Water Quality Prediction Using Decision Tree and KNN

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 work of our research aims to find out the Water Quality Index of bore water in our surrounding educational institutions using two machine learning algorithms. Our research work differentiates from other work by choosing Decision Tree, K-Nearest Neighbor 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). Using efficient chemical methods, the quality parameters of water were examined. We created our dataset by utilizing these metrics, and the dataset is given as our chosen algorithm's training and testing data. Finally, we got the WQI value with two different accuracies.

***Keywords**: Water quality Index, Decision Tree, KNN, Gini Index.*

I. INTRODUCTION

One of the major resources for human beings is water.. People use water frequently in their day-to-day lives. We should use pure water to avoid skin and lung diseases. For this purpose, we have calculated the value of the Water Quality Index[1] of water.

The methods of quality assessment of water differ in their methodologies as well as their input parameters.[1].The most frequent Water Quality Index Methods 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 the important parameters such as dissolved oxygen, salinity, acidity & alkalinity, total suspended solids (TDS), and biochemical oxygen demand (BOD) oxygen and tabulated it as a CSV file.

Nowadays 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 data and build the model.[2]. When it receives new data, it predicts the output for new data. The Volume of data determines the accuracy of the predicted output. If the volume of data is high only the model predicts the output as more accurate.

II. OBJECTIVES OF THE RESEARCH

- To collect periodically bore-well water from our surroundings
- To calculate water quality parameters TDS, pH, COD, BOD, F, Ca, and Mg hardness.
- To find out the Water Quality Index by taking an average of all the parameters.
- Based on the Gini Index Value, we construct a Decision Tree using its algorithm in Python.
- Decision Tree and K-Nearest Neighbor algorithms are used to find out the model performance.
- We developed a model using Python that uses the K-nearest neighbor and Decision Tree algorithms to predict water quality in real-time.

III. RESEARCH METHODOLOGY

Random water samples are collected from several areas around our village. We have collected water samples from various Educational Institutions like Schools, Colleges, and Universities. Nearly we collected 105 samples and physico-chemical characteristics of the collected water samples were examined and reported.

- *Using WAWQI Method to Calculate WQI*

Step 1: To calculate the value of various physico-chemical water quality parameters.
Step 2: To 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 Water Quality Index (WQI) using the formula, WQI = ((Σw_n* q_n)/Σw_n)

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 index

S.NO	SAMPLE NO	WQI	STATUS
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

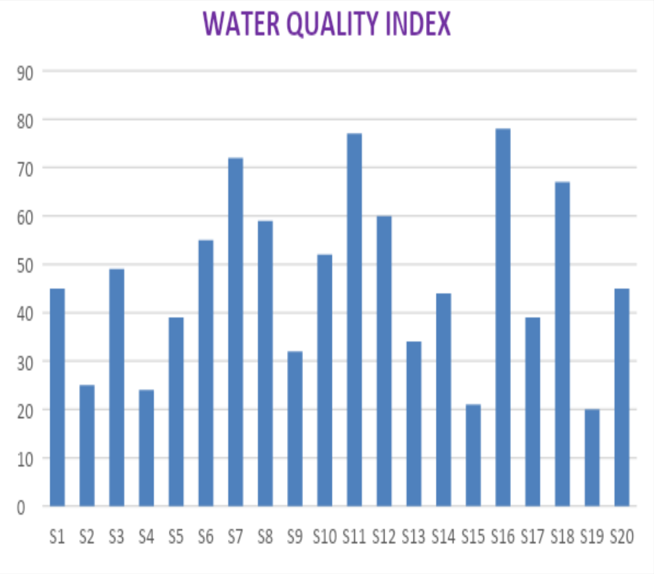

Fig 1: Water Quality Index

IV. ANALYSIS AND DISCUSSION

A. *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 like branches, roots, internal nodes, and leaf nodes. By using divide and conquer method only, Decision Tree searches to identify the root node within a tree. This process is continued until all 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 doesn’t require any normalization of data.
- Missing values in the dataset do not affect the construction of the Decision Tree.
- We can easily get the result from 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 of 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. This classifier predicts more accurate results, particularly when the individual trees are uncorrelated.

➢ *To choose the best attribute at each node*

We have to select the best attribute in each node among multiple ways like 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 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 lie between 0 and 1. The entropy value is zero when all samples in the data set, S, 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 Information gain. The attribute that has 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.

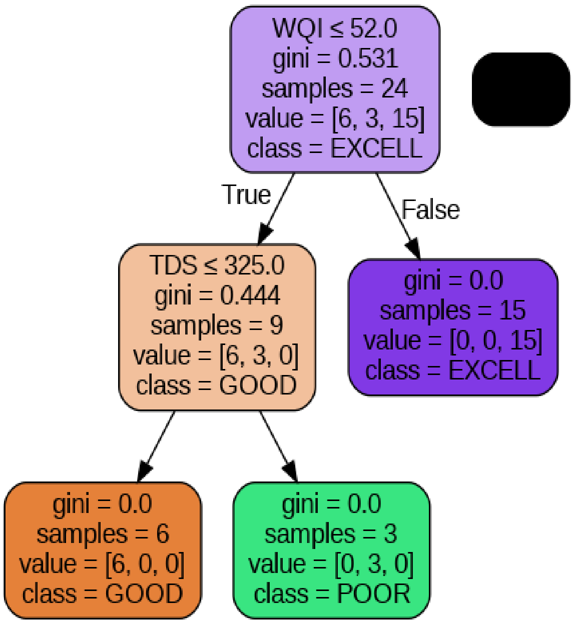

Fig 2: Decision Tree

B. K-means Clustering

Among several unsupervised machine learning algorithms, K-means clustering is one of the most effective ones. 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've covered 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 of each cluster.
Step 5: To take an average of samples from the same cluster.
Step-5: To reassign each data point to the new closest centroid of each cluster.
Step 6: If no new reassignment occurs, the model is ready. Else, go to step 4.

V. ANALYSIS AND DISCUSSION

We applied two algorithms K-Nearest Neighbor and Decision Tree for developing our classification model with our dataset as input. The WQI values of our samples were also calculated through the Models. We have utilized the Decision Tree and K-Nearest Neighbor classifiers. We got two different accuracies (K-Nearest Neighbor-High, Decision Tree-Low) using these two classifiers. From the result, we conclude that KNN has the highest accuracy at 95%, while the Decision Tree has the lowest accuracy at 93%. Figure 3 displays the performance of our applied models.

Table 3: Models Accuracy

S.NO	MODEL	ACCURACY (%)
1	Decision Tree	93
2	K-Nearest Neighbor	95

VI. CONCLUSION

The performance of machine learning techniques such as the Decision Tree and K-Nearest Neighbor Model to predict the water quality index of our surrounding educational institutions. The six important variables pH, TC, DO, BOD, Nitrate, and Temp for calculating the Water Quality Index were obtained from our dataset. We got the result by applying the two machine learning algorithms. We intimate the importance of water quality to the educational institutions of those who have bore water with low value. In the future, research will be carried out to build models that combine the proposed methods with deep learning approaches to improve efficiency.

REFERENCES

[1]. Valentina-Andreea Călmuc 1*, Mădălina Călmuc 1, Maria Cătălina Ţopa1, ela Timofti 1, Cătălina Iticescu 1, Lucian P. Georgescu 1 various methods for calculating the water quality index

[2]. Mehedi Hassan1, *, Md. Mahedi Hassan2, Laboni Akter3, Md. Mushfiqur Rahman4, Sadika Zaman1, Khan Md. Hasib5, Nusrat Jahan6, Raisun Nasa Smrity2, Jerin Farhana7, M. Raihan1, Swarnali Mollick8 - 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]