Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11469
e-ISSN : 2541-2019
p-ISSN : 2541-044X

K-NN Based Air Classification as Indicator of the Index of Air Quality in Palembang

Ahmad Sanmorino[1)*], Juhaini Alie[2)], Nining Ariati[3)], Sanza Vittria Wulanda[4)]
[1)2)3)4)]Universitas Indo Global Mandiri, Indonesia
[1)]sanmorino@uigm.ac.id, [2)]juhaini@uigm.ac.id, [3)]nining@uigm.ac.id, [4)]sanza@uigm.ac.id

Submitted : June 18, 2022 | **Accepted** : July 9, 2022 | **Published** : July 10, 2022

Abstract: Good air quality is something that is wanted by every human who lives in big cities. Clean air and no pollution is one of the proper environmental requirements. One of the most severe causes of air pollution is due to large-scale forest fires due to the long dry season or is carried out by irresponsible persons which they commonly refer to as land clearing in an easy and inexpensive way by utilizing the reason of the dry season. The purpose of this study is to classify air quality in Palembang using a data mining approach. Then use the results of the classification as an indicator of the level of air quality in the city of Palembang. The data mining approach that researchers use is the K-Nearest Neighbor algorithm. Based on the test results of K-NN calculations and measured using a confusion matrix produce an accuracy of 80 percent, 82.3 percent for precision, and 93.3 percent for recall. The measurement results show that the calculation using the K-NN algorithm can be used as an indicator in measuring air quality, of the 20 that have been trained and tested only 4 inaccurate data, this inaccuracy occurs because the source data has unbalanced classes such as unhealthy and very unhealthy healthy have 1 sample each. So it proves that the performance of classifiers using the K-NN algorithm relevant as an indicator of air quality levels in the city of Palembang.

Keywords: Air quality; classification; K-NN; Palembang; pollution

INTRODUCTION

Good air quality is something that is wanted by every human who lives in big cities. Quality air and not pollution is one of the proper environmental requirements. The creation of a proper environment is the mandate of the Constitution Article 28 paragraph 1: Everyone has the right to live in physical and spiritual prosperity, to live and to have a good and healthy environment and to have health services. However, the implementation of Article 28 paragraph 1 has not optimal.

One of the most severe causes of air pollution is due to large-scale forest fires due to the long dry season (Sannigrahi et al. 2022; Verma et al., 2022), or is carried out by irresponsible persons which they commonly refer to as land clearing in an easy and inexpensive way by utilizing the reason of the dry season (Kadir et al., 2022). This causes tremendous smoke pollution to the most dangerous levels. This happened on the island of Sumatra. Based on data from the National Disaster Management Agency (BNPB) until Monday, September 16, 2019, at 16:00, hotspots were found in Riau as many as 58, Jambi (62), South Sumatra (115), West Kalimantan (384), Central Kalimantan (513) and South Kalimantan (178).

As a result of the forest fires that occurred, many residents' activities were disrupted, Acute Canal Infection (ISPA) was widespread, children were unable to go to school, as usual, workers were postponed due to lack of visibility, and many flights were stopped and delayed due to the thickening of the smoke haze. In addition to pollutants contained in the air produced by human activity, it is also produced by natural processes such as volcanoes. according to the decision of the head of the Environmental Impact Management Agency number Kep-107/KABAPEDAL/11/1997 air pollutants containing various components of elemental compounds such as carbon monoxide (CO), nitrogen dioxide (NO2), sulfur dioxide (SO2), particulate matter (PM10), ozone (O3), and nitrogen monoxide (NO). If these compounds have high levels of content, they will be able to disrupt human health, especially in respiratory disorders that can cause death.

Peningkatan indeks kualitas udara telah dikerjakan oleh beberapa peneliti seperti Fung et al. (2022) melalui robust statistical approach, dan Jurado et al. (2022) menggunakan metode deep learning berbasis computational fluid dynamics. Existing air quality every day needs to be accurately measured and classified. Accurate

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11469
e-ISSN : 2541-2019
p-ISSN : 2541-044X

classification results will greatly assist the government in making policy. This policy aims to try to control pollution so that it is at the air quality standards that can benefit as well as possible for the survival of living things. At present information on the results of air quality classification has been presented as diverse as on the Public Data Display (PDD) board which is located in the intersection in Palembang while in the field of information technology the government has provided air quality data on the BMKG official website, but the air quality classification process must continue to be carried out by conducting research on air quality classification so that the classification process can be known more clearly for both the public and researchers other than that as a reference indicator of air quality levels for the government in air quality classification using a data mining approach (Toussaint, 2005).

LITERATURE REVIEW

The Air Quality Index is used as information material for the public about ambient air quality at specific locations and times (Sahraeia, Kuskapan, & Codur, 2022; Zhang et al., 2021). The Air Pollution Standards Index (APSI) is also used as a consideration for the central government and regional governments in implementing the management and control of air pollution that occurs. Table 1 shows the limit of the air pollution standard index:

Table 1. The Limit of Air Pollution Standard Index (APSI)

APSI			PM10 µg/m^3		SO_2 µg/m^3		CO µg/m^3		O_3 µg/m^3		NO_2 µg/m^3	
Categories	**Ib**	**Ia**	**Xb**	**Xa**	**Xb**	**Xa**	**Xb**	**Xa**	**Xb**	**Xa**	**Xb**	**Xa**
50	1	50	0	50	0	80	0	5	0	5	0	0
100	51	100	51	150	81	365	6	10	121	235	0	0
200	101	199	151	350	366	800	11	17	236	400	0	1130
300	200	299	351	420	801	1600	18	34	401	800	1131	2260
400	300	399	421	500	1601	2100	35	46	801	1000	2261	3000
500	400	500	501	600	2101	2620	47	57,5	1001	1200	3001	3750

APSI values are calculated for all measured parameters, so APSI values are obtained for each air quality parameter (Ho et al., 2018). The value taken as the final value of the APSI for the measurement of air quality that occurs is one of the highest APSI values from the calculation of all air quality parameters, see Table 2.

Table 2. Numbers and Categories of Air Pollution Standards Index (APSI)

Index	Categories
1 – 50	Good
51 – 100	Moderate
101 – 199	Not healthy
200 – 299	Very unhealthy
300 – more	Dangerous

METHOD

In this section, the author would explain the design of the research conducted (Fig. 1). The data collection phase is looking for basic materials, which are collecting data about Palembang City air quality taken from the Central Jakarta Meteorology, Climatology, and Geophysics Agency Air Laboratory. The process of collecting data is done through requests to obtain data via email. The literature study in this study was done by studying some literature, for example through internet media, journals, and books related to research which includes; the application of the k-nearest neighbor algorithm concept and case examples relating to air quality classification.

The stage of identification of this problem is the stage where the data collection has been selected. The dataset is selected because it is in accordance with the method to be used namely the data mining method to classify air quality based on the Air Pollution Standard Index after studying the existing literature. From the data collected, the selection is done by selecting and separating data based on specified criteria. The Disposed of unnecessary data on air quality characteristic data that will not affect the classification results. The data that has been through the selection and cleaning process is transformed into a form that can be applied to the system to be made. For classification, the data mining method that researchers use is the K-NN approach (Debnath, Sinha & Bhowmik, 2022; Shahbazi, Bagheri, & Gharehpetian, 2022).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11469
e-ISSN : 2541-2019
p-ISSN : 2541-044X

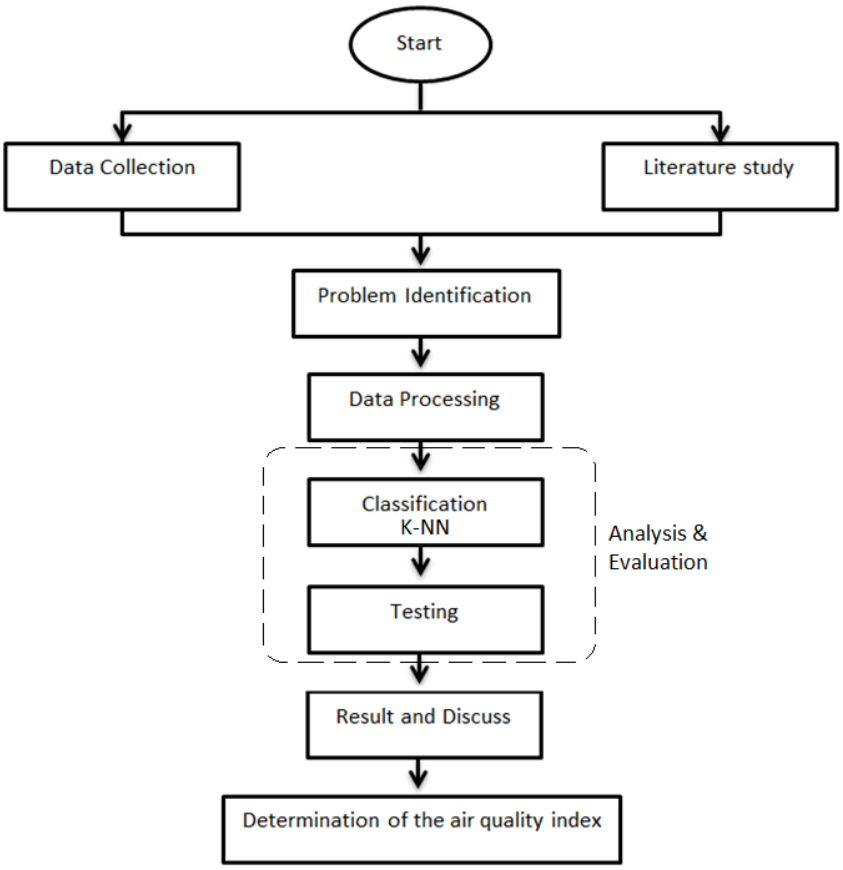


Fig. 1 Research Method

In the analysis phase, the input data is used in the air quality classification process. This stage explains that the existing input data will be divided into training data and testing data before the calculation process is carried out with the k-nearest neighbor model. At the application of the K-Nearest Neighbor, the analysis will be carried out on how to apply the K-NN algorithm to solve the problem in the form of dividing training data (samples) and testing data, then testing data is classified using the K-nearest neighbor (Guo et al., 2022; Nino-Adan et al., 2022; Urso et al., 2019). The initial step is to collect all air quality parameter data and then share the training data and testing data. The K-NN algorithm is very relevant to the type of data used in this study. The substances that makeup air have similarities to each other (neighborhood), so they can be classified using the K-NN algorithm.

The sample data and test data consist of No2, Co, Pm10, O3, So2, and air quality status. After the distance values in the sample data are known, clusters can be determined in the sample data based on the proximity of the values generated from these calculations. Then the distance data from all samples have been obtained in the third step and then the sorting results of the distance calculation (*di*) so that the obtained data with a sequence of distances from smallest to largest. After the distance calculation results have been sorted, we determine how much data the sample data wants to set (*k*) to find the composition of values based on sample data that has the closest distance to the test data to be assessed one by one as many variables are examined so that it gets the most air quality status matches the test data.

The validation method used is *K*-Fold cross-validation. Validation is a statistical method for evaluating and comparing learning algorithms by dividing data into two segments, one segment is used for learning or training data, and the other is used to validate the model (Refaeilzadeh, Tang, & Liu, 2009). In cross-validation, a collection of training and validation must be successive crossovers so that each data has a validated opportunity. K-fold cross-validation is a general technique for estimating classifier performance (Berrar, 2019). *K*-fold cross-validation is done by using the same dataset again, so that it results in *k* splitting of the data set into non-overlapping proportions of training (*k*-1)/*k* and 1/*k* for testing (Zhang & Liu, 2022; Wei & Chen, 2020).

To test the results of the k-nearest neighbor classification performed using a confusion matrix (Witten & Frank, 2011). This confusion matrix is a useful tool for analyzing how well the classifier we use can recognize patterns from different classes. Measurements in determining this classification use the value of accuracy and

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11469
e-ISSN : 2541-2019
p-ISSN : 2541-044X

error values. The accuracy value measures the success rate based on the closeness between the predicted value and the actual value. While the error value measures the average failure in making a classification between the predicted value and the actual value (Fig. 2).

		True Value	
		TRUE	FALSE
Value	TRUE	TP (*True Positive*) *Correct result*	FP (*False Positive*) *Unexpected result*
	FALSE	FN (*False Negative*) *Missing result*	TN (*True Negative*) *Correct absence of result*

Fig. 2 Model of Decision Making

RESULT

The validation technique used is 10-fold cross-validation, so the dataset is divided into 10 splits, then taken one part to be used as test data and the other used as training data. Intake is carried out in sequence starting from the first part until the tenth part (Table 3).

Table 3. Dataset for 10-Fold Cross-Validation Testing

Split	Categories	Co	So^2	No^2	Pm^{10}	O^3	APSI
1	Good	0	18.5	6.73	0	4.87	50
	Good	0	18.16	6.6	0	4.87	50
2	Good	0	17.25	6.75	0	5.24	50
	Good	0	16.16	7.12	0	2.95	50
3	Good	0	14.22	7.26	0	6.59	50
	Good	0	5.77	5.92	0	7.76	50
4	Moderate	5.14	10.74	6.43	12.67	6.41	51
	Good	4.82	11.59	6.15	12.68	7.76	48.2
5	Good	4.78	11.4	6.98	24.22	6.56	47.8
	Good	4.78	8.09	21.32	22.33	9.55	47.8
6	Good	4.96	7.73	22.04	29.69	9.28	49.6
	Good	4.84	4.66	20.22	25.61	11.08	48.4
7	Good	4.83	11.85	19.35	0.9	8.35	48.3
	Good	4.83	12.39	20.1	0	8.3	48.3
8	Moderate	1.13	4.62	16.93	73.58	107.52	86.79
	Good	0.59	20.69	0	30.94	56.86	30.94
9	Not healthy	16.11	23.58	15.68	68.22	92.74	187.28
	Dangerous	35.96	333.28	17.98	240.11	245.27	313.25
10	Very unhealthy	28.16	310.28	16.93	210.1	235.87	265.62
	Dangerous	35.96	315.09	16.98	212.59	248.23	313.25

The results of Euclidean distance calculations and the sequence of distances in Split 10, and Line 20 are shown in Table 4.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11469
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 4. The Euclidean distance calculations and the sequence of distances in Split 10, Line 20

No	Euclidean Distance	Sequence of Distances
1	428.9938694	11
2	429.2283016	12
3	429.645833	13
4	431.6128688	17
5	430.9866384	16
6	436.2532108	18
7	427.1004923	10
8	425.8056512	9
9	421.3489051	6
10	422.8339815	7
11	420.0171554	5
12	423.0247831	8
13	430.7730129	14
14	430.8685331	15
15	360.2118785	3
16	386.707791	4
17	351.010835	2
18	38.97515362	1

DISCUSSION

After performing the algorithm calculations, a manual calculation recapitulation is made for the K-Nearest Neighbor model in Table 5.

Table 5. The Recapitulation of Manual Calculations

Split	Row	Actual	Prediction
1	1	Good (G)	Good
	2	Good	Good
2	3	Good	Good
	4	Good	Good
3	5	Good	Good
	6	Good	Good
4	7	Moderate (M)	Good
	8	Good	Good
5	9	Good	Good
	10	Good	Good
6	11	Good	Good
	12	Good	Good
7	13	Good	Good
	14	Good	Good
8	15	Moderate	Not healthy
	16	Good	Good
9	17	Not healthy (NH)	Moderate
	18	Dangerous (D)	Dangerous
10	19	Very unhealthy (VU)	Dangerous
	20	Dangerous	Dangerous

Then the confusion matrix table is made and the model performance is calculated.

*name of corresponding author

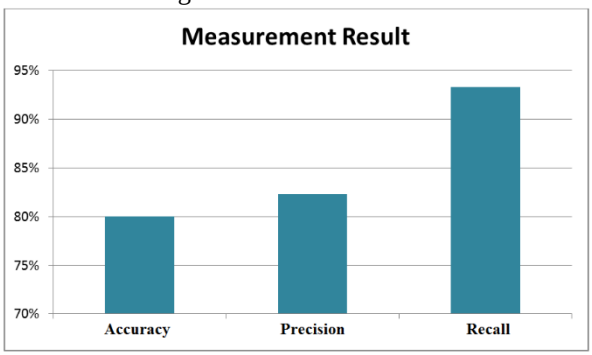
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 3, July 2022
DOI : https://doi.org/10.33395/sinkron.v7i3.11469
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Class		Prediction				
		G	M	NH	VU	D
Actual	G	14	0	0	0	0
	M	1	0	1	0	0
	NH	0	1	0	0	0
	VU	0	0	0	0	1
	D	0	0	0	0	2

Fig. 3 the Confusion Matrix

Based on Table 4 It can be seen that the sum of each of the confusion matrix values is True Positive (TP) for 14 data, True Negative for 2 data, False Positive (FP) for 3 data, and False Negative (FN) for 1 data. The known values are then calculated for accuracy, recall, and precision. Confusion matrix calculation results show that the accuracy of the K-NN calculation has a value of 80 percent; precision has a value of 82.3 percent, while the recall calculation results are 93.3 percent. To answer the formulation of the second problem, the classification results can be used as an indicator of the level of water quality in the city of Palembang. The results of these calculations can be illustrated in Fig. 4.

Fig. 4 Measurement result



CONCLUSION

K-NN calculations that have been carried out and measured using a confusion matrix produce an accuracy of 80 percent, a precision of 82.3 percent, and a recall of 93.3 percent. The measurement results show that the calculation using the K-NN algorithm can be used as an indicator in measuring air quality, of the 20 that have been trained and tested only 4 inaccurate data, this inaccuracy occurs because the source data has unbalanced classes such as unhealthy and very unhealthy healthy have 1 sample each. So it proves that the performance of classifiers using the K-NN algorithm relevant as an indicator of air quality levels in the city of Palembang.

ACKNOWLEDGMENT

The authors would like to thank you Universitas Indo Global Mandiri for supporting this study.

REFERENCES

Berrar, D. (2019). Cross-Validation. *Encyclopedia of Bioinformatics and Computational Biology*, vol. 1, pp. 542-545.

Debnath, S., Sinha, N., & Bhowmik, B. B. (2022). ML based modulation format identifier using K-NN algorithm. *Materialstoday: Proceedings*.

Fung, P. L. et al. (2022). Improving the current air quality index with new particulate indicators using a robust statistical approach. *Science of The Total Environment*.

Guo, J. et al. (2022). A representation coefficient-based k-nearest centroid neighbor classifier. *Expert Systems with Applications*, vol. 194.

Ho, A. F. W. et al. (2018). Health impacts of the Southeast Asian haze problem – A time-stratified case crossover study of the relationship between ambient air pollution and sudden cardiac deaths in Singapore. *International Journal of Cardiology*, vol. 271, pp. 352–358.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Jurado, X. et al. (2022). Deep learning methods evaluation to predict air quality based on Computational Fluid Dynamics. *Expert Systems with Applications*, vol. 203.

Kadir E. A. et al. (2022). Forest fire spreading and carbon concentration identification in tropical region Indonesia. *Alexandria Engineering Journal*, vol. 61, issue 2.

Nino-Adan, I. et al. (2022). Influence of statistical feature normalisation methods on K-Nearest Neighbours and K-Means in the context of industry 4.0. *Engineering Applications of Artificial Intelligence*, vol. 2022.

Sahraei, M. A., Kuskapan, E. & Codur, M. Y. (2021). Public transit usage and air quality index during the COVID-19 lockdown. *Journal of Environmental Management*, vol. 286.

Sannigrahi, S. et al. (2022). Examining the status of forest fire emission in 2020 and its connection to COVID-19 incidents in West Coast regions of the United States. *Environmental Research*, vol. 210.

Shahbazi, N., Bagheri, S., & Gharehpetian, G. B.(2022). Identification and classification of cross-country faults in transformers using K-NN and tree-based classifiers. *Electric Power Systems Research*, vol. 204.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. Encyclopedia of Database Systems, *Springer*, Berlin, pp. 532-538.

Urso, A. et al. (2019). Data Mining: Prediction Methods. *Encyclopedia of Bioinformatics and Computational Biology*, vol. 1.

Verma, S. et al. (2022). Characterizing aerosols during forest fires over Uttarakhand region in India using multi-satellite remote sensing data. *Advances in Space Research*, vol. 70, issue 4.

Wei, J. & Chen, H. (2020). Determining the number of factors in approximate factor models by twice K-fold cross validation. *Economics Letters*, vol. 191.

Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.). *Morgan Kaufmann*, Burlington.

Zhang, L. et al. (2021). Application of nonlinear land use regression models for ambient air pollutants and air quality index. *Atmospheric Pollution Research*, vol. 12, issue 10.

Zhang, X. & Liu, C. (2022). Model averaging prediction by K-fold cross-validation. *Journal of Econometrics*.

*name of corresponding author