International Journal of Sustainable Applied Sciences (IJSAS)
Vol.2, No.5, 2024: 379-390

Valuation K-Nearest Neighbors and Naïve Bayes for Dringking Water Potability Classification

Anisa Rahmawati[1*], Muhamad Fatchan [2], Wahyu Hadikristanto[3]
Universitas Pelita Bangsa
Corresponding Author: Anisa Rahmawati rahmawatianisa2000@gmail.com

A R T I C L E I N F O

Keywords: Drinking Water;K-Nearest Neighbors, Naïve Bayes, Machine Learning

Received : 03, April
Revised : 16, April
Accepted: 22, May

A B S T R A C T

The availability of drinking water that is safe and suitable for consumption is important to support health and development. This research emphasises the importance of handling the clean water crisis through the evaluation of drinking water quality using data mining algorithms. The dringking water quality evaluation method was selected using the K-Nearest Neighbors and Naive Bayes algorithms, replacing the manual method which is less responsive in predicting. The experimental process was conducted by utilising Kaggle website data by applying data processing and oversampling techniques to handle class imbalance in the dataset used. Bases on the research results, the accurancy of the K-Nearest Neighbors Algorithm reaches 65%, which is higher than the accuracy od the Naive Bayes Algorithm which is 64%. So it can be concluded that the K-Nearest Neighbors Algorithm is more effective in predicting the quality of water suitable for consumption. This research provides an in-depth insight into the use of technology and data analysis in dealing with the crisis in the availability of water suitable for consumption and offers suggestions for further research using more diverse methods and the use of more datasets to improve accuracy in evaluating the quality of potable water.

DOI: https://10.59890/ijsas.v2i5.1864
ISS-E: 3025-5597
https://journal.multitechpublisher.com/index.php/ijsas

INTRODUCTION

An indispensable essential for the survival of all creatures is dringking water, which plays an important role in maintaining health and bodily functions, especially for humans. The availability of drinking water that is safe and suitable for consumption is not only important for maintaining health, but also as a support for sustainable economic and social development (B. K. Mishra et al., 2021). Around 2.2 billion people worldwide still face challenges in gaining access to drinking water that meets safety standards (Salehi, 2022). The clean water crisis is a very worrying problem around the world, where the amount of clean water that can be consumed by humans is only 1%. According to WHO data, 633 million people have difficulty accessing clean water (Riyantoko et al., 2021). Water shortages are predicted to be faced by two-thirds of the world's population by 2025 according to estimates from the World Wide Fund (WFF). This situation has the potential to increase the suffering of the global ecosystem (Arora & Mishra, 2022). According to data from UNESCO, individuals living in areas experiencing water shortages number around 2 billion, while those without adequate access to safe dringking water sources account for more than 800 million people (Makarigakis & Jimenez-Cisneros, 2019).

Factors such as climate change, population growth, unsustainable land use, development, industrial and domestic effluents pose a serious threat to the availability of clean water quality (R. K. Mishra, 2023). Expanding regard for dringking water quality issues is boundless and worldwide because of the potential contamination that can be brough about by different hurtful substancea like pesticides, patogenic microorganisms, weighty metal, and other synthetic blends than can compromise the water supply required by the human body (P. Zhang et al., 2023). According, great administration and control are expected to protect the idea of clean water so the nature of water can be guaranteed to be appropriate for human use (Tangkelayuk, 2022). Assessment of drinking water quality is often done through manual laboratory testing which is time-consuming and less responsive in real time. With the development of technology especially in the field of machine learning comes the option to evaluate dringking water quality automatically, quickly and accurately (Rambe et al., 2024). Artificial intelligence (AI) methods such as machine learning are considered as effective tools for monitoring, data management and policy making in the context of water quality (Lowe & Qin, 2022).

Drinking water quality is measured by various variables and parameters that are used as the basis for model building such as ph, turbidity, metal content etc. (Shaibur et al., 2024). Classification is a stage of data pattern analysis that aims to identify the appropriate class or category in an unrecognised object based on the characteristics of the observed features (Ainurrohma, 2021). A few ordinarily utilized information characterization methods incorporate the K-Nearest Neighbors and Naive Bayes calulations which produce shifting exactness. The K-Nearest Neighbors calculation is most popular for its capacity to track down the nearest distance between the assessed informations and the closest neighbor in the preparations informational index (Saadatfar et al., 2020). In the mean time, the Naive Bayes calculation is utilized to work out the likelihood of each class and decide the most elevated likelihood worth to characterize the test information (Sendari et al., 2020). The use of algorithms in assessing the quality of potable water is increasingly attractive because it can provide effective and efficient solutions in facing global challenges related to the clean water crisis that is suitable for consumption (Ghina Annaifah, 2024). This research was conducted with the aim of analyzing and evaluating the performance of both algorithms in identifying the quality of dringking water. Through this research, it is expected to find the most effective method in achieving maximum accuracy in identifying the quality of drinking water.

LITERATURE REVIEW

Potable water must meet certain quality standards to be safe for consumption. The quality of drinking water can be influenced by several factors such as pollution, water sources and water treatment. Parameters used in assessing the quality of potable water include physical, chemical and microbial parameters such as colour, temperature, ph, turbidity, mercury, iron, metals, E.coli bacteria and many more (Silva et al., 2022). Data mining is the most common way of extricating data to acquire new bits of knowledge. This research utilises data mining techniques by implementing the K-Nearest Neighbors and Naïve Bayes methods to compare the most optimal accuracy levels of the two methods. Data mining is a process that involves utilizing data to detect relationships or patterns in large data sets with the goal of gaining new insights that may not have been revealed previously (Sree & Vardhani, 2015).Therefore, the development of effective classification models to assess water quality is very important in an effort to maintain public health (Park et al., 2020).

Classification is an important aspect of machine learning that aims to categorize data into specific groups. To create a model that may be used to categorize newly unlabeled data the process entails learning from an existing labeled collection of data. When it comes to treating drinking water resources, classification is used to identify based on physical, chemical and biological measurement characteristics. This allows for the quick and accurate decision making process (Zainurin et al., 2022).

The K-Nearest Neighbors (KNN) method is often the first choice in classification due to its similarity to the nearest neighbor class. However, its main drawback lies in its dependence on the K parameter, which can significantly affect classification results and its sensitivity to outlier data. So it is necessary to adjust the K parameter to improve optimal accuracy results (Wang et al., 2020). The basic principle of K-nearest Neighbors (KNN) is to find the closest data to the evaluation data based on the K nearest neighbors in the training dataset. Before searching for the closest distance, the K-Nearest Neighbors algorithm needs to do preprocessing or normalisation first which aims to equalise the standard values on all attributes or indicators used in the calculation (Song et al., 2022). The application of KNN to drinking water quality has an advantage in its ability to handle small and simple datasets very effectively (Juna et al., 2022).

Categorization approaches using Naïve bayes techniques based on Bayes principle are often used in various situations, including water quality assessment. Naïve Bayes assumes that all independent variables in the data have no relationship with each other. Thus, the measurement parameters can be studied separately, which speeds up the classification process. The drawback of Naïve Bayes lies in the dependence on the availability of consistent and good data to provide reliable results (Ilić et al., 2022). The advantage of using Niave Bayes lies in the need for little training data to determine the mean and variance parameters of the variables required for classification (Chen et al., 2021).

METHODOLOGY

This research uses an experimental and evaluation approach model that aims to compare and assess the effectiveness of data mining classification algorithms in analysing various aspects related to water quality.

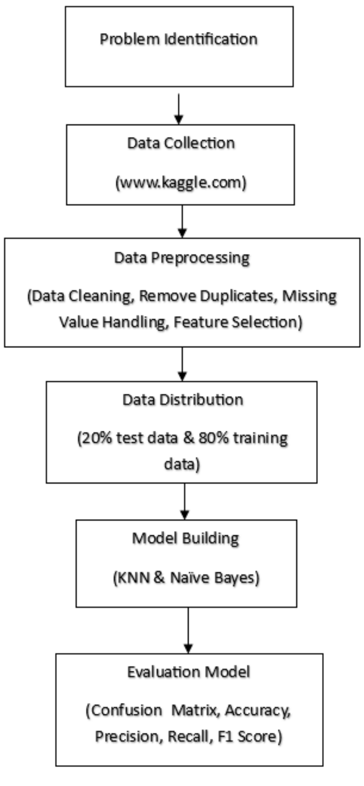

Figure 1. Research flow

Data Collections

Utilizing the Google Colab platfrom , analysis and processing were performed using the Python programming language. The scikit-learn library was used for machine learning, while pandas was used for data manipulation. The firsh step wa to collect data from a publick source Kaggle which contains various water samples with their quality parameters. Each sample acts as an attribute reflecting its reature while the ‘potability’ column serves as the target variable the predicts whether the water is safe for consumptions or not.

Data Preprocessing

The next process is data processing which consists of data cleaning, overcoming missing values by using a feature selection approach such as calculating the average to fill in the empty values with existing data, removing duplicate data which aims to ensure consistency and cleanliness of the data before proceeding to the next stage (Y. Zhang & Thorburn, 2022). In addition, class balencing is carried out using oversampling techniques to handle the imbalance of data classes that can improve the performance of the model to be tested. This is because of the inclination of the model to lean toward the larger part class which has more examples contrasted with the minority class. Thusly, an oversampling procedure where tests from the minority class are imitated to offset their number with the greater part class is required (Wongvorachan et al., 2023). Data visualisation uses bar charts to understand patterns and characteristics in the data.

Data Distribution

The next step is to divide the dataset into 20% testing data and 80% training data to train the model. In the process of data analysis and modelling we divide the dataset into two main part training data and testing data. This division aim to ensure that the developed model can be evaluated objectively on data that is has never seen during the training process. The training data is used to train the model where the model will learn the data to recognise patterns and make predictions. While the testing data is used to evaluate the performance of the model after the training process is complete. It gives an indication of how well the model can predict new data that is has never seen before. The trained model will be tested with the test data. The predicted results will be tested with the test data. The predicted results will be compared with the actual values to measure how well the model works.

Evaluation Model

The next step is to train the model using the K-Nearest Neighbors (KNN) and Naïve Bales algorithms, the evaluate the performance of both models. The evaluation is done using several metrics such as confusion matrix, accuracy, precision, recall and F1 score. Confusion matrix provides a detailed overview of how the model predicts each class to see the number of correct and incorrect predictions for each class. Accuracy measure how often the model makes correct predictions. Precision measure the accuracy of positive predictions. Recall measure the ability of the model to fins all positive examples. F1 score provides a balance between precision and recall, especially useful when there is an imbalance of classes in the dataset.

RESULTS

The data source is taken from the kaggle website https:// www.kaggle.com/datasets/uom190346a/water-quality-and- potability /data with the name 'water_potability' which is csv data type.

Table 1. Potability data

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0
...	...	...	...	...	...	...	...	...	...	...
3271	4.668102	193.681735	47580.991603	7.166639	359.948574	526.424171	13.894419	66.687695	4.435821	1
3272	7.808856	193.553212	17329.802160	8.061362	NaN	392.449580	19.903225	NaN	2.798243	1
3273	9.419510	175.762646	33155.578218	7.350233	NaN	432.044783	11.039070	69.845400	3.298875	1
3274	5.126763	230.603758	11983.869376	6.303357	NaN	402.883113	11.168946	77.488213	4.708658	1
3275	7.874671	195.102299	17404.177061	7.509306	NaN	327.459760	16.140368	78.698446	2.309149	1

3276 rows × 10 columns

Table 2. The Amount of Data Used was 3276 Rows of Data with Ten of Them:

Parameter	Descriptions
pH	The pH level of the water. The optimum pH required ranges from 6.5 to 8.5
Hardness	Water hardness, a measure of mineral content
Solids	Total dissolves solid in water
Chloramines	Cholramine concentration in water
Sulfate	Sulfate concentration in water
Conductivity	Electrical conductivity of water
Organic Carbon	Organic compounds dissolved in water
Trihalomethanes	The concentration of trihalomethanes in the water
Turbidity	Turbidity level , a measure of water clarity
Potability	Target variabel, indicates the potability of water with values of 1 (potable) and 0 (not potable)

Data analysis is needed to assess the feasibility of the data before proceeding to the next stage. There were data blanks in the NaN format, so it was necessary to fill in the blank values using the average value of each attribute.

```
ph                 491
Hardness             0
Solids               0
Chloramines          0
Sulfate            781
Conductivity         0
Organic_carbon       0
Trihalomethanes    162
Turbidity            0
Potability           0
dtype: int64
```

Figure 1.Data Missing Value

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	3276.000000	3276.000000	3276.000000	3276.000000	3276.000000	3276.000000	3276.000000	3276.000000	3276.000000	3276.000000
mean	7.080795	196.369496	22014.092526	7.122277	333.775777	426.205111	14.284970	66.396293	3.966786	0.390110
std	1.469956	32.879761	8768.570828	1.583085	36.142612	80.824064	3.308162	15.769881	0.780382	0.487849
min	0.000000	47.432000	320.942611	0.352000	129.000000	181.483754	2.200000	0.738000	1.450000	0.000000
25%	6.277673	176.850538	15666.690297	6.127421	317.094638	365.734414	12.065801	56.647656	3.439711	0.000000
50%	7.080795	196.967627	20927.833607	7.130299	333.775777	421.884968	14.218338	66.396293	3.955028	0.000000
75%	7.870050	216.667456	27332.762127	8.114887	350.385756	481.792304	16.557652	76.666609	4.500320	1.000000
max	14.000000	323.124000	61227.196008	13.127000	481.030642	753.342620	28.300000	124.000000	6.739000	1.000000

Figure 3. Average Value

Histogram graphs are presented as an overview of various measurement parameters in water that show how often the values of the parameters appear in the data.

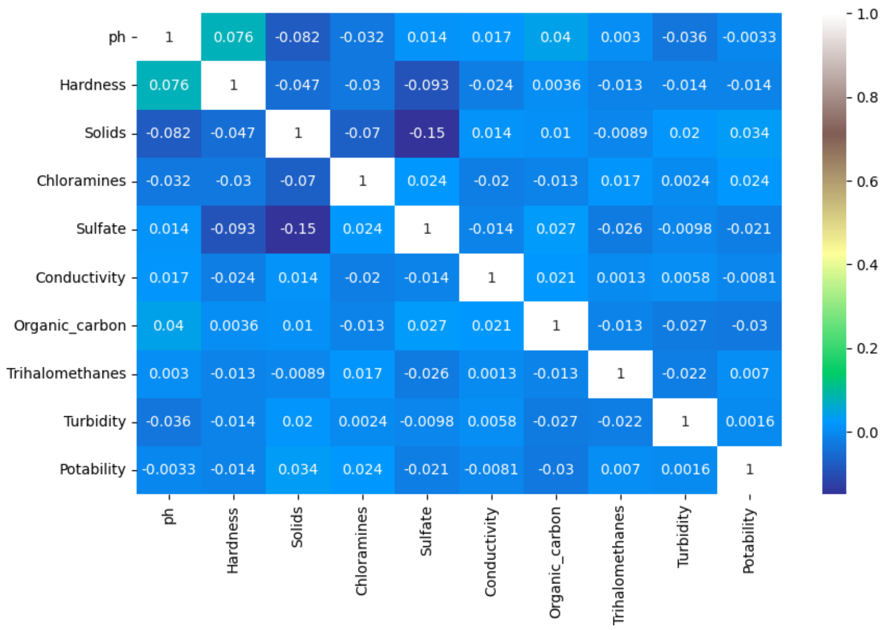

Figure 4. Histogram image

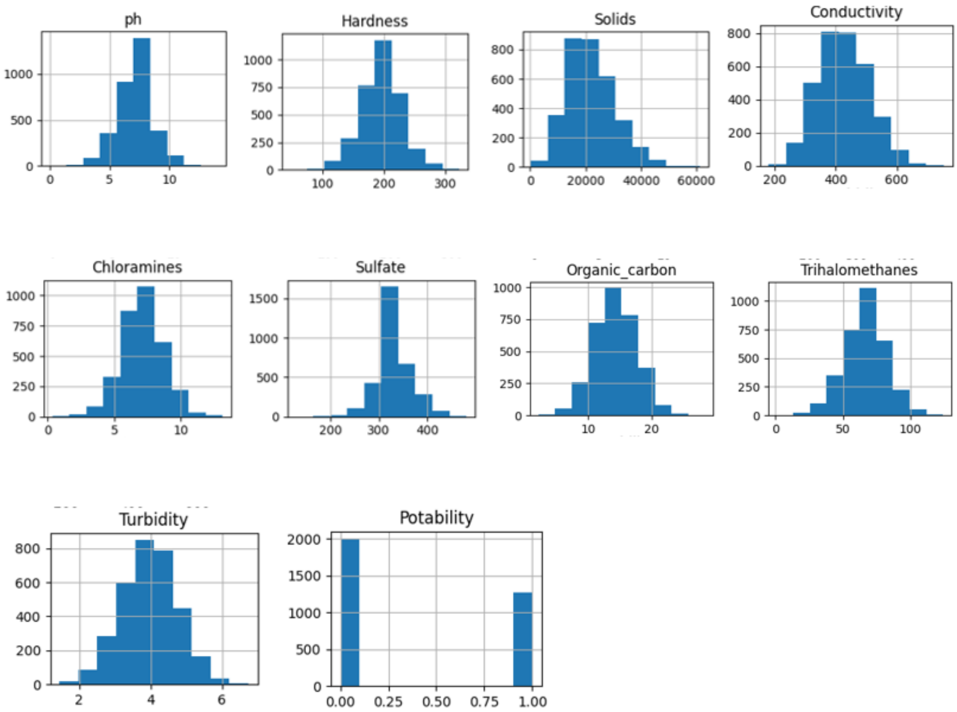

Figure 5.Metric Frequency Distribution

Understanding the distribution of the metrics is essential to improve the performance of the prediction model. The figure above helps detect class imbalance in the dataset. The next step involves dividing the water quality data represented in the potability column in the dataset using exploratory data analysis (EDA) which shows the difference that out of a total of 3276 data there is a total of 1998 non-potable data and 1278 potable data. The percentage indicates that 61% of the total data is non potable while 39% is potable.

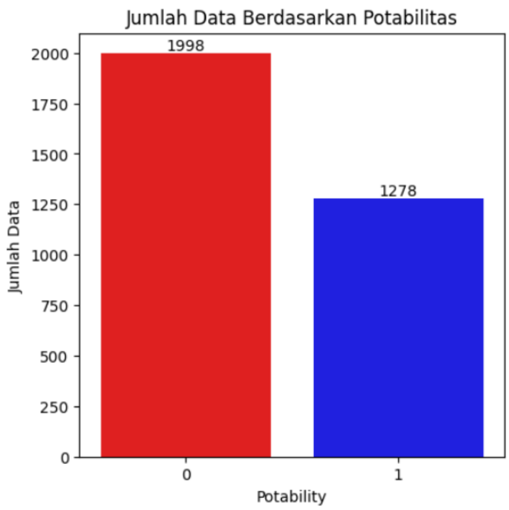

Figure 6.Comparison of the Amount of Data

Class balancing on the data is done after passing the data cleaning process that can affect the performance of the model. By using an oversampling technique that functions to duplicate the majority class sample (0) which has 1998 data records to be balanced with the minority class sample (1) so that it has the same number of data records.

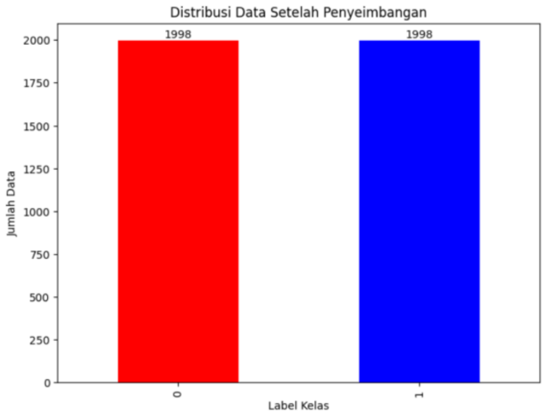

Figure 7.After Balencing Data

Data sharing was done with an allocation of 80% for model training and 20% for performance evaluations using test data. The number of training data entries was 2620, while the testing data had 656 entries. The modeling process involved two machine laerning algorithms that produced different values.

Table 3. The Number of Training Data

Algoritma		**Evaluation**				
KNN	Label	Precision	Recal	F1-score	Support	Accuracy
	0	0.65	0.64	0.76	396	0.65
	1	0.64	0.25	0.36	260	
Naïve Bayes	0	0.64	0.91	0.75	396	0.64
	1	0.63	0.22	0.33	260	

DISCUSSION

Based on the research results that have been presented previously, it shows that the K-Nearest Neighbors and Naïve Bayes algorithms provide fairly good predictions in evaluating the quality of potable water. However, it should be noted that classification performance can be affected by several factors such as parameter settings and model fit with the data used. The K-Nearest Neighbors algorithm tends to provide more accurate results, while Naïve Bayes shows a tendency to be faster in the classification process. Therefore, it is important to carefully consider the advantages and disadvantages of each algorithm according to the specific needs in assessing dringking water quality effectively and optimally.

CONCLUSIONS AND RECOMMENDATIONS

Based on the study's findings, it can be said that the K-Nearest Neighbors Algorithn outperforms the Naïve Bayes Algorithm in terms of accuracy. The K-Nearest neighbors algorithm achieved an accurancy rate of 65%, while the Naïve Bayes algorithm achieved an accuracy rate of 64%.

Therefore, it can be concluded that the K-Nearest Neighbors Algorithm is superior in modeling relationships and patterns in the tested data. The use of the K-Nearest Neighbors Algorithm by implementing oversampling and class balencing techniques results in a classification that is good enough to be applied to the case of drinking water quality classification. Suggestions for future research can use more methods in classification and the use of wider and more diverse data to increase the level of accuracy that is more optimal.

FURTHER STUDY

This research actually has constraints so further examinations is required for a more inside and out investigation of the poin 'Evaluation of K-Nearest Neighbors and Naïve Bayes for Drinking Water Quality Classification'.

ACKNOWLEDGMENT

In this study, the authors would like to thank the Supervisor who has directed so that this research can be completed. The author would also like to thank his parents for their support and encouragement. In addition, the author would like to express his deepest appreciation to colleagues who have provided advice and input that is very helpful in compiling this paper. Support and assistance from various parties are very meaningful to the author in completing this research.

REFERENCES

Ainurrohma. (2021). Akurasi Algoritma Klasifikasi pada Software Rapidminer dan Weka. PRISMA, Prosiding Seminar Nasional Matematika, 4, 493–499. https://journal.unnes.ac.id/sju/index.php/prisma/

Arora, N. K., & Mishra, I. (2022). Sustainable development goal 6: Global Water Security. Environmental Sustainability, 5(3), 271–275. https://doi.org/10.1007/s42398-022-00246-5

Chen, H., Hu, S., Hua, R., & Zhao, X. (2021). Improved naive Bayes classification algorithm for traffic risk management. Eurasip Journal on Advances in Signal Processing, 2021(1). https://doi.org/10.1186/s13634-021-00742-6

Ghina Annaifah, S. (2024). Tata kelola sumber daya air berkelanjutan-berkeadilan : Bagaimana Indonesia memperkuat poros maritim? EcoProfit: Sustainable and Environment Business, 1(2), 90–105. https://doi.org/10.61511/ecoprofit.v1i2.2024.331

Ilić, M., Srdjević, Z., & Srdjević, B. (2022). Water quality prediction based on Naïve Bayes algorithm. Water Science and Technology, 85(4), 1027–1039. https://doi.org/10.2166/wst.2022.006

Juna, A., Umer, M., Sadiq, S., Karamti, H., Eshmawi, A. A., Mohamed, A., & Ashraf, I. (2022). Water Quality Prediction Using KNN Imputer and Multilayer Perceptron. Water (Switzerland), 14(17), 1–19. https://doi.org/10.3390/w14172592

Lowe, M., & Qin, R. (2022). A Review on Machine Learning, Artificial Intelligence, and Smart Technology in Water Treatment and Monitoring. 1–28. https://www.mdpi.com/2073-4441/14/9/1384

Makarigakis, A. K., & Jimenez-Cisneros, B. E. (2019). UNESCO's contribution to face global water challenges. Water (Switzerland), 11(2). https://doi.org/10.3390/w11020388

Mishra, B. K., Kumar, P., Saraswat, C., Chakraborty, S., & Gautam, A. (2021). Water Security in a Changing Environment : Concept ,. Water, 13(4), 490.

Mishra, R. K. (2023). Fresh Water availability and It's Global challenge. Journal of Marine Science and Research, 2(1), 01–03. https://doi.org/10.58489/2836-5933/004

Park, J., Kim, K. T., & Lee, W. H. (2020). Recent advances in information and communications technology (ICT) and sensor technology for monitoring water quality. Water (Switzerland), 12(2). https://doi.org/10.3390/w12020510

Rambe, A. S., Tanjung, D., Octiara, E., Ridho, H., Yustina, I., Siregar, M., Lydia, M. S., Pasaribu, N., Zein, T. T., Supriana, T., & Hartono, R. (2024). Perkembangan teknologi digitaluntuk berbagai bidang kehidupan (digital teknologi for humanity). 1. usupress.usu.ac.id

Riyantoko, P. A., Fahrudin, T. M., Hindrayani, K. M., Data, S., & Timur, J. (2021). Analisis Sederhana Pada Kualitas Air Minum Berdasarkan Akurasi Model Klasifikasi Dengan Menggunakan Lucifer Machine Learning. Seminar Nasional Sains Data, 2(Senada), 12–18.

Saadatfar, H., Khosravi, S., Joloudari, J. H., Mosavi, A., & Shamshirband, S. (2020). A new k-nearest neighbors classifier for big data based on efficient data pruning. Mathematics, 8(2), 1–12. https://doi.org/10.3390/math8020286

Salehi, M. (2022). Global water shortage and potable water safety Today's concern and tomorrow's crisis. Environment International, 158, 106936. https://doi.org/10.1016/j.envint.2021.106936

Sendari, S., Zaeni, I. A. E., Lestari, D. C., & Hariyadi, H. P. (2020). Opinion Analysis for Emotional Classification on Emoji Tweets using the Naïve Bayes Algorithm. Knowledge Engineering and Data Science, 3(1), 50–59. https://doi.org/10.17977/um018v3i12020p50-59

Shaibur, M. R., Howlader, M., Ahmmed, I., Sarwar, S., & Hussam, A. (2024). Water quality index and health risk assessment for heavy metals in groundwater of Kashiani and Kotalipara upazila, Gopalganj, Bangladesh. Applied Water Science, 14(5). https://doi.org/10.1007/s13201-024-02169-4

Silva, G. M. E., Campos, D. F., Brasil, J. A. T., Tremblay, M., Mendiondo, E. M., & Ghiglieno, F. (2022). Advances in Technological Research for Online and In Situ Water Quality Monitoring—A Review. Sustainability (Switzerland), 14(9), 1–28. https://doi.org/10.3390/su14095059

Song, Y., Kong, X., & Zhang, C. (2022). A Large-Scale k -Nearest Neighbor Classification Algorithm Based on Neighbor Relationship Preservation.

Wireless Communications and Mobile Computing, 2022. https://doi.org/10.1155/2022/7409171

Sree, Y. U., & Vardhani, P. R. (2015). Pattern Finding in Large Datasets with Big Data Analytics Mechanism. 351 International Journal of Computer Engineering in Research Trends, 2(5), 2349–7084. http://www.ijcert.org

Tangkelayuk, A. (2022). The Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes, dan Decision Tree. JATISI (Jurnal Teknik Informatika Dan Sistem Informasi), 9(2), 1109–1119. https://doi.org/10.35957/jatisi.v9i2.2048

Wang, B., Ying, S., & Yang, Z. (2020). A Log-Based Anomaly Detection Method with Efficient Neighbor Searching and Automatic K Neighbor Selection. Scientific Programming, 2020. https://doi.org/10.1155/2020/4365356

Wongvorachan, T., He, S., & Bulut, O. (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. Information (Switzerland), 14(1). https://doi.org/10.3390/info14010054

Zainurin, S. N., Wan Ismail, W. Z., Mahamud, S. N. I., Ismail, I., Jamaludin, J., Ariffin, K. N. Z., & Wan Ahmad Kamil, W. M. (2022). Advancements in Monitoring Water Quality Based on Various Sensing Methods: A Systematic Review. International Journal of Environmental Research and Public Health, 19(21). https://doi.org/10.3390/ijerph192114080

Zhang, P., Yang, M., Lan, J., Huang, Y., Zhang, J., Huang, S., Yang, Y., & Ru, J. (2023). Water Quality Degradation Due to Heavy Metal Contamination: Health Impacts and Eco-Friendly Approaches for Heavy Metal Remediation. Toxics, 11(10). https://doi.org/10.3390/toxics11100828

Zhang, Y., & Thorburn, P. J. (2022). Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. Future Generation Computer Systems, 128, 63–72. https://doi.org/10.1016/j.future.2021.09.033