Assessing Water Quality For Drinking Purpose Using Machine Learning Models

Nilay Lilawala (20BCE141)
Institute of Technology,
Nirma University,
Ahmedabad, Gujarat, 382481.
Email: 20bce141@nirmauni.ac.in

Malav Gajera (20BCE146)
Institute of Technology,
Nirma University,
Ahmedabad, Gujarat, 382481.
Email: 20bce146@nirmauni.ac.in

Malav Shah (20BCE147)
Institute of Technology,
Nirma University,
Ahmedabad, Gujarat, 382481.
Email: 20bce147@nirmauni.ac.in

***Abstract*—Lifespan of people, animals, and plants depends on water. Pure water is sometimes not suitable for drinking, home, industrial, or other uses, despite its significance. The quality of water is affected by a variety of variables, including industrialisation, mining, pollution, and natural occurrences. These factors introduce or change numerous characteristics already existing in the water, which affects its acceptability by humans for drinking or any other type of usage. Rapid industrialization and urbanization have caused an alarming rate of water quality degradation, which has resulted in terrible illnesses. In water quality index prediction and its class here we investigated a number of supervised machine learning techniques.**

The proposed methodology employs 9 inputs, namely, pH, Degree of water hardness, Presence of solids, Chloroamines, Sulphates, Specific conductivity, Trihalomethanes, Organic Carbon percentage, and Turbidity. The data set used here is labelled i.e. Potability(1 or 0). This paper includes five machine learning classification models namely - Logistic Regression, Support vector machines, Decision Tree classifiers, Naive-Bayes and k nearest neighbour is used. The proposed methodology achieves reasonable accuracy using given parameters to check its usability in real time water quality determination operations.

I. Introduction

Water is one of the most vital part of human lives as it has innumerable uses in various sectors and fulfills majority of our everyday requirements. However, the purity of water is being compromised by life itself as the demand for resources is increasing day by day. Water is one of the most valuable tools of communication . Rapid development is causing a continuous deterioration in water quality. It is already well researched that the major cause of the hazardous illnesses and its spread is because of the degraded water quality. According to reports, billions of deadly illnesses and millions of deaths have been caused by diseases caused by degraded water quality, which accounts for more than 80% of infections in underdeveloped nations[1]. Specifically, the sixth objective, which is to guarantee and maintain everyone's access to water and sanitation[3]. Consuming this kind of water might have serious health effects or even be fatal. Therefore, it is important that a comprehensive methodology is implemented to guarantee careful monitoring of the water from its beginning till its final of use. Water samples must be regularly collected at each monitoring location in order to determine whether the water is appropriate for drinking, agricultural, and production uses.

Table 1. Feature Description

pH	pH of water (0 to 14).
Hardness	Capacity of water to precipitate soap in mg/L.
Solids	Total dissolved solids in ppm.
Chloramines	Amount of Chloramines in ppm.
Sulfate	Amount of Sulfates dissolved in mg/L.
Conductivity	Electrical conductivity of water in $\mu S/cm$.
Organic Carbon	Amount of organic carbon in ppm.
Trihalomethanes	Amount of Trihalomethanes in $\mu g/L$.
Turbidity	Measure of light emiting property of water in NTU.
Potability	Indicates if water is safe for human consumption. Potable - 1 and Not potable - 0

In order to evaluate the quality of water, a number of models have been created. All of these models take into account different variables, including Turbidity, Conductivity, Organic Carbon, Trihalmethanes, Solids, Suplhates etc. The

result of these models is the Potability Class. If it is 0 then is it not drinkable else it is drinkable. There are three phases involved in computing Water Quality Index. The first phase involved giving each of the 19 factors a weight (wi) based on how significant it is in relation to the overall quality of the water that may be used for drinking (Table 3). Due to its critical role in determining the quality of water, the parameter nitrate has been given a maximum weight of 5. Magnesium is assigned a weight(minimum) of 1, may not be detrimental when taken alone. Other In the subsequent phase, we use the given equation to find out the relative weight (Wi):

$$W_i = \frac{w_i}{\sum_{i=1}^{n} w_i}$$

Where, relative weight is given by Wi, weight of every factor is defined as wi and the number of parameters is n. Table 1 includes all relative weights for the factors. For the purpose of calculating WQI, various guidelines have been modified globally. For example, in India. standards are as below.

Table 2. Relative weight of chemical parameters.

Chemical parameters	Indian Standards	Weight (w_i)	Relative weight (W_i)
pH	6.5-8.5	4	0.09756
Total hardness (TH)	300-600	2	0.04878
Calcium	75-200	2	0.04878
Magnesium	30-100	2	0.02439
Bicarbonate	244-732	3	0.07317
Chloride	250-1,000	3	0.07317
Total dissolved solids (TDS)	500-2,000	4	0.09756
Fluoride	1-1.5	4	0.09756
Manganese	0.1-0.3	4	0.09756
Nitrate	45-100	5	0.12195
Iron	0.3-1.0	4	0.09756
Sulphate	200-400	4	0.09756
		$\sum w_i = 41$	$\sum W_i = 1.000$

This study on calculating WQI values includes five main categories, from ”good” to ”water unsuitable for consumption”, and range from 89.21 to 660.56. The percentage of water samples falling into each category of quality is shown in Table 3. It has been shown that the high WQI values at these stations are mostly caused by the groundwater’s higher concentrations of metallic iron, nitrates, dissolved solids, hard water, fluoride presence, bicarbonates, and manganese deposits.

Table 3. Qualitative classification of water samples based on their WQI values

WQI value	Water quality	Percentage of water samples (Pre-monsoon)
<50	excellent	00
50-100	good water	1.5
100-200	poor water	63.5
200-300	very poor water	22
>300	Water unsuitable for drinking	13.0

II. Literature Review

This study examines the approaches that have been used to address the issues with water quality on our planet. In the study, traditional and statistical analysis are frequently employed to help determine the quality of the water and can predict if the water is drinkable or not, while other analyses use machine learning approaches to help identify the best possible solution to the water quality problem. We were able to understand the water quality issue better becaus of the local studies that used lab analyses. In one such study, S.Acharya collected water samples from several Indian regions and evaluated them using manual lab analysis against various criteria.[9] They discovered a significant prevalence of E. coli and faecal coliform owing to industrial and sewage waste. As input, they employed 25 different water quality metrics. They obtained an R2 and MSE of 0.9270, 0.9390 and 0.1200, 0.1158 using a mix of backward elimination and forward selection selective combination approaches.

The majority of studies either utilised manual lab analysis, did not estimate the water quality index standard, or used too many factors to be sufficiently effective. This approach currently used and how the methodology improves on these ideas.

III. Data Pre-Processing

Kaggle provided the data for this water quality study and determining if it is drinkable or not. Labelled data distribution is shown using a pie-chart below.

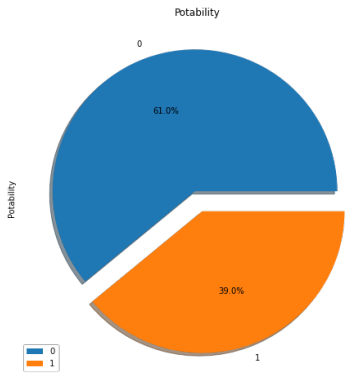

Figure 1. Potability Classes

A. Handling Null Values

Data was first analyzed and we found out that Sulphate, pH and Trihalomethanes contains NULL values. The missing percentage was 23.84%, 14.98% and 4.94% respectively. To avoid miss-classification and poor performance by machine learning models, the null values were filled with the mean values. The data has no null values now.

B. Data Cleaning and Outlier Detection

The section includes analysis of the box-plot. After there exist no null values, there were still outliers. Outliers can decrease the efficiency of the prediction model that were used to predict the water quality here.

For that, analysis on box-plot was brought up. Because the majority of the factors differed significantly and lied on the higher values, we used this analysis for outlier detection as it offers informative visuals to choose outlier detection threshold values based on the issue domain. 3 Following plot shows the range of the different parameters and the outliers that exist in the given data set.

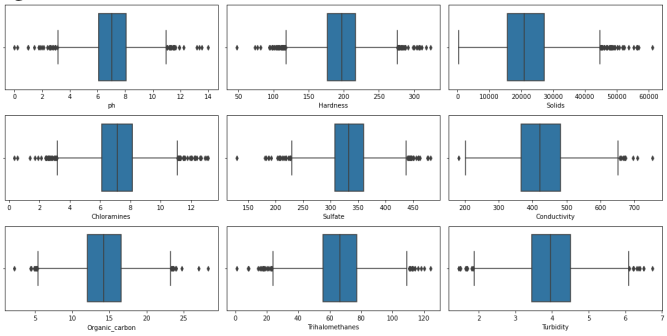

Figure 2. Box-plot before data cleaning

As we can see, pH, Chloramines, Trihalomehtanes and Solids have many outliers compared to other parameter. After applying proper treatment to the data set we can reduce the outliers and this increase in model efficiency. For that, upper bound and lower bound was calculated from the frequency distribution of the parameter using quantile() function. The data will be replaced with the upper bound value if the parameter value exceeds the upper bound, and the lower bound value if the parameter value exceeds the lower bound. Thus we modified an upper cap technique to filter out outliers after box-plot analysis revealed that the majority of parameters were outside the box, making outliers typical.

Given our limited data, we went through the same method again with all the parameters and individually eliminated the outliers . In order to avoid fudging the data set and only gently punish the values that were far outside of the range and unlikely to occur, we were also very lenient when determining the upper threshold of the parameters

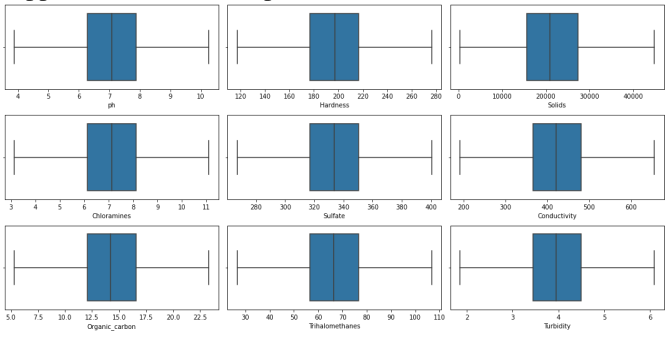

Figure 3. Boxplot after data cleaning

C. Correlation Analysis

We used correlation analysis to uncover potential correlations between the parameters in order to identify the dependent variables and forecast difficult-to-estimate variables using readily obtainable parameters. The Pearson correlation, which is the most widely used and reliable correlation approach, was employed.

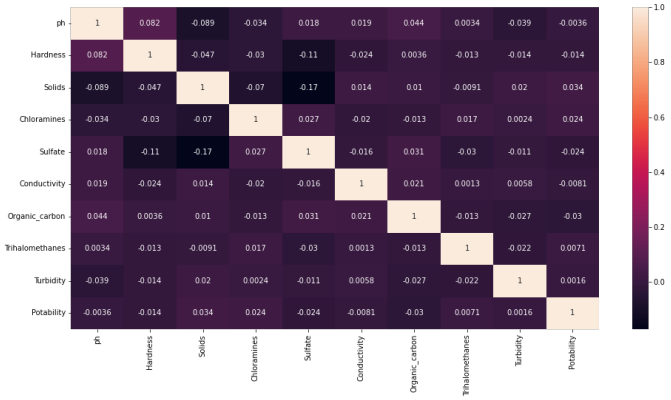

Figure 4. Correlation Heatmap

As the figure depicts, no feature is co-related with each other. All the features are independent. We choose all of the factors for the prediction of Water Potability to wrap up the correlation study.

D. Feature Scaling

As the figure suggest that the ranges of the different parameters are quite different. It affects the performance of the machine learning models. To avoid it, feature scaling is must done before model train.

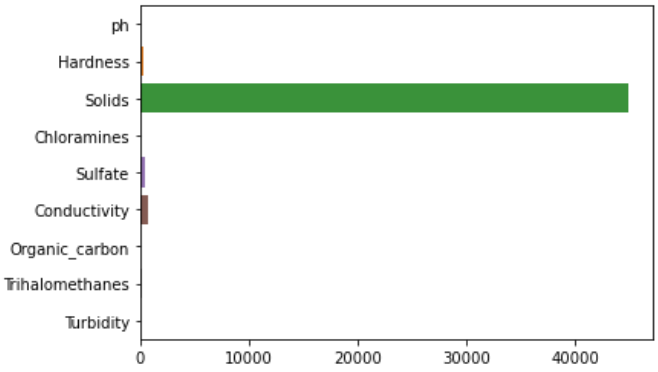

Figure 5. Value Range of Features

Sklearn library StandardScaler is used here to scale the features. It uses Z-score normalization to scale the features. For normalizing the given data through z-score normalization, subtract the population average from the data set and division with the standard deviation of that feature, which will predict the result usually lying between 3 and +3; To predict the standard score of the given sample x:

$$z = (x - \mu) / s$$

It converges the range to close values to each other.

where μ is the average of the training samples, and s can be defined as the standard deviation of the training samples.

IV. Machine Learning Model Training

A. Data Splitting

The initial step in using a machine learning model is to divide the given data set for model training, testing using a subset of the data, and accuracy metrics can be calculated to find out the model's efficiency.

Data should be split in the manner so that training data have a high amount of data. In this, data was being split at an 80-20 ratio of training vs. testing data. Doing this makes sure that an efficient train-test split data and use of uniform and predefined data for model fitting.

Also we have to make sure about that this splitted data should be randomly shuffled for better test prediction.

B. Machine Learning Algorithms

For water sample quality determination we have used regression as well as classification algorithms. We used five machine learning algorithms and compare the results and accuracy with the test data. The following algorithms were used in our comparative study: Logistic Regression, Decision Tree, Gaussian Naive Bayes, KNN and SVM. Sklearn module which is very well known for Data and Machine Learning is used here.

(1) Logistical Regression

In classification problems we use Logistic Regression. It is based on the mathematical logistic sigmoid function which ranges between 0 to 1. It is highly preferred algorithm employed for binary classification, Thus we used Logistic Regression here as we have only 2 classes namely, 0 or 1.

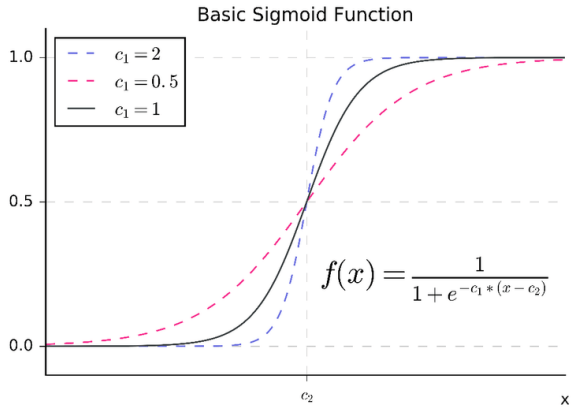

Figure 6. Logistic Regression

(2) Decision Tree

Trees are easy to understand and intuitive algorithms. It is the most powerful algorithm for classifying and predicting different classes. It is a flow-chart like tree structure. The decision are made using the values of all the useful input features fed to the model after training phase. It uses entropy(ID3 classifier). Through which it selects the base feature or attribute, based on the info gain of the attribute [5].

(3) Gaussian Naive Bayes

Naïve Bayes classification model has a easy and a quick implementation which uses the Bayes theorem for calculating probability of a given event with the prerequisite that the likelihood of two events is independent and irrespective of other events taking place [6].

(4) KNN

Finding the given points' nearest N neighbours and classifying them according to the majority of those neighbours is how the K nearest neighbour algorithm sorts data. One could use various strategies to resolve a draw, such as increasing n or adding bias toward one class. Large data sets shouldn't be used with K nearest neighbour because all processed occurs during testing phase and it is repeated through all of the training data, computing nearest neighbours each time. Our model implements this with the value of n as 7.

(5) Support Vector Machine

Although they may also be used for regression, support vector machines (SVMs) are often employed for classification. SVMs use data points plotted on a plane to visualise the data, defining a hyperplane between the classes and extending the margin to maximise the distinction between two classes, leading to fewer close calculation errors. [3].

V. Results

A. Accuracy Measure

(1) Accuracy

Accuracy can be defined as the right amount of predictions done using the machine learning model with respect to the observed values. Accuracy can be calculated using:

Accuracy = (True Positive + True Negative) / (True Positive + False Positive + True Negative + False Negative)

(2) Precision

Precision is the percentage of occurrences of a certain positive class that are properly classified out of all instances of that class that are classified. Precision can be determined using the following equation:

Precision = True Positive / (True Positive + False Positive)

(3) Recall

The ratio of any data set example of a specific positive class that were actually correctly classified is known as recall. Recall can be determined using the equation shown below:

Recall = True Positive / (True Positive + False Negative)

(4) F1 Score

We used their harmonic mean to depict the F1 score, which incorporates both characteristics and more accurately represents the accuracy measure as a whole because precision

and recall alone do not fully cover all aspects of accuracy. The f1 score lies between 0 and 1 and the closer it is to 1 the better.

F1 Score = (2 × Precision × Recall) / (Precision + Recall)

B. Result of various algorithm

Table 4 shows that all models performed poor w.r.t accuracy scores. Decision Tree had the least accuracy at 61%. This implies that Decision Tree mis-classified hazardous water samples as reliable for drinking about 4% of the time. GNB and SVM on the other hand resulted in better alternatives as it shows better accuracy score.

Table 4. Classification Result

Algorithm	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.63	0.37	0.00	0.22
Decision Tree	0.61	0.48	0.61	0.54
Guassian Naive Bayes	0.65	0.60	0.20	0.30
KNN	0.64	0.52	0.38	0.44
SVM	0.70	0.67	0.36	0.47

VI. Conclusion

WQI is used to assess the quality of water, one of the most vital resources for existence. Traditionally, expensive and time-consuming lab analysis is required to assess the quality of water. In this study, a different approach to machine learning was investigated to forecast water quality using basic, readily accessible water quality data. The study's data came from Kaggle, which was utilised to collect it. A selection of various supervised machine learning algorithms was used to predict Water Quality(Drinkable or not).

This work is mainly focused on the application of assessing water quality using machine learning models. Machine Learning models used here, which are Logistic Regression, Support Vector Machine, Decision Tree, KNN, Naive Bayes from which final results of the analysis depicts that SVM performed the best for drinking water potability prediction, as it gave the best test accuracy among all the models for classification.

The originator suggested the various application of deep learning models. But these are not supposed to use for this application. But in the future, deep learning models like various variants of the neural networks can be used to expand to this application. In future works, this idea can be integrated in a huge IoT-based online monitoring system. In which required parameters can be calculated using the sensors. The tested algorithms can predict the water quality or potability immediately on the basis of the real-time data fetched from the IoT system.

References

[1] PCRWR. National Water Quality Monitoring Programme, Fifth Monitoring Report (2005–2006); Pakistan Council of Research in Water Resources Islamabad: Islamabad, Pakistan, 2017.

[2] B. X. Lee, F. Kjaerulf, S. Turner, L. Cohen, P. D. Donnelly, R. Muggah, R. Davis, A. Realini, B. Kieselbach, L. S. MacGregor, I. Waller, R. Gordon, M. Moloney-Kitts, G. Lee, and J. Gilligan, "Transforming our world: Implementing the 2030 agenda through sustainable development goal indicators," J. Public Health Policy, vol. 37, no. S1, pp. 13–31, Sep. 2016.

[3] Integrated Approaches for Sustainable Development Goals Planning: The Case of Goal 6 on Water and Sanitation, U. ESCAP, Bangkok, Thailand,2017.

[4] Tong, S.; Koller, D. Support vector machine active learning with applications to text classification. J. Mach. Learn. Res. 2001, 2, 45–66.

[5] Quinlan, J.R. Decision trees and decision-making. IEEE Trans. Syst. Man Cybern. 1990, 20, 339–346. [CrossRef]

[6] Zhang, H. The optimality of naive Bayes. AA 2004, 1, 3.

[7] A. R. Finotti, R. Finkler, N. Susin, and V. E. Schneider, "Use of water quality index as a tool for urban water resources management," Int. J. Sustain. Develop. Planning, vol. 10, no. 6, pp. 781–794, Dec. 2018.

[8] R. Divahar, P. S. A. Raj, S. P. Sangeetha, T. Mohanakavitha, and T, 'Dataset for the analysis of water quality of ground water in suratkal, adajan district, Bangalore, India,' Data Brief, vol. 32, Oct. 2020, Art. no. 106112.

[9] S. Acharya, S. K. Sharma, and V. Khandegar, 'Assessment of water potability by WQI for irrigation and drinking in West Bangal, India,' Data Brief, vol. 18, pp. 2019–2028, Jun. 2018

[10] Ahmed, Umair & Mumtaz, Rafia & Anwar, Hirra & Shah, Asad & Irfan, Rabia & García-Nieto, José. (2019). Efficient Water Quality Prediction Using Supervised Machine Learning. Water. 11. 2210. 10.3390/w11112210.

[11] O. O. Ajayi, A. B. Bagula, H. C. Maluleke, Z. Gaffoor, N. Jovanovic and K. C. Pietersen, "WaterNet: A Network for Monitoring and Assessing Water Quality for Drinking and Irrigation Purposes," in IEEE Access, vol. 10, pp. 48318-48337, 2022, doi: 10.1109/ACCESS.2022.3172274.