

EFFICIENT WATER QUALITY ANALYSIS AND PREDICTION

BHARATHI G	(19I207)
DEVA DHARSHI M	(19I210)
RITHI RAJKUMAR	(19I249)
VIDYA SREE V S	(19I259)

Dissertation submitted in partial fulfilment of the requirements for
the degree of

BACHELOR OF TECHNOLOGY

Branch: INFORMATION TECHNOLOGY

Of Anna University



NOVEMBER 2022

DEPARTMENT OF INFORMATION TECHNOLOGY

PSG COLLEGE OF TECHNOLOGY

(Autonomous Institution)

COIMBATORE – 641 004

PSG COLLEGE OF TECHNOLOGY

(Autonomous Institution)

COIMBATORE – 641 004

EFFICIENT WATER QUALITY ANALYSIS AND PREDICTION

Bonafide record of work done by

BHARATHI G (19I207)

DEVA DHARSHI M (19I210)

RITHI RAJKUMAR (19I249)

VIDYA SREE V S (19I259)

Dissertation submitted in partial fulfillment of the requirements for
the degree of

BACHELOR OF TECHNOLOGY

Branch: INFORMATION TECHNOLOGY

Of Anna University

NOVEMBER 2022

INDUSTRY MENTORS NAME: Ms.LALITHA GAYATHRI

FACULTY MENTORS NAME: Ms.SENTHIL PRABHA RAJAGOPAL

Faculty Guide

Head of the Department

(Internal Examiner)

(External Examiner)

TABLE OF CONTENTS

S.no	Title	Pg no
1.	INTRODUCTION	5
1.1	Project overview	5
1.2	Purpose	5
2.	LITERATURE SURVEY	6
2.1	Existing problems	9
2.2	Reference	9
2.3	Problem statement definition	10
3.	IDEATION AND PROPOSED SOLUTION	11
3.1	Empathy Map Canvas	11
3.2	Ideation & Brainstorming	12
3.3	Proposed Solution	14
3.4	Problem Solution fit	16
4.	REQUIREMENT ANALYSIS	17
4.1	Functional requirements	17
4.2	Non-Functional requirements	18
5.	PROJECT DESIGN	18
5.1	Data Flow Diagrams	18
5.2	Solution & Technical Architecture	19
5.3	User Stories	23
6.	PROJECT PLANNING & SCHEDULING	24
6.1	Sprint Planning & Estimation	24
6.2	Sprint Delivery Schedule	25
6.3	Reports from JIRA	26
7.	CODING & SOLUTIONING (Explain the features added in the project along with code)	26

7.1	Feature 1	26
7.2	Feature 2	29
7.3	Database Schema (if Applicable)	29
8.	TESTING	33
8.1	Test Cases	33
8.	User Acceptance Testing	33
9.	RESULTS	34
9.1.	Performance Metrics	34
10.	ADVANTAGES & DISADVANTAGES	35
11.	CONCLUSION	36
12.	FUTURE SCOPE	36
13.	APPENDIX Source code Github & Project Demo Link	36

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

1.1PROJECT 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.

Massive increases in population, the industrial revolution, and the use of fertilizers and pesticides have led to serious effects on the Water Quality environments.

Thus, having models for the prediction of the Water Quality Index is of great help for monitoring water contamination.

1. LITERATURE SURVEY

Many papers had been referred to predict water quality using Machine Learning (ML) approaches. Some researchers used the traditional Machine Learning models, such as Decision Tree, Artificial Neural Network, Support Vector Machine, K-Nearest Neighbours and Naïve Bayes. However, in recent years, some researchers are moving towards more advanced ML ensemble models, such as Gradient Boosting and Random Forest. Traditional Machine Learning models, such as the Decision Tree model, are frequently found in the literature and performed well on water quality data.

1. Efficient Water Quality Prediction Using Supervised Machine Learning [2019] - Umair Ahmed Et al.

The proposed methodology has parameters: temperature, turbidity, pH and total dissolved solids. The proposed methodology achieves reasonable accuracy using a minimal number of parameters to validate the possibility of its use. This showed that polynomial regression with a degree of 2, and gradient boosting, with a learning rate of 0.1, outperformed other regression algorithms by predicting WQI most efficiently, while MLP with a configuration of (3, 7) outperformed other classification algorithms by classifying WQC most efficiently.

2. Prediction of Water Classification using Machine Learning [2015] - Wahab Et al.

Decision-tree-based models are more favourable to short-term prediction and may have a quicker calculation. Decision-tree-based ensemble models have their data, not being sensitive to missing values and being highly efficient when compared to other ML models. Based on the comparison among the 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.

3. Machine learning algorithms for efficient water quality prediction [2021] - Mourade Et al.

We take the advantages of machine learning algorithms to develop a model that is capable of predicting the water quality index and then the water quality class. The method we propose is based on four water parameters: temperature, pH, turbidity and coliforms. The use of the multiple regression algorithms has proven to be important and effective in predicting the water quality index. In addition, the adoption of the artificial neural network provides the most highly efficient way to classify the water quality.

4. A Review of the application of machine learning in water quality evaluation [2022] - Mengyuan Zhu Et al.

Machine learning is widely used in water quality monitoring and prediction. The performance of 45 machine learning algorithms is evaluated and discussed. More advanced sensors, including soft sensors are developed and applied in water quality. The feasibility and reliability of the algorithms and models are developed with an accuracy of over 90%.

5. Predicting and Analysing Water Quality using Machine Learning: A Comprehensive Model [2016] - Yafra Khan Et al.

This paper proposes the artificial neural network model to predict the quality of the water and also time series analysis is used. Data used for this model comes in the category of continuous-time time series, as it consists of the values of water quality factors observed with the time-interval of 6 minutes. ANN is used to interpret non-linear relationships of the data, the time series model used in this paper is Non-linear Autoregressive (NAR) model. After training the model, the evaluation parameters of Regression(R), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) have been calculated. The graphs for regression analysis demonstrate how well the training, testing, and validation sets of data match the function. The better the function fits, and as a result, the closer the value of Regression is to 1, the more accurate the prediction.

6. Water Quality Prediction Using Artificial Intelligence Algorithms [2020] - Theyazn H. H Aldhyani Et al.

In this paper, advanced artificial intelligence (AI) algorithms are developed to predict water quality index (WQI) and water quality classification (WQC). For the WQI prediction, artificial neural network models, namely nonlinear autoregressive neural networks (NARNET) and long short-term memory (LSTM) deep learning algorithm, have been used.. In addition, three machine learning algorithms, namely, support vector machine (SVM), -nearest neighbour (K-NN), and Naive Bayes, have been used for the WQC forecasting. Prediction results show that the NARNET model performed slightly better than the LSTM for the prediction of the WQI values and the SVM algorithm has achieved the highest accuracy (97.01%) for the WQC prediction. Furthermore, the NARNET and LSTM models have achieved similar accuracy for the testing phase with a slight difference in the regression coefficient (RNARNET = 96.17% and LSTM = 94.21%).

7. Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast) [2021] - Saber Kouadri Et al.

The Directorate of Water Resources (DRE) of the State of Illizi provided the water analysis results that were used to construct this paper. The 114 samples from 57 exploited wells of 6 different layers that made up the submitted data set were examined. In order to produce WQI predictions in the Illizi region of southeast Algeria, 8 artificial intelligence algorithms, including multilinear regression (MLR), random forest (RF), M5P tree (M5P), random subspace (RSS), additive regression (AR), artificial neural network (ANN), support vector

regression (SVR), and locally weighted linear regression (LWLR), were used. Correlation coefficient (R), mean absolute error (MAE), root mean square error (RMSE), relative absolute error (RAE), and root relative square error were some of the statistical measures used to evaluate the models (RRSE). The findings show that the main factors affecting WQI in the study area are TDS and TH.

8. Machine learning methods for better water quality prediction.[2019] - AliNajah Ahmed Et al.

In this paper, Radial Basis Function Neural Networks (RBF-ANN), Multi-Layer Perceptron Neural Networks, and Adaptive Neuro-Fuzzy Inference System (ANFIS) are a few of the several modelling strategies that have been used (MLP-ANN). Artificial intelligence (AI) implementation produces a flexible mathematical structure with the ability to recognise complicated and non-linear correlations between input and output data. Three evaluation strategies or assessment processes have been utilized to evaluate the model. The WDT-ANFIS model performed better than all the other models and showed a notable improvement in forecasting accuracy for all the water quality parameter. After the suggested model was validated, it was discovered that it accurately predicted all of the water quality characteristics.

9. Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment [2020] - AliEl Bilali Et al.

Irrigation water quality (IWQ) parameters, including the sodium absorption ratio (SAR), adjusted SARa, exchangeable sodium percentage (ESP), and percentage of sodium, have been predicted using machine learning (ML) models including the 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 Boost. According to the findings of the generalisation attempt, the ML models for the Cherrate watershed's TDS, SAR, and SARa parameters as well as the Nfifikh watershed's TDS, chloride, ESP, and %Na parameter are fairly generalised. The findings of this study also show that machine learning models are effective tools for precisely forecasting the quality of irrigation water by only using the characteristics that can be measured directly in a short amount of time.

10. Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index [2020] - Sani Isah Abba Et al.

In this paper, the ensemble method was proposed in order to increase the performance accuracy of the single models like support vector regression (SVR) and one multilinear regression. Several statistical criteria were used to assess the models performance. The collected results demonstrated that the data intelligence models could successfully predict the WQI using the superior modelling output of the NNE (Neural network ensemble).

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 3657 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. Efficient Water Quality Prediction Using Supervised Machine Learning [2019] - Umair Ahmed Et al
2. Prediction of Water Classification using Machine Learning [2015] - Wahab Et al.
3. Machine learning algorithms for efficient water quality prediction [2021] - Mourade Et al.
4. A Review of the application of machine learning in water quality evaluation [2022] - Mengyuan Zhu Et al.
5. Predicting and Analysing Water Quality using Machine Learning: A Comprehensive Model [2016] - Yafra Khan Et al.
6. Water Quality Prediction Using Artificial Intelligence Algorithms [2020] - Theyazn H. H Aldhyani Et al.
7. Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast) [2021] - Saber Kouadri Et al.

8. Machine learning methods for better water quality prediction.[2019] - AliNajah Ahmed
9. Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment [2020] - AliEl Bilali Et al
10. Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index [2020] - Sani Isah Abba Et al.

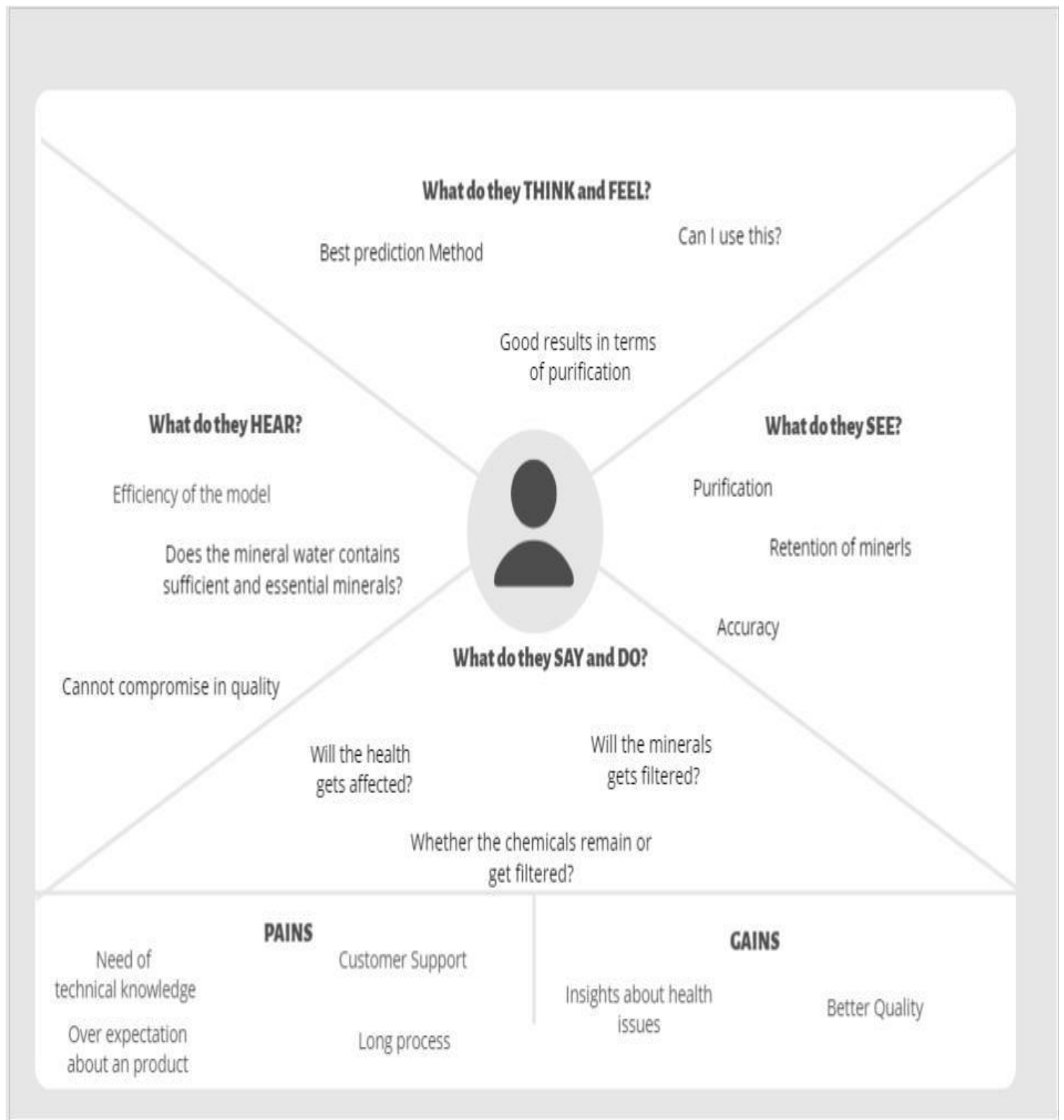
2.3 PROBLEM STATEMENT DEFINITION

Water makes up about 70% of the earth's surface and is one of the most vital sources. Everyday human activities and urbanization have led to an alarming deterioration of water quality, which in turn results in harmful diseases. Water quality has traditionally been estimated through statistical analysis, and it takes a long time to predict and estimate water quality. The alarming consequences of poor water quality require an alternative method that is faster and less expensive. The alarming consequences of poor water quality necessitate an alternative method that is faster and less expensive. For this reason, a set of machine learning algorithms has been developed to estimate water quality and describe the overall quality of the water.



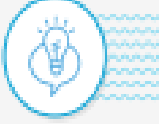
3. IDEATION AND PROPOSED SOLUTION

3.1 EMPATHY MAP CANVAS



3.2 IDEATION AND BRAINSTORMING

Temp to go



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
- 👥 4 hours to collaborate
- 👤 3.6 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

📄

Team gathering

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

🎯

Set the goal

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

📖

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

📝

How might we (your problem statement)?

🧠

Key rules of brainstorming

To run an smooth and productive session

🗣️ Stay on topic

💡 Encourage wild ideas

⏸️ Defer judgment

👂 Listen to others

🗳️ Go for volume

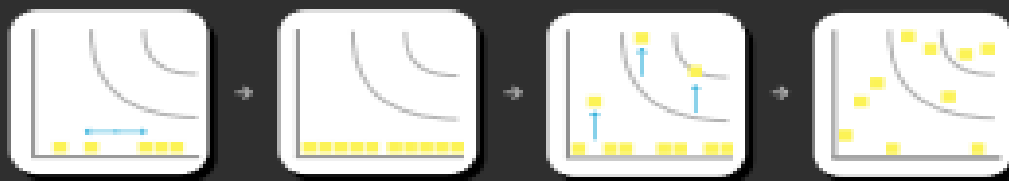
🖥️ If possible, be visual

4

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



3

Brainstorm

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

10 minutes

Tip

You can select a sticky note, unstick it, and then stick it to a new cluster.

Vijay Sree V S

Optimizing and ready availability of water by various parameters.

Optimizing and ready availability of water by various parameters.

Optimizing and ready availability of water by various parameters.

Dave Dharshini M

Optimizing and ready availability of water by various parameters.

Optimizing and ready availability of water by various parameters.

Optimizing and ready availability of water by various parameters.

Rishi Rajkumar

Optimizing and ready availability of water by various parameters.

Optimizing and ready availability of water by various parameters.

Optimizing and ready availability of water by various parameters.

Shreshth G

Optimizing and ready availability of water by various parameters.

Optimizing and ready availability of water by various parameters.

Optimizing and ready availability of water by various parameters.

Person 5

Optimizing and ready availability of water by various parameters.

Optimizing and ready availability of water by various parameters.

Optimizing and ready availability of water by various parameters.

Person 6

Optimizing and ready availability of water by various parameters.

Optimizing and ready availability of water by various parameters.

Optimizing and ready availability of water by various parameters.

Person 7

Optimizing and ready availability of water by various parameters.

Optimizing and ready availability of water by various parameters.

Optimizing and ready availability of water by various parameters.

Person 8

Optimizing and ready availability of water by various parameters.

Optimizing and ready availability of water by various parameters.

Optimizing and ready availability of water by various parameters.

4

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 can break it up into smaller sub-groups.

20 minutes

IDEAS

Optimizing and ready availability of water by various parameters.

The water quality should not be compromised.

Filtration to remove the impurities.

Oxygen level in water is the most important parameter.

Hardness of the water is measured.

The water should be in correct temperature.

The pH level of water should be between 6.5 and 8.5.

The calculation of water quality index to determine the quality of the water.

The amount of dissolved oxygen needs to be measured.

Turbidity.

3.3 PROPOSED SOLUTION

S. No	Parameter	Description
1.	Problem Statement (Problem to be solved)	Water is considered as a vital resource that affects various aspects of human health and lives. The quality of water is a major concern for people living in urban areas and the water is also most likely to become contaminated due to various factors including human, industrial and commercial activities as well as natural processes. In addition to that, poor sanitation infrastructure and lack of awareness also contributes Immensely to drinking water contamination.

2.	Idea / Solution description	It is not possible to check the quality of water manually every time. So an automatic real-time monitoring system is implemented based on machine learning technique to forecast the quality of water and to predict the health of Water according to its quality parameter level.
3.	Novelty / Uniqueness	<ul style="list-style-type: none"> • User Friendly • Determining the reuse and recycle of water • Detecting Quality parametric values.
4.	Social Impact / Customer Satisfaction	Customer satisfaction is an important factor to consider in total quality management. In order to achieve this goal, it is important that this project is used by all groups of people.
5.	Business Model (Revenue Model)	First the application is processed with real time data. Later it comes into the picture where everyone can see the Networking, conducting various activity and testing to them
6.	Scalability of the Solution	Helps in getting all required aspects regarding quality of water.

3.3 PROBLEM SOLUTION FIT

Define CS, fit into CC	<div>1. CUSTOMER SEGMENT(S)<div>CS</div></div> <div>Water is colorless substance required for the survival of most existing organisms and humans and consumed by all living creatures. Water is utilized for a variety of purposes, including drinking, agriculture, and industrial use.</div>	<div>6. CUSTOMER CONSTRAINTS<div>CC</div></div> <div>The purification of water should be done in the way that essential minerals are retained in the water. The water quality has influence on human health and environment.</div>	<div>5. AVAILABLE SOLUTIONS<div>AS</div></div> <div>The available solution is finding the water quality index (WQI) and water quality class(WQC).</div>	Explore AS, differentiate	
	<div>2. JOBS-TO-BE-DONE / PROBLEMS<div>J&P</div></div> <div>Now a days, due to urbanization water is getting contaminated. The contaminated water results in various waterborne diseases. The quality of water has a direct influence on both human health and the environment. Hence predicting and analyzing the water quality beforehand prevents many diseases.</div>	<div>9. PROBLEM ROOT CAUSE<div>RC</div></div> <div>Identify efficient and reliable solution. Collect sufficient amount of data. Identify the associated causal factor.</div>	<div>7. BEHAVIOUR<div>BE</div></div> <div>Water quality analyst predict the water quality patterns and it is very significant to include a temporal dimension to the analysis, so that the seasonal variation of water quality is addressed. They develop methods and identify the product which produces impurities.</div>		Focus on J&P, tap into BE, understand RC
	<div>3. TRIGGERS<div>TR</div></div> <div>This triggers to discover the pattern in user data and then make prediction based on intricate pattern for analyzing the quality of water. It also helps to improve the efficiency and more protected to drink.</div>	<div>10. YOUR SOLUTION<div>SL</div></div> <div>To predict the water quality, various supervised machine learning algorithms are employed. The models like linear regression, random forest classifier and support vector regression can be used. Hence supervised learning models are trained and developed to predict the quality of water.</div>	<div>8. CHANNELS of BEHAVIOUR<div>CH</div></div> <div>8.1 ONLINE Helps to notify the data preprocessing information.</div> <div>8.2 OFFLINE By attaining the standard quality of satisfy all parameterit is consider as pure water.</div>		
<div>4. EMOTIONS: BEFORE / AFTER<div>EM</div></div> <div>Before there is no efficient method to analyze the parameters. The statistical methods are time consuming. Now there is water purification systems because of those systems, diseases are prevented.</div>					
Identify strong TR & EM					

4. REQUIREMENT ANALYSIS

4.1 FUNCTIONAL REQUIREMENT

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User registration	Registration through Gmail Create an account Follow the instructions
FR-2	User Confirmation	Confirmation via Email and it is predicted by water level sensor
FR-3	Interface sensor	Interface sensor and Water level sensor produces the detection of clean drinking water
FR-4	Accessing datasets	Datasets are collected by data preprocessing method.
FR-5	Mobile application	The efficient of water quality is analyzed, the mobile application is not used .

4.2 NON-FUNCTIONAL REQUIREMENT

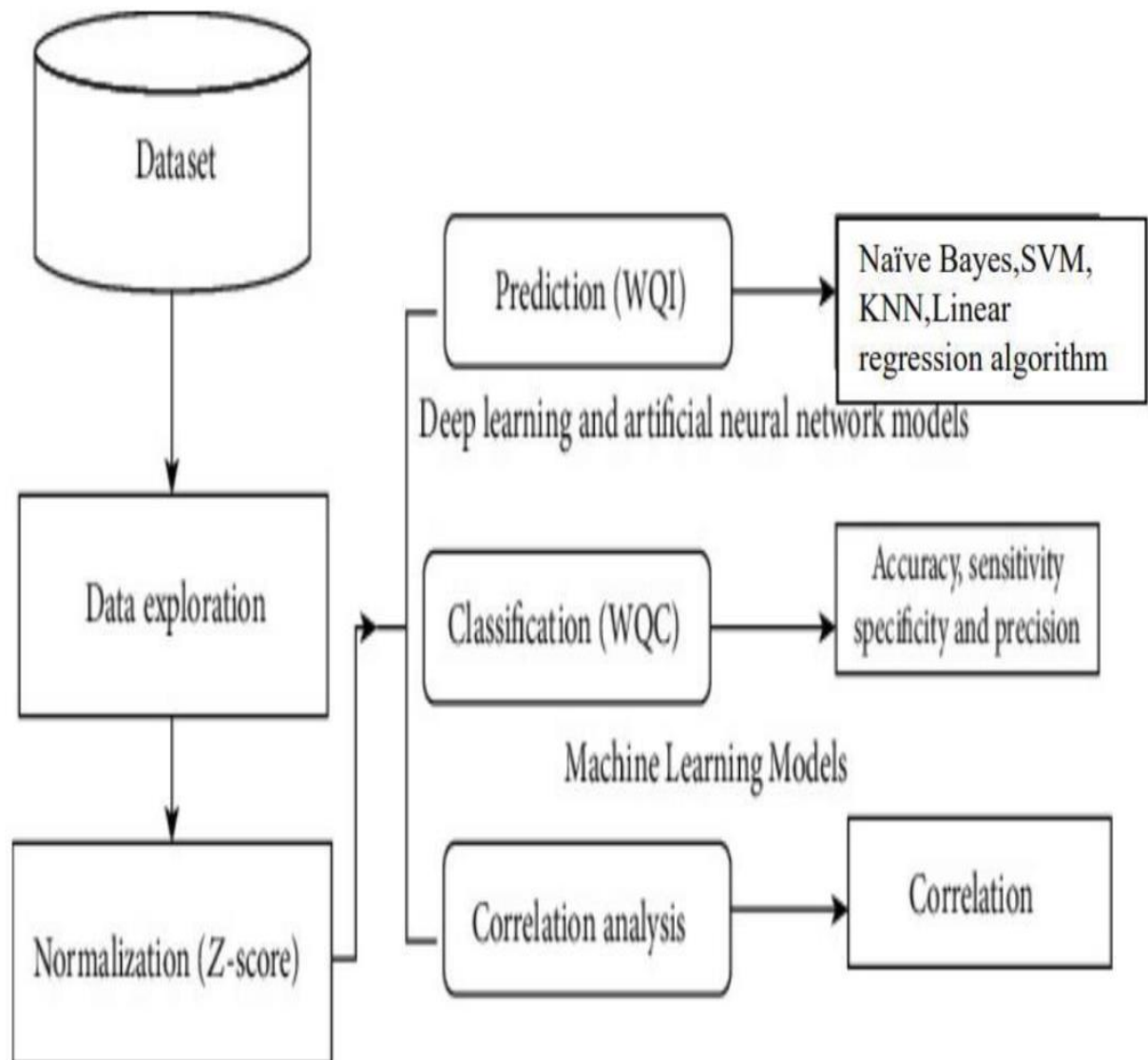
FR No.	Non-Functional Requirement	Description
NFR-1	Usability	The system provides a natural interaction with the users. Accurate water quality prediction with short time analysis and provide prediction safe to drink or not using some parameters and provide a great significance for water environment protection.
NFR-2	Security	The model enables with the high security system as the user's data will not be shared to the other sources. The system is protected with the user name and password throughout the process.
NFR-3	Reliability	The system is very reliable as it can last for long period of time when it is well maintained. The model can be extended in large scale by increasing the datasets.
NFR-4	Performance	Our system should run on 32 bit (x86) or 64 bit (x64) Dual-core 2.66-GHZ or faster processor. It should not exceed 2 GB RAM.
NFR-5	Availability	The system should be available for the duration of the user access the system until the user terminate the access. The system response to request of the user in less time and the recovery is done is less time.
NFR-6	Scalability	It provides an efficient outcome and has the ability to increase or decrease the performance of the system based on the datasets.

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.

DATAFLOW DIAGRAM:



5.2 SOLUTION AND TECHNICAL ARCHITECTURE

There are basically 8 steps for making our model predict the water quality of the water samples. Those steps are:-

1. 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.

2. 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.

3. Data Exploration:-

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

4. 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..

5. 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

6. 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.

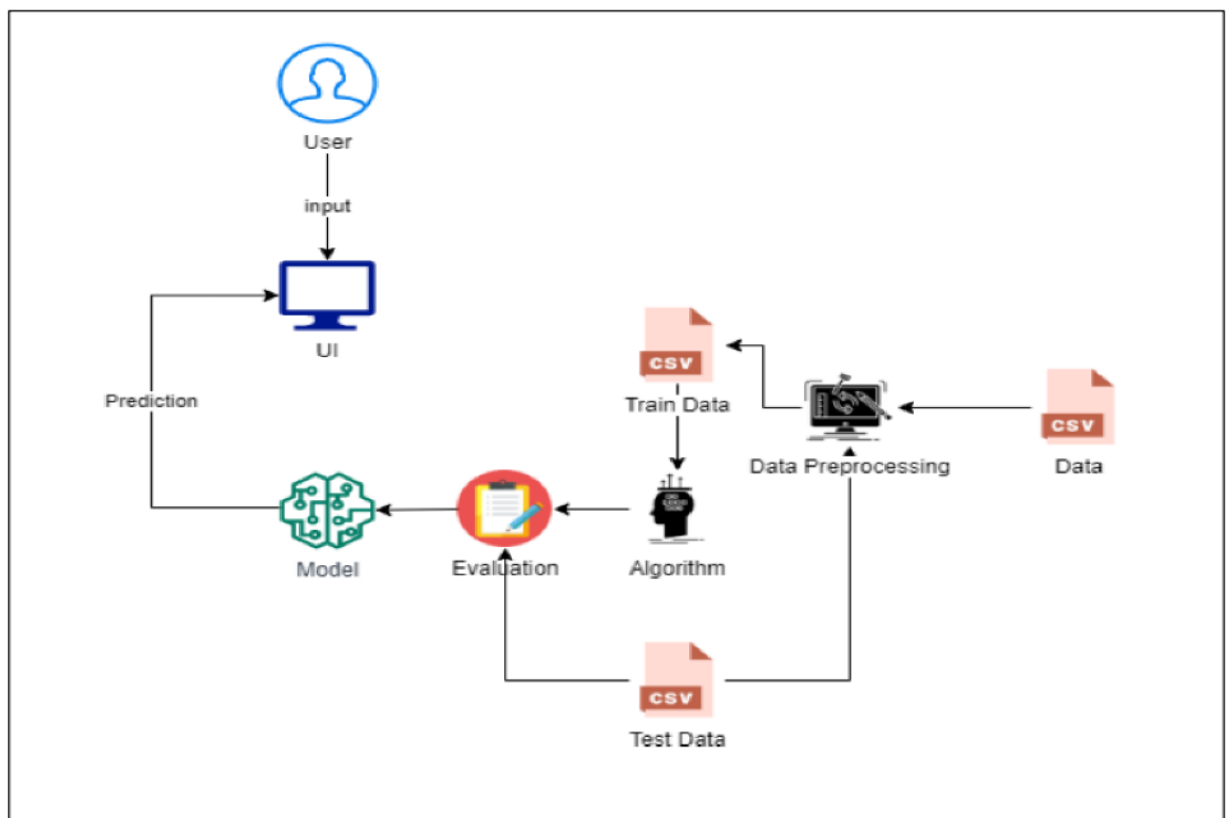
7. Data Modelling

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

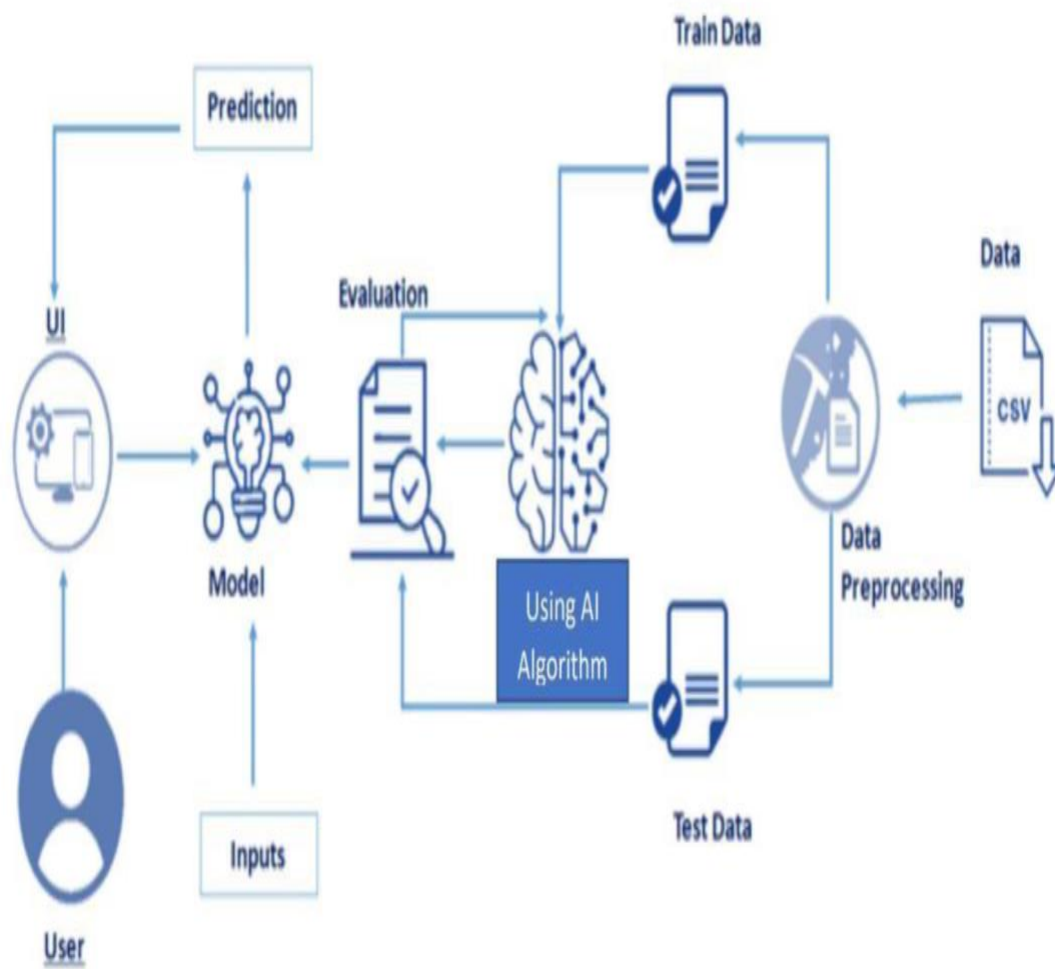
8. Model Evaluation

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:



Technical Architecture:



5.3 USER STORIES

 Scenario Browsing, booking, attending, and rating a local city tour	 Entice How does someone initially become aware of this process?	 Enter What do people experience as they begin the process?	 Engage In the core moments in the process, what happens?	 Exit What do people typically experience as the process finishes?	 Extend What happens after the experience is over?
 Steps What does the person (or group) typically experience?	User should know how the process works. User can use the provided manual to understand the process.	User should login to analyse the water quality. User should assess the terms and conditions of the process.	User should provide the parameters that are important to find the quality of water. User will be redirected to the result page.	The result will be displayed to the user. User can check whether the water is safe to drink or not. User can monitor the water quality again also.	User can provide a feedback or ratings about process. User can recommend others also.
 Interactions What Interactions do they have at each step along the way? <ul style="list-style-type: none"> People: Who do they see or talk to? Places: Where are they? Things: What digital touchpoints or physical objects would they use? 	Communication with experienced user.	Communication with service provider.	Interact with system for providing inputs.	Interaction through Email.	By using web application.
 Goals & motivations At each step, what is a person's primary goal or motivation? ("Help me..." or "Help me avoid...")	Help me to understand how the process works. Explain each and every step thoroughly.	Help me to analyse the quality of water.	Help me to provide the input successfully.	To provide the accurate measurement of water quality. To check whether the water is safe to drink or not.	To reduce the water borne diseases. To increase the efficiency.
 Positive moments What steps does a typical person find enjoyable, productive, fun, motivating, delightful, or exciting?	Safe to drink.	Simple and easy to analyse the water quality.	User feel relaxed about water quality.	Harmful substances are not present. Correct amount of substances are present.	Time consuming is less.
 Negative moments What steps does a typical person find frustrating, confusing, angering, costly, or time-consuming?	Provided measurements may not be accurate. Feeling insecure.	Feeling irritated when errors occur.	User feel frustrated if they don't get the expected results.	Datasets are difficult to collect.	Collecting the data from the water bodies can be expensive.
 Areas of opportunity How might we make each step better? What ideas do we have? What have others suggested?	Implement innovative ideas.	Identifying the common factors that affects the water quality and determining the best solution.	More number of water quality parameters should be analysed.	Automatic prediction and calculation.	Water quality analysis using WQI prediction.

6. PROJECT PLANNING AND SCDULING

6.1 SPRINT PLANNING AND ESTIMATION

Product Backlog, Sprint Schedule, and Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	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.	2	High	Deva Dharshini M
Sprint-1	User Confirmation	USN-2	As a user, I will receive confirmation email once I have registered for the application	1	Medium	Vidya Sree V S
Sprint-1	Login	USN-3	As a user, I can log into the application by entering email & password	2	High	Rithi Rajkumar Bharathi G
Sprint-2	Interface Sensor	USN-1	A sensor interface is a bridge between a device and any attached sensor. The interface takes data collected by the sensor and outputs it to the attached device	2	High	Rithi Rajkumar Bharathi G
Sprint-3	Coding (Accessing datasets)	USN-1	Coding is a set of instructions used to manipulate information so that a certain input results in a particular output.	2	High	Deva Dharshini M Vidya Sree V S Rithi Rajkumar Bharathi G
Sprint-4	Web Application	USN-1	As as user, I will show the current Information of the water quality.	1	Medium	Deva Dharshini M Vidya Sree V S

6.2 SPRINT SCHEDULE

Project Tracker, Velocity & Burn down Chart:

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	20	4 Days	24 Oct 2022	27 Oct 2022	20	29 Oct 2022
Sprint-2	20	5 Days	28 Oct 2022	01 Nov 2022	20	04 Nov 2022
Sprint-3	20	8 Days	02 Nov 2022	09 Nov 2022	20	11 Nov 2022
Sprint-4	20	9 Days	10 Nov 2022	18 Nov 2022	20	19 Nov 2022

Velocity:

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

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

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

Sprint 4: 1 user stories * 20 story points = 20

Total = 80

The average sprint velocity = $80/4 = 20$.

6.3 REPORTS FROM JIRA



7. CODING AND SOLUTIONS

7.1 FEATURE 1

Data collection and creation

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

In [2]: # importing libraries

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

In [3]: df = pd.read_csv('water_dataX.csv', encoding='ISO-8859-1', low_memory=False)

In [4]: df.head()

Out[4]:

	STATION CODE	LOCATIONS	STATE	Temp	D.O. (mg/l)	PH	CONDUCTIVITY (µmhos/cm)	B.O.D. (mg/l)	NITRATENAN N+ NITRITENANN (mg/l)	FECAL COLIFORM (MPN/100ml)	TOTAL COLIFORM (MPN/100ml)Mean	year
0	1393	DAMANGANGA AT D/S OF MADHUBAN, DAMAN	DAMAN & DIU	30.6	6.7	7.5	203	NAN	0.1	11	27	2014
1	1399	ZUARI AT D/S OF PT. WHERE KUMBARJRIA CANAL JOI...	GOA	29.8	5.7	7.2	189	2	0.2	4953	8391	2014
2	1475	ZUARI AT PANCHAWADI	GOA	29.5	6.3	6.9	179	1.7	0.1	3243	5330	2014
3	3181	RIVER ZUARI AT BORIM BRIDGE	GOA	29.7	5.8	6.9	64	3.8	0.5	5382	8443	2014
4	3182	RIVER ZUARI AT MARCAIM JETTY	GOA	29.5	5.8	7.3	83	1.9	0.4	3428	5500	2014

PROJECT REPORT

```
In [5]: df.dtypes
```

```
Out[5]: STATION CODE      object
LOCATIONS      object
STATE          object
Temp           object
D.O. (mg/l)    object
PH             object
CONDUCTIVITY (µmhos/cm)  object
B.O.D. (mg/l)  object
NITRATENAN N+ NITRITENANN (mg/l)  object
FECAL COLIFORM (MPN/100ml)  object
TOTAL COLIFORM (MPN/100ml)Mean  object
year           int64
dtype: object
```

```
In [6]: df.describe()
```

```
Out[6]:
```

	year
count	1991.000000
mean	2010.038172
std	3.057333
min	2003.000000
25%	2008.000000
50%	2011.000000
75%	2013.000000
max	2014.000000

```
In [7]: df.shape
```

```
Out[7]: (1991, 12)
```

```
In [8]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1991 entries, 0 to 1990
Data columns (total 12 columns):
STATION CODE      1991 non-null object
LOCATIONS         1991 non-null object
STATE             1991 non-null object
Temp              1991 non-null object
D.O. (mg/l)       1991 non-null object
PH                1991 non-null object
CONDUCTIVITY (µmhos/cm)  1991 non-null object
B.O.D. (mg/l)     1991 non-null object
NITRATENAN N+ NITRITENANN (mg/l)  1991 non-null object
FECAL COLIFORM (MPN/100ml)  1991 non-null object
TOTAL COLIFORM (MPN/100ml)Mean  1991 non-null object
year              1991 non-null int64
dtypes: int64(1), object(11)
memory usage: 101.2+ KB
```

Handling the Missing Data

```
In [9]: df.isnull().any()
```

```
Out[9]: STATION CODE      False
LOCATIONS      False
STATE          False
Temp           False
D.O. (mg/l)    False
PH             False
CONDUCTIVITY (µmhos/cm)  False
B.O.D. (mg/l)  False
NITRATENAN N+ NITRITENANN (mg/l)  False
FECAL COLIFORM (MPN/100ml)  False
TOTAL COLIFORM (MPN/100ml)Mean  False
year           False
dtype: bool
```

Data Pre-processing:

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.

Water Quality and 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 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 is the ideal value of parameter in pure water (0 for all parameters except 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.

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

PROJECT REPORT

In [17]: `df.head()`

Out[17]:

	station	location	state	Temp	do	ph	co	bod	na	tc	year
0	1393	DAMANGANGA AT D/S OF MADHUBAN, DAMAN	DAMAN & DIU	30.6	6.7	7.5	203.0	6.940049	0.1	27.0	2014
1	1399	ZUARI AT D/S OF PT. WHERE KUMBARJRIA CANAL JOI...	GOA	29.8	5.7	7.2	189.0	2.000000	0.2	8391.0	2014
2	1475	ZUARI AT PANCHAWADI	GOA	29.5	6.3	6.9	179.0	1.700000	0.1	5330.0	2014
3	3181	RIVER ZUARI AT BORIM BRIDGE	GOA	29.7	5.8	6.9	64.0	3.800000	0.5	8443.0	2014
4	3182	RIVER ZUARI AT MARCAIM JETTY	GOA	29.5	5.8	7.3	83.0	1.900000	0.4	5500.0	2014

Water Quality Index Calculation

In [18]: *#calculation of Ph*

```
df['npH']=df.ph.apply(lambda x:(100 if (8.5>=x>=7)
                                else(80 if (8.6>=x>=8.5) or (6.9>=x>=6.8)
                                else(60 if (8.8>=x>=8.6) or (6.8>=x>=6.7)
                                else(40 if (9>=x>=8.8) or (6.7>=x>=6.5)
                                else 0))))))
```

In [19]: *#calculation of dissolved oxygen*

```
df['ndo']=df.do.apply(lambda x:(100 if (x>=6)
                                else(80 if (6>=x>=5.1)
                                else(60 if (5>=x>=4.1)
                                else(40 if (4>=x>=3)
                                else 0))))))
```

In [20]: *#calculation of total coliform*

```
df['nco']=df.tc.apply(lambda x:(100 if (5>=x>=0)
                                else(80 if (50>=x>=5)
                                else(60 if (500>=x>=50)
                                else(40 if (10000>=x>=500)
                                else 0))))))
```

In [21]: *#calc of B.D.O*

```
df['nbdo']=df.bod.apply(lambda x:(100 if (3>=x>=0)
                                else(80 if (6>=x>=3)
                                else(60 if (80>=x>=6)
                                else(40 if (125>=x>=80)
                                else 0))))))
```

PROJECT REPORT

```
In [22]: #calculation of electrical conductivity
df['nec']=df.co.apply(lambda x:(100 if (75>-x>=0)
                        else(80 if (150>=x>=75)
                        else(60 if (225>=x>=150)
                        else(40 if (300>=x>=225)
                        else 0))))
```

```
In [23]: #Calculation of nitrate
df['nna']=df.na.apply(lambda x:(100 if (20>-x>=0)
                        else(80 if (50>-x>=20)
                        else(60 if (100>-x>=50)
                        else(40 if (200>=x>=100)
                        else 0))))
```

```
In [24]: df.head()
df.dtypes
```

```
Out[24]: station    object
location    object
state       object
Temp        float64
do           float64
ph           float64
co           float64
bod          float64
na           float64
tc           float64
year         int64
npH          int64
ndo          int64
nco          int64
nbdo         int64
nec          int64
nna          int64
dtype: object
```

```
In [25]: df['wph']=df.npH * 0.165
df['wdo']=df.ndo * 0.281
df['wbdo']=df.nbdo * 0.234
df['wec']=df.nec* 0.009
df['wna']=df.nna * 0.028
df['wco']=df.nco * 0.281
df['wqi']=df.wph+df.wdo+df.wbdo+df.wec+df.wna+df.wco
df
```



```

Random Forest Regression

In [41]: from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size = 0.2,random_state=10)

In [42]: #Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
x_train = sc.fit_transform(x_train)
x_test = sc.transform(x_test)

In [43]: from sklearn.ensemble import RandomForestRegressor
regressor = RandomForestRegressor(n_estimators = 10, random_state = 0)
regressor.fit(x_train, y_train)

F:\Users\DEVA DHARSHINI M\Anaconda3\lib\site-packages\ipykernel_launcher.py:3: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,), for example using ravel().
This is separate from the ipykernel package so we can avoid doing imports until

Out[43]: RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
                                max_features='auto', max_leaf_nodes=None,
                                min_impurity_decrease=0.0, min_impurity_split=None,
                                min_samples_leaf=1, min_samples_split=2,
                                min_weight_fraction_leaf=0.0, n_estimators=10,
                                n_jobs=None, oob_score=False, random_state=0, verbose=0,
                                warm_start=False)

In [44]: y_pred = regressor.predict(x_test)

Model Evaluation

In [45]: from sklearn import metrics
print('MAE:',metrics.mean_absolute_error(y_test,y_pred))
print('MSE:',metrics.mean_squared_error(y_test,y_pred))
print('RMSE:',np.sqrt(metrics.mean_squared_error(y_test,y_pred)))

MAE: 1.0048521303258215
MSE: 5.728742416040103
RMSE: 2.3934791446845955

In [46]: #accuracy of the model
metrics.r2_score(y_test, y_pred)

Out[46]: 0.9687726342666486

```

The model can be extended to function in the simulation of other machine learning problems. It uses the hyper plane to separate the points of the input vectors and finds the needed coefficients. The best hyper plane is the line with the largest margin, which is meant the distance between the hyper plane and the nearest input objects. The input points defined in the hyper plane 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

8.2 USER ACCEPTANCE TESTING

1. Purpose of Document

The purpose of this document is to briefly explain the test coverage and open issues of the [Product Name] 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

Therefore, Random Forest Regressor is used.

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

RANDOM FOREST REGRESSOR

Accuracy =96.8

Mean Absolute error = 1.00485

Mean squared error = 5.728

10. ADVANTAGES

1. 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.

2. 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.

3. 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

1. Training necessary somewhat difficult to manage over time and with large data sets

2. Requires manual operation to submit data, some configuration required

3. Costly, usually only feasible under Exchange Network grants Technical expertise and network server required

4. 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

Portability 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 diarrhoea. In this case, using a prescriptive analysis based on projected values would result in future capabilities to assist decision and policy makers.

12. SOURCE CODE

Machine learning has been widely used as a powerful tool to solve problems in the water environment because it can be applied to predict water quality, optimize water resource allocation, manage water resource shortages, etc. Despite this, several challenges remain in fully applying machine learning approaches in this field to evaluate water quality: 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. 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. 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: 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. 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. Interdisciplinary talent with knowledge in different fields should be trained to develop more advanced machine learning techniques and apply them in engineering practices.

