Estlett.0c00403

[9]. Dai, W., & Ji, W. (2014). A mapreduce implementation of C4. 5 decision tree algorithm. *International journal of database theory and application*, *7*(1), 49-60. https://doi.org/10.14257/ijdta.2014.7.1.05.

[10]. Damanik, I. S., Windarto, A. P., Wanto, A., Andani, S. R., & Saputra, W. (2019, August). Decision tree optimization in C4. 5 algorithm using genetic algorithm. In *Journal of Physics: Conference Series* (Vol. 1255, No. 1, p. 012012). IOP Publishing. https://doi.org/10.1088/1742-6596/1255/1/012012.

[11]. Deshmukh, M. A., & Gulhane, R. A. (2016). Importance of Clustering in Data Mining. International Journal of Scientific & Engineering Research, 7(2), 247–251.

[12]. Fang, C., Liu, H., Li, G., Sun, D., & Miao, Z. (2015). Estimating the impact of urbanization on air quality in China using spatial regression models. *Sustainability*, *7*(11), 15570-15592. https://doi.org/10.3390/Su71115570

[13]. Gavrilescu, M. (2016). Theoretical predictive air quality models. *The Quality of Air*, 97-116. https://doi.org/10.1016/Bs.Coac.2016.03.019

[14]. Gupta, B. C., Guttman, I., & Jayalath, K. P. (2020). Multiple Linear Regression Analysis. In Statistics And Probability With Applications For Engineers And Scientists Using Minitab, R And Jmp (Pp. 693–756). https://doi.org/10.1002/9781119516651.Ch16

[15]. Hecht-Nielsen, R. (1992). Theory Of The Backpropagation Neural Network**Based On "Nonindent" By Robert Hecht-Nielsen, Which Appeared In Proceedings Of The International Joint Conference On Neural Networks 1, 593–611, June 1989. © 1989 Ieee. In Neural Networks For Perception (Pp. 65–93). https://doi.org/10.1016/B978-0-12-741252-8.50010-8

[16]. Hedar, A. R., Ibrahim, A. M. M., Abdel-Hakim, A. E., & Sewisy, A. A. (2018). K-means cloning: adaptive spherical k-means clustering. *Algorithms*, *11*(10), 151. https://doi.org/10.3390/A11100151

[17]. Hope, T. M. (2019). Linear regression. *Machine Learning: Methods and Applications to Brain Disorders*, 67. https://doi.org/10.1016/B978-0-12-815739-8.00004-3

[18]. Katrina, W., Damanik, H. J., Parhusip, F., Hartama, D., Windarto, A. P., & Wanto, A. (2019, August). C. 45 Classification Rules Model for Determining Students Level of Understanding of the Subject. In *Journal of Physics: Conference Series* (Vol. 1255, No. 1, p. 012005). IOP Publishing. https://doi.org/10.1088/1742-6596/1255/1/012005

[19]. Kotu, V., & Deshpande, B. (2015). Clustering. In Predictive Analytics And Data Mining (Pp. 217–255). https://doi.org/10.1016/B978-0-12-801460-8.00007-0

[20]. Lee, S. H., McKeen, S. A., & Sailor, D. J. (2014). A regression approach for estimation of anthropogenic heat flux based on a bottom-up air pollutant emission database. *Atmospheric Environment*, *95*, 629-633. https://doi.org/10.1016/J.Atmosenv.2014.07.009

[21]. Li, Z., Chen, G., & Li, Q. (2021, March). Model to Predict Wartime Equipment Waste Based on Multiple Regression Analysis. In *Journal of Physics: Conference Series* (Vol. 1802, No. 4, p. 042044). IOP Publishing. https://doi.org/10.1088/1742-6596/1802/4/042044.

[22]. Maheshwari, K., & Lamba, S. (2019, September). Air quality prediction using supervised regression model. In *2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)* (Vol. 1, pp. 1-7). IEEE. https://doi.org/10.1109/Icict46931.2019.8977694

[23]. Hafez, S. M. (2018). Utilization of Artificial Neural Networks in Managing and Planning of Urban Projects. IEREK press. https://doi.org/10.21625/Archive.V2i1.232

[24]. Mustafidah, H., & Suwarsito, S. (2016). Testing Design of Neural Network Parameters in Optimization Training Algorithm. In *International Conference of Result and Community Services, 6th August 2016* (pp. 139-146).

[25]. Nagy, G., Kovács, R., Szőke, S., Bökfi, K., Gurgenidze, T., & Sahbeni, G. (2020). Characteristics of pollutants and their correlation to meteorological conditions in Hungary applying regression analysis. *Időjárás/Quarterly Journal of the Hungarian Meteorological Service*, *124*(1), 113-127. https://doi.org/10.28974/Idojaras.2020.1.6

[26]. Okwu, M. O., & Tartibu, L. K. (2021). Artificial neural network. In *Metaheuristic Optimization: Nature-Inspired Algorithms Swarm and Computational Intelligence, Theory and Applications* (pp. 133-145). Springer, Cham. https://doi.org/10.1007/978-3-030-61111-8_14

[27]. Patel, H. H., & Prajapati, P. (2018). Study and analysis of decision tree based classification algorithms. *International Journal of Computer Sciences and Engineering*, *6*(10), 74-78. https://doi.org/10.26438/ijcse/v6i10.7478.

[28]. Zhang, P., Dong, G., Sun, B., Zhang, L., Chen, X., Ma, N., ... & Chen, J. (2011). Long-term exposure to ambient air pollution and mortality due to cardiovascular disease and cerebrovascular disease in Shenyang, China. *PloS one*, *6*(6), e20827. doi: 10.1371/journal.pone.0020827.

[29]. Qi, J., Yu, Y., Wang, L., & Liu, J. (2016, October). K*-means: An effective and efficient k-means clustering algorithm. In *2016 IEEE international conferences on big data and cloud computing (BDCloud), social computing and networking (SocialCom), sustainable computing and communications (SustainCom)(BDCloud-SocialCom-SustainCom)* (pp. 242-249). IEEE. https://doi.org/10.1109/Bdcloud-Socialcom-Sustaincom.2016.46

[30]. Quirk, T. J., & Quirk, T. J. (2018). Multiple Correlation And Multiple Regression. In Excel 2016 In Applied Statistics For High School Students (Pp. 153–168). https://doi.org/10.1007/978-3-319-89993-0_7

[31]. Quirk, T.J., Rhiney, E. (2017). Multiple Correlation and Multiple Regression. In: Excel 2016 for Advertising Statistics. Excel for Statistics. Springer, Cham. https://doi.org/10.1007/978-3-319-72104-0_7

[32]. Rijayana, I., Fikri, M. I., Razaq, I. F., Achlafass, R. S., & Allshal, R. N. (2020). Using Data Mining with C45 Algorithm for Student Data Classification. *PalArch's Journal of Archaeology of Egypt/Egyptology*, *17*(10), 3827-3832.

[33]. Roy, S. S., Paraschiv, N., Popa, M., Lile, R., & Naktode, I. (2020). Prediction of air-pollutant concentrations using hybrid model of regression and genetic algorithm. *Journal of Intelligent & Fuzzy Systems*, *38*(5), 5909-5919. https://doi.org/10.3233/Jifs-179678

[34]. Sarma, K. V. S., & Vardhan, R. V. (2019). Multiple Linear Regression Analysis. In Multivariate Statistics Made Simple (Pp. 115–134). https://doi.org/10.1201/9780429465185-6

[35]. Shearin, S., Medley, A., Trudelle-Jackson, E., Swank, C., & Querry, R. (2021). Differences in predictors for gait speed and gait endurance in Parkinson's disease. *Gait & Posture*, *87*, 49-53. https://doi.org/10.1016/j.gaitpost.2021.04.019.

[36]. Na, S., Xumin, L., & Yong, G. (2010, April). Research on k-means clustering algorithm: An improved k-means clustering algorithm. In *2010 Third International Symposium on intelligent information technology and security informatics* (pp. 63-67). IEEE. https://doi.org/10.1109/Iitsi.2010.74

[37]. Son, Y., Osornio-Vargas, Á. R., O'Neill, M. S., Hystad, P., Texcalac-Sangrador, J. L., Ohman-Strickland, P., ... & Schwander, S. (2018). Land use regression models to assess air pollution exposure in Mexico City using finer spatial and temporal input parameters. *Science of the Total Environment*, *639*, 40-48. https://doi.org/10.1016/J.Scitotenv.2018.05.144

[38]. Soofastaei, A., Aminossadati, S. M., Arefi, M. M., & Kizil, M. S. (2016). Development of a multi-layer perceptron artificial neural network model to determine haul trucks energy consumption. *International Journal of Mining Science and Technology*, *26*(2), 285-293. https://doi.org/10.1016/J.Ijmst.2015.12.015

[39]. Widyastuti, M., Simanjuntak, A. G. F., Hartama, D., Windarto, A. P., & Wanto, A. (2019, August). Classification Model C. 45 on Determining the Quality of Custumer Service in Bank BTN Pematangsiantar Branch. In *Journal of Physics: Conference Series* (Vol. 1255, No. 1, p. 012002). IOP Publishing. https://doi.org/10.1088/1742-6596/1255/1/012002

[40]. Wu, Y. T., Huang, W. Y., Kor, C. T., Liu, K. H., Chen, T. Y., Lin, P. T., & Wu, H. M. (2021). Relationships between depression and anxiety symptoms and adipocyte-derived proteins in postmenopausal women. *Plos one*, *16*(3), e0248314. https://doi.org/10.1371/journal.pone.0248314.

[41]. Xie, Z., Dong, W., Liu, J., Liu, H., & Li, D. (2021, April). Tahoe: tree structure-aware high performance inference engine for decision tree ensemble on GPU. In *Proceedings of the Sixteenth European Conference on Computer Systems* (pp. 426-440). https://doi.org/10.1145/3447786.3456251.

[42]. Xu, Y., & Lan, S. (2019, December). Time Series Calibration Model for NO2 Based on Multiple Linear Regression. In *2019 International Conference on Economic Management and Model Engineering (ICEMME)* (pp. 313-316). IEEE. https://doi.org/10.1109/Icemme49371.2019.00068

[43]. Yi, X., Zhang, J., Wang, Z., Li, T., & Zheng, Y. (2018, July). Deep distributed fusion network for air quality prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 965-973). https://doi.org/10.1145/3219819.3219822