

**VIVEKANANDHA COLLEGE OF ENGINEERING FOR
WOMEN(AUTONOMOUS)**

Elayampalayam, Thiruchengode-637205

PROJECT

**EFFICIENT WATER QUALITY ANALYSIS AND
PREDICTION USING MACHINE LEARNING**

DONE BY

TEAM ID:PNT2022TMID23851

MYTHILI B (612919103064)

SWETHA R (612919103104)

SNEKA K (612919103091)

RESHMITHA K (612919103081)

GUIDED BY

Mr.M.THIRUPPATHI(MENTOR)

Ms. LALITHA GAYATHRI (INDUSTRY MENTOR)

INDEX

1. INTRODUCTION

1.1 Project Overview

1.2 Purpose

2. LITERATURE SURVEY

2.1 Existing problem

2.2 References

2.3 Problem Statement Definition

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas

3.2 Ideation & Brainstorming

3.3 Proposed Solution

3.4 Problem Solution fit

4. REQUIREMENT ANALYSIS

4.1 Functional requirement

4.2 Non-Functional requirements

5. PROJECT DESIGN

5.1 Data Flow Diagrams

5.2 Solution & Technical Architecture

5.3 User Stories

6. PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

6.2 Sprint Delivery Schedule

6.3 Reports from JIRA

7. CODING & SOLUTIONING (Explain the features added in the project along with code)

7.1 Feature 1

7.2 Feature 2

7.3 Database Schema (if Applicable)

8. TESTING

8.1 Test Cases

8.2 User Acceptance Testing

9. RESULTS

9.1 Performance Metrics

10. **ADVANTAGES & DISADVANTAGES**

11. **CONCLUSION**

12. **FUTURE SCOPE** 13. **APPENDIX** Source Code

GitHub & Project Demo

EFFICIENT 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. Water is the most important of sources, vital for sustaining all kinds of life; however, it is in constant threat of pollution by life itself. Water is one of the most communicable mediums with a far reach. Rapid industrialization has consequently led to deterioration of water quality at an alarming rate. Poor

water quality results have been known to be one of the major factors of escalation of harrowing diseases. As reported, in developing countries, 80% of the diseases are water borne diseases, which have led to 5 million deaths and 2.5 billion illnesses. The most common of these diseases in Pakistan are diarrhoea, typhoid, gastroenteritis, cryptosporidium infections, some forms of hepatitis and giardiasis intestinal worms. In Pakistan, water borne diseases, cause a GDP loss of 0.6–1.44% every year. This makes it a pressing problem, particularly in a developing country like Pakistan. Water quality is currently estimated through expensive and time-consuming lab and statistical analyses, which require sample collection, transport to labs, and a considerable amount of time and calculation, which is quite ineffective given water is quite a communicable medium and time is of the essence if water is polluted with diseaseinducing waste. The horrific consequences of water pollution necessitate a quicker and cheaper alternative. In this regard, the main motivation in this study is to propose and evaluate an alternative method based on supervised machine learning for the efficient prediction of water quality in real-time. A representative set of supervised machine learning algorithms were employed on the said dataset for predicting the water quality

index (WQI) and water quality class (WQC). The main contributions of this study are summarized as follows. A first analysis was conducted on the available data to clean, normalize and perform feature selection on the

water quality measures, and therefore, to obtain the minimum relevant subset that allows high precision with low cost. In this way, expensive and cumbersome lab analysis with specific sensors can be avoided in further similar analyses. A series of representative supervised prediction (classification and regression) algorithms were tested on the dataset worked here. The complete methodology is proposed in the context of water quality numerical analysis.

1.2 PURPOSE

Water makes up about 70% of the earth's surface and is one of the most important sources vital to sustaining life. Rapid urbanization and industrialization have led to a deterioration of water quality at an alarming rate, resulting in harrowing diseases. Water quality has been conventionally estimated through expensive and time-consuming lab and statistical analyses, which render the contemporary notion of real-time monitoring moot. The alarming consequences of poor water quality necessitate an alternative method, which is quicker and inexpensive. With this motivation, this research explores a series of supervised machine learning algorithms to estimate the water quality index (WQI), which is a singular index to describe the general quality of water, and the water quality class (WQC), which is a distinctive class defined on the basis of the WQI. The proposed methodology employs four input parameters, namely, temperature, turbidity,

Ph and total dissolved solids. Of all the employed algorithms, gradient boosting, with a learning rate of 0.1 and polynomial regression, with a degree of 2, predict the WQI most efficiently, having a mean absolute error (MAE) of 1.9642 and 2.7273, respectively. Whereas multi-layer perceptron (MLP), with a configuration of (3, 7), classifies the WQC most efficiently, with an accuracy of 0.8507. The proposed methodology achieves reasonable accuracy using a minimal number of parameters to validate the possibility of its use in real time water quality detection systems.

2. LITERATURE REVIEW

2.1 EXISTING PROBLEM

The basic idea of this research is to devise a comprehensive methodology that analyzes and predicts the water quality of particular regions with the help of certain water quality parameters. These parameters include physical, biological, or chemical factors which influence water quality. There are certain quality standards set up by international organizations like the World Health Organization (WHO) and the Environmental Protection Agency (EPA), which serve as a benchmark for determining the quality of water. In its document “Efficient Water Quality Analysis and Prediction using Machine Learning”, EPA mentions a total of 101 parameters that affect water quality in one way or another. However, some parameters have a greater and more visible effect on water quality than others.

TITLE: IMPROVING THE ROBUSTNESS OF BEACH WATER QUALITY MODELING USING AN ENSEMBLE MACHINE LEARNING

AUTHOR: Wang et al (2021)

This study demonstrates the utility of using a model stacking approach for predictive modeling of beach water quality. Since model stacking averages out noise from its base models, it is theoretically more promising than individual models in generating predictions with greater accuracy and robustness. The results from this study suggest that the model stacking algorithm has promise for improving the reliability of predictive modeling for beach microbial water quality of other sites with similar hydrogeological and environmental conditions such as other beaches along the Great Lakes. A comprehensive test needs to be done to understand the strength and weaknesses of individual base models and the stacking approach. This study indicated that the model stacking approach may improve the robustness of beach water quality modeling.

TITLE: ACCURATE PREDICTION SCHEME OF WATER QUALITY IN SMART MARICULTURE WITH A DEEP BI-S-SRU LEARNING NETWORK

AUTHOR: J. Liu, C. Yu, Z. Hu et al (2020)

This paper proposed the process and model for the accurate prediction of key water quality parameters (pH, water temperature, and dissolved oxygen). Firstly, the collected water quality data is repaired and corrected by the improved preprocessing method, and then the data is filtered and denoised by the wavelet transform method. After preprocessing, the data received by remote transmission can be recovered well. Next, we construct the Bi-S-SRU (Bi-directional Stacked SRU) deep learning prediction model by

importing a pretreated dataset weighted with the discovered correlation coefficients. The experimental results demonstrate that our proposed prediction model can achieve higher prediction accuracy and stability compared with RNN-based and SRU-based prediction models. The experimental results also show that the Bi-S-SRU-based prediction method is only slightly higher in time complexity than the traditional RNN-based or LSTM-based prediction method.

**TITLE: ASSESSMENT OF SURFACE WATER QUALITY BY
USING SATELLITE IMAGES FUSION BASED ON PCA
METHOD IN THE LAKE GALA, TURKEY**

AUTHOR: E. Batur and D. Makita (2019)

In this paper, the PCA model is presented to integrate surface water reflectance values from satellite images to monitor Gala Lake's surface water quality. The values of Chl-a, DO, TSS, SDD, TDS, and pH values calculated by the PCA method were found to be highly correlated with the measured water quality parameters. The results obtained were found to be directly proportional to the number of sensors. L8 OLI and S2A have higher spectral resolution than GK2 images. However, the high temporal resolution of GK2 allows the desired region to be displayed at more frequent intervals, allowing for better monitoring of the instantaneous changes in surface water quality. Therefore, longer measurements should be made and analyzed for a model covering all periods.

**TITLE: SURFACE WATER POLLUTION DETECTION
USING THE INTERNET OF THINGS**

AUTHOR: Shafi et al (2018)

In this paper, the proposed an IoT-based solution to monitor water quality in real-time. The proposed system provides remote monitoring of water quality assessment along with water flow control via a mobile app. Four machine learning algorithms including Support Vector Machine (SVM), k Nearest Neighbor (kNN), single layer neural network, and deep neural network have been applied for the classification of water quality and experimental results revealed that deep neural network outperforms all other algorithms with an accuracy of 93. This system has the potential to effectively utilize to overcome the challenges of water quality in the agriculture sector and various industries.

**TITLE: IMPROVING WATER QUALITY INDEX PREDICTION
IN PERAK RIVER BASIN MALAYSIA THROUGH A
COMBINATION OF MULTIPLE NEURAL NETWORKS**

AUTHOR: Ahmad et al (2017)

In this paper, they proposed a reliable real-time prediction model for WQI developed through a selective combination of multiple neural networks by excluding COD and BOD from model inputs as they cannot be measured in real-time. Single and multiple FANN are used in this paper to model the WQI in the Perak River basin. The selective combination schemes provide models with better generalization capability compared to combining all neural networks. The bootstrap aggregated models with selective combination provide a real-time WQI prediction tool without delay as only

real-time measurements are used as model inputs.

TITLE: ARTIFICIAL INTELLIGENCE FOR THE PREDICTION OF WATER QUALITY INDEX IN GROUNDWATER SYSTEMS

AUTHOR: Mohamad Sakizadeh (2016)

One of the problems of ANN's modeling in environmental studies which suffers from the problem of the small data records is the danger of over-fitting the model to the training data resulting in poor generalization of the model for the data out-of-the training data range. This study's results proved that this problem can be obviated by using some algorithms like Bayesian regularization and Ensemble methods. The prediction of water quality index (WQI) was successfully implemented by Bayesian regularization and Ensemble averaging methods, though the performance of Bayesian regularization was roughly better, with minimum test error indicating the good generalization ability of these methods in this field. The poor generalization ability is a problem that has been overlooked by most of the research all around the world although it is an important issue that should be taken into account.

TITLE: THE USE OF COMBINED NEURAL NETWORKS AND GENETIC ALGORITHMS FOR THE PREDICTION OF RIVER WATER QUALITY

AUTHOR: Ding et al (2014)

In this paper, they propose a water quality prediction model that combines PCA, BPNN, and GA. Using the BPNN model to study water classification and prediction can overcome disadvantages including the large workload of traditional evaluation methods and strong subjectivity. This model possesses

objectivity, universality, and practicality. PCA converts the multi-indices into a few aggregative indices with little original data information loss and reduces the input data to speed the training process. Using GA to optimize network parameters can effectively prevent the search process from converging to local optimum solutions, optimize global optimal network parameters, and significantly improve the accuracy of water quality prediction. This model can obtain high training speed and good prediction rate and can be extended to other classification problems.

Our Ideology

The estimated water quality in our work is based on nine parameters: ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic carbon, Trihalomethanes, Turbidity, and pH, which are tested according to World Health Organization (WHO) standards.

The proposed methodology improves on these notions and the methodology being followed is depicted in Figure 1.

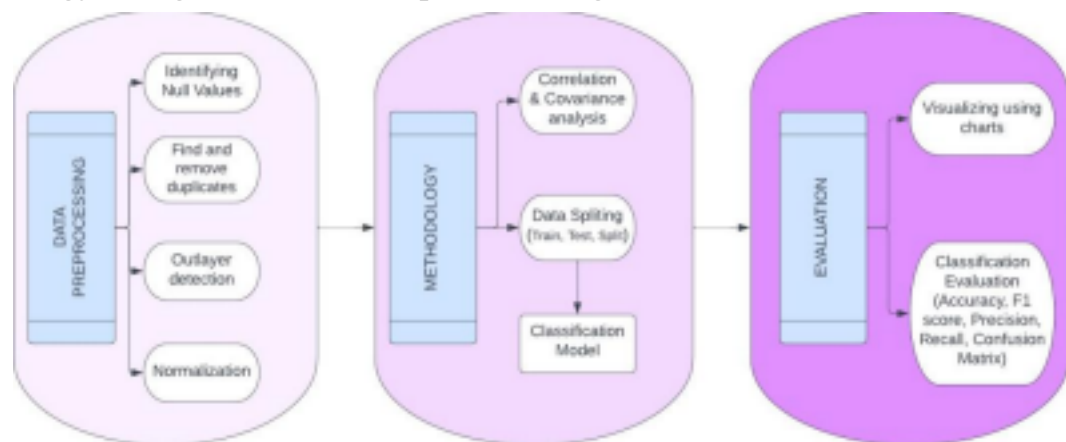


Figure 1

COMPARATIVE ANALYSIS OF LITERATURE SURVEY:

		Research er	Title	Parameters	Algorithm	Remarks
01		Wang et al	Improving the robustness of beach water quality modeling using an ensemble machine learning	turbidity, temperature, Culturable fecal indicator bacteria such as Escherichia coli (E. coli)	Partial least square, sparse partial least square, random forest, Bayesian network, Akhand linear regression	Highest accuracy of 82.3% with ensemble machine learning algorithm
02		Li, C. Z. Hu et al	Accurate prediction scheme of water quality in smart mariculture with a deep Bi-S SRU learning network	Salinity, chlorophyll, turbidity, Water Temperature, PH, Dissolved Oxygen(DO)	LSTM, SRU, RNN, LSTM, SRU and Bi-S SRU	Highest accuracy of 94.42% using a Bi-S-SRU
03	20	Batur and D. Makita	Assessment of surface water quality by using satellite images fusion based on PCA method in	DO, SDD, TDS, and pH Chl-a and TSS	MLR, SVM, ANN, AND PCA	Highest accuracy of 92% using a PCA-based RSR model

			the Lake Gala, Turkey			
04	3	Shafi et al	Surface Water Pollution Detection using the Internet of Things	turbidity, temperature and pH	Support Vector Machines (SVM), Neural Networks (NN), Deep Neural Networks (Deep NN), and k Nearest Neighbors (kNN)	Highest accuracy of 93% with Deep NN

05		17 had al	Improving water quality index prediction in Perak River basin Malaysia through a combination of multiple neural networks	Nitrate, PH, Electrical conductivity, Dissolved oxygen, total coliform, Biochemical Oxygen Demand	feedforward artificial neural network; forward selection; backward elimination; artificial neural network; multiple neural networks	Highest accuracy of 92.7% using a selective combination methods
06		adeh	Artificial intelligence for the prediction of water quality index in groundwater systems	EC, TDS, Mn, Cu, Cr(VI), Turbidity, pH, Ca, Mg, Total hardness, Sulfate, Fe, Fluoride Phosphate, Nitrate, Nitrite	ANN with Bayesian regularization	Highest accuracy of 80% using an Artificial Neural Network
07		Ding et	The Use of Combined Neural Networks and Genetic Algorithms for Prediction of River Water Quality	pH, NH ₃ -N, TN, Cr ⁶⁺ , TP, CODMn, BOD ₅ , TCN, COD, Cd, Cu, Zn, Pb, Hg, As, Se, F-, sulfide, dissolved oxygen, and LAS, etc.	Genetic Algorithm (GA), and Back Propagation Neural Network (BPNN)	The highest accuracy of Non polluted and polluted of 88.9% and 93.1% with PCA technique

2.2 REFERENCES

1. Wang Et Al (2021) - Improving the Robustness of Beach Water Quality Modeling using an Ensemble Machine Learning
2. J. Liu, C. Yu, Z. Hu Et Al (2020) - Accurate Prediction Scheme of Water Quality in Smart Mariculture with A Deep Bi-S-Sru Learning Network
3. E. Batur and D. Makita (2019) - Assessment of Surface Water Quality by using Satellite Images Fusion based on PCA Method in the Lake Gala, Turkey
4. Shafi Et Al (2018) - Surface Water Pollution Detection using The Internet of Things
5. Ahmad Et Al (2017) - Improving Water Quality Index Prediction in Perak River Basin Malaysia Through a Combination of Multiple Neural Networks
6. Mohamad Sakizadeh (2016) - Artificial Intelligence for the Prediction of Water Quality Index in Groundwater Systems
7. Ding Et Al (2014) - The Use of Combined Neural Networks and Genetic Algorithms for the Prediction of River Water Quality

2.3 PROBLEM STATEMENT DEFINITION

posed methodology achieves reasonable accuracy using a minimal number of parameters to validate the possibility of its use in real-time water quality detection systems.	
What are the boundaries of the problem?	There is no boundary limit for the issue because if anyone drinks unpurged or contaminated water, they

	will be affected.
What is the issue?	The most important behavioral risk factors of this disease can only be identified by taking samples of the contaminated water and then researching that water by using datasets and then only we can find the issue.
Where is the issue coming from?	It majorly occurs to the people on the riverside who use the river water. If the water had any harmful chemicals present, it would affect the people with a disease.

Why is it important that we fix the

problem?

It is very crucial to develop an

application that detects the disease because rapid urbanization and industrialization have led to a deterioration of water quality at an alarming rate, resulting in harrowing diseases. Water quality has been conventionally estimated through expensive, time-consuming lab and statistical analyses. In this, we are simply doing the project to find the chemicals using data science.

Which solution can be

used to address this

issue?

This study aims to predict water

quality components using Bi-S-SRU

(Bi-directional Stacked SRU) deep learning prediction model.

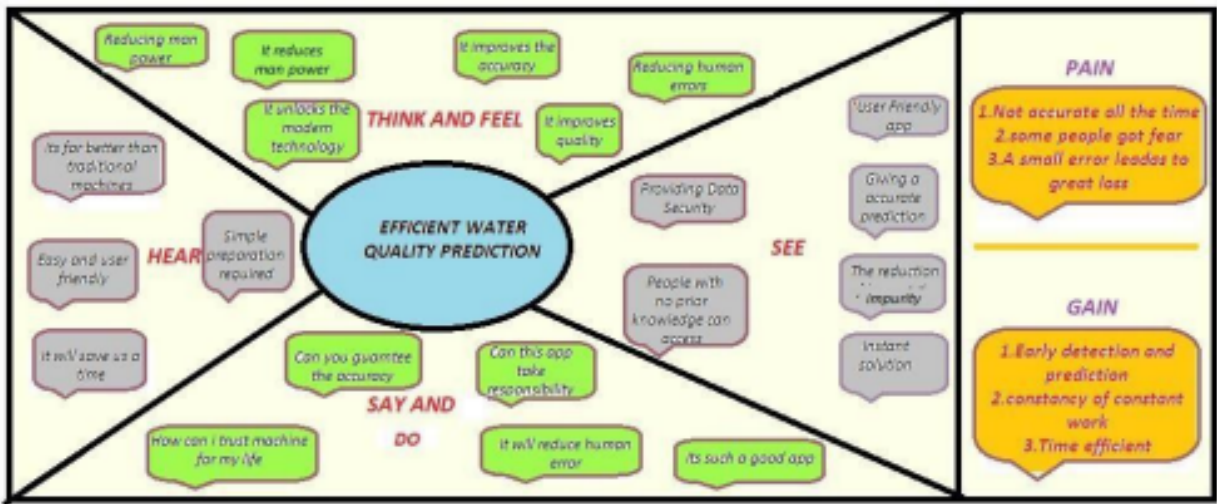
What methodology was used to solve the issue?

The estimated water quality in our work is based on nine parameters: pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic

	carbon, Trihalomethanes, Turbidity, and pH.
--	--


3. IDEATION & PROPOSED SOLUTION

3.1 EMPATHY MAP CANVAS






3.2 IDEATION & BRAINSTORMING

Brainstorm & Idea Prioritization Template: Step-1: Team Gathering, Collaboration and Select the Problem Statement




Brainstorm & idea prioritization

Use this template in your own brainstorming sessions so your team can unleash their imagination and start shaping concepts even if you're not sitting in the same room.

 10 minutes to prepare
 1 hour to collaborate
 3-8 people recommended

Before you collaborate
A little bit of preparation goes a long way with this session. Here's what you need to do to get going.

 10 minutes


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

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

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

[Open agenda](#)







1 Define your problem statement
What problem are you trying to solve? Frame your problem as a How Might We statement. This will be the focus of your brainstorm.

 5 minutes

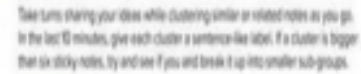
Exercise

How might we [your problem statement]?

Key rules of brainstorming
To run a smooth and productive session

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

Step-2: Brainstorm, Idea Listing and Grouping



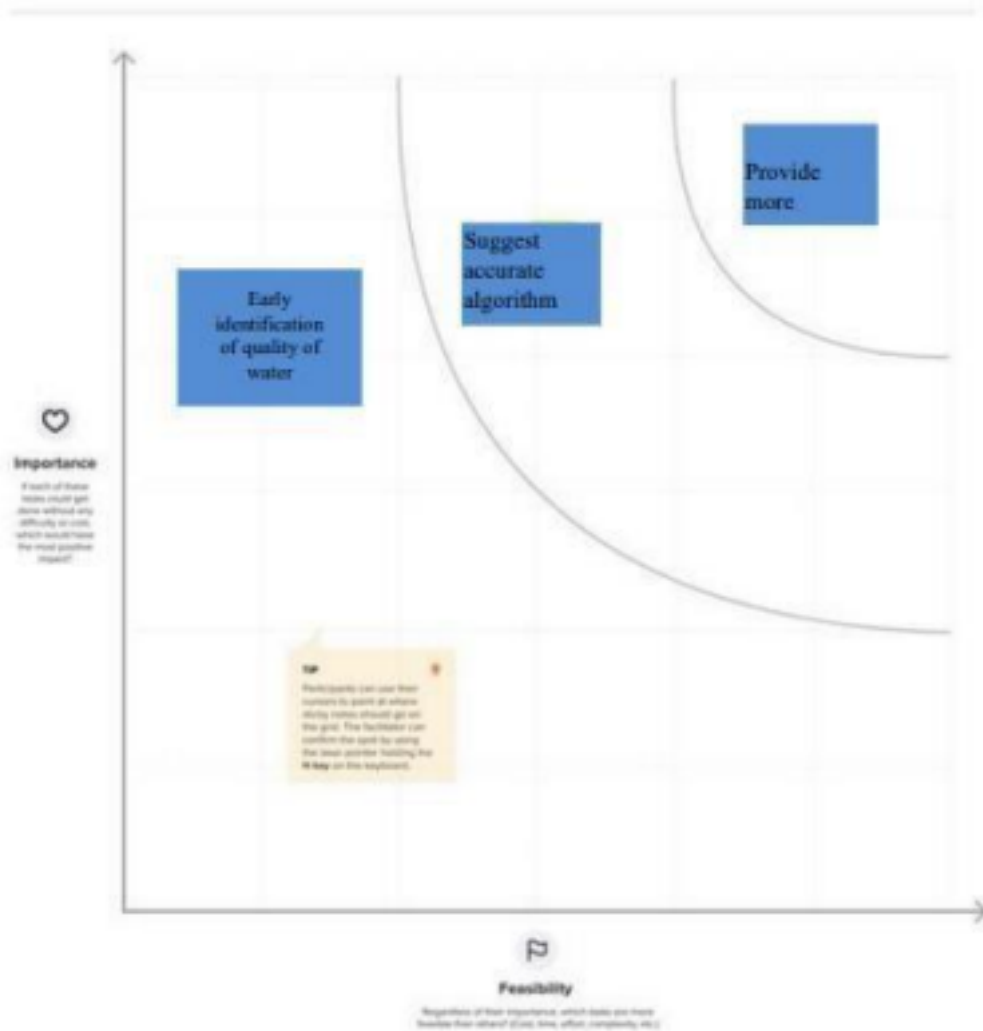
Step-3: Idea Prioritization



Prioritize

Your team should all be on the same page about what's important moving forward. Place your ideas on this grid to determine which ideas are important and which are feasible.

⌚ 30 minutes



Step

3.3 PROPOSED SOLUTION

3.	Novelty / Uniqueness	<ul style="list-style-type: none">• Using the model, it is possible to determine whether the water is suitable for drinking. Therefore, it contributes to the maintenance of health.
4.	Social Impact / Customer Satisfaction	<ul style="list-style-type: none">• Water makes up about 70% of the earth's surface and is one of the most important sources vital to sustaining life. Rapid urbanization and industrialization have led to a deterioration of water quality at an alarming rate, resulting in harrowing diseases.• Most of the research either employed manual lab analysis, not estimating the water quality index standard, or used too many parameters to be efficient enough.• With machine learning techniques, the implementation was done by the Water Quality Index (WQI).• Web app is developed as UI is provided for the customer/user where he has to enter the values for predictions.

5.	Business Model (Revenue Model)	<ul style="list-style-type: none"> • A web application that is integrated to the model built. A UI is provided for the uses where he has to enter the values for predictions. The enter values are given to the saved model and prediction is showcased on the UI and deploy it on IBM cloud. • We can sell it for the prediction of water in various environments if the model preforms well ,also can make the app as premium one.
----	--------------------------------	--

6.	Scalability of the Solution	<ul style="list-style-type: none"> • The proposed can be implemented in realtime water quality analysis by getting water sample using devices(Internet Of Things). • Real time applications can be used in various places like schools,colleges etc. • Machine learing model integrated with IOT can make users more comfortable and to use in realtime.
----	-----------------------------	---

3.4 PROBLEM SOLUTION FIT

<p>CS</p> <p>1.CUSTOMER SEGMENT(s)</p> <p>People, Residential, Commercial, Lab Testing</p>	<p>CC</p> <p>6.CUSTOMER CONSTRAINTS</p> <p>Water is essential for every one to sustain. If the water is impure it may cause diseases with this application it can be avoided.</p>	<p>AS</p> <p>5.AVAILABLE SOLUTION</p> <p>we need to train the datasets to run smoothly and see an incremental improvement in the prediction rate using Random Forest Regression algorithm on our dataset</p>
<p>J&P</p> <p>2.JOB-TO-BE DONE/PROBLEMS</p> <p>Check the quality of water, whether the water is drinkable, reason for un usability. Can verify the quality by themselves without expert</p>	<p>RC</p> <p>9.PROBLEM ROOT CAUSE</p> <p>The major cause of this problem is lack of drinking water and doesn't follow the proper diet and doesn't have proper awareness is also being a root cause.</p>	<p>BE</p> <p>7.BEHAVIOUR</p> <p>We will be building a web application that is integrated to the model built. The enter values are given to the saved model and prediction is showcased on the UI</p>
<p>TR</p> <p>3. TRIGGERS</p> <p>Using this application, user can avoid the fear of water quality. Since the user knows the quality of <u>water they are going to use.</u></p>	<p>SL</p> <p>10.YOUR SOLUTION</p> <p>The heart of the project depends upon the prediction of the quality of the water. As abundant as algorithms are</p>	<p>CH</p> <p>8.CHANNEL OF BEHAVIOUR</p> <p>Online: The application Notify the user with data preprocessing information</p>

4. EMOTIONS:BEFORE/AFTER

Before: There are no application to predict the water quality.

After: By using this easy to predict the quality of water using some a parameters. present in order to achieve such a goal, it is mandatory to select the best and the most efficient algorithm to finalize the predicted value.

Offline: Water quality has been conventionally estimated through expensive and time-consuming lab and statistical analyses, which render the contemporary notion of real-time monitoring moot.

4. REQUIREMENT ANALYSIS

4.1 Functional Requirements:

Following are the functional requirements of the proposed solution.

	R No. Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Interface	A detailed description of water quality should be provided.
FR-2	User Form	Values and measures require to predict the Water quality should be given as input in the form.
FR-3	Machine Learning Model Deployment	Develop the Machine Learning Regression Model to predict the Water Quality Index (WQI). Develop the Machine Learning Classification Model to predict the Water Quality Classification (WQC).
FR-4	Testing The Water Samples	Provides an option to test any kind of water samples with the required parameters and to calculate the Water Quality Index and impurities present
FR-5	Reporting	If any issues are faced by the customer or user it will be directly notified to the developer

4.2 Non-functional Requirements:

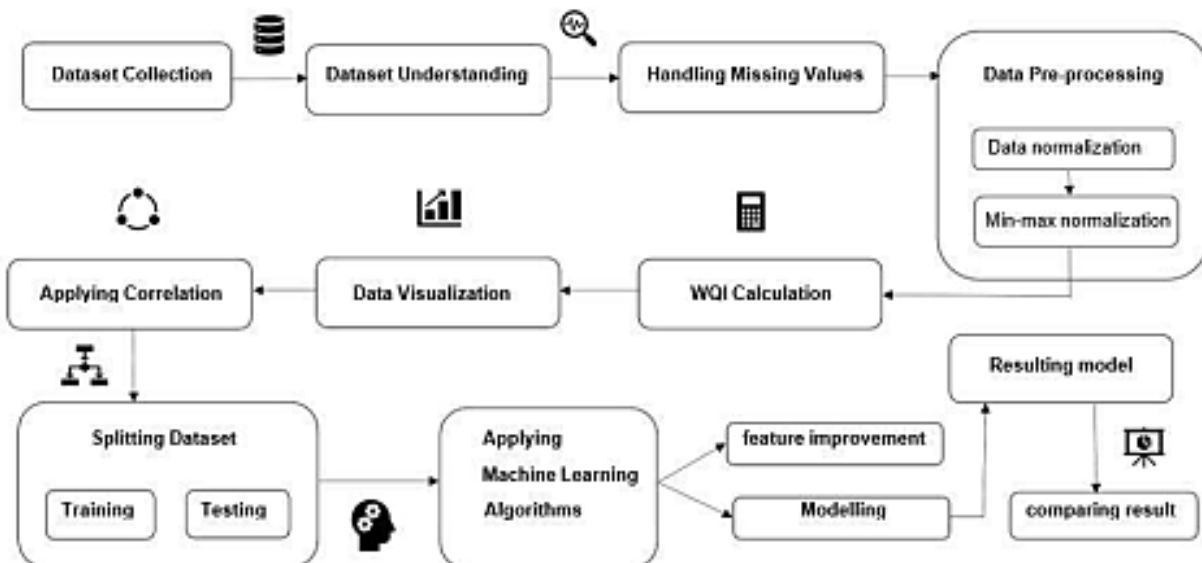
Following are the non-functional requirements of the proposed solution.

	Non-Functional Requirement	Description
NFR-1	Usability	Customers can access the system more efficiently and in a simpler way. The customers can have the opportunity to view a better interpretation of results. The customers are also recommended the purification techniques based on the impurities.
NFR-2	Security	All the predicted information is accessed only by the authenticated users
NFR-3	Reliability	It should be reliable in producing effective and efficient water quality prediction results. It should ensure the trust and belief among people that this water quality prediction system produces correct results when used.
NFR-4	Performance	The system should be consistent in producing the prediction results of the Water Quality Index (WQI) and also needs to ensure better throughput and response time compared to other systems.

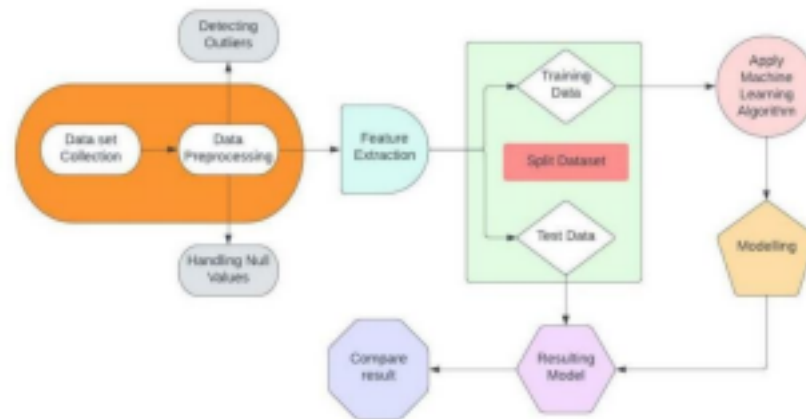
NFR-5	Availability	The system can be utilized by the customers 24/7 and it should be availed to test any kind of water samples anywhere
NFR-6	Scalability	It can be used by a wide variety of users like testing agencies, private and public laboratories, restaurants and hotels, and people who wish to test the quality of water they consume. The system should also be compatible enough to be integrated with future technologies also.

5.PROJECT DESIGN

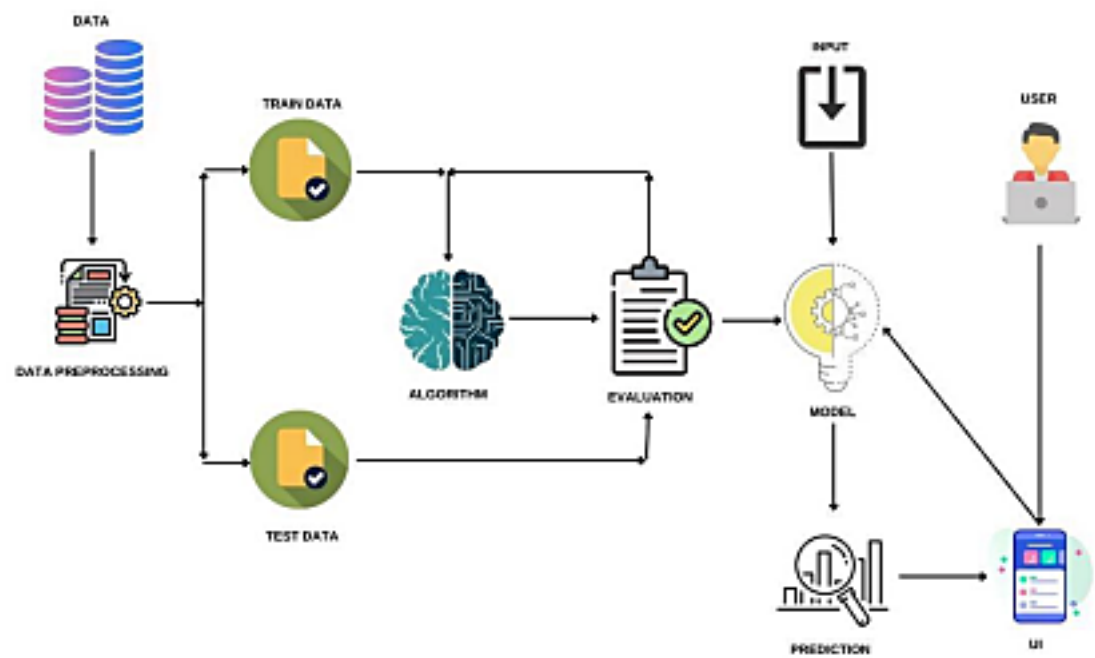
5.1 DATA FLOW DIAGRAMS



5.2 SOLUTION AND TECHNICAL ARCHITECTURE



Solution Architecture



Technical Architecture

Table-1: Components & Technologies:

S. No	Component	Description	Technology
1.	User Interface	User interacts by using web user interface.	HTML, CSS and Python Flask
2.	Application Logic-1 (Login)	User can able to login if that person is already registered to the site.	HTML, CSS and Python Flask
3.	Application Logic-2 (Register)	User needs to be registered if that person is new to the site.	HTML, CSS and Python Flask.
4.	Application Logic 3(Reporting Form)	User needs to click on the reporting form in order to get the prediction result	Front end- HTML, CSS and Python Flask. Back end – Query Languages, Python.
5.	Database	Data Type-String, Numeral values.	Query Languages such as MySQL, NoSQL etc.
6.	Cloud Database	Database Service on Cloud.	IBM DB2, IBM Cloud ant etc.
7.	File Storage	File storage requirements.	Local File-system.
8.	External API-1	Anyone can access the details with some restrictions to the personal	Web API.

		details of other users.	
9.	External API-2	Accessibility.	Aadhar API.
10.	Machine Learning Model	Predict the result based on the training and testing dataset.	Data Recognition Model, etc.
11.	Infrastructure (Server / Cloud)	Application Deployment on Local System.	Local System.

Table-2: Application Characteristics:

S. No	Characteristics	Description	Technology
1.	Open-Source Frameworks	Frameworks are used for predictive data analysis, providing clear and actionable error messages.	Tensor flow, Sci-kit learn, Keras.
2.	Security Implementations	OTP will be sent to the registered email id. Unauthorized users could not access the user's details.	Email Verification.
3.	Scalable Architecture	Scalability is improved for implementing the three tier architecture.	Three tier architecture.
4.	Availability	For enhancing the high availability, load balancer is needed.	Load Balancer.
5.	Performance	The model could be able to process large number of datasets.	Load Balancer.

5.3 USER STORIES

Us er Type	Functional Requireme nt (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priori ty	Relea se
------------------	--------------------------------------	-------------------------	----------------------	------------------------	--------------	-------------

Peop le (web user)		USN-1	As a user, I can understand the detailed description of water quality on the home page	I can access the web page	High	Sprint-1
	Input form	USN-2	As a user, I can enter the details required to analysis the water quality with use of form provided in the web page.	I can give inputs in the form and it is processed and visualize the water quality.	High	Sprint-2
		USN-3	As a user, I can contact the Customer care (people at the water resource organisation) to know the details of water	I can contact people with Whatsapp, instagram, twitter, mail and also I can make call		n Sprint-

6. PROJECT PLANNING AND SCHEDULING

6.1 SPRINT PLANNING AND ESTIMATION

	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint 1	Data Preparation	USN-1	Collecting water dataset and preprocessing it	10	High	AJAY G JANA R M
Sprint 1	Model Building	USN-2	Create an ML model to predict water quality	5	Medium	AJAY G JANA R M KAVIN N GOPI R

Sprint1	Model Evaluation	USN-3	Calculate the performance, error rate, and complexity of the ML model and evaluate the dataset based on the parameter that the dataset consists of.	5	Medium	
Sprint2	Model Deployment	USN-4	As a user, I need to deploy the model and need to find the results.	20	Medium	
Sprint3	Web page (Form)	USN-5	As a user, I can use the application by entering the water dataset to analyze or predict the results.	20	Medium	AJAY G JANA R M KAVIN N GOPI R

Sprint4	Dashboard	USN-6	As a user, I can predict the water quality by clicking the submit button and the application will show whether the water is efficient for use or not.	20	High	AJAY G JANA R M
---------	-----------	-------	---	----	------	--------------------

Project Tracker:

Sprint	Story Points	Duration	Sprint Start Date	Sprint End Date	Completed	Sprint Release Date
Sprint-1	20	6 Days	23 Oct 2022	28 Oct 2022	20	29 Oct 2022
Sprint-2	20	7 Days	29 Oct 2022	04 Nov 2022	20	05 Nov 2022
Sprint-3	20	7 Days	05 Nov 2022	11 Nov 2022	20	12 Nov 2022
Sprint-4	20	8 Days	12 Nov 2022	19 Nov 2022	20	19 Nov 2022

Total Story Points

Velocity:

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

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

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

4: 1 user stories x 20 story points = 20

Total = 80 The average sprint velocity is

$80 \div 4 = 20$.

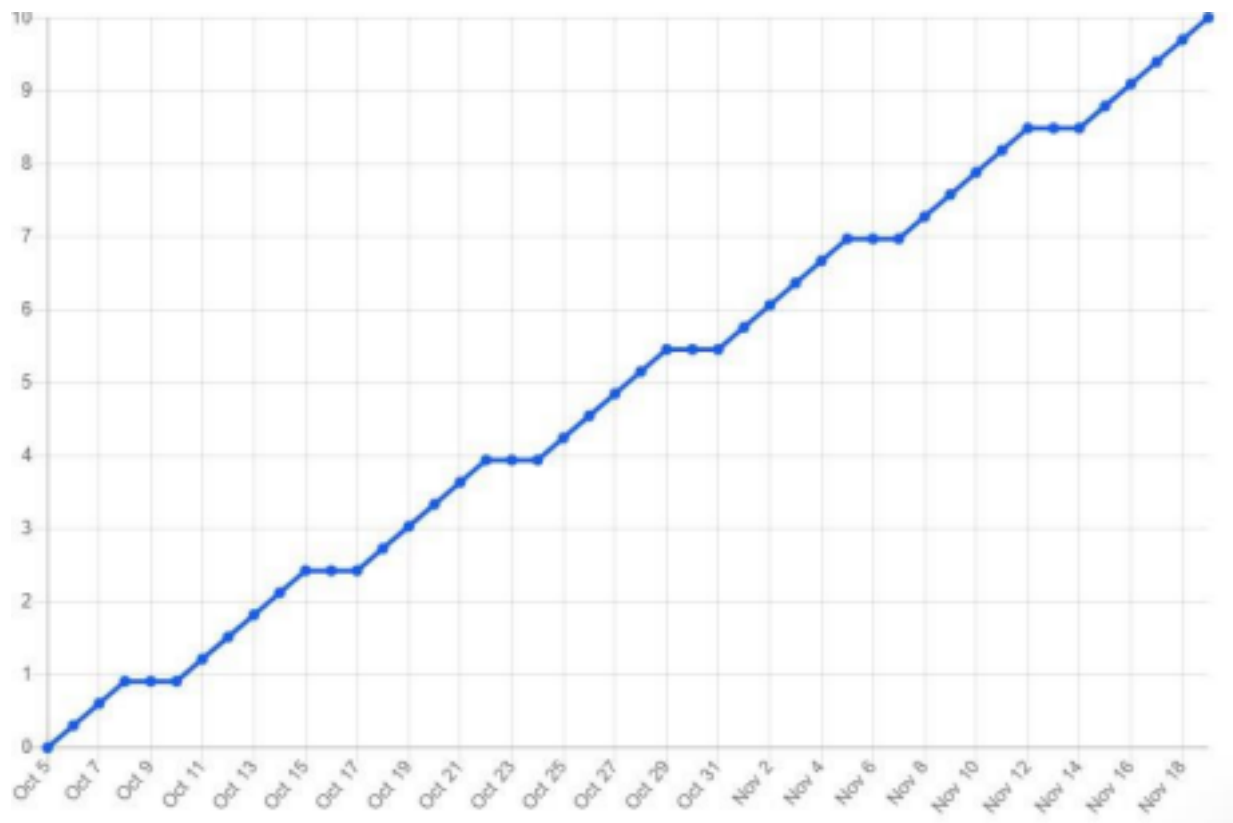
6.2 PROJECT DELIVERY SCHEDULE

TITLE	DESCRIPTION	DATE
Literature Survey & Information Gathering	Literature survey on the selected project & gathering information by referring the technical papers, research publications, journals etc.	1 SEPTEMBER 2022
Prepare Empathy Map	Prepare Empathy Map Canvas to capture the user Pains and Gains, prepare list of problem Statements that are to be solved by this project.	7 SEPTEMBER 2022 & 9 SEPTEMBER 2022

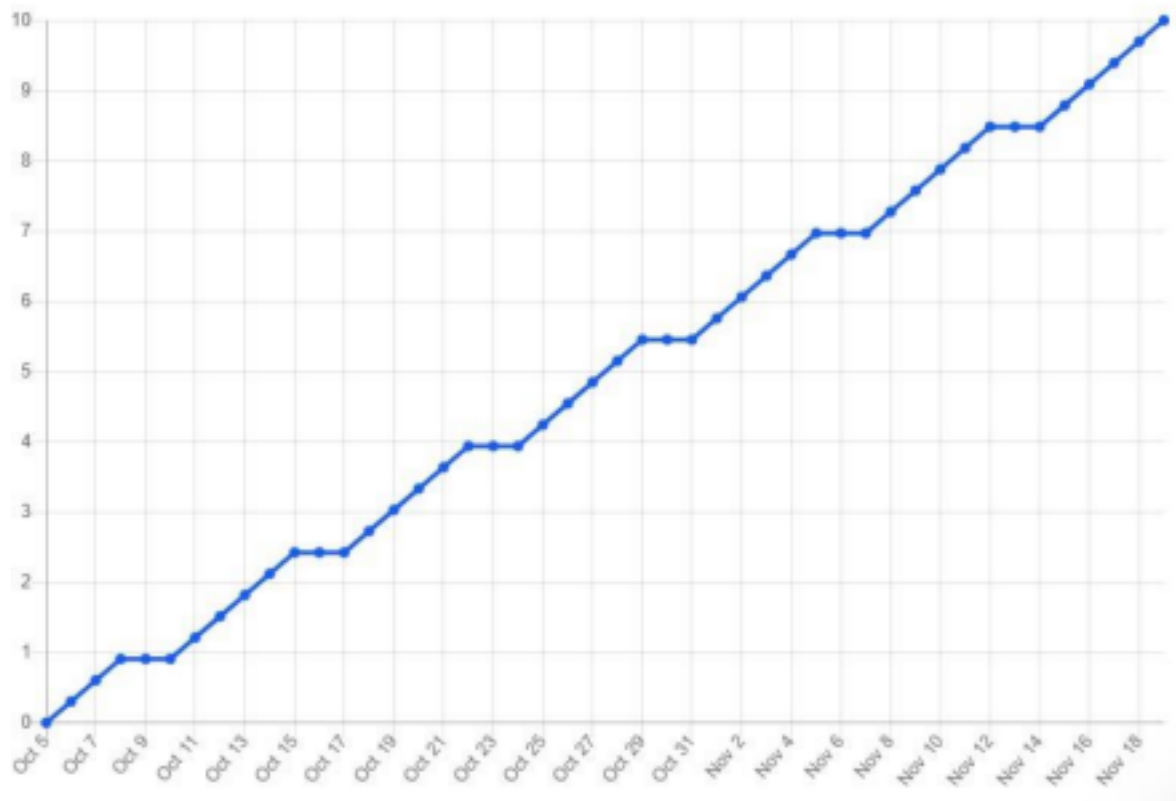
Ideation	List the ideas by organizing a brainstorming session and prioritize the top three ideas based on the feasibility and importance.	15 SEPTEMBER 2022
Proposed Solution	Prepare the proposed solution document, which includes novelty, feasibility of idea, revenue model, social impact, scalability of solution, etc.	22 SEPTEMBER 2022
Problem Solution Fit	Prepare problem - solution fit document.	30 SEPTEMBER 2022
Solution Architecture	Prepare solution architecture document.	30 SEPTEMBER 2022
Customer Journey	Prepare the customer journey maps to understand the user interactions and experiences with the application (entry to exit).	6 OCTOBER 2022
Functional Requirement	Prepare the functional requirement document.	11 OCTOBER 2022
Data Flow Diagrams and User_Stories	Prepare the Data flow diagrams and User Stories for the problem	14 OCTOBER 2022
Technology Stack Architecture	Prepare the Technology Stack Architecture	17 OCTOBER 2022

Prepare Milestone &Activity List	Prepare the milestones and activity list of the project.	21 OCTOBER 2022
Project Development Phase	Develop Project Development Phase which include Sprint 1, Sprint 2, Sprint 3, Sprint 4	ON PROGRESS....

6.3 REPORTS FROM JIRABURNDOWN CHART



BURNUP CHART



7.CODING AND SOLUTIONING

7.1 FEATURE 1 (RANDOM FOREST ALGORITHM MODEL)

Random Forest Classifier is used to train and test the model for detecting the Chronic Kidney Disease (CKD) with the help of collected and pre-processed dataset collections. NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of highlevel mathematical functions to operate on these arrays. Moreover, NumPy forms the foundation of the Machine Learning stack. Pandas is an open-source Python

package that is most widely used for data science/data analysis and machine learning tasks. Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. For a brief introduction to the ideas behind the library, you can read the introductory notes or the paper. Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible. Create publication quality plots. Make interactive figures that can zoom, pan, update. EDA is applied to investigate the data and summarize the key insights. It will give you the basic understanding of your data, its distribution, null values and much more. You can either explore data using graphs or through some python functions. There will be two types of analysis. Descriptive statistics are brief informational coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability. Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables, kurtosis, and Skewness. Label Encoding refers to converting the labels into a numeric form to convert them into the machine readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for

the structured dataset in supervised learning. “Pickling” is the process whereby a Python object hierarchy is converted into a byte stream, and “unpickling” is the inverse operation, whereby a byte stream is converted back into an object hierarchy. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework.

7.2 FEATURE 2(FLASK CONNECTIVITY)

The framework is the basis upon which software programs are built. It serves as a foundation for software developers, allowing them to create a variety of applications for certain platforms. It is a set of functions and predefined classes used to connect with the system software and handle inputs and outputs. It simplifies the life of a developer while giving them the ability to use certain extensions and makes the online applications scalable and maintainable. Flask is a web application framework written in Python. A Web Application Framework or simply a Web Framework represents a collection of libraries and modules that enable web application developers to write applications without worrying about low-level details such as protocol, thread management, among other examples. Flask is a web application framework written in Python. It was developed by Armin Ronacher, who led a team of international Python enthusiasts

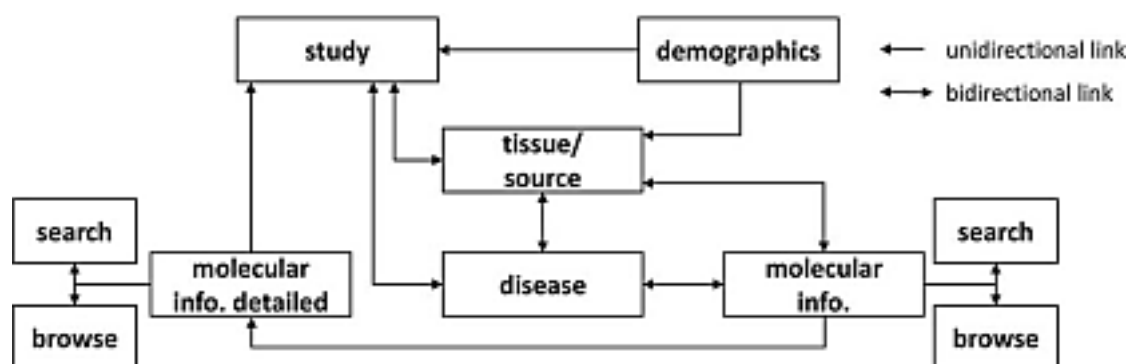
called Pocco. Flask is based on the Werkzeug WSGI toolkit and the Jinja2 template engine. Both are Pocco projects. The Web Server Gateway Interface (Web Server Gateway Interface, WSGI) has been used as a standard for Python web application development. WSGI is the specification of a common interface between web servers and web applications. Flask is often referred to as a micro-framework. It is designed to keep the core of the application simple and scalable. Instead of an abstraction layer for database support, Flask supports extensions to add such capabilities to the application. Unlike the Django framework, Flask is very Pythonic.

It's easy to get started with Flask, because it doesn't have a huge learning curve. HTML stands for Hyper Text Markup Language. HTML is the standard markup language for creating Web pages. HTML describes the structure of a Web page. HTML consists of a series of elements. HTML elements tell the browser how to display the content. Flask is used for developing web applications using python, implemented on Werkzeug and Jinja2. Advantages of using Flask framework are: There is a built-in development server and a fast debugger provided. The model deployed using Flask is used to predict the Chronic Kidney Disease. Hypertext markup language (HTML) is the basic language used to create documents for the Web and, along with HTTP (hypertext transfer protocol) and URLs (universal resource locators), is one of the three main

protocols of the Web. Hypertext is text that contains hyperlinks. A hyperlink is an automated cross-reference to another location on the same document or to another document which, when selected by a user, causes the computer to display the linked location or document within a concise period. A markup language is a set of tags that can be embedded in digital text to provide additional information about it, including its content, structure and appearance. This information facilitates automated operations on the text, including formatting it for display, searching it and even modifying it. Some type of markup language is employed by every word processing program and by nearly every other program that displays text, although such languages and their tags are typically hidden from the user. HTML consists of a set of predefined tags that can be embedded in text by web site designers in order to indicate the details of how web pages are rendered (i.e., converted into a final, easily usable, form) by web browsers. These details include paragraphing, margins, fonts (including style and size), columns, colors (background and text), links, the location of images, text flow around images, tables, and user input form elements (such as spaces for adding text and submit buttons).

7.3 DATABASE SCHEMA

In the recent decades, the evolution of omics technologies has led to advances in all



biological fields, creating a demand for effective storage, management and exchange of rapidly generated data and research discoveries. To address this need, the development of databases of experimental outputs has become a common part of scientific practice in order to serve as knowledge sources and data-sharing platforms, providing information about genes, transcripts, proteins or metabolites. In this review, we present omics databases available currently, with a special focus on their application in kidney research and possibly in clinical practice. Databases are divided into two categories: general databases with a broad

information scope and kidney-specific databases distinctively concentrated on kidney pathologies. In research, databases can be used as a rich source of information about pathophysiological mechanisms and molecular targets. In the future, databases will support clinicians with their decisions, providing better and faster diagnoses and setting the direction towards more preventive, personalized medicine. We also provide a test case demonstrating the potential of biological databases in comparing multi-omics datasets and generating new hypotheses to answer a critical and common diagnostic problem in nephrology practice. In the future, employment of databases combined with data integration and data mining should provide powerful insights into unlocking the mysteries of kidney disease, leading to a potential impact on pharmacological intervention and therapeutic disease management.

8.TESTING

8.1 TEST CASES

Test Case ID			Test Case Description	Test the Water quality Prediction Functionality		
Created By		AJAY G	Reviewed By	JANA R M		
Tester's Name		GOPI R	Date Tested	November 15, 2022	Test Case (Pass/Fail / Not Executed)	Pa s s
		KAVIN N		S #	Test Data	
S #	Prerequisites:			1	By Clicking the website link	
1	Access to Chrome Browser			2	Details should be in a integer format	
2	Entering the details required			3	Data should be filled	
3	check for correct values			4	Provide the datasets for model training	
4	Application to train the model					

Test Scenario Verify whether the deployed project predicts as per expected

Step #	Step Details	Expected Results	Actual Results	Pass / Fail / Not executed / Suspended
1	Navigate to corresponding website link	Site should open	As Expected	Pass
2	Enter the details	Details should be entered	As Expected	Pass

3	Click Submit	Check the result	As Expected	Pass
4	Output results	Results are generated	As Expected	Pass

8.2 USER ACCEPTANCE TESTING TEST

CASE 1:



The image shows a web application interface for "Water Quality Analysis". The background is a close-up of green leaves with water droplets. A semi-transparent white form is centered on the screen, containing the following fields and a button:

- Enter Year:
- Enter D.O:
- Enter PH:
- Enter Conductivity:
- Enter B.O.D:
- Enter Nitrate/n:
- Enter Total Coliforms:
- Predict:

TEST CASE 2:

A screenshot of a web application titled "Water Quality Analysis". The form is overlaid on a background image of green leaves with water droplets. The form contains several input fields with pre-filled values: "Enter Year" (2014), "Enter D.O" (6.5), "Enter PH" (7), "Enter Conductivity" (75), "Enter B.O.D" (20), "Enter Nitrogen" (3), and "Enter Total Coliform" (23). Below these fields is a "Predict" button. At the bottom of the form, it says "Fair The Predicted Value is 72.41".

Water Quality Analysis

Enter Year 2014

Enter D.O 6.5

Enter PH 7

Enter Conductivity 75

Enter B.O.D 20

Enter Nitrogen 3

Enter Total Coliform 23

Predict

Fair The Predicted Value is 72.41

9.RESULTS

9.1 PERFORMANCE METRICES

TITLE	DESCRIPTION	DATE
Literature Survey & Information Gathering	Literature survey on the selected project & gathering information by referring the technical papers, research publications, journals etc.	1 SEPTEMBER 2022
Prepare Empathy Map	Prepare Empathy Map Canvas to capture the user Pains and Gains, prepare list of problem Statements that are to be solved by this project.	7 SEPTEMBER 2022 & 9 SEPTEMBER 2022

Ideation	List the ideas by organizing a brainstorming session and prioritize the top three ideas based on the feasibility and importance.	15 SEPTEMBER 2022
Proposed Solution	Prepare the proposed solution document, which includes novelty, feasibility of idea, revenue model, social impact, scalability of solution, etc.	22 SEPTEMBER 2022
Problem Solution Fit	Prepare problem - solution fit document.	30 SEPTEMBER 2022
Solution Architecture	Prepare solution architecture document.	30 SEPTEMBER 2022
Customer Journey	Prepare the customer journey maps to understand the user interactions and experiences with the application (entry to exit).	6 OCTOBER 2022
Functional Requirement	Prepare the functional requirement document.	11 OCTOBER 2022
Data Flow Diagrams and User_Stories	Prepare the Data flow diagrams and User Stories for the problem	14 OCTOBER 2022
Technology Stack Architecture	Prepare the Technology Stack Architecture	17 OCTOBER 2022
Prepare Milestone &Activity List	Prepare the milestones and	21 OCTOBER 2022

	activity list of the project.	
--	-------------------------------	--

Project Development Phase	Develop Project Development Phase which include Sprint 1, Sprint 2, Sprint 3, Sprint 4	ON PROGRESS....
----------------------------------	--	-----------------

10. ADVANTAGES AND DISADVANTAGES

10.1 ADVANTAGES:

Whether it be for groundwater, surface water or open water, there are a number of reasons why it is important for you to undertake regular water quality testing. If you're wanting to create a solid foundation on which to build a broader water management plan, then investing in water quality testing should be your first point of action. This testing will also allow you to adhere to strict permit regulations and be in compliance with Australian laws. Identifying the health of your water will help you to discover where it may need some help. Ultimately, finding a source of pollution, or remaining proactive with your monitoring will enable you to save money in the long term. The more information that you can obtain will assist you with your decision on what product you may need to improve the condition of your water. Simply guessing and buying products based on a hunch or a general

trend is ill-advised, as each body of water has unique properties that can only be discovered through testing. Measuring the amount of dissolved oxygen in your water is another important advantage of water quality testing, as typically the less oxygen, the higher the water temperature, resulting in a more harmful environment for aquatic life. These levels do fluctuate slightly across the seasons, but regular monitoring of your water quality will allow you to discover trends over time, and whether there are other factors that may be contributing to the results you discover.

10.2 DISADVANTAGES

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

11. CONCLUSION

Water is one of the most essential resources for survival and its quality is determined through WQI. Conventionally, to test water quality, one has to go

through expensive and cumbersome lab analysis. This research explored an alternative method of machine learning to predict water quality using minimal and easily available water quality parameters. The data used to conduct the study were acquired from PCRWR and contained 663 samples from 12 different sources of Rawal Lake, Pakistan. A set of representative supervised machine learning algorithms were employed to estimate WQI. This showed that polynomial regression with a degree of 2, and gradient boosting, with a learning rate of 0.1, outperformed other regression algorithms by predicting WQI most efficiently, while MLP with a configuration of (3, 7) outperformed other classification algorithms by classifying WQC most efficiently. In this paper, the performance of artificial intelligence techniques were evaluated to predict the water quality components of Tیره River (Iran). To this end most dataset related well-known components, such as pH, SO₄, Na, Ca, Cl, Mg, HCO₃ etc., were collected. Results indicated that the applied models have suitable performance for predicting water quality.

12. FUTURE SCOPE

In future works, we propose integrating the findings of this research in a largescale IoT-based online monitoring system using only the sensors of the required parameters. The tested algorithms would predict the water quality immediately based on the real-time data fed from the IoT system. The proposed IoT system would employ the parameter sensors of pH, turbidity, temperature and TDS for parameter readings and communicate those readings

using an Arduino microcontroller and ZigBee transceiver. It would identify poor quality water before it is released for consumption and alert concerned authorities. It will hopefully result in curtailment of people consuming poor quality water and consequently de-escalate harrowing diseases like typhoid and diarrhea. In this regard, the application of a prescriptive analysis from the expected values would lead to future facilities to support decision and policy makers. More data sources are required to verify the reliability and robustness of the proposed models. So far, the water quality dataset from the LVW collected by Southern Nevada Water Authority and Las Vegas Wash Coordination Committee, and dataset collected from Boulder Basin have been used as the experimental dataset. In the future, more efforts will be made to find more datasets to build a more reliable water quality prediction model.

13.APPENDIX

SOURCE CODE

Machine learning has been widely used as a powerful tool to solve problems in the water environment because it can be applied to predict water quality, optimize water resource allocation, manage water resource shortages, etc. Despite this, several challenges remain in fully applying machine learning approaches in this field to evaluate water quality:

(1) Machine learning is usually dependent on large amounts of high-quality

data. Obtaining sufficient data with high accuracy in water treatment and management systems is often difficult owing to the cost or technology limitations.

(2) As the conditions in real water treatment and management systems can be extremely complex, the current algorithms may only be applied to specific systems, which hinders the wide application of machine learning approaches.

(3) The implementation of machine learning algorithms in practical applications requires researchers to have certain professional background knowledge.

To overcome the above-mentioned challenges, the following aspects should be considered in future research and engineering practices:

(1) More advanced sensors, including soft sensors, should be developed and applied in water quality monitoring to collect sufficiently accurate data to facilitate the application of machine learning approaches.

(2) The feasibility and reliability of the algorithms should be improved, and more universal algorithms and models should be developed according to the water treatment and management requirements.

(3) Interdisciplinary talent with knowledge in different fields should be trained to develop more advanced machine learning techniques and apply them in engineering practices.

REQUIREMENT.TXT

Flask = 2.2.2

Joblib = 1.2.0

Numpy = 1.23.4

Pandas =1.5.1

Scikit-learn =1.1.3

Xgboost = 1.7.1

Gunicorn= 20.1.0

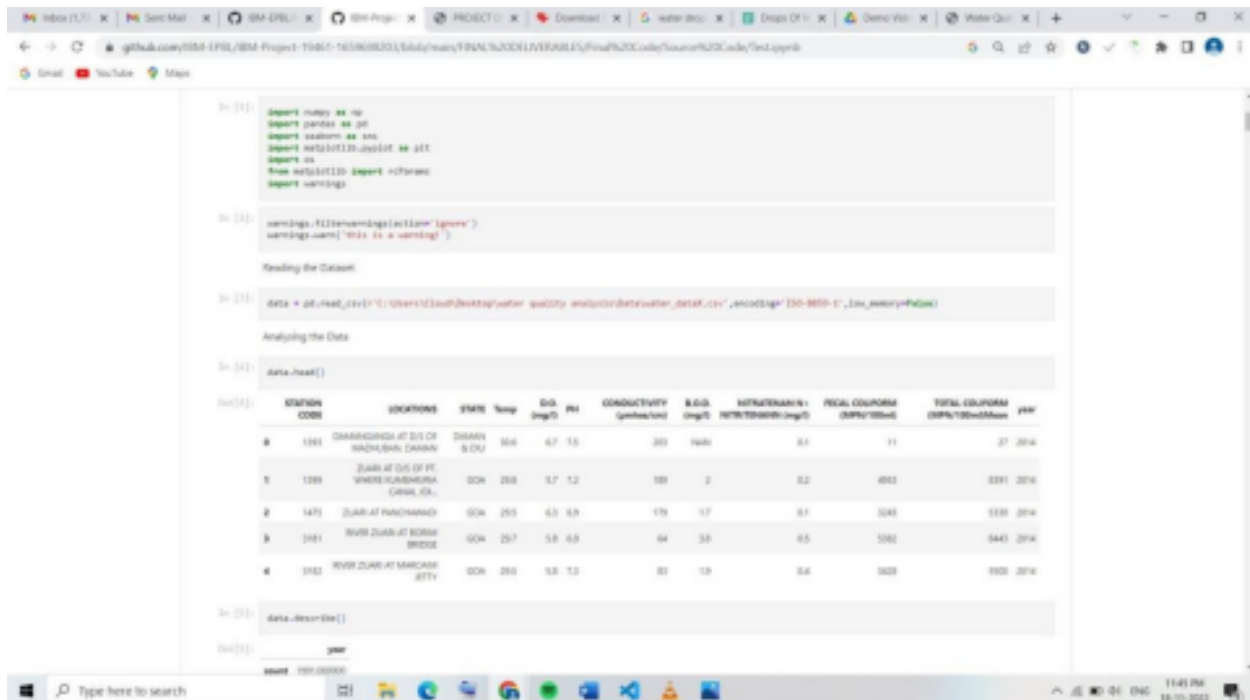
Matplotlib = 3.6.2

Seaborn = 0.12.1

APP.py:

```
1 import numpy as np
2 from flask import Flask, render_template, request
3 import pickle
4
5 app = Flask(__name__)
6 model = pickle.load(open('test.pkl', 'rb'))
7 @app.route('/', methods=['GET'])
8 def home():
9     return render_template("index.html")
10 @app.route('/login', methods=['POST'])
11 def login():
12     year = request.form['year']
13     do = request.form['do']
14     ph = request.form['ph']
15     co = request.form['co']
16     hot = request.form['hot']
17     sq = request.form['sq']
18     to = request.form['to']
19     total = [(int(year), float(do), float(ph), float(co), float(hot), float(sq), float(to))]
20     y_pred = model.predict(total)
21     y_pred = y_pred[0]
22     if(y_pred < 40 and y_pred < 400):
23         return render_template("index.html", placeholder = "Good!ent, The Predicted Value Is " + str(y_pred))
24     elif(y_pred < 40 and y_pred < 400):
25         return render_template("index.html", placeholder = "Very Good, The Predicted Value Is " + str(y_pred))
26     elif(y_pred < 40 and y_pred < 400):
27         return render_template("index.html", placeholder = "Good, The Predicted Value Is " + str(y_pred))
28     elif(y_pred < 40 and y_pred < 400):
29         return render_template("index.html", placeholder = "Fair, The Predicted Value Is " + str(y_pred))
30     elif(y_pred < 40 and y_pred < 400):
31         return render_template("index.html", placeholder = "Marginal, The Predicted Value Is " + str(y_pred))
32     else:
33         return render_template("index.html", placeholder = "Poor, The Predicted Value Is " + str(y_pred))
34
35 if __name__ == '__main__':
36     app.run(debug = True, port = 5000)
```


TEST.ipynb:



The screenshot shows a Jupyter Notebook with the following code cells:

```
In [1]: import numpy as np
import pandas as pd
import os
import warnings
import sys
from urllib.request import urlopen
import warnings

In [2]: warnings.filterwarnings('ignore')
warnings.warn('this is a warning')
```

Reading the Dataset

```
In [3]: data = pd.read_csv('C:\\Users\\User\\Desktop\\water_quality_analysis\\water_data.csv', encoding='ISO-8859-1', low_memory=False)
```

Analyzing the Data

```
In [4]: data.head()
```

```
In [5]:
```

	STATION CODE	STATIONS	STATE	Temp	D.O. (mg/l)	pH	CONDUCTIVITY (umhos/cm)	S.G.D. (mg/l)	NITRATE/NH4-N (mg/l)	TOTAL CHLORINE (MPN/100ml)	TOTAL CHLORINE (MPN/100ml)	PH
0	1391	CHANGINGHAT AT D/O OF NADH, BAN, DAKIN	BARAN	30.4	6.7	7.5	203	NaN	0.1	11	27	2016
1	1391	DUAR AT D/O OF PT. WARE, KUMHARA, CANAL, K.A.	ODK	28.0	5.7	7.2	109	2	0.2	493	2391	2016
2	1475	DUAR AT RANOHARDI	ODK	29.5	6.3	6.9	179	1.7	0.1	500	1139	2016
3	1481	RIVER DUAR AT BODAI BRIDGE	ODK	29.7	5.8	6.9	64	3.5	0.5	552	944	2016
4	1512	RIVER DUAR AT MARCAN, JTTV	ODK	28.0	5.5	7.0	83	1.9	0.4	303	953	2016

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

```
In [7]:
```

```
year
```

```
count    100.000000
```

The bottom of the notebook shows a search bar and system icons.

INDEX.html:

LINKS: GITHUB

:

[https://github.com/IBM-EPBL/IBM-Project-](https://github.com/IBM-EPBL/IBM-Project-19461-1659698203)

[19461-1659698203](https://github.com/IBM-EPBL/IBM-Project-19461-1659698203) DEMO VIDEO: