

# **EFFICIENT WATER QUALITY ANALYSIS AND PREDICTION A PROJECT REPORT**

Submitted by

<b>S.PRIYADARSHINI</b>	<b>-(310119104061)</b>
<b>S.PRIYADHARSHINI</b>	<b>-(310119104062)</b>
<b>S.PUSHPAROJA</b>	<b>-(310119104063)</b>
<b>S.SOWMYA</b>	<b>-(310119104076)</b>
<b>D.PRIYADARSHINI</b>	<b>-(310119104060)</b>

*in partial fulfillment for the award of the degree  
of*

**BACHELOR OF ENGINEERING**

*in*

**COMPUTER SCIENCE AND ENGINEERING**

**ANAND INSTITUTE OF HIGHER TECHNOLOGY**



**ANNA UNIVERSITY: CHENNAI 600 025**

**APRIL 2022**

# TABLE OF CONTENT

CHAPTER NO	TITLE	PAGE NO
<b>1</b>	<b>INTRODUCTION</b>	
	1.1 PROJECT OVERVIEW	4
	1.2 PURPOSE	4
<b>2</b>	<b>LITERATURE SURVEY</b>	5
	2.1 EXISTING PROBLEM	7
	2.2 REFERENCES	8
	2.3 PROBLEM STATEMENT DEFINITION	9
<b>3</b>	<b>IDEATION &amp; PROPOSED SOLUTION</b>	
	3.1 EMPATHY MAP CANVAS	9
	3.2 IDEATION & BRAINSTORMING	10
	3.3 PROPOSED SOLUTION	13
	3.4 PROBLEM SOLUTION FIT	15
<b>4</b>	<b>REQUIREMENT ANALYSIS</b>	
	4.1 FUNCTIONAL REQUIREMENT	16
	4.2 NON-FUNCTIONAL REQUIREMENTS	17
<b>5</b>	<b>PROJECT DESIGN</b>	
	5.1 DATA FLOW DIAGRAMS	19
	5.2 SOLUTION & TECHNICAL ARCHITECTURE	20
	5.3 USER STORIES	23
<b>6</b>	<b>PROJECT PLANNING &amp; SCHEDULING</b>	
	6.1 SPRINT PLANNING & ESTIMATION	24
	6.2 SPRINT DELIVERY SCHEDULE	25
	6.3 REPORTS FROM JIRA	26
<b>7</b>	<b>CODING &amp; SOLUTIONING (Explain the features added in the project along with code)</b>	

	7.1 FEATURE 1	27
	7.2 FEATURE 2	29
	7.3 DATABASE SCHEMA (if Applicable)	
<b>8</b>	<b>TESTING</b>	
	8.1 TEST CASES	32
	8.2 USER ACCEPTANCE TESTING	34
<b>9</b>	<b>RESULTS</b>	
	9.1 PERFORMANCE METRICS	34
<b>10</b>	<b>ADVANTAGES &amp; DISADVANTAGES</b>	37
<b>11</b>	<b>CONCLUSION</b>	38
<b>12</b>	<b>FUTURE SCOPE</b>	38
<b>13</b>	<b>APPENDIX</b>	
	SOURCE CODE	39
	GITHUB & PROJECT DEMO LINK	53

# 1.INTRODUCTION

Water quality has a direct impact on public health and the environment. Water is used for various practices, such as drinking, agriculture, and industry. Recently, development of water sports and entertainment has greatly helped to attract tourists ([Jennings 2007](#)). Among various sources of water supply, due to easy access, rivers have been used more frequently for the development of human societies. Using other water resources such as groundwater and seawater sometimes assisted with problems. For example, using groundwater without suitable recharge will lead to land subsidence

## 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

Hence, rapid industrial development has prompted the decay of water quality at a disturbing rate. Furthermore, infrastructures, with the absence of public awareness, and less hygienic qualities, significantly affect the quality of drinking water . In fact, the consequences of polluted drinking water are so dangerous and can badly affect health, the environment, and infrastructures. 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 analyze 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 modeling of the WQ. These methodologies include statistical approaches, visual modeling, analyzing 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 <sup>[13][14]</sup>, Artificial Neural Network <sup>[2][5][6][7]</sup>, Support Vector Machine <sup>[8][9][10]</sup>, K-Nearest Neighbors <sup>[21]</sup> and Naïve Bayes <sup>[18][22][23]</sup>. However, in recent years, some researchers are moving towards more advanced ML ensemble models, such as Gradient Boosting and Random Forest <sup>[1]</sup>

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 <sup>[4]</sup>. 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 <sup>[6]</sup>. Gakii and Jepkoech <sup>[3]</sup> compared five different decision tree classifiers, which are Logistic Model Tree (LMT), J48, Hoeffding 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 Jeihouni et al. <sup>[4]</sup> also compared five decision-tree-based models, which are Random Tree, Random Forest, Ordinary Decision Tree (ODT), Chi-square Automatic Interaction Detector and Iterative Dichotomiser 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 <sup>[8]</sup> used ANN algorithms to predict water quality. They found that Lavenberg Marquardt (LM) algorithm has a better performance than the Gradient Descent Adaptive (GDA) algorithm. Abyaneh <sup>[5]</sup> 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 <sup>[8]</sup>.

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 <sup>[11]</sup> 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 <sup>[12]</sup> 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 <sup>[12]</sup>. This is supported by another study in Eastern Azerbaijan, Iran <sup>[6]</sup>. 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 <sup>[4]</sup>. Naïve Bayes has also been widely used for predicting water quality. A study by Vijay and Kamaraj <sup>[2]</sup> 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 <sup>[22]</sup>. 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 <sup>[9]</sup>. 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 <sup>[30]</sup>. 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 <sup>[30]</sup>. 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 Low Tide at Kuantan River. Bachelor's Thesis, Universiti Malaysia Pahang, Gambang, Malaysia, 2010. Available online: [http://umpir.ump.edu.my/id/eprint/2449/1/JACKY\\_LING\\_KUO\\_BAO.PDF](http://umpir.ump.edu.my/id/eprint/2449/1/JACKY_LING_KUO_BAO.PDF) (accessed on 22 February 2022).
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.; Jamadon, N.K.; Mohmood, A.; Syahir, A. River Pollution Relationship to the National Health Indicated by Under-Five Child Mortality Rate: A Case Study in Malaysia. *Bioremediat. Sci. Technol. Res.* 2015, 3, 20–25.
4. Abbasi, T.; Abbasi, S.A. *Water Quality Indices*; Elsevier: Amsterdam, The Netherlands, 2012.
5. Abyaneh, H.Z. Evaluation of multivariate linear regression and artificial neural 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 Pengkalan Chepa Basin: Water Quality and Heavy Metal Analysis. *Sains Malays.* 2020, 49, 1787–1798.
7. Al-Badaai, F.; Shuhaimi-Othman, M.; Gasim, M.B. Water quality assessment of the Semenyih river, Selangor, Malaysia. *J. Chem.* 2013, 2013, 871056.
8. Asadollah, S.B.H.S.; Sharafati, A.; Motta, D.; Yaseen, Z.M. River water quality index prediction and uncertainty analysis: A comparative study of machine 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 machine 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.



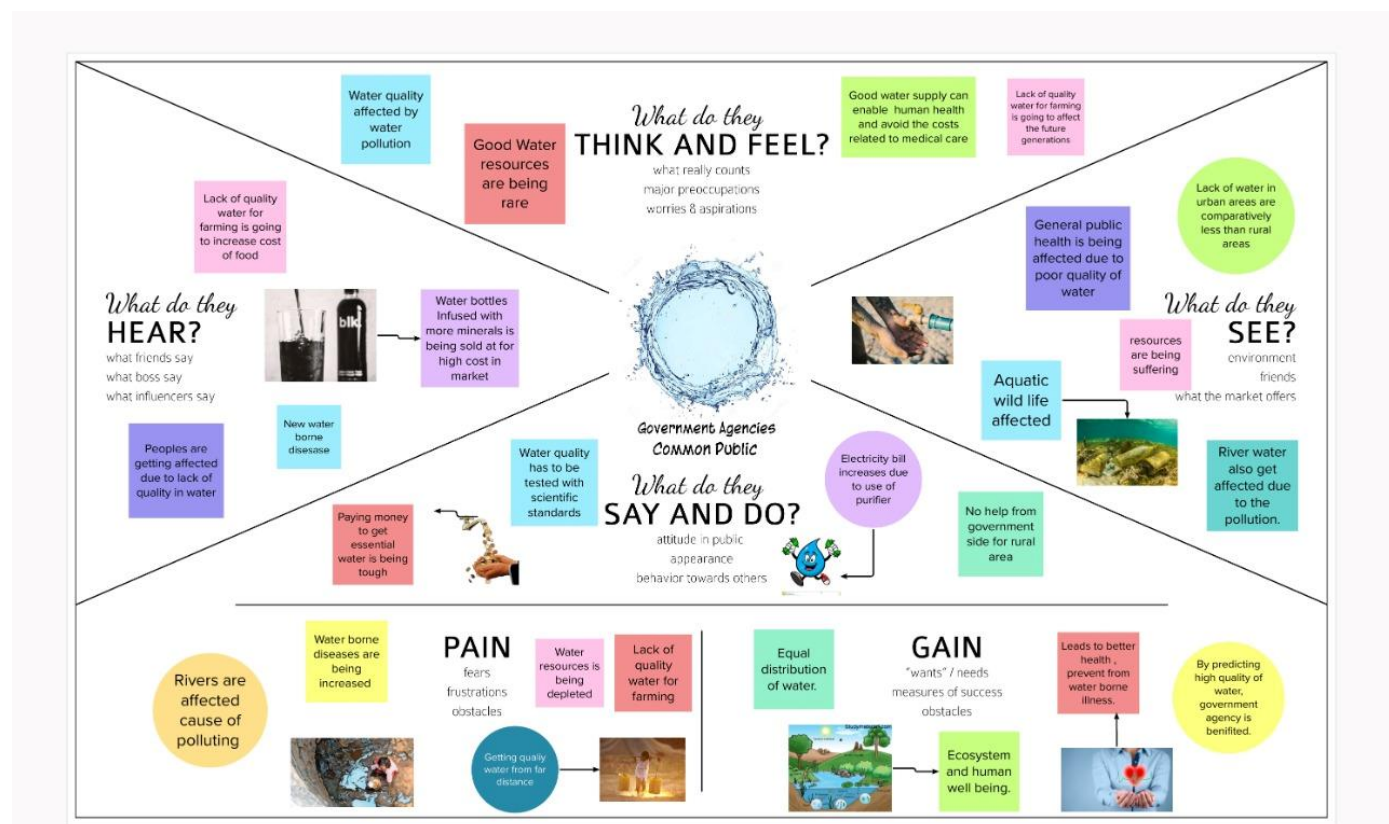
11. Sengorur, B.; Koklu, R.; Ates, A. Water quality assessment using artificial intelligence techniques: SOM and ANN—A case study of Melen River Turkey. *Water Qual. Expo. Health* 2015, 7, 469–490.
12. Aradhana, G.; Singh, N.B. Comparison of Artificial Neural Network algorithm for water quality prediction of River Ganga. *Environ. Res. J.* 2014, 8, 55–63.

## 2.3 PROBLEM STATEMENT DEFINITION

To predict the water safe or not for Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions.


## 3. IDEATION AND PROPOSED SOLUTION

### 3.1 EMPATHY MAP CANVAS






## 3.2 IDEATION AND BRAINSTORMING

Template




### Brainstorm & idea prioritization

Use this template in your own brainstorming sessions so your team can unleash their imagination and start shaping concepts even if you're not sitting in the same room.


 10 minutes to prepare  
 1 hour to collaborate  
 2-8 people recommended

[Share template feedback](#)



#### Before you collaborate

A little bit of preparation goes a long way with this session. Here's what you need to do to get going.

 10 minutes

A


**Team gathering**  
Define who should participate in the session and send an invite. Share relevant information or pre-work ahead.

B

**Set the goal**  
Think about the problem you'll be focusing on solving in the brainstorming session.


C

**Learn how to use the facilitation tools**  
Use the Facilitation Superpowers to run a happy and productive session.

[Open article](#) 


1


**Define your problem statement**  
What problem are you trying to solve? Frame your problem as a How Might We statement. This will be the focus of your brainstorm.


 5 minutes


PROBLEM


TO IMPROVE WATER QUALITY FOR GENERAL PUBLIC USE


**Key rules of brainstorming**  
To run a smooth and productive session


 Stay in topic.

 Encourage wild ideas.

 Defer judgment.

 Listen to others.

 Go for volume.

 If possible, be visual.

2

**Brainstorm**

Write down any ideas that come to mind that address your problem statement.

⌚ 10 minutes

**PRIYADARSHINI.S**

Use water purifiers to get quality water

Government agencies can check scientific standards of water (pH, turbidity etc)

New smart gadgets to test water quality should be used

Government give awareness to public to use quality water

People should properly dispose hazardous products in right place

Public and Government must clean up their surroundings water resources

**PUSHPAROJA.S**

People should be made aware of water borne diseases

Government should check the water quality level continuously

Government should clean up and restore existing waterways and water bodies

**PRIYADHARSHINI.S**

Toxic chemicals should not be disposed near water bodies

Using water quality test kits gives mostly accuracy value of quality in that water

Establish administrative checking system to check water quality

**PRIYADARSHINI.D**

Run cold water taps for 2 minutes before using to cooking and drinking purpose

Government should give more funds to improve and maintain water quality

Appoint officials for monitoring water quality management from government

3

**Group ideas**

Take turns sharing your ideas while clustering similar or related notes as you go. In the last 10 minutes, give each cluster a sentence-like label. If a cluster is bigger than six sticky notes, try and see if you can break it up into smaller sub-groups.

⌚ 20 minutes

Government agencies can check scientific standards of water (pH, turbidity)

Government should give awareness to public to use quality water

Government should check the water quality level continuously

Government give more funds to improve and maintain water quality

Government should clean up and restore existing waterways and water bodies

People should properly dispose hazardous products in right place

Public and Government must clean up their surroundings water resources

Toxic chemicals should not be disposed near water bodies

People should be made aware of water borne diseases

Using test kits gives mostly accuracy value of quality in that water

Establish administrative checking system to check water quality

Appoint officials for monitoring water quality management from government

New smart gadgets to test water quality should be used

Run cold water taps for 2 minutes before using to cooking and drinking purpose

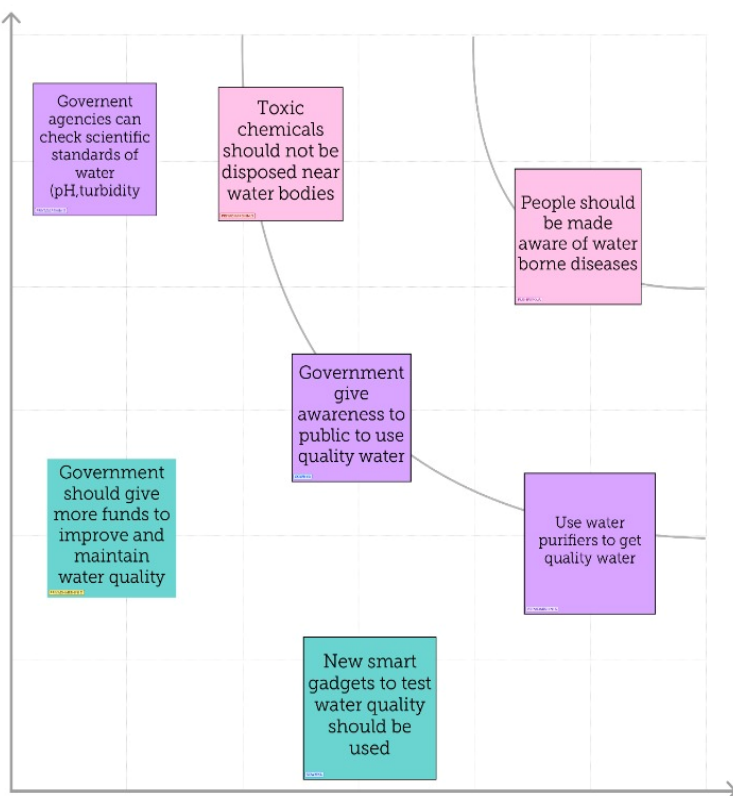
Use water purifiers to get quality water

## Prioritize

Your team should all be on the same page about what's important moving forward. Place your ideas on this grid to determine which ideas are important and which are feasible.

20 minutes

Importance  
 Which of these ideas could get us the most positive impact?



Feasibility  
 Regardless of their importance, which looks more feasible than others? (Cost, time, effort, complexity, etc.)



## After you collaborate

You can export the mural as an image or pdf to share with members of your company who might find it helpful.

## Quick add-ons

- A Share the mural**  
 Share a view link to the mural with stakeholders to keep them in the loop about the outcomes of the session.
- B Export the mural**  
 Export a copy of the mural as a PNG or PDF to attach to emails, include in slides, or save in your drive.

## Keep moving forward

- Strategy blueprint**  
 Define the components of a new idea or strategy.  
[Open the template →](#)
- Customer experience journey map**  
 Understand customer needs, motivations, and obstacles for an experience.  
[Open the template →](#)
- Strengths, weaknesses, opportunities & threats**  
 Identify strengths, weaknesses, opportunities, and threats (SWOT) to develop a plan.  
[Open the template →](#)

[Share template feedback](#)

### 3.3 PROPOSED SOLUTION

S NO	PARAMETER	DESCRIPTION
1.	Problem Statement (Problem to be solved)	<ul style="list-style-type: none"><li>• Water quality prediction using machine learning techniques. Our model predicts the drinkability of the water based parameters such as Ph value, conductivity, and hardness of the water,.</li></ul>
2.	Idea / Solution description	<ul style="list-style-type: none"><li>• Water quality prediction model using the principal component analysis followed by decision tree classification.</li><li>• Firstly, the water quality index (WQI) is calculated using the weighted arithmeticindex method.</li><li>• Secondly, the principal component analysis (PCA) is applied to the dataset, and the most dominant WQI parametershave been extracted.</li><li>• Thirdly, to predict the WQI, different regression algorithms are used to the PCAoutput.</li><li>• Finally, the decision tree classifier model is utilized to classify the waterquality status.</li></ul>
3.	Novelty / Uniqueness	<ul style="list-style-type: none"><li>• In this prediction, the main uniqueness isutilization of PCA and decision tree classifier model.</li></ul>

4.	Social Impact / Customer Satisfaction	<ul style="list-style-type: none"> <li>• This work can demonstrate how setting of more stringent water quality objectives can enhance and protect environmental assets of water resources.</li> <li>• This work can aid in justifying the range of water quality metrics set by government initiatives and to minimize further damages in water resources.</li> <li>• This work can help to quickly identify drinkability of water from new sources.</li> </ul>
5.	Business Model (Revenue Model)	<ul style="list-style-type: none"> <li>• For Analyzing the metrics of each water resource a charge of Rs 100 will be collected.</li> </ul>
6.	Scalability of the Solution	<ul style="list-style-type: none"> <li>• The solution is highly scalable as we use Machine learning technique .</li> <li>• A Automated system can be build to aid the government, to collect the water metrics and quickly analyze and predict the water quality.</li> </ul>

## 3.4 PROBLEM SOLUTION FIT

Project Title: Efficient water quality analysis and prediction using machine learning		Project Design Phase-I - Solution Fit Template		Team ID: PNT2022TMID37146	
Define CS, fit into CC	<b>1. CUSTOMER SEGMENT(S)</b> <span>CS</span> Who is your customer? i.e. working parents of 0-5 y.o. kids <ul style="list-style-type: none"> <li>Government agencies that provide water to general public</li> <li>General public looking for drinkable water with quality</li> </ul>	<b>6. CUSTOMER CONSTRAINTS</b> <span>CC</span> What constraints prevent your customers from taking action or limit their choices of solutions? i.e. spending power, budget, no cash, network connection, available devices. <ul style="list-style-type: none"> <li>Lack of Knowledge about Scientific standards</li> <li>No Tools and available devices to measure water quality</li> </ul>	<b>5. AVAILABLE SOLUTIONS</b> <span>AS</span> 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? i.e. pen and paper is an alternative to digital spreadsheet. <p>Installing RO and heating of water before using</p> <p><b>PROS:</b> Filters the most contaminants  <b>CONS:</b></p> <ul style="list-style-type: none"> <li>More water wasted</li> <li>Drinking water that's too hot can damage tissue</li> </ul>	Explore AS, differentiate	
	<b>2. JOBS-TO-BE-DONE / PROBLEMS</b> <span>J&amp;P</span> Which jobs-to-be-done (or problems) do you address for your customers? There could be more than one; explore different sides. <ul style="list-style-type: none"> <li>To analyze and predict the water quality based on scientific metrics</li> <li>To Check the scientific standards obtained from the water samples</li> <li>Awareness about water quality</li> </ul>	<b>9. PROBLEM ROOT CAUSE</b> <span>RC</span> What is the real reason that this problem exists? What is the back story behind the need to do this job? i.e. customers have to do it because of the change in regulations. <ul style="list-style-type: none"> <li>Water scarcity</li> <li>No availability of quality water</li> <li>We couldn't judge or find the water quality without scientifically testing</li> </ul>	<b>7. BEHAVIOUR</b> <span>BE</span> What does your customer do to address the problem and get the job done? i.e. exactly what? Find the right solar panel installer, calculate usage and benefits, indirectly associated: customers spend less time on housework work (i.e. housework). <ul style="list-style-type: none"> <li>Installing RO</li> <li>Purchasing Mineral water</li> <li>Checking hardness of water</li> </ul>		
Focus on J&P, fit into BE, understand RC	<b>3. TRIGGERS</b> <span>TR</span> What triggers customers to act? i.e. seeing their neighbour installing solar panels, reading about a more efficient solution in the news. <ul style="list-style-type: none"> <li>Water quality good in other countries</li> <li>Non drinkable water quality due to water scarcity in our country</li> </ul>	<b>10. YOUR SOLUTION</b> <span>SL</span> If you are working on an existing business, write down your current solution first, fill in the canvas, and check how much it fits reality. If you are working on a new business proposition, then keep it blank until you fill in the canvas and come up with a solution that fits within customer limitations, solves a problem and matches customer behaviour. <p>By collecting water samples from proven water and new water resources and using some parameters like Ph value, Hardness, Conductivity by using machine learning techniques to predict the water quality</p>	<b>8. CHANNELS of BEHAVIOUR</b> <span>CH</span> <b>8.1 ONLINE</b> What kind of actions do customers take online? Extract online channels from #7. <p>ONLINE: Browsing various ways for getting quality water</p> <b>8.2 OFFLINE</b> What kind of actions do customers take offline? Extract offline channels from #7 and use them for customer development. <p>OFFLINE: Boiling drinking water and Installing RO Process</p>	Identify strong TR & EM	
	<b>4. EMOTIONS: BEFORE / AFTER</b> <span>EM</span> How do customers feel when they face a problem or a job and afterwards? i.e. lost, insecure → confident, in control - use it in your communication strategy & design. <p><b>BEFORE:</b></p> <ul style="list-style-type: none"> <li>Feeling frustrated</li> <li>fear of water borne diseases</li> </ul> <p><b>AFTER:</b></p> <ul style="list-style-type: none"> <li>feeling safe</li> <li>Happy living</li> </ul>				



## 4. REQUIREMENT ANALYSIS

### 4.1 FUNCTIONAL REQUIREMENT

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	<ul style="list-style-type: none"><li>• Registration through Form</li><li>• Registration through Gmail</li><li>• Registration through LinkedIn</li></ul>
FR-2	User Confirmation	<ul style="list-style-type: none"><li>• Confirmation via Email</li><li>• Confirmation via OTP</li></ul>
FR-3	Select the water quality testing parameters	<ul style="list-style-type: none"><li>• Chemical contamination</li><li>• Microbial contamination</li><li>• Physical contamination</li></ul>
FR-4	Physical contamination	<ul style="list-style-type: none"><li>• <b>PH</b>-is important when disinfecting water with chloride</li><li>• <b>EC</b>-unusually high level may suggest chemical contamination.</li><li>• <b>Turbidity</b>-High turbidity decreases water acceptability.</li></ul>



FR-5	Chemical contamination	<ul style="list-style-type: none"> <li>• <b>Fluoride(1.5 mg/l)</b>-Fluoride is a naturally-occurring form of the element fluorine, which is sometimes found in groundwater at levels that exceed safe levels.</li> <li>• <b>Nitrate and Nitrite(50 mg/l)</b>-In most cases, these compounds aren't a serious health risk.</li> <li>• <b>Arsenic(10µg/l)</b>-The EPA says studies link long-term exposure of arsenic to certain cancers as well as cardiovascular, neurological, and other conditions.</li> <li>• <b>Chlorine(5 mg/L)</b>-This value is the health-based guideline. Chlorine is often used for water treatment.</li> </ul>
FR-6	Microbial contamination	<ul style="list-style-type: none"> <li>• <b>E. coli(0 MPN/100 ml)</b>-Provided indication of contamination by fecal coliforms or other harmful bacteria. This is important because fecal pollution is the major cause of water-borne diseases in humans.</li> </ul>

## 4.2 NON-FUNCTIONAL REQUIREMENT

FR No.	Non-Functional Requirement	Description
NFR-1	Usability	<ul style="list-style-type: none"> <li>• To improve usability of data provided to water quality exchange, to monitor nutrient record in water.</li> </ul>

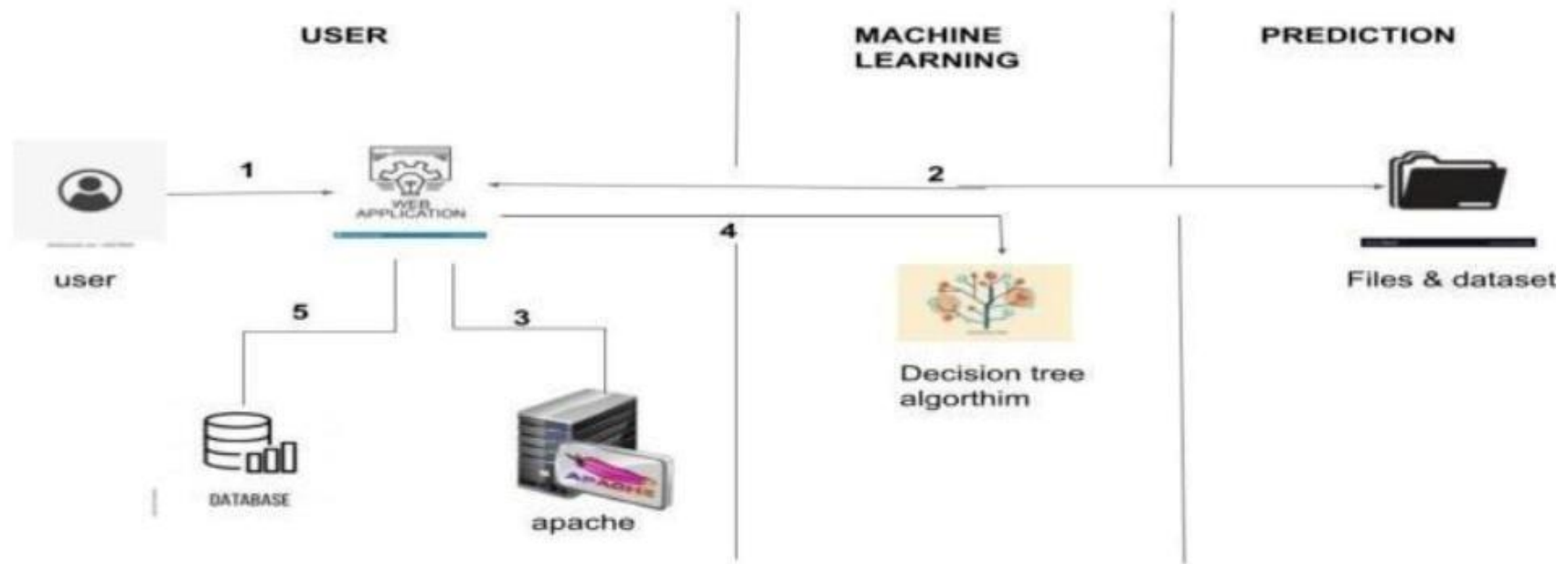
NFR-2	<b>Security</b>	<ul style="list-style-type: none"> <li>The capacity of a population to safeguard sustainable access to adequate quantities of acceptable quality for sustaining livelihoods, human well-being, socio-economic development, for ensuring protection against water borne population and water related diseases.</li> </ul>
NFR-3	<b>Reliability</b>	<ul style="list-style-type: none"> <li><b>System adequacy and system security-</b> A hierarchical framework approach to system adequacy evaluation is presented. Adequacy evaluation techniques for each hierarchical level associated with basic probabilistic indices.</li> </ul>
NFR-4	<b>Performance</b>	<ul style="list-style-type: none"> <li>The presence of certain contaminants in our water can lead to health issues, including gastrointestinal illness, reproductive problems, and neurological disorders. Infants, young children, pregnant women, the elderly, and people with weakened immune systems may be especially at risk for illness.</li> </ul>
NFR-5	<b>Availability</b>	<ul style="list-style-type: none"> <li>Low levels of rainfall and high temperatures lead to water deficits . When rainfall is low, there is less water available. When temperatures are high, water evaporates and so there is less available to use. Water surpluses are common where rainfall is high and temperatures are lower.</li> </ul>
NFR-6	<b>Scalability</b>	<ul style="list-style-type: none"> <li>Scaling occurs when water has high levels of minerals like calcium carbonate, which can build-up on surfaces. Slight scaling can be considered beneficial in that the inside surfaces of metal pipes become coated with harmless minerals that act as a barrier to corrosion.</li> </ul>

## 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.

Flow diagram:



## 5.2 SOLUTION AND TECHNICAL 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 analyze 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.

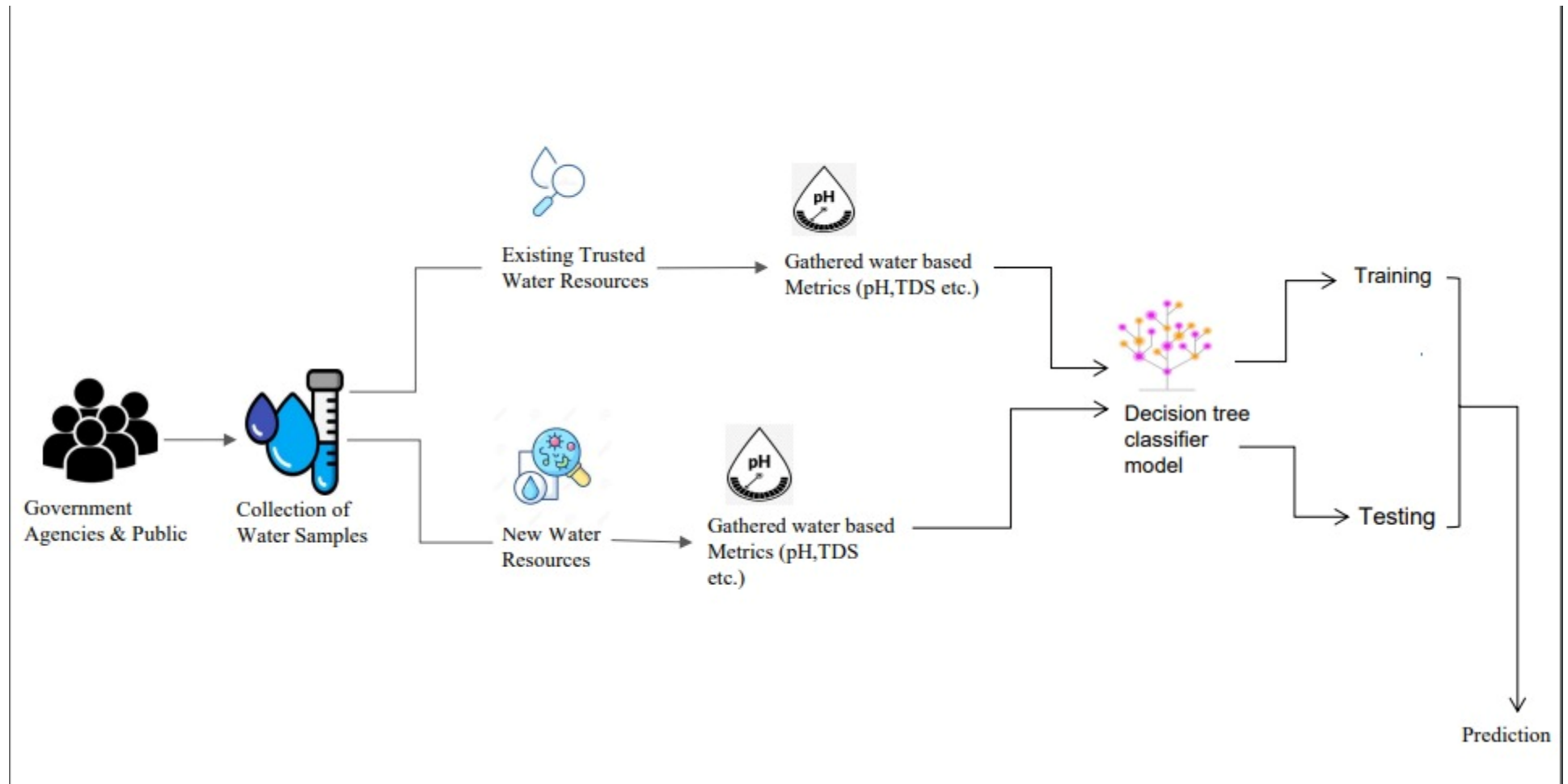
### *H. Data Modeling*

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

Use the below template to list all the user stories for the product.

User Type	Functional Requirement (Epic)	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 and can receive confirmationemail & click confirm	High	Sprint-1
	Login	USN-2	As a user, I can log into the application by entering email & password	I can log into the application by entering email & password	High	Sprint-1
	Dashboard	USN-3	As a user,I will receive the information about myself once I have login into the application	I can receive information about myself once login	High	Sprint-1
Customer (Web user)	Register	USN-4	As a web 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
	Login	USN-5	As a web user, I can log into the application by entering email & password	I can log into the application by entering email & password	High	Sprint-1
	Dashboard	USN-6	As a web user,I will receive the information about myself once I have login into the application	I can receive information about myself once login	High	Sprint-1
Customer Care Executive	Login	USN-7	Connect with the service by login	Can get connected with the server	Medium	Sprint-2
Administrator	Provide the service	USN-8	Data is given by the user to admin	Can add or update the data provided by the user	High	Sprint-1

## 6. PROJECT PLANNING AND SCDULING

### 6.1 SPRINT PLANNING AND ESTIMATION

#### Product Backlog, Sprint Schedule, and Estimation (4 Marks)

Use the below template to create product backlog and sprint schedule

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story points	Priority	Team Members
Sprint-1	Registration & Login	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	3	High	S.Priyadarshini
		USN-2	As a user , I will receive confirmation E-mail once I have registered for the application	2	High	S.Pushparoja
		USN-3	As a user, I can log into the application by entering email & password	1	High	S.Priyadharshini
Sprint-2	Dashboard	USN-4	As a user, I will able to see my profile information once I have login into the application	2	High	S.Sowmya
		USN-5	As a user, I will be able to logout from my account in dashboard from application	1	High	S.Pushparoja
Sprint-3	Data Collection & Processing	USN-6	As a user, I will give information about collected water samples.	3	High	D.Priyadarshini
		USN-7	User should give PH, temperature,conductivity,solid,hardness of the water	4	High	S.Priyadharshini
		USN-8	In data processing, data will get processed based on collected water sample	5	Medium	S.Priyadarshini



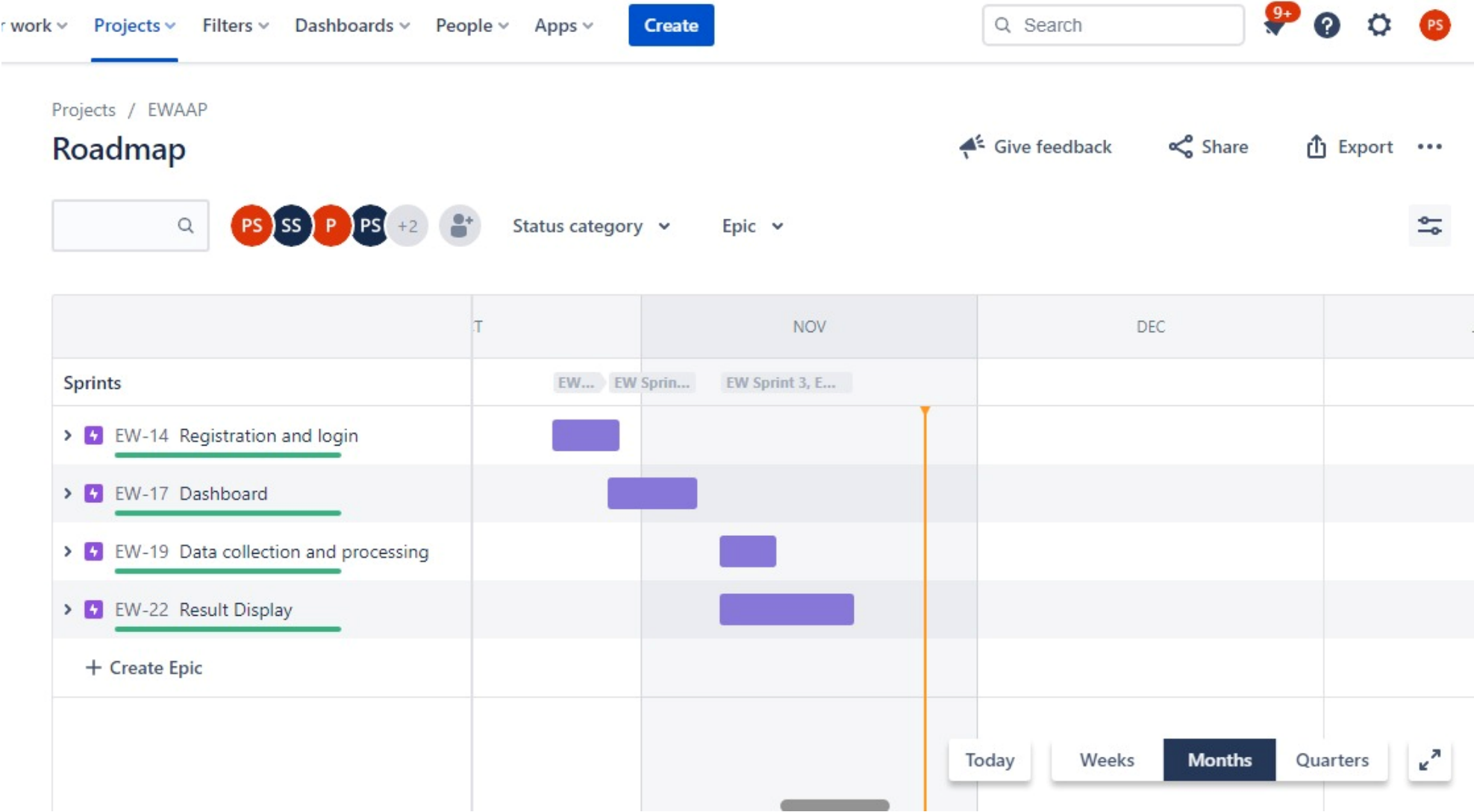
Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story points	Priority	Team Members
		USN-9	From the user given ph and temperature,conductivity,solid,hardness of the sample water and it will process by using machine learning technology to predict the water quality.	6	High	S.Pushparoja
Sprint-4	Result Display	USN-10	As a user, I can get the result of the given water sample are quality or not with accuracy in a result page in the app	3	High	S.Sowmya

## 6.2 SPRINT SCHEDULE

Project Tracker, Velocity & Burndown Chart: (4 Marks)

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	6	6 Days	24 Oct 2022	29 Oct 2022	6	19 Nov 2022
Sprint-2	3	6 Days	31 Oct 2022	05 Nov 2022	3	19 Nov 2022
Sprint-3	18	6 Days	07 Nov 2022	12 Nov 2022	18	19 Nov 2022
Sprint-4	3	6 Days	14 Nov 2022	19 Nov 2022	3	19 Nov 2022

# 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	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

#### 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

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

```
[42] 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:

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

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

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

where:  $V_i$  is the measured value of parameter in the tested water samples,  $V_{Ideal}$  is the ideal value of parameter in pure water (0 for all parameters except  $pH$  and  $DO$ ), and  $S_i$  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.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

```
# Random Forest Classifier
from sklearn.ensemble import RandomForestClassifier
rf_classifier = RandomForestClassifier(n_estimators = 20, criterion = 'entropy', class_weight = "balanced_subsample", random_state = 51)
rf_classifier.fit(X_train_final, y_train)
y_pred = rf_classifier.predict(X_test_final)
accuracy_score(y_test, y_pred)
```

0.635

```
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

```
# 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

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

	precision	recall	f1-score	support
0	0.64	0.90	0.75	497
1	0.50	0.17	0.25	303
accuracy			0.62	800
macro avg	0.57	0.53	0.50	800
weighted avg	0.59	0.62	0.56	800

## Support vector Machine

```
[53] # Support vector classifier
      from sklearn.svm import SVC
      svc_classifier = SVC(class_weight = "balanced" )
      svc_classifier.fit(X_train_final, y_train)
      y_pred_scv = svc_classifier.predict(X_test_final)
      accuracy_score(y_test, y_pred_scv)
```

0.6225

```
[54] print(classification_report(y_test, y_pred_scv))
```

	precision	recall	f1-score	support
0	0.70	0.69	0.70	497
1	0.50	0.50	0.50	303
accuracy			0.62	800
macro avg	0.60	0.60	0.60	800
weighted avg	0.62	0.62	0.62	800

The SVM model was developed in 1995 by Corinna Cortes and Vapnik. 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

The screenshot displays a web browser window with a single tab titled "WATER QUALITY". The address bar shows the URL "127.0.0.1:5000/predict". The main content area features a background image of a water droplet creating ripples. Overlaid on this are several input fields for water quality parameters, each with a numerical value entered:

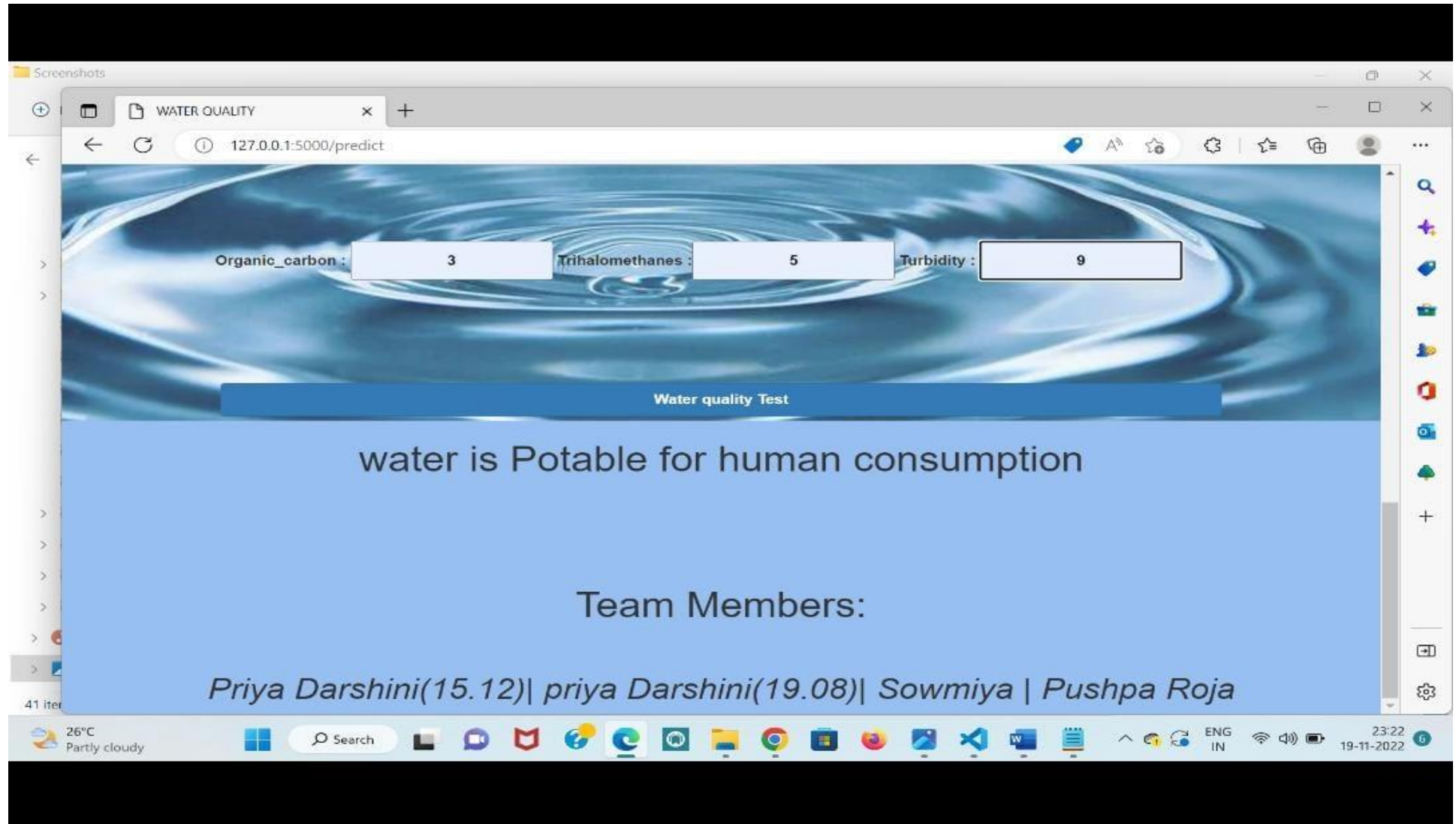
- pH value : 7
- Hardness : 7
- Solids : 0
- Chloramines : 5
- Sulfate : 7
- Conductivity : 8
- Organic\_carbon : 3
- Trihalomethanes : 5
- Turbidity : 9

Below the input fields, a blue bar contains the text "Water quality Test". At the bottom of the page, a large light blue box displays the prediction: "water is Potable for human consumption".

The Windows taskbar at the bottom shows the system clock as 23:22 on 19-11-2022, along with weather information (26°C, Partly cloudy) and various application icons.



## TEST CASE 2



## 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

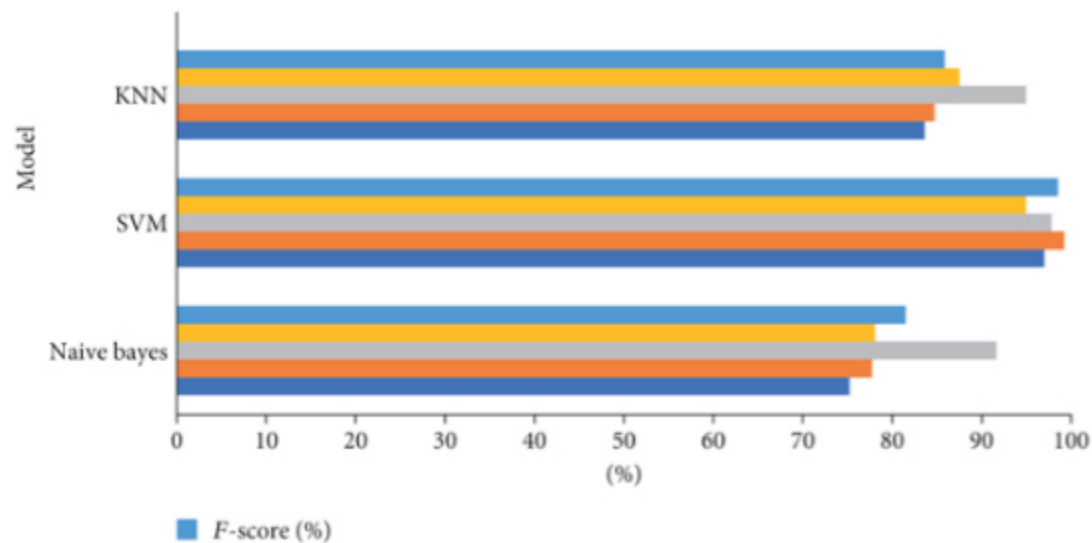
This report shows the number of test cases that have passed, failed, and untested

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 =  $\frac{TP+TN}{TP+FP+FN+TN}$**

**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 . FUTURE SCOPE

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

### SOURCE CODE

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv("water_potability.csv")
df.head()

df.shape

df.info()

df.describe()

sns.countplot(x='Potability',data=df )

df["Potability"].value_counts()

print(f"0 : {round(1998 /3276 * 100 , 2)}")
print(f"1 : {round(1278 /3276 * 100 , 2)}")

#EDA

df.isnull().sum()
```

```

for feature in df.columns:
    if df[feature].isnull().sum()>0:
        print(f"{feature} : {round(df[feature].isnull().mean(),4)*100}%")
## Fill missing values with median
for feature in df.columns:
    df[feature].fillna(df[feature].median() , inplace = True)

## find duplicate rows in dataset
duplicate = df[df.duplicated()]
duplicate

for i in df.columns:
    print(f" {i} : {len(df[i].unique())}")

for feature in df.columns:
    if feature == "Potability":
        pass
    else:
        bar = sns.histplot(df[feature] , kde_kws = {'bw' : 1} , )
        bar.legend(["Skewness: {:.2f}"].format(df[feature].skew()))
        plt.xlabel(feature)
        plt.ylabel("Probability density")
        plt.title(feature)
        plt.show()

# we don't have missing values in our dataset so we can skip if condition
for feature in df.columns:
    if 0 in df[feature].unique():# because log 0 is not defined thats why we are using this condition or we can also use log1p
        pass

```



```

else:
    df.boxplot(column=feature)
    plt.ylabel(feature)
    plt.title(feature)
    plt.show()

# removing outliers
Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)
IQR = Q3 - Q1
print(IQR)

df = df[~((df < (Q1 - 1.5 * IQR)) |(df > (Q3 + 1.5 * IQR))).any(axis=1)]
df.shape

df["Potability"].value_counts()

## Correlation
plt.figure(figsize=(25,25))
ax = sns.heatmap(df.corr(), cmap = "coolwarm", annot=True, linewidth=2)

Multivariate analysis
sns.pairplot(df , height=10 , size = 5 , hue = "Potability" )

### splitting data into x and y
X = df.iloc[:, :-1]
y = df.iloc[:, -1]

# split dataset into train and test

```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state= 5)
```

```
#from collections import Counter
#from imblearn.over_sampling import SMOTE

#counter = Counter(y_train)
#print(f"before oversampling: {counter}")
#smt = SMOTE()
#X_train , y_train = smt.fit_resample(X_train , y_train)
#counter = Counter(y_train)
#print(f"after oversampling : {counter}")
```

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
```

# Random Forest Classifier

```
from sklearn.ensemble import RandomForestClassifier
rf_classifier = RandomForestClassifier(n_estimators = 20, criterion = 'entropy', class_weight = "balanced_subsample", random_state = 51)
rf_classifier.fit(X_train_final, y_train)
```

```
y_pred = rf_classifier.predict(X_test_final)
accuracy_score(y_test, y_pred)
```

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

```
# 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)
```

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

Support vector Machine

```
# Support vector classifier
```

```
from sklearn.svm import SVC
svc_classifier = SVC(class_weight = "balanced" )
svc_classifier.fit(X_train_final, y_train)
y_pred_scv = svc_classifier.predict(X_test_final)
accuracy_score(y_test, y_pred_scv)
```

```
print(classification_report(y_test, y_pred_scv))
```

```
cm = confusion_matrix(y_test, y_pred_scv)
plt.title('Heatmap of Confusion Matrix', fontsize = 12)
sns.heatmap(cm, annot = True, fmt = "d")
```

```
plt.show()
```

### Hyperparameter Tuning with Support vector Machine

```
# defining parameter range
```

```
param_grid = {'C': [0.1, 1, 10, 100, 200, 400, 600, 800],  
              'gamma': [1, 0.1, 0.01, 0.001, 0.0001],  
              'kernel': ['rbf']}
```

```
from sklearn.model_selection import GridSearchCV
```

```
grid = GridSearchCV(SVC(), param_grid, refit = True, verbose = 3)
```

```
# fitting the model for grid search
```

```
grid.fit(X_train_final, y_train)
```

```
# print best parameter after tuning
```

```
print(grid.best_params_)
```

```
# print how our model looks after hyper-parameter tuning
```

```
print(grid.best_estimator_)
```

```
# Support vector classifier
```

```
from sklearn.svm import SVC
```

```
svc_classifier = SVC(class_weight = "balanced" , C=100, gamma=0.01)
```

```
svc_classifier.fit(X_train_final, y_train)
```

```
y_pred_scv = svc_classifier.predict(X_test_final)
```

```
accuracy_score(y_test, y_pred_scv)
```

```

print(classification_report(y_test, y_pred_xgb))

cm = confusion_matrix(y_test, y_pred_scv)
plt.title('Heatmap of Confusion Matrix', fontsize = 12)
sns.heatmap(cm, annot = True , fmt = "d")
plt.show()

## Pickle
from sklearn.svm import SVC
import pickle

# save model
pickle.dump(svc_classifier, open('model.pkl', 'wb'))

# load model
water_quality_model = pickle.load(open('model.pkl', 'rb'))

# predict the output
y_pred =water_quality_model.predict(X_test_final)

# confusion matrix
print('Confusion matrix of Support vector Machine : \n',confusion_matrix(y_test, y_pred),'\n')

```

## APP.PY

```
from flask import Flask, request, render_template
import pickle
import pandas as pd
import numpy as np
import joblib
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"
else:
    prediction = "not safe"

return render_template('home.html', prediction_text= "water is { } for human consumption ".format(prediction))

if __name__ == "__main__":
    app.run(debug=True)
```

## 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  home.html
app.py > hello
1  from flask import Flask, request, render_template
2  import pickle
3  import pandas as pd
4  import numpy as np
5  import joblib
6  scaler = joblib.load("my_scaler.save")
7
8
9  app = Flask(__name__)
10 model = pickle.load(open('model.pkl', 'rb'))
11
12 @app.route("/home")
13 @app.route("/")
14 def hello():
15     return render_template("home.html")
16
17 @app.route("/predict", methods = ["GET", "POST"])
18 def predict():
```

app.py X

sprint 4 > app.py > hello

```
13 @app.route("/")
14 def hello():
15     return render_template("home.html")
16
17 @app.route("/predict", methods = ["GET", "POST"])
18 def predict():
19     if request.method == "POST":
20         input_features = [float(x) for x in request.form.values()]
21         features_value = [np.array(input_features)]
22
23         feature_names = ["ph", "Hardness" , "Solids", "Chloramines", "Sulfate",
24                           "Conductivity", "Organic_carbon", "Trihalomethanes", "Turbidity"]
25
26         df = pd.DataFrame(features_value, columns = feature_names)
27         df = scaler.transform(df)
28         output = model.predict(df)
29
30         if output[0] == 1:
31             prediction = "safe"
```

## WATER QUALITY.IPYNB

### Support vector Machine

```
[53] # Support vector classifier
      from sklearn.svm import SVC
      svc_classifier = SVC(class_weight = "balanced" )
      svc_classifier.fit(X_train_final, y_train)
      y_pred_scv = svc_classifier.predict(X_test_final)
      accuracy_score(y_test, y_pred_scv)
```

0.6225

```
[54] print(classification_report(y_test, y_pred_scv))
```

	precision	recall	f1-score	support
0	0.70	0.69	0.70	497
1	0.50	0.50	0.50	303
accuracy			0.62	800
macro avg	0.60	0.60	0.60	800
weighted avg	0.62	0.62	0.62	800

# HOME.HTML

```
File Edit Selection View Go Run Terminal Help • app.py - sprint 4 - Visual Studio Code
```

EXPLORER

- SPRINT 4
  - env
    - Include
    - Lib
    - Scripts
    - share
    - xgboost
  - pyenv.cfg
  - static\CSS
    - images.jpeg
  - sea.jpg
  - templates
    - home.html
    - images.jpeg
    - sea.jpg
  - app.py
  - model.pkl
  - my\_scaler.save

app.py

```
1 from flask import Flask, request, render_template
2 import pickle
3 import pandas as pd
4 import numpy as np
5 import joblib
6 scaler = joblib.load("my_scaler.save")
7
8
9 app = Flask(__name__)
10 model = pickle.load(open('model.pkl', 'rb'))
11
12 @app.route("/home")
13 @app.route("/")
14 def hello():
15     return render_template("home.html")
16
17 @app.route("/predict", methods = ["GET", "POST"])
18 def predict():
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL JUPYTER

WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.

\* Running on http://127.0.0.1:5000

Press CTRL+C to quit

\* Restarting with stat

C:\Users\dhars\Downloads\sprint 4\env\lib\site-packages\sklearn\base.py:329: UserWarning: Trying to unpickle estimator StandardScaler from version 0.24.0 when using version 1.1.3. This might lead to breaking code or invalid results. Use at your own risk. For more info please refer to: [https://scikit-learn.org/stable/model\\_persistence.html#security-maintainability-limitations](https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations)

warnings.warn(

\* Debugger is active!

Ln 15, Col 40 Spaces: 4 UTF-8 CRLF Python 3.9.13 ('env': venv)

26°C Partly cloudy Search

23:15 19-11-2022

**LINKS:**

GITHUB :

<https://github.com/IBM-EPBL/IBM-Project-15900-1659606030>