RESEARCH ARTICLE OPEN ACCESS

Water Quality Prediction Using Machine Learning

Nikwade Sagar Ravindra*, Pandey Nilesh Munnu*,Patil Ganesh Bhagwat*, Bendale Sankalp Yogiraj

Department of Computer Engineering, K.V.N. Naik College of Engineering, Nashik

----------**********----------

Abstract:

Water quality deterioration poses a substantial environmental and health threat. Predicting water quality can be crucial for proactive management and mitigation strategies. This study explores the use of machine learning algorithms for water quality classification. The Water Quality Index (WQI), incorporating parameters like pH, temperature, and dissolved oxygen, will be used to evaluate water quality. Subsequently, water samples will be assigned distinct classes based on their WQI values. Machine learning algorithms, including K-Nearest Neighbors, Naive Bayes, Support Vector Machine, Decision Tree, and Random Forest, will be implemented for classification. Their performance will be evaluated using accuracy, confusion matrix, precision, recall, and F1-score metrics. This study aims to identify the optimal machine learning algorithm for water quality classification, contributing to the development of efficient and reliable water quality monitoring systems.

Keywords **—Machine Learning; WQI; WQC; K-NN; Decision Tree; Naïve Bayes; SVM; Random Forest ;**.

----------**********----------

I. INTRODUCTION

The critical role of water in sustaining human life and environmental well-being necessitates maintaining its quality. Rapid urbanization and industrial development have unfortunately led to water quality degradation, jeopardizing access to safe drinking water. Predicting water quality is crucial for ensuring the safety and sustainability of this vital resource. Water quality, encompassing its physical, chemical, and biological characteristics, is influenced by various factors, both natural and human-induced, such as climate change, land use patterns, population growth, and industrial activities. Therefore, monitoring and predicting water quality are essential for its safe and sustainable use. Traditional methods of water quality prediction rely on manual observations and laboratory analysis, which can be time-consuming, resource-intensive, and may not provide real-time information. Recently, machine learning techniques have emerged as a promising alternative approach. Machine learning algorithms can analyze historical data to identify patterns and trends, enabling them to make accurate predictions about future water quality parameters. This paper investigates the use of machine learning techniques for water quality prediction. Our aim is to demonstrate the effectiveness of these algorithms in predicting water quality, highlighting their potential to improve the accuracy and reliability of water quality monitoring. We will also discuss the challenges associated with this approach. Machine learning has proven its effectiveness in tackling various real-world challenges, including water quality prediction. The growing availability of water quality data empowers machine learning algorithms to develop accurate and reliable prediction models. The objective of this paper is to provide a comprehensive overview of cutting-edge machine learning techniques employed in water quality prediction. We begin by establishing the significance of water quality prediction and the inherent challenges associated with it. Water quality

prediction using machine learning involves building models capable of accurately predicting water quality based on various input parameters. The increasing popularity of machine learning techniques in this field stems from their ability to handle large and complex datasets. The initial step in water quality prediction using machine learning involves data collection. This entails gathering data on various water quality parameters, including pH, BOD (Biochemical Oxygen Demand), dissolved oxygen, nitrate levels, total coliform bacteria, conductivity, and various chemical concentrations.

II. MOTIVATION

The need for swift and accurate water quality information across diverse water sources drives the exploration of machine learning in this domain. Traditional monitoring methods, while crucial, are often costly, time-intensive, and lack real-time capabilities. Machine learning offers a promising alternative by effectively capturing the intricate interplay between various water quality parameters. This empowers authorities with real-time data, enabling them to make timely interventions and prevent potential contamination events. Consequently, public health risks are minimized, and the safety of drinking water supplies is bolstered. Furthermore, accurate predictions can optimize water treatment processes, leading to cost reductions. Additionally, valuable insights gained from these predictions can support conservation efforts for aquatic life and maintain the delicate ecological balance of water bodies. In essence, leveraging machine learning for water quality prediction represents a significant leap forward, paving the way for efficient, effective, and sustainable water resource management practices.

III. LITRATURE SURVEY

(1) COMPARISON OF WATER QUALITY CLASSIFICATION MODELS-

Water resources are often polluted by human intervention. Water pollution can be defined in terms of its quality which is determined by various features like pH, turbidity, electrical conductivity dissolved oxygen (DO), nitrate, temperature and biochemical oxygen demand (BOD). This paper presents a comparison of water quality classification models employing machine learning algorithms viz., SVM, Decision Tree and Naïve Bayes. The features considered for determining the water quality are: pH, DO, BOD and electrical conductivity. The classification models are trained based on the weighted arithmetic water quality index (WAWQI) calculated. After assessing the obtained results, the decision tree algorithm was found to be a better classification model with an accuracy of 98.50.

(2). DECISION TREE-BASED DATA MINING AND RULE INDUCTION FOR IDENTIFYING HIGH QUALITY GROUNDWATER ZONES TO WATER SUPPLY MANAGEMENT:

A Novel Hybrid Use of Data Mining and GIS-Groundwater is an important source to supply drinking water demands in both arid and semiarid regions. Nevertheless, locating high quality drinking water is a major challenge in such areas. Against this background, this study proceeds to utilize and compare five decision treebased data mining algorithms including Ordinary Decision Tree, Random Forest, Random Tree, Chi-square Automatic Interaction Detector, and Iterative Dichotomiser 3 for rule induction in order to identify high quality groundwater zones for drinking purposes. The proposed methodology works by initially extracting key relevant variables affecting water quality.

(3) CLASSIFIER FOR DRINKING WATER QUALITY IN REAL TIME-

Real time features are critical for automatic assessment of Drinking Water Quality. This paper explores the use of real time features to feed machine learning classifiers for DWQ.Two different representative datasets were used from: The Provincial Water Quality Monitoring Network from Ontario, Canada and National Hydrologic Information System from Central Region of

Portugal. The procedure followed in this study was: automatically computing a Water Quality Index to classify the datasets elements in five classes (excellent, good, medium, bad and very bad) using the Kumar method; selecting best performed real time features on results of classified datasets; and exploring machine learning algorithm for producing DWQ classifiers. In this work, we perform the classification of two classes (good and medium) out of the five possible categories, due to the absence of vectors in the datasets.

(4) PREDICTION OF IRRIGATION WATER QUALITY INDICES BASED ON MACHINE LEARNING AND REGRESSION

Assessing irrigation water quality is one of the most critical challenges in improving water resource management strategies. The objective of this work was to predict the irrigation water quality index of the Bahr El-Baqr, Egypt, based on nonexpensive approaches that requires simple parameters. To achieve this goal, three artifcial intelligence models and four multiple regression models ; potential of salinity, PS; permeability index. Electrical conductivity , sodium, calcium and bicarbonate were used as input exploratory variables for the models. The results indicated the water source is not suitable for irrigation without treatment. A good soil drainage system and salinity control measures are required to avoid salt accumulation within the soil.

(5) AN OPTIMIZED APPROACH FOR PREDICTING WATER QUALITY FEATURES BASED ON MACHINE LEARNING-

Traditionally, water quality is assessed using costly laboratory and statistical methods, rendering real-time monitoring useless. Poor water quality requires a more practical and cost-effective solution. The machine learning classification approach appears promising for rapid detection and prediction of water quality. Machine learning has been used successfully to predict water quality. However, research on machine learning for water quality index prediction is generally lacking. Therefore, this research aims to identify the important features for the WQI, which necessitated the classification of numerous indicators. This study develops four machine learning models based on the WQI and chemical parameters.

(6) NONLINEAR MAPPING APPROACH TO STAIN NORMALIZATION IN DIGITAL HISTOPATHOLOGY IMAGES USING IMAGE- SPECIFIC COLOR DENCONVULUTION-

Data pre-processing is considered as the core stage in machine learning and data mining. Normalization, discretization, and dimensionality reduction are well-known techniques in data pre-processing. This research paper seeks to examine the effects of Min-max, Z-score, Decimal Scaling, and Logarithm to the base 2 on the accuracy of J48 classifier using the NSL-KDD dataset. Experiments were conducted using the above-listed methods and their individual results were compared to each other. Principal component analysis and Linear Discriminant Analysis were tested for dimensionality reduction; furthermore, a hybrid combination of PCA and LDA was attempted and the performance showed an improved classification accuracy compared.

(7) Analysis of Physiochemical Parameters to Evaluate the Drinking Water Quality in the State of Perak, Malaysia- The drinking water quality was investigated in suspected parts of Perak state, Malaysia, to ensure the continuous supply of clean and safe drinking water for the public health protection. In this regard, a detailed physical and chemical analysis of drinking water samples was carried out in different residential and commercial areas of the state. A number of parameters such as pH, turbidity, conductivity, total suspended solids, total dissolved solids, and heavy metals such as Cu, Zn, Mg, Fe, Cd, Pb, Cr, As, Hg, and Sn were analysed for each water sample collected during winter and summer periods.

(8) Predicting and Analyzing Water Quality using Machine Learning: A Comprehensive Model- The deteriorating quality of natural water resources like lakes, streams and estuaries, is one of

the direst and most worrisome issues faced by humanity. The effects of un-clean water are far-reaching, impacting every aspect of life. Therefore, management of water resources is very crucial in order to optimize the quality of water. The effects of water contamination can be tackled efficiently if data is analyzed and water quality is predicted beforehand.

(9) An Innovative Index for Evaluating Water Quality in Streams- A water quality index expressed as a single number is developed to describe overall water quality conditions using multiple water quality variables. The index consists of water quality variables: dissolved oxygen, specific conductivity, turbidity, total phosphorus, and fecal coliform. The objectives of this study were to describe the preexisting indices and to define a new water quality index that has advantages over these indices. The new index was applied quantitative picture for the water quality situation. If the new water quality index for the impaired water is less than a certain number, remediation like in the form of total maximum daily loads or changing the management practices—may be neededto the Big Lost River Watershed in Idaho, and the results.

(10) Performance analysis of machine learning algorithms for water quality monitoring system Collection- Machine learning, a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from sensor data or databases. We have various machine learning algorithms for classification but choosing the best one is an important task. For selecting the best algorithm for our water quality monitoring system.

IV. METHODOLOGY

1. Data Gathering and Preparation:

Collect data related to factors influencing water quality, like pH levels dissolved oxygen concentrations, turbidity, temperature variations, etc. Process the data by addressing any missing values, outliers and adjusting the scale of features if needed.

2. Choosing and Enhancing Features:

Pick out features that play a role, in predicting water quality accurately.Carry out feature enhancement by either generating features or modifying existing ones to enhance the models effectiveness.

3. Selecting Models and Training Them:

Opt for machine learning algorithms for regression or classification tasks based on your prediction requirements.

Train the models using the prepared data and assess their performance using appropriate evaluation measures such, as Mean Absolute Error Mean Squared Error or R squared.

We test this five algorithm and use among one of them:

1)Decision tree learning.
2)Random Forest.
3)KNN(K-Nearest Neighbors).
4)Support Vector Machines(SVM).
5)Naïve Bayes.

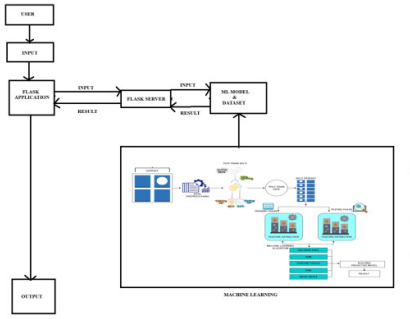

Fig: Architecture

4. WQI Calculation:

The quality of water can be concluded based on WQI value. The WQC class associated with each dataset is calculated using the value. Weighted Arithmetic Index Method is used to calculate the value of WQI. The equation for WQI calculation is

$$WQI=\frac{\sum qi \times wi}{\sum wi}$$

where 'qi' the quality estimate scale of each parameter in the range of 0 to 100 and 'wi' is the unit weight of the parameter. 'qi' values are calculated using the relationship

$$qi=100\times\frac{Vi-Vid}{Vs-Vid}$$

Vi-Observed value of i^{th} parameter
Vid-Ideal value of i^{th} parameter
Vs-Standard value of i^{th} parameter

The unit weight 'wi' is inversely proportional to the standard value recommended for each of the parameters

$$Wi=\frac{K}{Vs}$$

where K is the proportionality constant. WQC class is classified based on the tabulation given below

TABLE I. WQC CLASSIFICATION

Range of WQI	WQC Classification
0-25	Very Poor (1)
25-50	Poor (2)
50-70	Medium (3)
70-90	Good (4)
90-100	Excellent (5)

5. Min-Max Scaler:
Consequently, these algorithms often perform better after data scaling, leading to improved learning and testing outcomes This process involves splitting the data into dependent and independent variables, followed by applying a scaling technique like min-max scaling.

6.Train Test Split:
To train and evaluate a machine learning model, the data is typically split into two subsets. The training set, comprising 80% of the data in this project, is used to "fit" the model, meaning the model learns patterns and relationships within the data. The remaining 20%, called the testing set, is used to assess the model's generalizability and ability to make accurate predictions on unseen data. By comparing the model's predictions on the testing set with the actual known values, we can evaluate its performance and effectiveness.

7.Evaluation metrics:

a) Acuuracy Score: It is the ratio of correct prediction to the total number of predictions made.

b) Confusion matrix: It provides the precise number of predictions, including the number of accurate and erroneous ones

c) Classification Report: It is a summary of Confusion matrix and contains precision, recall, F1 Score, and support values of the model. Precision is the ratio of True Positives to total Positives and False Positives.

The weighted harmonic mean of precision and recall is F1 score.

$$F1\ score=2\times\frac{Precison\times Recall}{Precision+Recall}$$

8.Implementing Models with Flask:
Establish a Flask web application to act as the prediction platform. Define pathways to manage HTTP requests and responses. Integrate the trained machine learning model into the Flask application. Set up a form or an API endpoint where users can enter data for making predictions. Utilize the input provided by the users to predict outcomes using the model, in place and then share the results back, with the user.

V. RESULTS AND DISCUSSIONS

Using Flask Server the model is integrated with a user interface. The model is saved to a pickle file format and the user interface is developed using HTML. Users wil be able to input parameter values and receive the anticipated value of the WQC to which the given sample belongs, enabling them to evaluate the water quality.

VI. CONCLUSION

Our research involved conducting a comparative analysis of various machine learning classification algorithms to predict water quality classification.Derived from the Water Quality Index, incorporates all standard water quality indicators, providing a comprehensive assessment of water quality. Among the algorithms evaluated, the Random Forest Classifier emerged as the most effective, achieving an impressive accuracy score of 89.57% in predicting. Leveraging the best-performing model's evaluation metrics, a web application was developed using the Flask server

framework. The application's efficacy was tested with two water samples, both of which were accurately classified as 'good.' This developmentholds promise for efficient water quality management, enabling proactive measures to mitigate the detrimental effects of poor water quality and safeguard public health.

REFERENCES

1. N. Radhakrishnan and A. S. Pillai(2020) 'Comparison of Water Quality Classification Models using Machine Learning' 2020 5th International Conference on Communication and Electronics Systems (ICCES),pp. 1183-1188.
2. Jeihouni, M., Toomanian, A. & Mansourian, A. "Decision TreeBased Data Mining and Rule Induction for Identifying High Quality Groundwater Zones to Water Supply Management: a Novel Hybrid Use of Data Mining and GIS" Water Resour Manage 34, 139–154 (2020).
3. J. Camejo, O. Pacheco and M. Guevara(2013) 'Classifier for drinking water quality in real time', International Conference on Computer Applications Technology (ICCAT), pp. 1-5.
4. Mokhtar, A., Elbeltagi, A., Gyasi-Agyei, Y. et al. "Prediction of irrigation water quality indices based on machine learning and regression models" Appl Water Sci 12, 76 (2022).
5. Nur Afyfah Suwadi, Morched Derbali, Nor Samsiah Sani, Meng Chun Lam, Haslina Arshad, Imran Khan, Ki-Il Kim, "An Optimized Approach for Predicting Water Quality Features Based on Machine Learning", Wireless Communications and Mobile Computing, vol. 2022, Article ID 3397972, 20 pages, 2022.
6. H. S. Obaid, S. A. Dheyab and S. S. Sabry(2019) 'The Impact of Data Pre-Processing Techniques and Dimensionality Reduction on the Accuracy of Machine Learning', Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON), 2019, pp. 279-283
7. Khan, Y., See, C.S.: Predicting and analyzing water quality using machine learning: a comprehensive model. In: 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT), pp. 1–6 (2016).
8. Rahmanian, N., Ali, S.H.B., Homayoonfard, M., Ali, N.J., Rehan, M., Sadef, Y., Nizami, A.S.: Analysis of physicochemical parameters to evaluate the drinking water quality in the state of Perak, Malaysia. J. Chem. 2015, Article ID 716125, 10 pages, (2015).
9. Said,et al, "An innovative index for evaluating water quality in streams," Environment Management, vol.34, pp. 406-414, sep 2004.
10. D. Jalal and T.Ezzedine (2019) " Performance analysis of machine learning algorithm for water quality monitoring system." International Conference on Internet of Things, Embedded System and Communications, pp. 86-89.

