PROJECT REPORT

EFFICIENT WATER QUALITY ANALYSIS

AND PREDICTION USING MACHINE LEARNING

By Team – PNT2022TMID38737

Batch No – B3-3M5E

CK College of Engineering and Technology

Ajai B - 420719104003
Karan L - 420719104015
Harish T - 420719104013
Prem Kumar B - 420719104026

Under the guidance of,

Mr. Gopalakrishnan S (Mentor)

Ms. Lalitha Gayathri (Industry Mentor)

TABLE OF CONTENTS

S.no	Title	Pg no
1.	Introduction	3
1.a	Project overview	5
1.b	purpose	7
2.	Literature survey	9
2.1	Existing problems	11
2.2	Reference	13
2.3	Problem statement and definition	15
3.	IDEATION & PROPOSED SOLUTION	17
3.1	Empathy Map Canvas	18
3.2	Ideation & Brainstorming	19
3.3	Proposed Solution	21
3.4	Problem Solution fit	23
4.	REQUIREMENT ANALYSIS	25
4.1	Functional requirement	26
4.2	Non-Functional requirements	27
5.	PROJECT DESIGN	28
5.1	Data Flow Diagrams	39
5.2	Solution & Technical Architecture	30
5.3	User Stories	31
6.	PROJECT PLANNING & SCHEDULING	32
6.1	Sprint Planning & Estimation	34
6.2	Sprint Delivery Schedule	35
6.3	Reports from JIRA	36
7.	CODING & SOLUTIONING (Explain the features added in the project along with code)	37
7.1	Feature 1	38
7.2	Feature 2	39
7.3	Database Schema (if Applicable)	40
8.	TESTING	41

8.1	Test Cases	42
8.	User Acceptance Testing	43
9.	RESULTS	44
9.1.	Performance Metrics	44
10.	ADVANTAGES & DISADVANTAGES	45
11.	CONCLUSION	46
12.	FUTURE SCOPE	47
13.	APPENDIX	48

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 [13][14], Artificial Neural Network [2][5][6][7], Support Vector Machine [8][9][0], K-Nearest Neighbors [21] and Naïve Bayes [18][22][23]. However, in recent years, some researchers are moving towards more advanced ML ensemble models, such as Gradient Boosting and Random Forest [1]

Traditional Machine Learning models, such as the Decision Tree model, are frequently found in the literature and performed well on water quality data. However, decision-tree-based ensemble models, including Random Forest (RF) and Gradient Boosting (GB), always outperform the single decision tree [4]. Among the reasons for this are its ability to manage both regular attributes and data, not being sensitive to missing values and being highly efficient. Compared to other ML models, decision-tree-based models are more favorable to short-term prediction and may have a quicker calculation speed [6]. Gakii and Jepkoech [3] compared five different decision tree classifiers, which are Logistic Model Tree (LMT), J48, 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. [4] also compared five decision-tree-based models, which are Random Tree, Random Forest, Ordinary Decision Tree (ODT), Chi-square Automatic Interaction Detector and Iterative 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 [8] 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 [5] used ANN and multivariate linear regression models in his research and found that the ANN model outperforms the MLR model. However, the research only assessed the performance of the ANN model using root-mean-square error (RMSE), coefficient of correlation (r) and bias values. Although ANN models are the most broadly used, they have a drawback as the prediction power becomes weak if they are used with a small dataset and the testing data are outside the range of the training data [8].

Support Vector Machine has also been extensively used in water quality studies.

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

The ensemble method is a Machine Learning technique that combines several base learners' decisions to produce a more precise prediction than what can be achieved with having each base learner's decision [6]. This method has also gained wide attention among researchers recently. The diversity and accuracy of each base learner are two important features to make the ensemble learners work properly [7]. The ensemble method ensures the two features in several ways based on its working principle. There are two commonly used ensemble families in Machine Learning, which are bagging and boosting. Both the bagging and boosting
methods provide a higher stability to the classifiers and are good in reducing variance. Boosting can reduce the bias, while bagging can solve the overfitting problem [1]. A famous ensemble model that uses the bagging algorithm is Random Forest. It is a classification model that uses multiple base models, typically decision trees, on a given subset of data independently and makes decisions based on all models [5]. It uses feature randomness and bagging when building each individual

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

2.1 EXISITNG PROBLEM

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

2.2 REFERENCES

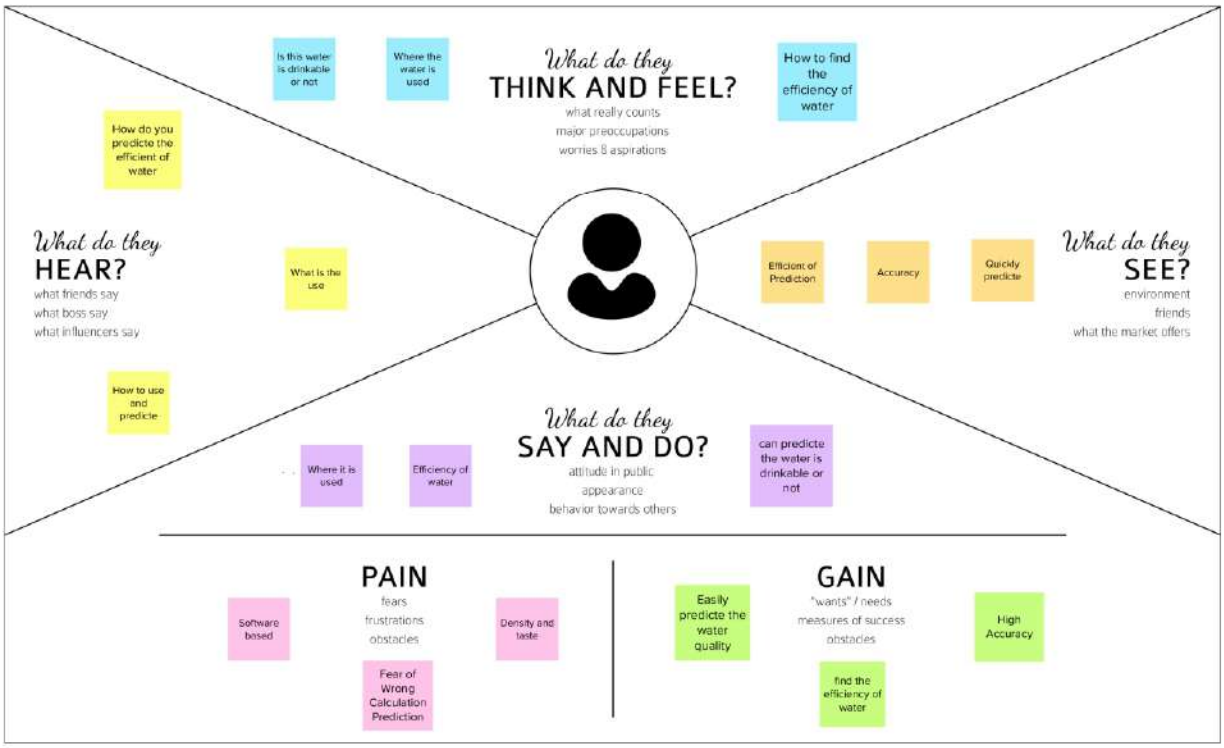
1. Ling, J.K.B. Water Quality Study and Its Relationship with High Tide and L ow Tide at Kuantan River. Bachelor's Thesis, Universiti Malaysia Pahang, Gambang, Malaysia, 2010. Available online: http://umpir.ump.edu.my/id/ep rint/2449/1/JACKY_LING_KUO_BAO.PDF (accessed on 22 February 202 2).
2. Xu, J.; 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 Mortalit y 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 neu ral networks in prediction of water quality parameters. J. Environ. Health Sc i. 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, 1 787–1798.
7. Al-Badaii, F.; Shuhaimi-Othman, M.; Gasim, M.B. Water quality assessme nt 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 qu ality index prediction and uncertainty analysis: A comparative study of mac hine learning models. J. Environ. Chem. Eng. 2021, 9, 104599.
9. Chen, K.; Chen, H.; Zhou, C.; Huang, Y.; Qi, X.; Shen, R.; Liu, F.; Zuo, M. ; Zou, X.; Wang, J. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different mac hine learning models based on big data. Water Res. 2020, 171, 115454.
10. Lerios, 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 i ntelligence techniques: SOM and ANN—A case study of Melen River Turk ey. Water Qual. Expo. Health 2015, 7, 469–490.
12. Aradhana, G.; Singh, N.B. Comparison of Artificial Neural Network algorit hm for water quality prediction of River Ganga. Environ. Res. J. 2014, 8, 55 –63.

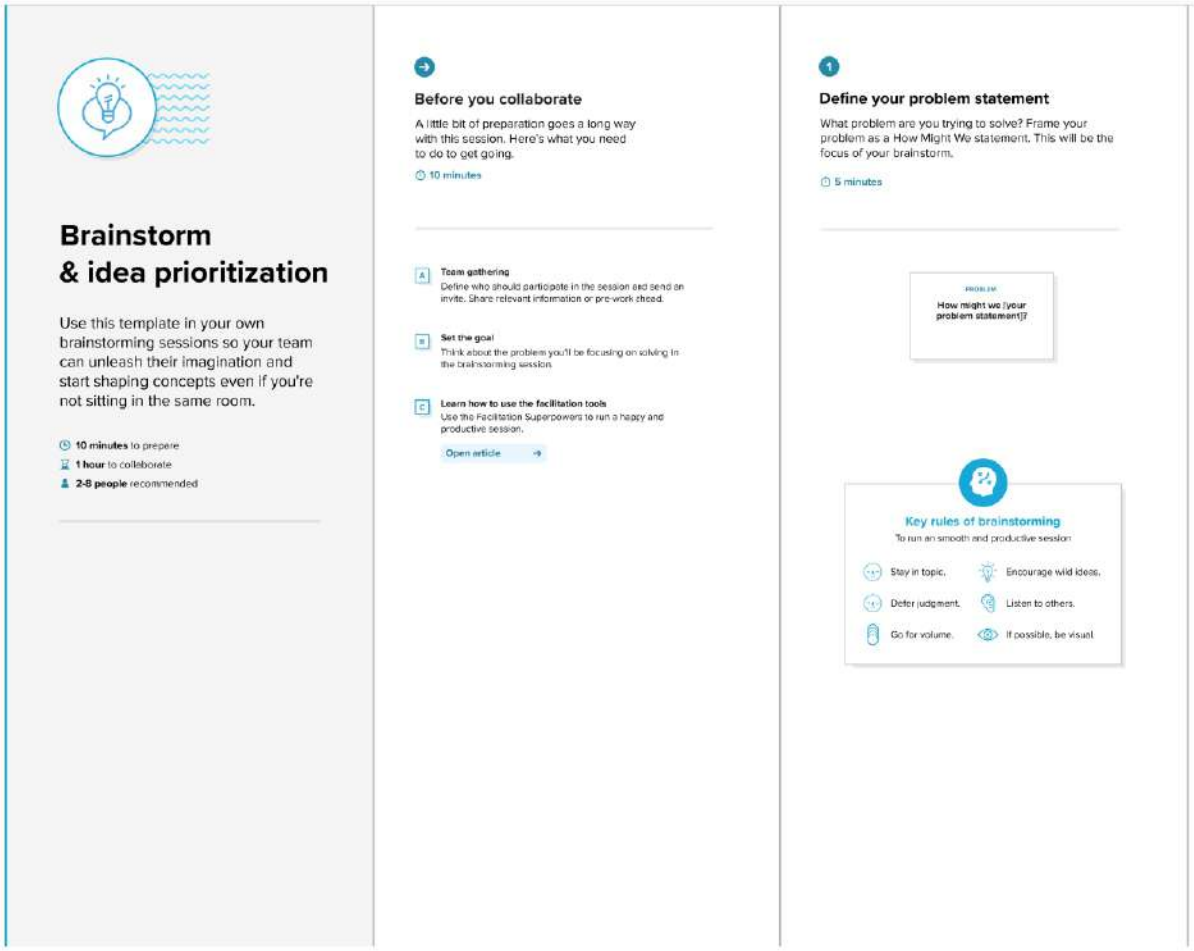
2.3 PROBLEM STATEMENT DEFINITION

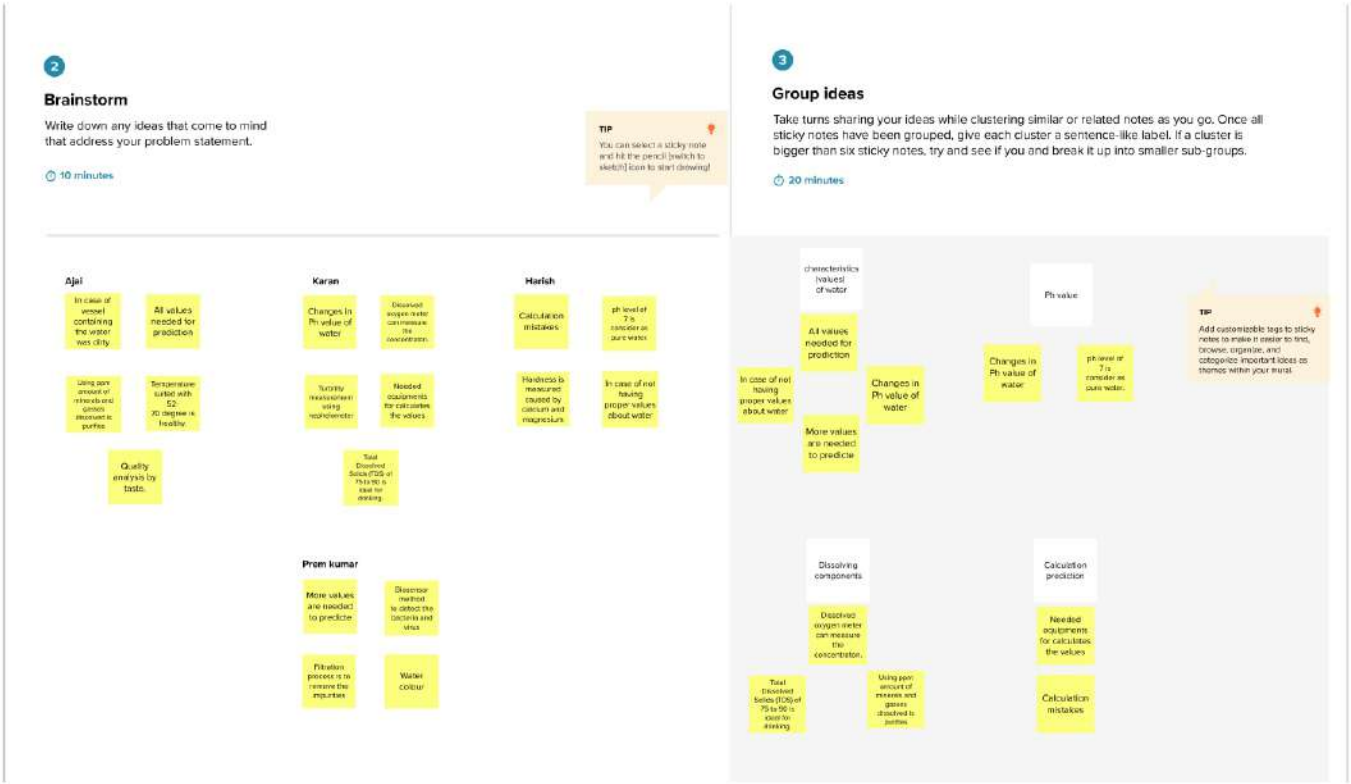
To predict the water safe or not for Access to safe drinking-water is essential to he alth, a basic human right and a component of effective policy for health protection . This is important as a health and development issue at a national, regional and lo cal level. In some regions, it has been shown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

Empathy map:


What do they
THINK AND FEEL?
what really counts
major preoccupations
worries & aspirations
Is this water is drinkable or not
Where the water is used
How to find the efficiency of water
How do you predicte the efficient of water
What do they
HEAR?
what friends say
what boss say
what influencers say
What is the use
How to use and predicte
What do they
SEE?
environment
friends
what the market offers
Efficient of Prediction
Accuracy
Quickly predicte
What do they
SAY AND DO?
attitude in public
appearance
behavior towards others
Where it is used
Efficiency of water
can predicte the water is drinkable or not
PAIN
fears
frustrations
obstacles
Software based
Density and taste
Fear of Wrong Calculation Prediction
GAIN
"wants" / needs
measures of success
obstacles
Easily predicte the water quality
High Accuracy
find the efficiency of water

Brainstorm & Idea Prioritization:



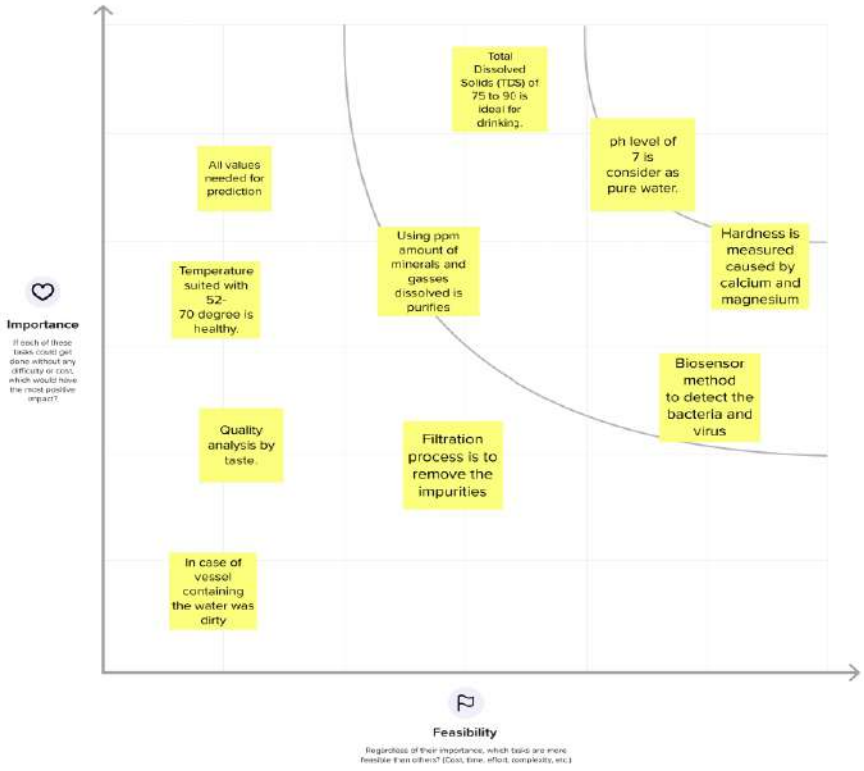


4

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



Importance

If each of these tasks could get done without any difficulty or cost, which would have the most positive impact?

Feasibility

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

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	Water is seen as a major resource that has an impact on many elements of human health and survival. People who live in metropolitan areas are often concerned about the quality of the water.
2.	Idea / Solution description	This project aims at building a machine learning model to predict a water quality by considering all water quality standard indicators.
3.	Novelty / Uniqueness	The proposed method is utilized to test portability. It has two phases: testing and training. working on past historical data.
4.	Social Impact / Customer Satisfaction	The quality of water services as a powerful environmental determinant and a foundation for the prevention and control of water borne diseases.
5.	Business Model (Revenue Model)	This model should be licensed by the machine learning as well as data analytics and make more impression among the people.
6.	Scalability of the Solution	A system that scales well will be able to maintain or increase its level of performance even as it is tested by larger than its operational demands.

1.CUSTOMER SEGMENTS There are various categories of customer of high-quality water in the public, private, and government sectors.	**6.CUSTOMER CONSTRAINS** Customers can use a web application to analyse the water quality by simply providing a few water characteristic data, but they do need some fundamental prerequisites, such as a network connection, a system or a mobile.	**5.AVAILABLE SOLUTIONS** Have information on the water's colour, odour, pH level, and other characteristics to evaluate if it is safe to drink or not. **PROS** Solution within a second **CONS** Accuracy is not 100 %
2.JOBS TO BE DONE/PROBLEM Gather the historical information about the quality of the water based on its many features and qualities in the chemical and physical compositions of nature.	**9.PROBLEM ROOT CAUSE** All living things are harmed by improper maintenance of rainwater and surface water from rivers that are combined with industrial waste and certain other human-generated contaminants.	**7.BEHAVIOUR** Customers must have up-to-date information about the water's status in order for machine learning models to accurately anticipate whether the water is excellent or harmful. Basic knowledge of water characteristics and web usage for the easy way to solution.
3.TRIGGERS General information about the water by using sensors and give those values to the application will give all the details of water quality.	**10.YOUR SOLUTION** **Simply entering the current water data to the web app which gives the analysis of water prediction.** **Using past historical data of water to predict and analyse the water in current scenarios.**	**8.CHANNELS OF BEHAVIOUR** **ONLINE** Customer can use the web app by simply entering the current URL of the website.
4.EMOTIONS:BEFORE/AFTER Without prior knowledge of water quality and drinking it leads to be causing various diseases and loss of life.		

Solution Requirements (Functional & Non-functional):

Functional Requirements:

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	Enter the Inputs	Input get through the form and Check the input data
FR-2	User Essential	User needs the result with correct accuracy
FR-3	Data Preprocessing	Get the tested and trained data from the raw dataset
FR-4	User input Evaluation	Evaluate the input with model and predicte the efficiency of water
FR-5	Prediction	Efficient for use or not

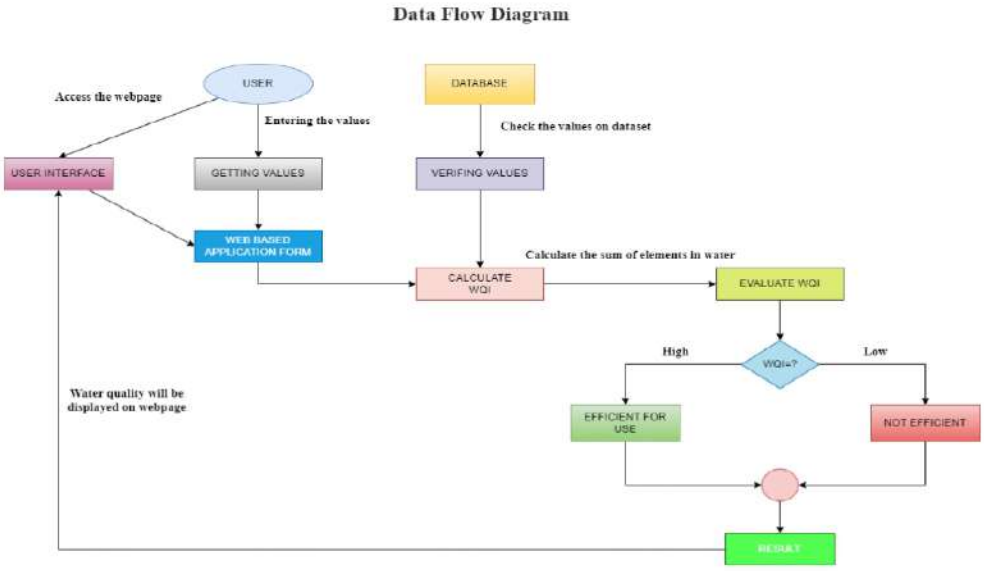
Non-functional Requirements:

Following are the non-functional requirements of the proposed solution.

FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	User friendly web application
NFR-2	**Security**	Website doesn't have any harmful virus and it didn't ask any permission
NFR-3	**Reliability**	Many type of water values are trained to model so prediction accuracy high
NFR-4	**Performance**	Quickly get the results
NFR-5	**Availability**	Available at any time in internet
NFR-6	**Scalability**	• Can access the website through computer and mobile • Lightweight web application

Data Flow Diagram &User Stories:

Data Flow Diagram:



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 (Web user)	Enter the values	USN-1	As a user , I can enter the values about the water	I can access the webpage through internet	High	Sprint-1
		USN-2	As a user, I can submit the values into the webpage	I can click the submit button	High	Sprint-1
		USN-3	As a user, I except correct efficient of The water		Medium	Sprint-2
	Data Preprocessing	USN-4	As a user, I can see the loading animation		Medium	Sprint-2
	User input Evaluation	USN-5	I can see the Evaluation quicly		High	Sprint-3
	Prediction	USN-6	As a user, I can see the result of the water efficient	It result visible on webpage	High	Sprint-3

Technology Stack (Architecture & Stack):

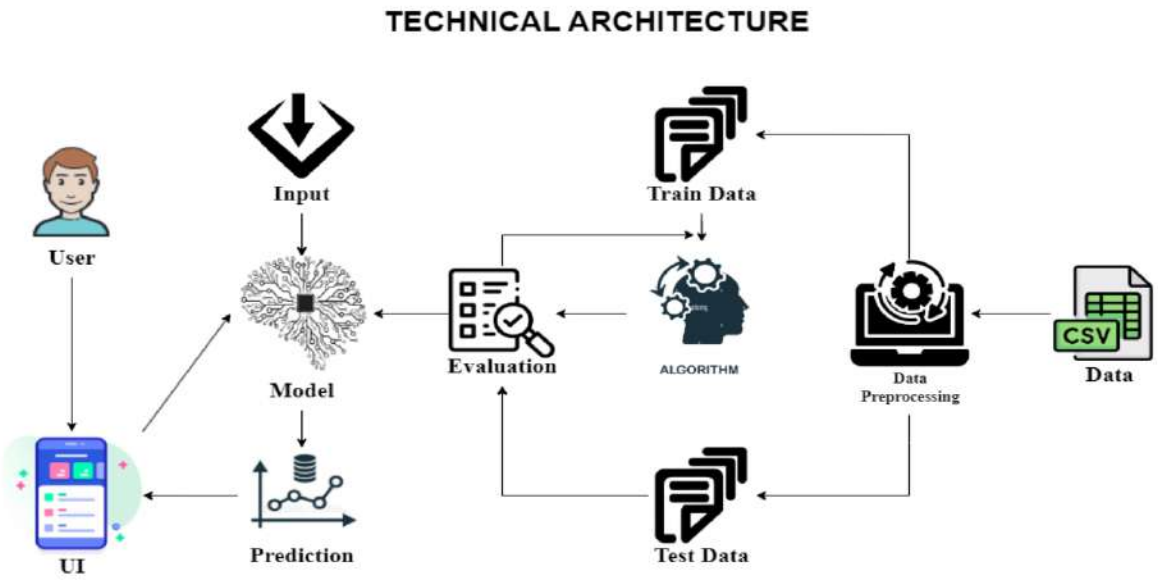


Table-1 : Components & Technologies:

S.No	Component	Description	Technology
1.	User Interface	Website UI	HTML, CSS, JavaScript / Angular Js / React Js etc.
2.	Application Logic-1	Logic for a process in the application	Python
3.	Application Logic-2	Logic for a process in the application	IBM Watson STT service
4.	File Storage	File storage requirements	IBM Block Storage or Other Storage Service or Local Filesystem
5.	External API-1	Purpose of External API used in the application	IBM cloud API
6.	Machine Learning Model	Purpose of Machine Learning Model	Water Quality Index Model
7.	Infrastructure (Server / Cloud)	Application Deployment on Cloud	Cloud Foundry

Table-2: Application Characteristics:

S.No	Characteristics	Description	Technology
1.	Open-Source Frameworks	List the open-source frameworks used	Angular.JS
2.	Security Implementations	Keep Software And Plugins Up-To-Date.	IAM Controls
3.	Scalable Architecture	Use of Webpage	js and Angular
4.	Availability	Use at any time	Webpage(HTML, CSS and JavaScript)
5.	Performance	Quickly display the result	Content Delivery Network

Project Planning :

Product Backlog, Sprint Schedule, and Estimation:

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Data Preparation	USN-1	Collecting water dataset and pre-processing it	20	High	Ajai B
Sprint-2	Model Building	USN-2	Create a ML model to predict water quality	10	Medium	Ajai B Karan L Harish T Prem Kumar B
Sprint-2	Model Evaluation	USN-3	Calculate the performance, error rate and complexity of ML model	10	Medium	
Sprint-3	Model Deployment	USN-4	Using flask and deploy model finally in IBM cloud using IBM storage and Watson Studio	10	Medium	
Sprint-3	Web page (Form)	USN-5	As a user, I can use the application by enter the water data in form	10	Medium	Ajai B Karan L Harish T Prem Kumar B
Sprint-4	Dashboard	USN-6	As a user, I can predict the water quality by click the submit button and the application will show the water is efficient for use or not	20	High	Ajai B

Project Tracker & Velocity:

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

Velocity:

Average Velocity = 80 / 20 = 4 Story Points per Day

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.

Water Quality Index Calculation

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

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

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

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

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

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

7.2 FEATURE 2

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

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

```
# 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
```

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

```
0.62125
```

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

```
              precision    recall  f1-score   support

           0       0.67      0.77      0.72       497
           1       0.50      0.38      0.43       303

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

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)
```

```
0.6225
```

```
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.

8.2 USER ACCEPTANCE TESTING

1. Purpose of Document

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

2. Defect Analysis

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

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

3. Test Case Analysis

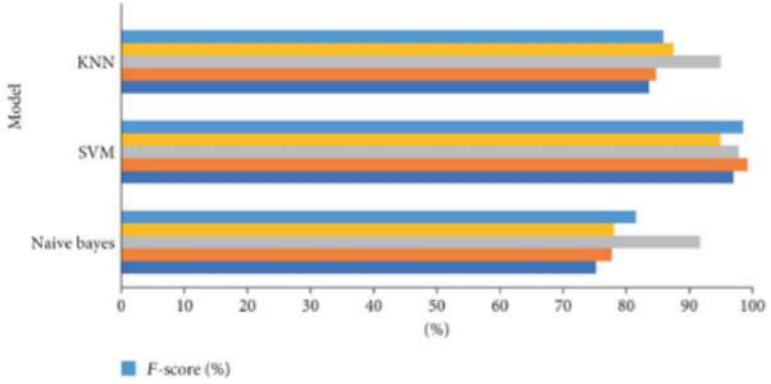
This report shows the number of test cases that have passed, failed, and untested

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

9.RESULT

9.1 PERFORMANCE METRICS

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



SO ,WE ARE GOING TO USE SVC

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

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

Table 1. Comparison of algorithms SN.

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

10. ADVANTAGES

Whether it be for groundwater, surface water or open water, there are a number of reasons why it is important for you to undertake regular water quality testing. If you're wanting to create a solid foundation on which to build a broader water management plan, then investing in water quality testing should be your first point of action. This testing will also allow you to adhere to strict permit regulations and be in compliance with Australian laws.
Identifying the health of your water will help you to discover where it may need some help. Ultimately, finding a source of pollution, or remaining proactive with your monitoring will enable you to save money in the long term. The more information that you can obtain will assist you with your decision on what product you may need to improve the condition of your water. Simply guessing and buying products based on a hunch or a general trend is ill-advised, as each body of water has unique properties that can only be discovered through testing.
Measuring the amount of dissolved oxygen in your water is another important advantage of water quality testing, as typically the less oxygen, the higher the water temperature, resulting in a more harmful environment for aquatic life. These levels do fluctuate slightly across the seasons, but regular monitoring of your water quality will allow you to discover trends over time, and whether there are other factors that may be contributing to the results you discover.

DISADVANTAGES

Training necessary Somewhat difficult to manage over time and with large data sets
Requires manual operation to submit data, some configuration required
Costly, usually only feasible under Exchange Network grants Technical expertise and network server required
Requires manual operation to submit data Cannot respond to data queries from other nodes, and therefore cannot interact with the Exchange Network Technical expertise and network server required

11. CONCLUSION

Potability determines the quality of water, which is one of the most important resources for existence. Traditionally, testing water quality required an expensive and time-consuming lab analysis. This study looked into an alternative machine learning method for predicting water quality using only a few simple water quality criteria. To estimate, a set of representative supervised machine learning algorithms was used. It would detect water of bad quality before it was released for consumption and notify the appropriate authorities It will hopefully reduce the number of individuals who drink low-quality water, lowering the risk of diseases like

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

12.SOURCE CODE

Machine learning has been widely used as a powerful tool to solve problems in the water environment because it can be applied to predict water quality, optimize water resource allocation, manage water resource shortages, etc. Despite this, several challenges remain in fully applying machine learning approaches in this field to evaluate water quality: (1) Machine learning is usually dependent on large amounts of high-quality data. Obtaining sufficient data with high accuracy in water treatment and management systems is often difficult owing to the cost or technology limitations. (2) As the conditions in real water treatment and management systems can be extremely complex, the current algorithms may only be applied to specific systems, which hinders the wide application of machine learning approaches. (3) The implementation of machine learning algorithms in practical applications requires researchers to have certain professional background knowledge.
To overcome the above-mentioned challenges, the following aspects should be considered in future research and engineering practices: (1) More advanced sensors, including soft sensors, should be developed and applied in water quality monitoring to collect sufficiently accurate data to facilitate the application of machine learning approaches. (2) The feasibility and reliability of the algorithms should be improved, and more universal algorithms and models should be developed according to the water treatment and management requirements. (3) Interdisciplinary talent with knowledge in different fields should be trained to develop more advanced machine learning techniques and apply them in engineering practices.

13. APPENDIX

REQUIREMENT.TXT

Flask == 2.2.2

joblib == 1.2.0

numpy == 1.23.4

pandas == 1.5.1

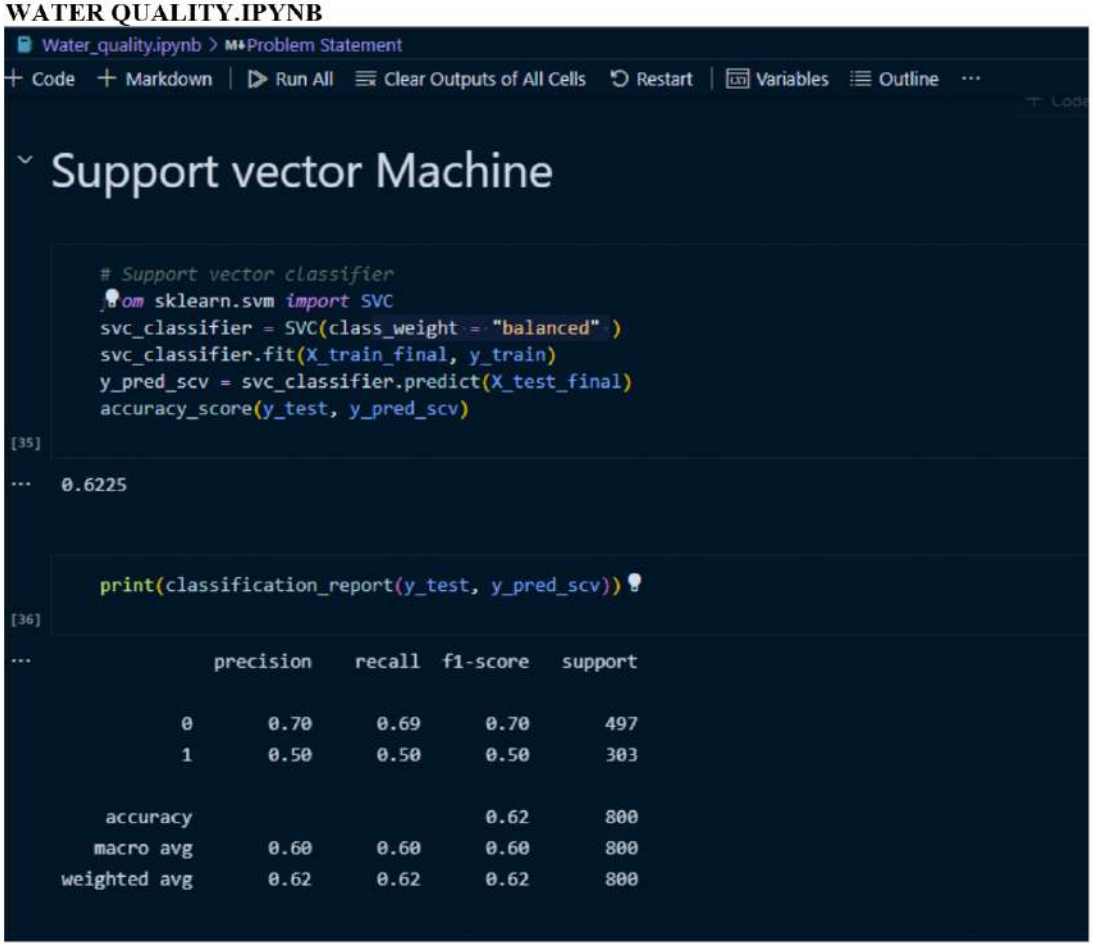
```
scikit-learn   ==  1.1.3
xgboost         == 1.7.1
gunicorn        == 20.1.0
matplotlib      == 3.6.2
seaborn         == 0.12.1
gevent
requests
flask-cors==3.0.10
```

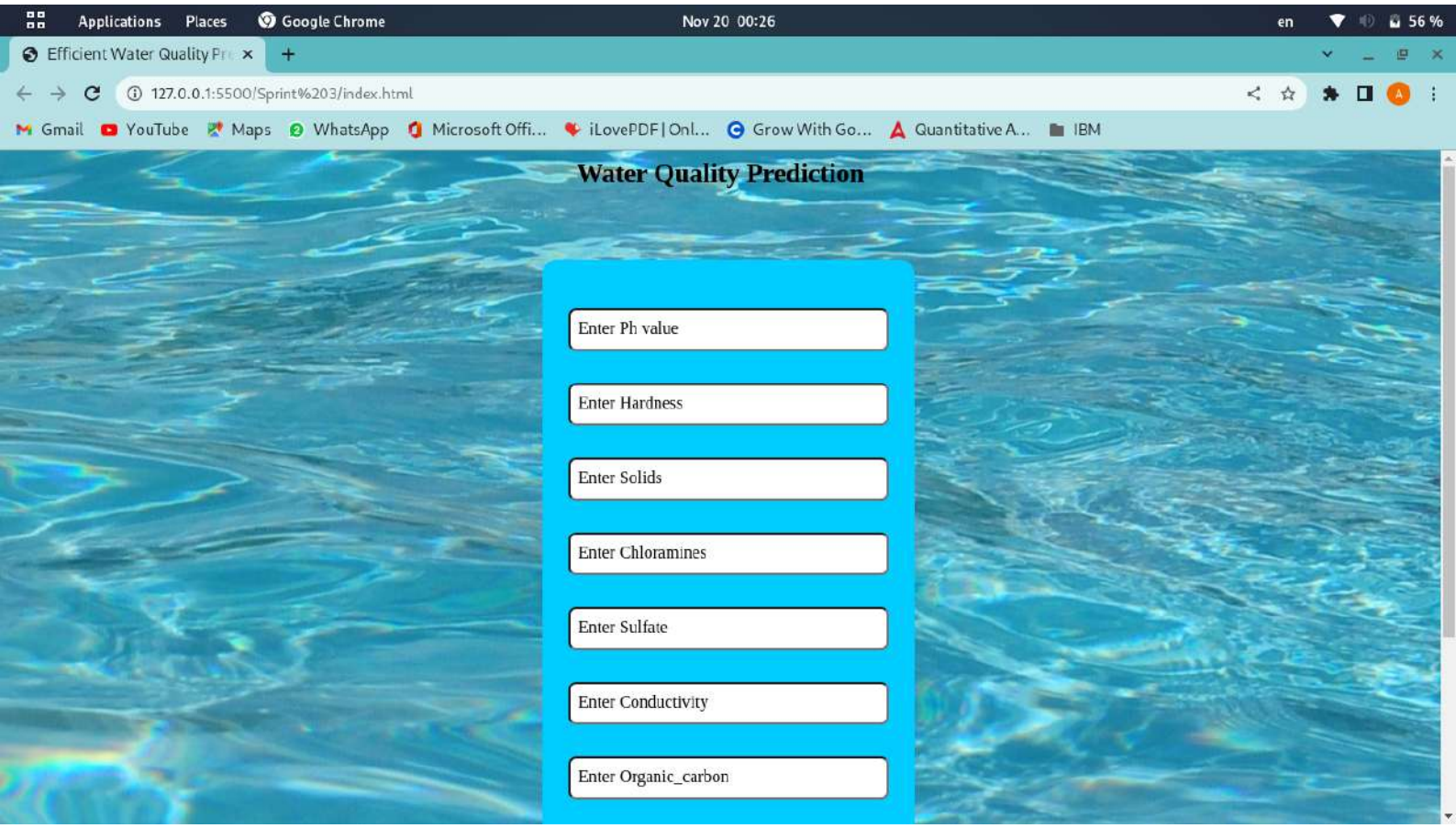
APP.PY

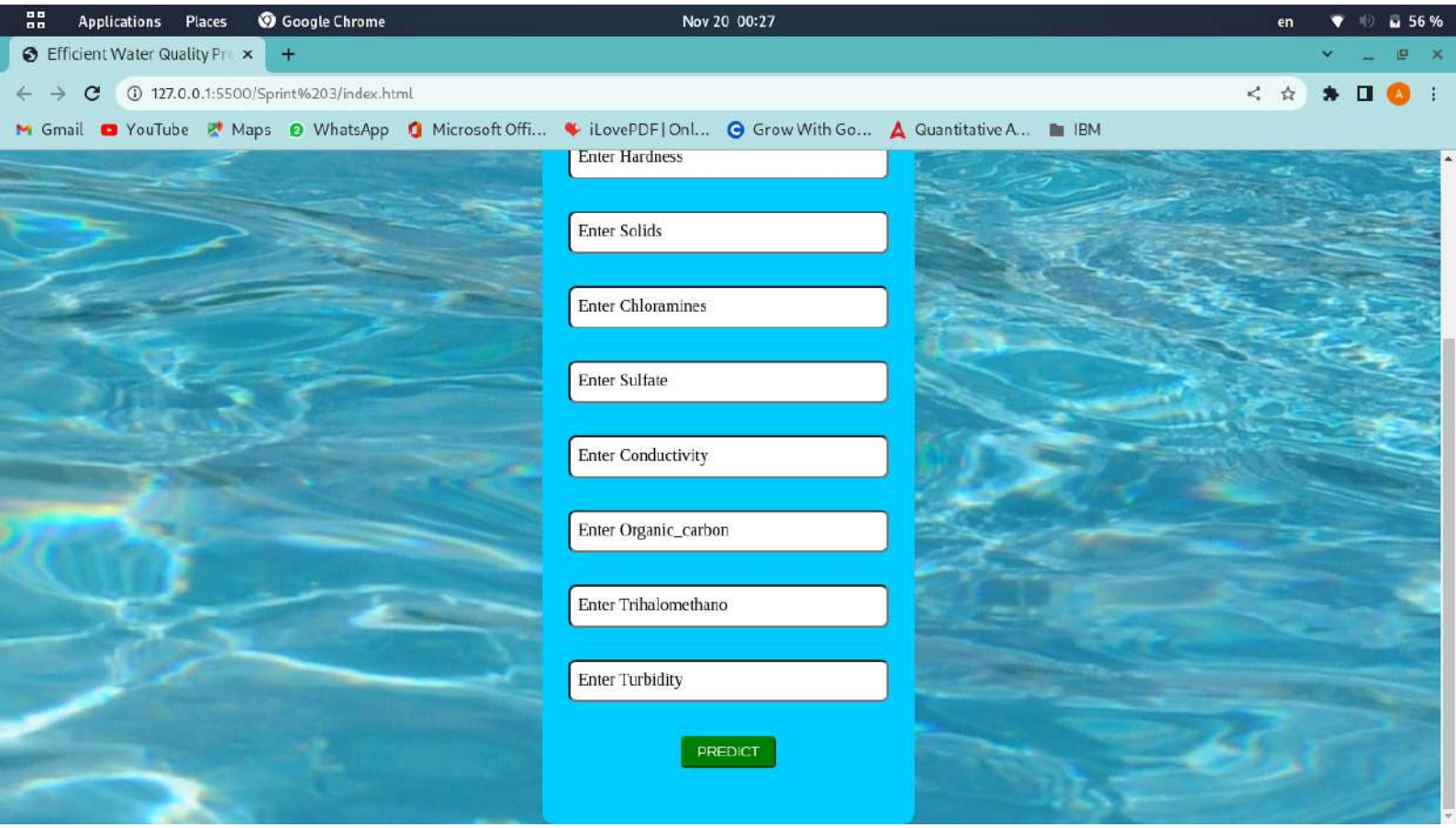
app.py > ...

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


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

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

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

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

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

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

WATER QUALITY.IPYNB




Applications
Places
Google Chrome
Nov 20 00:26
en
56 %
Efficient Water Quality Pr
127.0.0.1:5500/Sprint%203/index.html
Gmail
YouTube
Maps
WhatsApp
Microsoft Offi...
iLovePDF | Onl...
Grow With Go...
Quantitative A...
IBM
Water Quality Prediction
Enter Ph value
Enter Hardness
Enter Solids
Enter Chloramines
Enter Sulfate
Enter Conductivity
Enter Organic_carbon


Applications
Places
Google Chrome
Nov 20 00:27
en
56 %
Efficient Water Quality Pr
127.0.0.1:5500/Sprint%203/index.html
Gmail
YouTube
Maps
WhatsApp
Microsoft Offi...
iLovePDF | Onl...
Grow With Go...
Quantitative A...
IBM
Enter Hardness
Enter Solids
Enter Chloramines
Enter Sulfate
Enter Conductivity
Enter Organic_carbon
Enter Trihalomethano
Enter Turbidity
PREDICT