

Project Report
Efficient Water Quality Analysis & Prediction
using Machine Learning.



Thiagarajar College of Engineering,
Madurai-015.

Submitted by
Team ID: PNT2022TMID21280

Team Members:

917719C084_Sabareeswaran C

917719C126_Diwahar S

917719C132_Pavithra K

917719C137_Saravanan T G

CONTENTS

- 1. INTRODUCTION**
 1. Project Overview
 2. Purpose
- 2. LITERATURE SURVEY**
 1. Existing problem
 2. References
 3. Problem Statement Definition
- 3. IDEATION & PROPOSED SOLUTION**
 1. Empathy Map Canvas
 2. Ideation & Brainstorming
 3. Proposed Solution
 4. Problem Solution fit
- 4. REQUIREMENT ANALYSIS**
 1. Functional requirement
 2. Non-Functional requirements
- 5. PROJECT DESIGN**
 1. Data Flow Diagrams
 2. Solution & Technical Architecture
 3. User Stories
- 6. PROJECT PLANNING & SCHEDULING**
 1. Sprint Planning & Estimation
 2. Sprint Delivery Schedule
 3. Reports from JIRA
- 7. CODING & SOLUTIONING (Explain the features added in the project along with code)**
 1. Feature 1
 2. Feature 2
 3. Database Schema (if Applicable)
- 8. TESTING**
 1. Test Cases
 2. User Acceptance Testing
- 9. RESULTS**
 1. Performance Metrics
- 10.ADVANTAGES & DISADVANTAGES**
- 11.CONCLUSION**
- 12.FUTURE SCOPE**
- 13.APPENDIX**

Source Code
GitHub & Project Demo Link

1.Introduction:

1.1 Project Overview:

Water Quality can be defined as the chemical, physical and biological characteristics of water, usually in respect to its suitability for a designated use. Water can be used for recreation, drinking, fisheries, agriculture or industry. Each of these designated uses has different defined chemical, physical and biological standards necessary to support that use.

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.

In order to provide a better solution for the water quality analysis using machine learning algorithm, we implemented a model in this project.

1.2 Purpose:



Why Should I Test My Water Regularly

Water is perhaps the most precious natural resource after air. Though the surface of the earth is mostly consisting of water, only a small part of it is usable, which makes this resource limited. This precious and limited resource, therefore, must be used with care. As water is required for different purposes, the suitability of it must be checked before use. Also, sources of water must be monitored regularly to determine whether they are in sound health or not. Poor condition of water bodies is not only the indicator of environmental degradation, it is also a threat to the ecosystem. In industries, improper quality of water may cause hazards and severe economic loss. Thus, the quality of water is very important in both environmental and economic aspects.

2. Literature Survey:

2.1 Existing problem/solution:

1. The deteriorating quality of natural water resources like lakes, streams and estuaries, is one of the direst and most worrisome issues faced by humanity. The effects of un-clean water are far-reaching, impacting every aspect of life. Therefore, management of water resources is very crucial in order to optimize the quality of water. The effects of water contamination can be tackled efficiently if data is analysed and water quality is predicted beforehand. This issue has been addressed in many previous researches; however, more work needs to be done in terms of effectiveness, reliability, accuracy as well as usability of the current water quality management methodologies. The goal of this study is to develop a water quality prediction model with the help of water quality factors using Artificial Neural Network (ANN) and time-series analysis. This research uses the water quality historical data of the year of 2014, with 6-minutes time interval. Data is obtained from the United States Geological Survey (USGS) online resource called National Water Information System (NWIS). For this paper, the data includes the measurements of 4 parameters which affect and influence water quality. For the purpose of evaluating the performance of model, the performance evaluation measures used are Mean-Squared Error (MSE), Root Mean-Squared Error (RMSE) and Regression Analysis.

2. Water quality monitoring is a regular practice to assess the presence of pollutants in the water. The importance of monitoring is justified by the need to know the current state of aquatic ecosystems to design appropriate conservative and protective actions (Serrano Balderas et al., 2015). Data from water quality monitoring may be prone to have various problems (i.e., incomplete, inconsistent, inaccurate, or outlying data) that may result in misleading analysis interpretation (Berrahou et al., 2015). Incomplete data for instance, can be replaced by imputed values so that the statistical methods commonly used to describe patterns on water quality assessment (such as PCA, Hierarchical Classification, Kohonen-SOM) can be achieved. But imputation of missing values may impact statistical results. In this study, our goal is to assess the impact of imputation methods, and more generally of pre-processing, on the results of various statistical analyses. To this purpose, we studied five imputation methods (Mean, Hot-Deck, Sequential Imputation, Multiple Imputation and Iterative Stepwise Regression Imputation) on four statistical methods (Correlation, PCA, Kohonen-SOM and Hierarchical Classification) and developed a fully integrated analytics environment in R for statistical analysis of environmental data in general, and for water quality data analytics in particular.

The results obtained indicated that the imputation methods IRMI and MI generally improve the accuracy of the tested statistical methods when compared to methods without imputation. Our findings demonstrated that reliable results could be obtained when robust imputation methods are used to pre-process incomplete data.

3. Water quality becomes one of the important quality factors for the quality life in smart cities. Recently, water quality has been degraded due to diverse forms of pollution caused by disposal of human wastes, industrial wastes, automobile wastes. The increasing pollution affects water quality and the quality of people's life. Hence, water quality evaluation, monitoring, and prediction become an important and hot research subject. In the past, many environmental researchers have dedicated their research efforts on this subject using conventional approaches. Recently, many researchers begin to use the big data analytics approach to studying, evaluating, and predicting water quality due to the advances of big data applications and the availability of environmental sensing networks and sensor data. This paper reviews the published research results relating to water quality evaluation and prediction. Moreover, the paper classifies and compares the applied big data analytics approaches and big data-based prediction models for water quality assessment. Furthermore, the paper also discusses the future research needs and challenges.

2.2 References:

1. Predicting and analysing water quality using Machine Learning: A comprehensive model. Authors: Yafra Khan & Chai Soo See. Published in: 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT). Publisher: IEEE.

2. Water Quality Data Analytics. Authors: Eva Serrano, Laure Berti-Equille, Jean-Christophe Desconnets, Maria Aurora Armienta. Published in: iEMSs. International Environmental Modelling and Software Society Conference.

3. Data-Driven Water Quality Analysis and Prediction: A Survey. Authors: Jerry Gao. Published in: The 3rd IEEE International Conference on BIG DATA COMPUTING SERVICE AND APPLICATIONS.

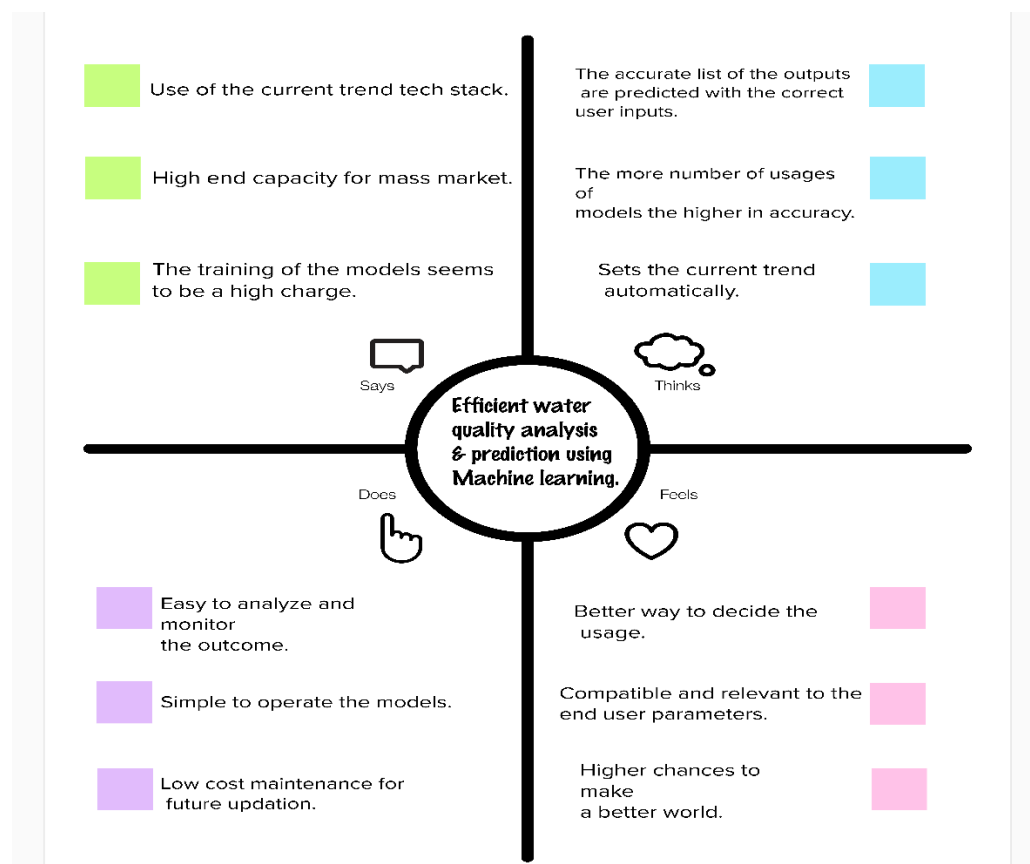
2.3 Problem Statement definition:

Water is perhaps the most precious natural resource after air. Though the surface of the earth is mostly consisting of water, only a small part of it is usable, which makes this resource limited. As water is required for different purposes, the suitability of it must be checked before use. Also, sources of water must be monitored regularly to determine whether they are in sound health or not. Thus, the quality of water is very important in both environmental and economic aspects.

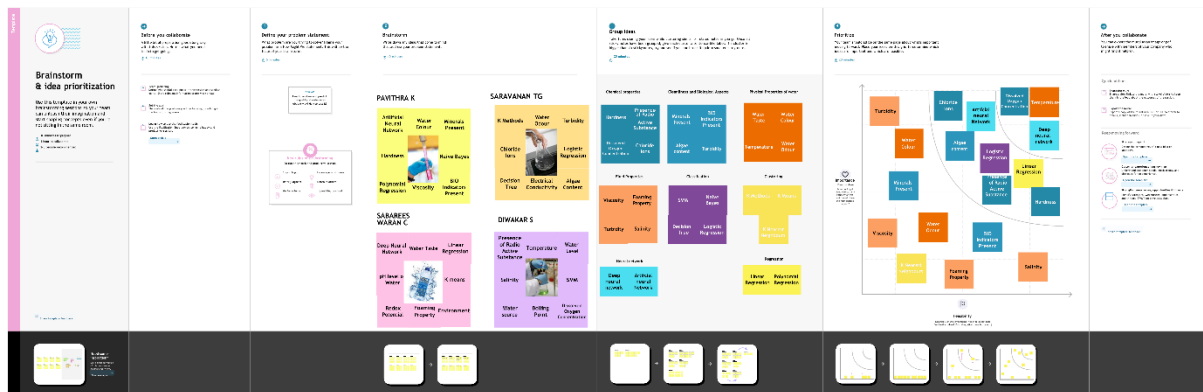
- To analyses the quality of water with its parameters.
- To perform a better analysis method using machine learning algorithms.
- To provide the valid results which is related to the real.
- To provide a user-friendly web application where the users can interact with it to analyses the water by providing the parameter values.

3. Ideation & Proposed Solution:

3.1 Empathy Map Canvas:



3.2 Ideation & Brainstorming:



3.3 Proposed Solution:

S.No	Parameter	Description
1.	Problem Statement (Problem to be solved).	Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions.
2.	Idea / Solution description.	To provide the better analysis of water quality, we use some prediction algorithms. The predicted value will determine the quality.
3.	Novelty / Uniqueness.	Water Quality Analysis holds crucial value as one of the sustaining factors of public health and sustenance. There have been numerous researches conducted in this premise, all evaluating upon selective parameters that promise to determine water quality. Feature engineering techniques like scaling and smote analysis can be done to elevate the quality so as to improve the final results.
4.	Social Impact / Customer Satisfaction.	Increased expenditure on recreational fishing, increased revenue from commercial fishing, reduced water treatment costs, average increase in property prices, tourism growth (with associated revenue and employment

		opportunities) and benefits of improved environmental and biodiversity protection.
5.	Business Model (Revenue Model).	This solution will have a major impact on the financial. Since, there is no better software to do water quality analysis, everything done physically which costs more. Definitely, our solution will show the reduction the costs.
6.	Scalability of the Solution.	The solution will be provided as a Software as a Service model which will be more scalable to use.

3.4 Problem Solution fit:

Project Title: Efficient Water Quality Analysis & Prediction using Machine Learning

Project Design Phase-I - Solution Fit Template

Team ID: PNT2022TMID21280

Define CS, fit into CC	1. CUSTOMER SEGMENT(S) <small>Regulators, local government employees, industry representatives, landowners, farmers and members of the general public.</small>	6. CUSTOMER CONSTRAINTS <small>Spending Power Network Connection Available devices</small>	5. AVAILABLE SOLUTIONS <small>Which solutions are available to the customers when they face the problem</small>	Explore AS, differentiate
	2. JOBS-TO-BE-DONE / PROBLEMS <small>Which jobs-to-be-done (or problems) do you address for</small>	9. PROBLEM ROOT CAUSE <small>What is the real reason that this problem exists? What is the back</small>	7. BEHAVIOUR <small>What do our customers do to address the problem and I.e. directly related: find the right solar panel installer, calculate</small>	
Focus on J&P, tap into BE, understand RC				Focus on J&P, tap into BE, understand RC
Identify triggers & dependencies	3. TRIGGERS <small>What triggers customers to act? I.e. seeing their neighbour installing solar panels, reading about a more efficient solution in the news.</small>	10. YOUR SOLUTION <small>If you are working on an existing business, write down your current solution first, fill in the canvas, and check how much it fits reality. If you are working on a new business proposition, then keep it blank until you fill in the canvas and come up with a solution that fits within customer limitations, solves a problem and matches customer behaviour.</small>	8. CHANNELS of BEHAVIOUR 8.1 ONLINE <small>What kind of actions do customers take online? Extract online channels from #7</small> 8.2 OFFLINE <small>What kind of actions do customers take offline? Extract offline channels from #7 and use them for customer development.</small>	Identify triggers & dependencies
	4. EMOTIONS: BEFORE / AFTER <small>How do customers feel when they face a problem or a job and afterwards? I.e. lost, insecure > confident, in control - use it in your communication strategy & design.</small>			

4. Requirement Analysis:

4.1 Function Requirements:

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User input	Users are required to give chemical components of their water which they need tested. The chemical components such as Temperature, pH, Dissolved Oxygen, Fecal Coliform, Biochemical oxygen demand, conductivity and Nitrate details are must.
FR-2	Output Display	Based on the range of water quality index available, given water samples are analyzed and predicted the final results.
FR-3	Model prediction	Confirming based on water quality index and shows the ML prediction with percent of various parameter.
FR-4	Data handling	File contains water quality metrics for different water bodies.
FR-5	Quality analysis	Analyze with acquired information of water across various water quality indication using different models.

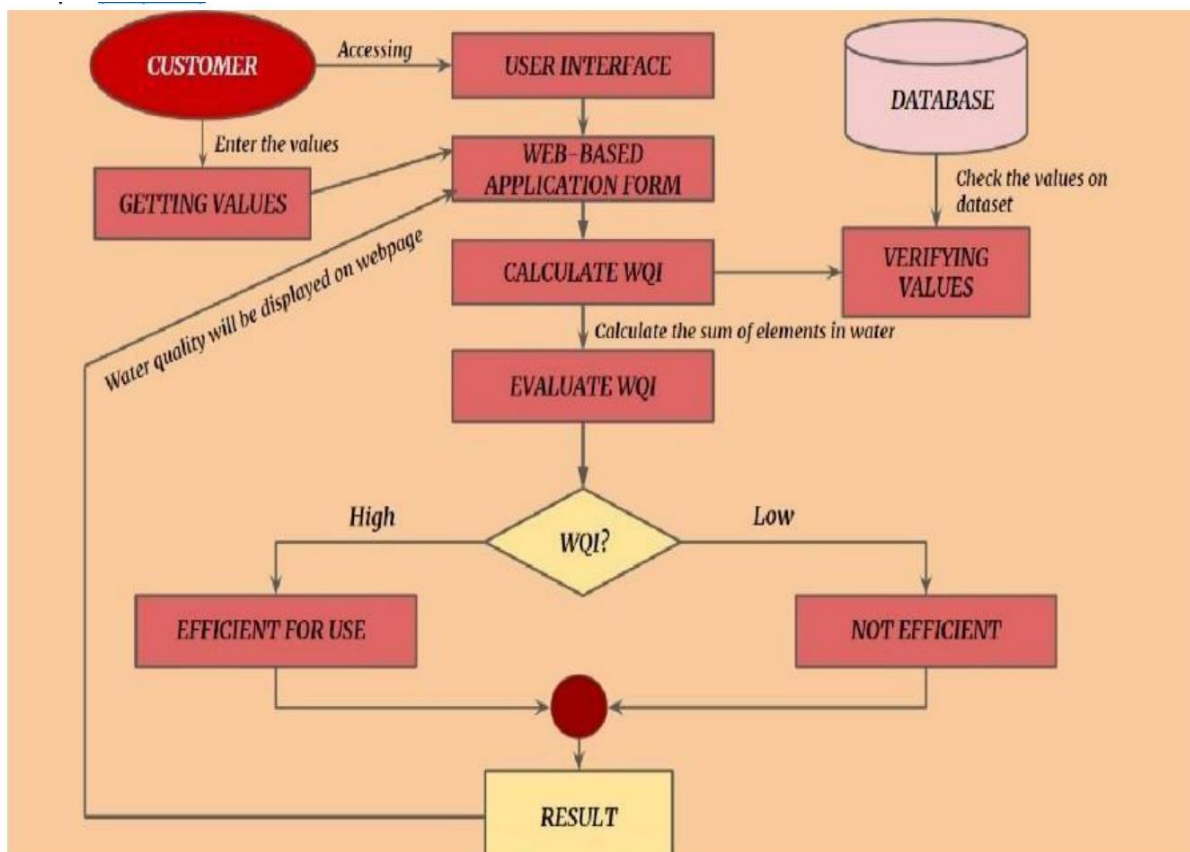
4.2 Non-Functional Requirements:

FR No.	Non-Functional Requirement	Description
NFR-1	Usability	System is stand up to the customer's expectations. When an application is, users can easily navigate its interface. The user can determine what feature and what it can do.
NFR-2	Security	Various forms of question for calculating the water quality index (wqi) and securely stored in database.
NFR-3	Reliability	If the number of failures is low, it means that the system operates properly. Reliability of Track the time between critical feature can help you understand

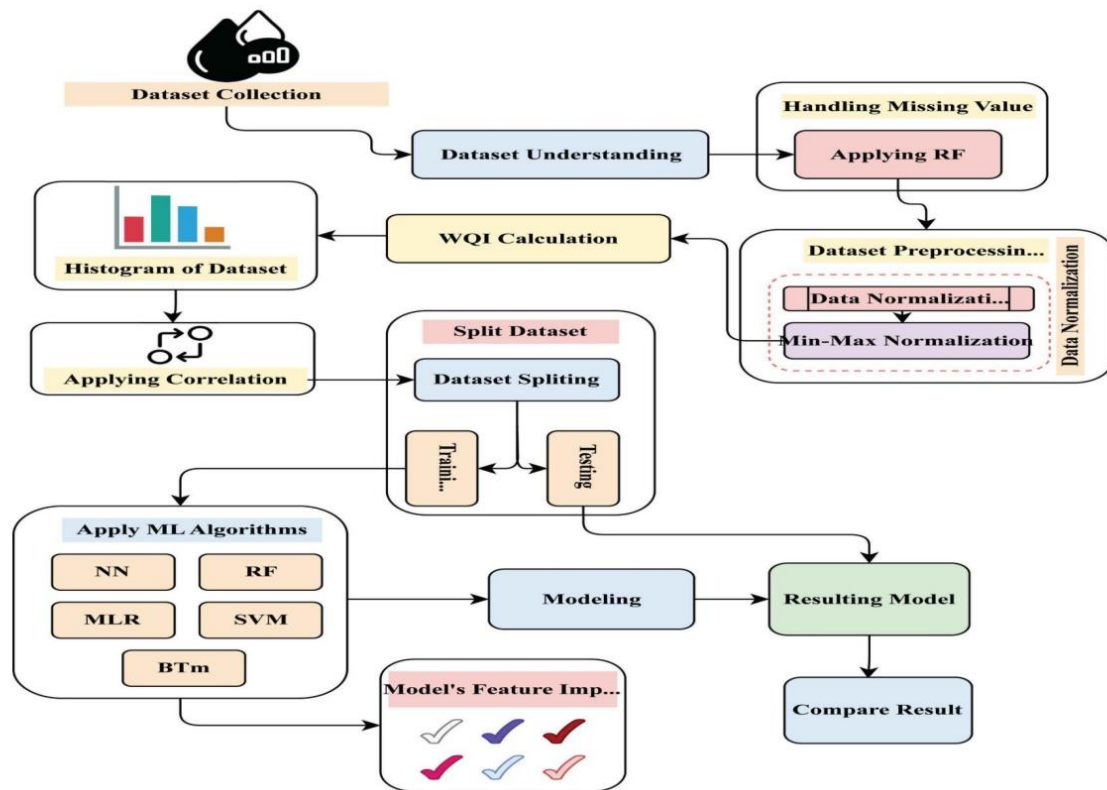
NFR-4	Performance	Our system should run on a 32 bit(x86) or 64 bit(x64) dual-core 2.66-GHZ or faster processor. It should not exceed 2 GB ram.
NFR-5	Availability	The system should be available for the duration of the user access to the system until the user terminates the access.
NFR-6	Scalability	It provides an efficient outcome and ability to increase or decrease the performance of a system based on datasets.

5. Project Design:

5.1 Data Flow Diagrams:



5.2 Solution & Technical Architecture:



5.3 User Stories:

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Web user)	Access web page	USN-1	As a user, anyone can access the web page to check the water quality.	I can access my webpage online at any time.	High	Sprint-1
Customer	Usage of water	USN-2	As per the usage of the user the quality of water should be predicted in an easy way.	Prediction can be done in an easy way.	High	Sprint-2
Customer	Accuracy of water	USN-3	Using a prediction model the user will know the quality of water on a daily basis.	The quality analysis of water will be accurate.	High	Sprint-3
Administrator	Manage the web page	USN-4	As admin, I can manage user details and update parameters essential for prediction.	Make changes on user interface (UI).	High	Sprint-3
Administrator	Calculation of WQI	USN-5	An admin can update the calculations for water quality index calculation.	Improves the accuracy of the calculation.	High	Sprint-3

6. Project Planning & Scheduling:

6.1 Sprint Planning & 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.	8	High	SABAREESWARAN C PAVITHRA K SARAVANAN TG DIWAHAR S
Sprint-1	User Confirmation	USN-2	As a user, I will receive confirmation email once I have registered for the application	8	Medium	SABAREESWARAN C PAVITHRA K SARAVANAN TG DIWAHAR S
Sprint-1	Login	USN-3	As a user, I can log into the application by entering email & password	4	High	SABAREESWARAN C PAVITHRA K SARAVANAN TG DIWAHAR S
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.	20	High	SABAREESWARAN C PAVITHRA K SARAVANAN TG DIWAHAR S
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.	20	High	SABAREESWARAN C PAVITHRA K SARAVANAN TG DIWAHAR S
Sprint-4	Web Application	USN-1	As user, I will show the current Information of the water quality.	20	Medium	SABAREESWARAN C PAVITHRA K SARAVANAN TG DIWAHAR S

6.2 Sprint Delivery Schedule:

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	25 Oct 2022	28 Oct 2022	20	01 Nov 2022
Sprint-2	20	4 Days	29 Oct 2022	1 Nov 2022	20	04 Nov 2022
Sprint-3	20	4 Days	02 Nov 2022	5 Nov 2022	20	11 Nov 2022
Sprint-4	20	4 Days	12 Nov 2022	15 Nov 2022	20	19 Nov 2022

6.3 Reports from JIRA:

The screenshot shows the Jira Backlog view. The left sidebar contains navigation options: PLANNING (Roadmap, Backlog, Board), DEVELOPMENT (Code), and Project pages. The main area displays the Backlog for the project 'Efficient Water Quality Analysis and Prediction using Machine Learning'. It shows two sprints: 'EWQAPUML Sprint 1' (25 Oct – 29 Oct) and 'EWQAPUML Sprint 2' (31 Oct – 5 Nov). Each sprint contains issues with labels like 'DATA COLLECTION' and 'MODEL BUILDING'. A search bar and filters are visible at the top.

The screenshot shows the Jira Board view for the same project. The left sidebar is similar to the Backlog view. The main area displays the 'EWQAPUML Sprint 1' board. It shows a Kanban board with columns: 'TO DO 2 ISSUES', 'IN PROGRESS', and 'DONE'. Issues are listed in the 'TO DO' column, including 'Data Collection' and 'Data Preprocessing'. A search bar and filters are visible at the top.

	SEP	OCT	NOV	
Sprints			EWQAPUML Sprint 1, EWQAPU...	
> EWQAPUML-1 Data Collection				
> EWQAPUML-2 Model Building				
> EWQAPUML-3 Training and Testing				
> EWQAPUML-4 Implementation of Web page				

7. Coding & Solutioning:

7.1 Feature 1:

Running Flask app.

```
PS C:\Users\LENOVO\OneDrive\Desktop\FlaskApplication> python -m flask run
C:\Python310\lib\site-packages\sklearn\base.py:329: UserWarning: Trying to unpickle estimator DecisionTreeRegressor from version 1.0.2 when using version 1.1.3. This might lead to breaking code or invalid results. Use at your own risk.
For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
warnings.warn(
C:\Python310\lib\site-packages\sklearn\base.py:329: UserWarning: Trying to unpickle estimator RandomForestRegressor from version 1.0.2 when using version 1.1.3. This might lead to breaking code or invalid results. Use at your own risk.
For more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
warnings.warn(
* Debug mode: off
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
* Running on http://127.0.0.1:5000
Press CTRL+C to quit
```

7.2 Feature 2:

Cloud deployment.

WQA ✔ Deployed Online

API reference Test

Code snippets

cURL Java JavaScript **Python** Scala

```
import requests


# NOTE: you must manually set API_KEY below using information retrieved from your IBM Cloud account
API_KEY = "<your API key>"
token_response = requests.post('https://iam.cloud.ibm.com/identity/token', data={"apikey":
    API_KEY, "grant_type": 'urn:ibm:params:oauth:grant-type:apikey'})
mltoken = token_response.json()["access_token"]

header = {'Content-Type': 'application/json', 'Authorization': 'Bearer ' + mltoken}

# NOTE: manually define and pass the array(s) of values to be scored in the next line
payload_scoring = {"input_data": [{"fields": [array_of_input_fields], "values": [array_of_values_

response_scoring = requests.post('https://us-south.ml.cloud.ibm.com/ml/v4/deployments/4f6df3d2-3e
    headers={'Authorization': 'Bearer ' + mltoken})
```

Efficient Water Quality Analysis Platform

 Enter the details:

StationID

Location

State

Temperature

Dissolved Oxygen(mg/l)

PH Value

Conductivity

B.O.D.(mg/l)


Nitratenan N + Nitritenann(mg/l)

Total Coliform(MPN/100ml)Mean

Year

Submit

Efficient Water Quality Analysis Platform

 Enter the details:

1339

Tamilnadu

Tamil

30.6

6.3

6.9

203.0
6.940049
0.1
27.0
2014
<input type="button" value="Submit"/>

Results:

Fair, The predicted value is [69.8402]

8. Testing:

8.1 Test Cases:

Test case id	Feature type	Component	Test Scenario	Steps to execute	Test data	Expected Result	Actual Result	Status
Home page_TC_001	Functional	Home page	Verify user can see the prediction button and the input columns for prediction.	Verify the prediction button to analyze the quality.	-	Input columns and the prediction button should be displayed.	Working as expected.	Pass.
Home page_TC_002	UI	Home page	Verify whether the font alignment and size are correct.	Verify whether the font alignment and size are correct.	-	The font alignment and size are correct.	Working as expected.	Pass.
Results page_TC_003	UI	Results page	Verify whether the font alignment and size are correct.	Verify whether the font alignment and size are correct.	-	The font alignment and size are correct.	Working as expected.	Pass.

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 Efficient Water Quality Analysis and Prediction using Machine Learning 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
Home Page	7	0	0	7
Client Application	51	0	0	51
Prediction	2	0	0	2
Pop ups	3	0	0	3
URL port	9	0	0	9
Final Report Output	4	0	0	4
Redirecting	2	0	0	2

9. Results:

9.1 Performance metrics:

```
In [37]: from sklearn import metrics
```

```
In [38]: 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: 0.8845729323308585  
MSE: 5.087236515388475  
RMSE: 0.9405173748160416
```

```
In [39]: metrics.r2_score(y_test, y_pred)
```

```
Out[39]: 0.9722694819035159
```

10. Advantages & Disadvantages:

Advantages: User can know the quality of water, no fear of water quality, helps to avoid impure water, helps to prevent disease caused from impure water as it shows the usability of the water.

Disadvantages: Requires lab values to calculate the purity.

11. Conclusion:

From this, we can conclude that we can analyze the quality of water from the values or data like BDO, Ph, Dissolved oxygen levels etc .By using this data we calculate the purity of water of by using some formulae. This application helps the user to know the purity of water they use ,components in the water, usability of water etc.

12. Future Scope:

It helps to calculate large data. Easily calculate the purity of water.

13. Appendix:

Github: <https://github.com/IBM-EPBL/IBM-Project-24129-1659938250>

Demo:<https://drive.google.com/file/d/1gxtRGfsxnq1V2STTziUfsZdrG5E84z9X/view?usp=sharing>