NALAIYA THIRAN

IBM

PROJECT REPORT

TEAM ID: PNT2022TMID52625

Efficient Water Quality Analysis and Prediction using Machine Learning

Presented by

AKASH RAJ N - CITC1905066

DEVATHARSHINI S - CITC1905075

PREM B - CITC1905106

VIJAY SHANMUGAN S - CITC1905122

INDEX

ABSTRACT

After air, water is arguably the most valuable natural resource. Although the majority of the earth's surface is made up of water, relatively little of it is really exploitable, making it an extremely scarce resource. Therefore, care must be taken when using this valuable and finite resource. Water must be suitable before usage since it is used for a variety of uses. Water bodies in poor condition pose a hazard to the ecology as well as being a sign of environmental deterioration. In industries, poor water quality can result in risks and significant financial loss. Water quality is therefore crucial for both environmental and economic reasons. Analysis of the water's purity is therefore necessary before utilising it for any purpose.

The quality of the water can be decided by calculating the Water Quality Index (WQI). Using the book, Field Manual for Water Quality Monitoring, the National Sanitation Foundation surveyed 142 people, representing a wide range of positions at the local, state, and national level, about 35 water quality tests for possible inclusions in an index. Among which six factors were chosen for Water Quality Index (WQI) calculation. So, we decided to develop a web application which takes in those six critical parameters (Conductivity, Nitratenan N + Nitrite Ann, pH, Dissolved Oxygen, Biochemical oxygen demand, Total coliform mean) as input which in turn is used by the model to calculate the Water Quality Index.

INTRODUCTION

1.1 Project Overview

Water is one of the most valuable natural resources that humans have gifted. One of humanity's most serious and concerning challenges is the diminishing condition of natural water resources such as lakes, streams, and estuaries. These resources are also at risk of contamination as a result of a variety of reasons, including human, industrial, and commercial activity, as well as natural processes. In addition, drinking water contamination is exacerbated by poor sanitation infrastructure and a lack of knowledge.

In truth, the repercussions of contaminated drinking water are quite harmful, posing a serious condition to human health, the environment, and infrastructure. According to a United Nations (UN) report, 1.5 million people expire each year as a result of illnesses that occurred in contaminated water. Water contamination is said to be the cause of 80% of health problems in impoverished countries. Every year, there are 5 million fatalities and 2.5 billion illnesses reported.

Water demand is increasing due to population growth, intense agriculture, urbanisation and industrial activities. Unclean water has far-reaching consequences that affect every part of existence. As a result, water resource management is critical if the water quality is to be maximised. Water contamination can be effectively addressed if data is examined and water quality can be fore casted ahead of time. Also, water resources play an important role in supplying drinking, industrial and agriculture domains. Water quality assessment and prediction are necessary to assess its suitability to serve a specific purpose and to determine appropriate treatments or precautions if it is not healthy.

Using different ANN models, several types of research have been conducted to simulate and predict water quality. The probability and usefulness of using ANN applications to forecast the quality of drinking water have been confirmed in these investigations. Measurements of different parameters like chemical oxygen demand (COD), dissolved oxygen (DO), electrical conductivity (EC), biochemical oxygen demand (BOD), temperature, pH, K, Na, Mg, and other water quality components have been proposed. Indeed, the good quality of water resources significantly reduces the cost of water treatment for drinking and industrial purposes and improves agricultural production.

This is mainly determined by the analysis of an important number of water quality parameters to quantify the dissolved substance. However, monitoring all parameters involved in a river or groundwater is often insufficient in developing countries as it is laborious and expensive. Hence, reducing the subjectivity and the effective cost for assessing water quality is a great challenge. However, in recent years, many water quality indexes (WQI) have been proposed and developed by several national and international organisations.

The adoption of the machine learning approach in water quality assessment has gained a lot of traction in recent years, possibly because it is accurate and flexible without requiring any parametric assumptions or complicated physical equations . The artificial neural network (ANN), in particular, has been used by countless researchers to predict water quality parameters in rivers. Other than having the ability to extract such complex relations between the predictor and the predicted variables, ANN models are also able to stimulate the nonlinear and time-varying features of the atmospheric variables at numerous scales.

1.2 Purpose

Water demand is increasing due to population growth, intense agriculture, urbanisation and industrial activities. Unclean water has far-reaching consequences that affect every part of existence. Although there are many machine learning models developed for water quality prediction, there is little to no research that integrates the models with a graphical user interface (GUI). The user is able to input real-time data into a GUI that has been connected to a machine learning model, thus minimising time consumption by reducing the number of procedures required to conduct the analysis. The GUI-based app in this study also enables a simpler and more trouble-free workflow in predicting water quality parameters and can be accessible to more people, especially on-site operators and trainees.

LITERATURE SURVEY

2.1 Existing Problem

Unclean water has far-reaching consequences that affect every part of existence. Although there are many machine learning models developed for water quality prediction, there is little to no research that integrates the models with a graphical user interface (GUI). The user is able to input real-time data into a GUI that has been connected to a machine learning model, thus minimising time consumption by reducing the number of procedures required to conduct the analysis. The GUI-based app in this study also enables a simpler and more trouble-free workflow in predicting water quality parameters and can be accessible to more people, especially on-site operators and trainees.

2.2 References

2.2.1 Prediction of irrigation water quality parameters using machine learning models in a semi arid environment.

AUTHORS: Ali El Bilali, Abdeslam Taleb

DESCRIPTION:

Evaluation of the water suitability for irrigation purposes using conventional approaches is generally expensive because it requires several parameters, particularly in developing countries. Therefore, developing accurate and reliable models may be valuable to overcome this issue in the management of the water used in agriculture. To achieve this purpose, 8 Machine Learning (ML) models namely: Artificial Neural Network (ANN), Multiple Linear Regression (MLR), Decision Tree, Random Forest (RF), Support Vector Regression (SVR), k-Nearest Neighbour (kNN), Stochastic Gradient Descent (SGD) and Adaptive Boosting (AdaBoost) have been developed and validated for predicting of 10 Irrigation Water Quality (IWQ) parameters such as Sodium absorption ratio (SAR), adjusted SARa, Exchangeable Sodium Percentage (ESP), percentage of Sodium (%Na), Residual Sodium Carbonate (RSC), Permeability Index (PI), Kelly Ratio (KR), Chloride Cl^-, Magnesium Absorption Ratio (MAR), and TDS dissolved in water surface of Bouregreg watershed in Morocco using electrical conductivity (EC) and pH as input variables. 300 samples are analysed at 9 monitoring stations across four main rivers, processed and selected to train and validate the models. The results have revealed that, except for SVR and k-NN models and MAR and PI parameters, all other models are highly accurate in predicting the other parameters with coefficients of correlations (r) with ranges of [0.56, 0.99], and [0.64, 0.99] for training and validation processes sequentially.

ADVANTAGE:

Adaboost prediction accuracy is high for seven parameters compared to other models.

LIMITATION:

SVM and KNN models showed low performance for predicting contents in water. Different models showed different accuracy.

2.2.2 A water quality prediction method based on the multi-time scale bidirectional LSTM network

AUTHORS: Qinghong Zou, Qingyu Xiong, Hualing Yi

DESCRIPTION:

As an important factor affecting the mangrove wetland ecosystem, water quality has become the focus of attention in recent years. Therefore, many studies have focused on the prediction of water quality to help establish a regulatory framework for the assessment and management of water pollution and ecosystem health. To make a more accurate and comprehensive forecast analysis of water quality, we propose a method for water quality prediction based on the multi-timescale bidirectional LSTM network. In the method, we improve data integrity and data volume through data preprocessing.And the network processes input data forward and backward and considers the dependencies at multiple time scales. Besides,we use the Box–Behnken experimental design method to adjust hyper-parameters in the process of modelling. In this study,we apply this method to the water quality prediction research of Beilun Estuary, and the performance of our proposed model is evaluated and compared with other models. The experiment results show that this model has better performance in water quality prediction than that of using LSTM or bidirectional LSTM alone.

ADVANTAGE:

MT-BLSTM had the highest accuracy compared to LSTM or bidirectional LSTM.

LIMITATION:

The reusability of the model needs to be improved.

2.2.3 Implementation of ML models for monitoring and predicting water quality parameters

AUTHORS: Gasim Hayder, Osman Kurniawan

DESCRIPTION:

Since clean water is well known as one of the crucial sources that all living things need in their daily lives, the demand for clean freshwater nowadays has increased. However, water quality is slowly deteriorating due to anthropogenic and natural sources of pollution and contamination. Therefore, this study aims to develop artificial neural network (ANN) models to predict six different water quality parameters in the Langat River, Malaysia. Moreover, an application (app) equipped with a graphical user interface (GUI) was designed and developed to conduct real-time prediction of the water quality parameters by using real-time data as inputs together with the ANN models. As for the results, all of the ANN models achieved high coefficients of determination (R2), which were between 0.9906 and 0.9998, as well as between 0.8797 and 0.9972 for training and testing datasets, respectively. The developed app successfully predicted the outcome based on the run models. The implementation of a GUI-based app in this study enables a simpler and more trouble-free workflow in predicting water quality parameters. By eliminating sophisticated programming subroutines, the prediction process becomes accessible to more people, especially on-site operators and trainees.

ADVANTAGE:

As only one method is used, the results were accurate.

LIMITATION:

This method takes only 6 water quality parameters. Other parameters are not considered.

2.2.4 Prediction of groundwater quality using efficient machine learning technique

AUTHORS: Sudhakar Singha, Srinivas Pasupuleti, Soumya S.Singha, Rambabu Singh, Suresh Kumar

DESCRIPTION:

To ensure safe drinking water sources in the future, it is imperative to understand the quality and pollution level of existing groundwater. The prediction of water quality with high accuracy is the key to control water pollution and the improvement of water management. In this study, a deep learning (DL) based model is proposed for predicting groundwater quality and compared with three other machine learning (ML) models, namely, random forest (RF), eXtreme gradient boosting (XGBoost), and artificial neural network (ANN). A total of 226 groundwater samples are collected from an agriculturally intensive area Arang of Raipur district, Chhattisgarh, India, and various physicochemical parameters are measured to compute entropy weight-based groundwater quality index (EWQI). Prediction performances of models are determined by introducing five error metrics. Results showed that DL model is the best prediction model with the highest accuracy in terms of R2, i.e., R2 = 0996 against the RF (R2 = 0.886), XGBoost (R2 = 0.0.927), and ANN (R2 = 0.917). The uncertainty of the DL model output is cross-verified by running the proposed algorithm with a newly randomised dataset for ten times, where minor deviations in the mean value of performance metrics are observed. Moreover, input variable importance computed by prediction models highlights that DL model is the most realistic and accurate approach in the prediction of groundwater quality.

ADVANTAGE:

Relatively higher prediction performance was observed in the XGBoost model than RF and ANN models.

LIMITATION:

A single monsoon dataset is considered in the prediction models as the limitation.

2.2.5 Efficient Prediction of Water Quality Index(WQI) Using Machine Learning Algorithms

AUTHORS: Md.Mahedi Hassan, Laboni Akter, Mushfiqur Rahman, Nusrat Jahan

DESCRIPTION:

The quality of water has a direct influence on both human health and the environment. Water is utilised for a variety of purposes, including drinking, agriculture, and industrial use. The water quality index (WQI) is a critical indication for proper water management. The purpose of this work was to use machine learning techniques such as RF, NN, MLR, SVM, and BTM to categorise a dataset of water quality in various places across India. Water quality is dictated by features such as dissolved oxygen (DO), total coliform (TC), biological oxygen demand (BOD), Nitrate, pH, and electric conductivity (EC). These features are handled in five steps: data pre-processing using min-max normalisation and missing data management using RF, feature correlation, applied machine learning classification, and model’s feature importance. The highest accuracy Kappa, Accuracy Lower, and Accuracy Upper findings in this research are 99.83, 99.17, 99.07, and 99.99, respectively. The finding showed that Nitrate, PH, conductivity, DO, TC, and BOD are the key qualities that contribute to the orderly classification of water quality, with Variable Importance values of 74.78, 36.805, 81.494, 105.770, 105.166, and 130.173, respectively.

ADVANTAGE:

Multinomial Logistic Regression has the highest accuracy at 99.83% Random forest has accuracy at 99.63%.

LIMITATION:

Only 7 parameters were considered to predict water quality.

2.2.6 Machine learning based marine water quality prediction for coastal hydro-environment management

AUTHORS: Tianan Deng, Kwok-Wing Chau, Huan-Feng Duan

DESCRIPTION:

During the past three decades, harmful algal blooms (HAB) events have been frequently observed in marine waters around many coastal cities in the world including Hong Kong. The increasing occurrence of HAB has caused acute influences and damages on water environment and marine aquaculture with millions of monetary losses. For example, the Tolo Harbour is one of the most affected areas in Hong Kong, where more than 30% HAB occurred. In order to forewarn the potential HAB incidents, the machine learning (ML) methods have been increasingly resorted to modelling and forecasting water quality issues. In this study, two different ML methods - artificial neural networks (ANN) and support vector machine (SVM) - are implemented and improved by introducing different hybrid learning algorithms for the simulations and comparative analysis of more than 30-year measured data, so as to accurately forecast algal growth and eutrophication in Tolo Harbour in Hong Kong. The application results show the good applicability and accuracy of these two ML methods for the predictions of both trend and magnitude of the algal growth. Specifically, the results reveal that ANN is preferable to achieve satisfactory results with quick response, while the SVM is suitable to accurately identify the optimal model but taking longer training time. Moreover, it is demonstrated that the used ML methods could ensure robustness to learn complicated relationships between algal dynamics and different coastal environmental variables and thereby to identify significant variables accurately. The results analysis and discussion of this study also indicate the potentials and advantages of the applied ML models to provide useful information and implications for understanding the mechanism and process of HAB outbreak and evolution that is helpful to improving the water quality prediction for coastal hydro-environment management.

ADVANTAGE:

ANN showed good predicting performances and there is no overfitting problem. SVM performance is better than all ANN models in terms of water quality prediction results.

LIMITATION:

It takes much longer to train the SVM than the ANN as the SVM takes a quadratic programming with time complexity of O(n3).

2.2.7 A novel approach for water quality classification based on the integration of deep learning and feature extraction techniques

AUTHORS: Smail Dilmi, Mohamed Ladjal

DESCRIPTION:

Water quality monitoring plays a vital role in the protection of water resources, environmental management, and decision-making. Artificial intelligence (AI) based on machine learning techniques has been widely used to evaluate and classify water quality for the last two decades. However, traditional machine learning techniques face many limitations, the most important of which is the inability to apply these techniques with big data generated by smart water quality monitoring stations to improve the prediction. Real-time water quality monitoring with high accuracy and efficiency for intelligent water quality monitoring stations requires new and sophisticated techniques based on machine and deep learning techniques. For this purpose, we propose a novel approach based on the integration of deep learning and feature extraction techniques to improve water quality classification. In this paper, the Tallest dam in Bouira (Algeria) as a case study. Moreover, we implemented the advanced deep learning method - Long Short Term Memory Recurrent Neural Networks (LSTM RNNs) to construct an intelligent model for drinking water quality classification. Furthermore, principal component analysis (PCA), linear discriminant analysis (LDA) and independent component analysis (ICA) techniques were used for feature extraction and data reduction from original features. Additionally, we used three methods of cross-validation and two methods of the out-of-sample test to estimate the performance of the LSTM RNNs model. From the results we found that the integration of LSTM RNNs with LDA, and LSTM RNNs with ICA yields an accuracy of 99.72%, using Random-Holdout technique.

ADVANTAGE:

Integration of LSTM with LDA/ICA gives 99.72% accuracy.

LIMITATION:

This method doesn’t measure chemical parameters that can be measured continuously.

2.2.8 Water quality prediction using machine learning and flask

AUTHORS: S.Sharath, R.Harish, V.Aishwarya and Dr.M.Preetha

DESCRIPTION:

Generally, Water pollution refers to the release of pollutants into the water that is detrimental to human health and the planet as whole. The aim is to investigate machine learning-based techniques for water quality forecasting by predicting results with the best accuracy. The analysis of the data set by supervised machine Learning techniques(SMLT) to capture information like variable identification, uni-variate analysis, bi-variate and multivariate analysis, missing value treatments and analysis data validation, data cleaning/preparation, and data visualisation will be done on the entire given data set. Our analysis provides a comprehensive guide to sensitivity analysis of model parameters with regard to performance in the prediction of water quality pollution by accuracy calculation. To propose a machine learning-based method to accurately predict the Water Quality Index value by prediction results in the form of best accuracy from comparing supervised classification machine learning algorithms. . Additionally, to compare and discuss the performance of various machine learning algorithms from the given transport traffic department data set with evaluation classification report, identify the confusion matrix and categorising data from priority and the result shows that the effectiveness of the proposed machine learning algorithm technique can be compared with the best accuracy with precision,Recall and F1 Score

ADVANTAGE:

Produces the appropriate results with the mentioned techniques.

LIMITATION:

Accuracy of the technique is not mentioned.

2.2.9 A novel machine learning application: Water quality resilience prediction Model

AUTHORS: Maryam Imani, Md Mahmudul Hasan, Luiz Fernando Bittencourt, Kent McClymont and Zoran Kapelan

DESCRIPTION:

Resilience-informed water quality management embraces the growing environmental challenges and provides greater accuracy by unpacking the systems' characteristics in response to failure conditions in order to identify more effective opportunities for intervention. Assessing the resilience of water quality requires complex analysis of influential parameters which can be challenging, time consuming and costly to compute. It may also require building detailed conceptual and/or physically process-based models that are difficult to build, calibrate and validate. This study utilises Artificial Neural Network (ANN) to develop a novel application to predict water quality resilience to simplify resilience evaluation. The Fuzzy Analytic Hierarchy Process method is used to rank water basins based on their level of resilience and to identify the ones that demand prompt restoration strategies. The commonly used 'magnitude * duration of being in failure state' quantification method has been used to formulate and evaluate resilience. A 17-years long water quality dataset from the 22 water basins in the State of São Paulo, Brazil, was used to train and test the ANN model. The overall agreement between the measured and simulated WQI resilience values is satisfactory and hence, can be used by planners and decision makers for improved water management. Moreover, comparative analyses show similarities and differences between the 'level of criticalities' reported in each zone by the Environment Agency of the state of São Paulo (CETESB) and by the resilience model in this study.

ADVANTAGE:

Resilience is a fast growing concept proved to be an effective approach in preparing engineering systems to tackle and cope with emerging challenges.

LIMITATION:

Model performance in validation dataset decreases abruptly in successive iterations.

2.2.10 Water Quality Prediction Method Based on LSTM-BP

AUTHORS: Huimin Jia, Xiaofeng Zhou

DESCRIPTION:

Water quality prediction is of practical significance not only for the planning, evaluation, and management of the water environment, but also for the prevention and control of water pollution. In order to improve the accuracy of water quality prediction, an LSTM-BP combined model algorithm based on Long Short Term Memory Neural Network (LSTM NN) and BP neural network is proposed. Taking the water temperature data of No.6 large-scale integrated observation buoy on the Yangtze estuary as an example, a time series prediction model framework is established, and the data processing to model simulation is completed with the help of Python to realise the water quality prediction based on LAST MBP. The method is compared with the LSTM model and BP model, the experimental results show that the time series predicted by LSTM-BP is more accurate. This LSTM-BP model can be effectively applied to the prediction of water quality indicators and the early warning and prediction system of water quality trends.

ADVANTAGE:

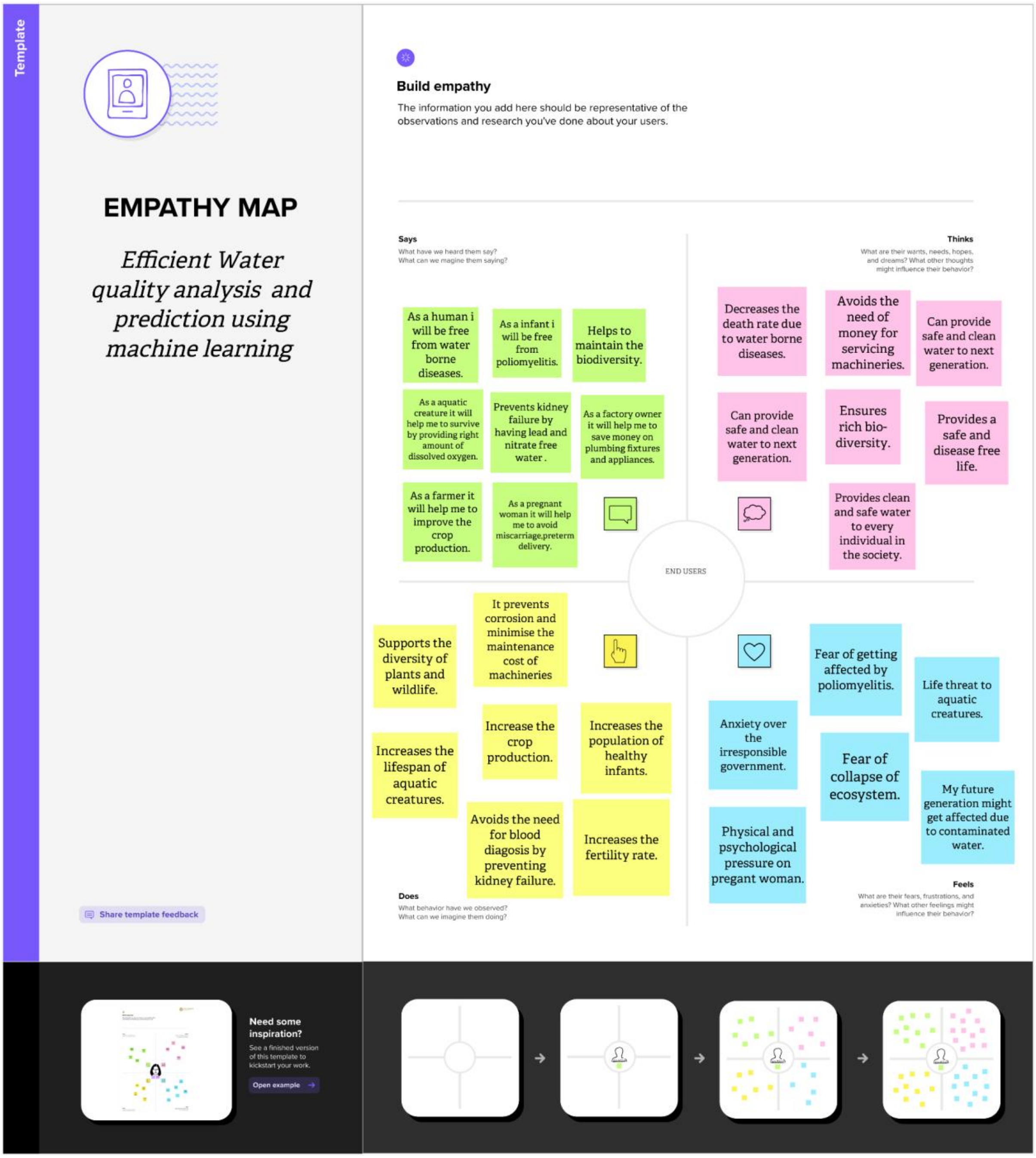
The integrated LSTM-BP has better prediction performance of time series and data generalisation ability than the single LSTM model and BP neural network model.

LIMITATION:

The Back propagation method's convergence speed is slow.

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas



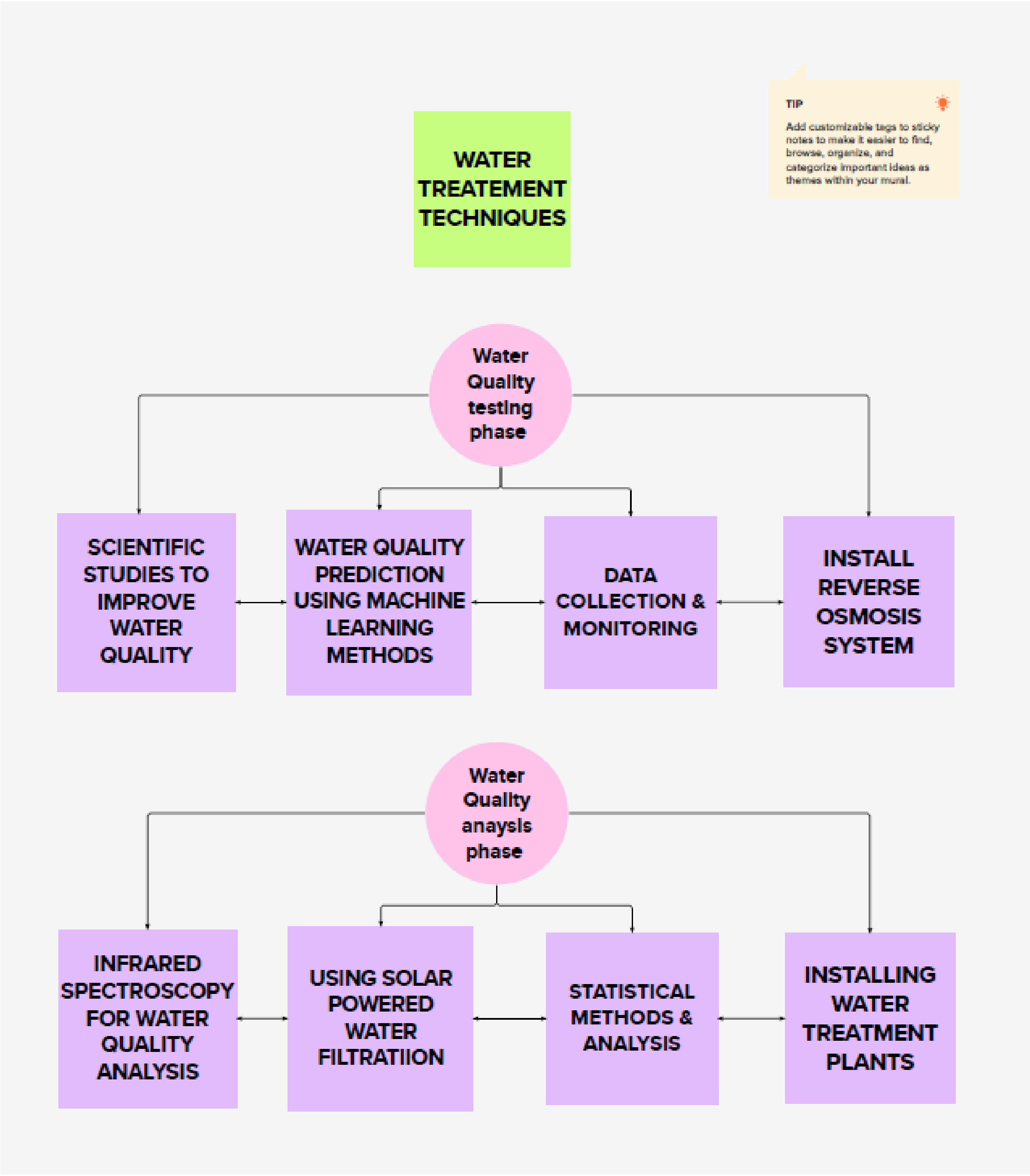
3.2 Ideation & Brainstorming

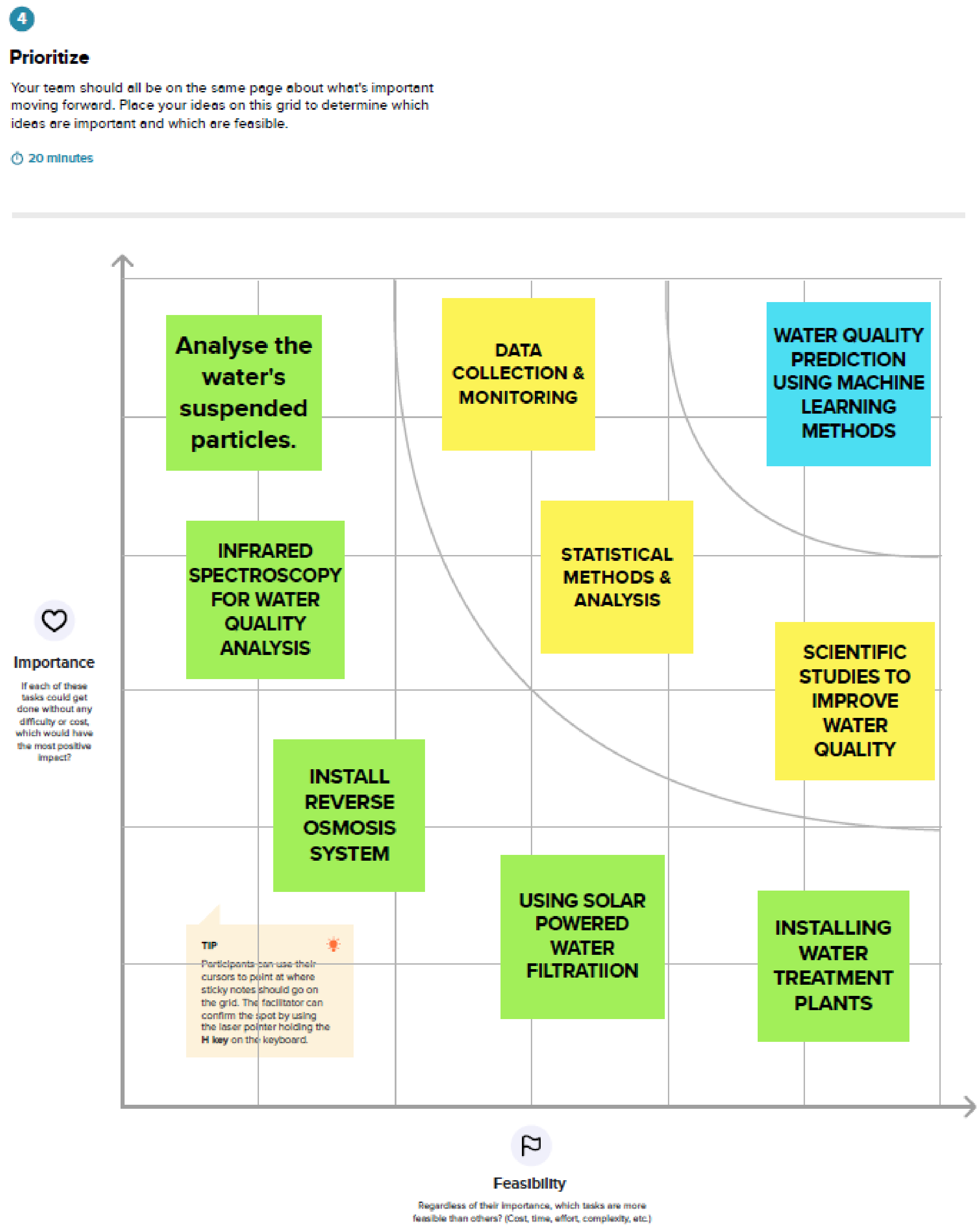
3

Group ideas

Take turns sharing your ideas while clustering similar or related notes as you go. Once all sticky notes have been grouped, give each cluster a sentence-like label. If a cluster is bigger than six sticky notes, try and see if you and break it up into smaller sub-groups.

20 minutes





3.3 Proposed Solution

3.3.1 Problem Statement

Water is seen as a significant resource that affects many aspects of human health and survival. Due to human activity as well as environmental factors, water contamination has increased recently. Numerous factors can taint water. As a result, drinking water is unsafe, and people should check the water's purity before using it for any reason.

3.3.2 Solution description

The goal of this project is to create a machine learning model that can forecast a water quality by taking into account all available standard indicators. The best prediction will be made using a large dataset and a good correlation between the factors.

3.3.3. Novelty

The portability of the suggested approach is tested. It comprises two stages: testing and training, which use historical data from the past. In comparison to other methods, this uses just the necessary data and produces predictions that are more accurate.

3.3.4 Social Impact

The effectiveness of water services as a key environmental factor and a base for the prevention and management of water-borne illnesses. Helps individuals better classify the available water for different uses based on analysis that may be used to practise water conservation.

3.3.5 Business Model

This concept has to be licensed by machine learning and data analytics in order to get greater traction with the public.

3.3.6 Scalability of the Solution

When put to the test by demands that are higher than its operating requirements, a system that scales effectively will be able to maintain or improve its level of performance. The model is safe and secure to use. It offers higher water quality accuracy. The temperature and pH level are easily discernible.

3.4 Problem Solution fit

Define CS, fit into CC

1. CUSTOMER SEGMENT(S) CS

Who is your customer?

1. Individuals
2. Farmers
3. Industrialist

6. CUSTOMER CONSTRAINTS CC

What constraints prevent your customers from taking action or limit their choices.

1. Unaware of water suspended particles.
2. Lack of knowledge on scientific methods to measure water quality.
3. Lack of efficient water Quality measure technique.

5. AVAILABLE SOLUTIONS AS

Which solutions are available to the customers when they face the problem or need to get the job done? What have they tried in the past? What pros & cons do these solutions have?

Reverse osmosis filters.
Copper filters.

PROS:
1. Removes harmful water suspended particles.
2. Removes salinity.
CONS:
1. High maintenance cost.

Explore AS, differentiate

Focus on J&P, tap into BE, understand RC

2. JOBS-TO-BE-DONE/COMPLETED J&P

Which jobs-to-be-done (or problems) do you address for your customers? There could be more than one; explore different sides.

1. To analyze and predict the water Quality using scientific metrics.

2. Awareness on scientific method to measure water Quality.

9. PROBLEM ROOT CAUSE RC

What is the real reason that this problem exists? What is the back story behind the need to do this job?

1. Water pollution.

2. Dumping of hazardous Industrial chemicals(waste) into river waters.

7. BEHAVIOUR BE

What does your customer do to address the problem and get the job done?

1. Installing RO systems.
2. Installing Copper filters.
3. Bleaching the water stored in tanks.

Focus on J&P, tap into BE, understand RC

3. TRIGGERS TR

What triggers customers to act?

Fear of water borne disease and frustration due to health issues.

4. EMOTIONS: BEFORE / AFTER EM

How do customers feel when they face a problem or a job and afterwards?

Before:
1. Frustration and Stress.
2 .Fear of water borne diseases.

After:
1. Happy and Secure.
2. Healthy life.

10. YOUR SOLUTION SL

Collecting water samples from different water sources and determining the essential components that improve water quality. Then, by using appropriate machine learning techniques on the data gathered, the water quality can be measured.

8. CHANNELS OF BEHAVIOUR CH

8.1ONLINE

What kind of actions do customers take online?

1. Browsing the web to learn about the techniques to measure water Quality.
2. Learning about hazardous water suspended particles.

8.2 OFFLINE

What kind of actions do customers take offline?

1. Installing Ro systems.
2. Drinking boiling water.

4. REQUIREMENT ANALYSIS

4.1 Functional Requirements

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	Enter the input	Get the input values via form and check the data
FR-4	Executive administration	Early warning/forecast monitoring is one of two separate roles that are included in the regulation of monitoring the water environment state and regulatory compliance, such as pollution event emergency management
FR-5	User Requirements	The user needs an accurate and exact result
FR-6	Data Preprocessing	From the raw dataset, obtain the tested and trained data
FR-7	Data Handling	Metrics for the various water bodies' water quality included in the file
FR-8	Quality analysis	Use multiple models to analyse the data on the water's obtained PH, TDS, and temperature levels, among other water quality indicators
FR-9	Model prediction	Based on the water quality index, the confirmation displays the machine learning

		prediction (Good, Partially Good, Poor) and the proportion of each parameter that is present
FR-10	Remote Visualisation	Visualisation of future forecasts using charts based on present and past values of all the parameters

4.2 Non functional requirements

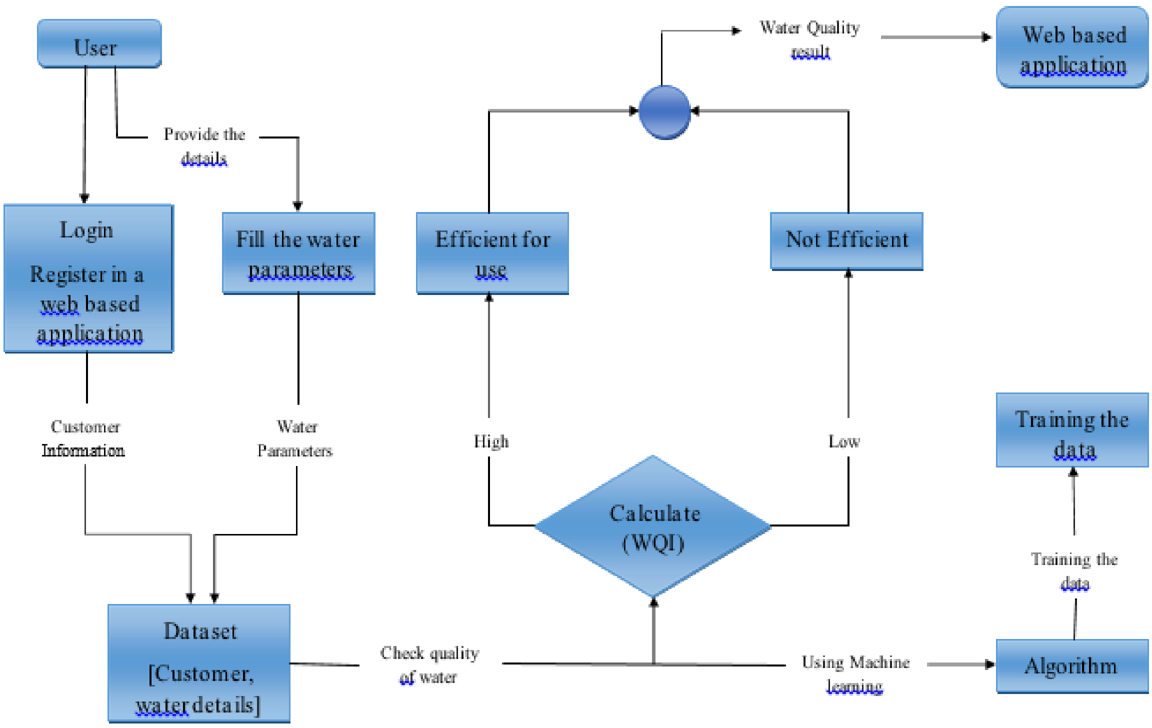
- The website should be fast enough to predict the output within seconds.
- The website will be safely encrypted.

FR No.	Non-Functional Requirement	Description
NFR-1	Usability	A user-friendly web application, the system provides natural interaction with the users
NFR-2	Security	The website is virus-free and did not request any authorization. The model has a strong security system since the user's information won't be shared with any other sources
NFR-3	Reliability	A wide variety of water values are trained in the model, increasing forecast accuracy. The model may be greatly expanded by adding more datasets
NFR-4	Performance	Get the results quickly
NFR-5	Availability	Available on the internet at any moment. As long as the user has access to the system, it should be accessible until the user terminates it. The system responds to user requests more quickly, and recovery is completed faster

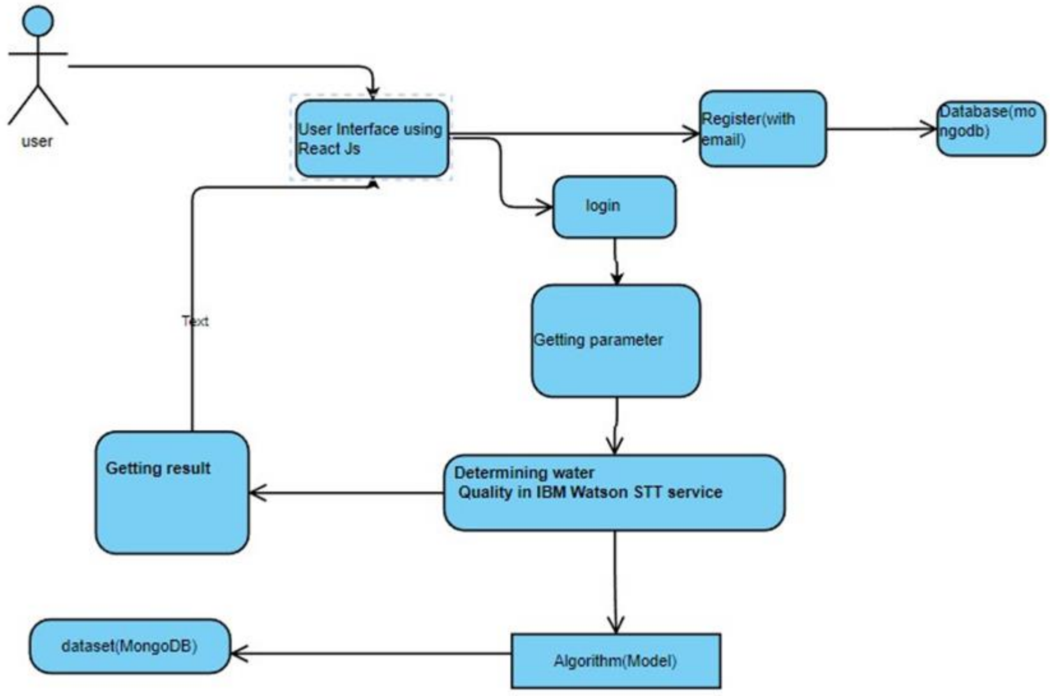
NFR-6	Scalability	It is a lightweight application, the users can access the website through mobile phones, tabs, desktop and laptop. It produces an effective result and has the capacity to alter the system's performance depending on the datasets

5. PROJECT DESIGN

5.1 Data Flow Diagrams



5.2 Solution & Technical Architecture



5.3 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 a 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 the 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 the correct coefficient		Medium	Sprint-3
	Data preprocessin g	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 results are visible	High	Sprint-4

			the water efficiency.	on the webpage.		
Customer Care Executive	Solving Customer issues.	USN-13	As a customer care executive, I solve the customer issues in using the application and web page.	It results in user interaction.	Medium	Sprint-5
Administrator		USN-14	I can manage the application.		Medium	Sprint-5

6. PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

Sprint	Milestone
Sprint 1	1. User Registers into the application by entering Email Id Password and re-enter Password for confirmation. 2. User Receives a confirmation mail for their registered Email. 3. User logs in into the website using Email Id and password.
Sprint 2	1. User can access the dashboard 2. User can enter the input parameters on water and get to know the water quality based on model's prediction
Sprint 3	1. Dataset is provided through which the model predicts the water quality based on input parameters. 2. The application must be secure.
Sprint 4	1. Administrator should properly maintain the website and update it when required.

6.2 Sprint Delivery Schedule

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password	5	High	Prem B Akash Raj N
Sprint-1		USN-2	As a user, I will receive confirmation email once I have registered for the application	5	High	Prem B Akash Raj N
Sprint-1		USN-3	As a user, I can register for the application through Phone number	2	Low	Prem B Akash Raj N
Sprint-1		USN-4	As a user, I can register for the application through Gmail	3	Medium	Prem B Akash Raj N
Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Login (User)	USN-5	As a user, I can log into the application by entering email & password	5	High	Prem B Akash Raj N

Sprint-2	Dashboard	USN-6	Once I have logged in, I can see my dashboard	6	Medium	Prem B Akash Raj N
Sprint-2	Web Access	USN-7	As a customer I can access the website to predict the water quality	7	High	Devatharshini S Vijay Shanmugan S
Sprint-2	Prediction	USN-8	As an end user, when I enter the input parameters of water, the website should predict the exact water quality	7	High	Devatharshini S Vijay Shanmugan S
Sprint-3	Analysis	USN-9	The end users need to analyse the water quality, which is achieved through the provided dataset by which the model is built and trained	10	Medium	Devatharshini S Vijay Shanmugan S
Sprint-3	Security	USN-10	Users expect their data and surfing to be secured	10	Medium	Devatharshini S Vijay Shanmugan S
Sprint-4	Database Access	USN-11	The administrator should maintain the website and update the website regularly.	20	Low	Devatharshini S Vijay Shanmugan S

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)
Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022

7. CODING & SOLUTIONING

7.1 Built a website

The entire website is hosted in the IBM cloud and it was created using ReactJs for the front end and Flask for the backend.

7.2 Get water quality standard inputs

The website is designed in such a way that it gets the six key parameters which are essential to calculate the water quality index from the user.

7.3 Get the Quality of water for the given inputs by using the model trained

By using the six parameters which is entered as input by the user the model in the backend predicts the WQI (Water Quality Index)

7.4 Return the result to the user

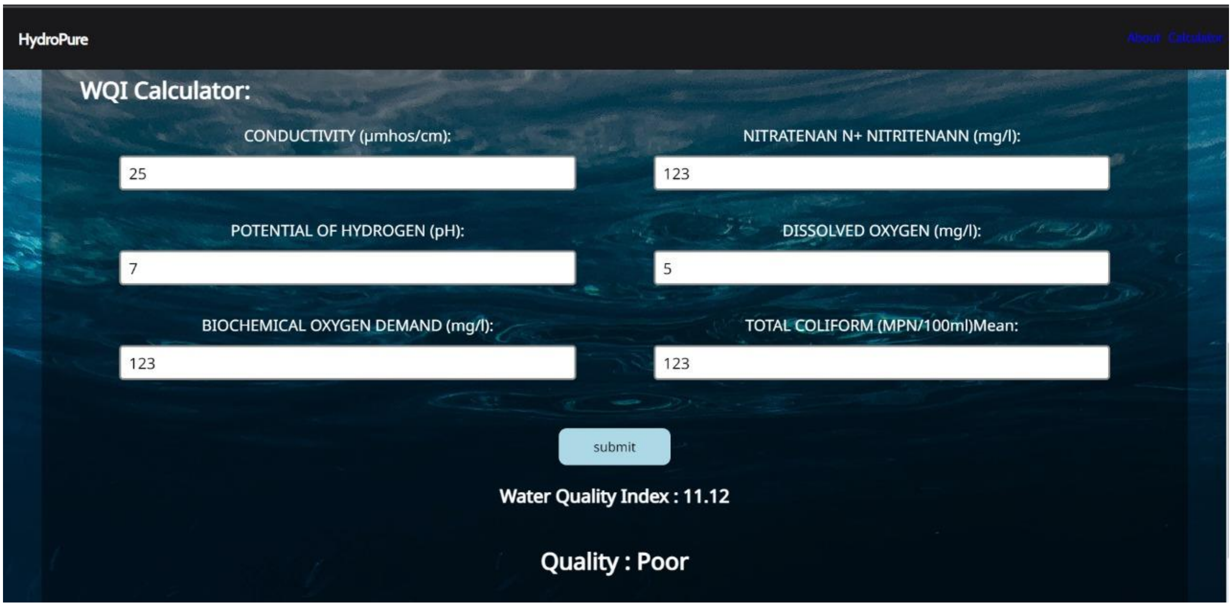
Finally the predicted WQI is displayed to the user.

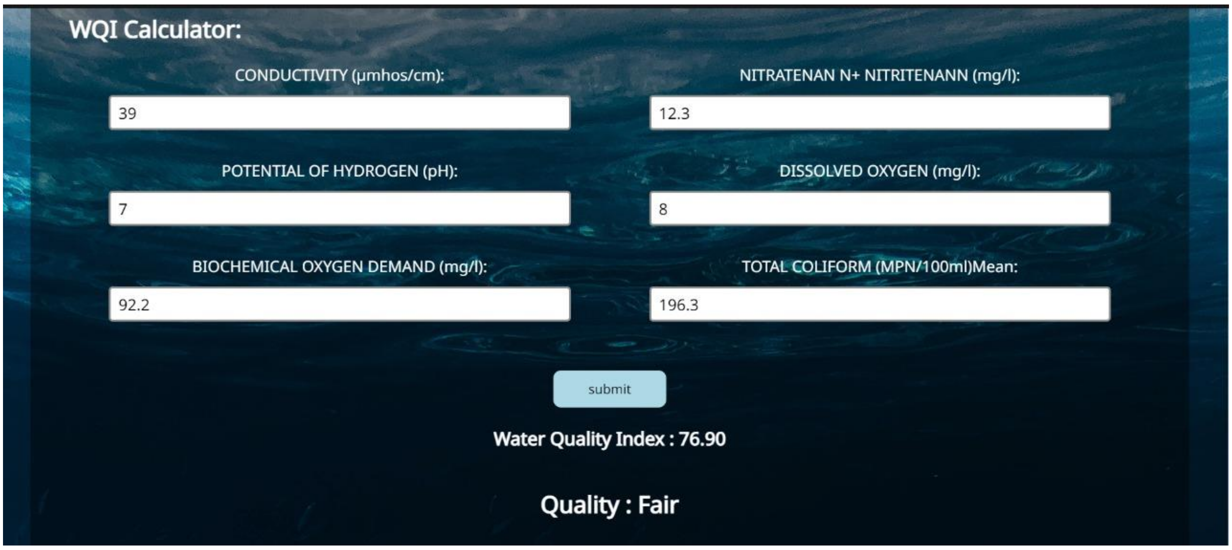
7.5 Workflow

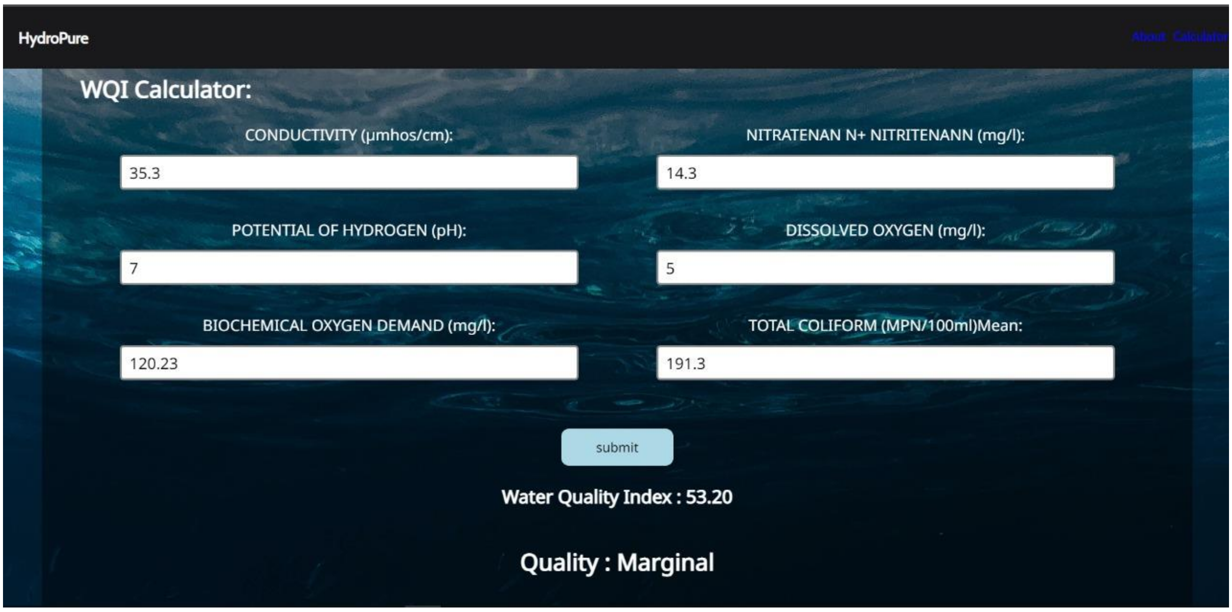
- Authentication screen is implemented to authenticate users by their registered Email id and verified by sending email to their mail.
- After their authentication they are redirected into a home page which consists of a dashboard(WQI calculator) and about page .
- In the home page, they are able to give input to the WQI calculator and get results according to their input with a good, fair tag with them.
- This WQI calculation is done using a machine learning model(Linear regression) . Every time, if a user submits their measurements, they are imputed into the model to get the most accurate WQI prediction value .

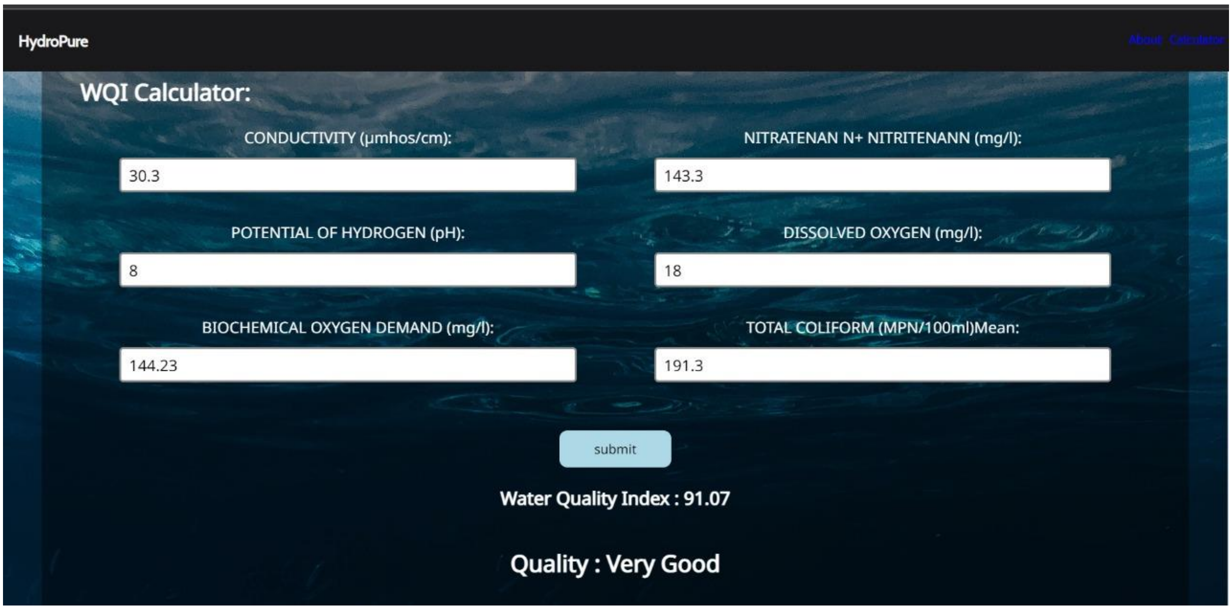
8. TESTING

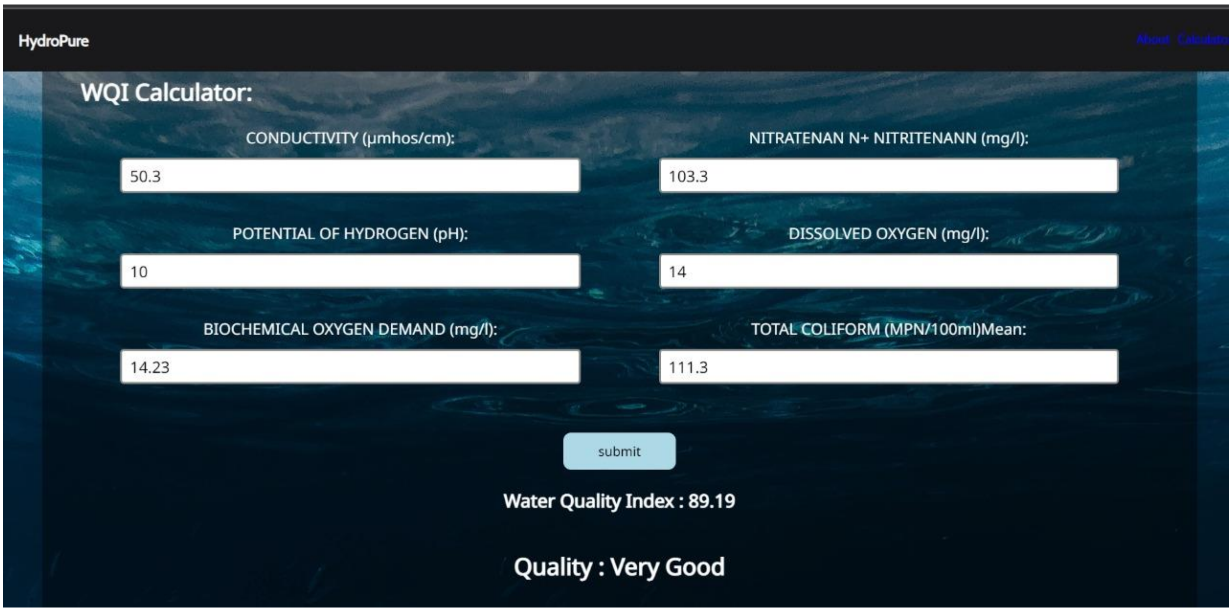
8.1 Test cases:

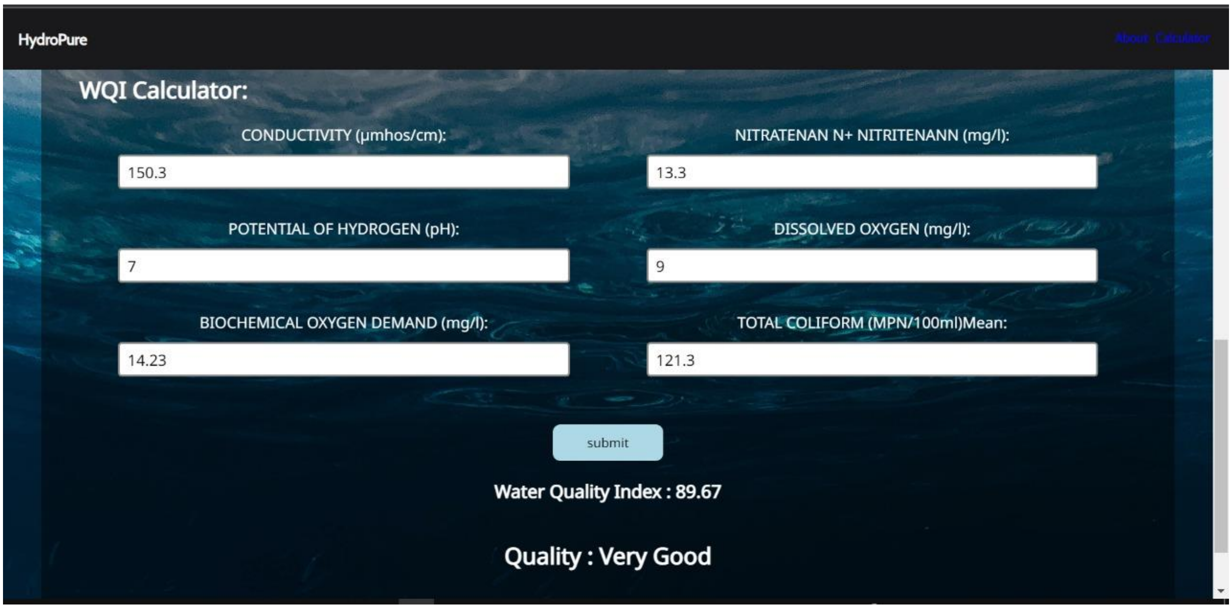





HydroPure
WQI Calculator:
CONDUCTIVITY (μmhos/cm):
35.3
NITRATENAN N+ NITRITENANN (mg/l):
14.3
POTENTIAL OF HYDROGEN (pH):
7
DISSOLVED OXYGEN (mg/l):
5
BIOCHEMICAL OXYGEN DEMAND (mg/l):
120.23
TOTAL COLIFORM (MPN/100ml)Mean:
191.3
submit
Water Quality Index : 53.20
Quality : Marginal


HydroPure
WQI Calculator:
CONDUCTIVITY (μmhos/cm):
30.3
NITRATENAN N+ NITRITENANN (mg/l):
143.3
POTENTIAL OF HYDROGEN (pH):
8
DISSOLVED OXYGEN (mg/l):
18
BIOCHEMICAL OXYGEN DEMAND (mg/l):
144.23
TOTAL COLIFORM (MPN/100ml)Mean:
191.3
submit
Water Quality Index : 91.07
Quality : Very Good


HydroPure
WQI Calculator:
CONDUCTIVITY (μmhos/cm):
50.3
NITRATENAN N+ NITRITENANN (mg/l):
103.3
POTENTIAL OF HYDROGEN (pH):
10
DISSOLVED OXYGEN (mg/l):
14
BIOCHEMICAL OXYGEN DEMAND (mg/l):
14.23
TOTAL COLIFORM (MPN/100ml)Mean:
111.3
submit
Water Quality Index : 89.19
Quality : Very Good



8.2 User acceptance testing

Purpose of Document

The purpose of this document is to briefly explain the test coverage and open issues of the r efficient water quality analysis and prediction project at the time of the release to User Acceptance Testing (UAT).

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**	**7**

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. RESULTS

9.1 Performance Metrics

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. 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)

LINEAR REGRESSION MODEL

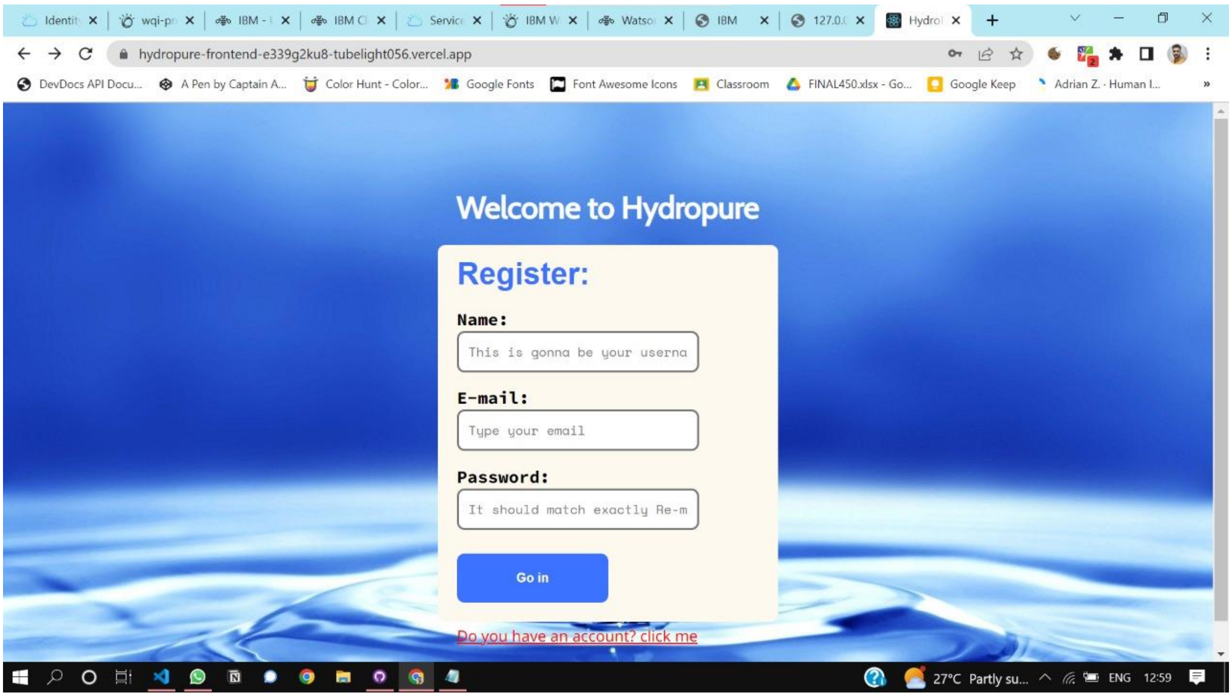
Accuracy =96.8

Mean Absolute error = 1.00485

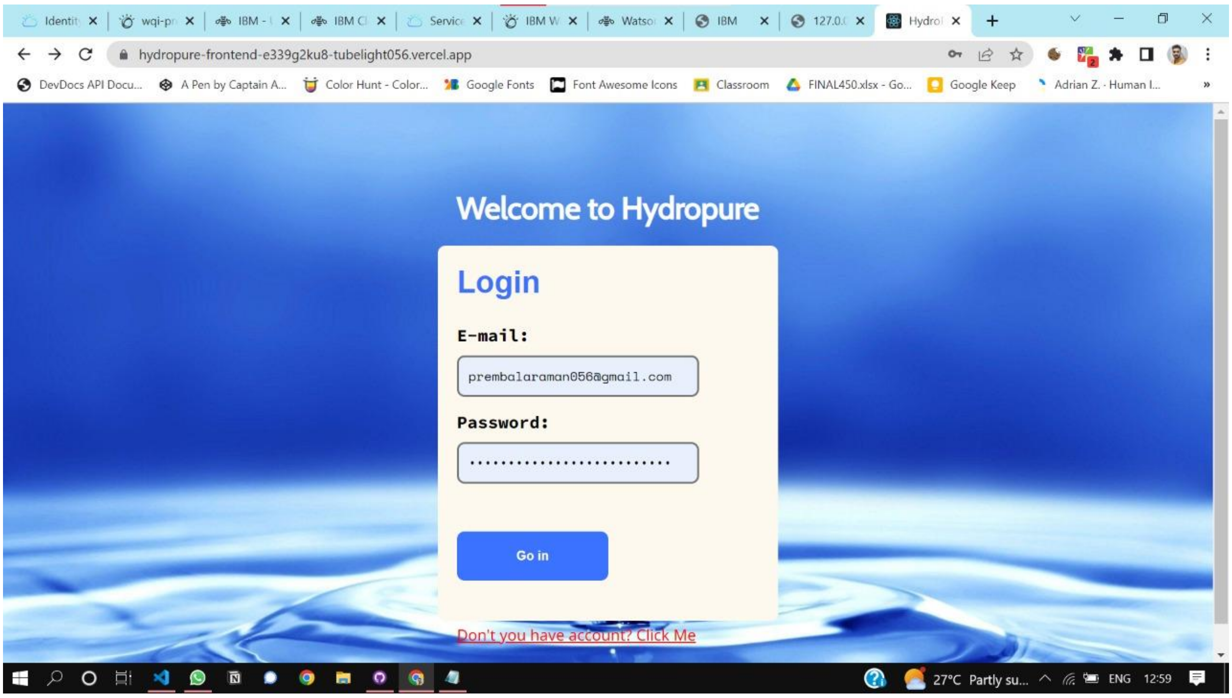
Mean squared error = 5.728

9.2 Web application

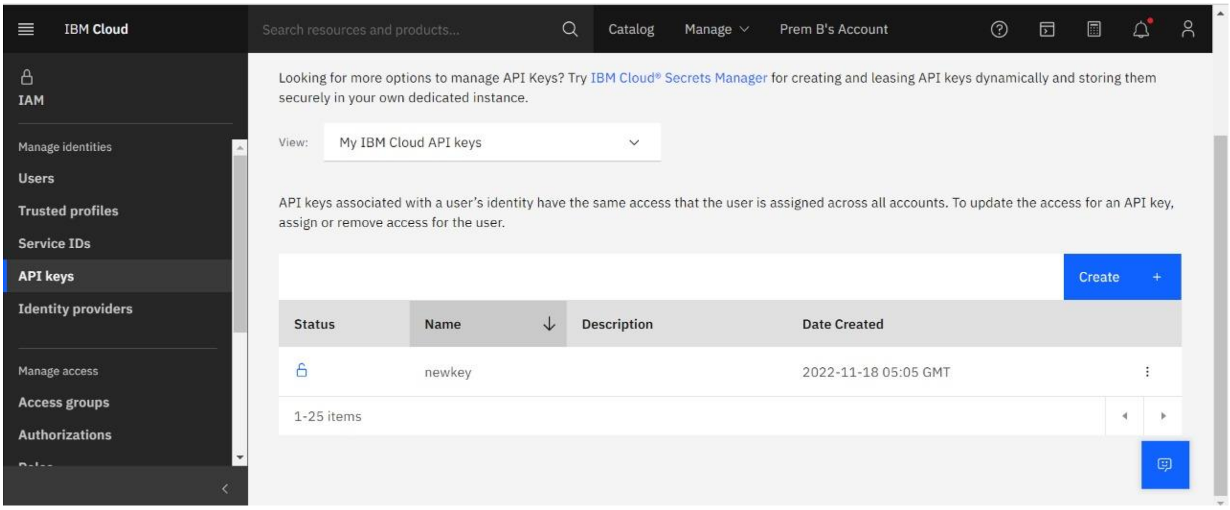
SIGN UP PAGE:

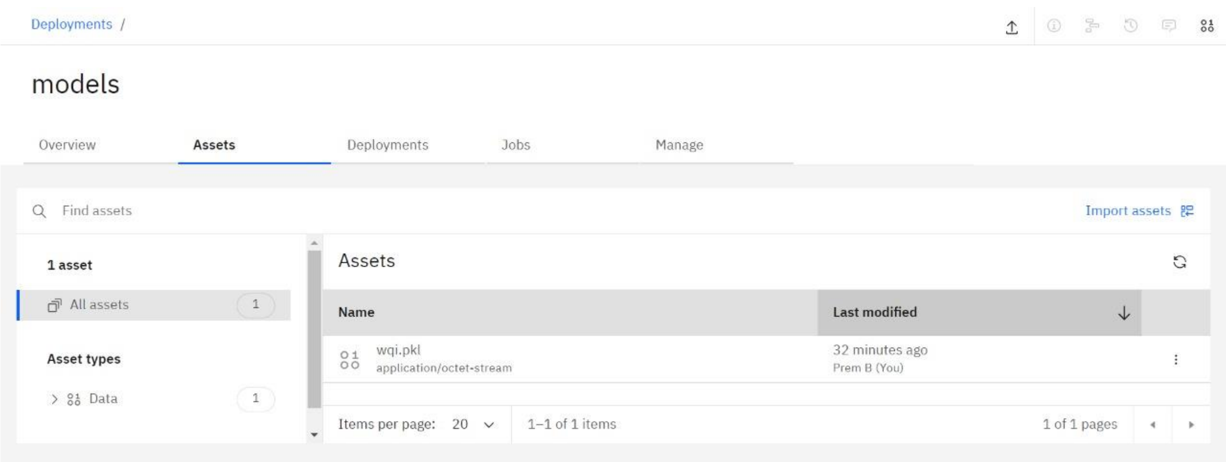


LOGIN PAGE:

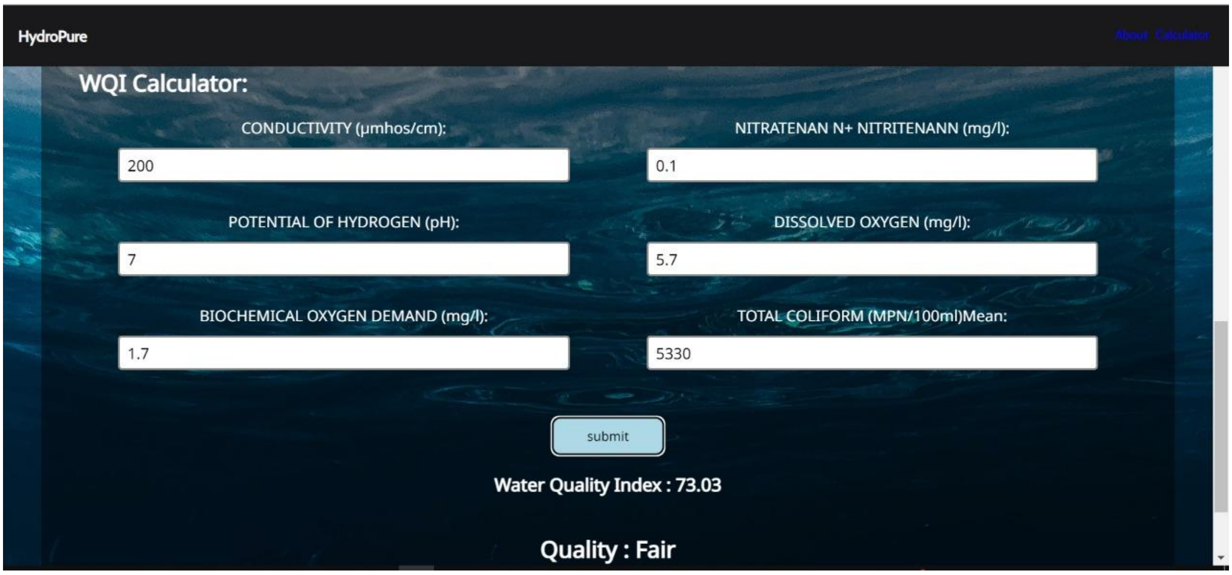


DEPLOYMENT AND KEY:




Deployments /
models
Overview
Assets
Deployments
Jobs
Manage
Find assets
Import assets
1 asset
All assets
1
Asset types
Data
1
Assets
Name
Last modified
wqi.pkl
application/octet-stream
32 minutes ago
Prem B (You)
Items per page: 20
1–1 of 1 items
1 of 1 pages

RESULT:


HydroPure
WQI Calculator:
CONDUCTIVITY (µmhos/cm):
200
NITRATENAN N+ NITRITENANN (mg/l):
0.1
POTENTIAL OF HYDROGEN (pH):
7
DISSOLVED OXYGEN (mg/l):
5.7
BIOCHEMICAL OXYGEN DEMAND (mg/l):
1.7
TOTAL COLIFORM (MPN/100ml)Mean:
5330
submit
Water Quality Index : 73.03
Quality : Fair

10. ADVANTAGES AND DISADVANTAGES

ADVANTAGES:

- The most obvious benefit of water quality analysis and prediction is that it helps to select the best waters for the ways that we want and need to use them — drinking water, swimming, fishing, irrigation, and much more.
- We make use of the linear regression model as it is easier to implement, interpret and efficiently train it.

DISADVANTAGES:

- Training is necessary but it seems a bit difficult to manage over time and with large data sets.
- 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

Water quality, one of the most crucial resources for life, is determined by potability. Traditionally, a costly and drawn-out lab analysis was needed to test the quality of water. This work investigated a different machine learning approach for forecasting water quality using just a few straightforward water quality variables. A group of representative supervised machine learning algorithms for learning were utilised. Before water was made available for consumption, it would find water of poor quality and alert the proper authorities. By decreasing the number of people who consume poor water, it should help prevent illnesses like typhoid and diarrhoea. Future skills to help decision and policy makers in this situation would arise from employing a prescriptive analysis based on projected values.

12. FUTURE SCOPE

Our future scope is to increase the efficiency of the model and also to enhance the security system by employing complex encryption- decryption method for storing and retrieving data.

13. APPENDIX

[1]. Ali El Bilali, Abdeslam Taleb, "Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment",Journal of the Saudi Society of Agricultural Sciences ,Volume 19, Issue 7, October 2020, Pages 439-451

[2]. Qinghong Zou, Qingyu Xiong, Hualing Yi, " A water quality prediction method based on the multi-time scale bidirectional LSTM network",Environ Sci Pollut Res Int. 2020 May.

[3]. Sudhakar Singha, Srinivas Pasupuleti,Soumya S. Singha, Rambabu Singh, Suresh Kumar"Prediction of groundwater quality using efficient machine learning technique",Chemosphere,March 2021.

[4]. Md.Mahedi Hassan, Laboni Akter, Mushfiqur Rahman, Nusrat Jahan, "Efficient Prediction of Water Quality Index(WQI) Using Machine Learning Algorithms",Human-Centric Intelligent Systems,Volume 1, Issue 3-4, December 2021.

[5].Tianan Deng, Kwok-Wing Chau, Huan-Feng Duan, "Machine learning based marine water quality prediction for coastal hydro-environment management",J Environ Manage,Apr 15 2021.

[6]. Smail Dilmi, Mohamed Ladjal, "A novel approach for water quality classification based on the integration of deep learning and feature extraction techniques",Chemometrics and Intelligent Laboratory Systems,July 2021.

[7]. S.Sharath, R.Harish, V.Aishwarya and Dr.M.Preetha, "Water quality prediction using machine learning and flask",International Research Journal of Modernization in Engineering Technology and Science,Volume:0,/Issue:06,June-2022.

[8]. Maryam Imani, Md Mahmudul Hasan, Luiz Fernando Bittencourt, Kent McClymont and Zoran Kapelan, "A novel machine learning application: Water quality resilience prediction Model",Science of The Total Environment,January 2021

[9]. Huimin Jia, Xiaofeng Zhou, "Water Quality Prediction Method Based on LSTM-BP",2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC),in IEEE,23 August 2020.

[10]. Al-Akhir Nayan, Muhammad Golam Kibria, Md. Obaidur Rahman and Joyeta Saha, "River Water Quality Analysis and Prediction Using GBM",2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT),29 November 2020.

GitHub Link: https://github.com/IBM-EPBL/IBM-Project-48462-1660807529

Video Demo Link:

https://drive.google.com/file/d/1hbxfD39oKoLVcyv6BEDI3gY9RMR1O1m5/view?usp=share_link

For Queries, Mail us at: 1905106cse@cit.edu.in