A Prediction of Water Quality Analysis Using Machine Learning

[1] Suma S
Assistant Professor ,Deptarment of Computer Science and Engineering ,CMR Technical Campus ,Affilitaed to JNTU, Hyderabad Telangana, India.
sn.suma05@gmail.com

[2] Rohit Moon
B.Tech Students , Deptarment of Computer Science and Engineering ,CMR Technical Campus ,Affilitaed to JNTU, Hyderabad Telangana, India.
177r1a05g4@cmrtc.ac.in

[3] Mohammed Umer
B.Tech Students , Deptarment of Computer Science and Engineering ,CMR Technical Campus ,Affilitaed to JNTU, Hyderabad Telangana, India
197r1a05q1@cmrtc.ac.in

[4] K.Srujan Raju
Professor ,Deptarment of Computer Science and Engineering ,CMR Technical Campus ,Affilitaed to JNTU, Hyderabad Telangana, India line
ksrujanraju@gmail.com

[5] Nuthanakanti Bhaskar
Assistant Professor ,Deptarment of Computer Science and Engineering ,CMR Technical Campus ,Affilitaed to JNTU, Hyderabad Telangana, India.
bhaskar4n@gmail.com

[6] Rakshita Okali
Assistant Professor ,Deptarment of Computer Science and Engineering ,CMR Technical Campus ,Affilitaed to JNTU, Hyderabad Telangana, India.
rakshitaokali1997@gmail.com

***Abstract*—Data on water quality in Kenya is analyzed using a decision tree classification model. Using data mining techniques based on parameters related to water quality, the decision tree algorithm helps predict clean water. A predictive model was developed to identify water samples requiring further analysis in order to streamline the work of laboratory technologists. WEKA software was used to implement the model based on secondary data collected from the Kenya Water Institute. Water samples were classified into clean and contaminated categories using the decision tree algorithm. A crucial factor for evaluating water quality is its alkalinity and conductivity. Public health and safety depend on access to clean drinking water. Researchers used five decision tree classifiers to evaluate the model's accuracy: J48, LMT, Random Forest, Hoeffding Tree, and Decision Stump**

Keywords— Decision tree, Machine learning, J48, LMT, Random Forest, and Decision Stump

I. INTRODUCTION

An algorithm for predicting water quality based on machine learning. Health risks and significant economic losses can result from poor water quality. Water quality must therefore be analyzed from both an environmental and economic perspective [6]. WEKA is an excellent tool for analyzing data for this purpose. WEKA is a JAVA-based data mining software developed by the University of Waikato in New Zealand. DM and ML communities have widely adopted it, and it represents a significant milestone. To ensure the safety of human consumption and environmental resources, water quality analysis is a vital task. It is often not feasible to monitor water quality in real time using traditional methods of analysis, which involve time-consuming and costly laboratory testing. Consequently, machine learning techniques have gained popularity in recent years for analyzing water quality efficiently and accurately. Water quality analysis is carried out using a decision tree algorithm based on the decision tree algorithm.

Different water quality parameters are used in the model to categorize water quality. Due to its interpretability and the ability to handle categorical and numerical data, the decision tree algorithm was chosen. Water resource managers can use the proposed model to monitor water quality in real time and make better decisions.

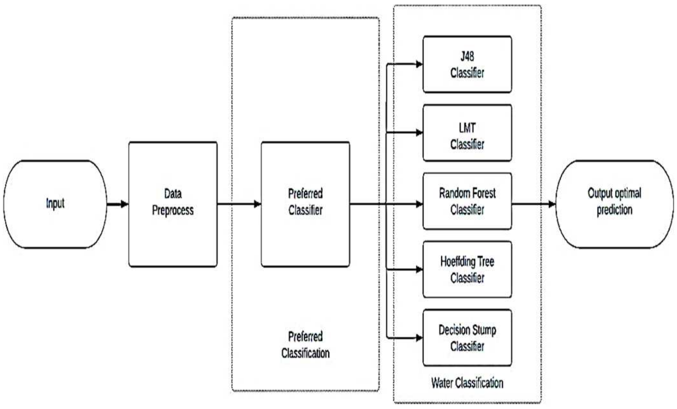


Fig.1: Classification of water quality according to its architecture

Five different classifiers were applied to assess the accuracy of the decision tree model in this architecture Fig.1. A decision tree classifier based on J48, LMT, Random Forest, Hoeffding Tree, and Decision Stump was used in this study. Water quality predictions were made using these classifiers to compare their accuracy.

a. J48 classifier: A popular implementation of the C4.5 algorithm is the J48 classifier, which uses the decision tree algorithm in order to classify data. Top-down greedy decision trees are built by the J48 algorithm from training data. A subset of the data is split recursively at each node of the tree according to the most significant attribute. In this case, the tree will reach a predefined depth when the subsets are homogeneous. Water quality analysis is one of the many applications of the J48 algorithm. In addition to its ability to handle categorical and numerical data, the J48 algorithm is easily interpretable, making its application in decision-making straightforward.

b. LMT Classifier: A logistic model tree classifier algorithm is a combination of a decision tree and a logistic regression algorithm. LMT algorithm generates prediction models using training data and fits logistic regression models at each node of the tree. By pruning the model, the algorithm

979-8-3503-4745-6/23/$31.00 ©2023 IEEE

avoids overfitting, resulting in a smaller and more interpretable tree. In classification problems with many attributes or noisy data, the LMT algorithm is particularly useful [8]. In addition to bioinformatics, finance, and environmental science, it has been applied to a variety of domains. The LMT algorithm can aid in decision-making due to its high accuracy and interpretability.

c. Random forest classifier: Multi-decision tree ensemble classifiers reduce accuracy and overfitting by using Random Forest Classifier. A subset of features and training data is randomly selected at each node of the forest to build a decision tree. By combining the predictions of all the trees, a final decision is made by combining the results of several decision trees. A feature importance ranking is also included in the algorithm, which indicates the importance of each feature in classification. Image recognition, bioinformatics, and finance are just a few of the tasks that the Random Forest algorithm can be used for. When compared to other classification algorithms, the Random Forest algorithm is less sensitive to outliers and is capable of handling high-dimensional and noisy data. Additionally, it provides insight into the most important features for classification tasks with reasonable accuracy.

d. Hoeffding Tree Classifier : An algorithm designed for handling continuous streaming data is this decision tree classifier. As updated data arrives, branches are added to the decision tree incrementally. In the algorithm, split nodes in a tree are determined by using the Hoeffding bound. With a small amount of data, Hoeffding bounds allow the algorithm to make decisions with high confidence. Internet traffic analysis, sensor networks, and financial applications are examples of applications where the Hoeffding Tree algorithm is particularly useful. With limited memory and computational resources, the Hoeffding Tree algorithm is capable of handling large amounts of data. Adaptability to changing data distributions is another advantage of the system.

e. Decision Stump Classifier : There are two leaf nodes and one root node in the decision stump classifier algorithm. Using the majority class in the corresponding subset of data, the algorithm divides the data based on a single attribute. When a single significant feature exists in the data or if a quick decision is required, the decision stump algorithm can be particularly useful. Since only one feature is considered in the classification task, it is less accurate than other more complex decision tree algorithms. Text classification and image recognition are two applications of the decision stump algorithm. When computational resources are limited, the decision stump algorithm is useful because of its simplicity and speed.

II. Related Work

Various machine learning techniques have been explored in previous studies to address water quality concerns. To deal with changes in drinking water quality, Muharemi et al. (2019) proposed the use of the K-Nearest Neighbor algorithm and Neural Network Classification based on Logistic Regression [1]. Similarly, Haghiabi et al. (2018) assessed the effectiveness of artificial intelligence techniques in predicting water quality components in Tireh River, Iran, such as Artificial Neural Networks, Group Data Management Methods, and Support Vector Machines[2]. Accordingly, Zhang et al. (2017) developed a statistical model and double-movement window algorithm for detecting anomalies in water quality data[3]. As compared to other algorithms such as AD and ADAM, the algorithm showed better anomaly detection performance on pH values. Several references have been published on the topic of water quality analysis classification models using decision trees, including:

TABLE 1. CLASSIFICATION MODELS FOR WATER QUALITY ANALYSIS USNING DESCISION TREE .

Algorithm	Authors	Application	Result
Classification- based Neural Network using Logistic Regression	Muharemi et al., 2019	Changes in the quality of drinking water	Adequate solution
Artificial Neural Network (ANN)	Haghiabi et al., (2018)	Water quality of the Tireh River	The Tansig And RBF functions have shown the best performance as transfer and core functions.
Support Vector Machine (SVM)	Haghiabi et.al., (2018)	Water quality of the Tireh River	The Tansig And RBF functions have shown the best performance as transfer and core functions.
Artificial Neural Networks (ANN)	Chou et al., (2018)	Water quality in the reservoir	It has been found to be more accurate than other models.
Support Vector Machines (SVM)	Mohammad pour et al. (2015)	Water quality	Competitive with neural networks

Applied data mining techniques can be used to predict water quality, as examined in this paper. In previous studies, in the above table 1 a wide range of data mining techniques have been used in the environmental domain, including the Nearest Neighbor Algorithm (KNN). The Artificial Neural Network (ANN)was proposed by Haghiabi et al. (2018) [2], the Artificial Neural Network (ANN) was proposed by Chou et al. (2018) [3], and the Support Vector Machine was proposed by Mohammadpour et al. (2015) [4].Two primary objectives are being pursued in this study:

1. By applying common data mining and classification techniques, a predictive model can bedeveloped to predict the type of data. Qualitative characteristics of river water.

2. Identify the most appropriate predictive model for the present study based on different evaluation metrics.

III. Proposed Work

Data mining algorithms were executed using WEKA, a data mining tool developed by the University of Waikato in New Zealand using the JAVA programming language. There is widespreadrecognition of WEKA in the machine learning and data mining communities. For classification, decision trees include ID3, AD trees, REP trees, J48 trees, FT trees, LAD trees, decision stumps, LMT trees, random forests, and random trees.

The following steps might be included in a proposed machine learning system for predicting water quality Fig.2:

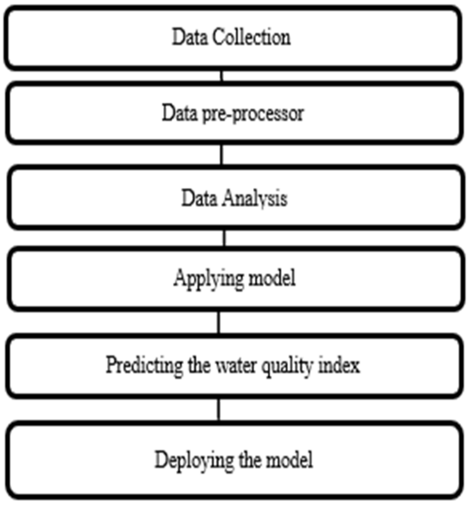


Fig.2: Steps for predicting water quality

1) Data collection: Obtaining data from various sources, including field measurements, laboratory analyses, and remote sensing.

2) Data pre-processing: The process of cleaning and preparing data for analysis, including removing incomplete or inaccurate data, scaling the data, and selecting relevant features from it, before it is ready to be analyzed.

3) Data splitting: To divide the training data into training and testing sets, based on the size and complexity of the dataset, in a ratio that is based on the size and complexity of the dataset.

4) Decision tree model building: Taking the training data and using one of the suitable algorithms,such as J48 or Random Forest, to build a decision tree model from it.

5) Model evaluation: The decision tree model is evaluated by taking the test data, comparing the results with various metrics, such as the accuracy, precision, recall, and F1 score, to compare how well the model performs.

6) Model optimization: Depending on the results of the evaluation, the parameters of the decisiontree model may be optimized or different algorithms may be selected to optimize the model.

7) Model deployment: These models are deployed to an appropriate platform with the goal of making predictions based on existing water quality data, such as a web-based application, in order to make recommendations.

Ultimately, the proposed system will facilitate decision-making and improve water management practices through reliable and efficient water quality analysis.

A. Decision Tree:

Classification and regression problems can be solved using Decision Trees as a supervised learning algorithm. An internal node represents a dataset's features, a branch represents decision rules, and a leaf represents outcomes Fig.3. Leaf nodes are the results of final decisions that have been made, while decision nodes are used to make any decisions that have been made. The features of the given dataset are used for decision-making or testing. Decision tree algorithms follow a tree-like structure, starting with a root node and expanding from there. Classification and regression tree algorithms, commonly referred to as CART, are used to build trees. Decision trees are based on questions that are asked and then subtrees are created accordingly (Yes/No).

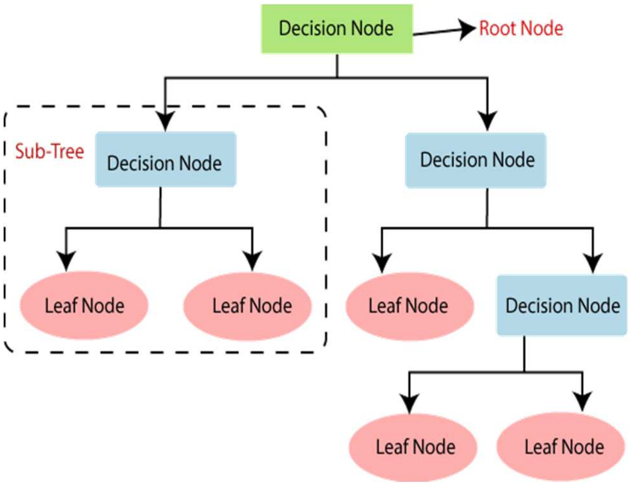


Fig.3: Decision tree

B. Different dimensions of sample data

There are different dimensions of the sample data to be used in the water quality indexing and to improve the prediction accuracy: a data normalization is initially carried out using Equation 1 to standardize the value of the measured parameters to be used in the calculations. The need to standardize the parameters arise from the fact that the parameters are all measured in various units. It is essential that a common denominator is established in order to realise a unitiess value that will cut across the various parameters

$$x' = \frac{x - x_{max}}{x_{max} - x_{min}} \qquad (1)$$

In equation (1) where x_{min} and x_{max} are the minimum and maximum values of the raw data obtained respectively. x is the raw data. x ' is the normalized data.

There has been a good number of research work with respect to water quality evaluation methods [11]. The single factor evaluation is used for rivers and lakes whereas the comprehensive water quality index methods have been developed by National Sanitation Foundation and Other groups based on the modified index method. There is also the water quality evolution trend analysis such as rank correlation method, time series analysis and parametric test method which many researchers have investigated as well [11]. These methods are weak due to the failure to deal with seasonal and missing values and the trend analysis reviewed hardly exhibit the distribution of the water quality indices in the water bodies. Having used the single factor evaluation for the calculation of different water quality index, it is seen as being more conservative than the comprehensive index method hence it is more suitable for evaluating water pollution of serious degree.

C. Pseudocode for a decision tree used to classify water quality:

```
# Step 1: Prepare the dataset
# Load the dataset from a CSV file or a database
# Preprocess of the dataset by handling missing values, encoding categorical variables, and normalizing numerical features
dataset = preprocess_dataset(raw_dataset)

# Step 2: Split the dataset in to the training set and testing sets train_set, test_set = split_dataset(dataset, test_ratio=0.3)

# Step 3: Train the decision tree model tree = decision_tree(train_set)

# Step 4: Evaluate the model on the testing set correct_count = 0
for example in test_set:
predicted_label = classify_example(tree, example) true_label = example[-1]
if predicted_label == true_label: correct_count += 1
accuracy = correct_count / len(test_set) print("Accuracy:", accuracy)
# Step 5: Use the model to classify new samples
new_sample = [6.8, 0.24, 0.35, 2.5, 0.045, 45, 170, 1.001, 3.0, 0.5, 10.8]
redicted_label = classify_example(tree, new_sample) print("Predicted label:", predicted_label)
end
```

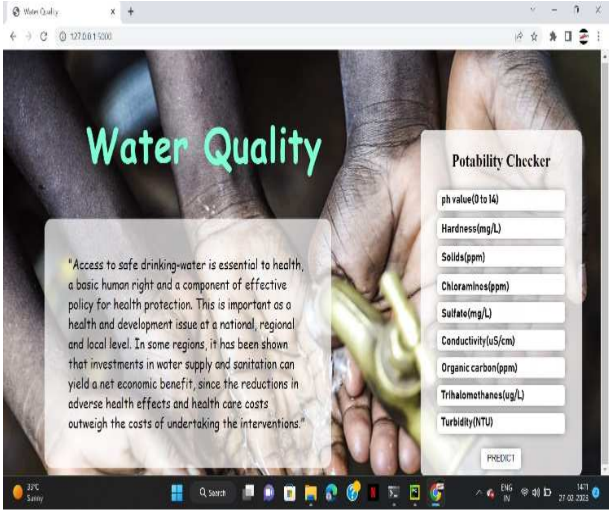
TABLE II WATER QUALITYPARAMETER

Water Quality Parameter	Permissible limit for surface waters
pH	6.5 – 9.0
Temperature	Increase of 10°C affects aquatic life
Dissolved oxygen	Fresh water: 7mg/-9mg/l Early life fishes: 9.5 mg/l in cold water and 6.0 mg/l in warm water
Ammonium	Greater than 0.1mg/l and less than 1mg/l
Nitrate	50mg
Turbidity	Based on dissolved solids
Phosphates	Less than 50pg/l at entry point and less than 25µg/l within the lake
Dissolved organic carbon	10mg/l threshold
Conductivity	Fresh water streams 150µS/cm to 500/cm

The above table2 shows that Water quality parameters and permissible limit for surface waters. Suppose that the training and testing sets of the water quality dataset have already been preprocessed. Missing values are handled, categorical variables are encoded, and numerical features are normalized by the preprocess_dataset function. Splitting the dataset into training and testing sets based on a specified test ratio is performed using the split_dataset function.

Training a decision tree model is accomplished using the decision_tree function. A decision tree and a sample are passed into the classify_example function, and the predicted label is returned for the sample.

IV. RESULTS

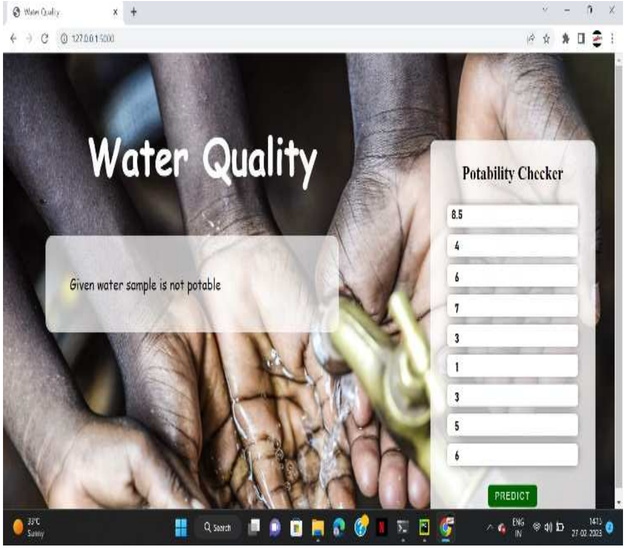


"Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions."

ph value(0 to 14)
Hardness(mg/L)
Solids(ppm)
Chloramines(ppm)
Sulfate(mg/L)
Conductivity(uS/cm)
Organic carbon(ppm)
Trihalomethanes(ug/L)
Turbidity(NTU)

Fig. 4: The Potability Checker assesses the potability of water.

The Potability Checker use various methods to assess the potability of water, including chemical testing, physical measurements, and/or microbial analysis. It may evaluate a wide range of parameters, such as pH, temperature, turbidity, dissolved solids, heavy metals, organic and inorganic contaminants, bacteria, viruses, and other pathogens as shown in Fig.4.These parameters are evaluated against the established guidelines and standards to determine if the water is safe and suitable for drinking.



Given water sample is not potable

8.5
4
6
7
3
1
3
5
6

Fig. 5: River water potability checker input values

In Fig.5 values of river water and after entering all values the given water sample is not potable.

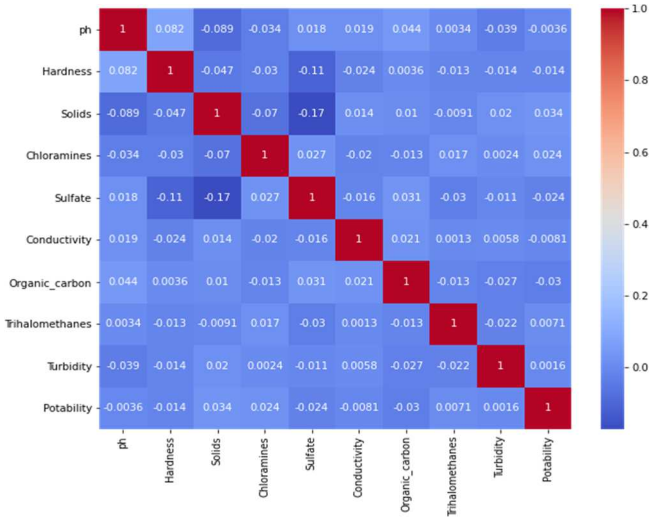


Fig. 6. Graphical representation

The above is the Graphical representation of the potable and non-potable water in Fig.6. Each water quality parameters have some limit for water quality by which the water quality is evaluated.

A. Portable

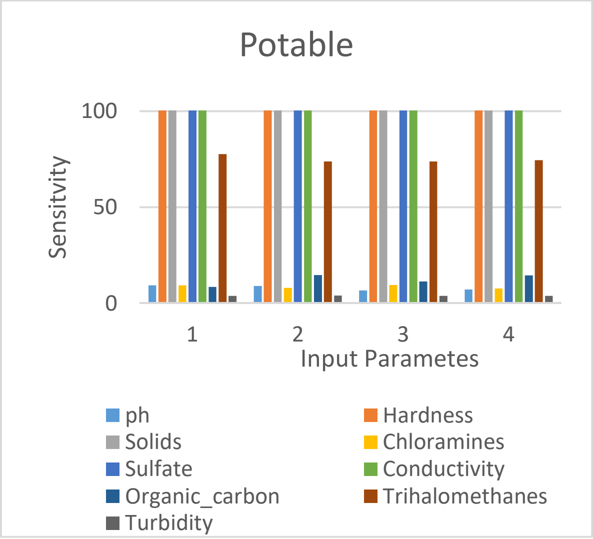


Fig.7. Graph on Potability

By measuring the pH of water, you can determine whether it is acidic or alkaline. From the Fig. 7 on a pH scale of 0 to 14, seven is neutral, seven is acidic, seven is alkaline, and seven is basic. "pH" measures the concentration of hydrogen ions in a solution.

In addition to hardness, water can also be classified by its quality by its hardness. Mineral-rich water, such as hard water or hard water with high levels of calcium and magnesium, has a high content of dissolved minerals. Water quality classification relies heavily on solids. Physical, chemical, and biological characteristics of water can be affected by solids, and overall quality can be affected by solids.

B. Non-Potable water:

A water supply with chloramines may be treated with disinfectants This treatment is done to maintain safe drinking water standards, so their presence is an indicator of water quality. Water salinity or hardness are typically measured by sulphate concentration when classifying water quality. Sulphates can cause bitter tasting water or cause laxative effects. As shown in the Fig.8.

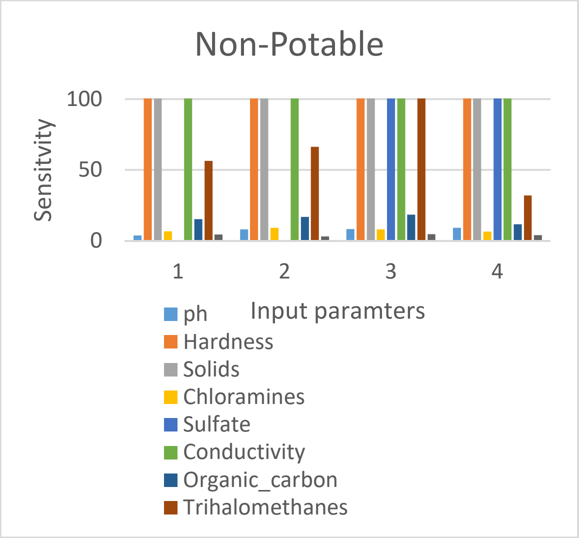


Fig. 8. Graph on Non-Potability

TABLE III. DATA SET VALUES FOR WATER PORTABILITY

Ph	Hardness	Solids	Chloramines	Sulphate	Conductivity	Organic carbon	Trihalomethanes	Turbidity	Potability
3.71608	129.422	18630.0	6.635246		592.8854	15.18001	56.32908	4.500656	0
8.09912	224.236	19909.5	9.275884		418.6062	16.86864	66.42009	3.055934	0
8.31676	214.373	22018.4	8.059332	356.886	363.2665	18.43652	100.3417	4.628771	0
9.09222	181.101	17978.9	6.5466	310.135	398.4108	11.55828	31.99799	4.075075	0
9.44513	145.805	13168.5	9.444471	310.583	592.659	8.606397	77.57746	3.875165	1
9.02484	128.096	19859.6	8.016423	300.150	451.1435	14.77086	73.77803	3.985251	1
6.800119	242.008	39143.4	9.501695	187.170	376.4566	11.43247	73.77728	3.85494	1
7.17413	203.408	20401.1	7.681806	287.085	315.5499	14.53351	74.40562	3.939896	1

In Table3 shows the data set values of potable and non potable water the above values are taken for water sample values from river water and lake water.

V. TESTING

Software testing is an important part of the software development life cycle because it aids in identifying bugs or defects before the software is released. Testing software is primarily concerned with ensuring that it meets all the specifications and behaves as expected. To test software, test cases must be created and executed, results must be compared with expected results, defects must be identified, and they must be reported to the development team.

a. Unit testing: Testing unit logic and ensuring that inputs produce valid outputs is an essential part of software development. Before integrating software units, all decision branches and internal code flow are validated via structural testing. An application, system configuration , or business process is tested using unit tests. By doing so, they ensure that every unique path of a business process follows the specifications as described in the documentation. In addition, unit tests require the user to know the construction of the software, so they are considered invasive since the inputs and outputs are well defined.

b. Integration Testing: Testing integrated software components is a crucial step in software development that determines whether they work together properly as one program. Despite passing unit tests successfully, integration tests ensure that the components are consistent and correct when combined. Software integration testing is primarily concerned with identifying any problems or issues resulting from the integration of software components. Integration testing is therefore crucial to ensuring that software systems function correctly and are reliable.

c. Functional Testing: The importance of functional testing in software testing cannot be overstated. Systematic testing provides proof that the tested features are available as specified in business and technical requirements, user manuals, and system documentation. It involves completing system procedures and interfacing with other systems and procedures, as well as verifying inputs, functions, and outputs. The key requirements, functions, and special test cases must be organized and prepared before performing functional testing. Software functional testing ensures that the system meets the specifications and performs its intended functions

VI. Conclusion

The analysis of water quality has been based on a decision tree-based classification model. An approach based on various input features can predict the quality of water. In order to provide clean and safe water for human consumption and tensure the sustainability of the environment, this approach can aid in decision-making and improve water management practices. A variety of decision tree algorithms are used for constructing classification models for water quality analysis, including Random Forest, LMT, Hoeffding Tree, J48 and Decision Stump. As part of the proposed system, data is collected and preprocessed, training and testing sets are separated, a decision tree model is built, its performance is evaluated and optimized, it is deployed to a suitable platform, and it is continuously monitored and updated. It is important to emphasize that accuracy and reliability of a classification model are dependent on many factors, such as data quality and quantity, feature selection, algorithm selection, and parameter optimization. By incorporating deep learning and ensemble methods, classification models can be enhanced for water quality analysis. It is possible to improve the quality of life for millions of people around the world by using classification models for water quality analysis.

VII. Futrure Scope

Water quality models should include biological and weather parameters, according to our research. By doing so, the possibility of contamination sources from a wide range of sources can be taken into consideration. In order to predict water quality accurately, hybrid models with minimal parameters are recommended. An essential information and strategy for early detection and mitigation of water contamination needs to be developed by stakeholders involved in water quality management. In this way, humans and the environment can both be protected from adverse effects.

VIII. Acknowledgment

We thank CMR Technical Campus for supporting this paper titled “A Prediction of Water Quality Analysis Using Machine Learning”, which provided good facilities and support to accomplish our work. Sincerely thank our Chairman, Director, Deans, Head Of the Department, Department Of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work

References

[1]. Azamathulla, H. M. 2013 2 – A Review on Application of Soft Computing Methods in Water Resources Engineering A2 – Yang, Xin-She. In: Metaheuristics in Water, Geotechnical and Transport Engineering (Gandomi, A. H., Talatahari, S. & Alavi, A. H., eds). Elsevier, Oxford, pp. 27–41.

[2]. Karanfil, O., & Konar, A. (2018). Development of a decision tree-based classification model for surface water quality. Environmental Monitoring and Assessment, 190(3), 171.

[3]. Saravi, S. S. S., & Malekian, A. (2019). Application of decision tree algorithm for prediction of water quality index (WQI). Journal of Water and Land Development, 43(1), 18-27.

[4]. Esakkirajan, M., & Thanushkodi, K. (2017). Decision tree based approach for water quality classification using WQI parameters. Journal of King Saud University-Engineering Sciences, 29(4), 318-324.

[5]. Ferreira, R. B., Lopes, J. A., & Ribeiro, R. (2018). Decision trees for monitoring and management of water quality in small and medium-size water supply systems. Water Science and Technology: Water Supply, 18(6), 2006-2013.

[6]. Jha, P., & Kumar, R. (2016). A novel decision tree approach for water quality prediction in rivers. Journal of Hydroinformatics, 18(1), 116-131.

[7]. Wang, Y., Zhan, C., Li, M., Chen, G., & Liu, Y. (2019). An improved decision tree method for water quality assessment of rivers. Journal of Cleaner Production, 210, 1001-1012.

[8]. Chatterjee, D., & Panchal, V. (2019). Water quality assessment of Ganga River basin using decision tree analysis. Environmental Earth Sciences, 78(19), 578.

[9]. Baskaran, R., & Muthusaravanan, S. (2018). An intelligent system for water quality analysis using decision tree algorithm. Applied Water Science, 8(5), 141.

[10]. Patil, S. S., & Jena, S. K. (2017). Decision tree algorithm based classification of water quality data. Procedia Engineering, 173, 506-513.

[11]. Abdellah El Hmaidi, Abdelghani Talhaoui, Imad Manssouri, Hajar Jaddi, Habiba Ousmana, "Contribution of the pollution index and GIS in the assessment of the physico-chemical quality of the surface waters of Moulouya River (NE, Morocco)", La Houille Blanche, vol.106, no.3, pp.45, 2020.

[12]. Sengorur, B.; Koklu, R.; Ates, A. Water quality assessment using artificial intelligence techniques: SOM and ANN—A case study of Melen River Turkey. Water Qual. Expo. Health 2015, 7, 469–490.

[13]. Aradhana, G.; Singh, N.B. Comparison of Artificial Neural Network algorithm for water quality prediction of River Ganga. Environ. Res. J. 2014, 8, 55–63

[14]. Muhammad, S.Y.; Makhtar, M.; Rozaimee, A.; Aziz, A.A.; Jamal, A.A. Classification model for water quality using machine learning techniques. Int. J. Softw. Eng. Its Appl. 2015, 9, 45–52.

[15]. Haghiabi, A.H.; Nasrolahi, A.H.; Parsaie, A. Water quality prediction using machine learning methods. Water Qual. Res. J. 2018, 53, 3–13.