06_JURNAL_S_2022_Implementation of a Breakpoint Halfway Discretization to Predict Jakarta's Air Quality

By Yulia Resti

WORD COUNT	3911	TIME SUBMITTED	23-APR-2023 04:30PM
		PAPER ID	99002147

Inovasi Matematika (Inomatika)
Vol.4 No.1, Januari 2022, hal.1 – 10
DOI: 10.35438/inomatika.v4i1.310
https://inomatika.unmuhbabel.ac.id/index.php

Implementation of a Breakpoint Halfway Discretization to Predict Jakarta's Air Quality

Winoto Chandra[1], Yulia Resti [2*], Bambang Suprihatin[3]

[1] Doctoral Program of Mathematics and Natural Science, Universitas Sriwijaya, Indonesia
[2,3] Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Sriwijaya,Indonesia
*yulia_resti@mipa.unsri.ac.id

Abstrak

Jakarta adalah salah satu kota paling tercemar di dunia walaupun dalam keadaan pandemi. Mengetahui prediksi kualitas udara harian memberikan banyak manfaat bagi masyarakat, khususnya bagi warga Jakarta. Di antaranya adalah kemampuan untuk mengambil tindakan pencegahan terhadap paparan udara berbahaya. Metode multinomial naïve Bayes dan metode pohon keputusan-ID3 merupakan metode-metode prediksi yang populer dan memiliki kinerja yang baik. Namun kedua metode ini menghendaki variable-variabelnya bertipe kategorik. Ketentuan ini menyebabkan perlunya proses diskritisasi variabel-variabel numerik. Penelitian ini bertujuan untuk memprediksi kualitas udara Jakarta berdasarkan *Particulate Matter* 10 μg (PM_{10}), Sulfur Dioksida (SO_2), Nitrogen Dioksida (NO_2), Ozon (O_3), dan Karbon Monoksida (CO) Jakarta menggunakan metode multinomial naïve Bayes dan pohon keputusan. Variabel-variabel kontinu ini didiskritisasi dengan mengimplementasikan pendekatan *breakpoint halfway* dalam dua cara, *all breakpoints halfway* dan *mixture breakpoints halfway*. Hasil penelitian menunjukkan bahwa metode pohon keputusan dengan pendekatan *mixture breakpoints halfway* memiliki kinerja yang lebih baik dibandingkan dengan kinerja metode multinomial naïve Bayes dengan akurasi 98,90%, spesifisitas 98,92%, sensitivitas 75,00%, presisi 74,00%, dan skor F1 sebesar 97,81%.

Kata kunci: Kualitas Udara, Jakarta, Multinomial Naïve Bayes, Pohon Keputusan-ID3.1

Abstract

Despite the pandemic, Jakarta is one of the most polluted cities in the world. Knowing the daily air quality forecast aids the community, particularly Jakarta residents. Among these is the ability to protect oneself from dangerous air. The multinomial naive Bayes and the decision tree-ID3 methods are popular and perform well. Both of these strategies, however, require categorical variables. This need necessitates the implementation of a discretization technique for numerical variables. The purpose of this study is to predict Jakarta's air quality using the multinomial naive Bayes and decision tree method based on Particulate Matter 10 µg (PM_{10}), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), Ozone (O_3), and Carbon Monoxide (CO). These continuous variables are discretized in two ways: using all midway breakpoints or halfway mixture breakpoints. The results indicated that the decision tree method with the mixture breakpoints halfway approach performed better than the multinomial nave Bayes method, with an accuracy of 98.90%, a specificity of 98.92%, a sensitivity of 75.00%, a precision of 75.00%, and an F1 score of 97.81%.

Keywords: *Air Quality, Jakarta, Multinomial Naïve BAYES, Decision Tree-ID3*

Received: October 14, 2021/ Accepted: December 12, 2021/ Published Online: January 31, 2022

INTRODUCTION

Jakarta has some of the worst air quality globally, even during a pandemic (Pranita, 2020). This assertion pertains to the Centre for Research on Energy and Clean Air (CREA) which measured air quality in Jakarta from January to May 2020. The curiosity in predicting air quality in metropolitan areas, particularly Jakarta, is growing in importance as the detrimental effects of air pollution on people's health and the environment. Air pollution is a global epidemic caused by chemical and biological compounds, as well as particulate matter (PM), and has a variety of negative effects on the environment and human health (Li et al., 2017). Particulate matter (PM) is a substantial component of urban air pollution that has a negative influence on the environment and human health. It has considerable potential for accumulation in the respiratory system of humans. The smaller the particulate size, the more likely it will settle in the human respiratory system and be breathed. There are two forms of particulate matter in general: PM less than 10 μg (PM_{10}) and PM less than 2.5 μg ($PM_{2.5}$). PM10 is inhaled and may enter the human respiratory system. Additionally, it might result in decreased sight, smoke, haze, and smog (Agustine et al., 2018).

The level of air quality can be indicated by the Air Pollution Index (API). In Indonesia, the API is called the Indeks Standar Pencemar Udara, shortened as ISPU. The ISPU was calculated by converting each of the five measured air pollution indicators to a dimensionless number; Particulate Matter 10 μg (PM_{10}), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), Ozone (O_3), and Carbon Monoxide (CO are the five indicators. For example, in Jakarta, the public can gain numerous benefits from understanding the daily air quality forecast. Among these is the ability to take precautions in the event of hazardous air quality.

In statistical learning, multinomial naïve Bayes (MNB) and decision trees-ID3 (DT) are two of the most widely used and straightforward classification methods. The MNB method is based on Bayes' theorem. Each variable must follow a multinomial distribution if there are more than two categories or a binomial distribution if there are only two categories (Pan et al., 2018; Chen & Fu, 2018). The DT method employs a tree structure representation in which each node represents a variable, each branch represents its value, and each leaf represents the class. Decision trees are pretty accurate in various situations (Han et al., 2012; Witten & Frank, 2005).

The MNB and DT methods are both popular and perform well. Both of these methods, however, require categorical variables. This need necessitates the implementation of a discretization technique for numerical variables (García et al., 2015). The purpose of this study is to predict Jakarta's air quality using the MNB and DT methods based on Particulate Matter 10 μg (PM10), Sulfur Dioxide (SO2), Nitrogen Dioxide (NO2), Ozone (O3), and Carbon

Monoxide (CO). These continuous variables are discretized in two ways: using all midway breakpoints or halfway mixture breakpoints (Witten & Frank, 2005). The performance measures: accuracy, sensitivity, specificity, precision, and F1 score are then compared-the greater the values of these indicators, the better the method's performance.

METHOD

The data used in this research is sourced from [http://data.jakarta.go.id/dataset/. The air quality is collected from January 1, 2017, to December 31, 2019, and monitored from five stations in Jakarta.

The multinomial naive Bayes (MNB) and decision tree-ID3 (DT) methods (Pan et al., 2018; Chen & Fu, 2018; James et al., 2013; Han et al., 2012) were employed in this study to develop a predictive model of Jakarta's air quality. These two methods are frequently successful in achieving accuracy while doing prediction/classification tasks (Kresnawat et al., 2021; Resti et al., 2021; Hussein et al., 2020; Chou & Lo, 2018). In addition, we proposed all midway breakpoints and halfway mixture breakpoints for the discretization of the numerical variables (Witten & Frank, 2005).

The MNB method is based on the Bayes theorem, which states that the maximum posterior probability of each observation is achieved by multiplying the prior probability by the likelihood probability. Suppose $P(X_k|Y_{healthy})$, $P(X_k|Y_{moderate})$, $P(X_k|Y_{unhealthy})$, $P(X_k|Y_{very\ unhealthy})$, and $P(X_k|Y_{hazardous})$ each is the likelihood probability of healthy air quality, moderate, unhealthy, and very unhealthy, each of which is written as follow,

$$P(X_k|Y_{healthy}) = \frac{\sum_c^m n_c(X_k|Y_{healthy}) + 1}{n(X_k|Y_{healthy}) + m} \quad (1)$$

$$P(X_k|Y_{moderate}) = \frac{\sum_c^m n_c(X_k|Y_{moderate}) + 1}{n(X_k|Y_{moderate}) + m} \quad (2)$$

$$P(X_k|Y_{unhealthy}) = \frac{\sum_c^m n_c(X_k|Y_{unhealthy}) + 1}{n(X_k|Y_{unhealthy}) + m} \quad (3)$$

$$P(X_k|Y_{very\ unhealthy}) = \frac{\sum_c^m n_c(X_k|Y_{very\ unhealthy}) + 1}{n(X_k|Y_{very\ unhealthy}) + m} \quad (4)$$

Posterior probability for Y_j, $j = healthy, moderate, unhealthy, very\ unhealthy$ is,

$$P(Y_j|X_1, \cdots, X_d) = \arg\max P(Y_j) \prod_{k=1}^{d} P(X_k|Y_j) \quad (5)$$

where the prior probability of each group is defined as,

$$P(Y_{healthy}) = \frac{\sum_{k=1}^{d} n(X_k|Y_{healthy}) + 1}{n + g} \quad (6)$$

$$P(Y_{moderate}) = \frac{\sum_{k=1}^{d} n(X_k|Y_{moderate}) + 1}{n + g} \quad (7)$$

$$P(Y_{unhealthy}) = \frac{\sum_{k=1}^{d} n(X_k|Y_{unhealthy}) + 1}{n + g} \quad (8)$$

$$P(Y_{very\ unhealthy}) = \frac{\sum_{k=1}^{d} n(X_k|Y_{very\ unhealthy}) + 1}{n + g} \quad (9)$$

For each Y_j, $n_c(X_k|Y_j)$ is the number of days predicted to have the j-th air quality in the variable X_k with category c, $n(X_k|Y_j)$ is the number of days predicted to have the j-th air quality in all variable X, $n(Y_j)$ is the number of days of observation, m is the number of categories in the variable X_k, and g is the total number of air quality categories.

The DT is a classification technique that utilizes a tree form similar to a flow chart (Han et al., 2012). The primary processes in building a decision tree are as follows: (1) determining a predictor variable as the root; (2) loading the branch for each value; (3) classifying each branch, and (4) repeating the process for each branch until all cases in that branch are classified as the same class. The most significant gain value of all variables is used to determine which variable is the root. Calculate the entropy of all values in the variable before calculating the maximum gain value. Entropy is a quantity used to determine the variance of the sample data. After determining the entropy value of the sample data, the most influential variable will be a measure of data classification, referred to as information gain. Consider successively S, k_s, Pi, X, k_X, $|S|$, and $|S_i|$ as a set of cases, the number of partitions/categories in S, the prior probability of each predictor variable, the number of partitions/categories in the variable X, the number of cases in S, and the number of cases in the m– partition, respectively. Entropy and gain are calculated by utilizing,

$$\text{Entropy (S)} = \sum_{m=1}^{k_s} -P_m \log_2 P_m \quad (10)$$

$$\text{Information Gain(S,X)} = \text{Entropy (S)} - \sum_{m=1}^{k_X} \frac{|S_m|}{|S|} \text{Entropy } (S_m) \quad (11)$$

The final stage is to evaluate the method's performance. Scalar values are used to quantify classification performance using a variety of metrics, including accuracy, recall-micro (μ), recall-macro (M), specificity-micro (μ), and specificity-macro (M). The values of TP_j, FP_j,

TN_j , and FN_j are determined for air quality degree, for each of $j = healthy, moderate, unhealthy, very\ unhealthy$, where TP is the true-positive, FP is the false-positive, FN is the true-negative, and TN is the true-negative. The performance measurements for the first air quality degree are listed in Table 1. Other air quality varieties have a comparable performance metric (Dinesh & Dash, 2016; Sokolova & Lapalme, 2009).

Table 1. Confusion Matrix the first air quality degree

	j	Actual			
		Health	**Moderate**	**Unhealthy**	**Very unhealthy**
Prediction	**Healthy**	TP	FN	FN	FN
	Moderate	FP	TN	TN	TN
	Unhealthy	FP	TN	TN	TN
	Very unhealthy	FP	TN	TN	TN

$$\text{Accuracy} = \frac{\sum_{j=1}^{4} \frac{TP_j + TN_j}{TP_j + FP_j + FN_j + TN}}{4} \quad (12)$$

$$\text{Specificity} = \frac{\frac{\sum_{j=1}^{4} TN_j}{\sum_{j=1}^{4} (FP_j + TN_j)}}{4} \quad (13)$$

$$\text{Sensitivity} = \frac{\sum_{j=1}^{4} \frac{TP_j}{TP_j + FN_j}}{4} \quad (14)$$

$$\text{Precision} = \frac{\sum_{j=1}^{4} \frac{TP_j}{TP_j + FP_j}}{4} \quad (15)$$

$$F_1\text{Score} = \frac{2\text{Precision}(\text{Sensitivity})}{(\text{Precision} + \text{Sensitivity})} \quad (16)$$

RESULTS

Daily air quality data were collected from January 1, 2017, to December 31, 2019. There were 47 missing data points in the 1095 observations during the 2017 period, resulting in the observation point being 1048. The data is separated into two portions during the prediction process, training data and test data. The training data are used to construct a prediction/classification learning model, while the test data are used to validate the previously constructed model. The data for learning was collected between 2017 and 2018 (65.17%), while the data for testing were collected in 2019 (34.83%). As demonstrated in Table 2, this data set is composed of four components.

Table 2. Composition Data Based on Target Variable

Data	Target Variable (Y)				Total	
	healthy	moderate	unhealthy	very unhealthy	Sum	%
Training	51	325	280	27	683	65.17
Testing	2	175	180	8	365	34.83

Assume X_1, X_2, X_3, X_4, dan X_5 represent particulate matter 10 μg (PM_{10}), Sulfur Dioxide (SO_2), Carbon Monoxide (CO), Ozone (O_3), and Nitrogen Dioxide (NO_2), respectively. The predictor variables in this study are all numerical types. The discretization of these numerical variables uses two ways, as presented in Table 3.

Table 3. Discretization of Predictor Variable

Variable	Value	Discretization			
		ABH		MBH	
		Category	Range	Category	Range
X_1	0-107	1	≤ 57	1	<40
		2	>57	2	40-80
				3	81-120
X_2	0-72	1	≤ 29	1	≤ 29
		2	>29	2	>29
X_3	0-88	1	≤ 18	1	≤ 18
		2	>18	2	>18
X_4	0-234	1	≤ 94	1	<50
		2	>94	2	50 – 100
				3	101-150
				4	151-300
X_5	0-34	1	≤ 12	1	≤ 12
		2	>12	2	>12

The first way, all variables are discretized using halfway breakpoints (all breakpoints halfway /ABH). The second way, not all variables are discretized using halfway breakpoints. Only variables that do not have information related to their categorization are discretized. So it is a mixture of breakpoint halfway with existing information (mixture breakpoint halfway/MBH). Here, the variables are X_2, X_3, and X_5.

The results of the 2019 data prediction using the MNB-ABH, the MNB-MBH, the DT-ABH, and the DT-MBH methods from the 2017-2018 data learning classification results are presented in Table 4, Table 5, Table 6, and Table 7.

Table 4. Air Quality Prediction using the MNB-ABH

Prediction	Actual			
	healthy	**moderate**	**unhealthy**	**very unhealthy**
healthy	2	0	0	0
moderate	57	94	6	18
unhealthy	0	0	28	152
very unhealthy	0	0	0	8

In the prediction results using MNB-ABH, the largest FN value is owned by the healthy group, followed by the very unhealthy group, and the unhealthy group. The moderate group has no FN scores. On the other hand, the largest FP value was owned by the unhealthy group and followed by the moderate group. The healthy and the very unhealthy groups do not have FP values. This prediction result makes MNB-ABH only obtain high specificity (more than 83%).

Table 5. Air Quality Prediction using the MNB-MBH

Prediction	Actual			
	healthy	**moderate**	**unhealthy**	**very unhealthy**
healthy	0	2	0	0
moderate	0	174	0	1
unhealthy	0	0	133	47
very unhealthy	0	0	0	8

Only the healthy group lacks a TP value in the MNB-MBH prediction results. The highly unhealthy group had the highest FN value, followed by the moderate group. Both the healthy and ill groups lack a FN value. As with the MNB-ABH prediction, the unhealthy group has the highest FP value, followed by the healthy and moderate groups. The group classified as extremely unwell has no FP value. As a result of this prediction, MNB-ABH only has a high specificity (more than 83%). As a result of this prediction outcome, MNB-MBH achieves a greater accuracy, F1Score, and specificity (above 86%) when compared to sensitivity and precision.

Table 6. Air Quality Prediction using the DT-ABH

Prediction	Actual			
	healthy	**moderate**	**unhealthy**	**very unhealthy**
healthy	0	2	0	0
moderate	0	151	24	0
unhealthy	0	0	180	0
very unhealthy	0	0	8	0

Additionally, the healthy group does not have a TP value as predicted by MNB-MBH in the prediction findings using DT-ABH. Additionally, the highly unhealthy group lacks a TP value, and both groups lack a FN value. The unhealthy group has the highest FN value, followed by the moderate group. The moderate group has the highest FP value, followed by the healthy group. As a result of the prediction results, DT-ABH achieves a greater accuracy, F1Score, and specificity (above 90%) than sensitivity and precision.

Table 7. Air Quality Prediction using the DT-MBH

Prediction	Actual			
	healthy	**moderate**	**unhealthy**	**very unhealthy**
healthy	2	0	0	0
moderate	0	175	0	0
unhealthy	0	0	180	0
very unhealthy	0	0	8	0

In the prediction results using DT-MBH, only the unhealthy group has an FN value and only the very unhealthy group has an FP value. This predictive result makes DT-MBH obtain the highest all sizes compared to other methods. Especially accuracy, F1Score, and specificity which is more than 97%. The performance of air quality prediction results using these methods is shown in Table 8.

Table 8. Performance of Air Quality Prediction

Method	Accuracy	Sens	Prec	F_1Score	Specificity
MNB-ABH	68.90	55.17	67.00	36.16	83.36
MNB-MBH	93.22	68.00	53.00	86.30	96.38
DT-ABH	93.70	46.57	45.90	90.68	95.41
DT-MBH	98.90	75.00	74.00	97.81	98.92

DISCUSSION

This article presents a breakpoint midway technique for predicting Jakarta's air quality using two statistical learning methods. Multinomial Naive Bayes and Decision Trees are the two techniques. Both of these methods require that all variables are of a categorical type. Therefore, the numerical variables in the research data need to be discretized first to obtain categorical type variables. The median value serves as the cutoff point for identifying halfway breakpoints. Discretization using a halfway breakpoint technique can be accomplished in two ways: using all breakpoints halfway (ABH) or using a mixture of halfway breakpoints (MBH). Two variables that have information related to the value level are PM_{10} (X_1) dan O_3 (X_4) so that

in MBH, only three variables are discretized using halfway breakpoints. The results of air quality prediction using two distinct discretization approaches demonstrate that the identical method produces significantly different predictive performance for both the MNB and DT methods. Overall, the DT technique outperforms the MNB method for air quality prediction when both ABH and MBH discretization is used, and MBH is implemented better in each method than ABH. The accuracy and specificity of air quality prediction utilizing the DT-MBH, DT-ADH, and MNB-MBH methods in this work are rated as very good (Mishra et al., 2016), with the accuracy and specificity of this prediction exceeding 93%. Simultaneously, the F1 score is greater than 86%. Air quality prediction results using the DT-MBH method have the best performance with accuracy and specificity of more than 98%, F1 score > 97%, and sensitivity and precision > 74%. This result is higher than (Castelli et al., 2020), who predicted air quality in California with 94.1% accuracy with the Radial Basis Function (RBF) method. This result is better than (Lin et al., 2018), who managed to achieve an accuracy of 72% in predicting air quality in Wuhan, China using Cloud Model Granulation. It is hoped that the relevance of the importance and success in predicting air quality in Jakarta will help reduce air pollution's critical impact on citizens and the environment. The results of this study indicate that predictions of air quality using other methods can get better results than those obtained by previous researchers and those carried out in this paper.

CONCLUSION

Air quality prediction has many benefits, especially for the surrounding community for better life management. The results of this research show that, in general, the performance of the proposed method is quite good, especially the DT-MNH method. However, the results show that three of the five measures have a value of > 97%, and two of the three performance measures have a value of > 74%.

REFERENCES

Agustine, I., Yulinawati, H., Gunawan, D., & Suswantoro, E. (2018). Potential impact of particulate matter less than 10 micron (PM10) to ambient air quality of Jakarta and Palembang. *IOP Conference Series: Earth and Environmental Science*, *106*(1). https://doi.org/10.1088/1755-1315/106/1/012057

Castelli, M., Clemente, F. M., Popovič, A., Silva, S., & Vanneschi, L. (2020). A Machine Learning Approach to Predict Air Quality in California. *Complexity*, *2020*(Ml). https://doi.org/10.1155/2020/8049504

Chen, H., & Fu, D. (2018). An Improved Naive Bayes Classifier for Large Scale Text. *Advances in Intelligent Systems Research*, *146*(Icaita), 33–36. https://doi.org/10.2991/icaita-18.2018.9

Chou, T., & Lo, M. (2018). Predicting Credit Card Defaults with Deep Learning and Other Machine Learning Models. *International Journal of Computer Theory and Engineering*, *10*(4), 105–110. https://doi.org/10.7763/ijcte.2018.v10.1208

Dinesh, S., & Dash, T. (2016). *Reliable Evaluation of Neural Network for Multiclass Classification of Real-world Data*. *1*. http://arxiv.org/abs/1612.00671

García, S., Luengo, J., & Herrera, F. (2015). Data Preprocessing in Data Mining. In J. Kacprzyk & L. C. Jain (Eds.), *Intelligent Systems Reference Library* (72nd ed., Vol. 72). Springer Cham Heidelberg New York Dordrecht London. https://doi.org/10.1007/978-3-319-10247-4

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: concepts and techniques*. Morgan Kaufmann Publishers. https://doi.org/https://doi.org/10.1016/C2009-0-61819-5

Hussein, A. M., Gheni, H. Q., Oleiwi, W. K., & Hasan, Z. Y. (2020). Prediction of credit card payment next month through tree net data mining techniques. *International Journal of Computing*, *19*(1), 97–105. https://doi.org/10.47839/ijc.19.1.1698

James, G., Daniela, W., Trevor, H., & Robert, T. (2013). An Introduction to Statistical Learning with Applications in R. In G. Casella;, S. Fienberg, & I. Olkin (Eds.), *Springer Texts in Statistics* (1st ed., Vol. 1). Springer New York Heidelberg Dordrecht London. https://doi.org/10.1007/978-1-4614-7138-7

Kresnawat, E. S. i, Resti, Y., Suprihatin, B., Kurniawan, M. R., & Amanda, W. A. (2021). Coronary Artery Disease Prediction Using Decision Trees and Multinomial Naïve Bayes with k-Fold Cross Validation. *Inomatika*, *3*(2), 174–189. https://doi.org/10.35438/inomatika.v3i2.266

Li, X., Chen, X., Yuan, X., Zeng, G., León, T., Liang, J., Chen, G., & Yuan, X. (2017). Characteristics of particulate pollution (PM2.5 and PM10) and their spacescale-dependent relationships with meteorological elements in China. *Sustainability (Switzerland)*, *9*(12), 1–14. https://doi.org/10.3390/su9122330

Lin, Y., Zhao, L., Li, H., & Sun, Y. (2018). Air quality forecasting based on cloud model granulation. *Eurasip Journal on Wireless Communications and Networking*, *2018*(1). https://doi.org/10.1186/s13638-018-1116-3

Mishra, S., Vanli, O. A., Huffer, F. W., & Jung, S. (2016). Regularized discriminant analysis for multi-sensor decision fusion and damage detection with Lamb waves. *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2016*, *9803*(850), 98032H. https://doi.org/10.1117/12.2217959

Pan, Y., Gao, H., Lin, H., Liu, Z., Tang, L., & Li, S. (2018). Identification of bacteriophage virion proteins using multinomial Naïve bayes with g-gap feature tree. *International Journal of Molecular Sciences*, *19*(6). https://doi.org/10.3390/ijms19061779

Pranita, E. (2020, August 12). Terkenal Buruk, Begini Kualitas Udara Jakarta Selama Pandemi Covid-19. *Kompas*, https://www.kompas.com/sains/read/2020/08/12/100200623/terkenal-buruk-begini-kualitas-udara-jakarta-selama-pandemi-covid-19?page=all

Resti, Y., Kresnawati, E. S., Dewi, N. R., Zayanti, D. A., & Eliyati, N. (2021). Diagnosis of diabetes mellitus in women of reproductive age using the prediction methods of naive bayes, discriminant analysis, and logistic regression. *Science and Technology Indonesia*, *6*(2), 96–104. https://doi.org/10.26554/STI.2021.6.2.96-104

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Witten, I. H., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. In *Complementary literature None* (2nd ed.). Morgan Kaufmann Publishers. https://www.elsevier.com/books/data-mining/witten/978-0-12-088407-0

06_JURNAL_S_2022_Implementation of a Breakpoint Halfway Discretization to Predict Jakarta's Air Quality

ORIGINALITY REPORT

12%
SIMILARITY INDEX

PRIMARY SOURCES

1	I Agustine, H Yulinawati, D Gunawan, E Suswantoro. " Potential impact of particulate matter less than 10 micron (PM) to ambient air quality of Jakarta and Palembang ", IOP Conference Series: Earth and Environmental Science, 2018 Crossref	55 words — 2%
2	sciencetechindonesia.com Internet	50 words — 1%
3	eprints.unhasy.ac.id Internet	37 words — 1%
4	Sugandi Yahdin, Anita Desiani, Shania Putri Andhini, Dian Cahyawati, Rifkie Primartha, Muhammad Arhami, Ditia Fitri Arinda. "COMBINATION OF KNN AND PARTICLE SWARM OPTIMIZATION (PSO) ON AIR QUALITY PREDICTION", BAREKENG: Jurnal Ilmu Matematika dan Terapan, 2022 Crossref	26 words — 1%
5	eprints.usm.my Internet	26 words — 1%
6	A Ridwan, T N Sari. "The comparison of accuracy between naïve bayes clasifier and c4.5 algorithm in classifying toddler nutrition status based on anthropometry index", Journal of Physics: Conference Series, 2021	25 words — 1%

Crossref

7	www.scilit.net Internet	18 words — 1%
8	jwoodscience.springeropen.com Internet	16 words — < 1%
9	www.lincoln.ne.gov Internet	16 words — < 1%
10	Robert Chinery, Randi Walker. "Development of Exposure Characterization Regions for Priority Ambient Air Pollutants", Human and Ecological Risk Assessment: An International Journal, 2010 Crossref	14 words — < 1%
11	storage.googleapis.com Internet	14 words — < 1%
12	www.answers.com Internet	14 words — < 1%
13	www.mdpi.com Internet	14 words — < 1%
14	ejournal.upsi.edu.my Internet	13 words — < 1%
15	V. Carollo, J. Reinoso, M. Paggi. "A 3D finite strain model for intralayer and interlayer crack simulation coupling the phase field approach and cohesive zone model", Composite Structures, 2017 Crossref	12 words — < 1%
16	www.tandfonline.com Internet	11 words — < 1%

17	essay.utwente.nl Internet	9 words — < 1%
18	www.msn.com Internet	9 words — < 1%
19	www.spiedigitallibrary.org Internet	9 words — < 1%
20	"Proceedings of the 11th International Conference on Robotics, Vision, Signal Processing and Power Applications", Springer Science and Business Media LLC, 2022 Crossref	8 words — < 1%
21	Yang, J.N.. "Sequential non-linear least-square estimation for damage identification of structures", International Journal of Non-Linear Mechanics, 200601 Crossref	8 words — < 1%
22	aaqr.org Internet	8 words — < 1%
23	ieomsociety.org Internet	8 words — < 1%
24	openaccess.marmara.edu.tr Internet	8 words — < 1%
25	repositori.usu.ac.id Internet	8 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	OFF