PROJECT REPORT

EFFICIENT WATER QUALITY ANALYSIS AND PREDICTION USING MACHINE LEARNING

By team – PNT2022TMID38291

Sri Venkateswaraa College Of Technology

Prabakaran S (412619104026)

Raja P (412619104030)

Kaviya N (412619104017)

Perumal K (412619104025)

Under the guidance of,

Ms. Preethisri C (Mentor)

Ms. Lalitha Gayathri (Industry Mentor)

Mr. Rakesh (Trainer)

TABLE OF CONTENTS

1.INTRODUCTION

Water is Basic necessity for all of the humans and lives. The quality of water is a major concern for people living in urban areas. The quality of water serves as a powerful environmental determinant and a foundation for the prevention and control of waterborne diseases. However, predicting the urban water quality is a challenging task since the water quality varies in urban spaces non-linearly and depends on multiple factors, such as meteorology, water usage patterns, and land uses, so this project aims at building a Machine Learning (ML) model to Predict Water Quality by considering all water quality standard indicators. Water is the most important of sources, vital for sustaining all kinds of life; however, it is in constant threat of pollution by life itself. Water is one of the most communicable mediums with a far reach. Rapid industrialization has consequently led to deterioration of water quality at an alarming rate. Poor water quality results have been known to be one of the major factors of escalation of harrowing diseases. As reported, in developing countries, 80% of the diseases are water borne diseases, which have led to 5 million deaths and 2.5 billion illnesses. The most common of these diseases in Pakistan are diarrhea, typhoid, gastroenteritis, cryptosporidium infections, some forms of hepatitis and giardiasis intestinal worms. In Pakistan, water borne diseases, cause a GDP loss of 0.6–1.44% every year. This makes it a pressing problem, particularly in a developing country like Pakistan. Water quality is currently estimated through expensive and time-consuming lab and statistical analyses, which require sample collection, transport to labs, and a considerable amount of time and calculation, which is quite ineffective given water is quite a

communicable medium and time is of the essence if water is polluted with disease- inducing waste. The horrific consequences of water pollution necessitate a quicker and cheaper alternative. In this regard, the main motivation in this study is to propose and evaluate an alternative method based on supervised machine learning for the efficient prediction of water quality in real-time. A representative set of supervised machine learning algorithms were employed on the said dataset for predicting the water quality index (WQI) and water quality class (WQC). The main contributions of this study are summarized as follows. A first analysis was conducted on the available data to clean, normalize and perform feature selection on the water quality measures, and therefore, to obtain the minimum relevant subset that allows high precision with low cost. In this way, expensive and cumbersome lab analysis with specific sensors can be avoided in further similar analyses. A series of representative supervised prediction (classification and regression) algorithms were tested on the dataset worked here. The complete methodology is proposed in the context of water quality numerical analysis.

1.1 PROJECT OVERVIEW

With the rapid increase in the volume of data on the aquatic environment, machine learning has become an important tool for data analysis, classification, and prediction. Unlike traditional models used in water-related research, data-driven models based on machine learning can efficiently solve more complex nonlinear problems. In water environment research, models and conclusions derived from machine learning have been applied to the construction, monitoring, simulation, evaluation, and optimization of various water treatment and management systems. Additionally, machine learning can provide solutions for water pollution control, water quality improvement, and watershed ecosystem security management. In this review, we describe the cases in which machine learning algorithms have been applied to evaluate the water quality in different water environments, such as surface water, groundwater, drinking water, sewage, and seawater. Furthermore, we propose possible future applications of machine learning approaches to water environments.

1.2. PURPOSE

As per the United Nations (UN) report, about 1.5 million people die each year because of contaminated water-driven diseases. In developing countries, it is announced that 80% of health problems are caused by contaminated water. Five million deaths and 2.5 billion illnesses are reported annually Such a mortality rate is higher than deaths resulting from accidents, crimes, and terrorist attacks.

Therefore, it is very important to suggest new approaches to analyse and, if possible, to predict the water quality (WQ). It is recommended to consider the temporal dimension for forecasting the WQ patterns to ensure the monitoring of the seasonal change of the WQ. However, using a special variation of models together to predict the WQ grants better results than using a single model. There are several methodologies proposed for the prediction and modelling of the WQ. These methodologies include statistical approaches, visual modelling, analysing

algorithms, and predictive algorithms. For the sake of the determination of the correlation and relationship among different water quality parameters, multivariate statistical techniques have been employed. The geostatistical approaches were used for transitional probability, multivariate interpolation, and regression analysis.

Massive increases in population, the industrial revolution, and the use of fertilizers and pesticides have led to serious effects on the WQ environments. Thus, having models for the prediction of the WQ is of great help for monitoring water contamination.

2. LITERATURE SURVERY

Many works had been conducted to predict water quality using Machine Learning (ML) approaches. Some researchers used the traditional Machine Learning models, such as Decision Tree [13][14], Artificial Neural Network [2][5][6][7], Support Vector Machine [8][9][0], K-Nearest Neighbors [21] and Naïve Bayes [18][22][23]. However, in recent years, some researchers are moving towards more advanced ML ensemble models, such as Gradient Boosting and Random Forest [1]

Traditional Machine Learning models, such as the Decision Tree model, are frequently found in the literature and performed well on water quality data. However, decision-tree-based ensemble models, including Random Forest (RF) and Gradient Boosting (GB), always outperform the single decision tree [4]. Among the reasons for this are its ability to manage both regular attributes and data, not being sensitive to missing values and being highly efficient. Compared to other ML models, decision-tree-based models are more favorable to short-term prediction and may have a quicker calculation speed [6]. Gakii and Jepkoech [3] compared five different decision tree classifiers, which are Logistic Model Tree (LMT), J48, Hoefflin tree, Random Forest and Decision Stump. They found that J48 showed the highest accuracy of 94%, while Decision Stump showed the lowest accuracy. Another study by Jeyhun et al. [4] also compared five decision-tree-based models, which are Random Tree, Random Forest, Ordinary Decision Tree (ODT), Chi-square Automatic Interaction Detector and Iterative Dichotomies 3 (ID3), to determine high water quality zones. They found that ODT and Random Forest produce higher accuracy compared to the other algorithms and the methods are more suitable for continuous datasets.

Another popular Machine Learning model to predict water quality is Artificial Neural Network (ANN). ANN is a remarkable data-driven model that can cater both linear and non-linear associations among output and input data. It is used to treat the non-linearity of water quality data and the uncertainty of contaminant source. However, the performance of ANN can be obstructed if the training data are imbalanced and when all initial weights of the parameter have the same value. In India, Aradhana and Singh [8] used ANN algorithms to predict water quality. They found that Levenberg Marquardt (LM) algorithm has a better performance than the Gradient Descent Adaptive (GDA) algorithm. Abyan [5] used ANN and multivariate linear regression models in his research and found that the ANN model outperforms the MLR model. However, the research only assessed the performance of the ANN model using root-mean-square error (RMSE), coefficient of correlation (r) and bias values. Although ANN models are the most broadly used, they have a drawback as the prediction power becomes weak if they are used with a small dataset and the testing data are outside the range of the training data [8].

Support Vector Machine has also been extensively used in water quality studies. Some studies proved that SVM is the best model in predicting water quality compared to other models. A study by Babbar and Babbar [11] found that Support Vector Machine and Decision Tree are the best classifiers because they have the lowest error rate, which is 0%, in classifying water quality class compared to ANN, Naive Bayes and K-NN classifiers. It also revealed that ML models can quickly determine the water quality class if the data provided represent an accurate representation of domain knowledge. In China, Liu and Lu [12] developed the SVM and ANN model to predict phosphorus and nitrogen. They found that SVM model achieves a better forecasting accuracy compared to the ANN model. This is because the SVM model optimizes a smaller number of parameters acquired from the principle of structural risk minimization, hence avoiding the occurrence of overtraining data to have a better generalization ability [12]. This is supported by another study in Eastern Azerbaijan, Iran [6]. They found that SVM has a better performance compared to the K-Nearest Neighbor algorithm in estimating two water quality parameters, which are total dissolved solid and conductivity. The results showed smaller error and higher R^2 than the results attained in Abbasi et al.'s report [4]. Naïve Bayes has also been widely used for predicting water quality. A study by Vijay and Kamaraj [2] found that Random Forest and Naïve Bayes produce better accuracy and low classification error compared to the C5.0 classifier. However, traditional ML models, for example, Decision Tree, ANN, Naïve Bayes and SVM, do not perform well. They have some weaknesses, such as a high tendency to be biased and a high variance [22]. For example, SVM uses the structural risk minimization principle to address overfitting problem in Machine Learning by reducing the model's complexity and fitting the training data successfully [9]. Meanwhile, the Bayes model uses prior and posterior probabilities in order to prevent overfitting problems and bias from using only sample information. In ANN, the training process takes a longer time and overfitting problems may occur if there are too many layers, while the prediction error may be affected if there are not enough layers [30]. Overfitting is a fundamental issue in supervised Machine Learning that prevents the perfect generalization of the model to fit the data observed on the training data, as well as unseen data on the testing set. Hence, overfitting occurs due to the presence of noise, a limited training set size, and classifier complexity [30]. One of the strategies considered by many previous works to reduce the effects of overfitting is to adopt more advanced methods, such as the ensemble method.

The ensemble method is a Machine Learning technique that combines several base learners' decisions to produce a more precise prediction than what can be achieved with having each base learner's decision [6]. This method has also gained wide attention among researchers recently. The diversity and accuracy of each base learner are two important features to make the ensemble learners work properly [7]. The ensemble method ensures the two features in several ways based on its working principle. There are two commonly used ensemble families in Machine Learning, which are bagging and boosting. Both the bagging and boosting

methods provide a higher stability to the classifiers and are good in reducing variance. Boosting can reduce the bias, while bagging can solve the overfitting problem [1]. A famous ensemble model that uses the bagging algorithm is Random Forest. It is a classification model that uses multiple base models, typically decision trees, on a given subset of data independently and makes decisions based on all

models [5]. It uses feature randomness and bagging when building each individual decision tree to produce an independent forest of trees. Random Forest carries all the advantages of a decision tree with the added effectiveness of using several models [2]. Another popular ensemble model is Gradient Boosting. Gradient Boosting is a Machine Learning technique that trains multiple weak classifiers, typically decision trees, to create a robust classifier for regression and classification problems. It assembles the model in a stage-wise way similar to other boosting techniques and it generalizes them by optimizing a suitable cost function. In the GB algorithm, incorrectly classified cases for a step are given increased weight during the next step. The advantages of GB are that it has exceptional accuracy in predicting and fast process [3]. Therefore, advanced models, such as Random Forest and Gradient Boosting, should be employed to cater for the lack of basic ML models.

2.1 EXISITNG PROBLEM

the main problem lies here. For testing the water quality, we have to conduct lab tests on the water which is costly and time-consuming as well. So, in this paper, we propose an alternative approach using artificial intelligence to predict water quality. This method uses a significant and easily available water quality index which is set by the WHO (World Health Organisation). The data taken in this paper is taken from the PCPB India which includes 3277 examples of the distinct wellspring. In this paper, WQI (Water Quality Index) is calculated using AI techniques. So, in future work, we can integrate this with IoT based framework to study large datasets and to expand our study to a larger scale. By using that it can predict the water quality fast and more accurately than any other IoT framework. That IoT framework system uses some limits for the sensor to check the parameters like ph., Temperature, Turbidity, and so on. And further after reading this parameter pass these readings to the Arduino microcontroller and ZigBee handset for further prediction

2.2 REFERENCES

1. Ling, J.K.B. Water Quality Study and Its Relationship with High Tide and L ow Tide at Kuantan River. Bachelor's Thesis, University Malaysia Pahang, Gam bang, Malaysia, 2010. Available online: http://umpir.ump.edu.my/id/e print/2449/1/JACKY_LING_KUO_BAO.PDF (accessed on 22 February 20 22).
2. Xu, J.; Gao, X.; Yang, Z.; Xu, T. Trend and Attribution Analysis of Runoff Changes in the Weihe River Basin in the Last 50 Years. Water 2022, 14, 47.
3. Wahab, M.A.A.; Jama don, N.K.; Mahmood, A.; Shahir, A. River Pollution Relationship to the National Health Indicated by Under-Five Child Mortalit y Rate: A Case Study in Malaysia. Bioremediate. Sci. Technol. Res. 2015, 3 , 20–25.
4. Abbasi, T.; Abbasi, S.A. Water Quality Indices; Elsevier: Amsterdam, The Netherlands, 2012.

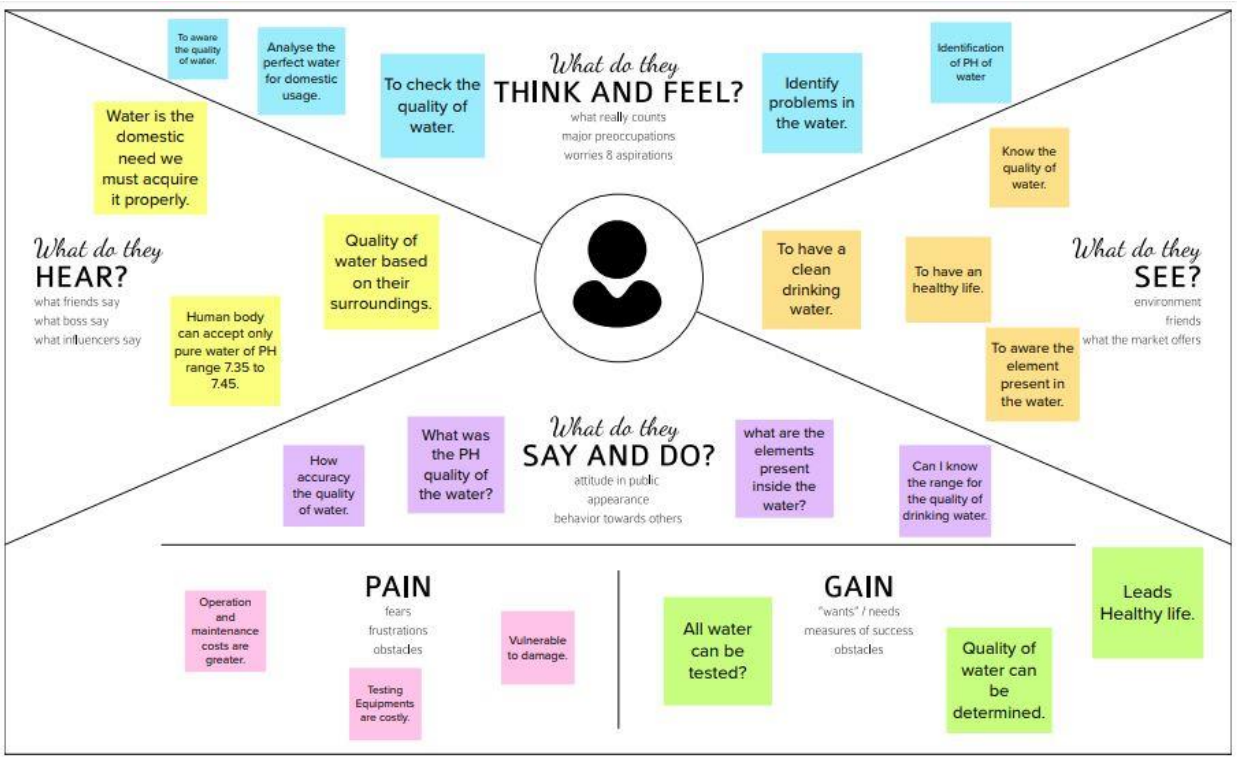
5. Abyan, H.Z. Evaluation of multivariate linear regression and artificial neura l networks in prediction of water quality parameters. J. Environ. Health Sci. Eng. 2014, 12, 40.
6. Alias, S.W.A.N. Ecosystem Health Assessment of Sungai Peng Kalan Chep a Basin: Water Quality and Heavy Metal Analysis. Sains Malays. 2020, 49, 1787–1798.
7. Al-Badain, F.; Shyheim-Othman, M.; Gasim, M.B. Water quality assessmen t of the Semenyih river, Selangor, Malaysia. J. Chem. 2013, 2013, 871056.
8. Asadullah, S.B.H.S.; Shariati, A.; Motta, D.; Yaseen, Z.M. River water qual ity index prediction and uncertainty analysis: A comparative study of machi ne learning models. J. Environ. Chem. Eng. 2021, 9, 104599.
9. Chen, K.; Chen, H.; Zhou, C.; Huang, Y.; Qi, X.; Shen, R.; Liu, F.; Zuo, M. ; Zou, X.; Wang, J. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different mac hine learning models based on big data. Water Res. 2020, 171, 115454.
10. Larios, J.L.; Villarica, M.V. Pattern Extraction of Water Quality Prediction Using Machine Learning Algorithms of Water Reservoir. Int. J. Mech. Eng. Robot. Res. 2019, 8, 992–997.
11. Sangrur, B.; Koku, R.; Ates, A. Water quality assessment using artificial int elligence techniques: SOM and ANN—A case study of Melen River Turkey . Water Qual. Expo. Health 2015, 7, 469–490.
12. Aradhana, G.; Singh, N.B. Comparison of Artificial Neural Network algorit hm for water quality prediction of River Ganga. Environ. Res. J. 2014, 8, 55 –63.

2.3 PROBLEM STATEMENT DEFINITION

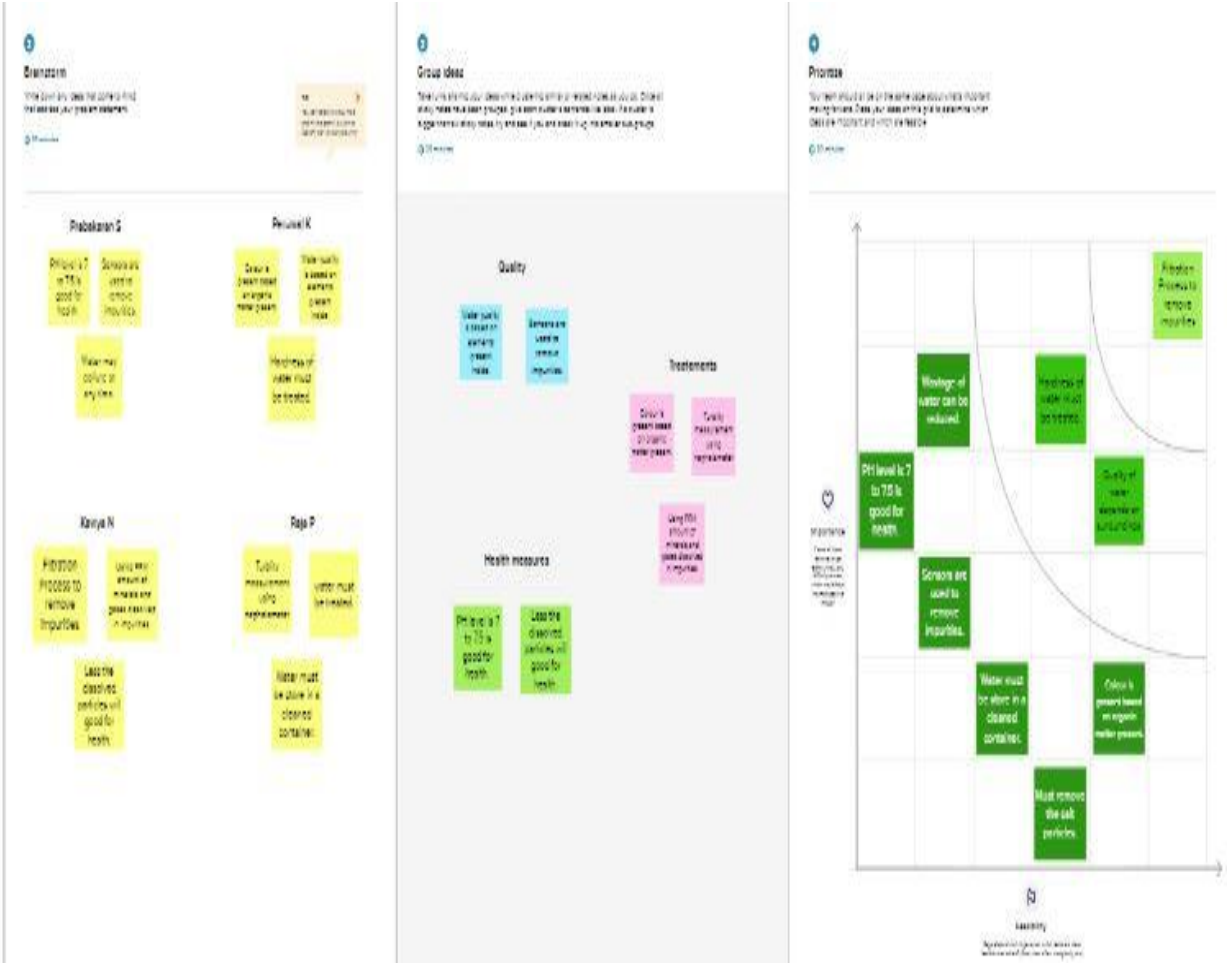
To predict the water safe or not for Access to safe drinking-water is essential to he alth, a basic human right and a component of effective policy for health protection . This is important as a health and development issue at a national, regional and lo cal level. In some regions, it has been shown 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.

3.IDEATION AND PROPOSED SOLUTION

3.1 EMPATHY MAP CANVAS



3.2 IDEATION AND BRAINSTORMING



3.3 PROPOSED SOLUTION

S. No	Parameter	Description
1.	Problem Statement (Problem to be solved)	Water is considered as a vital resource that affects various aspects of human health and lives. The quality of water is a major concern for people living in urban areas and the water is also most likely to become contaminated due to various factors including human, industrial and commercial activities as well as natural processes. In addition to that, poor sanitation infrastructure and lackof awareness also contributes immensely to drinking watercontamination.
2.	Idea / Solution description	It is not possible to check the qualityof water manually every time. So, anautomatic real-time monitoring system is implemented based on machine learning technique to forecast the quality of water and to predict the health of water according to its quality parameter level.
3.	Novelty / Uniqueness	• User Friendly • Determining the reuse and recycle of water • Detecting Quality parametric values.
4.	Social Impact / Customer Satisfaction	Customer satisfaction is an important factor to consider in total quality management. In order to achieve this goal, it is important one.
5.	Business Model (Revenue Model)	First the application is processed with real time data. Later it comes into the picture where everyone can see the networking, conducting various activity and testing to them.
6.	Scalability of the Solution	Helps in getting all required aspects regarding quality of water.

3.3 PROBLEM SOLUTION FIT

Problem - Solution Fit

Team Id: PNT2022TMID38291

IBM

to CL

1. CUSTOMER SEGMENT(S) CS

Water is the basic necessity for all kind of living beings. water is been used by every source of people in different areas such as Residential & commercial areas, testing purposes, etc.. All this we need quality and purified water. It impact the water quality monitoring management.

6. CUSTOMER LIMITATIONS EG. BUDGET, DEVICES CL

The quality testing required some basic set of budget required. If the water is not at standard quality it is an serious thread to all the people. Because water is essential one for all to sustain. Sometimes it may cause disease and it will affect the people,

5. AVAILABLE SOLUTIONS PLUSES & MINUSES AS

The available solution is finding water quality index (WQI) and water quality class (WQC).

Merits: It checks the turbitity, Ph, TDS, Hardness.

Demerits: It would identify the limited parameters in water.

Focus on PR, tap into BE, understand RC

2. PROBLEMS / PAINS + ITS FREQUENCY PR

It is very difficult to find the pure drinking water. Because it need more proof to be an qualified water. The rising water pollution ,resulting in lab testing to imperative reliability and accuracy and directly include the drinking water. The main problem is impurities present in the water.

9. PROBLEM ROOT / CAUSE RC

I Identify appropriate solution.

II Collect sufficient amount of data.

III Identify the associated casual factor.

Team no: PNT2022TMID38291

7. BEHAVIOR + ITS INTENSITY BE

Water quality analyst analyse the quality and develop policies and plans for control the factor which produce impurities.They conduct chemical,physical and biological test to define water quality standard.

to BE, understand RC

3. TRIGGERS TO ACT TR

This triggers to discover the pattern in user data and then make prediction based on intricate pattern for analyzing the quality of water. It also helps to improve the efficiency and more protected to drink

4. EMOTIONS BEFORE / AFTER EM

Before there is no technology to analyse the water quality so it cause problem in health issue. It cause disease such as diarrhea, dysentery, hepatitis, typhoid, polio and cholera. But now a days it is decreased because of Water monitoring system and methods of finding pure water.

10. YOUR SOLUTION SL

Using Advanced Artificial Intelligence seven significant parameters and developed models were evaluated based on some statistical parameters based on Naive Bayes algorithm, K Nearest Neighbor (KNN), Support Vector Machine (SVM) and Linear regression algorithm

8. CHANNELS of BEHAVIOR CH

ONLINE

Helps to notify the data preprocessing information.

OFFLINE

By attaining the standard quality of satisfy all parameterit is consider as pure water.

Extr

f BE

4. REQUIREMENT ANALYSIS

4.1 FUNCTIONAL REQUIREMEN

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Registration through Form Registration through Gmail Registration through LinkedIN
FR-2	User Confirmation	Confirmation via Email Confirmation via OTP
FR-3	Executive administration	Regulation of monitoring the water environment status and regulatory compliance like pollution event emergency management, and it includes two different functions: early warning/forecast monitoring.
FR-4	Data handling	File contains water quality metrics for different water bodies.
FR-5	Quality analysis	Analyze with the acquired information of the water across various water quality indicator like (PH, Turbidity, TDS, Temperature) using different models.
FR-6	Model prediction	Confirming based on water quality index and shows the machine learning prediction (Good, Partially Good, Poor) with the percentage of presence of various parameter.
FR-7	Remote Visualization	Visualization through charts based on present and past values of all the parameter for future forecast.
FR-8	Notification services	Confirming through notification of water status prediction with parameter presence along with timestamp.

4.2 NON-FUNCTIONAL REQUIREMENT

		sources. The system is protected with the user name and password throughout the process.
NFR-3	**Reliability**	The system is very reliable as it can last for long period of time when it is well maintained. The model can be extended in large scale by increasing the datasets.
NFR-4	**Performance**	Our system should run on 32 bit (x86) or 64 bit (x64) Dual-core 2.66-GHZ or faster processor. It should not exceed 2 GB RAM.
NFR-5	**Availability**	The system should be available for the duration of the user access the system until the user terminate the access. The system response to request of the user in less time and the recovery is done is less time.
NFR-6	**Scalability**	It provides an efficient outcome and has the ability to increase or decrease the performance of the system based on the datasets.

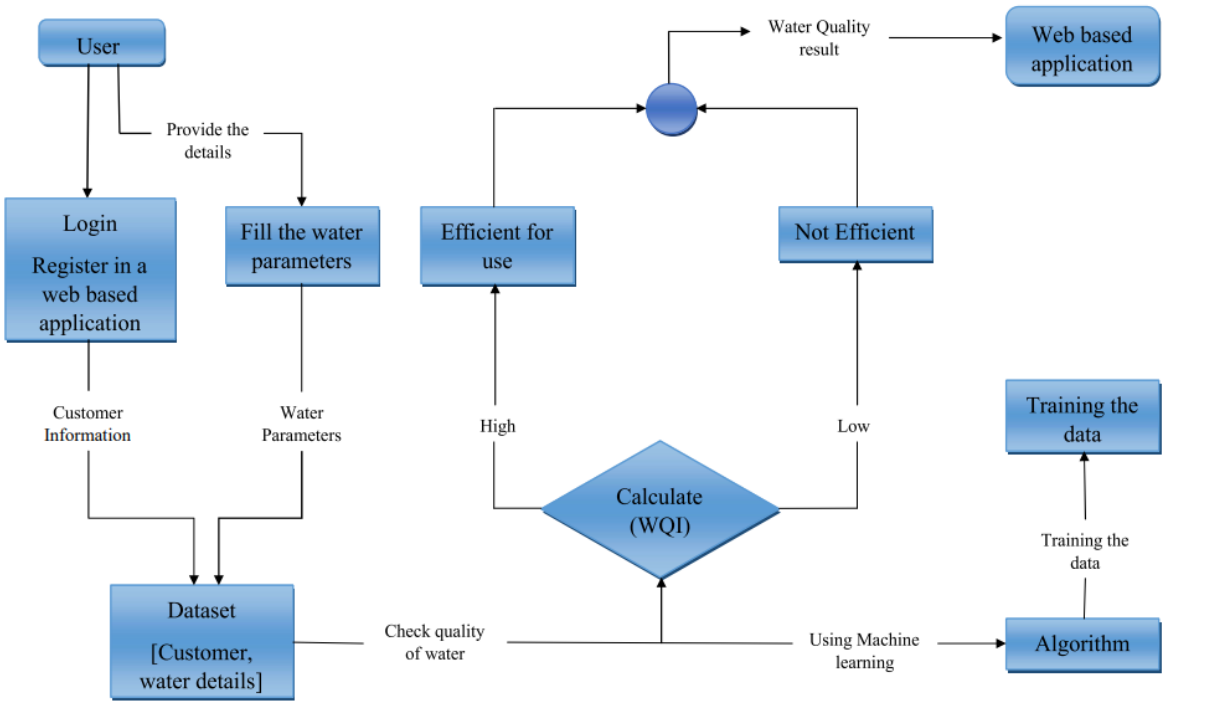
FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	The system provides a natural interaction with the users. Accurate water quality prediction with short time analysis and provide prediction safe to drink or not using some parameters and provide a great significance for water environment protection.
NFR-2	**Security**	The model enables with the high security system as the user's data will not be shared to the other

5.PROJECT DESIGN

5.1 DATA FLOW DIAGRAM

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.

DFD (Industry Standard):



5.2 SOLUTION AND TECHNINCAL ARCHITECTURE

There are basically 10 steps for making our model predict the water quality of the water samples. Those steps are: -

A. Problem Identification

In this step, we identify the problem which is solved by our model. So, the problem to be solved by our model is water quality prediction using a dataset.

B. Data Extraction: -

In this, we extract the data from the internet to train our data and predict the water quality. So, for that, we take the CPCB (Central Pollution Control Board India) dataset which contains 3277 instances of 13 different wellsprings which are collected between 2014 to 2020.

C. Data Exploration: -

In this step, we analyse the data visually by comparing some parameters of water with the WHO standards of water. It gives a slight overview of the data.

D. Data Cleaning

In this step, we clean that data like if there are some missing values in it so we replace them with mean and remove noise from the data.

F. Data Selection

In this step, we select the data types and source of the data. The essential goal of data selection is deciding fitting data type, source, and instrument that permit agents to respond to explore questions sufficiently

G. Data Splitting

In this step, we divide the dataset into smaller subsets for easing the complexity. Normally, with a two-section split, one section is utilized to assess or test the information and the other to prepare the model.

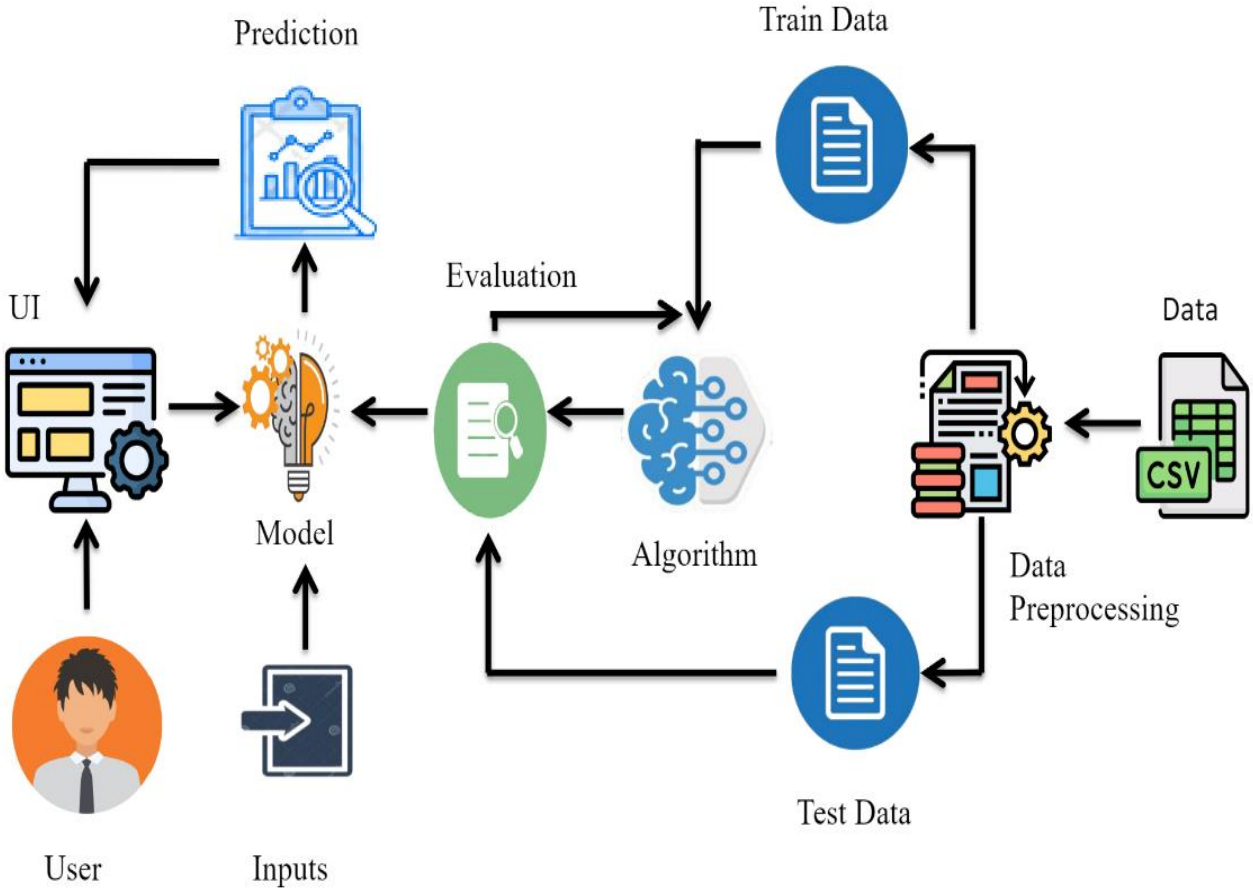
H. Data Modelling

In this step, we create a graph of the dataset for visual representation of data for better understanding. A Data Model is this theoretical model that permits the further structure of conceptual models and to set connections between data.

I. Model Evaluation

Model Evaluation is a fundamental piece of the model improvement process. In this step, we evaluate our model and check how well our model do in the future.

Solution Architecture



5.3 USER STORIES

User Stories

User Type	Functional Requirement	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Mobile user)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	I can access my account / dashboard.	High	Sprint-1
		USN-2	As a user, I will receive confirmation email once I have registered for the application.	I can receive confirmation email & click confirm.	High	Sprint-1
		USN-3	As a user, I can register for the application through Facebook.	I can register & access the dashboard with Facebook Login.	Low	Sprint-2
		USN-4	As a user, I can register for the application through Gmail		Medium	Sprint-1
	Login	USN-5	As a user, I can log into the application by entering email & password		High	Sprint-1
	Dashboard	USN-6	As a user, I can check my login details and work details		High	Sprint-1
Customer (Web user)	Web Access	USN-7	As a user, I can enter the values about the water.	I can access the webpage through internet.	High	Sprint-1
		USN-8	As a user, I can submit the values into the webpage.	I can click the submit button.	High	Sprint-2
		USN-9	As a user, I expect correct coefficient of water.		Medium	Sprint-3
	Data preprocessing	USN-10	As a user, I can see the loading information.		Medium	Sprint-3
	User Input Evaluation	USN-11	I can see the evaluation quickly.		High	Sprint-4
	Prediction	USN-12	As a user, I can see the result of the water efficient.	The results are visible on webpage.	High	Sprint-4

6. PROJECT PLANNING AND SCDULING

6.1 SPRINT PLANNING AND ESTIMATION

Product Backlog, Sprint Schedule, and Estimation:

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority
Sprint-1	Data Preparation	USN-1	Collecting water dataset and pre-processing it	20	High
Sprint-2	Model Building	USN-2	Create an ML model to predict water quality	5	Medium
Sprint-2	Model Evaluation	USN-3	Calculate the performance, error rate, and complexity of the ML model and evaluate the dataset based on the parameter that the dataset consists of.	5	Medium
Sprint-2	Model Deployment	USN-4	As a user, I need to deploy the model and need to find the results.	10	Medium
Sprint-3	Web page (Form)	USN-5	As a user, I can use the application by entering the water dataset to analyze or predict the results.	20	Medium
Sprint-4	Dashboard	USN-6	As a user, I can predict the water quality by clicking the submit button and the application will show whether the water is efficient for use or not.	20	High

6.2 SPRINT SCHEDULE

Project Tracker:

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date	Story Points Completed	Sprint Release Date
Sprint-1	20	6 Days	23 Oct 2022	28 Oct 2022	20	29 Oct 2022
Sprint-2	20	7 Days	29 Oct 2022	04 Nov 2022	20	05 Nov 2022
Sprint-3	20	7 Days	05 Nov 2022	11 Nov 2022	20	12 Nov 2022
Sprint-4	20	8 Days	12 Nov 2022	19 Nov 2022	20	19 Nov 2022

Velocity:

Sprint 1: 1 user stories x 20 story points = 20

Sprint 2: 1 user stories x 20 story points = 20

Sprint 3: 1 user stories x 20 story points = 20 Sprint

4: 1 user stories x 20 story points = 20

Total = 80 The average sprint velocity is

80 ÷ 4 = **20.**

6.3 REPORTS FROM JIRA

7. CODING AND SOLUTIONS

7.1 FEATURE 1

Data collection and creation

Data mining techniques require domain knowledge in order to generate predictions. For water quality applications, it is vital to understand how various water quality parameters influence water quality. This information can come from a domain expert or historical data collections. For the forecasting task, two types of data sets were used: a carefully created huge synthetic data set and an available real data set

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trih
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	

Data Preprocessing

The processing phase is very important in data analysis to improve the data quality. In this phase, the WQI has been calculated from the most significant parameters of the dataset. Then, water samples have been classified on the basis of the WQI values. For obtaining superior accuracy, the -score method has been used as a data normalization technique.

Feature scaling

```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train_final = sc.fit_transform(X_train)
X_test_final = sc.transform(X_test)
```

```
from sklearn.metrics import confusion_matrix, classification_report, accuracy_score
```

Water Quality Index Calculation

To measure water quality, WQI is used to be calculated using various parameters that significantly affect WQ [40–42]. In this study, a published dataset is considered to test the proposed model, and seven significant water quality parameters are included. The WQI has been calculated using the following formula:

$$\mathrm{WQI} = \frac{\sum_{i=1}^{N} q_i \times w_i}{\sum_{i=1}^{N} w_i},$$

where: is the total number of parameters included in the WQI calculations is the quality rating scale for each parameter calculated by equation (2) below, and is the unit weight for each parameter calculated by equation (3).

$$q_i = 100 \times \left(\frac{V_i - V_{\mathrm{Ideal}}}{S_i - V_{\mathrm{Ideal}}} \right),$$

where: is the measured value of parameter in the tested water samples is the ideal value of parameter in pure water (0 for all parameters except and), and is the recommended standard value of parameter (as shown in Table 1)

$$w_i = \frac{K}{S_i},$$

7.2 FEATURE 2

Performance Measures Results True Positives (TP) are when the model predicts the positive class properly. True Negatives (TN) is one of the components of a confusion matrix designed to demonstrate how classification algorithms work. Positive outcomes that the model predicted incorrectly are known as False Positives (FP). False Negatives (FN) are negative outcomes that the model predicts negative class. Accuracy is the most basic and intuitive performance metric, consisting of the ratio of successfully predicted observations to total observations.

Accuracy = TP+TN/(TP+FP+FN+TN)

```python
# Random Forest Classifier
from sklearn.ensemble import RandomForestClassifier
rf_classifier = RandomForestClassifier(n_estimators = 20, criterion = 'entropy', class_weight = "balance
rf_classifier.fit(X_train_final, y_train)
y_pred = rf_classifier.predict(X_test_final)
accuracy_score(y_test, y_pred)
```

```
0.635
```

```python
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.66      0.86      0.75       497
           1       0.54      0.26      0.35       303

    accuracy                           0.64       800
   macro avg       0.60      0.56      0.55       800
weighted avg       0.61      0.64      0.60       800
```

```python
# XGBoost Classifier
from xgboost import XGBClassifier
xgb_classifier = XGBClassifier(random_state=0)
xgb_classifier.fit(X_train_final, y_train)
y_pred_xgb = xgb_classifier.predict(X_test_final)
accuracy_score(y_test, y_pred_xgb)
```

```
0.62125
```

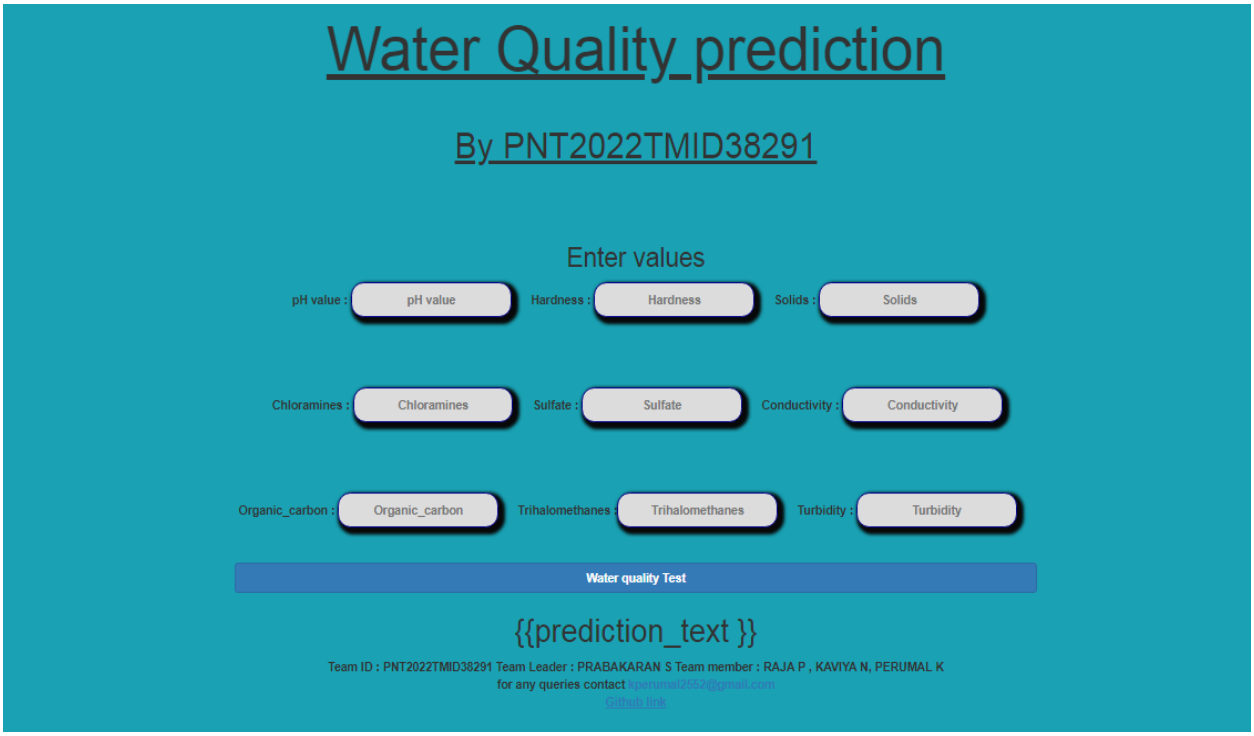
```python
print(classification_report(y_test, y_pred_xgb))
```

```
              precision    recall  f1-score   support

           0       0.67      0.77      0.72       497
           1       0.50      0.38      0.43       303

    accuracy                           0.62       800
   macro avg       0.59      0.57      0.57       800
weighted avg       0.61      0.62      0.61       800
```

The SVM model was developed in 1995 by Corinna Cortes and Vatnik. It has several unique benefits in solving small samples, and nonlinear and high-dimensional pattern recognition. It can be extended to function in the simulation of other machine learning problems. It uses the hyperplane to separate the points of the input vectors and finds the needed coefficients. The best hyperplane is the line with the largest margin, which is meant the distance between the hyperplane and the nearest input objects. The input points defined in the hyperplane are called *support vectors*. In this work, the linear SVM model along with the Gaussian radial basis function (equation (17)) is used to classify the tested water samples based on their quality.

8.TESTING

8.1 TEST CASES 1



8.2 USER ACCEPTANCE TESTING

1. Purpose of Document

The purpose of this document is to briefly explain the test coverage and open issues of the [ProductName] project at the time of the release to User Acceptance Testing (UAT).

2. Defect Analysis

This report shows the number of resolved or closed bugs at each severity level, and how they were resolved

Resolution	Severity 1	Severity 2	Severity 3	Severity 4	Subtotal
By Design	10	4	2	3	20
Duplicate	1	0	3	0	4
External	2	3	0	1	6
Fixed	11	2	4	20	37
Not Reproduced	0	0	1	0	1
Skipped	0	0	1	1	2
Won't Fix	0	5	2	1	8
Totals	24	14	13	26	77

3. Test Case Analysis

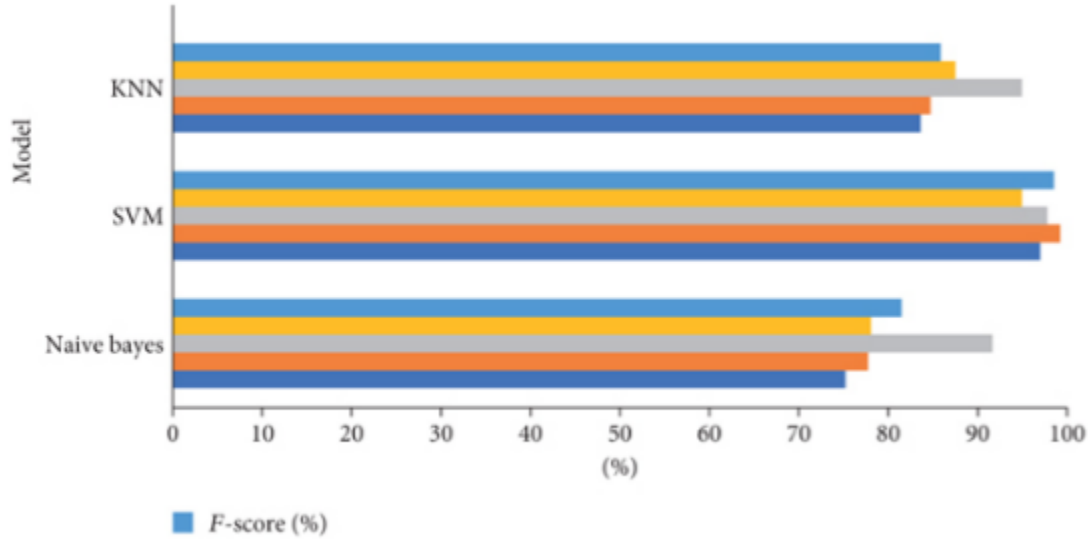
This report shows the number of test cases that have passed, failed, and untested

Section	Total Cases	Not Tested	Fail	Pass
Print Engine	7	0	0	7
Client Application	51	0	0	51
Security	2	0	0	2
Outsource Shipping	3	0	0	3
Exception Reporting	9	0	0	9
Final Report Output	4	0	0	4
Version Control	2	0	0	2

9.RESULT

9.1 PERFORMANCE METRICS

For validating the developed model, the dataset has been divided into 70% training and 30% testing subsets. While the ANN and LSTM models were used to predict the WQI, the SVM, KNN, and Naive Bayes were utilized for the water quality classification prediction



SO, WE ARE GOING TO USE SVC

Performance Measures Results True Positives (TP) are when the model predicts the positive class properly. True Negatives (TN) is one of the components of a confusion matrix designed to demonstrate how classification algorithms work. Positive outcomes that the model predicted incorrectly are known as False Positives (FP). False Negatives (FN) are negative outcomes that the model predicts negative class. Accuracy is the

most basic and intuitive performance metric, consisting of the ratio of successfully predicted observations to total observations.

Accuracy = TP+TN/(TP+FP+FN+TN)

Table 1. Comparison of algorithms

SN.

SN.	Algorithm	Type	ACCURACY	Precision	Recall f1-Score
1	RANDOM FOREST	58.5	0.42	0.38	0.40
2	XGBOOST	61.7	0.43	0.12	0.18

10. ADVANTAGES

Whether it be for groundwater, surface water or open water, there are a number of reasons why it is important for you to undertake regular water quality testing. If you're wanting to create a solid foundation on which to build a broader water management plan, then investing in water quality testing should be your first point of action. This testing will also allow you to adhere to strict permit regulations and be in compliance with Australian laws.

Identifying the health of your water will help you to discover where it may need some help. Ultimately, finding a source of pollution, or remaining proactive with your monitoring will enable you to save money in the long term. The more information that you can obtain will assist you with your decision on what product you may need to improve the condition of your water. Simply guessing and buying products based on a hunch or a

general trend is ill-advised, as each body of water has unique properties that can only be discovered through testing.

Measuring the amount of dissolved oxygen in your water is another important advantage of water quality testing, as typically the less oxygen, the higher the water temperature, resulting in a more harmful environment for aquatic life. These levels do fluctuate slightly across the seasons, but regular monitoring of your water quality will allow you to discover trends over time, and whether there are other factors that may be contributing to the results you discover.

DISADVANTAGES

Training necessary Somewhat difficult to manage over time and with large data sets

Requires manual operation to submit data, some configuration required

Costly, usually only feasible under Exchange Network grants technical expertise and network server required

Requires manual operation to submit data Cannot respond to data queries from other nodes, and therefore cannot interact with the Exchange Network

Technical expertise and network server required

11. CONCLUSION

Potability determines the quality of water, which is one of the most important resources for existence. Traditionally, testing water quality required an expensive and time-consuming lab analysis. This study looked into an

alternative machine learning method for predicting water quality using only a few simple water quality criteria. To estimate, a set of representative supervised machine learning algorithms was used. It would detect water of bad quality before it was released for consumption and notify the appropriate authorities It will hopefully reduce the number of individuals who drink low-quality water, lowering the risk of diseases like typhoid and diarrhea. In this case, using a prescriptive analysis based on projected values would result in future capabilities to assist decision and policy makers.

12.SOURCE CODE

Machine learning has been widely used as a powerful tool to solve problems in the water environment because it can be applied to predict water quality, optimize water resource allocation, manage water resource shortages, etc. Despite this, several challenges remain in fully applying machine learning approaches in this field to evaluate water quality: (1) Machine learning is usually dependent on large amounts of high-quality data. Obtaining sufficient data with high accuracy in water treatment and management systems is often difficult owing to the cost or technology limitations. (2) As the conditions in real water treatment and management systems can be extremely complex, the current algorithms may only be applied to specific systems, which hinders the wide application of machine learning approaches. (3) The implementation of machine learning algorithms in practical applications requires researchers to have certain professional background knowledge.

To overcome the above-mentioned challenges, the following aspects should be considered in future research and engineering practices: (1) More advanced sensors, including soft sensors, should be developed and applied in water quality monitoring to collect sufficiently accurate data to facilitate the application of machine learning approaches. (2) The feasibility and reliability of the algorithms should be improved, and more universal algorithms and models should be developed according to the water treatment and management requirements. (3) Interdisciplinary talent with knowledge in different fields should be trained to develop more advanced machine learning techniques and apply them in engineering practices.

13. APPENDIX

REQUIREMENT.TXT

Flask == 2.2.2

joblib == 1.2.0

numpy == 1.23.4

pandas == 1.5.1

scikit-learn == 1.1.3

xgboost == 1.7.1

gunicorn == 20.1.0

matplotlib == 3.6.2

seaborn == 0.12.1

gevent

requests

flask-cors==3.0.10

APP.PY

app.py > ...

```python
from flask import Flask, request, render_template
import pickle
import pandas as pd
import numpy as np
import joblib
import os
from gevent.pywsgi import WSGIServer
scaler = joblib.load("my_scaler.save")


app = Flask(__name__)
model = pickle.load(open('model.pkl', 'rb'))

@app.route("/home")
@app.route("/")
def hello():
    return render_template("home.html")

@app.route("/predict", methods = ["GET", "POST"])
def predict():
    if request.method == "POST":
        input_features = [float(x) for x in request.form.values()]
        features_value = [np.array(input_features)]

        feature_names = ["ph", "Hardness" , "Solids", "Chloramines", "Sulfate",
                         "Conductivity", "Organic_carbon","Trihalomethanes", "Turbidity

        df = pd.DataFrame(features_value, columns = feature_names)
        df = scaler.transform(df)
        output = model.predict(df)

        if output[0] == 1:
            prediction = "safe"
```

WATER QUALITY.IPYNB

HOME.HTML

```
<!doctype html>
<html>
<head>

<title>  Water Quality </title>
<!-- BootStrap -->
            <link
href="https://maxcdn.bootstrapcdn.com/bootstrap/3.3.5/css/bootstrap.min.css"
            rel="stylesheet"
            integrity="sha256-MfvZlkHCEqatNoGiOXveE8FIwMzZg4W85qfrfIFBfYc=
sha512-
dTfge/zgoMYpP7QbHy4gWMEGsbsdZeCXz7irItjcC3sPUFtf0kuFbDz/ixG7ArTxmDjLXDmezHubeNiky
KGVyQ=="
            crossorigin="anonymous">
<style>
input[type=text]
  {
   border:1px solid darkblue;
   border-radius:15px;
   box-shadow:4px 4px 4px 4px rgb(8, 8, 8);
  }
input{
    text-align: center;
    width: 20%;
    height: 70px;
    font-size: 14px;
    padding-top:0px ;
}
.thick {
  text-decoration-line: underline;
  text-decoration-style: solid;
  text-decoration-color: rgb(32, 159, 163)
  text-decoration-thickness: 2px;
}
</style>
</head>

<body  style="background-color:rgb(26, 162, 180);">


 <div class="login">
```

```
 <form action="{{ url_for('predict')}}"method="post">


<center>
<b>
<h1 style="font-size:70px;" class = thick id="heading">Water Quality
prediction</h1><br>
<h1 style="font-size:40px;" class = thick id="heading">By
PNT2022TMID38291</h1><br><br><br>

</b>
</center>


<center>
  <h3 style="font-size:30px;" style="color:rgb(1, 3, 8);">Enter  values</h3>
<div id="bodycontent">

<label>
                    pH value  :   <input type="text" name="ph" placeholder="pH
value" style="background-color:#DCDCDC; height:40px"  required="required"
/>     
                  Hardness :   <input type="text" name="Hardness"
placeholder="Hardness"  style="background-
color:#DCDCDC;  height:40px"required="required" />     
                    Solids :  <input type="text" name="Solids"
placeholder="Solids"  style="background-color:#DCDCDC;
height:40px"required="required" /><br><br><br><br><br>     

                   Chloramines :   <input type="text" name="Chloramines"
placeholder="Chloramines"  style="background-color:#DCDCDC;
height:40px"required="required" />     
                  Sulfate :    <input type="text" name="Sulfate"
placeholder="Sulfate" style="background-color:#DCDCDC; height:40px"
required="required" />     
               Conductivity :      <input type="text" name="Conductivity"
placeholder="Conductivity"  style="background-color:#DCDCDC; height:40px"
required="required" />     <br><br><br><br><br>

                    Organic_carbon :  <input type="text" name="Organic_carbon"
placeholder="Organic_carbon" style="background-color:#DCDCDC; height:40px"
required="required" />     
                      Trihalomethanes :<input type="text" name="Trihalomethanes"
placeholder="Trihalomethanes" style="background-color:#DCDCDC;
height:40px"  required="required" />     
```

```
                    Turbidity :   <input type="text" name="Turbidity"
placeholder="Turbidity" style="background-color:#DCDCDC;
height:40px"  required="required" />     <br><br><br>
          <!-- Show button -->
                     <div class="button_cont" ><a
class="button_css"  target="_blank" rel="nofollow noopener">
                   <button type="submit" class="btn btn-primary btn-block btn-
large"><strong>Water quality Test</strong></button></a>
</center>
</label>
</div>
</form>
<center>
                                                                        <h1>{{pre
diction_text }}</h1>
  </center>
  </div>
  <footer>
    <center>
    <p><b>Team ID : PNT2022TMID38291

      Team Leader : PRABAKARAN S

      Team member : RAJA P , KAVIYA N,  PERUMAL K</i><br>
    for any queries contact
    <a href="kperumal2552@gmail.com">kperumal2552@gmail.com</a><br>
    <a href="https://github.com/IBM-EPBL/IBM-Project-3498-1658570711"><u>Github
link</u></a></p></center>
  </footer>

</body>
</html>
```

LINKS:

Github*:* ***https://github.com/IBM-EPBL/IBM-Project-3498-1658570711***