ISSN No:-2456-2165

Detrimental Object Detection in Water using Machine Learning: A Review

Rinkal Chauhan*, Bela Shrimali
LDRP Institute of Technology and Research,
KSV, SVKM, Gandhinagar, India

Abstract Water sources are often polluted due to human intervention. Water pollution may be classified according to its quality, which is governed by factors such as pH, turbidity, the conductivity of dissolved oxygen (DO), nitrate, temperature, and biological oxygen demand (BOD). This research compares water quality categorization methods that use machine learning techniques, namely SVM, Decision Tree and Naïve Bayes. The characteristics considered to determine the quality of water are:

pH, DO, BOD and electrical conductivity. Classification models are formed based on the numerical water quality index (WAWQI). After evaluating the results, With an accuracy of 98.50%, the decision tree approach was determined to be a better categorization model.

Keywords:- *Classification model; decision tree; support vector machine; naïve bayes; water quality index.*

I. INTRODUCTION

Water resources play an important and indispensable role in the lives of the people. The main water sources are rivers, lakes and ponds that provide adequate water for irrigation, electricitysupply, traffic and daily-life needs. To meet these needs, it must be clean and free from contamination. In a developing country like india, due to industrialization, all types of pollution growing swiftly. Recent trends show that water quality particularly in India has dropped to a very large extent, resulting in large amounts of water that are not suitable for use.

According to World Health Organization (WHO) records, water challenges such as lack of clean and safe water for domestic purposes, increasing urban pollution and water scarcity are increasing at an alarming rate. dynamic [1]. As a result, the decade 2018-2028 has been declared by the United Nations General Assembly as the International Decade of Action, “Water for Sustainable Development” [2]. Increasing population rate is another factor causing water scarcity. Water quality is controlled by its certain constituent parameters, so when wastewater discharges into water, the concentration of these water parameters changes resulting in a decrease in waterquality.

In recent years, many investigative efforts have been made in water quality monitoring and a variety of water quality models have emerged. Conventional methods include manual data collection and statistical evaluation. GHOST. Tirabassi et al. developed a statistical model to predict water quality without reference to chemical, biological and physical relationships [3]. In this paper, the black box concept is used, i.e. with a known input, a relatively reliable output can be predicted. According to Gaganjot Kaur Kang et al., when there is a large amount of data, big data analysis can be applied [4]. One challenge encountered in this method is the accuracy of the water quality assessment and prediction model. H.C. Guo et al. propose a random water quality prediction system established to reveal the hazard characteristics of many properties based on Kalman filtering and self-adaptation techniques. The system predicted the BOD and DO levels of the Yilou River [5].

Water quality is determined by different levels of different parameters. Amit Sinha et al. proposed a fuzzy model that inputthree parameters (conductivity, pH and hardness) and the modelwas then simulated using MATLAB [6]. The dataset is generated from multiple water samples collected from different regions of Uttar Pradesh. Artificial Neural Network (ANN) algorithms have been widely used to predict water quality, an example of an algorithm is the back-propagation algorithm (BP). One problem in this BP algorithm is the low percentage of accuracy, so an improved artificial bee colony (IABC) algorithm was presented. When comparing the two algorithms, the model increased its accuracy by 25%. In this algorithm, the connectionweight values between the network layers and the threshold value of each layer are improved first. Amir Hamzeh Haghiabi et al. propose a study comparing the performance of data processing group (GMDH), ANN and SVM methods to predict water quality of the Tireh River in southwestern Iran [9]. By analyzing the results of the three algorithms, the SVM model has a better performance in terms of accuracy. Wang Xuan et al. propose a proposal to solve the problem of classification, non- linearity and incomplete data prediction using SVM. However, the practicality of SVM suffers due to the difficulty of choosing appropriate SVM parameters. This paper presents the combination of SVM and swarm optimization to decide the parameters without SVM for better model accuracy [7]. Salisu Yusuf Muhammad et al. proposed a suitable classification model based on machine learning techniques [8].

Comparison of five classification algorithms such as Naïve Bayes, K star, Bagging, J48 and Conjunctiva rule was performedto find the important factors that help to classify the water quality of Kinta River, Perak Malaysia. Among the five models, Lazy model using K Star algorithm is considered the best algorithm with 86.67% accuracy.

Four sections make up the organization of the paper. The resources and techniques employed for the research analysis are presented in Part II. In part III, the experiment's findings are discussed, and in section IV, the paper is concluded.

ISSN No:-2456-2165

II. MATERIALS AND METHOD OF ANALYSIS

A. Experimental Dataset

Two data sets were reviewed for water quality testing in the proposed facilities. The first dataset covers twenty-eightdifferent water quality parameters or characteristics of the Narmada River, which flows through the state of Madhya Pradesh, during 2017-2018 [14]. Values are taken monthly at different stations on the river. Some examples of properties are pH, chloride, BOD, potassium, nitrate, DO, conductivity, etc. The second dataset is aggregated historical water quality data for selected locations in India. The values of the parameters in each column are values for the years 2003 to 2014 [15]. The dataset includes eight parameters such as total coliform count, temperature, conductivity, fecal coliform, BOD, DO, pH and nitrate. There are approximately 1991 row values, some of which are invalid values that will be discarded during thisprocess. Both datasets are provided by Indian government websites.

B. Water quality parameters

Water quality depends on physical, chemical and biological factors of water content. Changing the value of parameters such as turbidity, temperature, biological oxygen demand, dissolved oxygen, conductivity, nitrate and pH leads to changes in water quality. Each parameter has a maximum allowable level defined by WHO and ICMR. Based on the availability of data, four water features were selected for the proposed study from the two datasets as presented in Table I [10].

Parameter	Standard values	Unit weights
pH	6.5-8.5	0.2190
BOD	5 mg/L	0.3723
DO	5 mg/L	0.3723
EC	250 S/cm	0.3710

Table 1: Standards for Water Quality Parameters

The water utilities industry has developed a set of standards for water quality parameters. These standards are intended to provide guidance for the development of safe drinking water programs, with the goal of protecting customers' health. The following is a summary of these standards.

- *pH:*

One of the most crucial characteristics of water quality is this. The appropriate pH range, according to WHO guidelines, is between 6.5 and 8.5. When the pH is below 6.5, water loses its ability to form vitamins and minerals in the body, and when the pH is above 8.5, it causes skin irritation and the water tastes salty. Aquatic life cannot survive when in range.

- *Biochemical Oxygen Demand*

BOD is the oxygen demand determination for industrial and domestic waste stabilization. These wastewaters deposited in rivers pollute water quality that can be determined by BOD. 3 mg is the maximum acceptable limit for BOD. According to WHO, BOD should not exceed 6 mg.

- *Dissolved Oxygen (DO)*

DO indicates changes caused by aerobic and anaerobic processes and also provides information on river conditions.

A good range for DO is 5 to 14.6 mg depending on temperature,altitude and salinity.

- *Electrical Conductivity (EC)*

A measure of the ability to conduct electricity through water.

As the number of dissolved solids and inorganic chemicals increases, so does the electrical conductivity. Conductivity measurements are therefore used to determine the amount of these chemicals. Maximum allowed value is 250.

C. Supporting Vector Machine

In machine learning, there are two types of learning: a supervised learning type and an unsupervised learning type. SVM is a type of supervised learning model that is combined with data analysis for classification or regression. Support Vector Classifier (SVC) based on SVM, to classify data into two or more classes. This algorithm was proposed by V. Vapnik and his team. SVC is a discriminant classifier well defined by a partition hyperplane. Given the labeled training data, the algorithm generates the best possible hyperplane that groups thenew instances. Simply put, given a set of training data, each of which has been marked in one of two classes, the SVM algorithm trains the model in such a way that a new example arrives, then it is classified into either group as shown in Fig. first.

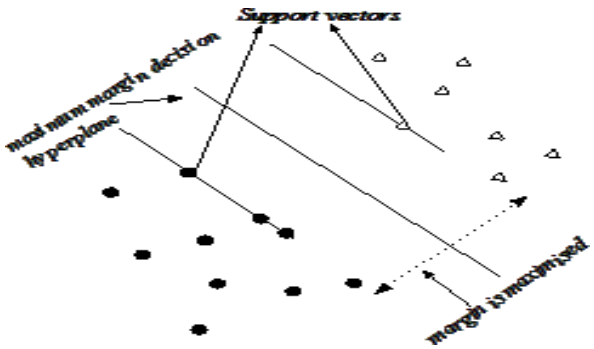

Fig. 1: Support Vector Machine: Linearly Separable Data The Support Vector Machine is a machine learning technique that maximizes the margin between a set of training data and a set of targetlabels

The advantages of SVM are: Can be used for bothlinear and non-linear data classification.

- It is also memory efficient as it uses a subset of the training data known as support vectors.
- For classification, the data can be divided into multiple classes if desired.

SVC has many core features determined from classificationdata. Some main features of the kernel are [17]:

ISSN No:-2456-2165

I. Linear kernel function: $K(x_i, x_j) = x_i^T x_j$ (1)

II. Polynomial Kernel function: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$ (2)

III. Radial Basis function (RBF):

$$(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right), \gamma > 0 \quad (3)$$

IV. Sigmoid Kernel: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$ (4)

Where the inputs are xi and xj, and γ is the regularization factor. The model's efficiency may be increased by selecting the suitable kernel and various parameters such as γ, C and ε

D. Decision Tree Classifier

A decision tree classifier (DTC) is a classification algorithm that further divides the data set as in a recursive algorithm intosmaller sets based on several tests performed at each tree node.

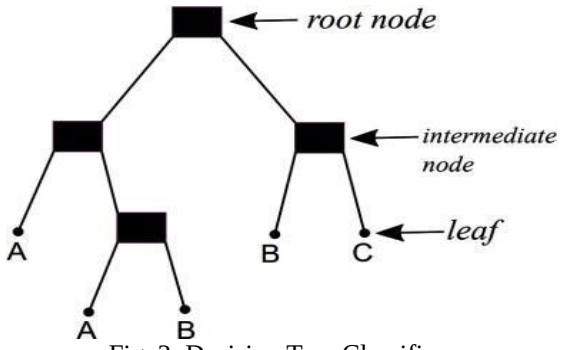

Fig. 2: Decision Tree Classifier:

The Decision Tree Classifier is a machine learning algorithm that is used to classify data. It can be used to perform classification and regression tasks. The classification task involves predicting whether a specific instance belongs to one of the categories or classes in the training set. Regression tasks involve predicting the value for one or more features for each new instance in the training set.

The tree in the DTC consists of a root node (original dataset), intermediate nodes (split datasets), and leaf (final sets of data) as represented in Fig.2. Each tree comprises one parent node andtwo or more child nodes. Decision trees have more advantages when compared to traditional classifying methods based on maximum livelihood theory concepts. The decision treeclassification is non-parametric, and it does not require assumptions on the distribution of input data. Finally, the structure and framework of the decision tree are simple and easy to interpret in the classification. Additional advantages are it needs minimal knowledge for data preparation and also achieves good outcomes for large datasets.

The three main algorithms used are Iterative Dichotomiser 3, C4.5 and the CART algorithm. Among these, the most commonly used algorithm is the CART (Classifier and Regression Tree) algorithm, introduced by Breiman et al. [11]. This algorithm divides a node into two nodes iteratively based on a predictor variable until the resulting nodes are homogeneous enough for the process to complete. Thus, these identity nodes represent class labels, and intermediate nodes arethe features that lead to class labels. However, the separation criterion corresponds to an increase in the purity of a node. There are three types of decomposition criteria, namely Gini criterion, Twoing criterion and Ordered Twoing criterion. Some stopping rules for node splitting are that if the node becomes pure, the tree depth reaches the user-defined tree depth and the node size is larger than the user-defined size [11]. The scikit- learning package in python uses an enhanced version of Breiman's CART algorithm.

➢ *Naive Bayes Classifier*

Naive Bayes classification is also an example of supervised learning. This is one of the most efficient algorithms. The basic premise of Naive Bayes is that each feature is independent and of equal importance. Naive Bayes relies onBayes' probability theorem [16]:

$$P(h|X) = \frac{P(h|Y)P(Y)}{P(X)} \quad (5)$$

Where X is the data set, h is the hypothesis such that X falls into a certain class C, and P (h|X) is the posterior probabilityof X [14].

Assume a set of n samples S = {S1, S2,Sn}. where the data Si have m-dimensional feature vectors X = {X1, X2, ...
...
Xm}. This forms the training data set. Also let be the numberof classes k, c = {c1,c2, ck} and each sample belongs to

$$P(C_i|X) = \prod_{t=1}^{n} P(X_t|C_i) \quad (6)$$

one of these classes. By introducing an additional data sample X, predict the class to which this particular data sample belongs using the highest conditional probability P(Ci|X),i =1,2 k I can do it.

Using Assumptions of Independence Between Traits or Attributes:

Where Xt are the values for the attributes of X. Naive Bayes uses a density function such as normal or gauss, log-normal, Poisson, and gamma to calculate the P(Xt|Ci) value for each attribute. The Scikit-learning package in python software uses Gaussian distribution and the equation is:

$$P(X_t|C_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{(X_t-\mu_i)^2}{2\sigma_i^2}} \quad (7)$$

Where $-\infty < x, \mu < +\infty, \sigma > 0, \mu$ is mean, σ is the standard deviation.

ISSN No:-2456-2165

E. Determination of Water Quality Index

There are four main types of water: water. Surface water, groundwater, wastewater, rainwater. Various methods are available to determine the Water Quality Index (WQI) of different types of water. One method for rainwater is the spatial distribution of WQI. Wastewater Quality Index (WWQI) can bedetermined using a variety of methods. B. Total, Health, and Acceptance WWQI. It can be used in various aggregate functions such as arithmetic and geometry. Also, the WQI of waste water can be obtained from the arithmetically weighted WQI (WAWQI). The National Sanitation Foundation WQI, Canadian Council of Environment Ministers WQI and WAWQIare used to find his WQI on and above the ground. WQI focuses on surface water as this study is relevant to surface water. Comparing among the above three methods, WAWQI showed superior results [12]. WAWQI was proposed by Horton in his 1965 and later by Brown et al. introduced. Unit weight (Wi) is inversely proportional to the proposed standard value for the corresponding property. Values for each parameter are shown in Table 1:

$$W_i = K\sum\frac{1}{S_{standard}} \quad (8)$$

The proportionality factor K is determined by:

$$K = \frac{1}{\sum\frac{1}{S_1}+\frac{1}{S_2}+\cdots+\frac{1}{S_n}} \quad (9)$$

WQI	Class
0-25	Excellent
26-50	Good
51-75	Fair
76-100	Poor
Greater than 100	Unfit for consumption

Table 2: Classification Of Water Ased On Wawqi

There are three main types of water quality: Non-point source pollution refers to the discharge of pollutants from a point source, such as a factory or wastewater treatment plant. Point source pollution is the discharge of pollutants from a single point source(such as a factory) into a water body.

III. RESULTS AND DISCUSSIONS

As already mentioned, the purpose of this work is to compare the performance of three machine learning models: SVM, Decision Tree, and Naive Bayes for water quality classification based on computed WQI. When implementing a machine learning algorithm, its performance is validated using two data sets.

Two performance evaluation parameters are defined to comparethe efficiency of the three algorithms. **Balanced Accuracy** Score: Equilibrium accuracy is the arithmetic mean of the accuracies achieved in each class. It ranges from 0 to 1, with 1 being the highest score. Mainly used when classes are skewed. A few classes are defined.

A. Confusion Matrix:

This is a performance measure for a classification problem in machine learning. The number of properly classified and falselyclassified datasets were aggregated with the appropriate values. As the name suggests, a confusion matrix is a matrix with rows showing the predicted class and columns showing the actual class of the corresponding data.

As mentioned in the above section, there are many factors that contribute to the optimization of classifier performance, they are called tuning parameters. In the SVM classifier, the two main parameters that define it are the kernel and the C parameter. During the training phase, different kernels as well as different values of C parameters were tested. Among the four kernels, the linear multiplication function is said to be more efficient for thisdata model. In a decision tree, the tree depth is checked with different values to avoid overfitting.

The obtained results are presented in Tables III and IV. Table III presents a summary of the results obtained using dataset 1 (Narmada river parameters) with four water quality parameters. From this table, we see that among the three algorithms, SVM and DT have better performance than Naïve Bayes. that is, they have the highest accuracy of 87.10% with 132 correctly

Classifiers	SVM	Decision Tree	Naïve Bayes
Accuracy	87.10	87.10	74.60
Correctly classified	132	132	129
Incorrectly classified	2	2	5

Table 3: Classification Sultsusing dataset

A classification results using dataset is the process of transforming raw data into a form that can be used by a machine learning algorithm. The classification results using dataset classifies the data based on its features and labels, also known as features and classes. The idea behindthis technique is to classify objects from one set of features into another set of features.

Classifiers	SVM	Decision Tree	Naïve Bayes
Accuracy	95.63	98.50	95.17
Correctlyclassified	540	544	531
Incorrectlyclassified	10	6	19

Table 4: Classification Results Using Dataset 2

The classification results obtained from the dataset were then compared with those obtained from other datasets in order to identify any differences. The first step of this analysis is to compare the mean, standard deviation and variance of each classifier output with those obtained from other trained classifiers.

Classified data out of a total of 134 data. Number of misclassifications found in the confusion matrix. Table IV illustrates the results obtained when testing data set 2 with the above 4 parameters. From the analysis of the results, it is foundthat the decision tree algorithm has the highest accuracy of 98.50% of the three algorithms. The number of

misclassified cases in DT was low compared with cases in SVM and Naïve Bayes.

By comparing the two tables, Naïve Bayes is not a suitable algorithm for this problem. The number of misclassified data was higher for the Naïve Bayes model in both data sets. In Table III, we see that the SVM and the decision tree have the same results as in Table IV; less amount of misclassified data in decision tree model. Thus, by analyzing the obtained results, it can be concluded that the decision tree corresponds well to the two data sets.

IV. CONCLUSION

The paper describes the evaluation of water high-satisfactory detection the usage of diverse device getting to know fashions along with SVM, DT, and Naive Bayes primarily based totally on computed weighted mathematics WQI. The version has been examined and confirmed towards 4 primary water high- satisfactory elements along with water pH, DO, conductivity and BOD. Extensive simulation evaluation is finished on real- time statistics units and the outcomes of 3 device-getting to know algorithms are presented. Based at the outcomes obtained; the choice tree set of rules changed into observed to be the maximum appropriate class version to symbolize the water high- satisfactory classes. The paintings may be similarly prolonged through education device getting to know fashions on big datasets and figuring out most effective high-satisfactory parameters for calculating water high-satisfactory. Water pollutants may be described in phrases of first-rate as decided via way of means of numerous traits consisting of pH, turbidity, the conductivity of dissolved oxygen (DO), nitrate, temperature,and biochemical oxygen demand (BOD).

REFERENCES

[1.] Forde, Martín, Ricardo Izurieta, and Banu Ôrmeci. "Water and health." Water Quality in the Americas, pp.27, 2019.
[2.] Rahmon, E., "Water for sustainable development", UNChronicle, vol. 55/1, pp.9-12, 2018.
[3.] M. A. Tirabassi, "A statistically based mathematical water quality model for a non-estuarine river system1." JAWRA Journal of the American Water Resources Association, Vol. 7, December 1971, pp. 1221 -1237.
[4.] Kang, Gaganjot, Jerry Zeyu Gao, and Gang Xie. "Data-Driven water quality analysis and prediction: A survey." IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService), 2017, pp. 224-232.
[5.] Guo, H. C., L. Liu, and G. H. Huang. "A stochast ic water quality forecasting system for the Yiluo River." Journal of Environmental Informatics, vol.1, no. 2, pp.18-32, 2003.
[6.] Sinha and R. K. Isaac, "An analytical FISmodel to check the quality of drinking water,"3rdInternational Conference on Computat ional Intelligence & Communication Technology (CICT), Ghaziabad, 2017, pp. 1-6.
[7.] W. Xuan, L. Jiake and X. Deti, "A hybrid approach of support vector machine with particle swarm optimization for water quality prediction," 5th International Conference on Computer Science & Education, Hefei, 2010, pp. 1158-1163.
[8.] Muhammad, Salisu Yusuf, Mokhairi Makhtar, Azilawati Rozaimee, Azwa Abdul Aziz, and Azrul Amri Jamal. "Classification model for water quality using machine learning techniques." International Journal of software engineering and its applications, vol 9, no. 6, 2015, pp. 45 - 52.
[9.] Haghiabi, A.H., Nasrolahi, A.H. and Parsaie, A., Water quality prediction using machine learning methods. Water Quality Research Journal, vol.53, no.1, pp.3-13, 2018.
[10.] Singh, Gurdeep, and Rakesh Kant Kamal. "Application of water quality index for assessment of surface water quality status in Goa." Current World Environment vol. 9, no. 3, pp. 994, 2014.
[11.] Breiman L, Friedman J, Stone CJ, Olshen RA, Classification and regression trees, CRC press, 1984.
[12.] Gupta, Nidhi, Pankaj Pandey, and Jakir Hussain. "Effect of physicochemical and biological parameters on the quality of river water of Narmada, Madhya Pradesh, India." Water Science, vol 31, no. 1, pp. 11-23, 2017.
[13.] Brown, Robert M., Nina I. McClelland, Rolf A. Deininger, and Michael F. O'Connor. "A water quality index—crashing the psychological barrier." In Indicators of environmental quality, pp. 173-182. Springer, Boston, MA, 1972.
[14.] Mppcb.nic.in. [online] Available at: <http://www.mppcb.nic.in/proc/narmada-report 2017-18.pdf> [Accessed 17 May 2020].
[15.] Anbarivan.N.L, "Indian water quality data," Kaggle, 23-Oct-2018. [Online].
[16.] Available: https://www.kaggle.com/anbarivan/indian-waterquality-data. [Accessed: 17-May-2020].
[17.] Aiswarya Vijayakumar and A. S. Mahesh, "Quality Assessment of Ground Water on Small Dataset", International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 5, 2019.
[18.] Aiswarya Vijayakumar and A. S. Mahesh, "Quality Assessment of Ground Water in Pre and Post-Monsoon Using Various Classification Technique", International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 2, 2019