

## IMPLEMENTATION OF THE K-NEAREST NEIGHBOR METHOD IN KNOWING WATER QUALITY

<sup>1</sup>David Nico L Nainggolan, <sup>2</sup>Eko Hariyanto, <sup>3</sup>Arpan  
System Computer, Universitas Pembangunan Pancabudi Medan  
[davidnico19961996@gmail.com](mailto:davidnico19961996@gmail.com), [arsevent@pancabudi.ac.id](mailto:arsevent@pancabudi.ac.id)

### Abstract

**Article Info**  
Received, 01/06/2022  
Revised, 28/06/2022  
Accepted, 29/06/2022

The function of classification is the process carried out in predicting a data that has a class that is still unknown, the pattern that is owned is also already regular in a classification method. K-NN is a group that has an instances-based learning system, in conducting group searches by performing the value of k objects into the test with the closest value to the value of other data. KNN uses the closest distance value to the tested dataset in carrying out the classification process. Drinking water is very important for health and is a very effective component for the health of the human body. Health is very influential on the country's economy, it is necessary to invest in water that is very beneficial for the community. This study conducted a search for accuracy of water quality with data as many as 3276 different bodies of water in order to know which water can be drunk and not drinkable. The results of the accuracy of the KNN classification model that can increase the level of accuracy better for the data used. So the research on water quality has an accuracy of 56.40% with 370 data on drinkable water. Researchers hope that accuracy can be improved again by combining the optimization of the classification model in future studies.

Keywords: Classification, K-NN, Accuracy, Water.

### 1. INTRODUCTION

According to (Raviya, 2013) the function of classification is a process carried out in predicting a data that has a class that is still unknown, the pattern that is owned is also already regular in a classification method. The classification method also discusses other algorithms such as SVM, KNN, Nave Bayes, and so on (Sahu, 2011).

According to (Pan, 2016) KNN is very influential on non-parametric techniques in the form of classification but the level of performance depends on the point of equalization of variables which are generally correlated with points that are far away. The distance between the measured values of the given limits of the standard deviation values. The KNN method (Tan, 2006) K is used in each class that has a large influence on the K value. If k is smaller than the classification that is useful for the data, it is not met, if a larger K value can more easily cause outliers in k neighbors who are close to the class center (Gou, 2014). K-Nearest Neighbor (KNN) is a method that implements a lazy learning system that finds groups of k objects in the closest training data or similar to the test data objects.

In the study (Toto, 2017) K-Nearest Neighbor produced 159 data whose prediction results were almost the same by only having a difference between the first choice and the second choice and some data that had predicted similar things between the second choice and the first choice which caused the attribute with the number of that has been used will result in less accurate predictions.

### 2. LITERATURE REVIEW

K-NN is a group that has an instances-based learning system, in conducting group searches by performing the value of k objects into the test with the closest value to the value of other data. KNN uses the closest distance value to the tested dataset in carrying out its classification process. The approach is carried out in finding a problem in the calculation at the closest distance between the new problem and the previous one by weighting it by equating it with the number of existing features. (Dhany, 2021).

In the K-NN method there is a feature in which there are vectors and a data group from the sample data in the training data. These features have an unknown classification data process, because

the distance between the other vectors and the training vector is calculated based on the value of the closest k to be taken.

Below are the steps in the calculation of the KNN algorithm as follows:

- K determined by the parameter value of the closest number of values
- Perform calculations with Euclidean distances with an existing dataset.
- Process data sequences that have low distance groups
- Category N collected
- K-NN uses categories in predicting what is needed with the calculated value of query instances.

In the classification using KNN with test data and training data, calculations are carried out in carrying out test data which will then be displayed in the confusion table. matrix (Witten, 2015). The classification generated by the test data has different classes, namely positive and negative. The following is the confusion.matrix table as follows:

Table 1. Confusion Matrix

Two Class Prediction		Class Prediction	
		Yes	No
Actual Class	Yes	True Positif	False Negatif
	No	False Positif	True Negatif

(Sumber : Witteen, 2005)

In the table above, it is explained about the yes and no classes in the classification method with values True Positive (True Positive) and True Negatives (true negative). Predictions that are not the same or that have a positive value when carrying out the expected prediction process are False Positive and vice versa. The Confusion Matrix equation can be seen below: (Witteen, 2005).

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

In the data in which it has a positive class value whose classification prediction results are in accordance with the actual, it has a positive value called True Positive. In data that has a class that has a negative value whose classification prediction results are in accordance with the actual, it has a negative value called True Negative.

The data in which the class has a negative value but the prediction results have a positive value is called False Positive. In data that has a class that has a positive value but the prediction result has a negative value, it is called False Negative

### 3. METHODS

This study aims to analyze the performance of the K-NN method so that observational data is needed which can be obtained from (<https://www.kaggle.com/>) then the authors choose to analyze using the Kaggle dataset, namely water quality. The research process can be seen in the following chart:



Figure 1. Research Process Chart

### 4. Research Results and Discussion

Below is a description of the water quality dataset which is summarized in the following table:

No	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
1	0	204.8	20791.	7.300	368.5	564.309	10.37	86.9	2.96	0

			3							
2	3.716	129.4	18630. 1	6.635	0	592.885	15.18	56.3	4.50	0
3	8.099	224.2	19909. 5	9.275	0	418.606	16.86	66.4	3.05	0
4	8.316	214.3	22018. 4	8.059	356.8	363.267	18.43	100.3	4.62	0
5	9.092	181.1	17979	6.546	310.1	398.411	11.55	31.9	4.07	0
6	5.584	188.3	28748. 7	7.544	326.6	280.468	8.39	54.9	2.55	0
7	10.22	248.0	28749. 7	7.513	393.6	283.652	13.78	84.6	2.67	0
8	8.635	203.3	13672. 1	4.563	303.3	474.608	12.36	62.7	4.40	0
9	0	118.9	14285. 6	7.804	268.6	389.376	12.76	53.9	3.59	0
...	...	...	...	...	...	...	...	...	...	...
3276	7.87	195.1	17404. 2	7.509		327.46	16.14	78.6	2.30	1

Information:

1. pH value: acid-base balance of water.
2. Hardness: from calcium and magnesium salts.
3. Solids (Total dissolved solids TDS): Water with a high TDS value indicates that the water is highly mineralized.
4. Chloramine: the main disinfectant used in public water systems.
5. Sulfates: naturally occurring substances in minerals, soil, and rocks.
6. Conductivity: i.e. Pure water with a good insulator, not a good conductor of electricity.
7. Organic Carbon: in raw water comes from decaying natural organic matter (NOM) and synthetic sources.
8. Trihalomethanes: chemicals that can be found in water that contains chlorine.
9. Turbidity: Turbidity of water depends on the amount of solids present in the suspended state.
10. Potability: if the water is safe for human consumption, 1 means it can be drunk and 0 is not drinkable.

The results of the research that have been carried out have found that the processes that have been tested and carried out by the author, as well as the process of sharing training data and test data. So then carry out the classification process by looking for the accuracy value of the K-NN method.

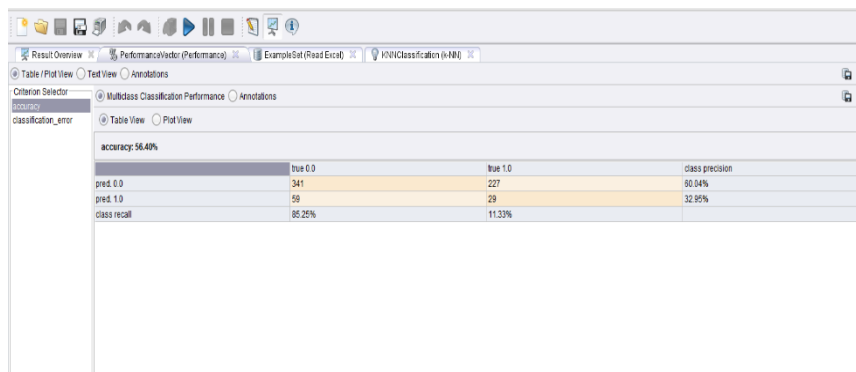
	Class 0	Class 1
pred. 0	341 (False Negative)	227 (True Positive)
pred. 1	59 (True Negative)	29 (False Positive)

The calculations in the confusion matrix are as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{341+29}{341+29+59+227} = \frac{370}{656} = 0.5640 * 100\% = 56.40\%$$

$$\text{Classification\_error} = \frac{FP+FN}{TP+TN+FP+FN} = \frac{59+227}{341+29+59+227} = \frac{286}{656} = 0.4359 * 100\% = 43.59\%$$

The results of the program can be seen as follows:



accuracy: 56.40%			
	true 0.0	true 1.0	class precision
pred 0.0	341	227	80.04%
pred 1.0	59	29	32.95%
class recall	85.25%	11.33%	

## 5. Conclusion

This study resulted in the accuracy of the KNN classification model which can increase the level of accuracy better for the data used. So the research on water quality has an accuracy of 56.40% with 370 data on drinkable water. Researchers hope that accuracy can be improved again by combining the optimization of the classification model in future studies.

## REFERENCE

- [1]. Aryza, S., Irwanto, M., Lubis, Z., Siahaan, A. P. U., Rahim, R., & Furqan, M. (2018). A Novelty Design Of Minimization Of Electrical Losses In A Vector Controlled Induction Machine Drive. In IOP Conference Series: Materials Science and Engineering (Vol. 300, No. 1, p. 012067). IOP Publishing.
- [2]. Gou, J., Yi, Z., Du, L., & Xiong, T. 2012. A Local Mean-Based k-Nearest Centroid Neighbor Classifier. The Computer Journal 55(6):pp.1058-1071.
- [3]. Nur Anisah. 2019. Analisis Algoritma Support Vector Machine Learning Dan K-Nearest Neighbor Dalam Akurasi Data. Universitas Sumatera Utara
- [4]. Pan, Z., Wang, Y. & Ku, W. 2017. A New General Nearest Neighbor Classification Based On The Mutual Neighborhood Information. Knowledge Based Systems 121: 142-152.
- [5]. Tan, P., Steinbach, M., & Kumar, V. 2006. Introduction to Data Mining. Boston: Pearson Education.