Jurnal Mantik, 6 (1) (2022) 30-35

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik/index

Implementation of K-Nearest Neighbors (KNN) Algorithm in Classification of Data Water Quality

Adli Abdillah Nababan[1], Muhammad Khairi[2], Bayu Samudera Harahap[3]

[1]Bisnis Digital, STMIK Pelita Nusantara, Jl. Iskandar Muda No 1 Medan, 20154, Indonesia
[2,3]Teknik Informatika, STMIK Pelita Nusantara, Jl. Iskandar Muda No 1 Medan, 20154, Indonesia

E-mail: adliabdillahnababan@gmail.com

A R T I C L E I N F O

Article history:

Received: Mar 14, 2022
Revised: Apr 08, 2022
Accepted: May 06, 2022

Keywords:

Data Mining,
K-Nearest Neighbors (KNN),
Water Quality

A B S T R A C T

Data mining is a process of extracting useful information and patterns from a very large data set. Data mining is also a process of finding useful information that can be used as a supporting tool in decision making. Data that is processed using data mining is able to produce knowledge in accordance with the expectations of technological development. Many techniques can be used in data mining, one of which is data classification techniques. Classification is usually used to obtain patterns or models by going through the process of using existing algorithms. Like the K-Nearest Neighbors algorithm. K-Nearest Neighbors is a case-based reasoning methodology that is trained with a stored case, and can be accessed to perform new solutions. There is a lot of data that can be used in the implementation of classification, but in this study the data used is a collection of water data to determine the quality and quality.

1. Introduction

Water is a basic need in daily life that must be met, so it takes a proper management in determining the level of cleanliness of water. To test the level of cleanliness of the water required a clear laboratory test. This takes a very long time because the process of analyzing water quality and water feasibility is complicated so that the process of distributing water information to the community takes a long time. So we need an alternative to help solve the water quality problems.

The development of information technology is growing rapidly, making the user's need for information increasing. Information is the result of processing a set of data that gives meaning. Many problems arise regarding data processing, such as not all data stored in the database is useful and can be used in decision making, so the data must be processed so that it can be utilized optimally. One of the methods in processing the data is the data mining approach.

Data mining is an excavation of data in information processing with the aim of finding important patterns in the pile of data in the database so that it becomes a knowledge representation. One of the techniques used in data mining is data classification. Many methods can be used in classification, one of which is the K-Nearest Neighbors algorithm.

Yofi Firdan, et al (2018) conducted a study entitled "K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor" with the results of the KNN algorithm performance of 64%. Research conducted by Ashari et al (2013) compared the performance of Naïve Bayes, Decision Tree and K-Nearest Neighbor (KNN).

The results show that Naïve Bayes has the best accuracy in classification compared to Decision Tree and K-Nearest Neighbor (KNN) with an average accuracy of 73.7%, while the average accuracy of Decision Tree and KNN is 58.9% and 56.7, respectively. %. The specific purpose of this research is to apply the KNN method in classifying data to determine water quality.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

2. Method

In this study, all the initial research procedures must first have been carried out such as a library study which was carried out by collecting and reading and understanding references related to the problem. collect and read and understand references related to the development problem of the Human Nose Pattern recognition application. The research framework is as follows:. The workflow diagram of this research is illustrated in the following:

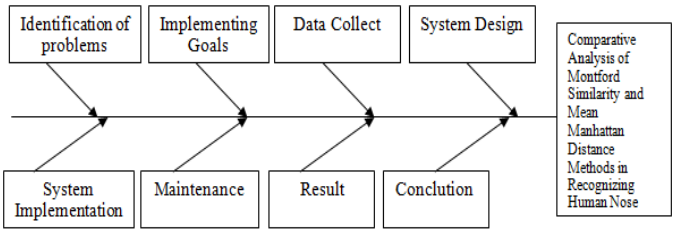

Figure 1. ResearchFlow Chart

In problem identification is done by looking for references from data mining by reading research related to the problem under study. This research is to carry out a classification technique on a set of water data to determine the quality of the water so that water providers can easily classify water quality according to categories, Good Condition, Lightly Polluted, Medium Polluted, Heavily Polluted. The previous research can be seen in the following table:

TABLE 1
PREVIOUS RESEARCH

No	Author/Year	Result
1	Maniya, dkk (2011)	compared K-Nearst Neighbor (KNN) and Naïve Bayes for the prediction of tuberculosis. The results of his research, classification accuracy with KNN reached 58%, while classification with Naïve Bayes reached 79.9%
2.	Mustafa, dkk (2012)	comparing K-Nearst Neighbor (KNN) with Artificial Neural Network (ANN) in the classification of spectrogram images in brain balancing. The results showed that KNN and ANN were able to classify spectrogram images with an accuracy of 87.5% to 90% for brain balancing applications.
3	Mahdi, dkk (2012)	comparing K-Nearst Neighbor (KNN), Naïve Bayes and Fuzzy in disease diagnosis. The results of the percentage of correct diagnosis using Fuzzy diagnosis, K-Nearst Neighbor (KNN), and Naïve Bayes were 91%, 87.9% and 72.4%, respectively.
4	Kim, dkk (2012)	comparing K-Nearst Neighbor (KNN) and Support Vector Machines (SVM) in image classification. The results of this study, the classification with KNN reached 78.03% while with SVM it reached 92%.
5	Tomar & Nagpal (2016)	comparing K-Nearst Neighbor (KNN) and Support Vector Machines (SVM) in age estimation classification. The results of the study were the classification with KNN reached 58.1% while the classification with SVM reached 61.33%.
6	Ni Luh Gede Pivin (2017)	The application of the "K-Nearest Neighbor Method For Car Selection Recommendation System" is able to help provide a shadow or reference to the user or prospective buyer in determining the selection of a car as needed.
7	Yahya & Winda (2020)	applying the K-Nearest Neighbor Algorithm for the classification of the Effectiveness of Vape Sales (Electric Cigarettes) on "Lombok Vape On" and getting the level of accuracy using the KNearest Neighbor algorithm with K-Fold Validation 6 is 86.48%

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Data Mining is the process of extracting data into information that has not previously been conveyed, with the right techniques the data mining process will provide optimal results. Some people argue that data mining is nothing more than machine learning or statistical analysis running on databases. In scientific journals, data mining is also known as Knowledge Discovery in Databases. then set the objectives of the problem to be studied, The next step is to collect data in the form of water data, The next stage is collecting data in the form of water data, designing the system construction and implementing the KNN algorithm in the system. **K**-Nearest Neighbor (KNN) is the simplest method used for classification.

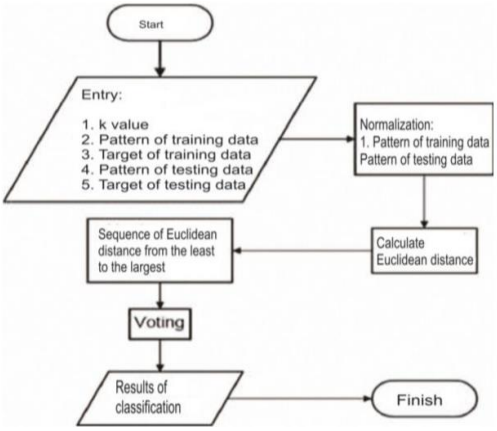

Figure 2. KNN flow chart

The steps in the classification of the K-Nearest Neighbor (KNN) are:

a. Input the calculated convolution image value
b. Specifies the parameter k (number of closest neighbors).
c. Calculating proximity based on the Euclidean distance model to the given training data, with the equation:

$$D(x,y) = ||x-y||_2\sqrt{\sum_{j=1}^{N}|x-y|^2} \quad \text{.....(1)}$$

d. Sort the distance results obtained in ascending order (sequentially from high to low value).
e. Count the number of each class based on the k nearest neighbors
f. The majority class is used as the class for the test data.

For the measurement of performance indicators used in this study is the Confusion matrix by conducting tests to estimate the right and wrong objects

TABLE 2
CONVOLUTION MATRIKS

Predcition Value	Actual Value	
	TP	TN
	FP	FN

Information:

TP = True Positive TN= True Negative
FB = False Positive FN = False Negative
The formula for calculating the confusion matrix is written as below:

a. Precision is useful for measuring the level of accuracy between the information requested by the user and the answer given by the system with the equation

Pre = TP / TP + FP .....(2)

b. Recall is useful for measuring the level of success of the system in retrieving an information, in the equation:

Re = TP / TP + FN .....(3)

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

c. Accuracy is useful for measuring the performance of a method, with the equation
Acu = TP + TN/ TP + TN+ FP+ FN .....(4)

3. Result and Discussion

The results discussed are in the form of performance and appearance of the system built. The appearance of In this study uses a water quality dataset derived from the research of Denades et al. (2016). The details of the data used can be seen in table :

TABLE 3
DATA OF WATER QUALITY

No	TSS (mg/L)	DO (mg/L)	COD (mg/L)	…	Quality Status
1	2	4	8	…	*Good Condition*
2	3	4.5	19.2	…	*Good Condition*
3	3	4.4	16	…	*Good Condition*
4	4	4.1	4.793	…	*Good Condition*
…	…	…	…	…	...
120	97	6.3	55.4	…	*Heavily Polluted*

In this section, we refer to the previous sub-chapters implemented in the Python programming language. The author presents the results and discussion of research regarding the application of the K-Nearest Neighbors (K-NN) method in water quality data classification. Measurement of method performance is based on the level of accuracy, precision, recall and the resulting f1-score. then do the test. The water quality data set has 8 attributes, 4 classes and 120 instances. Class distribution is in the form of good condition (30 instances), lightly polluted (30 instances), medium polluted (30 instances) and heavily polluted (30 instances).

TABLE 4
WATER QUALITY DATASET ATTRIBUTE INFORMATION

No.	Atribut	Nilai
1	TSS (mg/L)	[2-266]
2	DO (mg/L)	[0.02-8.43]
3	COD (mg/L)	[1.7-416]
4	BOD (mg/L)	[0.6-150]
5	Total phospat (mg/L)	[0.0016-1.23]
6	Fecal Coliform (mg/L)	[27-2800000]
7	Total Coliform (mg/L)	[74-53000000]
8	Pij	[0.54-15.31]
9	Quality Status	{good condition, lightly polluted, medium polluted, heavily polluted}

The syntax used to display the dataset used in Python can be seen in the picture bellow

```
In [3]: dataset.head()
Out[3]:
```

	TSS	DO	COD	BOD	Total_Phospat	Fecal_Coliform	Total_Coliform	Pij	Class
0	2.0	4.0	8.000	2.60	0.10	92	150	0.76	Good Condition
1	3.0	4.5	19.200	3.10	0.14	92	150	0.88	Good Condition
2	3.0	4.4	16.000	2.90	0.12	930	2400	0.91	Good Condition
3	4.0	4.1	4.793	1.32	0.18	1100	1400	0.87	Good Condition
4	4.0	4.2	8.000	2.50	0.11	230	750	0.81	Good Condition

Figure 3. Displaying Datasets in Python

Based on table 3, a comparison of method performance measurements will be made based on the level of accuracy, precision, recall and f1-score, to see more clearly the performance measurement of the

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

application of the K-Nearest Neighbors (K-NN) method in the classification of water quality data presented in the form of tables and figures. the following graph.

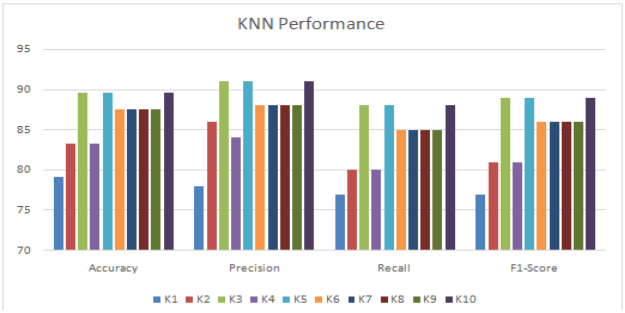

Figure 4. KNN Method Performance Measurement

TABLE 5
MEASURING THE PERFORMANCE OF THE KNN METHOD IN CLASSIFYING

K	Accuracy	Precision	Recall	F1-Score
1	79.16%	78.00%	77.00%	77.00%
2	83.33%	86.00%	80.00%	81.00%
3	89.58%	91.00%	88.00%	89.00%
4	83.33%	84.00%	80.00%	81.00%
5	89.58%	91.00%	88.00%	89.00%
6	87.5%	88.00%	85.00%	86.00%
7	87.5%	88.00%	85.00%	86.00%
8	87.5%	88.00%	85.00%	86.00%
9	87.5%	88.00%	85.00%	86.00%
10	83.33%	84.00%	80.00%	81.00%
Avg	85.83%	86.6%	83.3%	84.2%

From 4 it can be seen that the KNN method is able to provide a better accuracy value with an average accuracy value of 85.83%. Where the highest accuracy value of the KNN method is obtained when k is worth 3 and 5, which is 89.58%, while the lowest accuracy value is obtained when k is worth 1, which is 79.16%. The average precision value of all k in KNN is 86.6%. Meanwhile, the average recall value of all k in KNN is 83.3%. Meanwhile, the average f1-score of all k in KNN is 84.2%. Based on the tests that have been carried out from the water quality data set, it can be seen that KNN is able to provide better accuracy values in classifying water quality data.

4. Conclusion

The conclusions of this study are: Classification of data using the KNN method is able to provide a good accuracy value with an average accuracy value of 85.83%. Where the highest accuracy value of the KNN method is obtained when k is worth 3 and 5, which is 89.58%, while the lowest accuracy value is obtained when k is worth 1, which is 79.16%. Apart from being viewed in terms of accuracy, the performance measurement method is also very good based on the average precision value of all k in KNN, which is 86.6%. Meanwhile, the average recall value of all k in KNN is 83.3%. Meanwhile, the average f1-score of all k in KNN is 84.2%. Based on the tests that have been carried out from the water quality data set, it can be seen that KNN is able to provide better accuracy values in classifying water quality data.

5. References

[1] Firdaus, D. (2017). Penggunaan Data Mining dalam Kegiatan Sistem Pembelajaran Berbantuan Komputer. Jurnal Format, 93-97
[2] Mardi, Y. (2017) "Data Mining : Klasifikasi Menggunakan Algoritma C4.5" Jurnal Edik Informatik, 213-219

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

[3] Delima. E. S. (2018). Penerapan Data Mining Penjualan Sepatu Menggunakan Metode Algoritma Apriori. Jurnal Teknik. 156-161
[4] Chen, Y., Hao, Y. (2017) A Feature Weighted Support Vector Machine and K-Nearest Neighbor Algorithm for Stock Market Indices Prediction. Expert Systems With Applications 340-355.
[5] Danades, A., Pratama, D., Anggraini, D., Anggriani, D. (2016). Comparison of Accuracy Level K-Nearest Neighbor Algorithm and Support Vector Machine Algorithm in Classification Water Quality Status. International Conference on System Engineering and Technology, 137-141
[6] Gou, J. & Xiong, T. (2011) A Novel Weighted Voting for K-Nearest Neighbor Rule. Journal of Computer, 833-840.
[7] Kim, J., Kim, B., Savarese, S. (2012) Comparing Image Classification Methods: K- Nearest- Neighbor and Support-Vector-Machines. Proceedings of the 6th WSEAS International Conference on Computer Engineering and Applications, and Proceedings of the 2012 American Conference on Applied Mathematics. 133- 138.
[8] Kuhkan, M. (2016) A Method to Improve the Accuracy of K-Nearest Neighbor Algorithm. Internatonal Journal of Computer Engineering and Information Technology, 90- 95.
[9] Mahdi, A., Razali, A., Alwakil, A (2012) Comparison of Fuzzy Diagnosis with K-Nearest Neighbor and Naïve Bayes Classifiers in Disease Diagnosis. Broad Research in Artificial Intelligence and Neuroscience (BRAIN), 58-66.
[10] Maniya, H., Hasan, I.M., Patel, K.P. 2011. Comparative study of Naïve Bayes Classifier and KNN for Tuberculosis. International Conference on Web Services Computing (ICWSC) 2 22-26.
[11] Moosavian, A., Ahmadi, H., Tabatabaeefar, A., Khazaee, M. 2012. Comparison of two classifiers; K-nearest neighbor and artificial neural network, for fault diagnosis on a main engine journal-bearing. Shock and Vibration 20(2): 263-272.
[12] Mustafa, M., Taib, N.M., Murat, J.H. Sulaiman, N. (2012) Comparison between KNN and ANN Classification in Brain Balancing Application via Spectrogram Image. Journal of Computer Science & Computational Mathematics 2(4): 17-22.
[13] Luh.N.G. 2017. Penerapan Metode K-Nearest Neighbor Untuk Sistem Rekomendasi Pemilihan Mobil. Techno.COM1,6
[14] Yahya & Winda. P.H. 2020. Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Efektivitas Penjualan Vape (Rokok Elektrik) pada "Lombok Vape On". Infotek : Jurnal Informatika dan Teknologi.
[15] Firdan. Y.S, Riza.A, Much. A.M. 2018. K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor. Scientific Journal of Informatics