Water Quality Prediction Using Machine Learning

P. Nisha Priya B.E., M.E., (Ph.D.,)[1], V.Priya B.E., M.E[2], Sindhumathi G B.E, (M.E)[3]
[1]*Asst. Professor &Computer Science and Engineering,Head of Department, Computer Science and Engineering CSI College of Engineering, Ketti*
[2]*Assistant. Professor, CSI College of Engineering, Ketti*
[3]*Student, Computer Science and Engineering, CSI College of Engineering, Ketti*

***Abstract*—The preservation of water quality is vital to human well-being since it is an essential and necessary resource for maintaining human life. The water contamination presents serious health dangers, such as illnesses. In such as cholera, diarrhea, and other watery illnesses. Therefore, maintaining clean and safe water becomes essential to advancing public health. According to recent research, it is estimated that water-related ailments claim the lives of 3,575,000 people annually. As a result, precise water quality forecasting could significantly lower the prevalence of these illnesses. Algorithms for machine learning have become extremely good at forecasting water quality, allowing for accurate and timely monitoring of water resources. In this project, we used Machine Learning models including the Random Forest Classifier, Decision Tree, Support Vector Machine, and K-Nearest Neighbor Classifier. A dataset comprising parameters such as pH, hardness, total dissolved solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity is used to train the models. Using measures like precision and recall, the algorithms are evaluated to determine the Water Quality Index with accuracy. When compared to other models, the Random**

I. INTRODUCTION

A. Need for The Project

Water is a vital natural resource that is essential for the survival of all living things on Earth, as it makes up 71% of the planet's surface. It is not only necessary for drinking but also plays a crucial role in industries, agriculture, and global trade via oceans and seas. Due to the importance of water for human life, research has focused on preserving water quality and preventing pollution to meet international standards.

Different water sources, such as groundwater, springs, rivers, lakes, and streams, have specific quality standards based on their intended use, such as for agricultural, industrial, or human purposes. For example, drinking water should be fresh and unpolluted, while irrigation water should not be too saline or toxic. Water used in industries has different requirements based on the nature of the industrial processes.

To ensure water quality, it is necessary to understand and predict potential risks of pollution. The aims to address this challenge by exploring the use of machine learningtechnique for water quality classification. By leveraging machine learning model, analyse andclassify water quality based on various parameters and features.

B. Objective

Human activities and practices, as well as natural processes, can significantly and alarmingly impact water quality, particularly for human. These practices and activities can lead to pollution by improperly disposing of waste and pollutants, posing significant threats to aquatic ecosystems and human well-being. Industrial plants and vehicles, for instance, are major contributors to water pollution, causing adverse effects on surface water and groundwater. Industrial activities can result in the release of various pollutants into water bodies, altering their chemical composition and overall quality. This includes the emission of harmful substances that contribute to the acidification of water sources, leading to decreased pH levels, reduced acid-neutralizing capacity, and elevated concentrations of aluminium. The acidification of water bodies not only affects the availability of clean water but also has detrimental effects on aquatic organisms and their habitats. It disrupts the ecological balance and can lead to the decline of sensitive species.

C. Project Overview

Water quality is assessed based on various features, including pH value, hardness resulting from calcium and magnesium salts, total dissolved solids (TDS),

chloramines, sulphate, electrical conductivity (EC), total organic carbon (TOC), turbidity, and trihalomethanes. To predict water quality classification (WQC), machine learning algorithm offer valuable tools for data Pre-processing, handling missing data, removing feature correlations, applying classification techniques, and analysing the significance of feature selection.

1.1 Background and Motivation

Water is a vital natural resource that is essential for the survival of all living things on Earth, as it makes up 71% of the planet's surface. It is not only necessary for drinking but also plays a crucial role in industries, agriculture, and global trade via oceans and seas. Due to the importance of water for human life, research has focused on preserving water quality and preventing pollution to meet international standards.

Different water sources, such as groundwater, springs, rivers, lakes, and streams, have specific quality standards based on their intended use, such as for agricultural, industrial, or human purposes. For example, drinking water should be fresh and unpolluted, while irrigation water should not be too saline or toxic. Water used in industries has different requirements based on the nature of the industrial processes.

To ensure water quality, it is necessary to understand and predict potential risks of pollution. The aims to address this challenge by exploring the use of machine learning technique for water quality classification. By leveraging machine learning model, analyse and classify water quality based on various parameters and features.

1.2 purpose and goal of the project

Human activities and practices, as well as natural processes, can significantly and alarmingly impact water quality, particularly for human. These practices and activities can lead to pollution by improperly disposing of waste and pollutants, posing significant threats to aquatic ecosystems and human well-being. Industrial plants and vehicles, for instance, are major contributors to water pollution, causing adverse effects on surface water and groundwater. Industrial activities can result in the release of various pollutants into water bodies, altering their chemical composition and overall quality. This includes the emission of harmful substances that contribute to the acidification of water sources, leading to decreased pH levels, reduced acid-neutralizing capacity, and elevated concentrations of aluminium. The acidification of water bodies not only affects the availability of clean water but also has detrimental effects on aquatic organisms and their habitats. It disrupts the ecological balance and can lead to the decline of sensitive species.

1.3 organization of the project

Water quality is assessed based on various features, including pH value, hardness resulting from calcium and magnesium salts, total dissolved solids (TDS), chloramines, sulphate, electrical conductivity (EC), total organic carbon (TOC), turbidity, and trihalomethanes. To predict water quality classification (WQC), machine learning algorithm offer valuable tools for data Pre-processing, handling missing data, removing feature correlations, applying classification techniques, and analysing the significance of feature selection.

II. LITERATURE REVIEW

A. Problems in The Existing System

Water is a valuable resource and sustains almost all lives on earth. The depletion of fresh water is a serious concern. Water is essential for the continuation of life and ensuring the safety and accessibility of drinking water is a pressing global issue. Potable water is water that is obtained from the surface and ground sources. It is also called drinking water and is treated to the levels that are suitable for drinking.Water potability testing looks for coliform bacteria, improper pH, sodium, chloride nitrate, sulfate, manganese, iron, water hardness, and the total dissolved solids in the water.There has been a lot of research on using machine learning in the water quality index (WQI), water quality classification (WQC), and wastewater treatment. In a study by various machine learning techniques, including random forests, neural network, multinomial logistics regression, support vector machine, and bagged tree models, were applied to classify a dataset of water quality in India. Their results showed that nitrate, pH, conductivity, dissolved oxygen, total coliform, and biological oxygen demand are the main factors that affect WQC.The used machine learning technique to select quality features for the system model performed the best in predicting features such as electrical conductivity, sodium absorption ratio, and total hardness. It used data

collected through the Internet of Things and a neural network machine learning technology to forecast water pollution in residential overhead tanks.

The techniques used and their advantages and disadvantages of references papers are,

Water quality prediction and classification based on principal component regression and gradient boosting classifier approach

Authors: Md. Saikat Islam Khan a,d , Nazrul Islam b,d,⇑ , Jia Uddin c , Sifatul Islam a,d , Mostofa Kamal Nasir a,d

Publisher: Journal of King Saud University - Computer and Information Sciences · June 2021

Abstract

Estimating water quality has been one of the significant challenges faced by the world in recent decades. This paper presents a water quality prediction model utilizing the principal component regression technique. Firstly, the water quality index (WQI) is calculated using the weighted arithmetic index method. Secondly, the principal component analysis (PCA) is applied to the dataset, and the most dominant WQI parameters have been extracted. Thirdly, to predict the WQI, different regression algorithms are used to the PCA output. Finally, the Gradient Boosting Classifier is utilized to classify the water quality status. The proposed system is experimentally evaluated on a Gulshan Lake-related dataset. The results demonstrate95% prediction accuracy for the principal component regression method and 100% classification accuracy for the Gradient Boosting Classifier method, which show credible performance compared with the stateof-art model.

Research Paper on Analysing impact of Various Parameters on Water Quality Index Authors: Divya Bhardwaj and Neetu Verma 1 M. TECH Scholar

Publisher: International Journal of Advanced Research in Computer Science.

https://www.researchgate.net/publication/351344665

Abstract

Water is a limited natural resource. Therefore, preserving water is very important for protection of our environment. Various water quality monitoring systems have been developed to measure concentration of the constituents in quantity for characterisation of water fordifferent uses. Water quality can be estimated through quality index which in turn is analysed through various parameters such as pH level, Turbidity, Dissolved Oxygen, Conductivity etc. This paper addresses the impact of parameters on water quality index. Moreover, the paper also depicts how water can be utilised based on various values of parameters.

Comparison of Water Quality Classification Models using Machine Learning Authors: Neha Radhakrishnan, Anju S Pillai.

Publisher:International Conference on Communication and Electronics Systems (ICCES 2020).

https://www.researchgate.net/publication/342852709

Abtract

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. Afterassessing the obtained results, the decision tree algorithm was found to be a better classification model with an accuracy of 98.50%.

III. SYSTEM REQUIREMENT

3.1 HARDWARE REQUIREMENT

System -Windows 10/11 Speed
-2.4GHZ
Hard disk -80 GB Monitor
-15VGA Color Ram
-8 GB

3.2 SOFTWARE REQUIREMENT

Front End -Python Backend
-Weka Tool

IV. SYSTEM DESIGN

In our proposed system they have following modules are

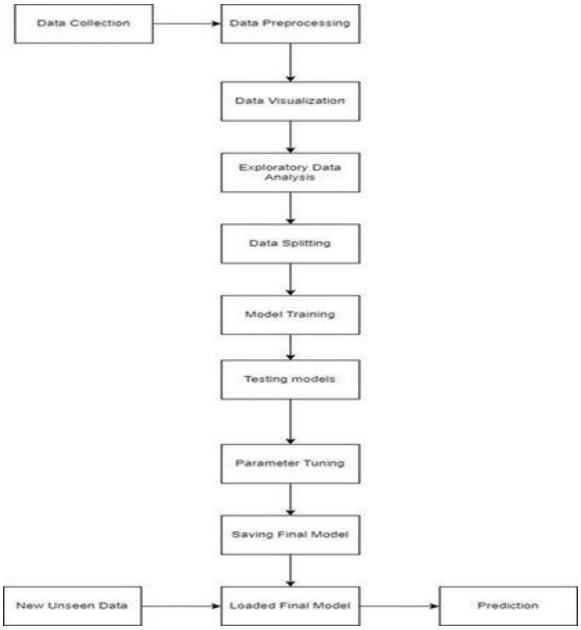

Fig. 4.1. System Design

V. CHAPTER 5 DATASET

5.1 Dataset Description:

The dataset used in this study is sourced from Kaggle and contains comprehensive information onwater samples. A total of 3,276 samples were collected and analyzed,making it a substantial dataset for conducting a robust analysis. The dataset consists of 3,276 water samples, each containing information on nine important parameters. The chapter discusses the significance of these parameters in evaluating water qualityand ensuring its safety for human consumption.
The dataset link:
https://www.kaggle.com/datasets/adityakadiwal/water-potability
The dataset includes the following hydro-chemical parameters and portability labels:

5.1.1. pH Value:

The pH value is an essential parameter for assessing the acid-base balance of water. Itserves as anindicator of the water's acidic or alkaline condition. The World Health Organization (WHO)

has recommended a maximum permissible pH range of 6.5 to 8.5 for drinking water.

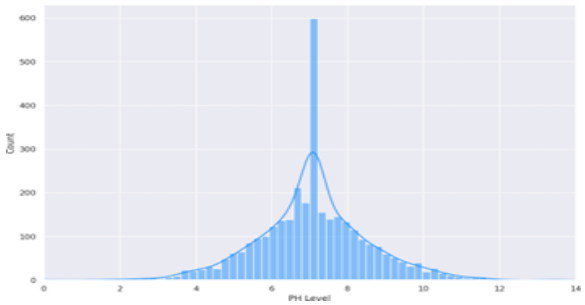

Figure: PH Level

5.1.2. Hardness:

Hardness is primarily caused by calcium and magnesium salts dissolved from geological deposits.It is a measure of water's capacity to precipitate soap due to the presence of these minerals.

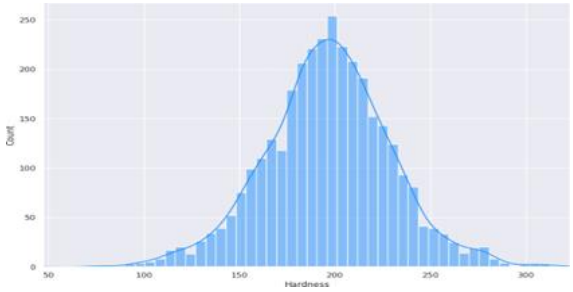

Fig: Hardness

5.1.3. Total Dissolved Solids (TDS):

TDS refers to the ability of water to dissolve inorganic and some organic minerals orsalts such aspotassium, calcium, sodium, and bicarbonates. High TDS levels can leadto an undesirable taste and diluted color in water. The WHO has set a desirable limitof 500 mg/l and a maximum limit of 1000 mg/l for TDS in drinking water.

7.1.4. Chloramines:

Chlorine and chloramine are commonly used disinfectants in public water systems. Chloramines are formed when ammonia is added to chlorine for water treatment. The concentration of chlorinein drinking water is considered safe up to 4 milligrams per liter (mg/L) or 4 parts per million (ppm).

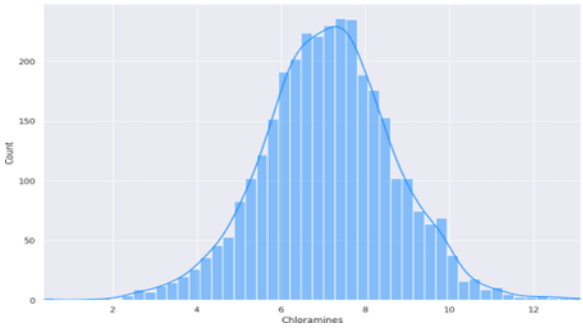

Fig: Chloramines

5.1.5. Sulfate:

Sulfates are naturally occurring substances found in minerals, soil, rocks, groundwater, plants, andfood. They have various industrial uses and can be present in different concentrations in freshwatersupplies.

The concentration of sulfates in seawater is around 2,700 mg/L, while freshwater supplies typicallyrange from 3 to 30 mg/L. In certain locations, sulfate concentrationscan be as high as 1000 mg/L.

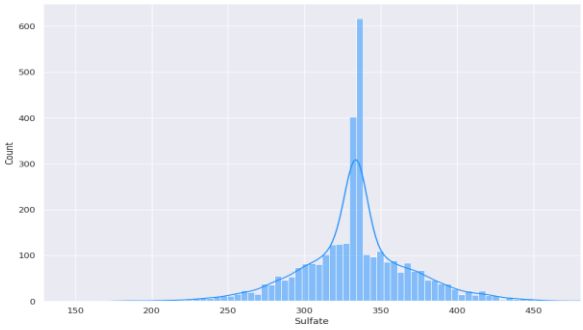

Fig: Sulfate

5.1.6. Conductivity:

Conductivity measures the ability of water to conduct electrical current. It is influenced by the concentration of ions present in the water, which in turn determinesthe level of dissolved solids. The WHO recommends that the electrical conductivity (EC) value should not exceed 400 µS/cm.

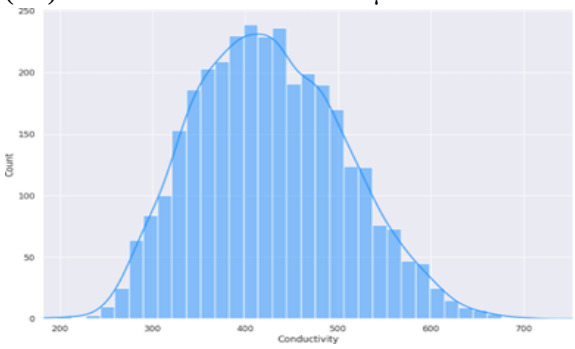

Fig: Conductivity

5.1.7. Organic Carbon:

Total Organic Carbon (TOC) in water originates from decaying natural organic matter and synthetic sources. It represents the total amount of carbon in organic compoundsdissolved in water. The US Environmental Protection Agency (EPA) has set limits of<2 mg/L as TOC in treated/drinking water and <4 mg/L in source water used for treatment.

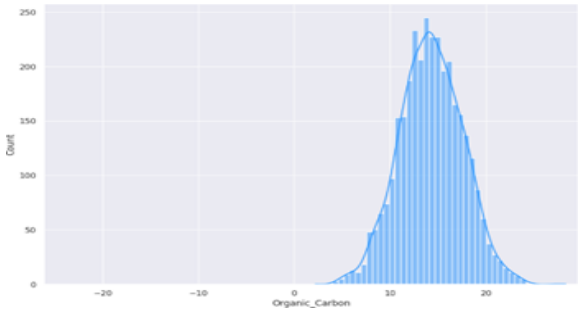

Fig: Organic Carbon

5.1.8. Trihalomethanes:

Trihalomethanes (THMs) are chemicals that can be found in water treated with chlorine. The concentration of THMs depends on factors such as the level of organicmaterial in the water, the amount of chlorine used for treatment, and the water's temperature.

THM levels up to 80 ppm are considered safe in drinking water.

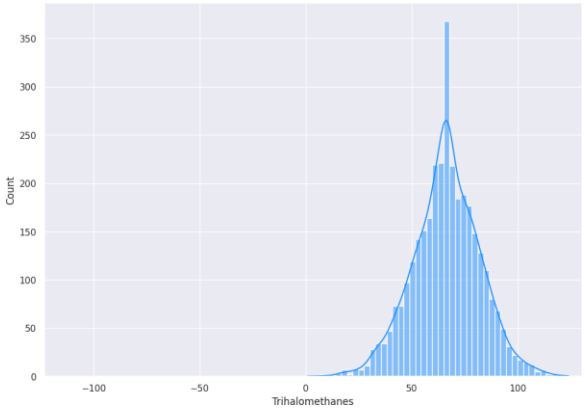

Fig..Trihalomethanes

5.1.9. Turbidity:

Turbidity is a measure of the quantity of solid matter present in water in a suspendedstate. It reflects the water's light-emitting properties and is often used as an indicator of the quality of wastewater discharge in terms of colloidal matter. The WHO recommends a maximum turbidity value of 5.00 NTU (Nephelometric Turbidity Units

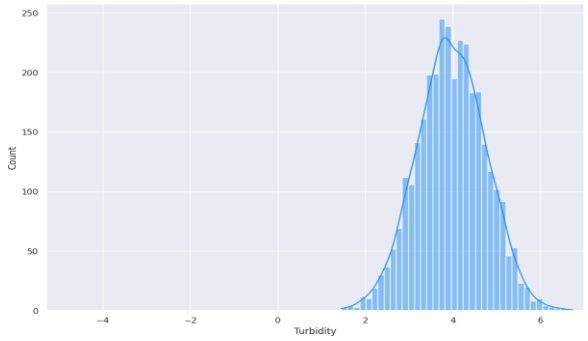

Fig: Turbidity

7.2 Dataset Split:

To perform the analysis, the dataset was divided into training and testing sets. Approximately 75%of the samples (2,457) were allocated for training, while the remaining 25% (819 samples) were used for testing the models. This division ensuresthe reliability and generalizability of the resultsobtained from the analysis.

7.3 Missing Value Imputation:

To handle missing values in the dataset, single imputation was employed. Specifically, missing values labeled as "potable" were imputed using the mean of allnon-missing "potable" samples. Similarly, missing values labeled as "non-potable" were imputed using the mean of non- missing "non-potable" samples. This imputationtechnique ensures that the missing values are replaced with reasonable estimates based on the respective potability labels.

The pH value of water is essential for assessing its acid-base balance and determiningits acidic oralkaline nature. The World Health Organization (WHO)

REFERENCES

[1] Md. Saikat Islam Khan a,d , Nazrul Islam b,d,⇑, Jia Uddin c , Sifatul Islam a,d , Mostofa Kamal Nasir .” Water quality prediction and classification based on principal component regression and gradient boosting classifier approach” Water Resources ResearchGate (2021): DOI: 10.1016/j.jksuci.2021.06.003.

[2] Divya Bhardwaj and Neetu Verma M. TECH Scholar,” Research Paper on Analysing impact of Various Parameters on Water Quality Index.” International Journal of Advanced Research in Computer Science Volume 8, No. 5, (2017): 0976-5697.

[3] Neha Radhakrishnan, Anju S Pillai., “Comparison of Water Quality Classification Models” ResearchGate (2020): ISBN: 978-1-7281-5371-1.

[4] Nur Hanisah Abdul Malek, Wan Fairos Wan Yaacob, Syerina Azlin Md Nasir and Norshahida Shaadan..” Prediction of Water Quality Classification of the Kelantan RiverBasin, Malaysia, Using Machine Learning Techniques” (2018): 33-47.

[5] MoslehHmoud Al-Adhaileh, * and FawazWaselallahAlsaade.” Modelling and Prediction of Water Quality by Using Artificial Intelligence”,MPMD- 2021, 13(8), 4259; https://doi.org/10.3390/su13084259.

[6] K.Kalaivanan and J. Vellingiri.,” Survival Study on Different Water Quality Prediction Methods Using Machine Learning”, Nature Environment and Pollution Technology an International QuarterlyScientific Journal., (2021): vol (21): https://doi.org/10.46488/NEPT.2022.v21i03.032.

[7] Dr. Sanjeev Singh, Dr. Dilkeshwar Pandey Shashwat Singh, Anurag Shrivastava, Pankaj Kumar,Prajwal Upman.,” Water Quality Prediction Using Machine Learning”, Section A-Researchpaper(2023),vol(1502-1509): doi: 10.48047/ecb/2023.12.si6.138.

[8] Vijay Anand M1, Chennareddy Sohitha1, Galla Neha Saraswathi1 and Lavanya G,”Water quality prediction using CNN”, Journal of Physics: Conference Series, 2484 (2023) 012051, doi:10.1088/1742-6596/2484/1/012051.

[9] Md. Saikat Islam Khan, Nazrul Islam b,, Jia Uddin c , Sifatul Islam , Mostofa Kamal Nasir Santosh, Tang,” Water quality prediction and classification based on principal component regressionand gradient boosting classifier approach”,https://doi.org/10.1016/j.jksuci.2021.06.003.

[10] Dakhaz Mustafa Abdullah, Adnan Mohsin Abdulazeez.” Machine Learning Applications based on SVM Classification: A Review”, Qubahan academic journals, Doi: 10.48161/Issn.2709-8206,2021.

[11] Mahmoud Y. Shams, Ahmed M. Elshewey, El-Sayed M. El-kenawy3, Abdelhameed Ibra him4, Fatma M. Talaat1, Zahraa Tarek,” Water quality prediction using machine learning models based on grid search method”, Multimedia Tools and

Applications https://doi.org/10.1007/s11042-023-16737-4

[12] Dao Nguyen Khoi, Nguyen Trong Quan, Do Quang Linh, Pham Thi Thao Nhi and NguyenThi Diem Thuy," Using Machine Learning Models for Predicting the Water Quality Index in the La Buong River, Vietnam",MDPI (2022),vol- 14, 1552. https://doi.org/ 10.3390/w14101552.

[13] Chao-ying joanne peng Kuk lida lee Gary m. Ingersoll," An Introduction to Logistic RegressionAnalysis and Reporting",ResearchGate, September 2002,DOI: 10.1080/00220670209598786.

[14] Jehad Ali1, Rehanullah Khan, Nasir Ahmad," Random Forests and Decision Trees", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012, ISSN (Online): 1694-0814.

[15] Gongde Guo1, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer," KNN Model Based Aproach in Classification", https:// www.researchgate. net/ public ation/ 2948052.