PREDICTION OF AIR QUALITY INDEX USING DECISION TREE WITH DISCRETIZATION

NING ELIYATI[1], MAUIZZATIL RAHMAYANI[1], SHOHIF WIJAYA[1], DES A. ZAYANTI[1], ENDANG S. KRESNAWATI[1], YULIA RESTI[1*]

[1]*Departement of Mathematics, Faculty of Mathematics and Natural Science, Universitas Sriwijaya, Inderalaya, Indonesia*

**Corresponding author: yulia_resti@mipa.unsri.ac.id*

(Received: 28 July 2022; Accepted: 11 October 2022; Published online: 01 November 2022)

ABSTRACT: Air quality is indicated by the Air Quality Index (AQI). Prediction or classification of AQI is an important research issue because it can impact many factors, such as the environment, health, transportation, agriculture, plantations, tourism, and education. The purpose of this study is to predict AQI using a decision tree. The results of calculating the performance of the decision tree method that implements the discretization technique show that this method is very good at predicting air quality, as indicated in particular by the Average Accuracy value of 99.05%, Macro Precision of 78.59%, and Macro Recall of 77.46%.

KEYWORDS: Air Quality Index, Decision Tree, and Repeated k-Fold Cross Validation.

1. INTRODUCTION

Air is a colorless mixture of gases found on the earth's surface. Air is invisible to the eye, has no smell, and has no taste [1]. Air is one type of natural resource because it has many functions for living things. Air contains oxygen (O2) for breathing, carbon dioxide (CO2) for photosynthesis, and ozone (O3) to block ultraviolet rays from the sun [2]. Air quality is indicated by the Air Quality Index (AQI). AQI can fluctuate rapidly due to weather factors. The increase in AQI occurred due to industrial growth, increasing population, and the number of motorized vehicles that were not comparable with reforestation and preservation of trees. The better the AQI, the lower the adverse effects on the health of living things, and vice versa [3].

The prediction or classification of AQI is an important research issue because it can impact many factors, such as the environment, health, transportation, agriculture, plantations, tourism, and education [4, 5]. The level of performance of prediction methods can depend on many factors, and the preprocessing of raw data is one factor that often affects classification performance [6, 7]. Some classification methods require the same type of predictor variables, while others can implement variables of different types, both numerical and categorical. Several techniques that can be implemented in the preprocessing process so that data has the same type are transformation, normalization, and discretization [8].

The decision tree is a classification method that determines the gain information for each predictor variable to form a tree structure [9]. IF-THEN logic is implemented in each decision branching [10]. If the initial processing of raw data is done correctly, this method often results in high classification performance [7, 11-13]. The results of air quality prediction (classification) are helpful for the government and the community to take policies or actions to

reduce/avoid the impact of poor air quality. This study aims to predict air quality using a decision tree by applying discretization techniques in the initial processing of raw data.

2. RESEARCH METHOD

2.1. Data

The data used in this research is secondary data from kaggle.com. The data consists of 2502 observations with 18 predictor variables, all numerical in type and target variables representing air quality in five categories: hazardous, very unhealthy, unhealthy, unhealthy for sensitive groups, and moderate). These variables are weather factors that affect air quality in Shanghai, China, in 2014-2021. The weather factors are maximum temperature, minimum temperature, snow intensity, sunlight intensity, ultraviolet index, moon illumination, dew point combination (a combination of temperature, relative humidity, and dew point), cold wind intensity, strong wind speed, heat index, wind intensity, cloud cover, air humidity, rainfall, air pressure, average temperature, visibility, degrees of wind direction, and wind speed.

2.2. Methodology

The steps taken in this study are:

1. Describe and visualize research data to find out the patterns and characteristics of the data
2. Partition the data into training and test data where observation data in 2014-2019 is implemented as training data and observation data in 2020-2021 as test data.
3. Building a computational system that predicts/classifies air quality using a tree-based method. The main steps of building predictions with a decision tree are selecting a variable as the root node, determining a branch for each value, dividing the unit into classes, and fourthly repeating the process for each branch until all cases on each branch have the same class. The basis for choosing a variable as the root node is all variables' highest information gain value. The entropy value is first determined to obtain the highest information gain value. Entropy is a parameter that measures the variance of the sample data so that it becomes a determinant in determining the variables that influence decision-making. The Entropy and Information Gain are determined by [7, 13],

$$Entropy\ (S) = \sum_{j=1}^{k_s} -P_j\ log_2 P_j \quad (1)$$

$$Entropy\ (S_c) = \sum_{j=1}^{k_s} -P_c\ log_2 P_c \quad (2)$$

$$Information\ Gain(S, X) = entropy\ (S) - \sum_{c=1}^{k_X} \frac{|S_c|}{|S|} Entropy\ (S_c) \quad (3)$$

For S and S_c be the total sample and the total sample in the c-th category of the predictor variable X. P_j, and P_c be the prior probability in the j-th air quality and the prior probability in the c-th category of the predictor variable X.

4. Determine the performance of air quality prediction using average accuracy, macro precision, and macro recall [14, 15].

$$average\ accuracy\ = \frac{\sum_{j=1}^{J} \frac{TP_j+TN_j}{TP_j+FP_j+TN_j+FN_j}}{J} \quad (4)$$

$$macro\ precision\ = \frac{\sum_{j=1}^{J} \frac{TP_j}{TP_j+FP_j}}{J} \quad (5)$$

$$macro\ recall\ = \frac{\sum_{j=1}^{J} \frac{TP_j}{TP_j+FN_j}}{J} \quad (6)$$

For TP_j, TN_j, FP_j, and FN_j each one represents true positive, true, negative, false positive, dan false negative pada confusion matrix [16, 17].

3. RESULT AND DISCUSSION

A summary of air quality data in this study is presented in Table 1.

Table 1: Summary of research data

Definition	Scale	Data
Maximun Temperature (X_1)	Interval	(-3)-40˚C
Minimum Temperature (X_2)	Interval	(-6)-3˚C
Total Snow (X_3)	Interval	0-1,7mm
Sun Hour (X_4)	Interval	3,8-14,5h
UV Index (X_5)	Interval	1-9nm%
Moon Ilumination (X_6)	Interval	0-100 ˚C
Dew Point (X_7)	Interval	(-23)-28 ˚C
Feels Like (X_8)	Interval	(-9)-45 ˚C
Heat Index (X_9)	Interval	(-3)-45 ˚C
Wind Chill (X_{10})	Interval	(-9)-36 ˚C
Wind Gust (X_{11})	Interval	4-82km/h
Cloud Cover (X_{12})	Interval	0-100okta
Humidity (X_{13})	Interval	18-97%
Precipitation (X_{14})	Interval	0-127mm
Pressure (X_{15})	Interval	986-1039mb
Temperature (X_{16})	Interval	(-3)-40 ˚C
Visibillity (X_{17})	Interval	3-20m
Winddir Degree (X_{18})	Interval	8-347˚
Wind Speed (X_{19})	Interval	3-51km/h
Air Quality Index (Y)	Ordinal	Moderate Unhealthy for sensitive groups Unhealthy Very Unhealthy Hazardous

In this study, interval scale data were discretized using common references in grouping weather variables, as presented in Table 2

Table 2: Variable discretization

Variable	Discretization	Interval	
X_1	1	freezing	<-0,1˚C
	2	Cold	0˚C-20,4˚C
	3	Cool	20,5˚C-23,9˚C
	4	Warm	24˚C-29,9˚C
	5	Hot	30˚C- 37,9˚C
	6	Very hot	>38˚C
X_2	1	freezing	<-0,1˚C
	2	Cold	0˚C-20,4˚C
	3	Cool	20,5˚C-23,9˚C
	4	Warm	24˚C-29,9C
	5	Hot	30˚C- 37,9˚C
	6	Very hot	>38˚C
⋮			
X_{10}	1	Very low	<-0,1
	2	Low	0-0,33
	3	Normal	0,34-0,67
	4	high	0,68-1,01
	5	Very high	>1,02
⋮			
X_{18}	1	North	0-23
	2	Northeast	24-68
	3	East	69-113
	4	Southeast	114-158
	5	South	159-203
	6	Southwest	204-248
	7	West	249-293
	8	Northwest	294-336
	9	North	>337

Tables 3 and 4 each present the results of entropy and information gain calculations for each category on each variable that affects AQI.

Table 3: Initial entropy

Independent Variable	Initial Entropy	Independent Variable	Initial Entropy
[X_1=1]	0	[X_6=1]	1,9160
[X_1=2]	1,7054	[X_6=2]	1,9807
[X_1=3]	2,0095	[X_6=3]	1,9733
[X_1=4]	1,9967	[X_6=4]	1,9274
[X_1=5]	1,9692	[X_6=5]	1,8469
[X_1=6]	1,3516	[X_7=1]	1,3723
[X_2=1]	1,2950	[X_7=2]	1,7336
[X_2=2]	1,8215	[X_7=3]	1,6594
[X_2=3]	1,9980	[X_7=4]	1,8597
[X_3=4]	1,8323	[X_7=5]	2,0140
[X_2=5]	1,5000	[X_7=6]	1,9982
[X_2=6]	0	[X_{78}=7]	1,9957
[X_3=1]	1,9411	[X_7=8]	1,8808
[X_3=2]	0	[X_8=1]	1,8033
[X_3=3]	1,0000	[X_8=2]	1,6513
[X_3=4]	0	[X_8=3]	1,8778
[X_3=5]	0	[X_8=4]	1,9866
[X_4=1]	1,9531	[X_9=5]	1,9381
[X_4=2]	2,0267	[X_9=1]	1,8606
[X_4=3]	1,8458	[X_9=2]	1,9902

Independent Variable	Initial Entropy	Independent Variable	Initial Entropy
[X_4=4]	1,9762	[X_9=3]	1,8227
[X_4=5]	1,9306	[X_{10}=4]	1,8319
[X_5=1]	1,7937	[X_{10}=5]	0
[X_5=2]	1,8785	[X_9=1]	1,9347
[X_5=3]	1,9834	[X_9=2]	1,5546
[X_5=4]	1,9654	⋮	⋮
[X_5=5]	0	[X_{19}=12]	0

Table 4 shows that X_8 (dew point) has the highest information gain, 0.1233, so X_8 is the root node. The next step is to create a branch of this variable with 8 branches representing each category: very dry, comfortable drier air, very comfortable, comfortable, slightly humid, moderately humid, very humid, and extremely humid. X19 (wind speed) has the highest information gain for the very dry category, so this variable becomes an internal node.

Table 4: Initial information gain

Variable	Initial Information Gain	Variable	Initial Information Gain
(Y,X_1)	0,0773	(Y,X_{10})	0,0025
(Y,X_2)	0,0999	(Y,X_{11})	0,0344
(Y,X_3)	0,0035	(Y,X_{12})	0,0501
(Y,X_4)	0,0183	(Y,X_{13})	0,0351
(Y,X_5)	0,0383	(Y,X_{14})	0,0192
(Y,X_6)	0,0139	(Y,X_{15})	0,0141
(Y,X_7)	0,1233	(Y,X_{16})	0,0773
(Y,X_8)	0,0890	(Y,X_{17})	0,0028
(Y,X_9)	0,0679	(Y,X_{18})	0,0649
		(Y,X_{19})	0,0270

Furthermore, X_{18} branching produces X_{10} (wind gust) as an internal node. Calculations are performed on all branches similarly; all nodes end in the AQI prediction or classification. Figure 1 shows the branching for the internal node $X_{18,}$ which is followed by the internal node X_{10} and ends in AQI predictions for the first class (moderate), second (unhealthy for sensitive groups), and fifth (hazardous). On the internal node X_{10}, there are only three AQI classes; for other AQI classes, there are other internal nodes. The results of AQI predictions using a decision tree are presented in Table 5.

Table 5: AQI prediction results using decision trees

	AQI Prediction				
Actual	Moderate	Unhealthy for sensitive groups	Unhealthy	Very Unhealthy	Hazardous
Moderate	0	0	0	0	0
Unhealthy for sensitive groups	0	82	1	1	2
Unhealthy	0	0	120	3	2
Very Unhealthy	0	0	0	207	8
Hazardous	0	0	1	0	330

The results of calculating the performance of the decision tree method that implements the discretization technique show that this method is very good at predicting AQI, which is indicated in particular by the average accuracy value of 99.05%, macro precision of 78.59%, and macro recall of 77.46%.

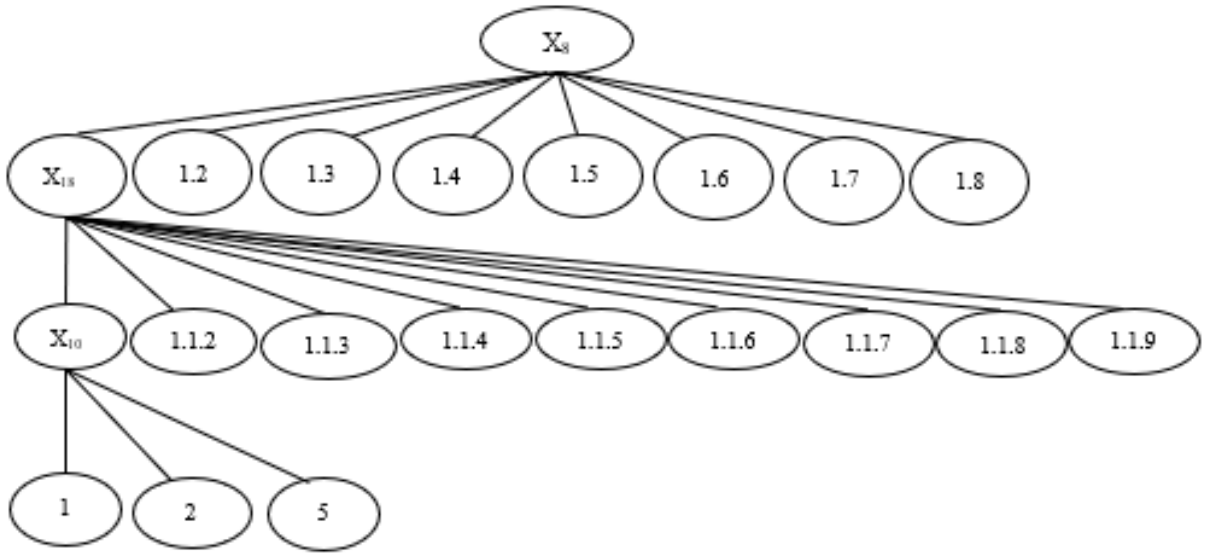

Fig. 1. Branching on variable wind speed and wind gust

4. CONCLUSION

The prediction or classification of the Air Quality Index is an important research issue because it significantly impacts many factors in human life. This study implements the decision tree method to predict or classify air quality. Each predictor variable is discretized using common references in grouping weather variables. Decision-making at each node is determined based on the highest information gain in all categories on the predictor variable to form a tree structure. Air quality prediction using the decision tree method in this study shows very good performance at average accuracy. This also indicates that the proposed discretization technique is appropriate.

ACKNOWLEDGEMENT

This publication of this article was funded by DIPA od Public Service Agency of Universitas Sriwijaya 2022. SP DIPA-023.17.2.677515/2022, On December 13, 2021. In according with the Rector's Decree Number: 0019/UN9/SK.LP2M.PT/2022, on June 15, 2022

REFERENCES

[1] A. Mayssara A. Abo Hassanin Supervised, "Udara," Pap. Knowl. . Towar. a Media Hist. Doc., pp. 6–35, 2014.
[2] H. W. Isramadhanti, “Gambaran Kualitas Udara di Kota Yogyakarta Berdasarkan Pemantauan Air Quality Monitoring System tahun 2019-2020,” Skripsi. Politek. Kesehat. Kemenkes Yogyakarta, pp. 30–48, 2019.
[3] J. Prayudha, A. Pranata, and A. Al Hafiz, “Implementasi Metode Fuzzy Logic Untuk Sistem Pengukuran Kualitas Udara Di Kota Medan Berbasis Internet of Things (Iot),” Jurteksi, vol. 4, no. 2, pp. 141–148, 2018, doi: 10.33330/jurteksi.v4i2.57.
[4] Liu, H., Li, Q., Yu, D., Gu, Y., 2019. Air quality index and air pollutant concentration prediction based on machine learning algorithms, Applied Scince, 9, 4069, doi: 10.3390/app9194069.
[5] Jia, M., Cheng, X., Zhao, T., Yin, C., Zhang, X., Wu., X, Wang, L., Zhang, R. (2019), Regional air quality forecast using a machine learning method and the WRF model over the Yangtze River Delta, East China. Aerosol and Air Quality Research, 19, pp. 1602-1613
[6] Podviezko, A., Podvezko, V., 2015. Influence of dta transformation on multicriteria evaluation,

Procedia Engineering, 122, pp.151-157.
[7] Kresnawati, E. S., Resti, Y., Suprihatin, B., Kurniawan, M. R., & Amanda, W. A. (2021a). 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
[8] 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
[9] 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
[10] 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
[11] Chandra, W., Resti, Y., Suprihatin, B., 2022. Implementation of a Breakpoint Halfway Discretization to Predict Jakarta's Air Quality, Inomatika.
[12] Resti, Y., Irsan, C., Putri, M. T., Yani, I., Anshori, and Suprihatin, B. (2022a). Identification of Corn Plant Diseases and Pests Based on Digital Images using, Science and Technology Indonesia, 7(1), 29–35, https://doi.org/10.26554/sti.2022.7.1.29-35.
[13] Resti, Y., Irsan, C., Amini, M., Yani, I., Passarella, R., Zayanti, D. A. (2022b). Performance Improvement of Decision Tree Model using Fuzzy Membership Function for Classification of Corn Plant Diseases and Pests, Science and Technology Indonesia, 7(3), 284–290, https://doi.org/10.26554/sti.2022.7.3.284-290.
[14] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Inf. Process. Manag., vol. 45, no. 4, pp. 427–437, 2009.
[15] Dinesh, S., & Dash, T. (2016). Reliable Evaluation of Neural Network for Multiclass Classification of Real-world Data. 1. http://arxiv.org/abs/1612.00671
[16] Burger, S.V., Introduction to Machine Learning with R, 2018. Oreilly. United State of America. 978-1-491-97644-9 [LSI].
[17] 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