IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

WATER QUALITY PREDICTION USING MACHINE LEARNING ALGORITHMS

V. Queen Jemila[1], M. Dhanalakshmi [2] and M.Amutha[3]

[1]Assistant Professor of Computer Applications (PG)

[2,3]Associate Professors of Chemistry

[1, 2,3]V.V.Vanniaperumal College for Women Virudhunagar

ABSTRACT

The main task of our research work is to calculate the Water Quality Index of bore water in our surrounding educational institutions using three machine learning algorithms. Our research work differentiates from other work by choosing Decision Tree, K-Nearest Neighbor, and Naive Bayes Algorithms and their accuracy. We collected water samples from various resources and calculated the six important factors: salinity, total suspended solids (TDS), dissolved oxygen (DO), acidity and alkalinity (pH), and biochemical oxygen demand (BOD). The water quality parameters were analyzed using standard chemical methods. Using these parameters we created our own dataset and the dataset is given as our chosen algorithms train and test data. Finally, we got the WQI value with three different accuracies.

Keywords: Water quality Index, Decision Tree, KNN, and Naive Bayes

1. Introduction

Water is the major important resource of mankind. In everyday life, people use water frequently. It is one of the most needs of human beings to avoid skin and lung diseases, we must use good-quality water. For this purpose, we have to calculate the value of the Water Quality Index of our daily usage water,

Water quality assessment methods differ in their methodology as well as their input parameters.[1].The most frequent Water Quality Index Methods are the National Sanitation Foundation Method, Oregon Water Quality Index Method, Weighted Arithmetic Water Quality Index Method, and the Canadian Council of Ministers of the Environment Water Quality Index Method CCME-WQI)...In this research paper, we adopted the Weighted Arithmetic Water Quality Index Method. We calculated the important parameters: salinity, total suspended solids (TDS), dissolved oxygen (DO), acidity and alkalinity (pH), and biochemical oxygen demand (BOD) and tabulated as a CSV file.

Nowadays many problems are solved by machine learning algorithms in an efficient way. The most important algorithms are Regression, Decision Tree, Random forest, clustering, and Support Vector Machine. A Machine Learning model system learns from data and builds the model.[2]. When it receives new data, it predicts the output for new data. The accuracy of predicted output depends upon the volume of data. If the volume of data is high only the model predicts the output as more accurate.

2. Objectives of the Research

- To collect periodically bore-well water from of our surroundings
- To calculate water quality parameters TDS, pH, COD, BOD, F, Ca and Mg hardness.
- To make use of the Decision Tree algorithm, construct a decision tree based on Gini Index in python.
- To find out the Water Quality Index (WQI) by taking an average of all the parameters.
- Decision Tree, K-Nearest Neighbor and Naïve Bayes algorithms are used to optimize model performance.
- The proposed model approaches: develop a software application that uses the Decision Tree, K-Nearest Neighbor and Naïve Bayes algorithms to predict water quality in real time.

3. Research Methodology

Random water samples are taken from several areas around our village. We have collected water samples from various Educational Institutions like Schools, Colleges and Universities. Nearly we collected 94 samples and physico-chemical characteristics of the collected water samples were examined and reported.

3.1 Methodology in Calculating WQI Using WAWQI Method

Step 1: T find out the various physico-chemical water quality parameters.

Step 2: Calculate Proportionality constant K by using the formula K $= (1/(1/ \sum^{n})$ i

Step 3: calculate a quality rating for the nth parameter (q_n) where there are n 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 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 Water Quality Index (WQI) using formula, WQI = $((\Sigma w_n * q_n)/\Sigma w_n)$

Table 1 Water Quality Index (WQI) and Status of water quality

LEVEL OF WQI	STATUS OF WATER
0 -25	Excellent
26-50	Good
51-75	Poor
76-100	Very Poor
>100	Unsuitable for usage

Table 2 Water Quality Index of our samples

S.NO	SAMPLE NO	WQI	STATUS
1	S1	45	GOOD
2	S2	25	EXCELLENT
3	S3	49	GOOD
4	S4	24	EXCELLENT
5	S5	39	GOOD
6	S6	55	POOR
7	S7	72	VERY POOR
8	S8	59	POOR
9	S9	32	GOOD
10	S10	52	POOR
11	S11	77	VERY POOR
12	S12	60	POOR
13	S13	34	GOOD
14	S14	44	GOOD
15	S15	21	EXCELLENT
16	S16	78	VERY POOR
17	S17	39	GOOD
18	S18	67	POOR

19	S19	20	EXCELLENT
20	S20	45	GOOD

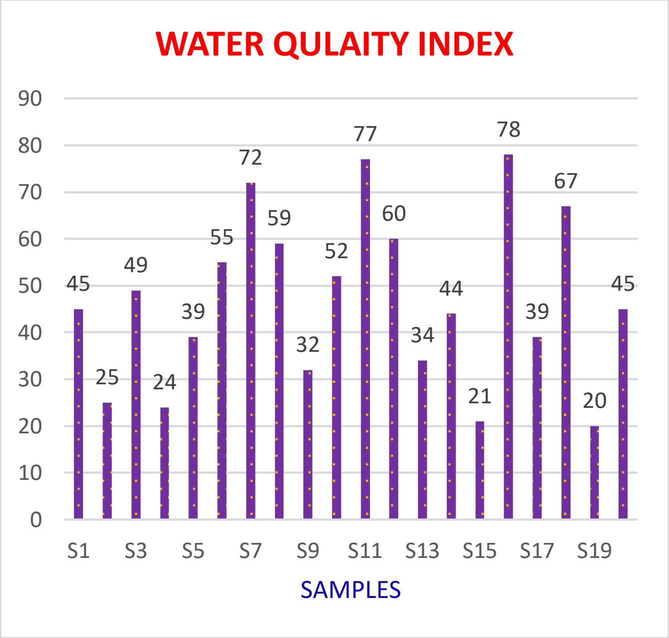

Figure 1: Water Quality Index

4. Analysis and Discussion

4.1 Decision Tree

A decision tree is a quantile supervised learning algorithm. It is used for both classification as well as regression tasks. It has a hierarchical tree structure, which consists of a root node, branches, internal nodes, and leaf nodes. Decision tree uses divide-and-conquer method by conducting a greedy search to identify the root node within a tree. This process is continued until all node's gini[4] values are calculated using Entropy.

The important features of decision tree algorithms are it requires less effort for data preprocessing, It doesn't require any normalization of data, Missing values in dataset does not affect the construction of Decision Tree and We can easily explain the Decision Tree model. When this occurs, it is known as data fragmentation, and it can often lead to overfitting. To reduce the complexity and prevent overfitting, pruning is usually employed; this is a process, which removes branches that split on features with low importance. Pruning is the process of removing connections from a network to increase the speed inference and reduce its storage size. Pruning of a network deletes the unneeded parameters from an overly parameterized network. The model's fit can then be evaluated through the process of cross-validation. Another way that decision trees can maintain their accuracy is by forming an ensemble via a random forest algorithm; this classifier predicts more accurate results, particularly when the individual trees are uncorrelated with each other.

How to choose the best attribute at each node?

There are multiple ways to select the best attribute at each node, we use 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 in a dataset. It is an essential metric that helps to evaluate the quality of a model and its ability to make accurate predictions. Here we use this entropy to determine the best split at each node. By understanding the concept of entropy, data scientists and machine learning engineers can build more robust and accurate models. Information gain is related to Entropy. It measures the impurity of the sample values. It is defined by the following formula [7]

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

Entropy values can fall between 0 and 1. If all samples in the data set, S, belong to one class, then entropy will equal zero. If half of the samples are classified as one class and the other half are in another class, entropy will be at its highest at 1. In order 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. Information gain represents the difference in entropy before and after a split on a given attribute. The attribute with the highest information gain will produce the best split as it’s doing the best job at classifying the training data according to its target classification.

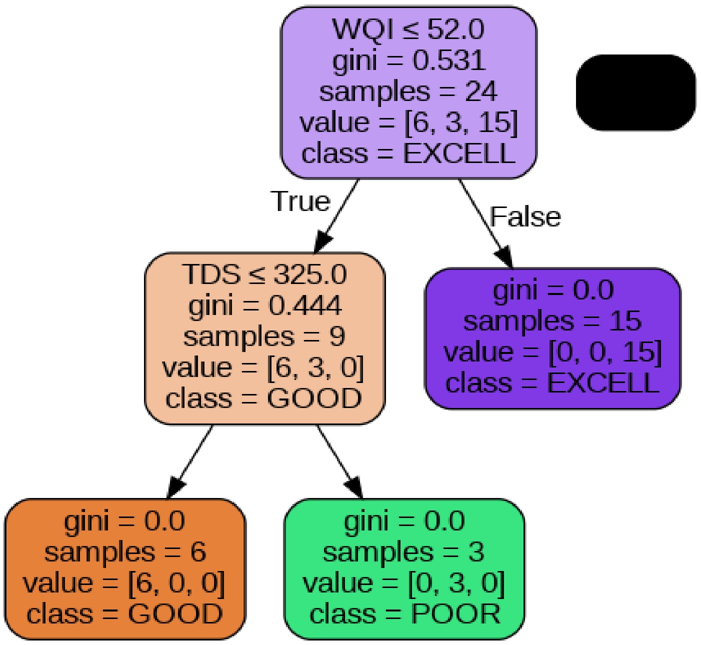

Figure 2: Decision Tree

4.2 K-means clustering

K-means clustering is one of the most effective unsupervised machine learning algorithms. 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 value of K is one of the key points of the K-means algorithm. Here, we've covered a common technique for choosing K in the machine learning K-means algorithm. Numerous applications of the technique exist, such as document classification, image segmentation, and recommendation engines.

Steps for K-Nearest Neighbor

Step 1: First, Choose the number of clusters as K.

Step 2: Select random K points or centroids. The centroids may not be from the input dataset.

Step 3: Assign each data point to its closest centroid. It will form the predefined K clusters.

Step 4: Calculate a new centroid of each cluster, taking an average of samples belonging to the same cluster.

Step-5: Repeat step 3, which means reassigning each data point to the new closest centroid of each cluster.

Step 6: If no new reassignment occurs, then the model is ready. Else, go to step 4.

4.3 Naïve Bayes Algorithm

The Bayes Theorem is the foundation of the probabilistic machine learning method known as Naive Bayes, which is utilized for a variety of classification problems. We shall learn about the Naive Bayes algorithm in this essay. A straightforward mathematical procedure for computing conditional probabilities is known as Bayes' Theorem. The probability of an event happening given that another event has already happened is known as conditional probability.

$$P(h|D)=(P(D|h).P(h))/P(D)$$

5. Analysis and Discussion

We have applied three algorithms for developing our classification model based on our dataset Using five classifiers, we have utilized the Decision Tree, K-Nearest Neighbor, and Naïve Bayes Model (BT). Various levels of accuracy with the WQI values have been achieved by the classifiers we have examined for our classification. We have applied each of the classifiers mentioned before to our dataset and after applying those algorithms, we have achieved the highest accuracy, and the lowest accuracy and have found out the average accuracy by utilizing those algorithms. Among the Decision Trees, K-Nearest Neighbor, and Naïve Bayes, Naïve Bayes R has the highest accuracy at 99%, while Decision Tree has the lowest accuracy at 97%. The highest WQI value has also been obtained through the Models. Figure 3 shows the classification model performances of our applied models.

Table 3: Models Accuracy

S.NO	MODEL	ACCURACY (%)
1	Decision Tree	94
2	K-Nearest Neighbor	95
3	Naïve Bayes	97

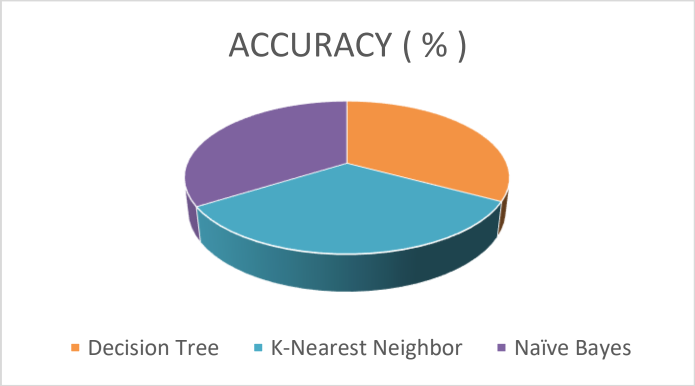

Figure 3 : Accuracy of Models

6. Conclusion

The performance of machine learning techniques such as Decision Tree, K-Nearest Neighbor, and Naïve Bayes Model to predict the water quality components of an Indian water quality dataset was evaluated in this work. The most well-known dataset variables, such as BOD, DO, TC, Nitrate, pH, and Temp, were obtained for this purpose. The findings revealed that the applied models performed well in forecasting water quality parameters; however, the greatest performance was linked with the Naïve Bayes with Accuracy Upper. Further research will be done to build models that combine the proposed method with other techniques and deep learning approaches to improve the efficacy of the selection process.

7. References

1.	entina-Andreea Călmuc 1*, Mădălina Călmuc 1, Maria Cătălina Ţopa1, aela Timofti 1, Cătălina Iticescu 1, Lucian P. Georgescu 1 various methods for ulating the water quality index
2.	Mehedi Hassan[1, *], Md. Mahedi Hassan[2], Laboni Akter[3], Md. Mushfiqur Rahman[4], Sadika Zaman[1], Khan Md. Hasib[5], Nusrat Jahan[6], Raisun Nasa Smrity[2], Jerin Farhana[7], M. Raihan[1], Swarnali Mollick[8] - Efficient Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms
3.	Amir Hamzeh Haghiabi;Ali Heidar Nasrolahi;Abbas Parsaie Water quality prediction using machine learning methods
4.	T.Suryakanthi, Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm, International Journal of Advanced Computer Science and Applications January 2020
5.	Abazi A.M.S., Durmishi B.H., Sallaku F.S., Cadraku H.S., Fetoshi O.B., Ymeri P.H., Bytyci P.S. Assessment of water quality of sitnica river by using water quality index (WQI) RASAYAN J. Chem. 2020;13(1):146–159. [Google Scholar]
6.	Badan Pengendalian Lingkungan Hidup Kabupaten Bandung . Pemerintah Kabupaten Bandung Provinsi Jawa Barat: Bandung; Indonesia: 2015. Laporan Status Lingkungan Hidup Daerah Kabupaten Bandung.
7.	Cude C.G. Oregon water quality index: a tool for evaluating water quality management effectiveness. *J. Am. Water Resour. Assoc.* 2001;37(1):125–
8.	Darvishi G., Kootenaei F.G., Ramezani M., Lotfi E., Asghamia H. Comparative investigation of river water quality by OWQI, NSFWQI, and wilcox indexes (case study: the Talar River – Iran) *Arch. Environ. Protect.* 2016;42(1):41–48]
9.	Davies J. Application and test of the Canadian water quality index for assessing changes in water quality in lakes and rivers of central north America. *Lake Reservoir Manag.* 2016;22(4):308–320]