Project Report

Team ID	PNT2022TMID21489
Project Name	Essential Water Quality Analysis and Prediction using Machine learning

1. INTRODUCTION

1.1 PROJECT OVERVIEW

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.

1.2 PURPOSE

The purpose of water quality analysis is to obtain quantitative information on the physical, chemical, and biological characteristics of water via statistical sampling. Water quality analysis is required mainly for monitoring purposes. A few important assessments include:

- To check whether the water quality follows the standards, and hence, is suitable or not for the designated use.
- To monitor the efficiency of a system, working for water quality maintenance
- To check whether upgradation/change of an existing system is required and to decide what changes should take place.

2. LITERATURE SURVEY

2.1 EXISTING PROBLEM

Pesticides and fertilizers can be carried into lakes and streams by rainfall runoff or snowmelt or percolate into aquifers. Human and animal waste. Human wastes from sewage and septic systems can carry harmful microbes into drinking water sources, as can waste from animal feedlots and wildlife. Surface waters and aquifers are contaminated by various chemicals, microbes, and radionuclides. If drinking water contains unsafe levels of contaminants, it can cause health effects, such as gastrointestinal illnesses, nervous system or reproductive effects, and chronic diseases such as cancer. Factors that can influence whether a contaminant will lead to health effects include the type of contaminant, its concentration in the water, individual susceptibility, the amount of water consumed, and the duration of exposure.

2.2 REFERENCES

#LITERATURE 1

TITLE: River water quality index prediction and uncertainty analysis: A comparative study of machine learning models

AUTHOR: Seyed Babak Haji, SeyedAsadollah, AhmadSharafati, DavideMotta, Zaher Mundher Yaseen

The most used indicator for describing the quality of surface water is the Water Quality Index (WQI). Extra Tree Regression (ETR), a new ensemble machine learning model for forecasting monthly WQI values at the Lam Tsuen River in Hong Kong, is introduced in this study. The performance of the traditional standalone models, Support Vector Regression (SVR) and Decision Tree Regression, is contrasted with that of the ETR model (DTR).

For the purpose of developing the prediction models, monthly input data on the quality of the water, including Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Dissolved Oxygen (DO), Electrical Conductivity (EC), Nitrate-Nitrogen (NO3 -N), Nitrite-Nitrogen (NO2 -N), Phosphate (PO43-), potential for Hydrogen (pH), Temperature (T), and Turbidity (Using graphical comparisons and numerical indices, different input data combinations are examined and evaluated in terms of prediction performance.

The analysis demonstrates that the ETR model typically generates WQI predictions for both training and testing periods that are more accurate. Even though a combination of input parameters only containing BOD, Turbidity, and Phosphate concentration yields the second-best prediction accuracy (R2test=0.97, RMSEtest=3.74), including all ten input variables yields the highest prediction performance (R2test=0.98, RMSEtest=2.99).

The model structure and input parameter uncertainty analysis shows that the prediction results are more sensitive to the former. In terms of prediction accuracy and a reduction in the number of input parameters, the ETR model generally outperforms earlier methods for WQI prediction.

LINK: https://www.sciencedirect.com/science/article/abs/pii/S2213343720309489

#LITERATURE 2

TITLE: Predictive Modeling Approach for Surface Water Quality: Development and Comparison of Machine Learning Models

AUTHOR: Muhammad Izhar Shah, Wesam Salah Alaloul, Abdulaziz Alqahtani, Ali Aldrees, Muhammad Ali Musarat and Muhammad Faisal Javed

Human health, environmental services, and agricultural production are all at risk due to the growing global problem of water pollution. A detailed understanding of the developing concerns about water quality can be gained through the unique characteristics of artificial intelligence (AI) based modelling. The current study examines the accuracy of monthly total dissolved solids (TDS) and specific conductivity (EC) models for the upper Indus River at two outlet stations using gene expression programming (GEP), artificial neural networks (ANN), and linear regression models (LRM). 360 TDS and EC monthly records spanning 30 years of historical water quality data were utilised to train and evaluate the models. The TDS and EC modelling were associated with seven input factors based on a substantial correlation. Various performance measure indicators, error evaluation, and external criteria were all used to analyse the results. When the models' simulated results were compared to actual data, they showed a good correlation, with both TDS and EC showing correlation coefficients above 0.9.

The GEP and ANN models continued to be the effective methods for foretelling TDS and EC. The mathematical equations for the formulated GEP show its originality as compared to ANN and LRM. The results of the parametric analysis showed that the modelling procedure had taken into account the effects of all the input factors. The generalised outcome and robustness of the suggested approaches were confirmed by the external assessment standards. In conclusion, the findings of this study showed that developing AI- based models for river water quality evaluation, management, and policy making is both affordable and beneficial.

LINK: https://www.mdpi.com/2071-1050/13/14/7515

#LITERATURE 3

TITLE: Comparative Assessment of Individual and Ensemble Machine Learning Models for Efficient Analysis of River Water Quality

AUTHOR: Abdulaziz Alqahtani, Muhammad Izhar Shah, Ali Aldrees, and Muhammad Faisal Javed

The outcomes of the parametric analysis demonstrated that the modelling process had accounted for the effects of every input element. The external assessment criteria confirmed the overall result and reliability of the offered approaches. The results of this study demonstrated that creating AI-based models for river water quality assessment, management, and policy making is both feasible and advantageous.

Seven input parameters were chosen for the dataset of the projected models' training and testing on the basis of their significant association. The ensemble RF model was optimised by creating 20 sub-models and selecting the most accurate one. Known statistical measures including the coefficient of determination (R2), mean absolute error (MAE), root mean squared error (RMSE), and Nash-Sutcliffe efficiency were used to evaluate the models' goodness-of-fit (NSE). The R2 value for the GEP, RF, and ANN models, respectively, was found to be 0.96, 0.98, and 0.92, demonstrating a significant correlation between inputs and modeling outputs. The comparative effectiveness of the suggested methodologies demonstrated the RF's relative superiority over GEP and ANN. The most accurate model among the 20 RF sub-models produced R2 values of 0.941 and 0.938, with 70 and 160 numbers corresponding estimators. On training and testing data, respectively, the ensemble RF model produced the lowest RMSE values of 1.37 and 3.1.

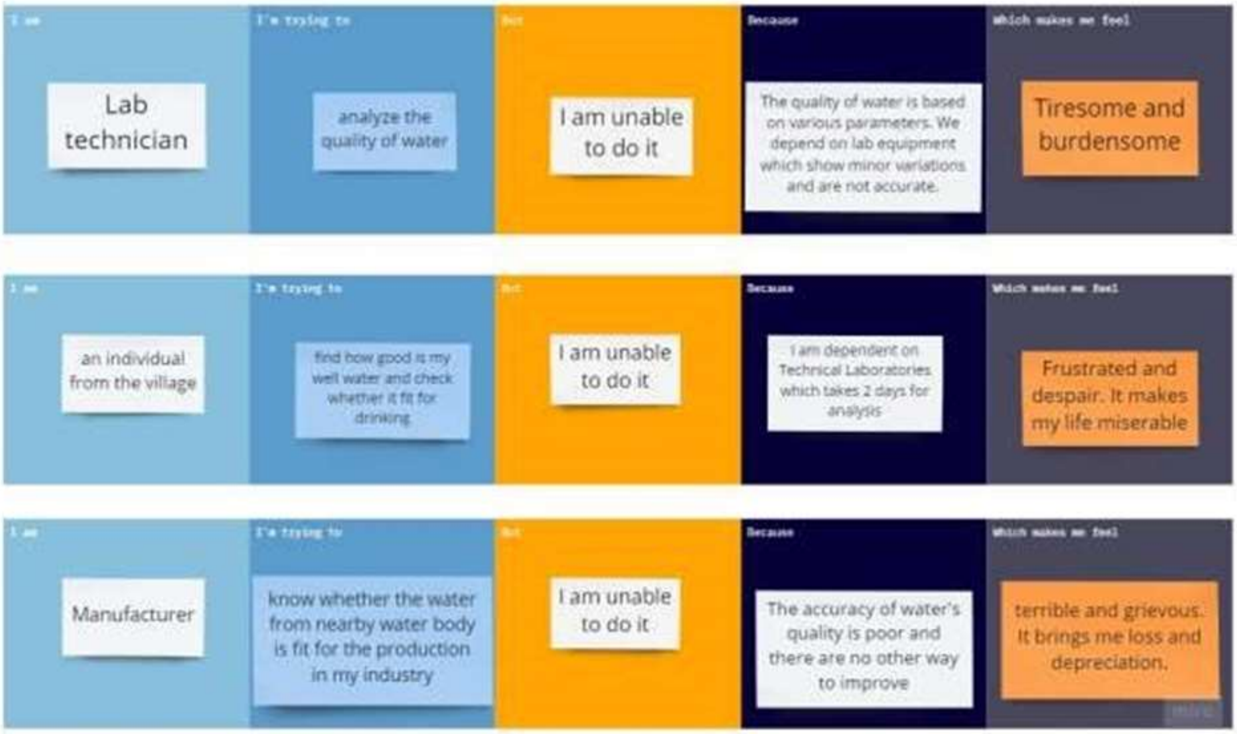
The generalized outcomes of all the procedures were ensured by the models' evaluation of outside criteria. As a result of the current study, it was concluded that the RF model with a few essential parameters might be prioritized for water quality assessment and management.

LINK: https://www.mdpi.com/2071-1050/14/3/1183

2.3 Problem Statement Definition

The main aim of the project is to analyze the quality of water based on various parameters such as Dissolved oxygen, BOD, Conductivity, nitrate level, and coliform level. Different algorithms are executed to obtain maximum accuracy and the best algorithm is chosen. The ML model obtained is saved and integrated with a flask application. The application is then deployed in IBM Watson.

PROBLEM STATEMENTS

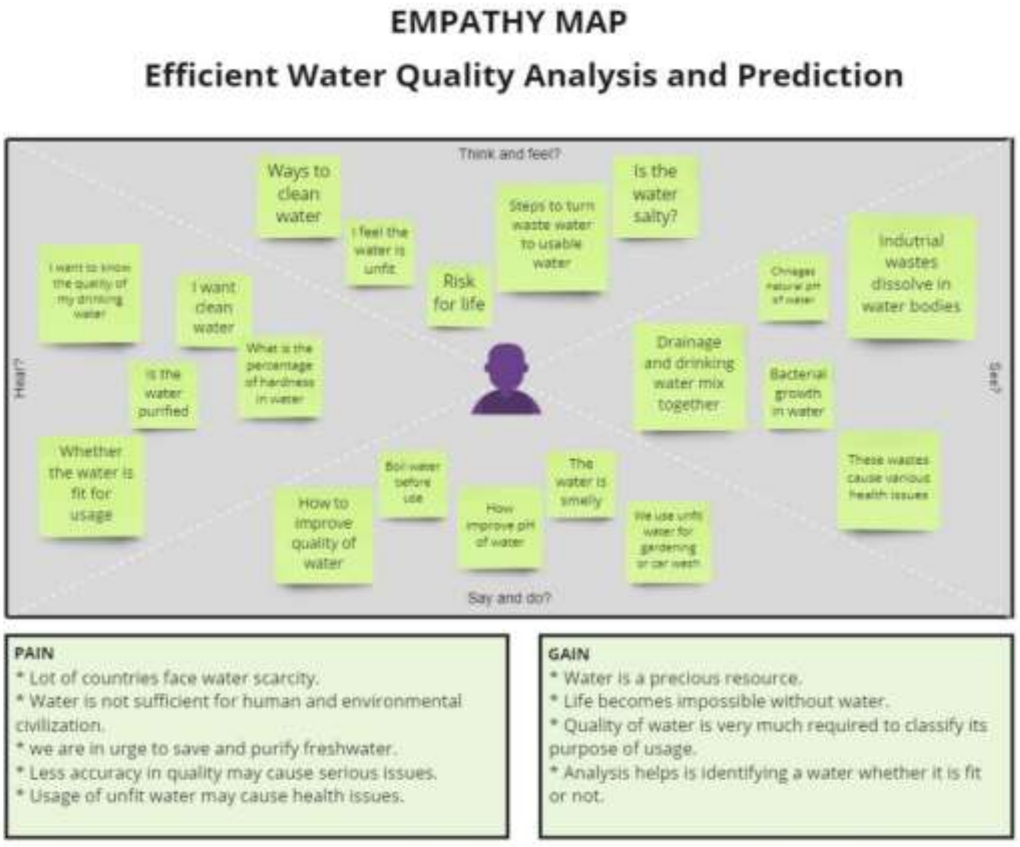


I am	I'm trying to	But	Because	Which makes me feel
Lab technician	analyze the quality of water	I am unable to do it	The quality of water is based on various parameters. We depend on lab equipment which show minor variations and are not accurate.	Tiresome and burdensome
an individual from the village	find how good is my well water and check whether it fit for drinking	I am unable to do it	I am dependent on Technical Laboratories which takes 2 days for analysis	Frustrated and despair. It makes my life miserable
Manufacturer	know whether the water from nearby water body is fit for the production in my industry	I am unable to do it	The accuracy of water's quality is poor and there are no other way to improve	terrible and grievous. It brings me loss and depreciation.

Problem Statement (PS)	I am (Customer)	I'm trying to	But	Because	Which makes me feel
PS-1	Lab technician	analyse the quality of water	I am unable to do it	The quality of water is based on various parameters. We depend on lab equipment which show minor variations and is not accurate.	Tiresome and burdensome
PS-2	an individual from the village	find how good is my well water and check whether it fit for drinking	I am unable to do it	I am dependent on Technical Laboratories which takes 2 days for analysis	Frustrated and despair. It makes my life miserable
PS-3	Manufacturer	know whether the water from nearby water body is fit for the production in my industry	I am unable to do it	The accuracy of water's quality is poor and there are no other way to improve	terrible and grievous. It brings me loss and depreciation.

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas



EMPATHY MAP

Efficient Water Quality Analysis and Prediction

Think and feel?
Ways to clean water
I feel the water is unfit
Risk for life
Steps to turn waste water to usable water
Is the water salty?
Hear?
I want to know the quality of my drinking water
I want clean water
Is the water purified
What is the percentage of hardness in water
Whether the water is fit for usage
See?
Chnages natural pH of water
Indutrial wastes dissolve in water bodies
Drainage and drinking water mix together
Bacterial growth in water
These wastes cause various health issues
Say and do?
How to improve quality of water
Boil water before use
How improve pH of water
The water is smelly
We use unfit water for gardening or car wash

PAIN

* Lot of countries face water scarcity.
* Water is not sufficient for human and environmental civilization.
* we are in urge to save and purify freshwater.
* Less accuracy in quality may cause serious issues.
* Usage of unfit water may cause health issues.

GAIN

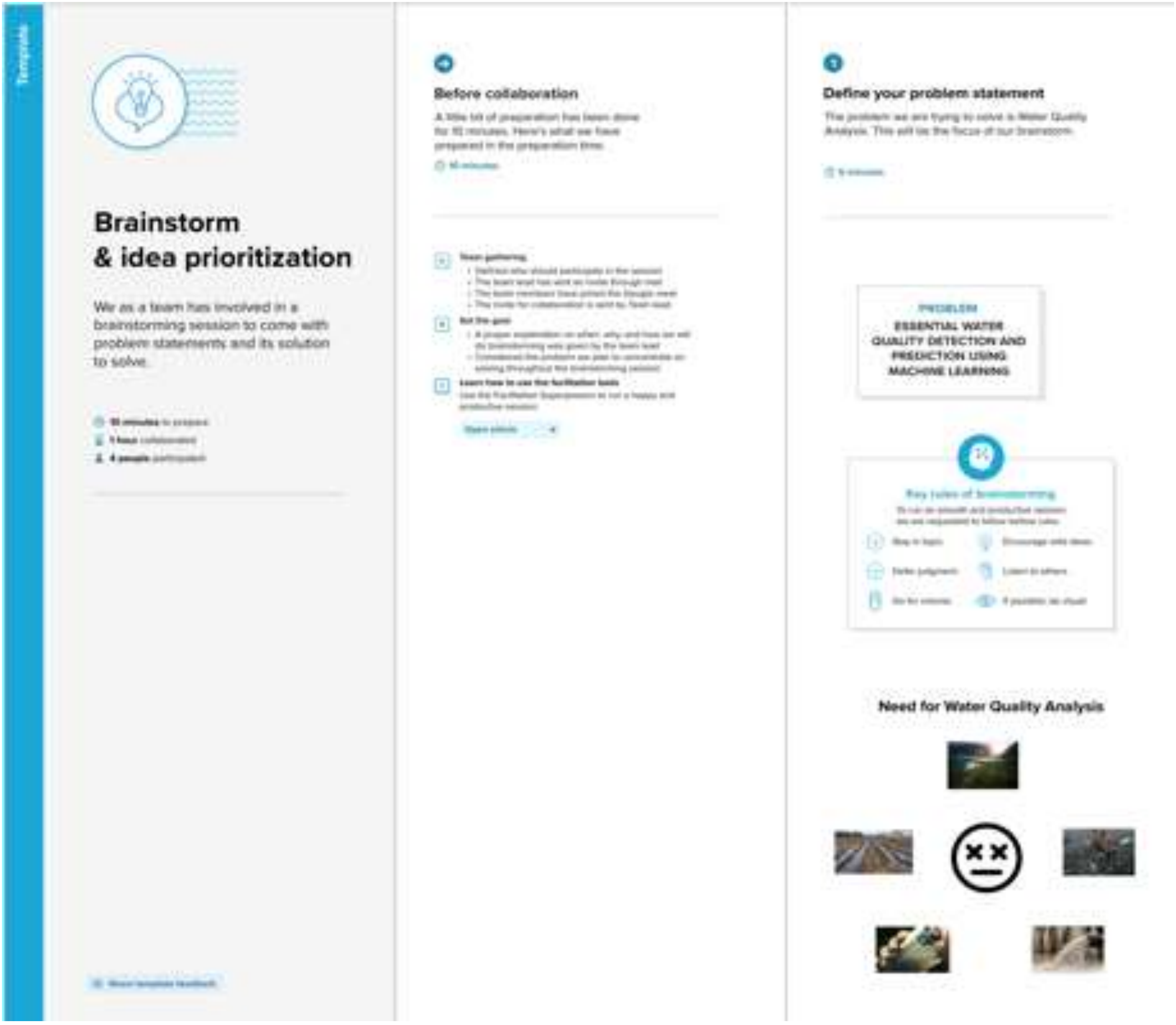
* Water is a precious resource.
* Life becomes impossible without water.
* Quality of water is very much required to classify its purpose of usage.
* Analysis helps is identifying a water whether it is fit or not.

3.2 IDEATION & 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 welcome and built upon, and all participants are encouraged to collaborate, helping each other develop a rich number of creative solutions.

Brainstorming steps

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



Template

Brainstorm & idea prioritization

We as a team has involved in a brainstorming session to come with problem statements and its solution to solve.

- 10 minutes to prepare
- 1 hour collaborated
- 4 people participated

Share template feedback

Before collaboration

A little bit of preparation has been done for 10 minutes. Here's what we have prepared in the preparation time.

10 minutes

☐ **Team gathering**
- Defined who should participate in the session
- The team lead has sent an invite through mail
- The team members have joined the Google meet
- The invite for collaboration is sent by Team lead

☐ **Set the goal**
- A proper explanation on what, why and how we will do brainstorming was given by the team lead
- Considered the problem we plan to concentrate on solving throughout the brainstorming session

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

Open article →

1 Define your problem statement

The problem we are trying to solve is Water Quality Analysis. This will be the focus of our brainstorm.

5 minutes

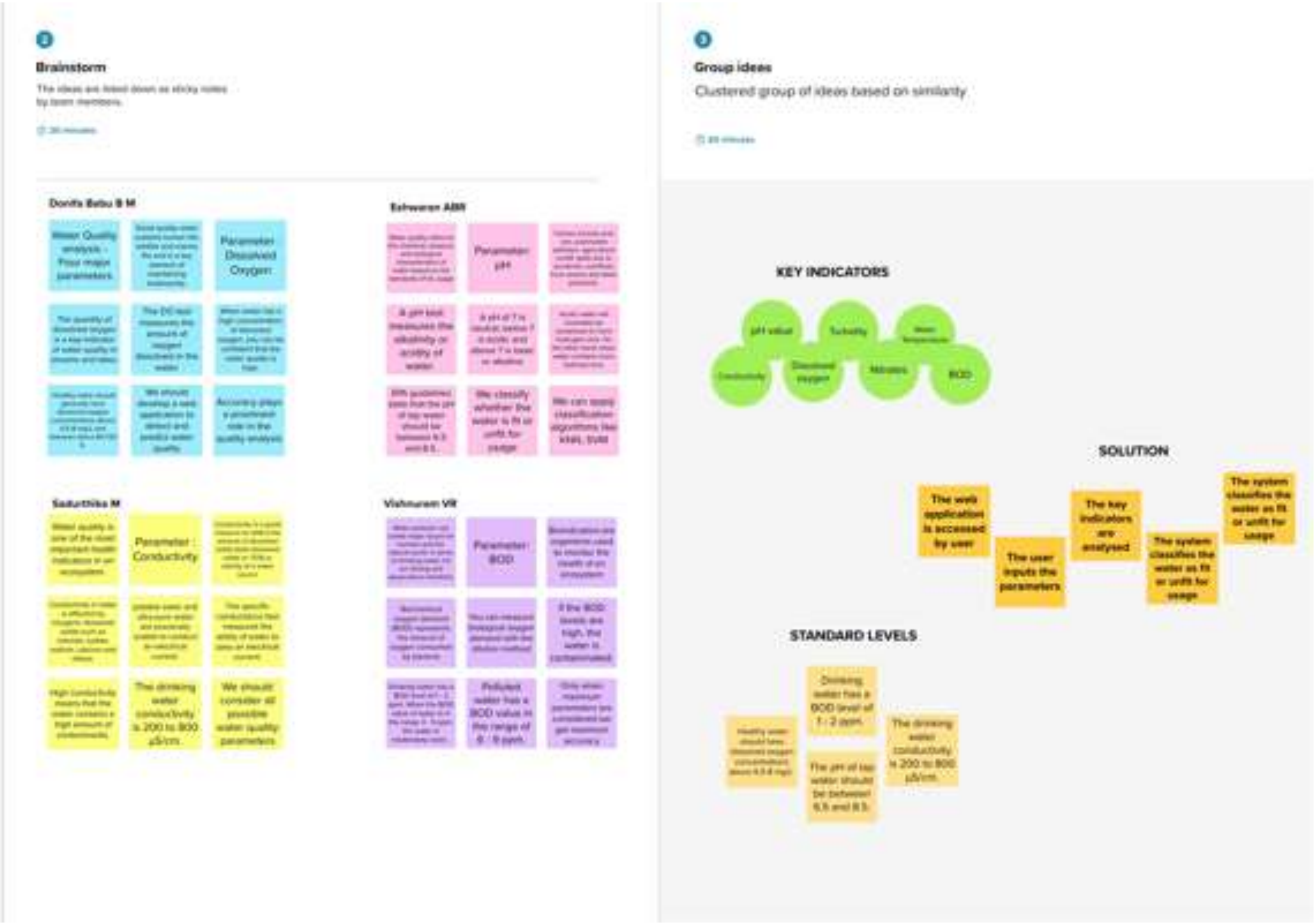
PROBLEM
ESSENTIAL WATER QUALITY DETECTION AND PREDICTION USING MACHINE LEARNING

Key rules of brainstorming
To run an smooth and productive session we are requested to follow below rules.

Stay in topic.	Encourage wild ideas.
Defer judgment.	Listen to others.
Go for volume.	If possible, be visual.

Need for Water Quality Analysis

Step 2: Brainstorm, Idea Listing, and Grouping



2

Brainstorm

The ideas are listed down as sticky notes by team members.

20 minutes

Donifa Babu B M

- Water Quality analysis - Four major parameters
- [illegible]
- Parameter: Dissolved Oxygen
- The quantity of dissolved oxygen is a key indicator of water quality in streams and lakes
- The DO test measures the amount of oxygen dissolved in the water
- When water has a high concentration of dissolved oxygen, you can be confident that the water quality is high
- Healthy water should generally have dissolved oxygen concentrations above 6.5-8 mg/L and between about 80-120 %
- We should develop a web application to detect and predict water quality
- Accuracy plays a prominent role in the quality analysis

Eshwaran ABR

- Water quality refers to the chemical, physical and biological characteristics of water based on the standards of its usage
- Parameter: pH
- [illegible]
- A pH test measures the alkalinity or acidity of water
- A pH of 7 is neutral, below 7 is acidic and above 7 is basic or alkaline
- [illegible]
- WHO guidelines state that the pH of tap water should be between 6.5 and 8.5.
- We classify whether the water is fit or unfit for usage
- We can apply classification algorithms like KNN, SVM

Sadurthika M

- Water quality is one of the most important health indicators in an ecosystem.
- Parameter : Conductivity
- [illegible]
- Conductivity in water is affected by inorganic dissolved solids such as chloride, sulfate, sodium, calcium and others
- potable water and ultra-pure water are practically unable to conduct an electrical current
- The specific conductance test measures the ability of water to pass an electrical current
- High conductivity means that the water contains a high amount of contaminants
- The drinking water conductivity is 200 to 800 µS/cm
- We should consider all possible water quality parameters

Vishnuram VR

- [illegible]
- Parameter : BOD
- Bioindicators are organisms used to monitor the health of an ecosystem
- Biochemical oxygen demand (BOD) represents the amount of oxygen consumed by bacteria
- You can measure biological oxygen demand with the dilution method
- If the BOD levels are high, the water is contaminated
- Drinking water has a BOD level of 1 - 2 ppm. When the BOD value of water is in the range 3 - 5 ppm the water is moderately clean.
- Polluted water has a BOD value in the range of 6 - 9 ppm.
- Only when maximum parameters are considered we get maximum accuracy

3

Group ideas

Clustered group of ideas based on similarity

20 minutes

KEY INDICATORS

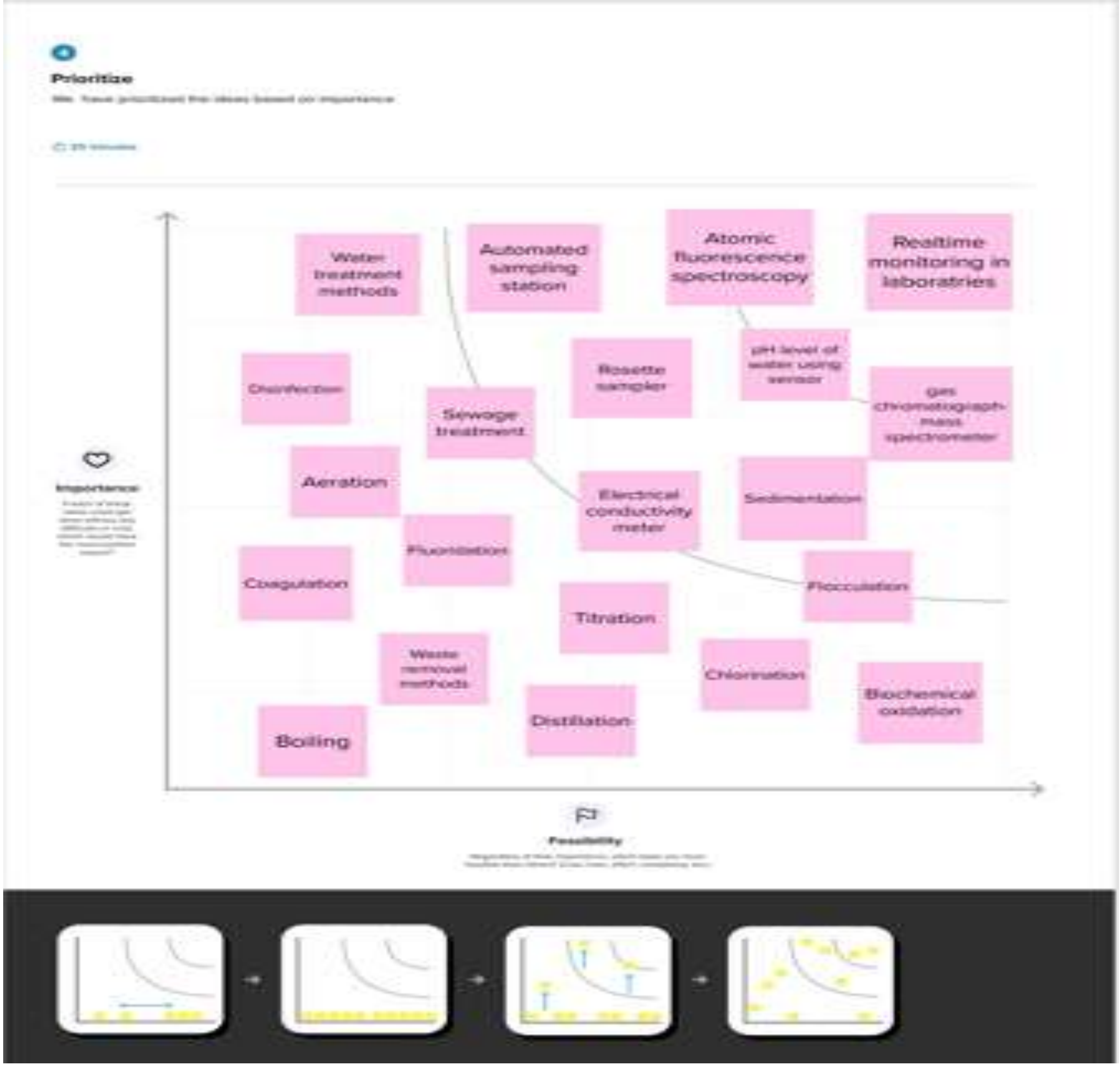
SOLUTION

- The web application is accessed by user
- The user inputs the parameters
- The key indicators are analysed
- The system classifies the water as fit or unfit for usage
- The system classifies the water as fit or unfit for usage

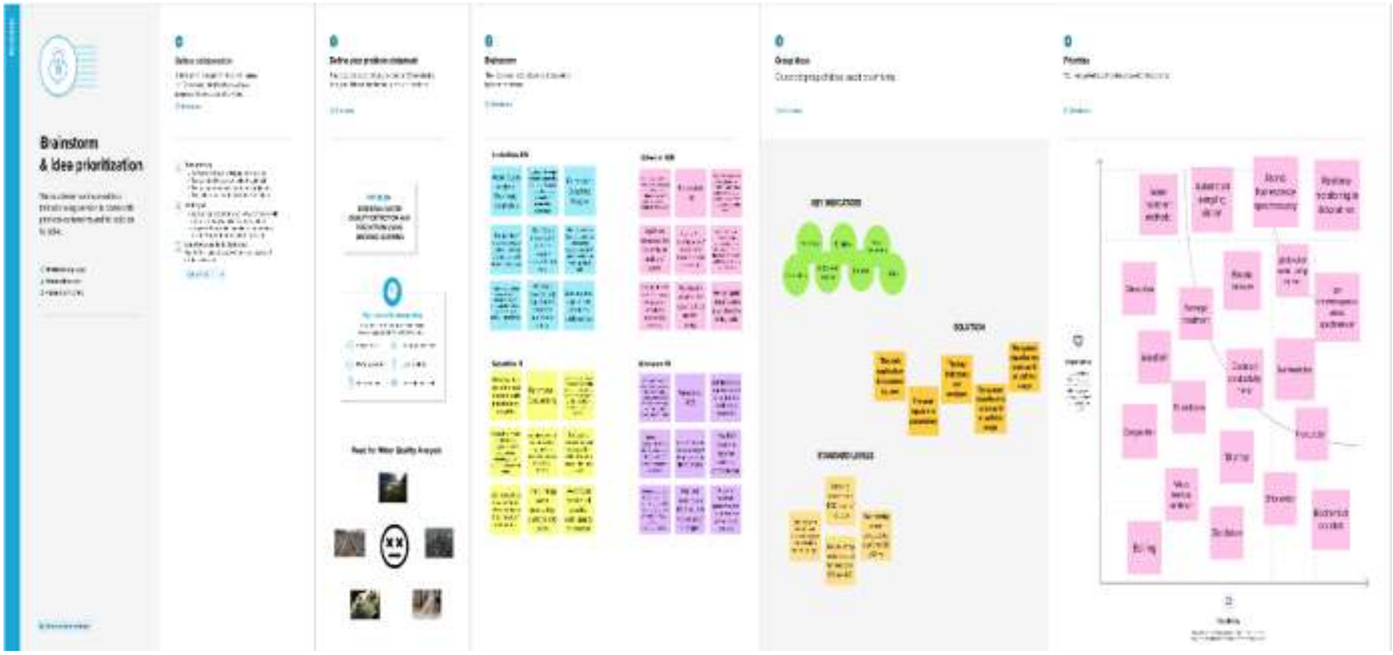
STANDARD LEVELS

- Healthy water should have dissolved oxygen concentrations above 6.5-8 mg/L
- Drinking water has a BOD level of 1 - 2 ppm.
- The pH of tap water should be between 6.5 and 8.5.
- The drinking water conductivity is 200 to 800 µS/cm.

Step 3: Idea Prioritization


Prioritize
We have prioritized the ideas based on importance
Water treatment methods
Automated sampling station
Atomic fluorescence spectroscopy
Realtime monitoring in laboratries
Disinfection
Rosette sampler
pH level of water using sensor
gas chromatograph-mass spectrometer
Sewage treatment
Importance
Aeration
Electrical conductivity meter
Sedimentation
Fluoridation
Coagulation
Flocculation
Titration
Waste removal methods
Chlorination
Biochemical oxidation
Distillation
Boiling
Feasibility

Overall Brainstorming And Idea Prioritization



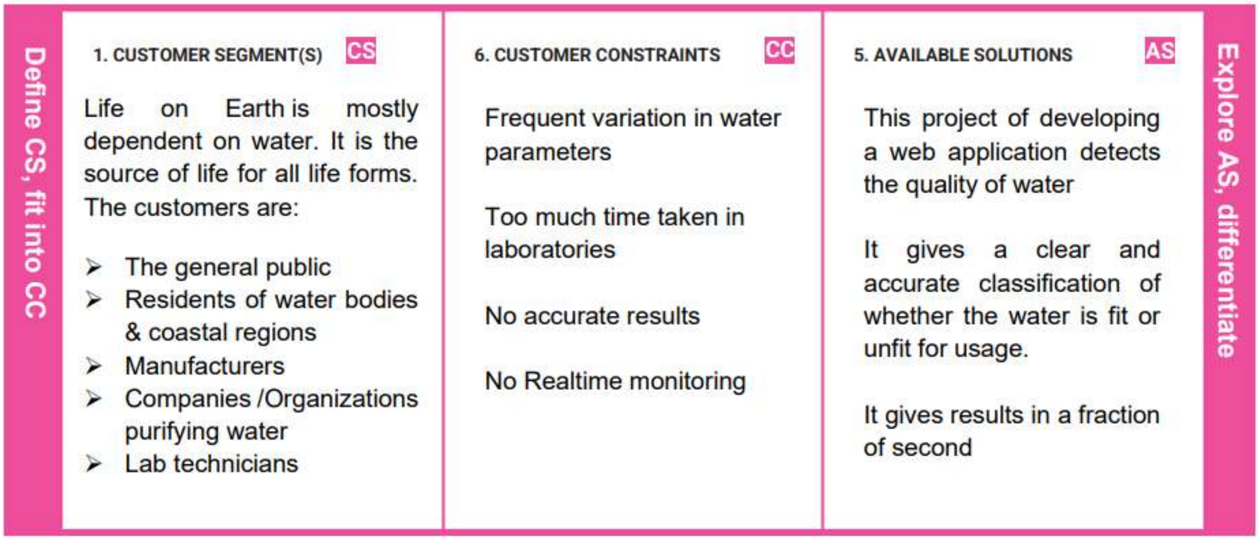
3.3 PROPOSED SOLUTION

S. No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	WATER QUALITY DETECTION AND PREDICTION Around the world, 80% of wastewater is dumped back into the environment, mostly untreated, damaging rivers, lakes, and seas. Our health is in danger as a result of this widespread issue with water pollution. More people die each year from unsafe water than from war and all other types of violence combined. Access to clean drinking water is crucial for health, a fundamental human right, and a component of successful health protection policies. On a national, regional, and local level, this is significant as a health and development issue. Investments in water supply and sanitation have been shown to produce a net economic advantage in some areas because they reduce negative health effects and medical expenses more than they cost to implement.

		Water can be utilized for drinking, cleaning, fishing, farming, industry, and enjoyment. Different defined chemical, physical, and biological standards are required to support each of these designated purposes. For instance, water used for swimming or drinking has higher criteria than water used for business or agriculture. Based on purpose, the properties of water may vary. The chemical, physical, and biological properties of water are referred to as its quality.
2.	Idea / Solution description	The Quality of water is predicted and based on it, water can be treated and used for each purpose. Water Quality prediction plays a major role in categorizing the water to be fit or unfit for usage.
3.	Novelty / Uniqueness	**#NOVELTY 1** Major parameters affecting quality such as • Dissolved Oxygen • pH • Temperature • Conductivity • BOD are considered for analysis **#NOVELTY 2** Web application with simple user interface does not exist
4.	Social Impact / Customer Satisfaction	All forms of life on earth are more significantly impacted by this idea. To continuously evaluate the quality of the water coming from the various supply sources, both qualitative and quantitative measurements are occasionally required. The sophistication of their consumers with regard to water quality data and laboratory tests may or may not be known to dealers. They are specifically interested in finding out if the equipment they choose will successfully remove pollutants that are known to be present in their drinking water. Product "certifications" are excellent, but they can also be perplexing, particularly when numerous businesses

		make promises without any supporting scientific data. The user can more easily determine the quality of his water with the aid of this application. With the stroke of a button, detailed reports on domestic water quality down to the user's system can be generated. Users can search and analyse local water quality data more precisely than ever before.
5.	Business Model (Revenue Model)	In general, ISO-certified laboratories should sample and analyse water. Only trained technicians should conduct sampling and monitoring tests. However, it is a laborious process that could take up to two days to provide an accurate report. This application would provide the result more precisely and in a fraction of a second. Consequently, the procedure of analysing the water quality is made simpler, and lab personnel's lives are made simpler.
6.	Scalability of the Solution	Precisely defined parameters are crucial to the model's accuracy. The quality assurance of your analytical result will help you gain or maintain user trust. The ability of your web application to handle an increasing number of people interacting with the app at once is known as scalability. Scalability is highly impacted by the quality of the code as well. The entire cycle of developing a web application must include testing. A smooth growth of your application is guaranteed by proper load and performance testing.

3.4 PROBLEM SOLUTION FIT



Define CS, fit into CC

1. CUSTOMER SEGMENT(S) CS

Life on Earth is mostly dependent on water. It is the source of life for all life forms. The customers are:

- The general public
- Residents of water bodies & coastal regions
- Manufacturers
- Companies /Organizations purifying water
- Lab technicians

6. CUSTOMER CONSTRAINTS CC

Frequent variation in water parameters

Too much time taken in laboratories

No accurate results

No Realtime monitoring

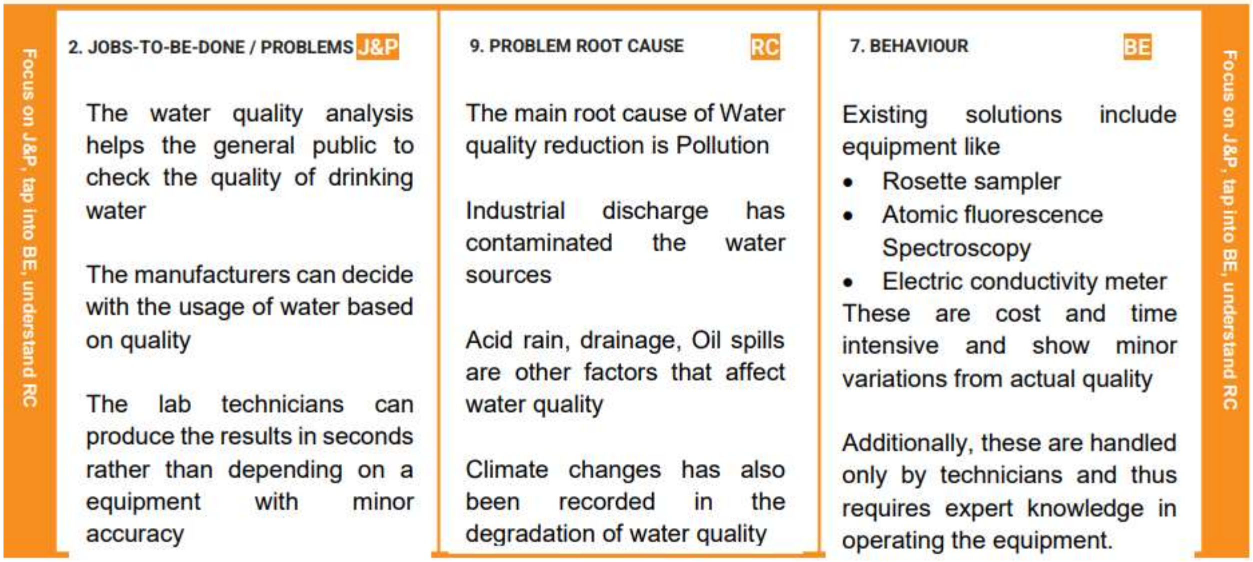
5. AVAILABLE SOLUTIONS AS

This project of developing a web application detects the quality of water

It gives a clear and accurate classification of whether the water is fit or unfit for usage.

It gives results in a fraction of second

Explore AS, differentiate



Focus on J&P, tap into BE, understand RC

2. JOBS-TO-BE-DONE / PROBLEMS J&P

The water quality analysis helps the general public to check the quality of drinking water

The manufacturers can decide with the usage of water based on quality

The lab technicians can produce the results in seconds rather than depending on a equipment with minor accuracy

9. PROBLEM ROOT CAUSE RC

The main root cause of Water quality reduction is Pollution

Industrial discharge has contaminated the water sources

Acid rain, drainage, Oil spills are other factors that affect water quality

Climate changes has also been recorded in the degradation of water quality

7. BEHAVIOUR BE

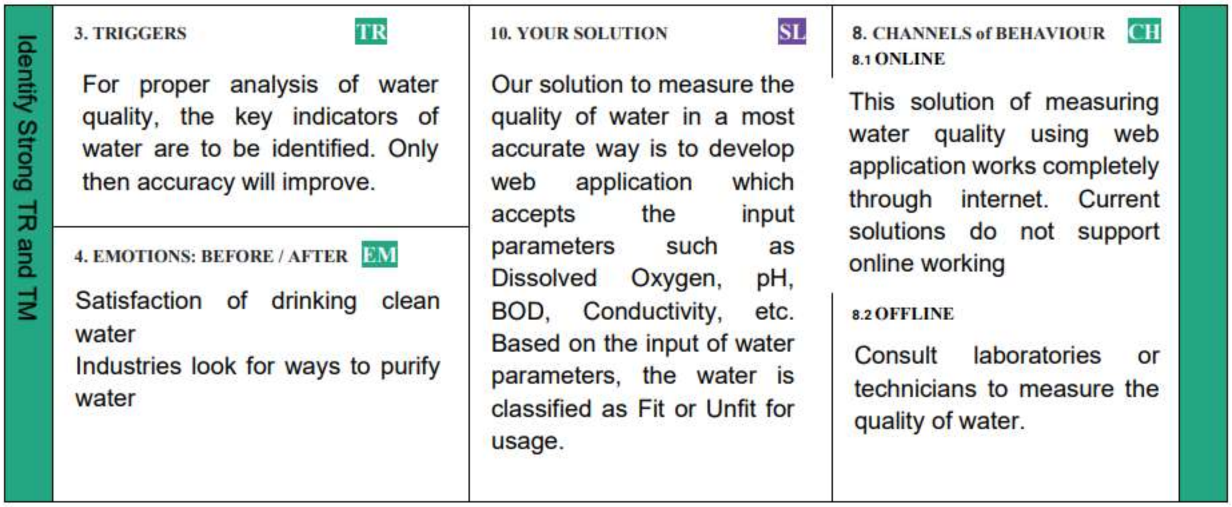
Existing solutions include equipment like

- Rosette sampler
- Atomic fluorescence Spectroscopy
- Electric conductivity meter

These are cost and time intensive and show minor variations from actual quality

Additionally, these are handled only by technicians and thus requires expert knowledge in operating the equipment.

Focus on J&P, tap into BE, understand RC



Identify Strong TR and TM

3. TRIGGERS TR

For proper analysis of water quality, the key indicators of water are to be identified. Only then accuracy will improve.

4. EMOTIONS: BEFORE / AFTER EM

Satisfaction of drinking clean water
Industries look for ways to purify water

10. YOUR SOLUTION SL

Our solution to measure the quality of water in a most accurate way is to develop web application which accepts the input parameters such as Dissolved Oxygen, pH, BOD, Conductivity, etc. Based on the input of water parameters, the water is classified as Fit or Unfit for usage.

8. CHANNELS of BEHAVIOUR CH

8.1 ONLINE

This solution of measuring water quality using web application works completely through internet. Current solutions do not support online working

8.2 OFFLINE

Consult laboratories or technicians to measure the quality of water.

4. REQUIREMENT ANALYSIS

4.1 FUNCTIONAL REQUIREMENT

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	Training	dataset to train the model
FR-2	Detection and Prediction	Selection of appropriate parameters Clustering based on similarity Classification as fit or unfit for usage
FR-3	test and output	Based on the parameters and the trained model, the result is displayed

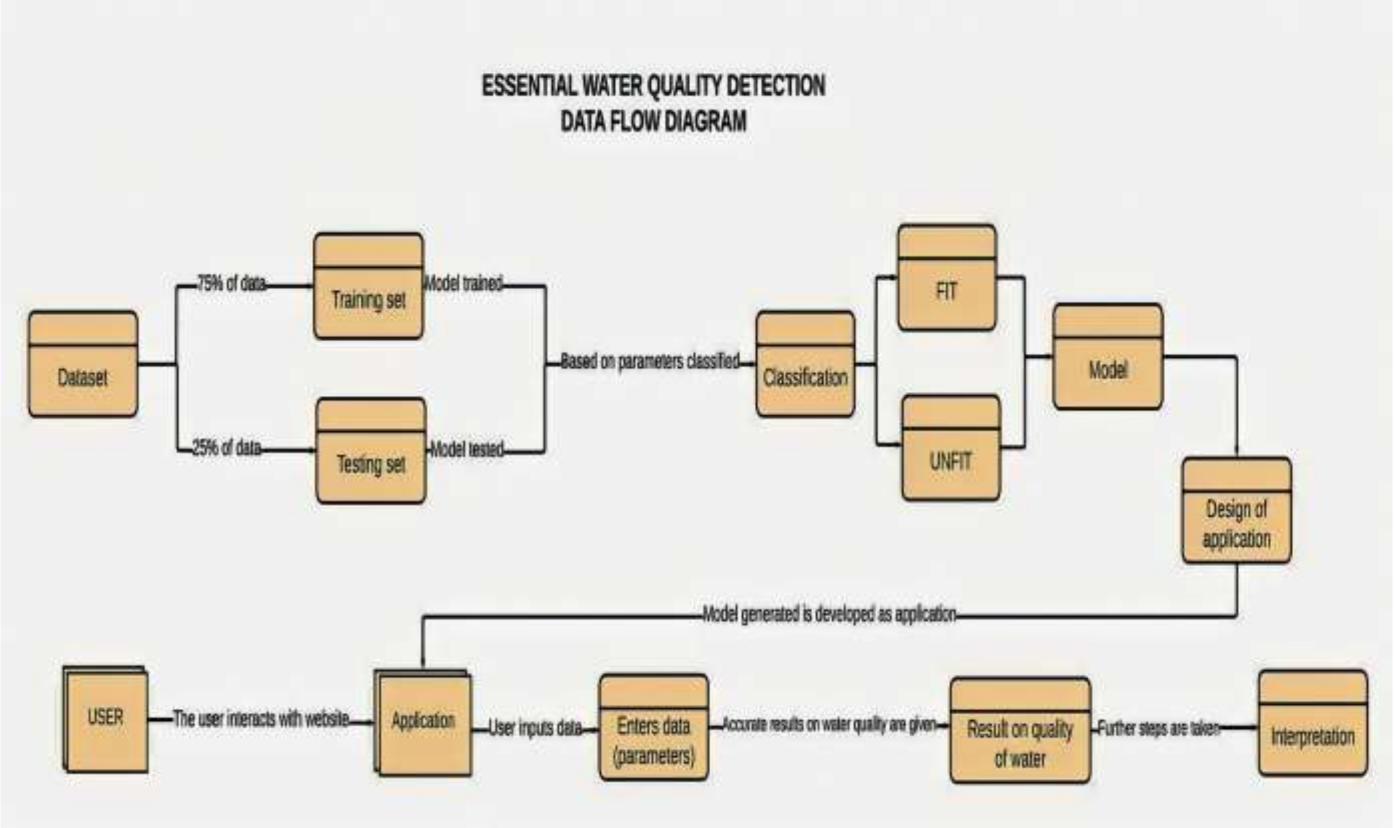
4.2 NON-FUNCTIONAL REQUIREMENTS

FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	Users should feel comfortable. All the pages on a website or mobile application should have the same look and feel. Navigations and UI should be simple and easy to use.
NFR-2	**Security**	The user credentials and their login information should be kept confidential.
NFR-3	**Reliability**	Our system needs to be operational at least 99.5% of the time in order to respond to requests from websites and mobile devices. Fault tolerance should be high.

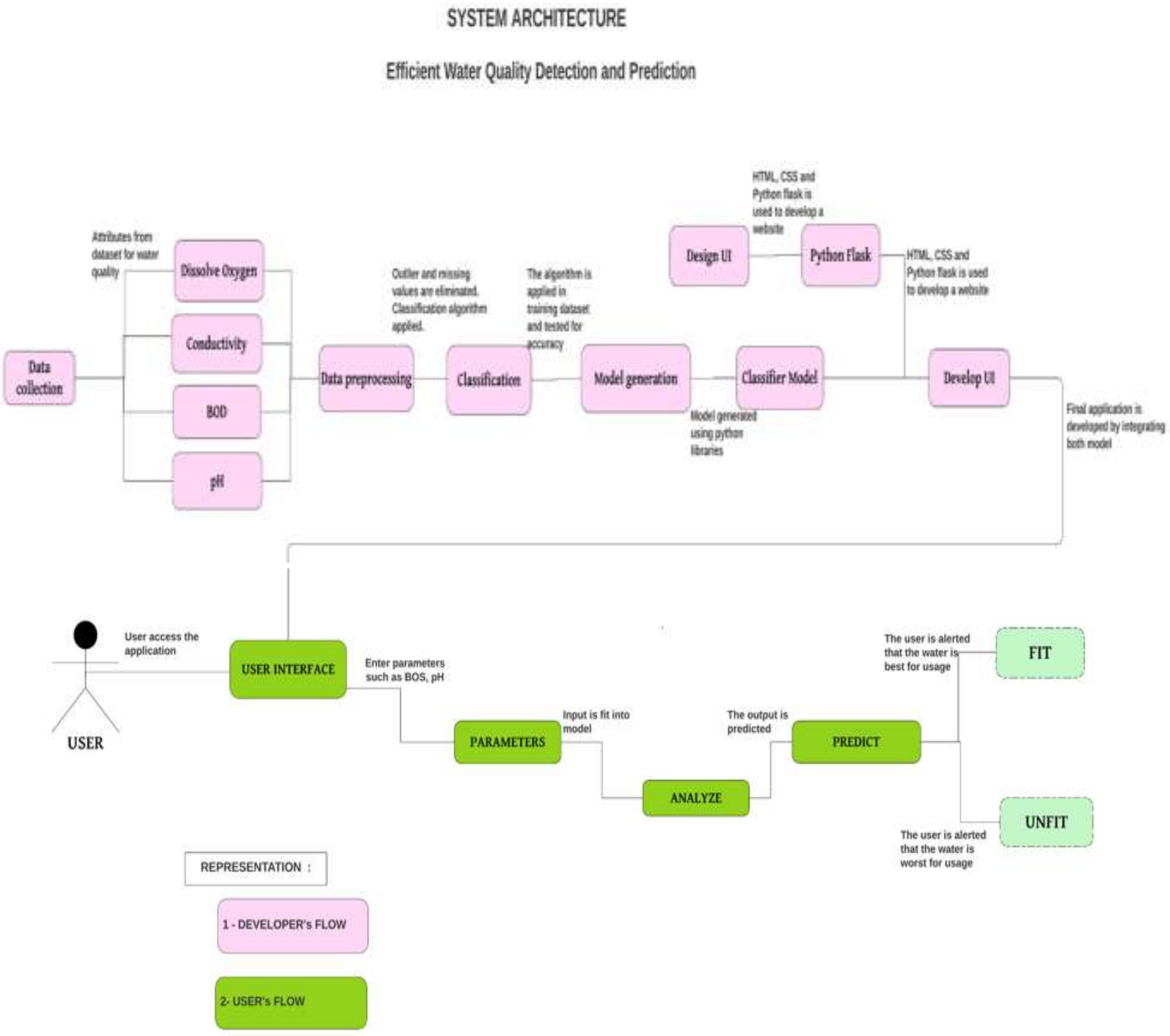
NFR-4	**Performance**	Visitors expect a good-looking website that is easy to use and loads fast. Hence the application should accept several requests and respond in fraction of a second.
NFR-5	**Availability**	The system's monitoring and maintenance should be fundamentally focused. It should not be the case that there are too many jobs running on several machines, making it difficult to monitor if they are uninterrupted
NFR-6	**Scalability**	Maintaining multiple users' data, accuracy of results, data transmission rate, and increase or decrease of storage are monitored.

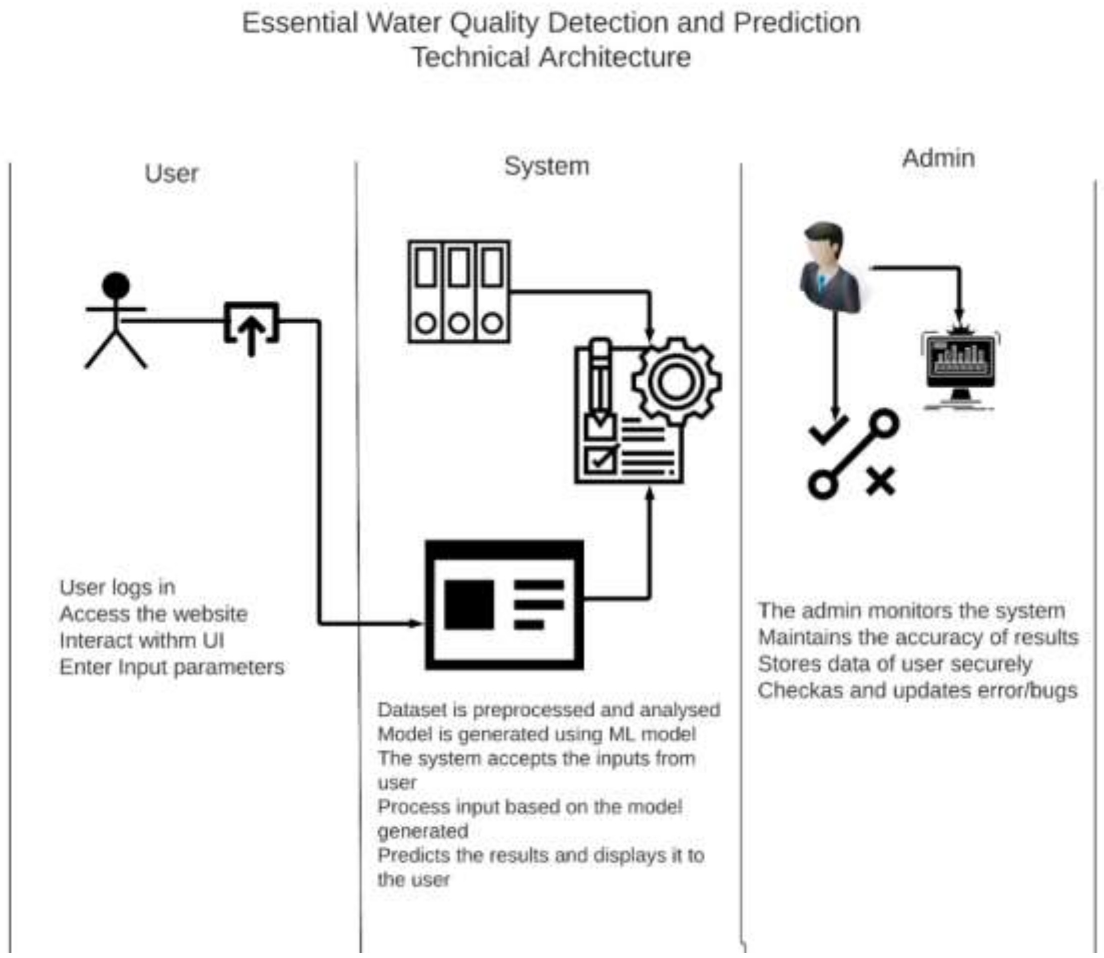
5. PROJECT DESIGN

5.1 DATA FLOW DIAGRAMS



5.2 SOLUTION & TECHNICAL ARCHITECTURE


SYSTEM ARCHITECTURE
Efficient Water Quality Detection and Prediction
Data collection
Attributes from dataset for water quality
Dissolve Oxygen
Conductivity
BOD
pH
Outlier and missing values are eliminated. Classification algorithm applied.
Data preprocessing
The algorithm is applied in training dataset and tested for accuracy
Classification
Model generation
Model generated using python libraries
Classifier Model
Design UI
HTML, CSS and Python flask is used to develop a website
Python Flask
HTML, CSS and Python flask is used to develop a website
Develop UI
Final application is developed by integrating both model
USER
User access the application
USER INTERFACE
Enter parameters such as BOS, pH
PARAMETERS
Input is fit into model
ANALYZE
The output is predicted
PREDICT
The user is alerted that the water is best for usage
FIT
UNFIT
The user is alerted that the water is worst for usage
REPRESENTATION :
1 - DEVELOPER's FLOW
2- USER's FLOW



Essential Water Quality Detection and Prediction
Technical Architecture

User logs in
Access the website
Interact withm UI
Enter Input parameters

Dataset is preprocessed and analysed
Model is generated using ML model
The system accepts the inputs from user
Process input based on the model generated
Predicts the results and displays it to the user

The admin monitors the system
Maintains the accuracy of results
Stores data of user securely
Checkas and updates error/bugs

Table-1 : Components & Technologies:

S.No	Component	Description	Technology
1.	User Interface	The user interacts with the website/ application through web UI	HTML, CSS, Python
2.	Application Logic-1	Logic for a process in the application	Python
3.	Application Logic-2	Logic for detection of water quality in the application	Python
4.	Application Logic-3	Logic for prediction of water quality in the application	Python
5.	Database	Data format for processing	MySQL
6.	Cloud Database	Database Service on Cloud	IBM DB2.

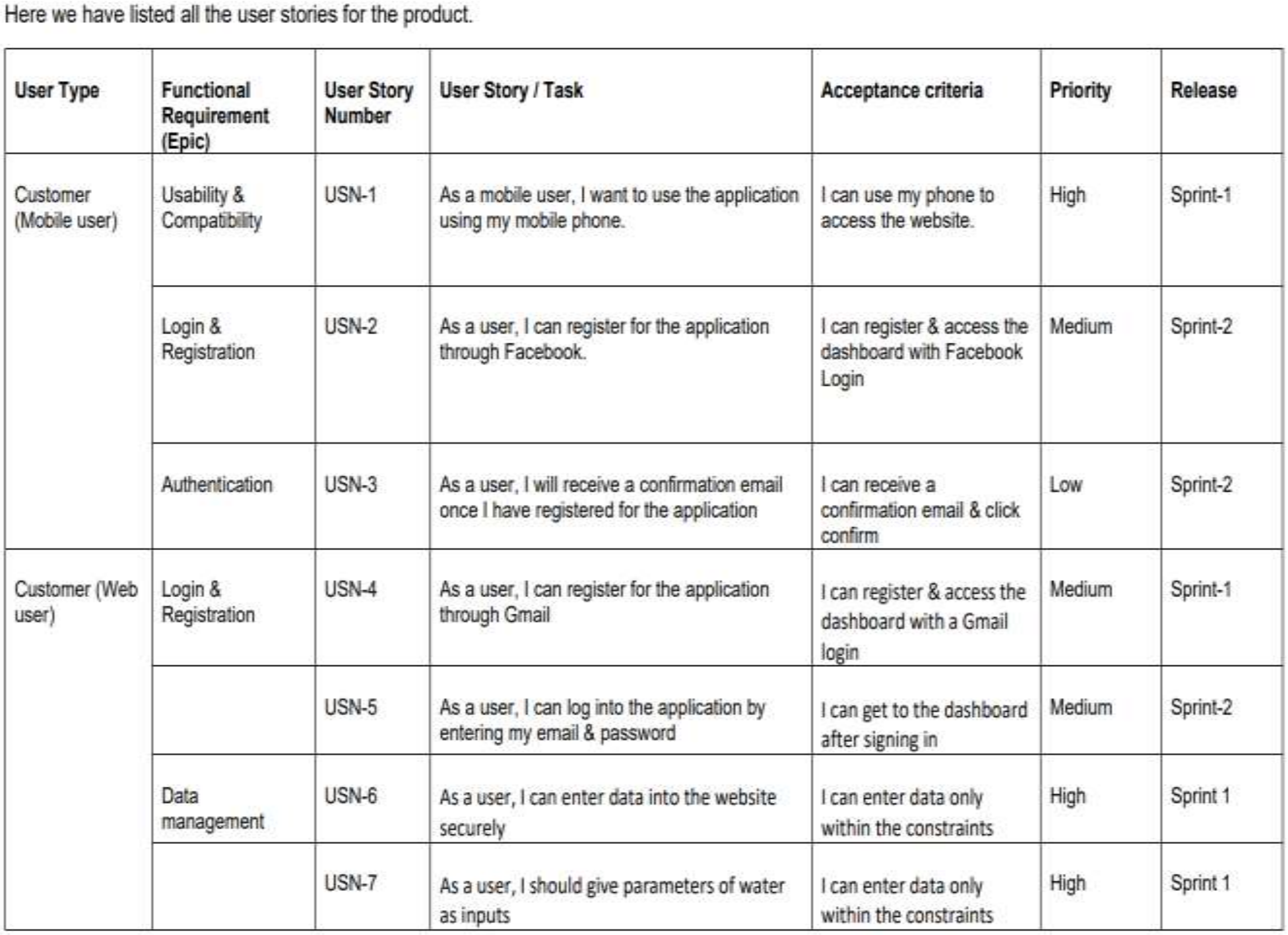
7.	File Storage	To store files/ data for the process	Local Filesystem
8.	Machine Learning Model	Classification and clustering are techniques used in data mining to analyze collected data. Classification is used to label data, while clustering is used to group similar data instances together.	Clustering and classification Model.
9.	Infrastructure (Server / Cloud)	Application Deployment on Local System / Cloud Local Server Configuration: built-in flask web server	Local web server

Table-2: Application Characteristics:

S.No	Characteristics	Description	Technology
1.	Open-Source Frameworks	Micro web framework written in python	Flask
2.	Security Implementations	Flask-Security allows you to quickly add common security mechanisms to your Flask application. They include Session-based authentication, Role management.	Flask Security & Validation
3.	Scalable Architecture	Flask is also highly scalable as it can process a high number of requests each day. This micro-framework modularize the entire code and let developers work on independent chunks and use them as the code base grows.	Flask
4.	Availability	High compatibility with the latest technologies and allows customization.	Flask
5.	Performance	➔ Integrated support for unit testing. ➔ RESTful request dispatching. ➔ Uses Jinja templating.	Flask

S.No	Characteristics	Description	Technology
		➔ Support for secure cookies (client-side sessions) 100% WSGI 1.0 compliant.	

5.3 USER STORIES

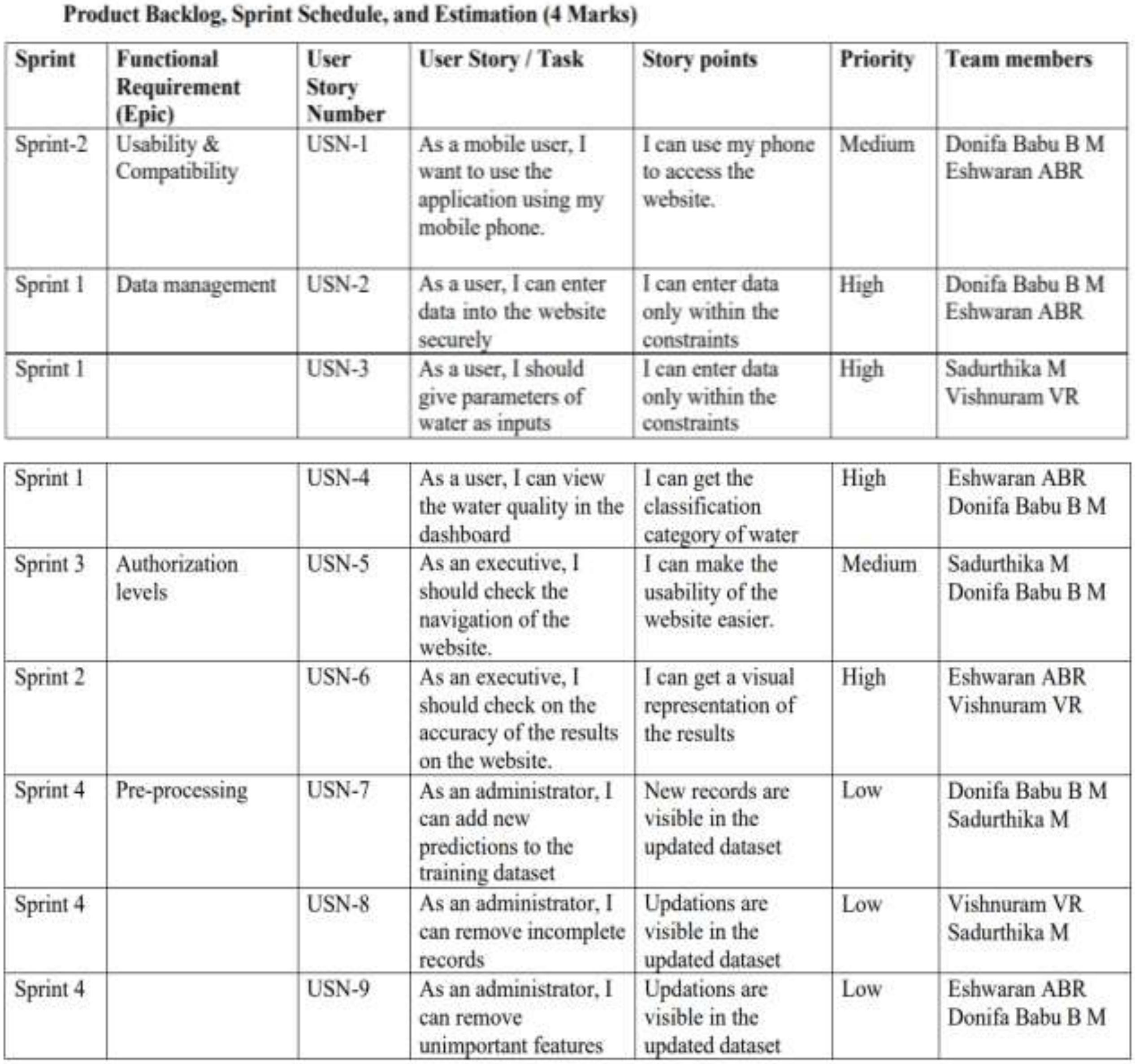


Here we have listed all the user stories for the product.

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Mobile user)	Usability & Compatibility	USN-1	As a mobile user, I want to use the application using my mobile phone.	I can use my phone to access the website.	High	Sprint-1
	Login & Registration	USN-2	As a user, I can register for the application through Facebook.	I can register & access the dashboard with Facebook Login	Medium	Sprint-2
	Authentication	USN-3	As a user, I will receive a confirmation email once I have registered for the application	I can receive a confirmation email & click confirm	Low	Sprint-2
Customer (Web user)	Login & Registration	USN-4	As a user, I can register for the application through Gmail	I can register & access the dashboard with a Gmail login	Medium	Sprint-1
		USN-5	As a user, I can log into the application by entering my email & password	I can get to the dashboard after signing in	Medium	Sprint-2
	Data management	USN-6	As a user, I can enter data into the website securely	I can enter data only within the constraints	High	Sprint 1
		USN-7	As a user, I should give parameters of water as inputs	I can enter data only within the constraints	High	Sprint 1

6. PROJECT PLANNING & SCHEDULING

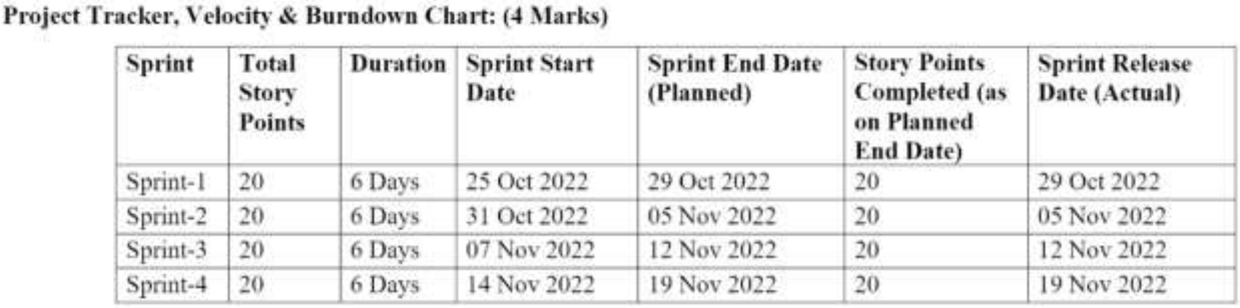
6.1 Sprint Planning & Estimation



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

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story points	Priority	Team members
Sprint-2	Usability & Compatibility	USN-1	As a mobile user, I want to use the application using my mobile phone.	I can use my phone to access the website.	Medium	Donifa Babu B M Eshwaran ABR
Sprint 1	Data management	USN-2	As a user, I can enter data into the website securely	I can enter data only within the constraints	High	Donifa Babu B M Eshwaran ABR
Sprint 1		USN-3	As a user, I should give parameters of water as inputs	I can enter data only within the constraints	High	Sadurthika M Vishnuram VR
Sprint 1		USN-4	As a user, I can view the water quality in the dashboard	I can get the classification category of water	High	Eshwaran ABR Donifa Babu B M
Sprint 3	Authorization levels	USN-5	As an executive, I should check the navigation of the website.	I can make the usability of the website easier.	Medium	Sadurthika M Donifa Babu B M
Sprint 2		USN-6	As an executive, I should check on the accuracy of the results on the website.	I can get a visual representation of the results	High	Eshwaran ABR Vishnuram VR
Sprint 4	Pre-processing	USN-7	As an administrator, I can add new predictions to the training dataset	New records are visible in the updated dataset	Low	Donifa Babu B M Sadurthika M
Sprint 4		USN-8	As an administrator, I can remove incomplete records	Updations are visible in the updated dataset	Low	Vishnuram VR Sadurthika M
Sprint 4		USN-9	As an administrator, I can remove unimportant features	Updations are visible in the updated dataset	Low	Eshwaran ABR Donifa Babu B M

6.2 SPRINT DELIVERY SCHEDULE

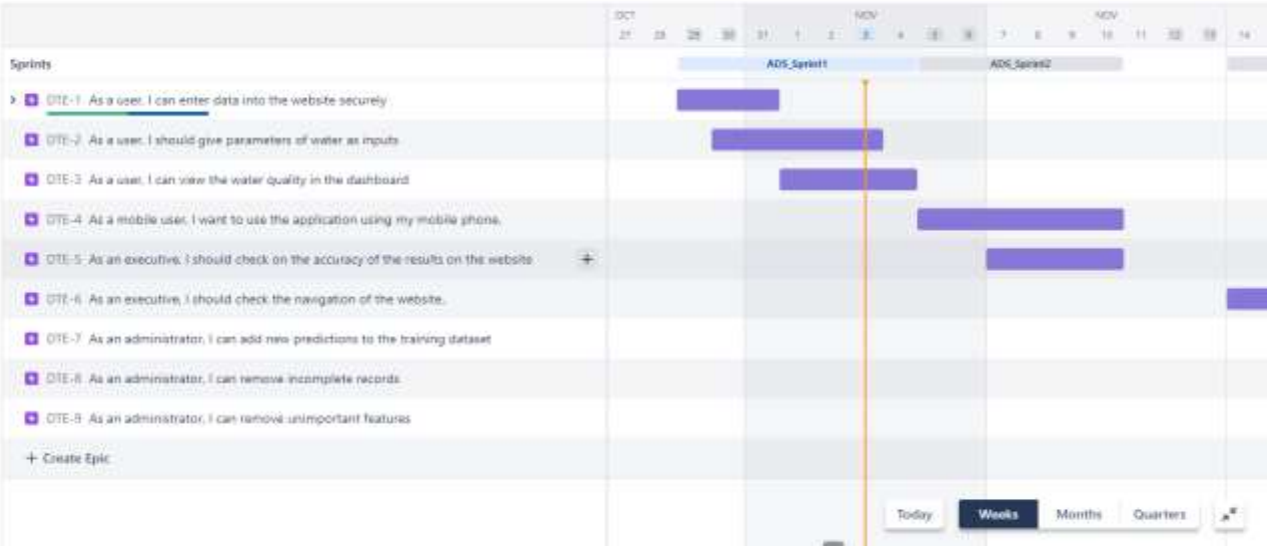


Project Tracker, Velocity & Burndown Chart: (4 Marks)

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	25 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

6.3 REPORTS FROM JIRA

Road map:

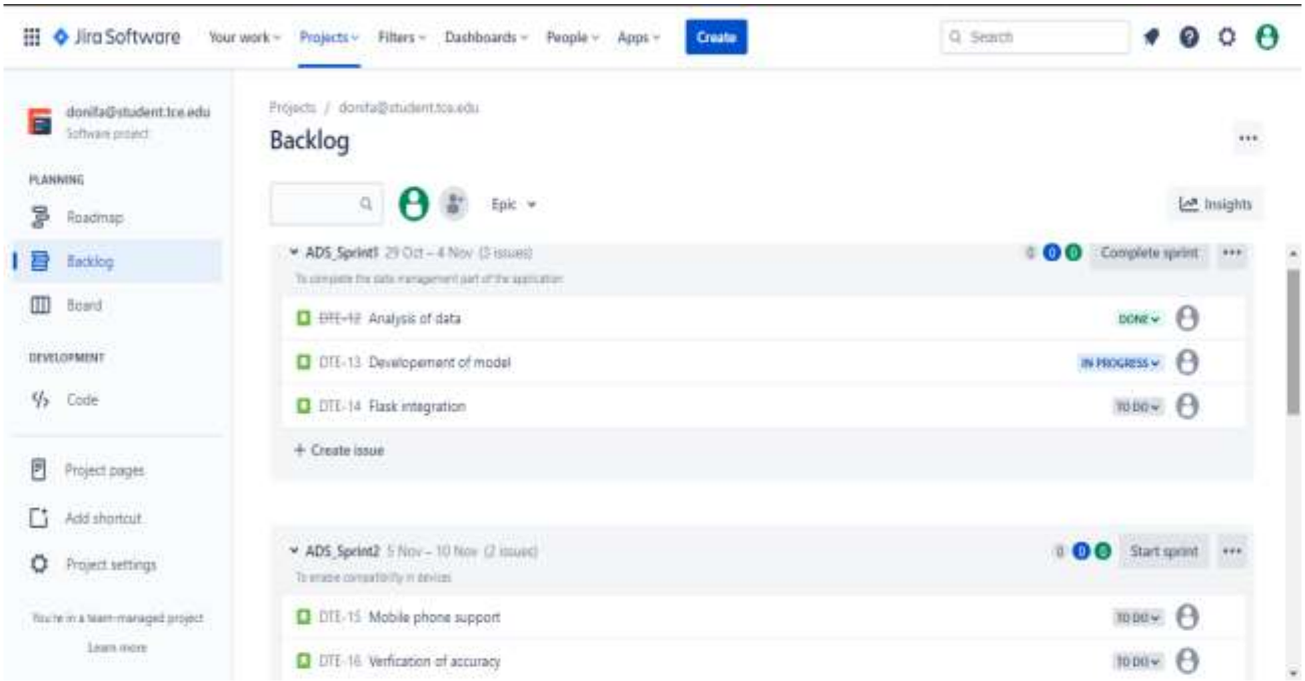


Sprints

- DTE-1 As a user, I can enter data into the website securely
- DTE-2 As a user, I should give parameters of water as inputs
- DTE-3 As a user, I can view the water quality in the dashboard
- DTE-4 As a mobile user, I want to use the application using my mobile phone.
- DTE-5 As an executive, I should check on the accuracy of the results on the website
- DTE-6 As an executive, I should check the navigation of the website.
- DTE-7 As an administrator, I can add new predictions to the training dataset
- DTE-8 As an administrator, I can remove incomplete records
- DTE-9 As an administrator, I can remove unimportant features

+ Create Epic

Backlog:



Projects / donita@student.tce.edu

Backlog

ADS_Sprint1 29 Oct – 4 Nov (3 issues)

To complete the data management part of the application

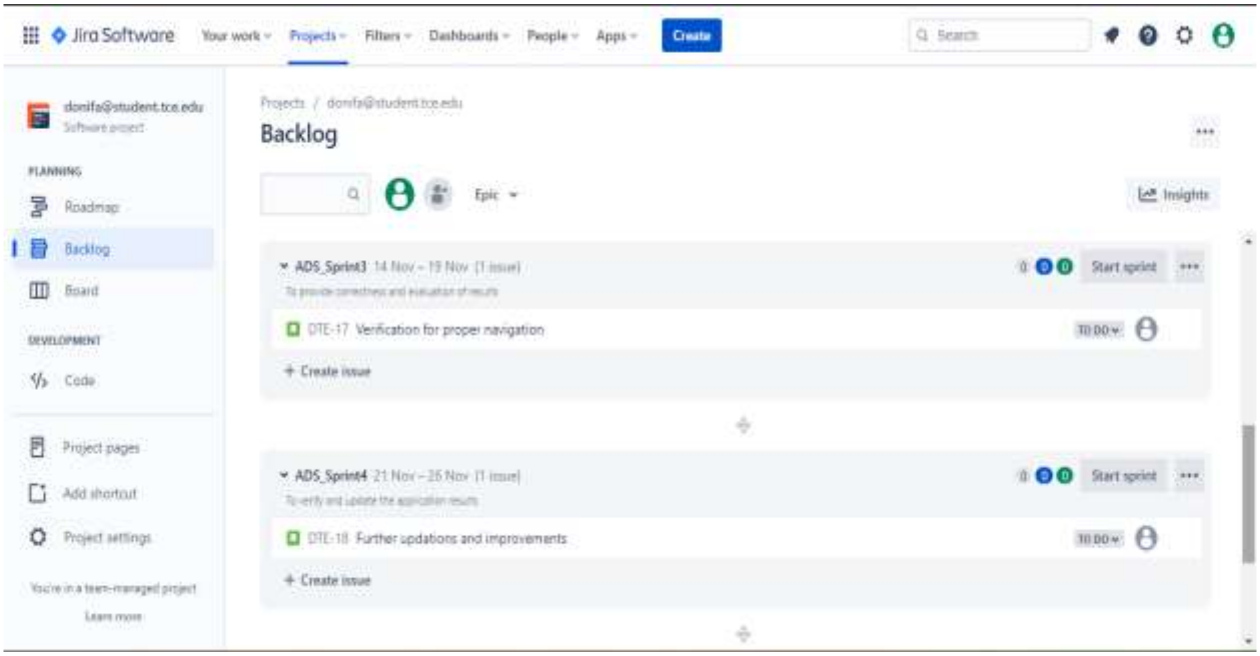
~~DTE-12~~	Analysis of data	DONE
DTE-13	Developement of model	IN PROGRESS
DTE-14	Flask integration	TO DO

+ Create issue

ADS_Sprint2 5 Nov – 10 Nov (2 issues)

To enable compatibility in devices

DTE-15	Mobile phone support	TO DO
DTE-16	Verfication of accuracy	TO DO



Backlog

ADS_Sprint3 14 Nov – 19 Nov (1 issue) 0 0 0

To provide correctness and evaluation of results

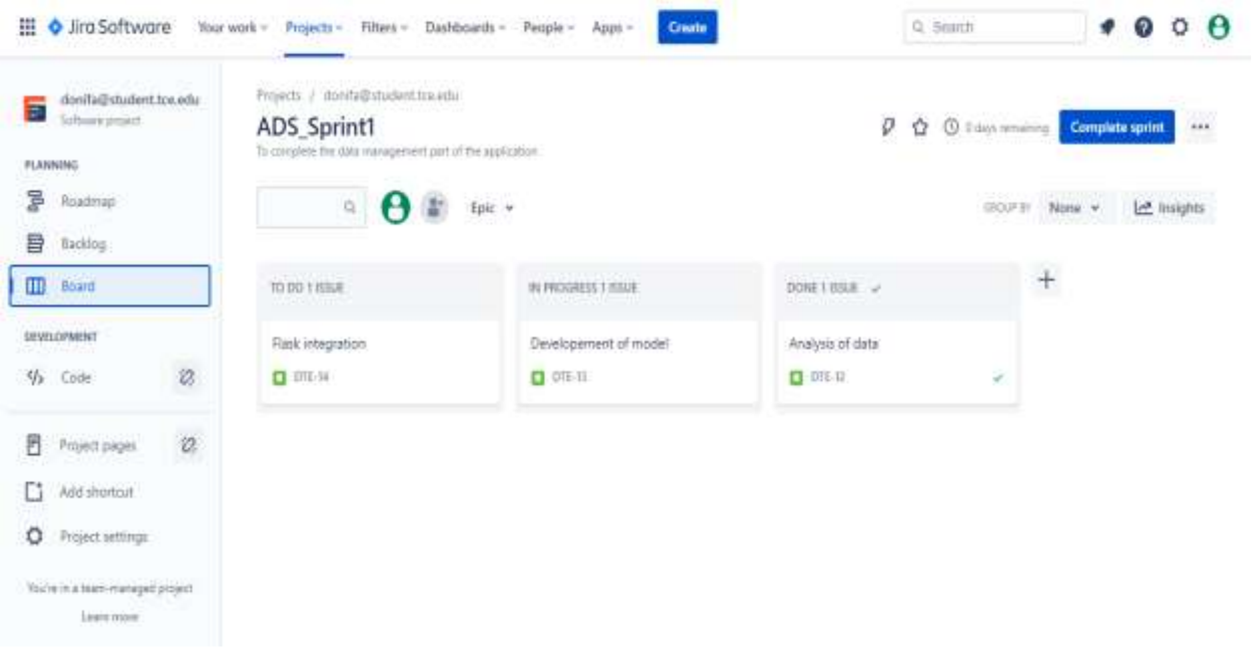
DTE-17 Verification for proper navigation

ADS_Sprint4 21 Nov – 26 Nov (1 issue) 0 0 0

To verify and update the application results

DTE-18 Further updations and improvements

Board:



ADS_Sprint1

To complete the data management part of the application

TO DO 1 ISSUE

Flask integration

DTE-14

IN PROGRESS 1 ISSUE

Developement of model

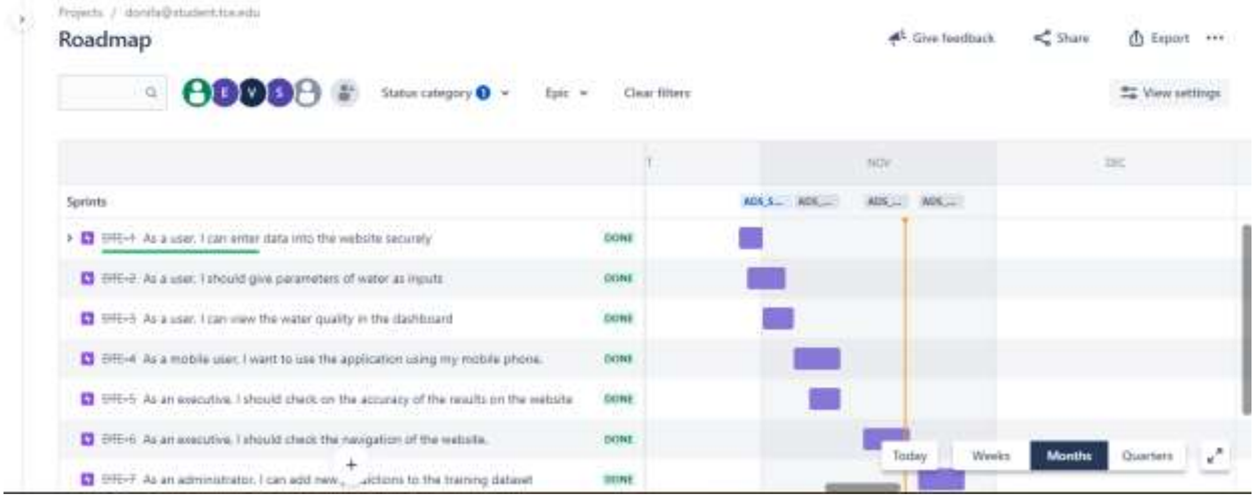
DTE-13

DONE 1 ISSUE

Analysis of data

DTE-12

AFTER COMPLETION OF PROJECT :



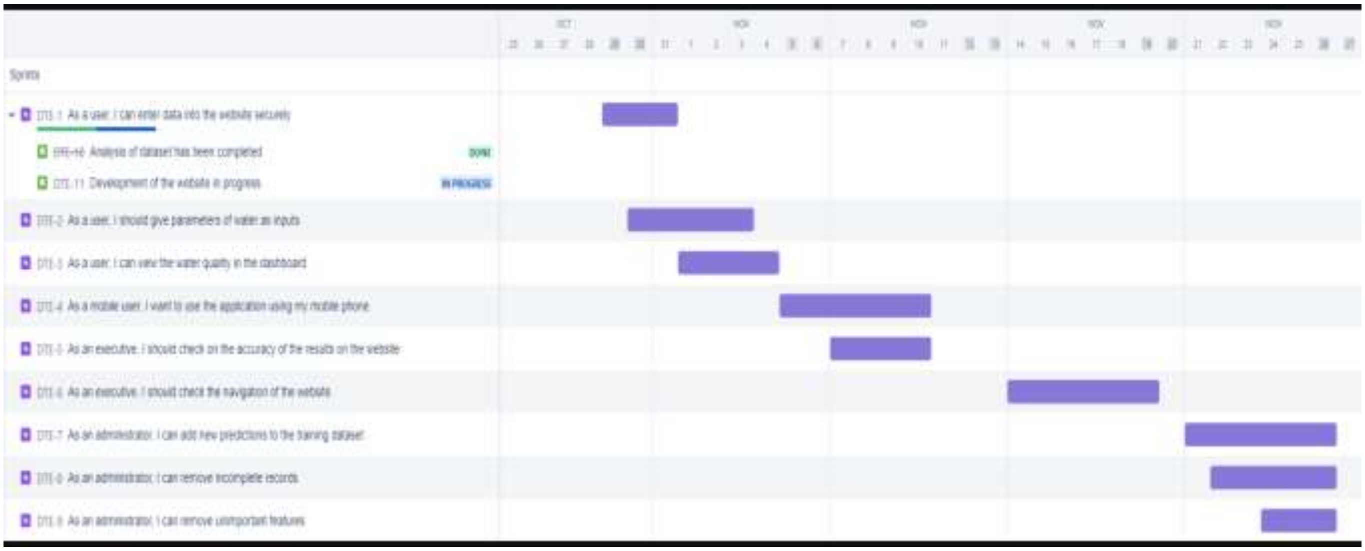
Projects / donita@student.tce.edu

Roadmap

Status category 1 Epic Clear filters

Sprints

- ~~EFE-1~~ As a user, I can enter data into the website securely DONE
- ~~EFE-2~~ As a user, I should give parameters of water as inputs DONE
- ~~EFE-3~~ As a user, I can view the water quality in the dashboard DONE
- ~~EFE-4~~ As a mobile user, I want to use the application using my mobile phone. DONE
- ~~EFE-5~~ As an executive, I should check on the accuracy of the results on the website DONE
- ~~EFE-6~~ As an executive, I should check the navigation of the website. DONE
- ~~EFE-7~~ As an administrator, I can add new ...ictions to the training dataset DONE



Sprints

- DTE-1 As a user, I can enter data into the website securely
  - ~~EFE-10~~ Analysis of dataset has been completed DONE
  - DTE-11 Development of the website in progress IN PROGRESS
- DTE-2 As a user, I should give parameters of water as inputs
- DTE-3 As a user, I can view the water quality in the dashboard
- DTE-4 As a mobile user, I want to use the application using my mobile phone
- DTE-5 As an executive, I should check on the accuracy of the results on the website
- DTE-6 As an executive, I should check the navigation of the website
- DTE-7 As an administrator, I can add new predictions to the training dataset
- DTE-8 As an administrator, I can remove incomplete records
- DTE-9 As an administrator, I can remove unimportant features

7. CODING & SOLUTIONING

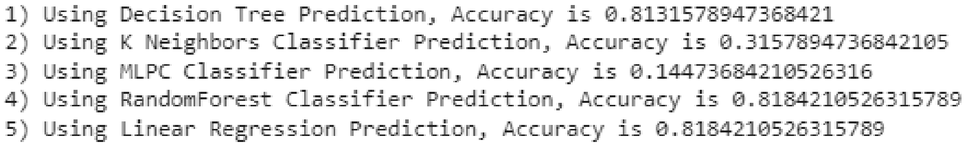
7.1 Feature 1: Predictive model

```
#Splitting the data into dependent and independent variables
X= df[['year', 'DO', 'PH', 'CO','BOD','NI','Tot_col']]
df['wqi']=df['wqi'].astype('int')
Y= df[['wqi']]

X.shape

Y.shape
```

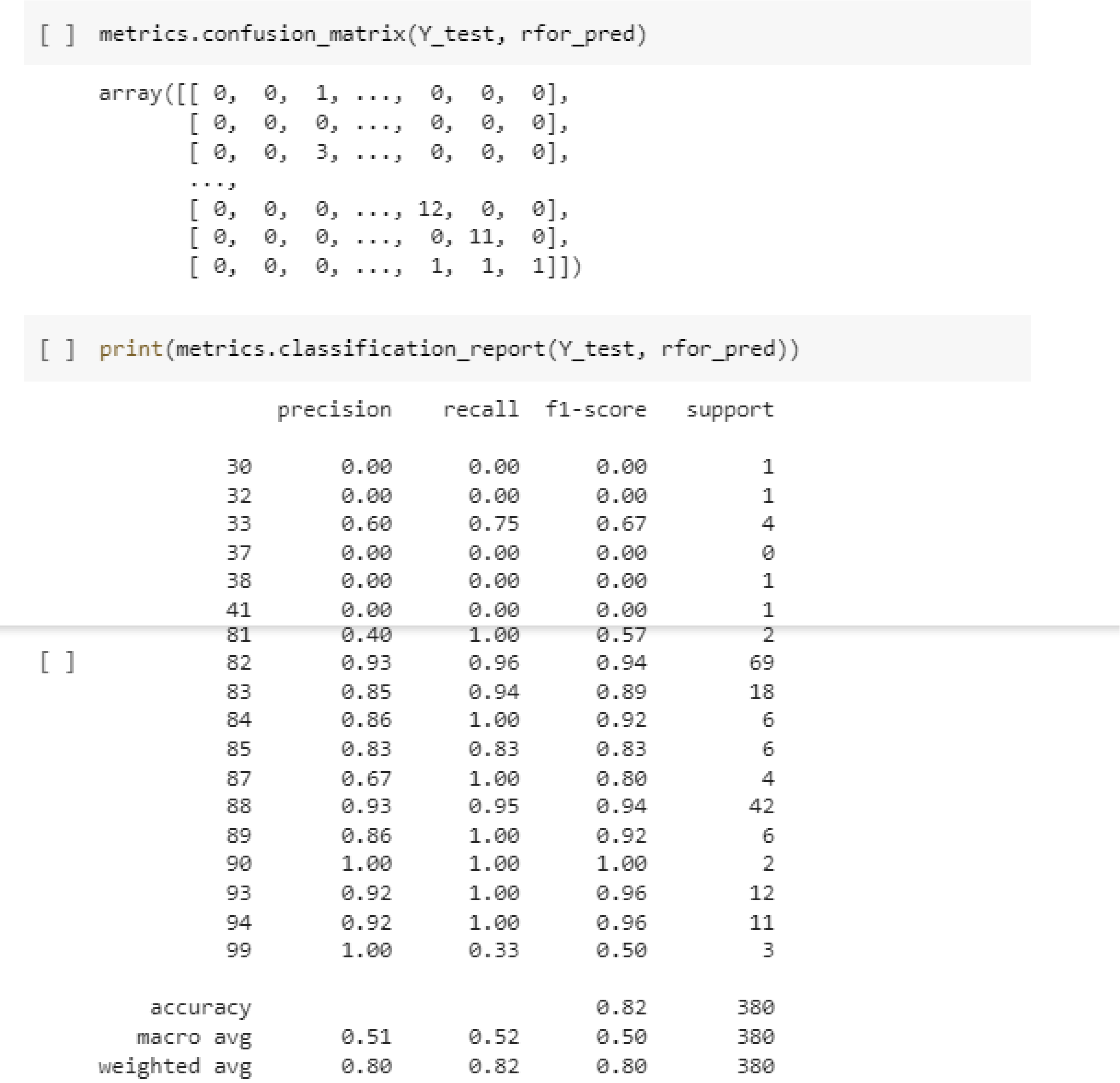
```
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn import linear_model
from sklearn import metrics
import math
from sklearn.metrics import mean_squared_error

X_train, X_test, Y_train, Y_test = train_test_split( X, Y,test_size=0.2, random_state=10)

#from sklearn.preprocessing import StandardScaler
#sc_X = StandardScaler()
#X_train = sc_X.fit_transform(X_train)
#X_test = sc_X.transform(X_test)

#{Decision Tree Model}
clf = DecisionTreeClassifier()
clf = clf.fit(X_train,Y_train)
clf_pred=clf.predict(X_test)
clf_accuracy=metrics.accuracy_score(Y_test,clf_pred)
print("1) Using Decision Tree Prediction, Accuracy is " + str(clf_accuracy))

#{K Neighbors Classifier}
knn = KNeighborsClassifier(n_neighbors=7)
knn=knn.fit(X_train,Y_train.values.ravel())
knn_pred=knn.predict(X_test)
knn_accuracy=metrics.accuracy_score(Y_test,knn_pred)
print ("2) Using K Neighbors Classifier Prediction, Accuracy is " + str(knn_accuracy))

#{using MLPClassifier}
mlpc = MLPClassifier()
mlpc.fit(X_train,Y_train.values.ravel())
mlpc_pred=mlpc.predict(X_test)
mlpc_accuracy=metrics.accuracy_score(Y_test,mlpc_pred)
print ("3) Using MLPC Classifier Prediction, Accuracy is " + str(mlpc_accuracy))

#{using MLPClassifier}
rfor = RandomForestClassifier()
rfor.fit(X_train,Y_train.values.ravel())
rfor_pred=rfor.predict(X_test)
rfor_accuracy=metrics.accuracy_score(Y_test,rfor_pred)
print ("4) Using RandomForest Classifier Prediction, Accuracy is " + str(rfor_accuracy))
```

```
#{using Linear Regression}
linreg=linear_model.LinearRegression()
linreg.fit(X_train,Y_train)
linreg_pred=rfor.predict(X_test)
linreg_accuracy=metrics.accuracy_score(Y_test,linreg_pred)
rmse = math.sqrt(mean_squared_error(Y_test,linreg_pred))
print ("5) Using Linear Regression Prediction, Accuracy is " + str(linreg_accuracy))

# Accuracy found maximum in RandomForest and Linear Regression


metrics.confusion_matrix(Y_test, rfor_pred)
print(metrics.classification_report(Y_test, rfor_pred))

# Saving the model
# loading library
import pickle
with open('WaterQuality_RFModel.pkl', 'wb') as files:
    pickle.dump(rfor, files)
```

OUTPUT:



```
1) Using Decision Tree Prediction, Accuracy is 0.8131578947368421
2) Using K Neighbors Classifier Prediction, Accuracy is 0.3157894736842105
3) Using MLPC Classifier Prediction, Accuracy is 0.14473684210526316
4) Using RandomForest Classifier Prediction, Accuracy is 0.8184210526315789
5) Using Linear Regression Prediction, Accuracy is 0.8184210526315789
```



```
[ ]  metrics.confusion_matrix(Y_test, rfor_pred)

     array([[ 0,  0,  1, ...,  0,  0,  0],
            [ 0,  0,  0, ...,  0,  0,  0],
            [ 0,  0,  3, ...,  0,  0,  0],
            ...,
            [ 0,  0,  0, ..., 12,  0,  0],
            [ 0,  0,  0, ...,  0, 11,  0],
            [ 0,  0,  0, ...,  1,  1,  1]])
```

```
[ ]  print(metrics.classification_report(Y_test, rfor_pred))

                   precision    recall  f1-score   support

               30       0.00      0.00      0.00         1
               32       0.00      0.00      0.00         1
               33       0.60      0.75      0.67         4
               37       0.00      0.00      0.00         0
               38       0.00      0.00      0.00         1
               41       0.00      0.00      0.00         1
               81       0.40      1.00      0.57         2
[ ]            82       0.93      0.96      0.94        69
               83       0.85      0.94      0.89        18
               84       0.86      1.00      0.92         6
               85       0.83      0.83      0.83         6
               87       0.67      1.00      0.80         4
               88       0.93      0.95      0.94        42
               89       0.86      1.00      0.92         6
               90       1.00      1.00      1.00         2
               93       0.92      1.00      0.96        12
               94       0.92      1.00      0.96        11
               99       1.00      0.33      0.50         3

         accuracy                           0.82       380
        macro avg       0.51      0.52      0.50       380
     weighted avg       0.80      0.82      0.80       380
```

Link for Colab : https://colab.research.google.com/drive/1XuAKvroTOVjJUDhBld0QEW6uIxMCM422#scrollTo=t5cM9GC9W1ex

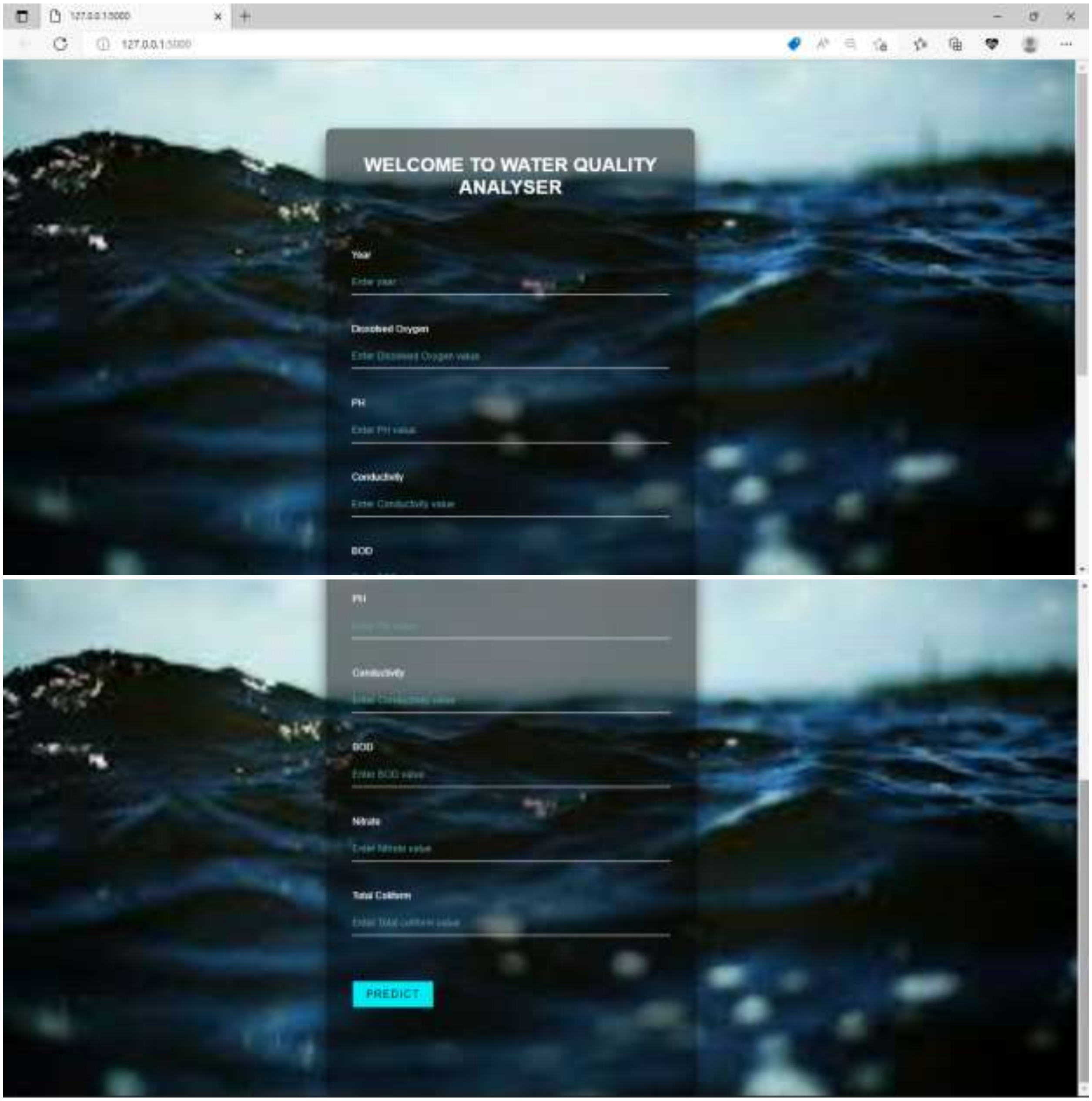
7.2 Feature 2: Input from users

```
from flask import Flask, render_template, request

app = Flask(__name__)

# interface between my server and my application wsgi
import pickle
```

```
model = pickle.load(open('WaterQuality_RFModel.pkl','rb'))

@app.route('/')#binds to an url
def helloworld():
    print("KIDDO>>")
    return render_template("index.html")

@app.route('/login', methods=['POST','GET'])#binds to an url
def login():
    print("Welcome to login")
    aa= request.form["year"]
    q= request.form["DO"]
    r= request.form["PH"]
    s= request.form["Conductivity"]
    pp = request.form["BOD"]
    qq = request.form["NI"]
    ss = request.form["Tot_col"]

    #t=request.form
    t=[[int(aa),float(q),float(r),float(s),float(pp),float(qq),float(ss)]]
    output= model.predict(t)
    print(output)
    output=output[[0]]
    if (output>=95 and output<=100):
        return render_template("index.html", y="Excellent. The WQI predicted value is " + str(output))  # we can only concatenate string to str
    elif (output>=89 and output<=94):
        return render_template("index.html", y="Very Good. The WQI predicted value is " + str(output))  # we can only concatenate string to str
    elif(output>=80 and output<=88):
        return render_template("index.html", y="Good. The WQI predicted value is " + str(output))  # we can only concatenate string to str
    elif (output >=65 and output <=79):
        return render_template("index.html", y="Fair. The WQI predicted value is " + str(output))  # we can only concatenate string to str
    elif (output >=45 and output <=64):
        return render_template("index.html", y="Marginal. The WQI predicted value  is " + str(output))  # we can only concatenate string to str
    else:
        return render_template("index.html", y="Poor. The WQI predicted value is " + str(output))  # we can only concatenate string to str


@app.route('/admin')#binds to an url
```

```
def admin():
    return "Hey Admin How are you?"

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

OUTPUT:



127.0.0.1:5000

WELCOME TO WATER QUALITY ANALYSER

Year

Enter year

Dissolved Oxygen

Enter Dissolved Oxygen value

PH

Enter PH value

Conductivity

Enter Conductivity value

BOD

PH

Enter PH value

Conductivity

Enter Conductivity value

BOD

Enter BOD value

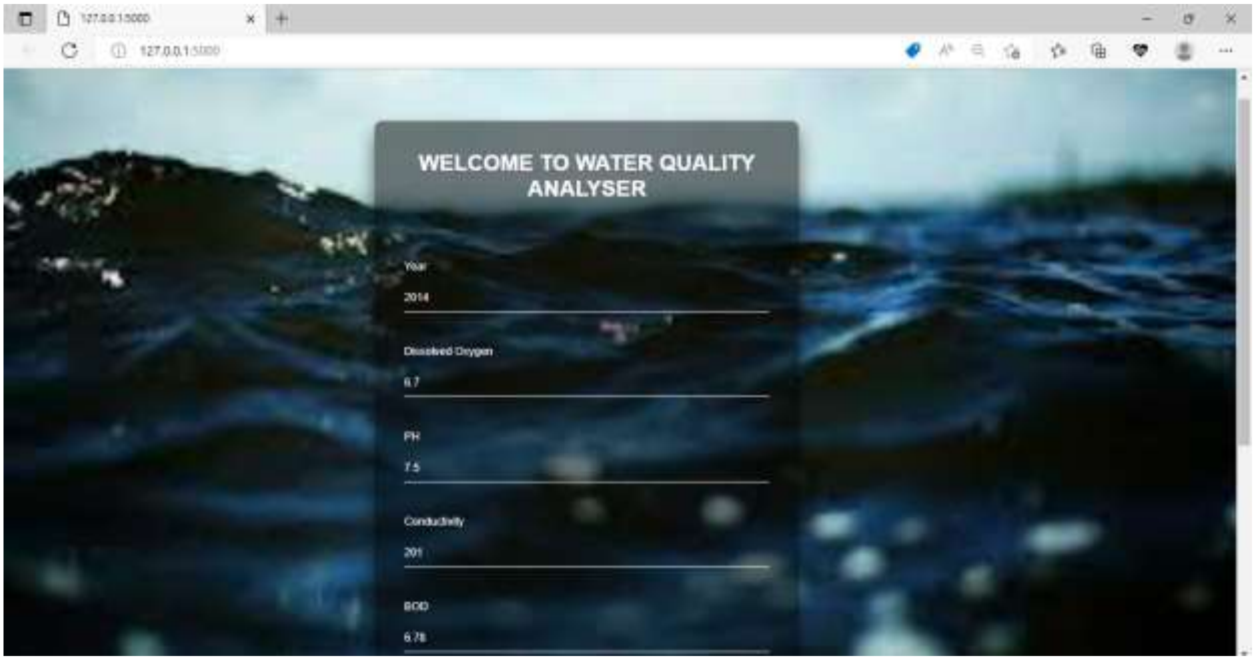
Nitrate

Enter Nitrate value

Total Coliform

Enter total coliform value

After inputs are being entered,



WELCOME TO WATER QUALITY ANALYSER

Year

2014

Dissolved Oxygen

6.7

PH

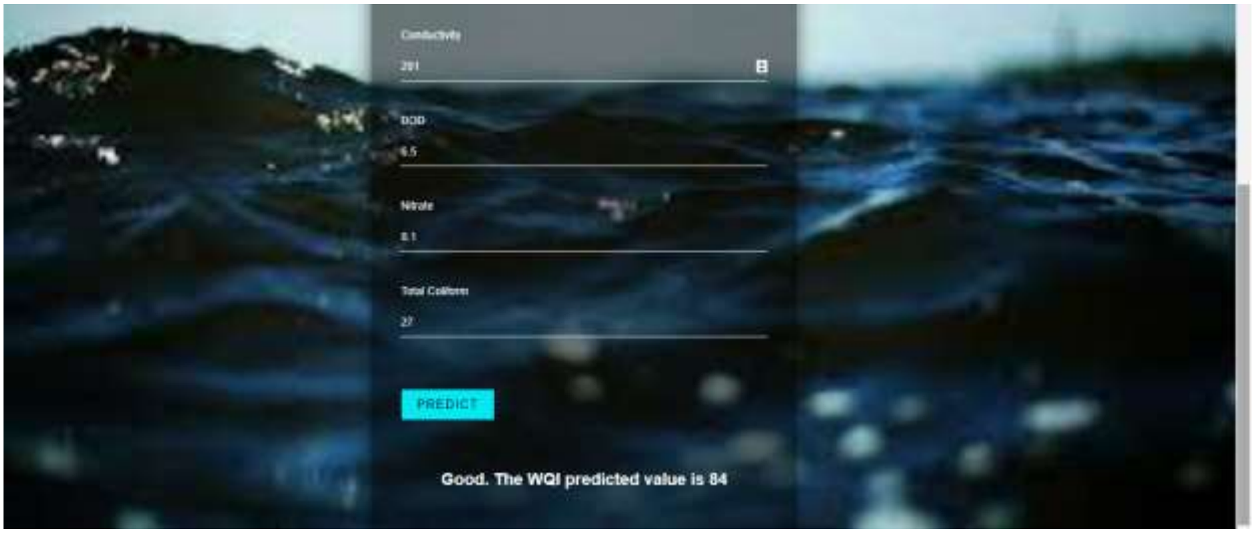
7.5

Conductivity

201

BOD

6.78



Conductivity

201

BOD

6.5

Nitrate

0.1

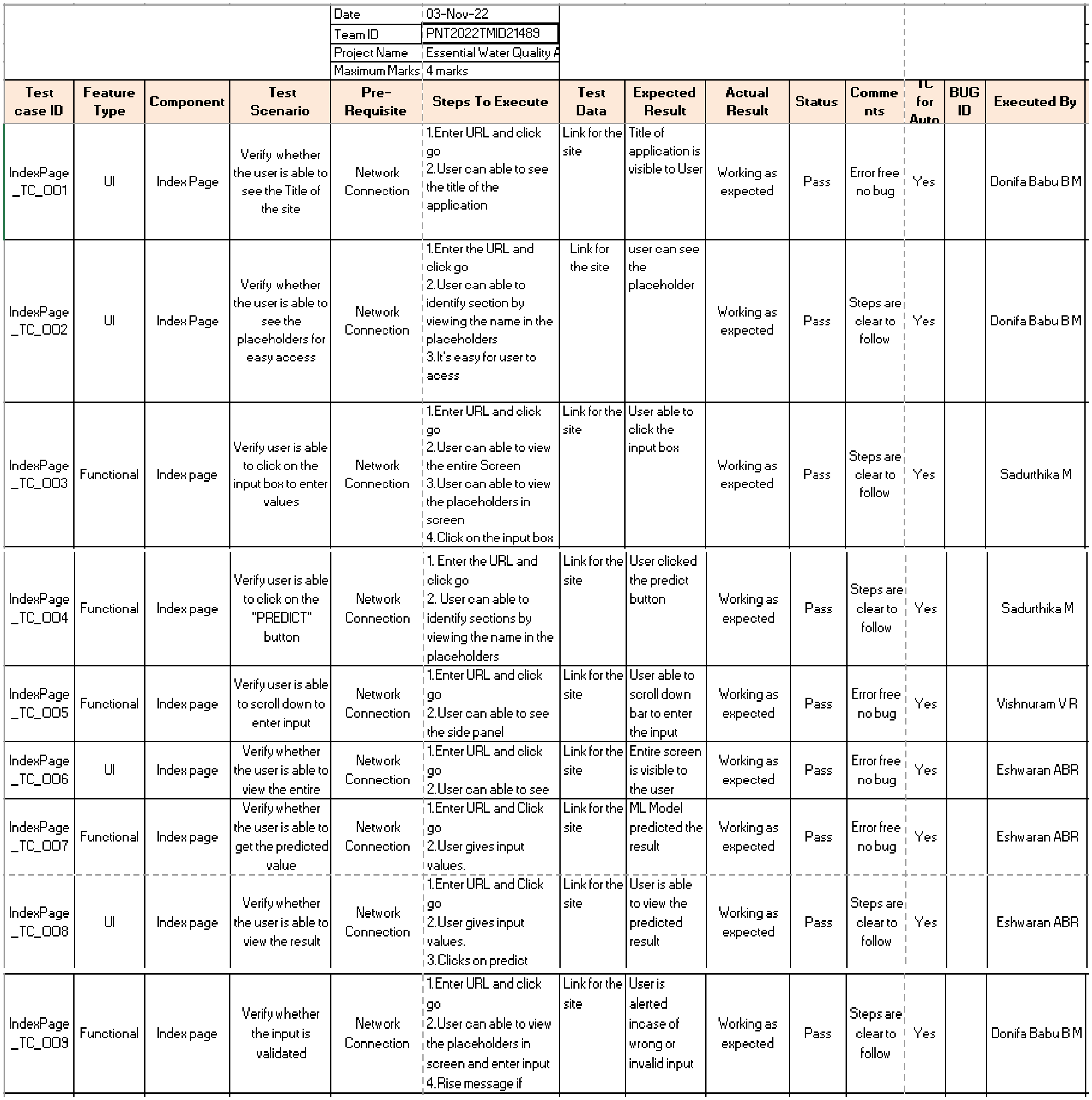
Total Coliform

27

Good. The WQI predicted value is 84

8. TESTING

8.1 TEST CASES



				Date	03-Nov-22								
				Team ID	PNT2022TMID21489								
				Project Name	Essential Water Quality A								
				Maximum Marks	4 marks								
Test case ID	**Feature Type**	**Component**	**Test Scenario**	**Pre-Requisite**	**Steps To Execute**	**Test Data**	**Expected Result**	**Actual Result**	**Status**	**Comments**	**TC for Auto**	**BUG ID**	**Executed By**
IndexPage _TC_001	UI	Index Page	Verify whether the user is able to see the Title of the site	Network Connection	1.Enter URL and click go 2.User can able to see the title of the application	Link for the site	Title of application is visible to User	Working as expected	Pass	Error free no bug	Yes		Donifa Babu B M
IndexPage _TC_002	UI	Index Page	Verify whether the user is able to see the placeholders for easy access	Network Connection	1.Enter the URL and click go 2.User can able to identify section by viewing the name in the placeholders 3.It's easy for user to acess	Link for the site	user can see the placeholder	Working as expected	Pass	Steps are clear to follow	Yes		Donifa Babu B M
IndexPage _TC_003	Functional	Index page	Verify user is able to click on the input box to enter values	Network Connection	1.Enter URL and click go 2.User can able to view the entire Screen 3.User can able to view the placeholders in screen 4.Click on the input box	Link for the site	User able to click the input box	Working as expected	Pass	Steps are clear to follow	Yes		Sadurthika M
IndexPage _TC_004	Functional	Index page	Verify user is able to click on the "PREDICT" button	Network Connection	1. Enter the URL and click go 2. User can able to identify sections by viewing the name in the placeholders	Link for the site	User clicked the predict button	Working as expected	Pass	Steps are clear to follow	Yes		Sadurthika M
IndexPage _TC_005	Functional	Index page	Verify user is able to scroll down to enter input	Network Connection	1.Enter URL and click go 2.User can able to see the side panel	Link for the site	User able to scroll down bar to enter the input	Working as expected	Pass	Error free no bug	Yes		Vishnuram V R
IndexPage _TC_006	UI	Index page	Verify whether the user is able to view the entire	Network Connection	1.Enter URL and click go 2.User can able to see	Link for the site	Entire screen is visible to the user	Working as expected	Pass	Error free no bug	Yes		Eshwaran ABR
IndexPage _TC_007	Functional	Index page	Verify whether the user is able to get the predicted value	Network Connection	1.Enter URL and Click go 2.User gives input values.	Link for the site	ML Model predicted the result	Working as expected	Pass	Error free no bug	Yes		Eshwaran ABR
IndexPage _TC_008	UI	Index page	Verify whether the user is able to view the result	Network Connection	1.Enter URL and Click go 2.User gives input values. 3.Clicks on predict	Link for the site	User is able to view the predicted result	Working as expected	Pass	Steps are clear to follow	Yes		Eshwaran ABR
IndexPage _TC_009	Functional	Index page	Verify whether the input is validated	Network Connection	1.Enter URL and click go 2.User can able to view the placeholders in screen and enter input 4.Rise message if	Link for the site	User is alerted incase of wrong or invalid input	Working as expected	Pass	Steps are clear to follow	Yes		Donifa Babu B M

8.2 USER ACCEPTANCE TESTING

Defect Analysis:

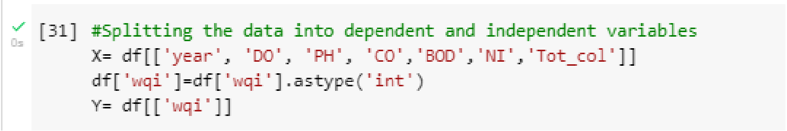
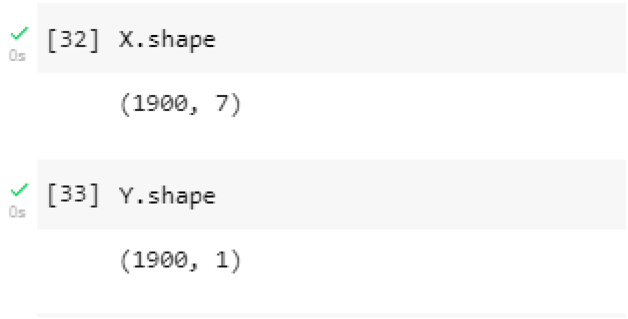
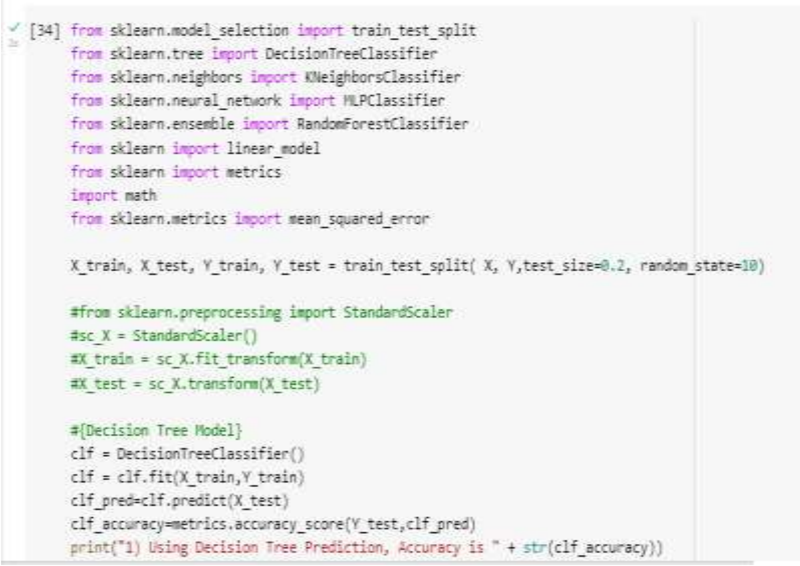
Resolution	Severity 1	Severity 2	Severity 3	Severity 4	Subtotal
By Design	10	4	2	3	19
Duplicate	1	0	3	0	4
External	2	3	0	1	6
Fixed	11	2	4	20	37
Not Reproduced	0	0	0	0	0
Skipped	0	0	0	0	0
Won't Fix	0	0	0	0	0
Totals	24	9	9	24	66

Test Case analysis:

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

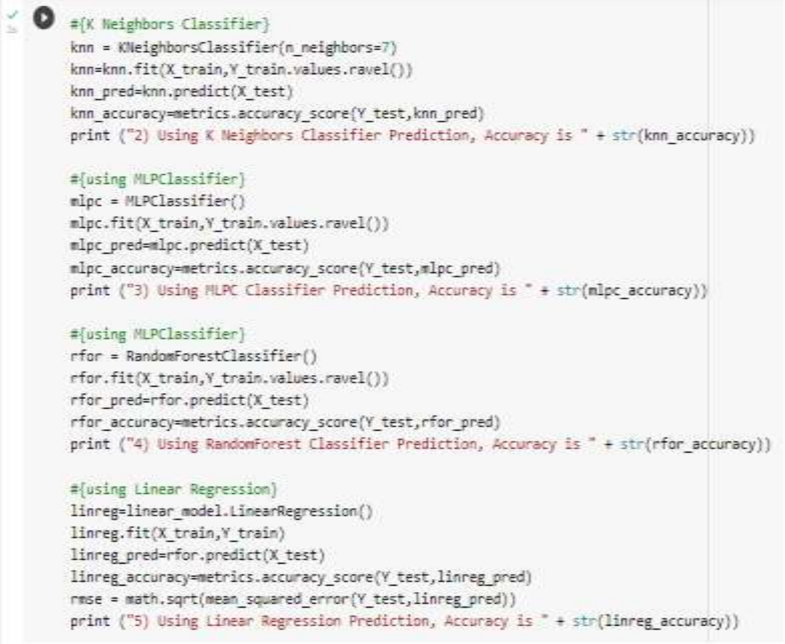
9.1 PERFORMANCE METRICS

No.	Parameter	Values	Screenshot
1.	Metrics	Classification Model:	  

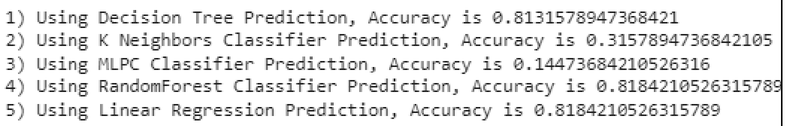
```
[31] #Splitting the data into dependent and independent variables
     X= df[['year', 'DO', 'PH', 'CO','BOD','NI','Tot_col']]
     df['wqi']=df['wqi'].astype('int')
     Y= df[['wqi']]
```

```
[32] X.shape

     (1900, 7)

[33] Y.shape

     (1900, 1)
```

```
[34] from sklearn.model_selection import train_test_split
     from sklearn.tree import DecisionTreeClassifier
     from sklearn.neighbors import KNeighborsClassifier
     from sklearn.neural_network import MLPClassifier
     from sklearn.ensemble import RandomForestClassifier
     from sklearn import linear_model
     from sklearn import metrics
     import math
     from sklearn.metrics import mean_squared_error

     X_train, X_test, Y_train, Y_test = train_test_split( X, Y,test_size=0.2, random_state=10)

     #from sklearn.preprocessing import StandardScaler
     #sc_X = StandardScaler()
     #X_train = sc_X.fit_transform(X_train)
     #X_test = sc_X.transform(X_test)

     #{Decision Tree Model}
     clf = DecisionTreeClassifier()
     clf = clf.fit(X_train,Y_train)
     clf_pred=clf.predict(X_test)
     clf_accuracy=metrics.accuracy_score(Y_test,clf_pred)
     print("1) Using Decision Tree Prediction, Accuracy is " + str(clf_accuracy))
```

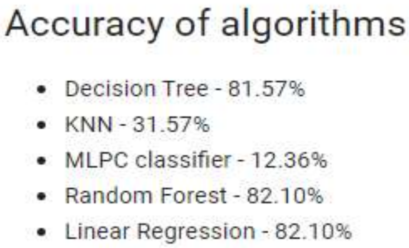


```
#{K Neighbors Classifier}
knn = KNeighborsClassifier(n_neighbors=7)
knn=knn.fit(X_train,Y_train.values.ravel())
knn_pred=knn.predict(X_test)
knn_accuracy=metrics.accuracy_score(Y_test,knn_pred)
print ("2) Using K Neighbors Classifier Prediction, Accuracy is " + str(knn_accuracy))

#{using MLPClassifier}
mlpc = MLPClassifier()
mlpc.fit(X_train,Y_train.values.ravel())
mlpc_pred=mlpc.predict(X_test)
mlpc_accuracy=metrics.accuracy_score(Y_test,mlpc_pred)
print ("3) Using MLPC Classifier Prediction, Accuracy is " + str(mlpc_accuracy))

#{using MLPClassifier}
rfor = RandomForestClassifier()
rfor.fit(X_train,Y_train.values.ravel())
rfor_pred=rfor.predict(X_test)
rfor_accuracy=metrics.accuracy_score(Y_test,rfor_pred)
print ("4) Using RandomForest Classifier Prediction, Accuracy is " + str(rfor_accuracy))

#{using Linear Regression}
linreg=linear_model.LinearRegression()
linreg.fit(X_train,Y_train)
linreg_pred=rfor.predict(X_test)
linreg_accuracy=metrics.accuracy_score(Y_test,linreg_pred)
rmse = math.sqrt(mean_squared_error(Y_test,linreg_pred))
print ("5) Using Linear Regression Prediction, Accuracy is " + str(linreg_accuracy))
```



```
1) Using Decision Tree Prediction, Accuracy is 0.8131578947368421
2) Using K Neighbors Classifier Prediction, Accuracy is 0.3157894736842105
3) Using MLPC Classifier Prediction, Accuracy is 0.14473684210526316
4) Using RandomForest Classifier Prediction, Accuracy is 0.8184210526315789
5) Using Linear Regression Prediction, Accuracy is 0.8184210526315789
```

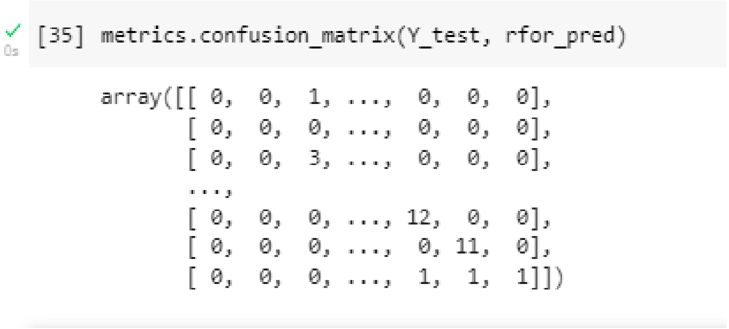
Accuracy Score



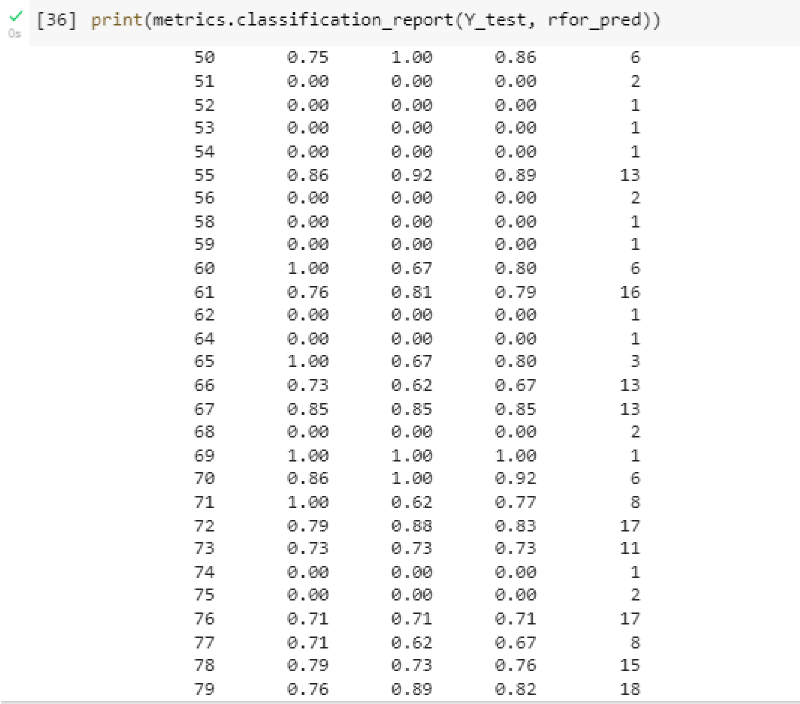
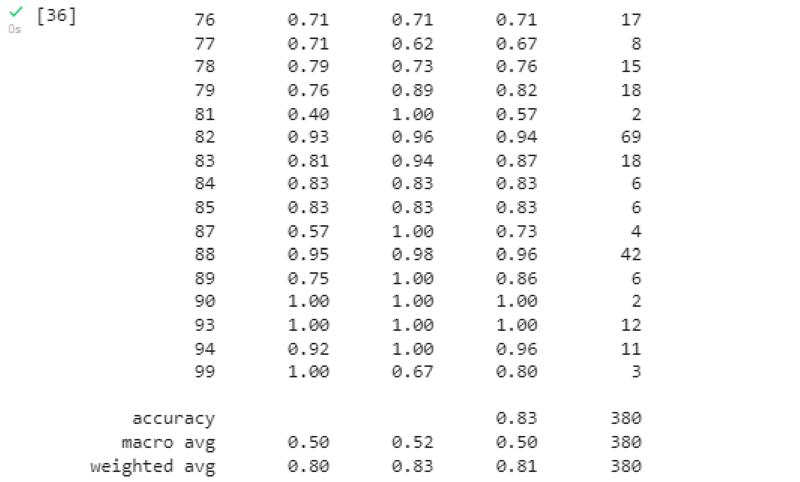
Accuracy of algorithms

- Decision Tree - 81.57%
- KNN - 31.57%
- MLPC classifier - 12.36%
- Random Forest - 82.10%
- Linear Regression - 82.10%

Confusion Matrix

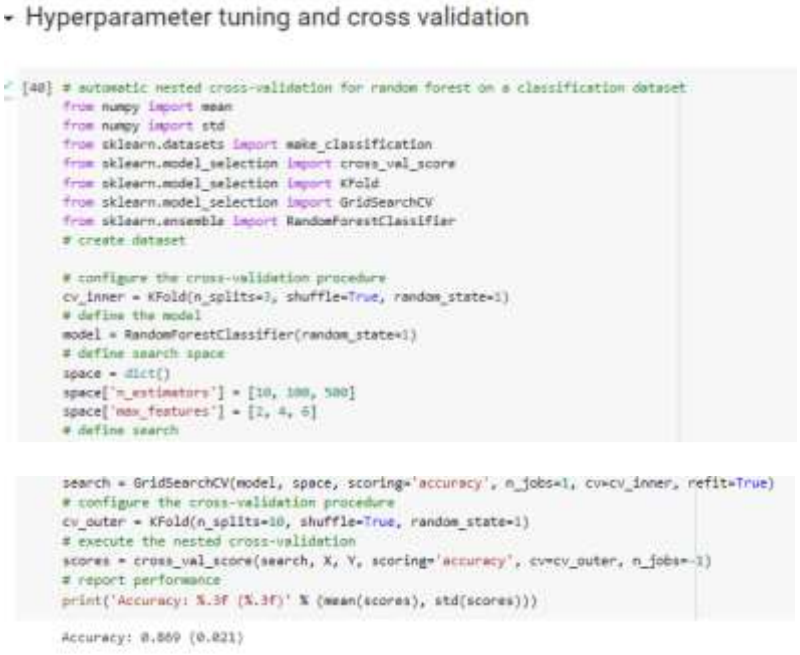


```
[35] metrics.confusion_matrix(Y_test, rfor_pred)

array([[ 0,  0,  1, ...,  0,  0,  0],
       [ 0,  0,  0, ...,  0,  0,  0],
       [ 0,  0,  3, ...,  0,  0,  0],
       ...,
       [ 0,  0,  0, ..., 12,  0,  0],
       [ 0,  0,  0, ...,  0, 11,  0],
       [ 0,  0,  0, ...,  1,  1,  1]])
```

		Classification Report	 

```
[36] print(metrics.classification_report(Y_test, rfor_pred))
                  50       0.75      1.00      0.86         6
                  51       0.00      0.00      0.00         2
                  52       0.00      0.00      0.00         1
                  53       0.00      0.00      0.00         1
                  54       0.00      0.00      0.00         1
                  55       0.86      0.92      0.89        13
                  56       0.00      0.00      0.00         2
                  58       0.00      0.00      0.00         1
                  59       0.00      0.00      0.00         1
                  60       1.00      0.67      0.80         6
                  61       0.76      0.81      0.79        16
                  62       0.00      0.00      0.00         1
                  64       0.00      0.00      0.00         1
                  65       1.00      0.67      0.80         3
                  66       0.73      0.62      0.67        13
                  67       0.85      0.85      0.85        13
                  68       0.00      0.00      0.00         2
                  69       1.00      1.00      1.00         1
                  70       0.86      1.00      0.92         6
                  71       1.00      0.62      0.77         8
                  72       0.79      0.88      0.83        17
                  73       0.73      0.73      0.73        11
                  74       0.00      0.00      0.00         1
                  75       0.00      0.00      0.00         2
                  76       0.71      0.71      0.71        17
                  77       0.71      0.62      0.67         8
                  78       0.79      0.73      0.76        15
                  79       0.76      0.89      0.82        18
```

```
[36]
                  76       0.71      0.71      0.71        17
                  77       0.71      0.62      0.67         8
                  78       0.79      0.73      0.76        15
                  79       0.76      0.89      0.82        18
                  81       0.40      1.00      0.57         2
                  82       0.93      0.96      0.94        69
                  83       0.81      0.94      0.87        18
                  84       0.83      0.83      0.83         6
                  85       0.83      0.83      0.83         6
                  87       0.57      1.00      0.73         4
                  88       0.95      0.98      0.96        42
                  89       0.75      1.00      0.86         6
                  90       1.00      1.00      1.00         2
                  93       1.00      1.00      1.00        12
                  94       0.92      1.00      0.96        11
                  99       1.00      0.67      0.80         3

            accuracy                           0.83       380
           macro avg       0.50      0.52      0.50       380
        weighted avg       0.80      0.83      0.81       380
```

2.	Tune the Model	Hyperparameter Tuning and Validation Method	

- Hyperparameter tuning and cross validation

```
[40] # automatic nested cross-validation for random forest on a classification dataset
     from numpy import mean
     from numpy import std
     from sklearn.datasets import make_classification
     from sklearn.model_selection import cross_val_score
     from sklearn.model_selection import KFold
     from sklearn.model_selection import GridSearchCV
     from sklearn.ensemble import RandomForestClassifier
     # create dataset

     # configure the cross-validation procedure
     cv_inner = KFold(n_splits=3, shuffle=True, random_state=1)
     # define the model
     model = RandomForestClassifier(random_state=1)
     # define search space
     space = dict()
     space['n_estimators'] = [10, 100, 500]
     space['max_features'] = [2, 4, 6]
     # define search

     search = GridSearchCV(model, space, scoring='accuracy', n_jobs=1, cv=cv_inner, refit=True)
     # configure the cross-validation procedure
     cv_outer = KFold(n_splits=10, shuffle=True, random_state=1)
     # execute the nested cross-validation
     scores = cross_val_score(search, X, Y, scoring='accuracy', cv=cv_outer, n_jobs=-1)
     # report performance
     print('Accuracy: %.3f (%.3f)' % (mean(scores), std(scores)))

     Accuracy: 0.869 (0.021)
```

10. ADVANTAGES & DISADVANTAGES

11. CONCLUSION

The proposed system helps to predict water quality based on water parameters. One of the major issues faced by the world is water scarcity. Another one is water contamination. Reasons for these two can be increased population, growth of urbanization to its peak, and the tremendous increase of industrialization. The proposed system can be implemented by including basic parameters and can be expanded by incorporating various features associated with water quality. These kinds of quality monitoring systems will help society to achieve a more secure future as water pollution can be controlled to a great extent through continuous monitoring. With minimal features, implementation will be much easier

12. FUTURE SCOPE

The proposed system uses machine learning algorithms to forecast potential water quality, The system can be improved by performing various algorithms to increase accuracy. The project’s future scope is to gather water parameters, the proposed system comprises multiple sensors interfaced with Node MCU. The user is notified before the water turns contaminated. The implemented solution protects the water from pollution while still being cost-efficient.

13. APPENDIX

Source Code:

https://colab.research.google.com/drive/1XuAKvroTOVjJUDhBld0QEW6uIxMCM422#scrollTo=fNr8TIntIDks

GitHub:

https://github.com/IBM-EPBL/IBM-Project-40383-1660628866

Project Demo Link:

https://drive.google.com/file/d/1YIep5dqNpxQI293g3EhDN3aNlCgEF6lW/view?usp=sharing