19CSP14 - PROFESSIONAL READINESS FOR INNOVATION, EMPLOYABILITY AND ENTREPRENEURSHIP

UNIVERSITY ELIGIBITY CRITERIA PREDICTOR

Team ID: PNT2022TMID18285

Team Leader: VEVINYA A

Team member 1: ALFINA GRACELINE J

Team member 2: SOWPARNIKA R S

Team member 3: JAISHA G

1 . INTRODUCTION

a. Project Overview

Water is one of the most important natural resources for all living organisms on earth. The monitoring of treated wastewater discharge quality is vitally important for the stability and protection of the ecosystem. Collecting and analyzing water samples in the laboratory consumes much time and resources. In the last decade, many machine learning techniques, like multivariate linear regression (MLR) and artificial neural network (ANN) model, have been proposed to address the problem. However, simple linear regression analysis cannot accurately forecast water quality because of complicated linear and nonlinear relationships in the water quality dataset. The ANN model also has shortcomings though it can accurately predict water quality in some scenarios. So AutoMl and Random Forest algorithm has been proposed for accurate results. Random Forest algorithm with hyper paramters has shown good and improved accuracy in water quality prediction

b. Purpose

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 analyzed and water quality is predicted beforehand. So the purpose of this study is to develop a water quality prediction model with the help of water quality factors using Random Forest and AutoMl Algorithm.

2. LITERATURE SURVEY

2.1 EXISTING PROBLEM

1) Automating water quality analysis using ML and auto ML techniques

This paper evalutes traditional and AutoML techniques within the avenue of water quality analysis by collecting the dataset from the Korattur Lake, Chennai. The dataset consists of observations of over a ten-year period, starting from 2009 until 2019. Under 9 parameters, around 5000 records are existent. The 9 parameters specified are Total Dissolved Solids (TDS), Turbidity, pH, Chemical Oxygen Demand (COD), Iron, Phosphate, Sodium, Chloride and

Nitrate. From the preliminary stages, data proved to have a profound impact upon the both models. Use of SMOTE increased accuracy, reinforcing the fact that AutoML, efficient as it might be, provides better results when data is cleaned, handled and moulded to suit the purpose. The factors such as time taken, academic experience required are all extremely less in the case of AutoML.

2) WaterNet: A Network for Monitoring and Assessing Water Quality for Drinking and Irrigation Purposes.

In this paper they have expressed water quality in terms of WQI (Water Quality Index) and IWQI (Irrigation Water Quality Index). Collecting water samples from different sources, measuring the various parameters present, and bench-marking these measurements against pre- set standards, while adhering to various guidelines during transportation and measurement can be extremely daunting. This uses network architecture to collect data on water parameters in real-time and use Machine Learning (ML) tools to automatically determine suitability of water samples for drinking and irrigation purposes. The developed monitoring network is based on LoRa and takes the land topology into consideration. Results of the test showed that LR performed best for drinking water, as it gave the highest classification accuracy and lowest false positive and negative values, while SVM was better suited for irrigation water.

3) Evaluation and Analysis of Goodness of Fit for Water Quality Parameters using Linear Regression through the Internet of Things (IoT) based Water Quality Monitoring System.

In this paper they have used IoT help to obtain real-time data, in the river basin region. To implement this we make use of WQM system. WQM consists of sensors such as T, pH, dissolved oxygen (DO), electrical conductivity (EC), biochemical oxygen demand (BOD), NO3, and total dissolved solids (TDSs) to monitor water quality . The Smart WQM is used for ecological monitoring of the water environment. An IoT system based on low-cost Raspberry Pi for WQM that controls the flow of water. The monitored parameters are physicochemical parameters. The WQM uses linear regression that helps to estimate the relationship between two parameters. After linear regression apply one-way ANOVA to the samples. It is used to compare two or more sample means by the F distribution method. Overall, we can see that all of the water quality parameters are within the normal range prescribed, and the water can be used for daily purposes.

4) Multiparametric System for Measuring Physicochemical Variables Associated to Water Quality Based on the Arduino Platform.

In this paper they have used pH, ORP, turbidity and TDS sensors provide an analog output. A 16-bit-ADC module increases the resolution of 10 bits offered by the Arduino Mega native ADC. ORP is an electrochemical parameter that is measured similarly to pH, but the electrode uses a

noble metal as a measurement element. TDS provides a measure of the water salinity, and it is related to the EC of water. Turbidity is an optical property describing how much light is scattered for a water sample. An IR light source like LED, sends light into a water sample. The Arduino Mega has an ADC of 10 bits. Incorporating the module ADS1115 from Adafruit, having a 16-bit ADC with a programmable gain amplifier improves the system resolution. Dissolved oxygen provides the magnitude of the oxygen gas dissolved in water. Overall the system exhibited a good performance with low-cost and readily available elements.

5) Predicting and analyzing water quality using Machine Learning: a comprehensive model.

In this paper they have developed a water quality prediction model with the help of water quality factors using Artificial Neural Network(ANN) with Nonlinear Autoregressive(NAR) time series. Time series has been used with Scaled Conjugate Gradient(SCG) as training algorithm. Time Series Data's are collected from United States Geological Survey(USGS) online resource called NWIS .The samples include the data ranging from January to March 2014,with 6-minute time interval. Four water quality factors Turbidity, Dissolved Oxygen Concentration, Chlorophyll and Specific Conductance have been measured using four ANN models. The performance of 4 models have been analyzed using Mean Square Error(MSE) andRoot Mean Square Error(RMSE).The ANN-NAR model provides best accuracy with lowest MSE od 3.7x10^-4 for turbidity and best Regression Value for Specific Conductance(0.99).

6) Flexible RFID tag for sensing the total minerals in drinking water via smartphone tapping.

In this project, they have designed and implemented RFID sensor tag for evaluating total minerals in drinking water. The sensor reading can be obtained through smartphone tapping and the results are received in 1 second the reading range between smart phone and sensor tag is 1-3cm.The developed RFID sensors exhibits particular superiority in flexibility and convenience of use due to advantages in wireless power, data transfer, no added hardware and software for smartphones.

2.2 References

[1] Prasad, D. Venkata Vara, P. Senthil Kumar, Lokeswari Y. Venkataramana, G. Prasannamedha, S. Harshana, S. Jahnavi Srividya, K. Harrinei, and Sravya Indraganti. "Automating water quality analysis using ML and auto ML techniques." Environmental Research 202 (2021): 111720.

[2] Ajayi, Olasupo O., Antoine B. Bagula, Hloniphani C. Maluleke, Zaheed Gaffoor, Nebo Jovanovic, and Kevin C. Pietersen. "WaterNet: A Network for Monitoring and

Assessing Water Quality for Drinking and Irrigation Purposes." IEEE Access 10 (2022): 48318-48337.

[3] Kenchannavar, Harish H., Prasad M. Pujar, Raviraj M. Kulkarni, and Umakant P. Kulkarni. "Evaluation and Analysis of Goodness of Fit for Water Quality Parameters using Linear Regression through the Internet of Things (IoT) based Water Quality Monitoring System." IEEE Internet of Things Journal (2021).

[4] Fonseca-Campos, Jorge, Israel Reyes-Ramirez, Lev Guzman-Vargas, Leonardo Fonseca- Ruiz, Jorge Alberto Mendoza-Perez, and P. F. Rodriguez-Espinosa. "Multiparametric System for Measuring Physicochemical Variables Associated to Water Quality Based on the Arduino Platform." IEEE Access 10 (2022): 69700-69713.

[5] Khan, Yafra, and Chai Soo See. "Predicting and analyzing water quality using Machine Learning: a comprehensive model." In 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT), pp. 1-6. IEEE, 2016.

6] Qian, Xueqing, Zhen Li, Zhaozong Meng, Nan Gao, and Zonghua Zhang. "Flexible RFID tag for sensing the total minerals in drinking water via smartphone tapping." IEEE Sensors Journal 21, no. 21 (2021): 24749-24758.

2.3 Problem Statement Definition

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. The quality of water serves as a powerful environmental determinant and a foundation for the prevention and control of waterborne diseases. However predicting the urban water quality is a challenging task since the water quality varies in urban spaces non-linearly and depends on multiple factors, such as meteorology, water usage patterns, and land uses, so this project aims at building a Machine Learning (ML) model to Predict Water Quality by considering all water quality standard indicators

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas

An empathy map is a simple, easy-to-digest visual that captures knowledge about a user’s behaviors and attitudes. It is a useful tool to helpsteams better understand their users.

Creating an effective solution requires understanding the true problem and the person who is experiencing it. The exerciseof creating the map helps participants considerthings from the user’s perspective along with his or her goals and challenges.

Link:https://github.com/IBM-EPBL/IBM-Project-213-1658224391/blob/main/Project%20Design%20%26%20Planning/Ideation%20Phase/Empathy%20Map.pdf

3.2 Ideation and Brainstorming

Brainstorming provides a free and open environment that encourages everyone within a team to participate in the creative thinking process that leads to problem solving. Prioritizing volume over value, out-of-the-box ideas are welcomeand built upon, and all participants are encouraged to collaborate, helping each other develop a rich amount of creative solutions.

Step-1: Team Gathering, Collaboration and Select the Problem Statement

Step-2: Brainstorm, Idea Listing and Grouping

Step-3: Idea Prioritization

Link: https://github.com/IBM-EPBL/IBM-Project-213-1658224391/blob/main/Project%20Design%20%26%20Planning/Ideation%20Phase/Brainstorming-%20Idea%20Generation-%20Prioritizaation%20Template.pdf

3.3 Proposed Solution

S.No.	Parameter	Description
1.	Problem Statement (Problem to besolved)	Efficient Water Quality Analysis & Prediction Using MachineLearning
2.	Idea / Solution description	Dataset has to be pre-processed and suitable ML algorithm has to be applied andit has to be finetuned to improve the accuracy.
3.	Novelty / Uniqueness	After pre-processing the imported dataset, Random Forest and AutoML algorithm is applied andfine tuned to improve the accuracy.

4.	Social Impact / Customer Satisfaction	By adopting thismethod, people cometo know thecontent present in the waterthey use and they findit safe.
5.	Business Model (Revenue Model)	This method has the benefit of reusing the water by analyzing the content present in thatwater.
6.	Scalability of the Solution	Further afteranalysing the content present in the water, sufficient nutrients can be added which lacks in the water..

3.4 Problem Solution fit

Project Title: Efficient WaterQuality Analysis and Prediction using Machine Learning

Link: https://github.com/IBM-EPBL/IBM-Project-213-1658224391/blob/main/Project%20Design%20%26%20Planning/Ideation%20Phase/Problem_solution_fit_canvas.pdf

4. REQUIREMENT ANALYSIS

4.1 Functional requirements

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Registration through Form Registration through Gmail Registration through LinkedIN
FR-2	User Confirmation	Confirmation via EmailConfirmation via OTP
FR-3	Enter the input	Get the inputvalues via formand check the data.
FR-4	Executive administration	Two separate roles: Early warning/forecast monitoring - that are included in the regulation of monitoring thewater environment state and regulatory compliance, such as pollution event emergency management.

FR-5	User Requirements	The user needsan accurate and exact result.
FR-6	Data Preprocessing	From the raw dataset, obtainthe tested and trainedData.
FR-7	Data Handling	Metrics for thevarious water bodies'water quality included in the file.
FR-8	Quality analysis	Use multiple models to analysethe data on thewater's obtained PH, TDS,and temperature levels,among other water quality indicators.
FR-9	Model prediction	Based on the water qualityindex, the confirmation displays the machine learning prediction (Good, Partially Good, Poor) and the proportion of eachparameter that is present.
FR-10	Remote Visualization	Visualisation of futureforecasts using chartsbased on present and past valuesof all the parameters.

4.2 Non-functional Requirements:

Following are the non-functional requirements of the proposedsolution.

FR No.	**Non-Functional Requirement**	**Description**
NFR-1	**Usability**	A user-friendly web application, the system provides natural interaction with the users.
NFR-2	**Security**	The website is virus-free and did not request any authorization. The model hasa strong security system since the user'sinformation won't be shared withany other sources.
NFR-3	**Reliability**	A wide varietyof water valuesare trained in the model, increasing forecast accuracy. Themodel may begreatly expanded by adding moredatasets.
NFR-4	**Performance**	Get the results quickly.

NFR-5	**Availability**	Available on the internet at any moment.As long asthe user has access to the system, it should be accessible until theuser terminates it. The system responds to user requests more quickly, andrecovery is completed faster.
NFR-6	**Scalability**	It is a lightweight application, the users can access the website through mobile phones, tabs, desktop and laptop.It produces an effective result and has the capacity to alter the system's performancedepending on the datasets.

5. PROJECT DESIGN

5.1 Data Flow Diagram

A Data Flow Diagram(DFD) is a traditional visualrepresentation of the information flows within a system. A neat and clear DFD can depictthe 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.

5.2 Solution & Technical Architecture

Project flow from user side:

The user will give the details about the content present in the water in the web user interface. The website will then processthe data through the Machine Learningmodel deployed in it by us. Then it shows the predicted result in the web page.

Table-1 : Components & Technologies:

Component	**Description**	**Technology**
User Interface	How userinteracts with application e.g. Web UI, Mobile App, Chatbot.	HTML, CSS,JavaScript
Application Logic-1	Logic fora process in the application	Java / Python

Application Logic-2	Logic fora process in the application	IBM WatsonSTT service
Application Logic-3	Logic fora process in the application	IBM Watson Assistant
Database	Data Type,Configurations etc.	MySQL, NoSQL, etc.
Cloud Database	Database Service on Cloud	mongoDB atlas.
File Storage	-	-
External API-1	Purpose of External API used in the application	NPM package encryption
External API-2	Purpose of External API used in the application	Aadhar API,etc.
Machine Learning Model	Purpose of Machine Learning Model	Object Recognition Model, etc.
Infrastructure (Server/ Cloud)	-	-

Table-2: Application Characteristics:

S.No	**Characteristics**	**Description**
1.	Open-Source Frameworks	List the open-source frameworks used
2.	Security Implementations	List all the security
3.	Scalable Architecture	Justify the scalability of architecture
4.	Availability	Justify the availability of application
5.	Performance	Design consideration for the performance of theapplication.

5.3 User Stories

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Mobileuser)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirmingmy password.	I can access my account /dashboard	High	Sprint-1
		USN-2	As a user, I will receive confirmation emailonce I have registered for the application	I can receive confirmation emailand &click confirm	High	Sprint-1
		USN-3	As a user, I can register for the applicationthrough Facebook	I can register & accessthe dashboard with Facebook Login	Low	Sprint-2
		USN-4	As a user, I can register for the applicationthrough Gmail		Medium	Sprint-1

	Login	USN-5	As a user, I can log into the application byentering email & password		High	Sprint-1
	Dashboard					
Customer (Webuser)	Register	USN-7	As a web user, I can register for the application by entering my email, password,and confirming my password.	I can access my account /dashboard	High	Sprint-1
	Login	USN-8	As a web user, I can log intothe	I can log intothe	High	Sprint-1
			application by entering email &	application by		
			password	entering email & password		
USN-12	As a customer careexecutive, I will receivethe information aboutissues of the customer once I can login into the	I can receive issuefrom customer	High	Acceptance criteria	Priority	Release

	application					
Administrator	Login	USN-13	As a administrator, I can login using Admin user name and password	Sprint-1		
	Dashboard	USN-14	As a administrator, I can access the dashboard of the customer	I can login to the application	High	Sprint-1

6. PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story/ Task	Story Points	Priority	Team Members

Sprint1	Data Collection	USN-1,2	Collecting/downloading datasetfor pre-processing .	10	High	VevinyaA Alfina Graceline J Sowparnika R S Jaisha G
Sprint1		USN-1,2	Data pre-processing-formats the data and handlesthe missing data in the dataset..	10	Medium	VevinyaA Alfina Graceline J Sowparnika R S Jaisha G
Sprint2	Model Building	USN-1,2	Calculate the Water Quality Index (WQI) using specified formula for every parameter.	10	High	VevinyaA Alfina Graceline J Sowparnika R S JaishaG
Sprint2		USN-1,2	Splitting the data into training and testing datasetfrom the entire dataset.	10	High	VevinyaA Alfina Graceline J Sowparnika R S JaishaG
Sprint3	Training and Testing	USN-1,2	Training the model using AutoMLalgorithm and testing the performance of the model (accuracy rate)	20	High	VevinyaA Alfina Graceline J Sowparnika R S JaishaG

Sprint4	Implementation ofWeb page	USN-1,2	Implementing the web page for collecting the data from user	10	High	VevinyaA Alfina Graceline J Sowparnika R S JaishaG
Sprint4		USN-1,2	Deploying the model using IBM Cloud and IBM Watson Studio	10	Medium	VevinyaA Alfina Graceline J Sowparnika R S JaishaG

6.2 Sprint Delivery Schedule

Project Tracker, Velocity& Burndown Chart:

Sprint	**Total StoryPoints**	**Duration**	**SprintStart Date**	**Sprint End Date(Planned)**	**Story Points Completed (as on PlannedEnd Date)**	**Sprint ReleaseDate (Actual)**
Sprint1	20	6 Days	24 Oct 2022	29 Oct 2022	20	29 Oct 2022
Sprint2	20	6 Days	31 Oct 2022	05 Nov 2022	20	05 Nov 2022
Sprint3	20	6 Days	07 Nov 2022	12 Nov 2022	20	12 Nov 2022
Sprint4	20	6 Days	14 Nov 2022	19 Nov 2022	20	19 Nov 2022

Velocity:

Sprint 1 Average Velocity:

Average Velocity= 20/6 = 3.3

Sprint 2 Average Velocity:

Average Velocity= 20/6 = 3.3

Sprint 3 Average Velocity:

Average Velocity= 20/6 = 3.3

Sprint 4 Average Velocity:

Average Velocity= 20/6 = 3.3

7. CODING & SOLUTIONING

7.1 Feature 1

As we have used cloud depolyment model, anywhere from the world people can check the quality of water before drinking it.

7.2 Feature 2

We have fine tuned the parameters , in such a way the user can easily measure the parameter values. The parameters we have used is pH, Turbidity, Hardness, Chloramines, Sulfate, Conductivity, Organic carbon , Trihalomethane. These parameters can be easily found and measured even by the normal people without high knowledge.

8.TESTING

Here the Water quality prediction is the home page and Portable water prediction is the predicted output page.

8.1 TEST CASES

Compone	Test Scenario	Steps To	Test	ExpectedResult	Actual	Status

nt		**Execute**	**Data**		**Result**	
Home Page	Verify user can see the submit button and the input columns for prediction	Verify the submit button to analyze the quality.	-	Input columns and thepredicti on buttonshou ld be displayed	Working as expect ed	Pass
Home Page	Verify whetherthe page redirection is correct	Verify whetherthe redirection of page to predicted page is correct.	-	Redirection to predicted page should be correct	Working as expect ed	Pass
Home Page	Verifywhether the Heading, font alignment and size are correct	Verifywhether the Heading , font alignment and size are correct	-	The Heading, font alignment and size should be displayed correctly.	Working as expect ed	Pass
Predicted Page	Verifywheth er the predicted page displays the predicted value correctly	Verifywheth er the predicted page displays the data correctly	-	The predicted page displays the data correctly	Working as expect ed	Pass

Test Scenarios:

1. Verify user can see home page?
2. Verify user can enter values to input field?
3. Verify user can see predicted output page?
4. Verify Predicted data is displayed or not?

8.2 USER ACCEPTANCE TESTING

Purpose of Document

The purpose of this documentis to briefly explain the test coverageand open issues of the [ProductName] projectat the time of the release to User Acceptance Testing (UAT).

Defect Analysis

This reportshows the numberof resolved or closedbugs at each severity level, and howthey 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

Test Case Analysis

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

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

9. RESULTS

9.1 Performance Metrics

Evaluation metrics:

Algorithms	Random Forest Classifier	AutoML	Random Forest Classifier with Hyperparameters
Accuracy	0.6587225929456625	0.6549094375595805	0.6679387312944022

Random Forest Classifier:

Confusion Matrix:

[[563, 74],

[284, 128]]

Classifciation Report:

precision recall f1-score support

0	0.88	0.66	0.76	847
1	0.31	0.63	0.42	202
accuracy			0.66	1049
macro avg	0.60	0.65	0.59	1049
weighted avg	0.77	0.66	0.69	1049

10. ADVANTAGES & DISADVANTAGES

ADVANTAGES:

1)People who are undergoing Chemotherapy, transplant , pregnant women , infants can drink the safe water after testing the quality.

2) We have upload our trained model in IBM Watson cloud, So it is easy for accessing and testing the water quality for people anywhere in the world provided with internet connection.

3) We can avoid many water borne diseases like Cholera , Diarrhea, Hepatitis A etc. .

DISADVANTAGES:

1)People with internet connectivity can only make use of this water quality analyzer.

11. CONCLUSION

Water is one of the most important natural resources for all living organisms on earth. The monitoring of treated wastewater discharge quality is vitally important for the stability and protection of the ecosystem. Collecting and analyzing water samples in the laboratory consumes much time and resources. So machine Learning techniques like Random Forest and Auto Ml algorithm is proposed and model is trained using these algorithm and accuracy is predicted. It is observed that Random Forest algorithm with hyper paramters gives better accuracy of 66.79%.

12. FUTURE SCOPE

Testing the quality of water before using prevents many water borne diseases. People at Anytime and at Anywhere can use this and can get benefit from it. Machine learning algorithm like Random forest with hyper paramters gives better accuracy of 66.79%.In future we will extend our project by increasing accuracy with the help of other machine learning algorithms.