Contents lists available at www.journal.unipdu.ac.id
Register
Journal Page is available to www.journal.unipdu.ac.id/index.php/register

Research article

The Application of Modified K-Nearest Neighbor Algorithm for Classification of Groundwater Quality Based on Image Processing and pH, TDS, and Temperature Sensors

Hasna Shafa Amalia[a], Ummi Athiyah[b], Arif Wirawan Muhammad[c]*

[a,c] *Department of Informatics, Insitut Teknologi Telkom Purwokerto, Indonesia*
[b] *Department of Data Science, Insitut Teknologi Telkom Purwokerto, Indonesia*
[c] *Department of Information Security and Web Technology, Universiti Tun Hussein Onn Malaysia (UTHM), Malaysia*
email: [a] 18102160@ittelkom-pwt.ac.id, [b] ummi@ittelkom-pwt.ac.id, [c] arif@ittelkom-pwt.ac.id*
** Correspondence*

ARTICLE INFO

Article history:
Received 3 June 2022
Revised 21 October 2021
Accepted 20 December 2022
Available online 14 February 2023

Keywords:
classification
groundwater
Modified K-Nearest Neighbor
image processing

Please cite this article in IEEE style as:
H. S. Amalia, U. Athiyah, and A. W. Muhammad, "Application of Modified K-Nearest Neighbor Algorithm for Classification of Groundwater Quality Based on Image Processing and pH, TDS, and Temperature Sensors," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 9, no. 1, pp. 42-54, 2023.

ABSTRACT

The limited availability of water in remote areas makes rural communities pay less attention to the water quality they use. Water quality analysis is needed to determine the level of groundwater quality used using the Modified K-Nearest Neighbor Algorithm to minimize exposure to a disease. The data used in this study was images combined with sensor data obtained from pH (Potential of Hydrogen), TDS (Total Dissolved Solids) sensors and Temperature Sensors. The test used the Weight voting value as the highest class majority determination and was evaluated using the K-Fold Cross Validation and Multi Class Confusion Matrix algorithms, obtaining the highest accuracy value of 78% at K-Fold = 2, K-Fold = 9, and K- Fold = 10. Meanwhile, the results of testing the effect of the K value obtained the highest accuracy value at K = 5 of 67.90% with a precision value of 0.32, 0.37 recall, and 0.33 F1-Score. From the results of the tests carried out, it can be concluded that most of the water conditions are suitable for use.

1. Introduction

Water is one of the essential basic needs for living things, especially humans, for their daily needs and other activities. More than 50% of the human body consists of water, with a daily consumption of about 1.5 liters. The problem that often occurs today is the availability of water unavailable in remote areas, such as rural areas. The limited amount of water makes rural communities pay less attention to water quality, resulting in many cases of disease caused by the quality of the water they use [1]. Cibangkong Village, located in Pekuncen District, Banyumas Regency, is one of the remote villages where people still struggle to access clean water. Most people still use river water and well water as the primary water source for their daily needs. However, in its utilization, the public generally does not understand the water quality that is often used by referring to the requirements and standards set by the Ministry of Health, Regulation Number 492 of 2010, concerning Drinking Water Quality Requirements [2].

Along with the current technological advances, artificial intelligence in the form of machine learning can help people analyze, test and determine the quality of water used. Currently, to determine the quality of the water, we can use a pH (Potencial of Hydrogen) sensor, a TDS (Total Dissolved Solid) sensor, and a temperature sensor [3]. In addition, Machine Learning is also an algorithm that is developed so that computers can act or behave like humans based on databases and sensors [4]. K-

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 42-54

Nearest Neighbor (KNN) is a classification method in machine learning commonly used because it is simple and practical to implement [5]. The KNN method was developed into the Modified K-Nearest Neighbor (MKKN) method so that the classification results are more accurate, and it is expected that higher accuracy results compared to the KNN accuracy results can be obtained [6]. The K-Nearest Neighbor modification is also considered capable of classifying the results of Image Processing [7].

Based on these problems, it is necessary to analyze the groundwater quality in Cibangkong Village, Pekuncen District, Banyumas Regency, Central Java Province, in light of the predetermined drinking water quality requirements. The analysis was carried out using the Modified K-Nearest Neighbor (MKKN) algorithm method, with the parameters used in the form of water images processed using color feature extraction. Regarding the sensor parameters employed, they were pH, TDS and temperature sensors. Subsequently, the Modified K-Nearest Neighbor model was evaluated using the K-Fold Cross Validation algorithm. Therefore, the selection of the three sensors can be considered representative for determining water quality in accordance with the Ministry of Health Regulation Number 492 of 2010 concerning drinking water quality requirements.

2. Materials and Methods

The researchers conducted the current research in order to test water quality using the Modified K-Nearest Neighbor and evaluate the model using the K-Fold Cross Validation. There are several aspects that distinguish the current study from previous studies. These differences can be seen in terms of the problems raised, the types of methods used in the research, and the components used as supporting materials. The following are some previous studies related to the current research

The first research discusses the problem of river water quality in Riau Province. This study uses the Modified K-Nearest Neighbor with 14 parameters. The data was classified into three categories: good, slightly polluted, and moderately polluted. This test produces the highest accuracy value at k = 3 and k = 5 with a value of 97.44% using a confusion matrix. Meanwhile, system testing was carried out using a black box so that the system could run properly and an optimal output could be produced [8]. However, the current research only employed two different types of data from unbalanced categories, which may potentially have an impact on the accuracy value obtained.

The second study discusses the classification of coal at PT. Eternal Sun Radiation in Anggana District, Kutai Kartanegara Regency, East Kalimantan Province. The study implemented the Modified K-Nearest Neighbor with 3 categorization, namely Anthracite, Bituminous, and Sub-Bituminous. The dataset used was 37 pieces of data, and the data was dominated by Sub-Bituminous coal (67%), with the determination of optimal K using 10-Fold Cross-Validation. The test results obtained an accuracy percentage of 100% with a value of K = 3 [9]. The difference between this previous study and this current one is in the amount of data used affecting the accuracy results obtained. Furthermore, this study initially processed image data to determine the accuracy value.

In the third relevant study, the K-Nearest Neighbor was used in a case-based reasoning methodology that is trained with a stored case, which can be accessed to perform new solutions. There was various types of data that can be adopted in the implementation of classification, but in the current study the data used was a collection of water data to determine the water quality [10].

In the fourth study, K-Nearest Neighbor (k-NN) was employed to categorize documents based on study interests with information gain features selection to handle unbalanced data and cosine similarity to measure the distance between test and training data. Based on the results of tests conducted with 276 training data, the highest result was the use of the information gain selection feature, which used 80% training data and 20% test data, producing an accuracy of 87.5% with a parameter value of k = 5. The highest accuracy result of 92.9% was achieved without information gain feature selection, with the proportion of training data of 90% and 10% test data and parameter values of k = 5, 7, and 9 [11].

The fifth study discusses a comparison of water quality classification models by employing machine learning algorithms viz., SVM, Decision Tree, and Naïve Bayes. The features implemented for determining the water quality were: pH, DO, BOD, and electrical conductivity. The classification models were trained based on the weighted arithmetic water quality index (WAWQI) calculated. After assessing the results obtained, the decision tree algorithm was found to be a better classification model with an accuracy of 98.50%. [12]. Meanwhile, the current study has a balanced amount of data, especially in each class.

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 42-54

2.1. Research Method

The research was begun with the formulation of the problems, research significance, objectives, literature study, and data collection in the form of images and sensors. To collect image data, the processing was carried out, first, by converting to images to grayscale. Then the Contrast Limited Adaptive Histogram Equalization (CLAHE) process and first-order feature extraction were carried out. The results obtained from image processing were in the form of parameter means, standard deviations, skewness, and kurtosis. Meanwhile, the data obtained from the sensor was in the form of pH, TDS, and temperature. Then, these parameters were used as a dataset. The implementation process was carried out through several stages: data selection, data cleaning, transformation, and then classification using Modified K-Nearest Neighbor. At the classification stage, the use of Modified K-Nearest Neighbor involves several processes: Manhattan Distance, validity, and Weighted Voting. The last stage was testing the K-Fold Cross Validation and the effect of the K value using the Confusion Matrix. The research flow chart can be seen in figure 1 below.

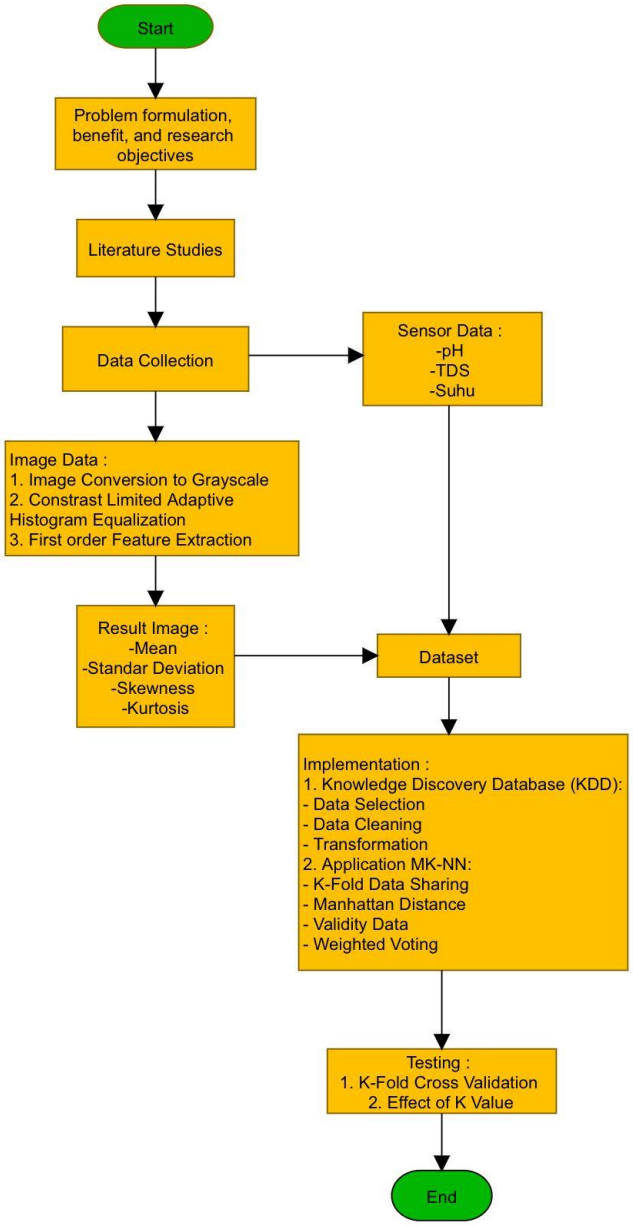

Fig. 1. Cognitive process dimension

2.2. Data Collection Techniques

The data in this study was collected through literature studies and direct data collection. In the literature study, the researchers collected data or references relevant to the research to be carried out. These

 http://doi.org/10.26594/register.v9i1.3045

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 42-54

references could be obtained from books, journals, and the internet. Meanwhile, the direct data collection was carried out in Cibangkong Village using image data and sensor data. An example of the image data taken can be seen in figure 2.

Fig. 2. An Example of Image Data

2.3. Image Processing

After the data was collected, it needs to be processed. The image data was processed through several stages as follows:

a. Image Conversion to Grayscale

Grayscale, often called a gray image, is a digital image processing technique with one channel for each pixel intensity. The obtained image is converted into grayscale form because the gray color has the same intensity in the RGB (Red, Green, Blue) [13].

b. Contrast Limited Adaptive Histogram Equalization or CLAHE is an image improvement method of Adaptive Histogram Equalization (AHE), which was used to reduce exposure in an image. CLAHE can also be used to increase the contrast in an image [14].

c. First-order feature extraction is a feature-taking method based on particular characteristics. The characteristics used in this study are the means, standard deviations, skewness, and kurtosis [15].

1) Mean

Mean is the median value or the average value of the distribution on the intensity of an image. The mean can be calculated using the following equations 1 and 2:

$$\mu = \sum_n f_n \, p(f_n) \quad (1)$$

$$p(f_n) = \frac{f}{\sum f} \quad (2)$$

where:

μ = is the average value
n = is the number of pixels
f = is the frequency value
f_n = is the gray intensity value
$p(f_n)$ = is the probability value of the occurrence of the value intensity on a histogram

2) Standard Deviation

Standard deviation is the distribution of the intensity values of a gray image. The standard deviation can be calculated using the following equation 3:

$$s = \sqrt{\sigma^2 = \sum_n (f_n - \mu)^2 \, p(f_n)} \quad (3)$$

Where:

s = is the standard deviation value

3) Skewness

Skewness is the value of the level of asymmetry on the histogram curve of an image. Skewness can be calculated using the following equation 4:

$$a_3 = \frac{1}{\sigma^3} \sum_n (f_n - \mu)^3 \, p(f_n) \quad (4)$$

4) Kurtosis

Kurtosis is the value of a curve in the histogram of an image. Kurtosis can be calculated using the following equation 5:

$$a_4 = \frac{1}{s} \sum_n (f_n - \mu)^4 \, p(f_n) \quad (5)$$

 http://doi.org/10.26594/register.v9i1.3045

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 42-54

Where:

a_4 = is the value of *kurtosis*

2.4. The Application of Knowledge Discovery Database (KDD)

Knowledge Discovery in Databases (KDD) is a structured process to analyze and obtain new, precise, helpful information and find patterns from complex data. Data Mining is a core part of the Knowledge Discovery in the Databases process to build a model, find a pattern, to explore data using specific algorithms [16]. The stages of KDD are:

a. Data Selection determines the database related to groundwater quality in Cibangkong Village. At the data selection stage, operational data was selected. The data selected is the attributes of the image processing, such as means, standard deviations, skewness, and kurtosis, and the attributes of the sensor are TDS values, pH, and temperature.
b. Pre-processing is the process where data cleaning is carried out. Unnecessary attributes can be removed at this stage.
c. Transformation stage is where the data is transformed for adjustments so that the data can be processed.

2.5. The Application of Modified K-Nearest Neighbor (MK-NN)

Modified K-Nearest Neighbor (MK-NN) is an algorithm developed from the K-Nearest Neighbor (K-NN) method. The mechanism of the MK-NN method is still the same as the K-NN method, that is, grouping new data through the *K* nearest neighbors. However, the MK-NN method has a number differences that distinguish it from the K-NN method. The MK-KN method involves calculation for the validity data training and calculation of *weight voting* for all test data using validity data [17]. The stages in the Modified K-Nearest Neighbor are:

a. Manhattan Distance

   Manhattan Distance is a method for calculating distance. Manhattan Distance is used to find the minimum distance of two points with the calculation formula as follows [18]:

   $$d(x,y) = \sum_{i=1}^{n} |x_i - y_i| \quad (6)$$

   Where:

   d = is the distance between x and y
   x = is cluster center data
   y = is data on attribute
   i = is every data
   n = is the number of data
   x_i = is the data at the center of the i-th cluster
   y_i = is the data on every data it-i

b. Data Validation

   Each training data to be processed using the Modified K-Nearest Neighbor (MK-NN) algorithm must go through a validation stage and depend on its closest neighbors. The results of the validity can be used as new information. The formula used to calculate the validity value is as follows :

   $$Validity(x) = \frac{1}{H} \sum_{i=1}^{H} S(lbl(x), lbl(Ni(x))) \quad (7)$$

   Where:

   Validity = is the validity between training data
   H = is the number of nearest neighbors
   $lbl(x)$ = is the class label x
   $lbl(Ni(x))$ = is the class label of the closest point to x.

   The function of S is used to calculate the similarity between two points, namely x and the i-th data from the nearest neighbor using the equation 8:

   $$S(a,b) = \begin{cases} 1a = b \\ 0a \neq b \end{cases} \quad (8)$$

   Where:

   S = is *similarity*
   a = is a in the *training*
   b = is a class other than a in the *training*

 http://doi.org/10.26594/register.v9i1.3045

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 42-54

c. Weighted Voting

The calculation of Weighted Voting uses the value of the results of the validity process and distance calculations on training data and test data. Each data is to be assigned a weighted value. From the results of the calculation on the Weighted Voting, the highest value is taken as a class determination. The value that has been determined is 0.5. Then, the value of each class is added up to select the class with the largest number. The calculation of *Weighted Voting* can be seen in equation 9 [19]:

$$W_{(X)} = \text{Validity(x)} \times \frac{1}{d+0,5} \quad (9)$$

Where:

$W_{(X)}$ = is *Weighted Voting* x
$Validity(x)$ = is the value of validity
d = is a distance value of

2.6. Testing

After the classification process was completed, the next process was the testing stage. The tests were carried out to measure how well the system functions. In the test, the model was validated using K-fold Cross Validation, while the level of accuracy was tested using the Multi Class Confusion Matrix. K-fold Cross Validation is one of the testing techniques used to measure process performance on a model or algorithm by dividing the data randomly to be grouped into several sets according to the number of K [21]. Meanwhile, Multi Class Confusion Matrix is a testing method commonly used to measure or calculate the value of precision, recall, error rate, and accuracy. The calculation of the value of precision, recall, and accuracy used in this study consists of three categories: feasible, less feasible, and not feasible. The calculation of the value of precision, recall, and accuracy is presented in the following table [20].

Table 1. Multi Class Confusion Matrix

		PREDICTION		
		Positive	**Negative**	**Netral**
ACTUAL	**Positive**	TPos	FPosNeg	FPosNet
	Negative	FNegPos	TNeg	FNegNet
	Netral	FNetpos	FNetNeg	TNet

Based on Table 1 above, the accuracy value can be calculated using the formula presented in equation 10 below.

$$accuracy = \frac{TPos+TNeg+Tnet}{\sum Data} \quad (10)$$

3. Results and Discussion

This chapter presents the process of image data processing and the application of the Modified K-Nearest Neighbor in classifying groundwater quality in Cibangkong Village, Pekucen District, Banyumas Regency. The stages of applying the procedures are carried out based on the design outlined in the previous chapter. To obtain the desired results, the test was carried out in several stages, starting from the image processing process, the application of the Knowledge Discovery Database (KDD), and the application of the Modified K-Nearest Neighbor (MK-NN).

3.1. Image Processing

At this stage, the images were processed to obtain results in the forms of means, standard deviations, skewness, and kurtosis. Several steps were carried out, starting from converting the images to grayscale, conducting Contrast Limited Adaptive Histogram Equalization (CLAHE), and first-order feature extraction.

a. Image Conversion to Grayscale

Grayscale contains information about the lighting in an image. The obtained image was converted into a grayscale form because the gray color has the same intensity within the RGB (Red, Green, Blue). The results of the grayscale image can be seen in figure 3.

 http://doi.org/10.26594/register.v9i1.3045

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 42-54

(a) (b)

Fig. 3. (a) An original image that has not been processed (b) The image that has been processed.

b. Contrast Limited Adaptive Histogram Equalization (CLAHE)

The CLAHE process was performed to improve the quality of the captured image to produce a more precise value based on the shape of the histogram. The results of the CLAHE process on the image data, along with the pixel intensity graph, can be seen in figure 4.

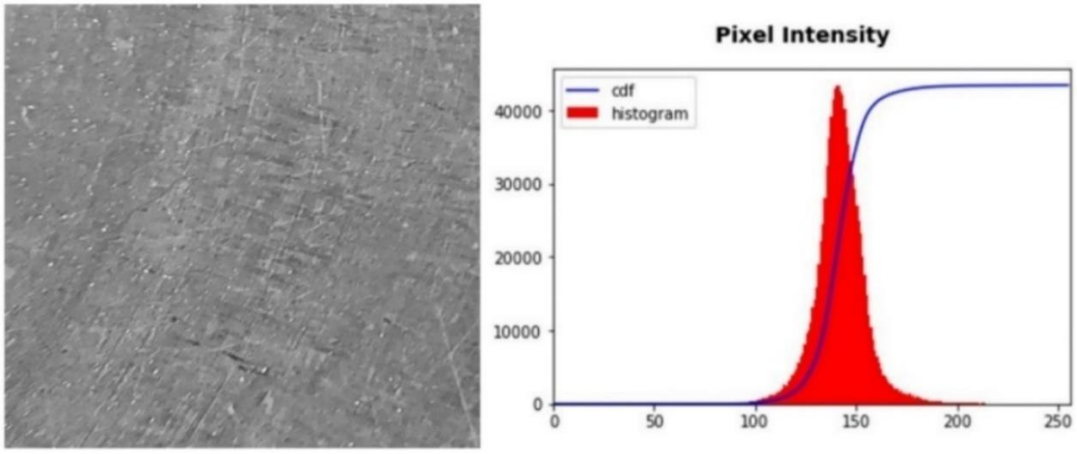

Fig. 4. Contrast Limited Adaptive Histogram Equalization

c. First-Order Feature Extraction

Feature extraction was performed to obtain the means, standard deviations, entropy, and skewness values. The results of the first-order feature extraction process for the top 5 data can be seen in table 2.

Table 2. Results of First-Order Feature Extraction

Citra	Mean	Standard Deviation	Skewness	Kurtosis
1	139.6453	15.55067704	0.041497389	5.819171
2	143.1097	15.69580687	0.07070109	4.505447
3	149.5215	14.6380436	0.103939693	6.140854
4	177.7366	12.85191205	0.057317065	3.398456
5	184.2149	11.40779317	-0.06881971	4.660153

3.2. The Application of Knowledge Discovery Database (KDD)

At the Knowledge Discovery Database (KDD) stage, the process started with the data selection process and data cleaning, and finally, the transformation was carried out.

a. Data Selection

The Data selection stage aims to select the parameters to be used in the next stage. The initial groundwater quality data has 12 parameters obtained from the images with four parameters, namely the means, standard deviations, skewness, and kurtosis. Meanwhile, there were three sensors to obtain pH, TDS, and temperature. The three physical parameters were in the form of color, smell, and taste. The other two parameters were the image and house description. These parameters were then selected to determine the requirements of water quality standards, such as physical, chemical,

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 42-54

and biological parameters. Parameters irrelevant to the study's purpose were removed in this stage, such as image description and house parameters.

b. Data Cleaning

The Data Cleaning process aims to examine inconsistent data, correct errors, and eliminate duplicate data. The initial data that has been selected was then cleaned using parameters that have a value of 'string' or character since the data could not be calculated. The results of the data cleaning for the top 5 data can be seen in Table 3 below.

Table 3. Results of Data Cleaning

Temperature	pH	TDS	Mean	Standard Deviation	Skewness	Kurtosis	Label
26	7.8	260	139.6453	15.5506	0.041497	5.819171	L
27	7.6	236	143.1097	15.6958	0.070701	4.505447	L
28	7.6	216	149.5215	14.6380	0.103939	6.140854	L
27	7	145	177.7366	12.8519	0.057317	3.398456	KL
27	7.6	211	184.2149	11.4077	-0.068819	4.660153	KL

In the label attribute, L indicates Eligibility, KL indicates Less Eligibility, and TL stands for Ineligibility.

c. Transformation

At this stage, the data transformation was carried out by changing the string type data to numeric data. As described in the "Label" attribute in Table 3 above, the value of KL (Less Eligible) was transformed into a numerical value of 0; the value of L (Eligible) was transformed into a numerical value of 1; and the value of TL (Ineligible) was transformed into a value of 2. The results of the data transformation of the top 5 data can be seen in Table 4 as follows.

Table 4. The Results of Data Cleaning

Temperature	pH	TDS	Mean	Standard Deviation	Skewness	Kurtosis	Label
26	7.8	260	139.6453	15.5506	0.041497	5.819171	1
27	7.6	236	143.1097	15.6958	0.070701	4.505447	1
28	7.6	216	149.5215	14.6380	0.103939	6.140854	1
27	7	145	177.7366	12.8519	0.057317	3.398456	0
27	7.6	211	184.2149	11.4077	-0.068819	4.660153	0

3.3. The Application of Modified K-Nearest Neighbor (MK-NN)

The stages of applying the MK-NN method were divided into several calculation processes, starting from calculating the distance (Manhattan Distance), validating the data, and calculating Weight Voting. However, before conducting these procedures, it is necessary to share training data and test data based on K-Fold Cross Validation.

a. The Distribution of Data Based on K-Fold Cross Validation

In this study, 100 pieces of data in the dataset will be divided into several sets based on the K-Fold value. The K value used in this study was K-1 to K-10, while the K-Fold value was from 2-Fold to 10-Fold. In this study, the researcher used a 10-Fold Cross Validation value with a K value of 2. Then, the 100 total data held would be divided into 10 data subsets. Each subset has ten total data and was tested ten times. Of the 10 subsets, 9 subsets were used as the training data, and one subset was used as the test data. The distribution of train and test data Fold-1 at K-Fold=10 is as follows.

Table 5. Distribution of Train Data for Each

No	Temperature	pH	TDS	Mean	Standard Deviation	Skewness	Kurtosis	Label
11	28	7.5	236	144.6765	23.4689	0.028825	3.261815	1
12	28	7.3	143	139.6577	16.3159	0.040312	6.334641	1
13	27	7.4	215	141.7302	22.3669	-0.012060	3.537585	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
98	27	6.6	294	175.115	13.5353	0.008497	46.51708	0
99	27	6.4	300	179.9523	13.4371	-0.003547	42.28206	2
100	26	6.9	234	147.5042	13.6818	0.036848	5.867746	1

 http://doi.org/10.26594/register.v9i1.3045

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 42-54

Table 6. Distribution of Test Data for Each Subset

No	Temperature	pH	TDS	Mean	Standard Deviation	Skewness	Kurtosis	Label
1	26	7.8	260	139.6453	15.5506	0.041497	5.819171	1
2	27	7.6	236	143.1097	15.6958	0.070701	4.505447	1
3	28	7.6	216	149.5215	14.6380	0.103939	6.140854	1
4	27	7.0	145	177.7366	12.8519	0.057317	3.398456	0
5	27	7.6	211	184.2149	11.4077	-0.068819	4.660153	0
6	26	7.6	232	137.9577	15.4276	0.062074	8.053961	1
7	26	7.5	239	165.5105	15.6763	-0.031227	3.368098	1
8	26	7.5	223	144.7662	22.6578	0.077950	4.292749	1
9	26	7.5	217	140.2694	13.2085	0.020393	6.146558	1
10	27	7.4	225	143.4012	14.4242	0.027817	4.784989	1

b. Manhattan Distance

The distance calculation was carried out using Manhattan Distance. The calculation was conducted using equation 2.14. The results of the calculation of the Manhattan distance to the training data can be seen in Table 7 below.

Table 7. Manhattan Distance Data Training at Fold-1

D	11	12	13	…	98	99	100
11	0	108.456	26.4649	…	143.547	150.460	19.828
12	108.456	0	84.0728	…	231.152	238.064	104.35
13	26.4649	84.0728	0	…	165.016	171.904	37.338
…	…	…	…	…	…	…	…
98	143.547	231.1521	165.016	…	0	15.3825	129.73
99	150.460	238.0646	171.904	…	15.3825	0	136.64
100	19.8286	104.3508	37.3380	…	129.735	136.647	0

After obtaining the Manhattan distance results for the training data, the following is the result of the Manhattan Distance for the test data at fold-1.

Table 8. Manhattan Distance Data Train Fold-1

D	1	2	3	…	8	9	10
11	41.8194	11.7254	36.730	…	16.9807	38.5606	23.9441
12	120.794	100.231	85.0990	…	95.7299	80.1270	90.2973
13	57.6363	30.3011	20.4393	…	13.2720	16.3606	20.9009
…	…	…	…	…	…	…	…
98	114.415	135.239	147.167	…	154.665	154.454	144.154
99	121.328	142.152	154.080	…	161.577	161.171	151.066
100	36.6808	11.5045	24.0137	…	24.930	25.6033	16.4370

c. Data Validity

The data validity process was carried out on each parameter contained in the training data subsets. The following is the formula for calculating the value of data validity using equation 1 with K = 2.

$$Validity(x) = \frac{1}{H}\sum_{i=1}^{H} S\left(lbl(x), lbl\left(Ni(x)\right)\right) \quad (7)$$

Train data 11 = $\frac{1}{2}(1+1) = 1$

Train data 12 = $\frac{1}{2}(1+1) = 1$

Train data 13 = $\frac{1}{2}(1+1) = 1$

The validity of the training data at fold-1 is described in Table 9 as follows.

Table 9. Validity of the Training Data

Fold-1	Test										
	11	12	13	14	15	…	96	97	98	99	100
Validity	1	1	1	1	1	…	0	1	0	2	1

d. Weight Voting

Weight voting was carried out on the test data and on all of the training data using equation 2. The following is an example of the calculation to find the Weight Voting.

$$W11{,}1(train\ data\ 11, test\ data\ 1) = 1\ x\ \frac{1}{41.81947} = 0.0236297 \quad (11)$$

 http://doi.org/10.26594/register.v9i1.3045

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 42-54

The calculation was carried out for all training data and the test data. After obtaining all the weight voting results, the largest value from the Weight Voting was identified, and all the results were sorted, descending from the largest to the smallest value. After obtaining the largest K results from the Weight Voting and its class, the researchers looked for the majority of the class from the K value. The results from the majority were used as the results of the prediction class. The results of the Weight Voting of test data at fold-1 can be seen in Table 10. The results of the prediction class and the actual class can are presented in Table 11.

Table 10. Weight Voting at Fold-1

Train Data	Test Data	Weight Voting	Class
76		0.06584098	L
86	1	0.05631559	L
63		0.03855382	L
91		0.12800727	L
40	2	0.08939296	L
100		0.08330186	L
⋮	⋮	⋮	⋮
21		0.12772796	L
15	10	0.09211361	L
16		0.08053460	L

3.4. Analysis

An analysis was carried out to obtain the results of testing the K-Fold Cross Validation value and the effect of the K value using the Multi-Class Confusion Matrix method.

3.4.1. K-Fold Cross Validation Testing

The K-Fold Cross Validation test was carried out using the Multi-Class Confusion Matrix method with a K-Fold value of 2-Fold to 10-Fold. The results of the K-Fold value tests that have been carried out produce the scores for precision, recall, f1-score and accuracy from K-Fold=2 to K-Fold=10 on K-1 to K-10. The graph of the K-Fold test results is shown in Figure 5 below.

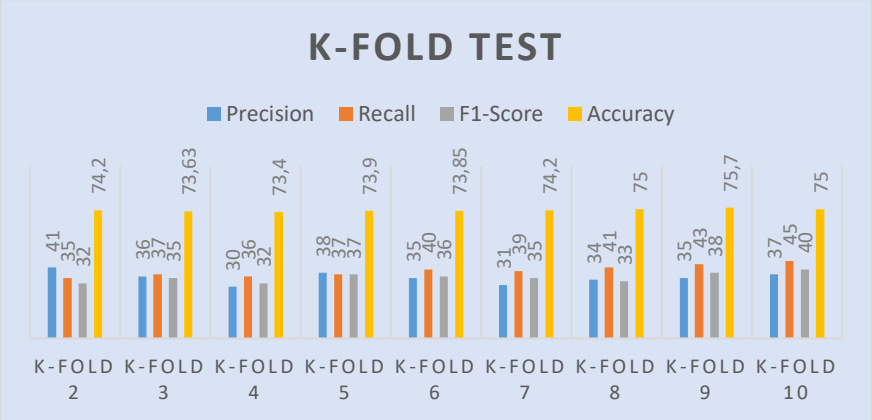

Figure 5. K-Fold Test Results

It can be seen in the graph of the K-Fold test results above that the highest average accuracy value obtained was at K-Fold = 9, with a percentage of 75.70% and a precision value of 0.35; the highest recall value was 0.43, and the highest F1-Score was 0.40. The lowest average value obtained was at K-Fold = 4 with a percentage of 73.40%, with a precision value of 0.30, recall value of 0.36, and F1-Score of 0.32.

3.4.2. Testing the Effect of K Value

Testing the K value is carried out using various K values, including from K = 1 to K = 10. The results of testing the effect of the K value can be seen in Figure 6. From the graph of the results of testing the effect of the K value above, the highest average accuracy value is obtained at K = 5 with a percentage of 67.90% with a precision value of 0.32, recall of 0.37, and F1-Score by 0.33. The lowest average value was obtained at K=3 with a percentage of 60.30% with a precision value of 0.39, recall of 0.38, and F1-Score of 0.35.

 http://doi.org/10.26594/register.v9i1.3045

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 42-54

Table 11. 10-Fold Prediction Results

Fold	Class Result	Test- 1	2	3	4	5	6	7	8	9	10		Precision	Recall	F1-Score	Accuracy
1	Prediction	1	1	1	0	1	1	1	1	1	1	0	1.00	0.50	0.67	90%
	Actual	1	1	1	0	0	1	1	1	1	1	1	0.89	1.00	0.94	
	Average												0.94	0.75	0.80	
2												0	0.33	1.00	0.50	80%
	Prediction	1	1	1	1	1	1	0	1	0	0	1	1.00	0.88	0.93	
	Actual	1	1	1	1	1	1	1	1	0	2	2	0.00	0.00	0.00	
	Average												0.44	0.62	0.48	
3												0	1.00	1.00	1.00	90%
	Prediction	1	1	1	1	1	1	1	1	0	1	1	0.89	1.00	0.94	
	Actual	1	1	1	1	1	1	1	1	0	2	2	0.00	0.00	0.00	
	Average												0.63	0.67	0.65	
4												0	0.00	0.00	0.00	90%
	Prediction	1	1	1	1	1	1	1	1	1	2	1	0.89	1.00	0.94	
	Actual	1	1	1	1	1	0	1	1	1	2	2	1.00	1.00	1.00	
	Average												0.63	0.67	0.65	
5												0	0.00	0.00	0.00	60%
	Prediction	1	1	0	1	1	1	0	1	0	2	1	0.71	0.71	0.71	
	Actual	2	1	1	1	1	1	1	0	0	2	2	1.00	1.00	1.00	
	Average												0.57	0.57	0.57	
6												0	1.00	0.50	0.67	80%
	Prediction	1	1	1	1	1	1	1	1	0	1	1	0.78	1.00	0.88	
	Actual	1	1	1	1	1	1	0	1	0	2	2	0.00	0.00	0.00	
	Average												0.59	0.50	0.51	
7												0	0.00	0.00	0.00	70%
	Prediction	1	1	1	1	1	1	1	1	0	1	1	0.75	0.86	0.80	
	Actual	1	1	1	1	1	1	0	1	0	2	2	1.00	1.00	1.00	
	Average												0.58	0.62	0.60	
8												0	0.00	0.00	0.00	70%
	Prediction	1	1	1	1	1	1	1	1	1	1	1	0.70	1.00	0.82	
	Actual	1	1	0	1	1	0	1	1	1	2	2	0.00	0.00	0.00	
	Average												0.23	0.33	0.27	
9												0	0.50	0.50	0.50	80%
	Prediction	1	0	1	1	1	1	1	1	1	0	1	0.88	1.00	0.93	
	Actual	0	0	1	1	1	1	1	1	1	2	2	0.00	0.00	0.00	
	Average												0.46	0.50	0.48	
10												0	0.50	0.50	0.50	70%
	Prediction	1	1	1	2	1	0	0	2	2	1	1	1.00	0.71	0.83	
	Actual	1	1	1	1	1	0	1	0	2	1	2	0.33	1.00	0.50	
	Average												0.61	0.74	0.61	

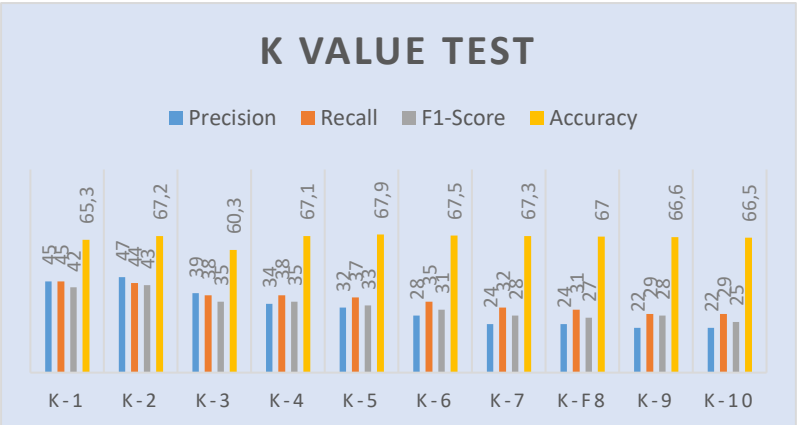

Figure 6. The Results of Testing the Effect of K Value

 http://doi.org/10.26594/register.v9i1.3045

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 42-54

4. Conclusion

In this study, the Modified K-Nearest Neighbor Algorithm was employed to determine groundwater quality. The results of the Weight Voting value was used to determine the highest class majority, and they were evaluated using the K-Fold Cross Validation and Multi Class Confusion Matrix algorithms. The test obtained the highest accuracy result of 78% in K-Fold = 2, K-Fold = 9, and K-Fold = 10 with the highest average accuracy obtained at K-Fold = 9 with a percentage of 75.70% with a precision value of 0.35, recall value of 0.43, and F1-Score of 0.40. The lowest average value was obtained at K-Fold = 4 with a percentage of 73.40% with a precision value of 0.30, recall value of 0.36, and F1-Score of 0.32. Meanwhile, the results of testing the effect of the K value obtained the highest accuracy value at K = 5 with a percentage of 67.90% with a precision value of 0.32, recall value of 0.37, and F1-Score of 0.33. The lowest average value was obtained on K3 with a percentage of 60.30% with a precision value of 0.39, recall value of 0.38, and F1-Score of 0.35. The large amount of training data greatly affects the accuracy results obtained, that is, the more training data used, the better the accuracy results obtained. The use of the number of K-Folds also affects the results obtained, that is, the higher the K-Fold used, the higher the impact of accuracy. For the result of Multi Class Confusion Matrix test, it was at 0 because of the influence of the data imbalance and the number of each class classification. However, the results of the study were considered relatively good, even though they had not reached 80%. From the results of the tests carried out, it can be concluded the groundwater quality is in a usable condition. The Modified K-Nearest Neighbor algorithm also works relatively well for the classification procedure.

Authors' Contributions

H. S. Amalia: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, visualization, writing – original draft, and writing - review & editing. U. Athiyah: Conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, resources, validation, writing – original draft, and writing – review & editing. A. W. Muhammad: Conceptualization, funding acquisition, investigation, methodology, supervision, writing – original draft, and writing - review & editing.

Acknowledgment

The author's acknowledgment should be conveyed to IT Telkom Purwokerto for providing facilities and knowledge during education. We would like to thank our supervisor who has directed and guided us during this research activity. We also thank the people of Cibangkong Village, Pekuncen District, Banyumas Regency who had provided access for the researchers to cunduct this research, which is an effort to bridge technology to humanity.

Declaration of Competing Interest

We declare that we have no conflict of interest.

References

[1] World Health Organization, Water for health: taking charge, World Health Organization (WHO), 2001.

[2] Zamroni A., Trisnaning P.T., Prasetya H.N.E., Sagala S.T., and Putra A.S. (2022). Geochemical Characteristics and Evaluation of the Groundwater and Surface Water in Limestone Mining Area around Gunungkidul Regency, Indonesia. The Iraqi Geological Journal, 189-198.

[3] P. Rekha, K. Sumathi, S. Samyuktha, A. Saranya, G. Tharunya and R. Prabha, "Sensor Based Waste Water Monitoring for Agriculture Using IoT," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020.

[4] S. Nashif, R. Raihan, R. Islam and M. H. Imam, "Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system," *World Journal of Engineering and Technology,* vol. 6, no. 4, pp. 854-873, 2018.

ISSN 2502-3357 (online) | ISSN 2503-0477 (print)
9 (1) January 2023 42-54

[5] Boateng, E. , Otoo, J. and Abaye, D. (2020) Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review. *Journal of Data Analysis and Information Processing*, **8**, 341-357. doi: 10.4236/jdaip.2020.84020.

[6] H. Shahabi *et al.*, "Flood Detection and Susceptibility Mapping Using Sentinel-1 Remote Sensing Data and a Machine Learning Approach: Hybrid Intelligence of Bagging Ensemble Based on K-Nearest Neighbor Classifier," *Remote Sensing*, vol. 12, no. 2, p. 266, Jan. 2020, doi: 10.3390/rs12020266.

[7] Okfalisa, I. Gazalba, Mustakim and N. G. I. Reza, "Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification," in *017 2nd international conferences on information technology, information systems and electrical engineering (ICITISEE)*, 2017.

[8] Y. Lee, A. Scolari, B.-G. Chun, M. D. Santambrogio, M. Weimer, and M. Interlandi, "Pretzel: Opening the Black Box of Machine Learning Prediction Serving Systems," in Proceedings of the 13th USENIX Symposium on Operating Systems Design and Implementation (OSDI '18), Carlsbad, CA, USA, Oct. 8-10, 2018.

[9] B. G. Marcot and A. M. Hanea, "What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?," *Computational Statistics,* vol. 36, no. 3, pp. 2009-2031, 2021, https://doi.org/10.1007/s00180-020-00999-9.

[10] A. A. Nababan, M. Khairi, and B. S. Harahap, "Implementation of K-Nearest Neighbors (KNN) Algorithm in Classification of Data Water Quality", *Mantik*, vol. 6, no. 1, pp. 30-35, Mar. 2022.

[11] R. I. Perwira, B. Yuwono, R. I. P. Siswoyo, F. Liantoni and H. Himawan, "Effect of information gain on document classification using k-nearest neighbor," Register: Jurnal Ilmiah Teknologi Sistem Informasi, vol. 8, no. 1, pp. 50-57, 2022.

[12] N. Radhakrishnan and A.S. Pillai, "Comparison of Water Quality Classification Models using Machine Learning," in Proceedings of the Fifth International Conference on Communication and Electronics Systems (ICCES), 2020, pp. 409-413.

[13] C.-M. Hsu, C.-C. Hsu, Z.-M. Hsu, F.-Y. Shih, M.-L. Chang, and T.-H. Chen, "Colorectal Polyp Image Detection and Classification through Grayscale Images and Deep Learning," *Sensors*, vol. 21, no. 18, p. 5995, Sep. 2021, doi: 10.3390/s21185995.

[14] V. Stimper, S. Bauer, R. Ernstorfer, B. Schölkopf, and R.P. Xian, "Multidimensional Contrast Limited Adaptive Histogram Equalization," IEEE Access, vol. 7, pp. 150834-150846, 2019, doi: 10.1109/ACCESS.2019.2952899.

[15] S. Nalband, C.A. Valliappan, A. Prince, and A. Agrawal, "Time-frequency based feature extraction for the analysis of vibroarthographic signals," Comput. Electr. Eng., vol. 67, pp. 196-208, Jul. 2018, doi: 10.1016/j.compeleceng.2018.02.009.

[16] M. M. Ghazala and A. Hammad, "Application of knowledge discovery in database (KDD) techniques in cost overrun of construction projects," *International Journal of Construction Management,* vol. 22, no. 9, pp. 1632-1646, 2022.

[17] S. M. Ayyad, A. I. Saleh and L. M. Labib, "Gene expression cancer classification using modified K-Nearest Neighbors technique," *Biosystems,* vol. 176, pp. 41-51, 2019.

[18] M. Faisal, E.M. Zamzami, and Sutarman, "Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance," J. Phys.: Conf. Ser., vol. 1566, article 012112, Nov. 2019, doi: 10.1088/1742-6596/1566/1/012112.

[19] V. C. Osamor and A. F. Okezie, "Enhancing the weighted voting ensemble algorithm for tuberculosis predictive diagnosis," *Scientific Reports,* vol. 11, article 14806, Jul. 2021, doi: 10.1038/s41598-021-94279-w.

[20] I. Markoulidakis, I. Rallis, I. Georgoulas, G. Kopsiaftis, A. Doulamis, and N. Doulamis, "Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem," *Technologies*, vol. 9, no. 4, p. 81, Nov. 2021, doi: 10.3390/technologies9040081.