ANALYZING THE PERFORMANCE OF MACHINE LEARNING ALGORITHMS FOR PREDICTING WATER QUALITY INDEX

V. Queen Jemila, Assistant Professor of Computer Applications (PG) V.V. Vanniaperumal College for Women Virudhunagar, India - 626001.
M. Dhanalakshmi Associate Professors of Chemistry V.V. Vanniaperumal College for Women Virudhunagar, India - 626001.
M. Amutha Associate Professors of Chemistry V.V. Vanniaperumal College for Women Virudhunagar, India - 626001.

ABSTRACT
Our research aimed to calculate the water quality indices of the bore water in our surrounding educational institutions using three machine learning algorithms. Our research differs from other related works by choosing decision tree, K-nearest neighbor, and naive Bayes methods and analyzing their performance accurately. We collected water samples from our neighboring educational institutions such as Schools and Colleges and calculated six important factors: salinity, total suspended solids (TDS), dissolved oxygen (DO), acidity and alkalinity (pH), and biochemical oxygen demand (BOD). Using the efficient chemical method weighted arithmetic water quality index (WAWQI), the quality parameters of the water samples were examined. We created our dataset by utilizing these metrics, and the dataset is used as our chosen algorithm's training and testing data. We implemented these machine learning algorithm models using Google Colab. We created three separate models using the above algorithms and analyzed their performance. Finally, we obtained the WQI values of our dataset and three different accuracies.

Keywords:
Decision Tree, Gini Index, KNN, Naive Bayes Water Quality Index

1. INTRODUCTION
One of the major resources for human beings is water. People use water frequently in their day-to-day lives. Pure water was used to avoid skin and lung diseases. For this purpose, we calculated the value of the water quality index [1] of the water.
The methods of assessing the quality of water differ in terms of their methodologies and input parameters [1]. The most common Water Quality Index Methods used are the National Sanitation Foundation Method, the Oregon Water Quality Index Method, the Weighted Arithmetic Water Quality Index Method, and the Canadian Council of Ministers of the Environment Water Quality Index Method [2]. We adopted the weighted arithmetic water quality index method in this research paper. We calculated important parameters, such as dissolved oxygen content, salinity, acidity and alkalinity, total suspended solids (TDS), and biochemical oxygen demand (BOD) and tabulated he results in a CSV file.
Currently, many problems are efficiently solved by machine learning algorithms. The most important algorithms are decision tree, support vector machine, regression, random forest, and clustering. The key behind the machine learning model is to learn from the data and build the model [2]. When new data are received, the output of the new data is predicted. The volume of data determines the accuracy of the predicted output. When the volume of data is high, only the model predicts the output as more accurate.

2. OBJECTIVES OF THE RESEARCH
- ❖ To collect periodic amounts of bed-well water from our surroundings
- ❖ The water quality parameters TDS, pH, COD, BOD, F, Ca, and Mg hardness were calculated.
- ❖ The water quality indices were determined by averaging all the parameters.
- ❖ Based on the Gini index, we construct a decision tree using the algorithm in Python.

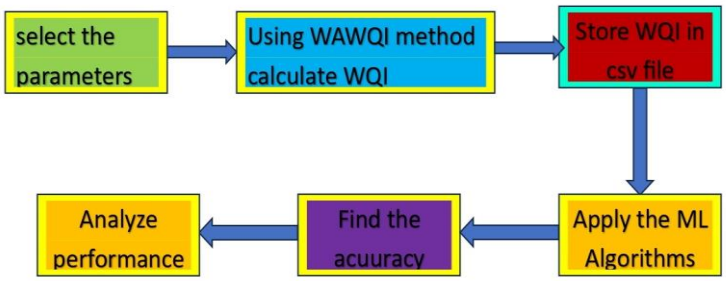
❖ The decision tree and K-nearest neighbor algorithms were used to determine the model performance.

❖ We developed models using Google Colab that use the K-nearest neighbor and decision tree algorithms to predict water quality in real-time.

3. RESEARCH METHODOLOGY

Random water samples were collected from several areas around our village. We collected water samples from various educational institutions, such as schools, colleges, and universities. We collected nearly 105 samples, and the physicochemical characteristics of the collected water samples were examined and reported. The flowchart of our methodology is shown in Figure 1.

Figure 1: WQI PROCESSING FLOW



3.1 Using the WAWQI Method to Calculate the WQI

Step 1: Calculate the values of various physicochemical water quality parameters.

Step 2: Find the proportionality constant K by using the formula $K = \mathbf{(1/(1/\sum^{n})}$

Step 3: To calculate a quality rating for the nth parameter (q_n)

where n=number of parameters

using formula $\mathbf{q_n = 100\ \{ (v_n - v_{io})/(s_n - v_{io})\}}$

v_n = Estimated value of the n^{th} parameter of the given sampling station.

v_{io}= Ideal value of the n-th parameter in pure water

s_n=Standard permissible value of the n^{th} parameter.

Step 4: Calculate the unit weight for the n^{th} parameter. $W_n=(k/s_n)$.

Step 5: Calculate the water quality index (WQI) using the formula, $WQI = ((\Sigma w_n * q_n)/\Sigma w_n)$

We calculated the water quality indices of our samples stored in our dataset. The pictorial form of this index refers to some of the samples given in Figure 2. Based on the water quality indices, the status of the water samples is tabulated in Table 1, and the water quality indices of our selected samples are tabulated in Table 2.

Table 1 Water quality index (WQI) and status of water quality

LEVEL OF WQI	STATUS OF WATER
from 0 to 25	High
from 26 to 50	Medium
from 51 to 75	Low
from 76 to 100	Very Low
greater than 100	Unsuitable to use

Table 2 Status of water quality indices

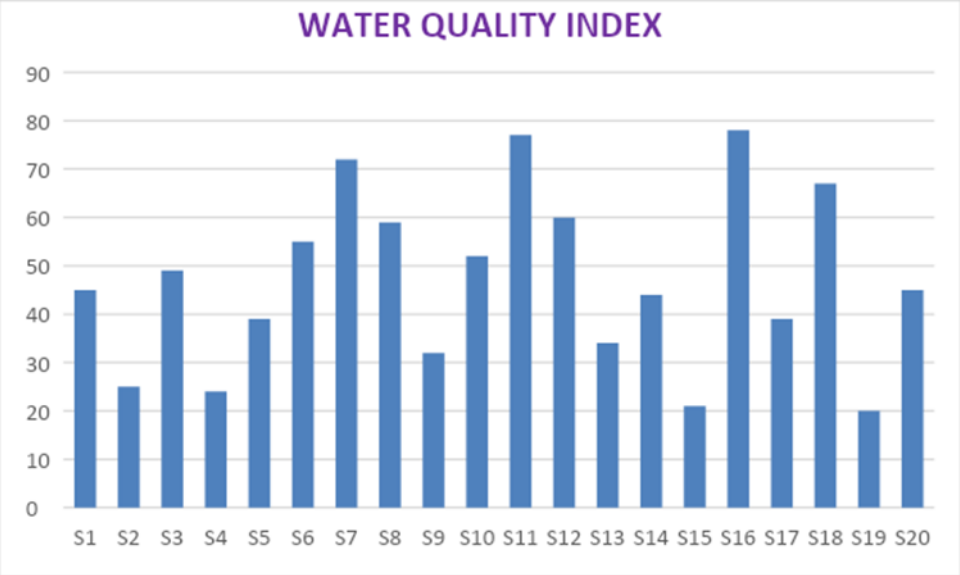
S.NO	SAMPLE NO	WQI	STATUS

ISSN: 2278-4632
Vol-14, Issue-2, No.03, February: 2024

1	S1	45	Medium
2	S2	25	High
3	S3	49	Medium
4	S4	24	High
5	S5	39	Medium
6	S6	55	Low
7	S7	72	Low
8	S8	59	Low
9	S9	32	Medium
10	S10	52	Low
11	S11	77	Very Low
12	S12	60	Low
13	S13	34	Medium
14	S14	44	Medium
15	S15	21	High
16	S16	78	Very Low
17	S17	39	Medium
18	S18	67	Low
19	S19	20	High
20	S20	45	Medium

Figure 2: Water quality indices

ISSN: 2278-4632
Vol-14, Issue-2, No.03, February: 2024



4. ANALYSIS AND DISCUSSION

A pictorial representation of our samples is given in Figure 2. These samples are categorized into training and test data from our three selected models.

4.1 Decision Tree

One of the quantile-supervised learning algorithms is the decision tree. This algorithm is mainly used for regression and classification tasks. The decision tree has different parts, such as branches, roots, internal nodes, and leaf nodes. By using the divide and conquer method only, decision tree searches to identify the root node within a tree. This process is continued until all the node's Gini [4] values are calculated using the entropy formula.

The salient features of decision tree algorithms

➔ They require less effort for data preprocessing.

➔ It does not require any normalization of the data.

➔ Missing values in the dataset do not affect the construction of the decision tree.

➔ We can easily obtain results from the decision tree model.

When this occurs, it is known as data fragmentation, and it can often lead to fragmentation overfitting. To reduce the complexity and prevent overfitting, pruning is usually employed; this is a process, in which branches that split features with low importance are removed.

Pruning is the process of removing connections from a network to increase the speed of inference and reduce its storage size. Pruning of a network deletes the unnecessary parameters from an overly parameterized network. The model fit can then be evaluated through the process of cross-validation. This classifier predicts more accurate results, particularly when the individual trees are uncorrelated.

Choosing the best attribute at each node

We must select the best attribute in each node among multiple ways such as information gain and Gini impurity. They help to evaluate the quality of each test condition and how well it will be able to classify samples into a class.

Entropy and Information Gain

Entropy is used to measure the uncertainty of data. It is an essential metric that helps to evaluate the quality of a model and its ability to make accurate predictions. Here, we used this entropy to determine the best split at each node. By using entropy only, we can construct more robust and accurate models. Information gain is related to entropy. It measures the impurity of the sample. It is defined by the following formula [7]:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

The entropy values lie between 0 and 1. The entropy value is zero when all the samples in the dataset belong to the same class. If half of the samples are classified under one class and the other half of the samples are in another class, then the entropy value is 1. To select the best feature to split on and find the optimal decision tree, the attribute with the smallest amount of entropy should be used. The difference in entropy before and after a split on a given attribute is represented by the information gain. The attribute that has the highest information gain will produce the best split asbecause it is doing the best job at classifying the training data according to its target classification.

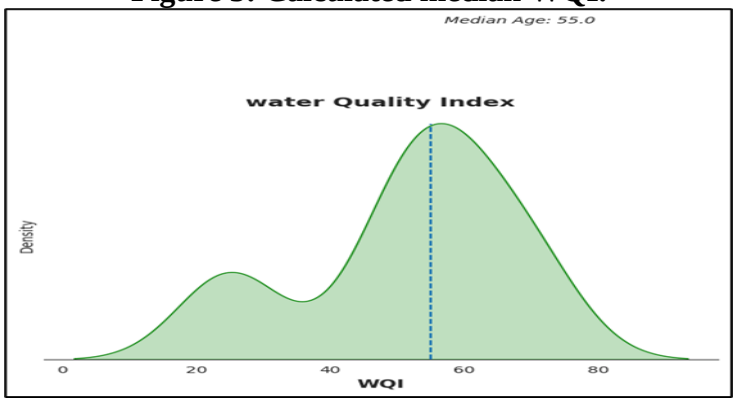
4.2 K-means clustering

Among several unsupervised machine learning algorithms, K-means clustering is one of the most effective. K-means clustering assigns data points to clusters based on which reference point is closest after constructing a centroid for the appropriate number of classes. Choosing the K value is the key point of the K-means algorithm. Here, we cover a common technique for choosing K in the machine learning K-means algorithm.

K-nearest neighbor algorithm steps

Step 1: Choose the number of clusters as K.
Step 2: Select random K points or centroids.
Step 3: Assign each data point to its closest centroid. It forms the predefined K clusters.
Step 4: Calculate a new centroid for each cluster.
Step 5: Take an average of samples from the same cluster.
Step- 5: Each data point is reassigned to the new closest centroid of each cluster.
Step 6: If no new reassignment occurs, the model is ready. Otherwise, go to step 4.

Figure 3: Calculated median WQI.



4.3 Naive Bayes

The naive Bayes algorithm is based on the Bayes theorem. It is also one of the simplest supervised learning algorithms. The naive Bayes classifier is a fast, accurate, and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets. Prior and posterior probabilities are used in this algorithm.– Figure 3 shows the median value of the water quality indices of our samples calculated using this algorithm. The steps used in the naive Bayes algorithm are listed below.

Ø– For the given class labels, calculate the prior probability.
Ø– Apply the Bayes formula, and find the posterior probability.
Ø The given input belongs to the class that has a higher probability.

5. Analysis and Discussion

We applied three algorithms, naive Bayes, K-nearest neighbor, and decision tree algorithms, to develop our classification model with our dataset as input. The WQI values of our samples were also calculated through the models. We utilized the naive Bayes, decision tree, and K-nearest neighbor classifiers. We obtained three different accuracies naive Bayes-high and decision tree-low) using the three classifiers shown in Table 3. From the results, we conclude that naive Bayes has the highest accuracy at 95%, while the decision tree has the lowest accuracy at 91%. Figure 3 displays the performance of our applied models.

Table 3: Model accuracy

S.NO	MODEL	ACCURACY (%)
1	Decision Tree	91
2	K-Nearest Neighbor	92
3	Naive Bayes	95

6. CONCLUSION

The performance of machine learning techniques such as the naive Bayes, decision tree, and K-nearest neighbor models in predicting the water quality indices of our surrounding educational institutions. The six important variables, pH, TC, DO, BOD, nitrate, and temperature, for calculating the water quality index were obtained from our dataset. We obtained the results by applying the three machine learning algorithms. We intimate the importance of water quality to the educational institutions of those who have low-value water. In the future, research will be carried out to construct models that combine the proposed methods with deep learning approaches to improve efficiency.

7. REFERENCES

1. Jain D, Shah S, Mehta H et al (2021) A Machine Learning Approach to Analyze Marine Life Sustainability. In: Proceedings of International Conference on Intelligent Computing, Information and Control Systems. Springer, pp 619–632
2. Clark RM, Hakim S, Ostfeld A (2011) Handbook of water and wastewater systems protection. In: Protecting Critical Infrastructure. Springer, pp 1–29. https://doi.org/10.1007/978-1-4614-0189-6
3. Hu Z, Zhang Y, Zhao Y et al (2019) a water quality prediction method based on the deep LSTM network considering correlation in smart mariculture. Sensors 19:1420
4. Zhou J, Wang Y, Xiao F et al (2018) Water quality prediction method based on IGRA and LSTM. Water 10:1148
5. Waqas M, Tu S, Halim Z et al (2022) the role of artificial intelligence and machine learning in wireless networks security: principle, practice and challenges. Artif Intell Rev 55:5215–5261. https://doi.org/10.1007/s10462-022-10143-2
6. Halim Z, Waqar M, Tahir M (2020) A machine learning-based investigation utilizing the in-text features for the identification of dominant emotion in an email. Knowl Based System 06443. https://doi.org/10.1016/j.knosys.2020.106443
7. Wu J, Wang Z (2022) A Hybrid Model for Water Quality Prediction Based on an Artificial Neural Network, Wavelet Transform, and Long Short-Term Memory. Water 14:61
8. Lee S, Lee D (2018) Improved prediction of harmful algal blooms in four Major South Korea's Rivers using deep learning models. Int J Environ Res Public Health 15:1322
9. Liu P, Wang J, Sangaiah AK et al (2019) Analysis and prediction of water quality using LSTM deep neural networks in IoT environment. Sustainability 11:2058
10. Hmoud Al-Adhaileh M, Waselallah Alsaade F (2021) Modeling and prediction of water quality by using artificial intelligence. Sustainability 13:4259
11. Bhardwaj D, Verma N (2017) Research paper on analyzing impact of various parameters on water quality index. Int J Adv Res Comput Sci 8(5):2496–498
12. Malek NHA, Wan Yaacob WF, Md Nasir SA, Shaadan N (2022) Prediction of Water Quality Classification of the Kelantan River Basin, Malaysia, Using Machine Learning Techniques. Water 14:1067
13. Slatnia A, Ladjal M, Ouali MA, Imed M (2022) Improving prediction and classification of water quality indices using hybrid machine learning algorithms with features selection analysis. In: Online International Symposium on Applied Mathematics and Engineering (ISAME22), vol 1. ISAME22, Istanbul-Turkey, pp 16–17
14. Deng T, Chau K-W, Duan H-F (2021) Machine learning based marine water quality prediction for coastal hydro environment management. J Environ Manage 284:112051

15. Boehm, A. B., S. B. Grant, J. H. Kim, S. L. Mowbray, C. D. McGee, C. D. Clark, D. M. Foley, and D. E. Wellman. 2002. Decadal and shorter period variability of surf zone water quality at Huntington Beach, California. Environ. Sci. Technol. 36:3885-3892. [PubMed] [Google Scholar]
16. Brenniman, G. R., S. H. Rosenberg, and R. L. Northrop. 1981. Microbial sampling variables and recreational water quality standards. Am. J. Public Health 71:283-289. [PMC free article] [PubMed] [Google Scholar]
17. Cabelli, V. J., A. P. Dufour, M. A. Levin, L. J. McCabe, and P. W. Haberman. 1979. Relationship of microbial indicators to health effects at marine bathing beaches. Am. J. Public Health 69:690-696. [PMC free article] [PubMed] [Google Scholar]
18. Cabelli, V. J., A. P. Dufour, L. J. McCabe, and M. A. Levin. 1983. A marine recreational water quality criterion consistent with indicator concepts and risk analysis. J. Water Pollut. Control Fed. 55:1306-1314. [Google Scholar]
19. Cabelli, V. J., A. P. Dufour, L. J. McCabe, and M. A. Levin. 1982. Swimming-associated gastroenteritis and water quality. Am. J. Epidemiol. 115:606-616. [PubMed] [Google Scholar]
20. Cheung, W. H. S., K. C. K. Chang, and R. P. S. Hung. 1991. Variations in microbial densities in beach waters and health-related assessment of bathing water quality. Epidemiol Infect. 106:329-344. [PMC free article] [PubMed] [Google Scholar]
21. Collison, P. 1998. Of bombers, radiologists, and cardiologists: time to ROC. Heart 80:215-217. [PubMed] [Google Scholar]
22. DeLong, E. R., D. M. DeLong, and D. L. Clarke-Pearson. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837-845. [PubMed] [Google Scholar]
23. Dufour, A. P. 1984. Bacterial indicators of recreational water quality. Can. J. Public Health 75:49-56. [PubMed] [Google Scholar]
24. Haile, R. W., J. S. Witte, M. Gold, R. Cressey, C. McGee, R. C. Milikan, A. Glasser, N. Harawa, C. Ervin, P. Harmon, J. Harper, J. Dermand, J. Alamillo, K. Barrett, M. Nides, and G.-Y. Wang. 1999. The health effects of swimming in ocean water contaminated by storm drain runoff. Epidemiology 10:355-363. [PubMed] [Google Scholar]
25. Messer, J. W., and A. P. Dufour. 1998. A rapid, specific membrane filtration procedure for enumeration of enterococci in recreational waters. Appl. Environ. Microbiol. 64:678-680. [PMC free article] [PubMed] [Google Scholar]
26. Metz, C. 1978. Basic principles of ROC analysis. Semin. Nucl. Med. 8:283-. [PubMed] [Google Scholar]
27. U.S. Environmental Protection Agency. 1986. Ambient water quality criteria for bacteria. EPA440/5-84-002. U.S. EPA Office of Water, Washington, D.C.
28. U.S. Environmental Protection Agency. 2000. Implementation guidance for ambient water quality criteria for bacteria—1986. Draft EPA-823-D-00-001. U.S. EPA Office of Water, Washington, D.C.
29. U.S. Environmental Protection Agency. 1999. Review of potential modeling tools and approaches to support the BEACH program. 823-R-99-002. U.S. EPA Office of Science and Technology, Washington, D.C.

