PROJECT

EFFICIENT WATER QUALITY ANALYSISAND PREDICTION USING MACHINE LEARNING

DONE BY

TEAM ID: PNT2022TMID53671

VIGNESH S
SYED FARDEEN ALTHAF
SREE MURARI
PRASSANNA G

GUIDED BY

Dr. MALATHI S.R (MENTOR)
Ms. LALITHA GAYATHRI (INDUSTRY MENTOR)

INDEX

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 diarrhea, 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 disease- inducing 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 machines.

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 QUALITYMODELING USINGAN ENSEMBLE MACHINELEARNING

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 WATERQUALITY INSMARTMARICULTURE 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 orLSTM-based prediction method.

TITLE: ASSESSMENT OF SURFACEWATER QUALITY BY USING SATELLITEIMAGES 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 behighly 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 GK2allows 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 POLLUTIONDETECTION 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 asthey 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 WATERQUALITYINDEXIN 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-fittingthe 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 COMBINEDNEURAL NETWORKS AND GENETICALGORITHMS FORTHE 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 searchprocess 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.

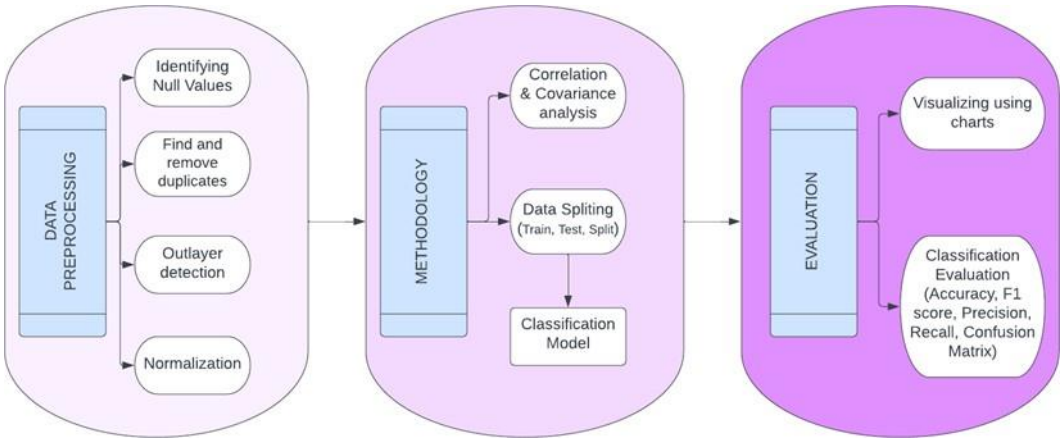


Figure 1

COMPARATIVE ANALYSISOF LITERATURE SURVEY:

S. No	Year	Researcher	Title	Parameters	Algorithm	Remarks
01	2021	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	2020	J. Liu, C. Yu, Z. Huetal	Accurate prediction schemeof water quality in smart mariculture with a deep Bi-S-SRUlearning 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	2019	E. Batur and D. Makita	Assessment of surface water quality by using satellite images fusion based on PCA method in the Lake Gala, Turkey	DO, SDD, TDS, and pH Chl-a and TSS	MLR, SVM, ANN,AND PCA	Highest accuracyof 92% using a PCA-based RSR model

04	2018	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 (DeepNN), and k Nearest Neighbors (kNN)	Highest accuracy of 93% with DeepNN

05	2017	Ahmadet 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 method
06	2016	Sakizadeh	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	2014	Ding et al	The Use of Combined Neural Networks and Genetic Algorithms for Prediction of River Water Quality	pH, NH3-N, TN, Cr6+, TP, COD Mn,BOD5, TCN,COD, Cd, Cu, Zn, Pb, Hg, As, Se, F-, sulfide, dissolved	Genetic Algorithm (GA), and Back Propagation Neural Network (BPNN)	The highest accuracy of Non polluted and polluted of88.9% and 93.1%wi thPCA technique

				oxygen, and LAS, etc.		

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 Methodin 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

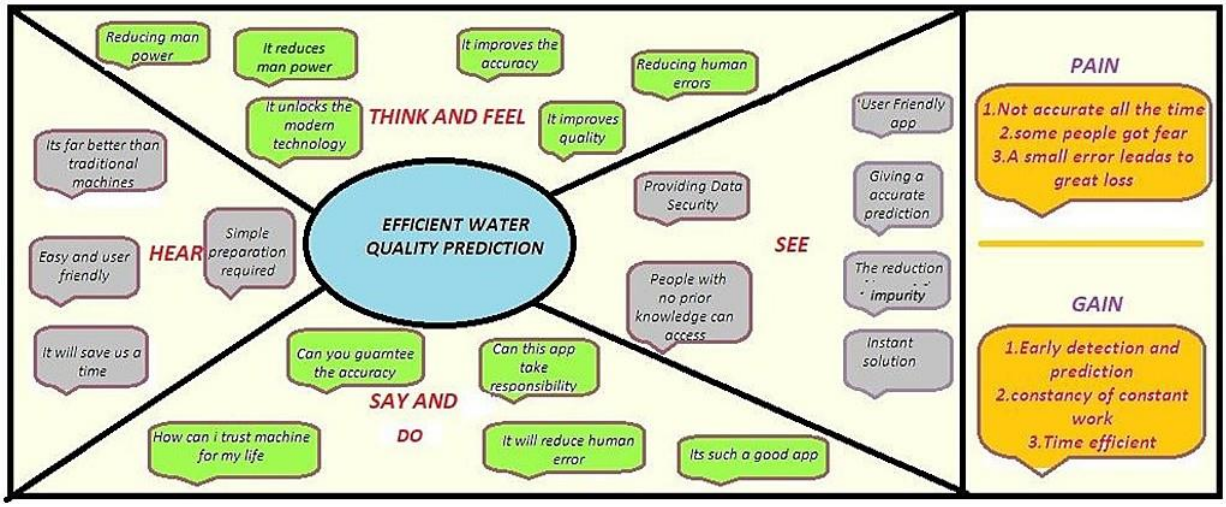
What is the aim of the project?	The proposed methodology achieves reasonable accuracy using a minimal number of parameters to validate the possibility ofits use in real-time water quality detection systems.
What are the boundaries of theproblem?	There is no boundary limit forthe issue because if anyone drinks unpurged or contaminated water, they willbe affected.
What is the issue?	The most important behavioral riskfactors of this disease can only be identified by taking samples of thecontaminated water and then researching that water by using datasets and then only we can find the issue.
Where is the issue comingfrom?	It majorly occurs to the people on the riverside who use the river water. If the waterhad any harmfulchemicals present, it would affect the peoplewith a disease.

Why is it important that we fix theproblem?	It is very crucial to develop an application that detects the diseasebecause rapid urbanization andindustrialization have led to a deterioration of water qualityat analarming 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 usingdata science.

Which solution can beused 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 tosolve 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

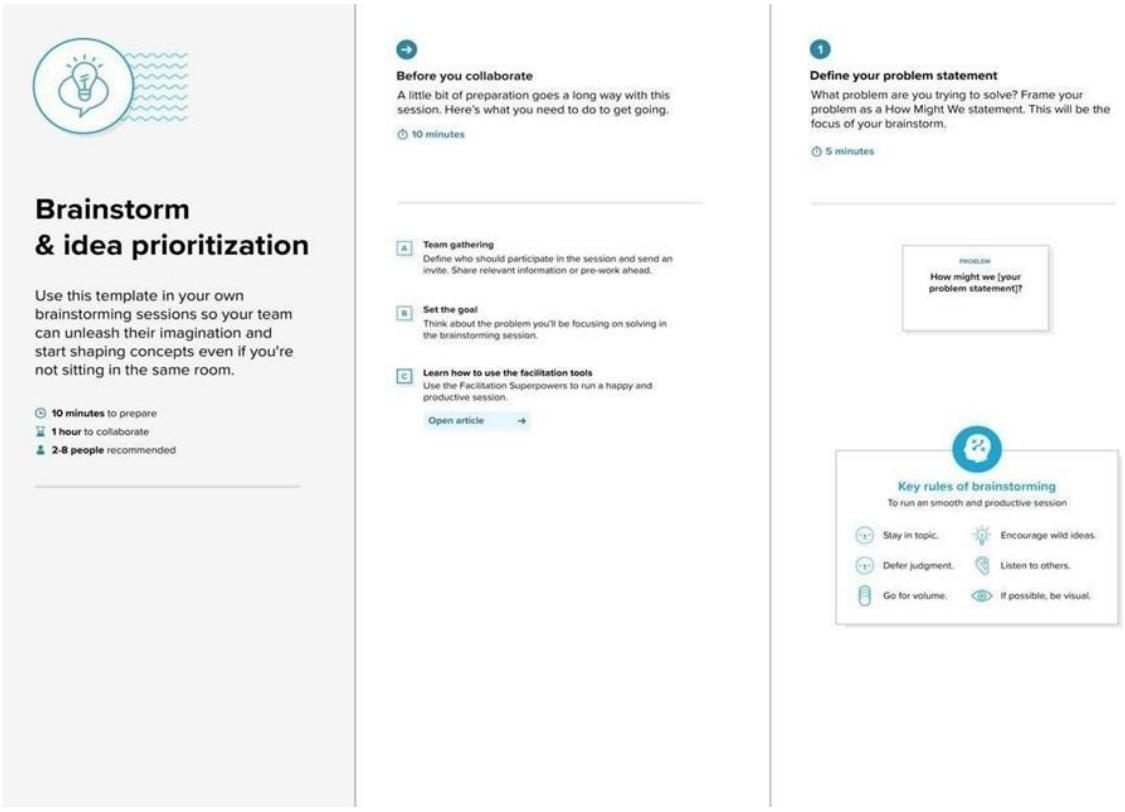
3.1 EMPATHY MAP CANVAS



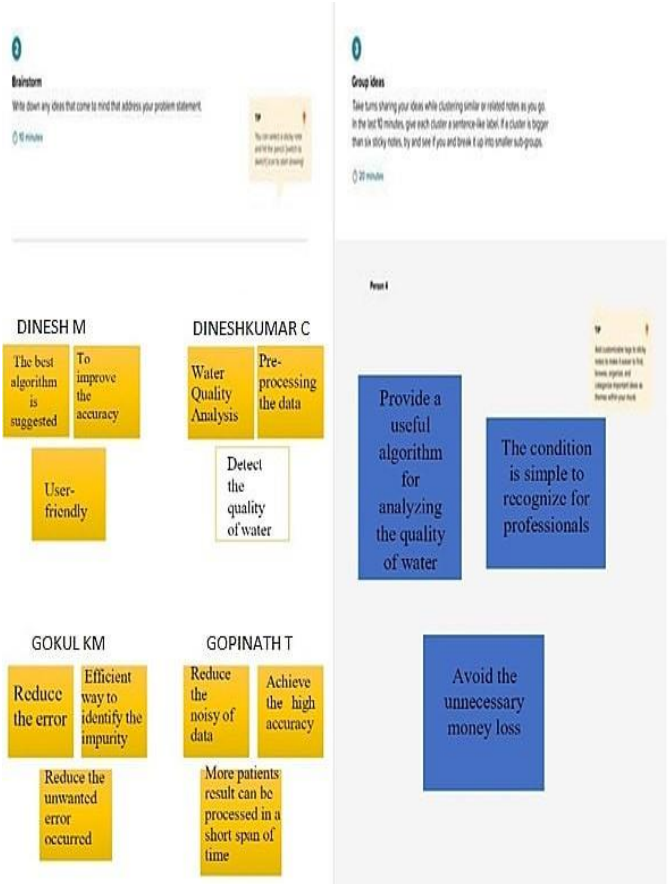
3.2 IDEATION & BRAINSTORMING

Brainstorm & Idea Prioritization Template:

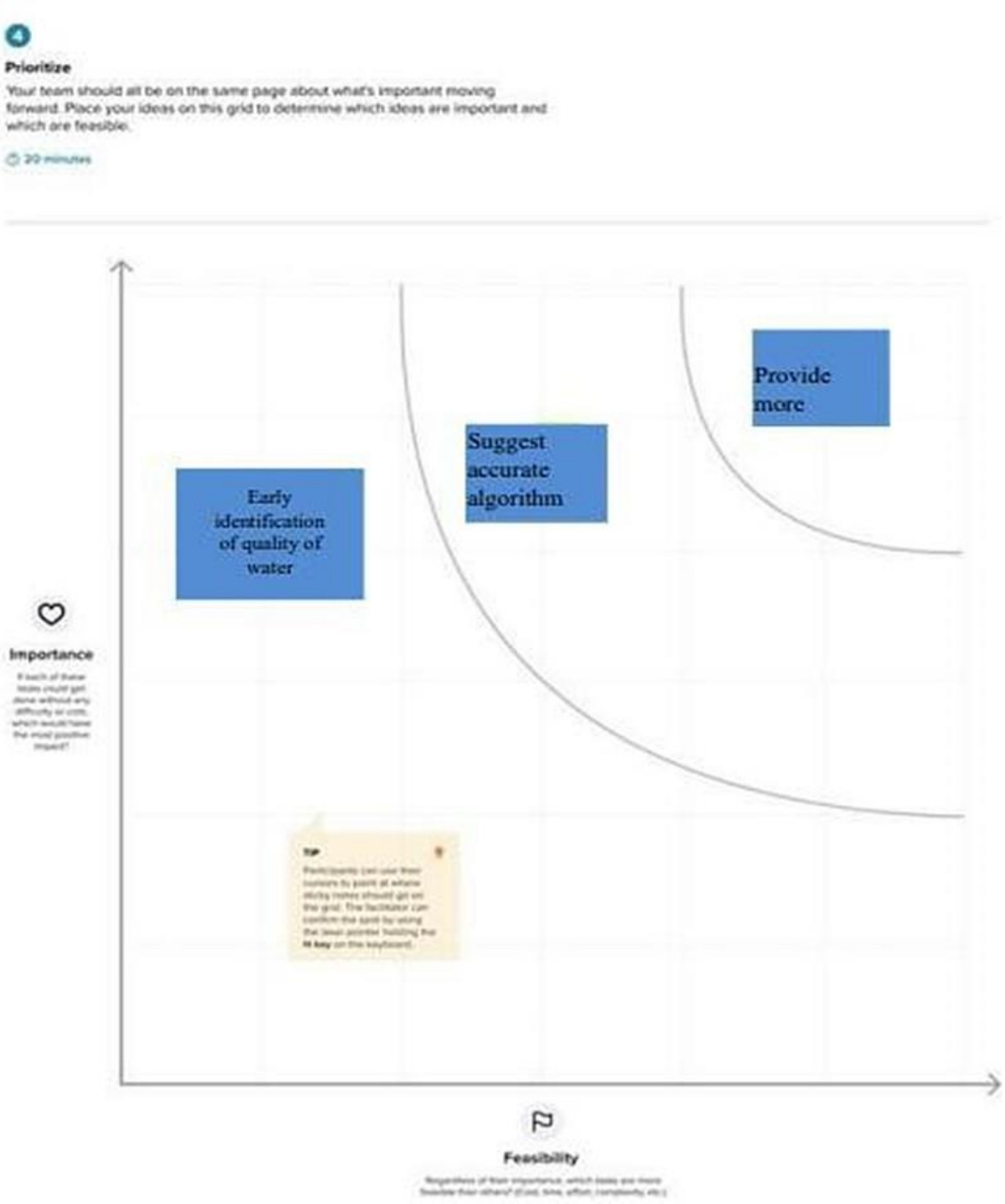
Step-1: Team Gathering, Collaboration and Select the Problem Statement



Step-2: Brainstorm, Idea Listing and Grouping



Step-3: Idea Prioritization



3.3 PROPOSED SOLUTION

S.No.	Parameter	Description
1.	Problem Statement (Problem to besolved)	1. People living in urban areasare concerned about the quality of water.Prevention and control of waterborne diseases are dependent onthe quality of water as an environmental determinant.The water on the planet makes up about 70% of its surface and is one of the most important sources forlifeWater qualityhas deteriorated at an alarmingrate because of rapid urbanization and industrialization. Water qualityhas been conventionally estimated throughexpensive and time- consuming lab andstatistical analyses, whichrender the contemporary notionof real-time monitoring moot. 2. It is, however, challenging to predict the quality of urban water since it varies nonlinearly in urban spaces and depends on a range of factors, including meteorology, water use patterns, and land uses. Hence, this project aims at developing a Machine Learning (ML) model that predicts water quality by

		taking into account allthe indicators of water quality.
2.	Idea / Solution description	1. A proposed model uses PH, DO, andother water quality standard indicators to predictwater quality. 2. On our dataset, we need to train thedatasets to run smoothly and see an incremental improvementin prediction rate using Random Forest Regression 3. Our plan is to build a web application that is

		integrated with the model. It provides a userinterface where the usercan enterpredictions.Onthe UI, predictionsare displayed based on the values entered into the saved model.

3.	Novelty / Uniqueness	• Using the model, it is possibleto determine whether the water is suitable for drinking.Therefore, it contributes tothemaintenance of health.

4.	Social Impact/ Customer Satisfaction	1. Water makes up about 70% of the earth's surface and is one of the most important sources vital to sustaining life. Rapid urbanizationand industrialization have led to a deterioration of water quality at analarming rate, resulting in harrowing diseases. 2. Most of the research either employed manual lab analysis, not estimating the water quality index standard,or used too many parametersto be efficient enough. 3. With machine learningtechniques, the implementation was done bythe WaterQuality Index(WQI). 4. 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)	1. A web application that is integrated tothe 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 predictionis showcased on the UIand deployit on IBM cloud. ● We can sell it for the prediction of water in various environments if the model preforms well ,also can maketheapp as premium one.
6.	Scalability of the Solution	1. The proposed can be implemented inrealtime water quality analysis by getting water sample using devices(Internet Of Things). 2. Real time apllications canbe used invarious placeslike schools,colleges etc. 3. Machine learing model integrated with IOT can make users more comfortable and to use in realtime.

3.4 PROBLEM SOLUTION FIT

CS	CC	AS
1.CUSTOMER SEGMENT(s) People, Residential, Commercial,Lab Testing	6.CUSTOMER CONSTRAINTS Water is essential for every one to sustain. If the water isimpure it may cause diseases withthis application it can be avoided.	5.AVAILABLE SOLUTION we need to trainthe datasetsto run smoothly and see an incremental improvement inthe prediction rate using Random Forest Regression algorithm on our dataset
J&P	**RC**	**BE**
2.JOB-TO-BE DONE/PROBLEMS Check the quality of water, whetherthe wateris drinkable, reason for unusability. Can verify the quality by themselves without expert	9.PROBLEMROOT CAUSE The major cause of this problemislack of drinking water and doesn't follow the proper diet anddoesn't have proper awareness is alsobeing a root cause.	7.BEHAVIOUR We will be building a web applicationthat is integrated to themodelbuilt. The enter values are given to the saved model and predictionis showcased on the UI

TR

3. TRIGGERS

Using this application, user can avoid the fear of water quality. Since the userknowsthe quality of water they are going to use.

EM

4.

EMOTIONS:BEFORE/AF

TERBefore: Thereare no

application to predictthe water quality.

After: By using thiseasy topredictthe quality of water usingsome a parameters.

SL

10. YOUR SOLUTION

The heart of the project dependsupon the prediction ofthe quality of thewater. As abundant as algorithms are present in order to achieve sucha goal, it is mandatory to select the best and the most efficient algorithm to finalize the predicted value.

C

H

8.CHANNEL OF BEHAVIOUR

Online: The application Notifythe user withdata preprocessing information

Offline: Water quality has been conventionally estimated through expensiveand 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.

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

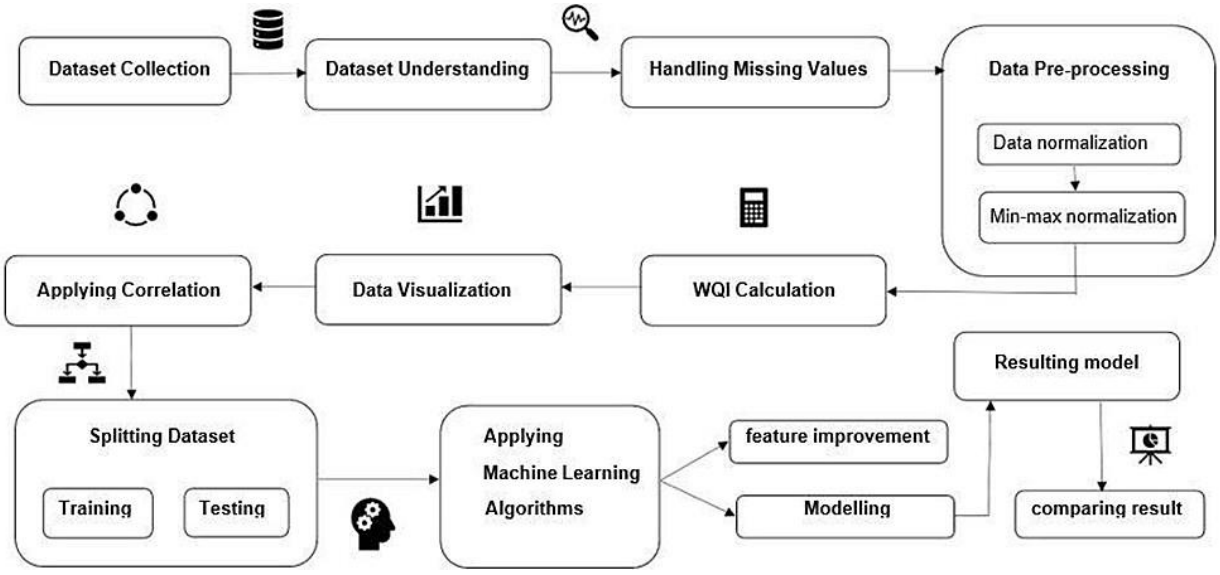
4.2 Non-functional Requirements:

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

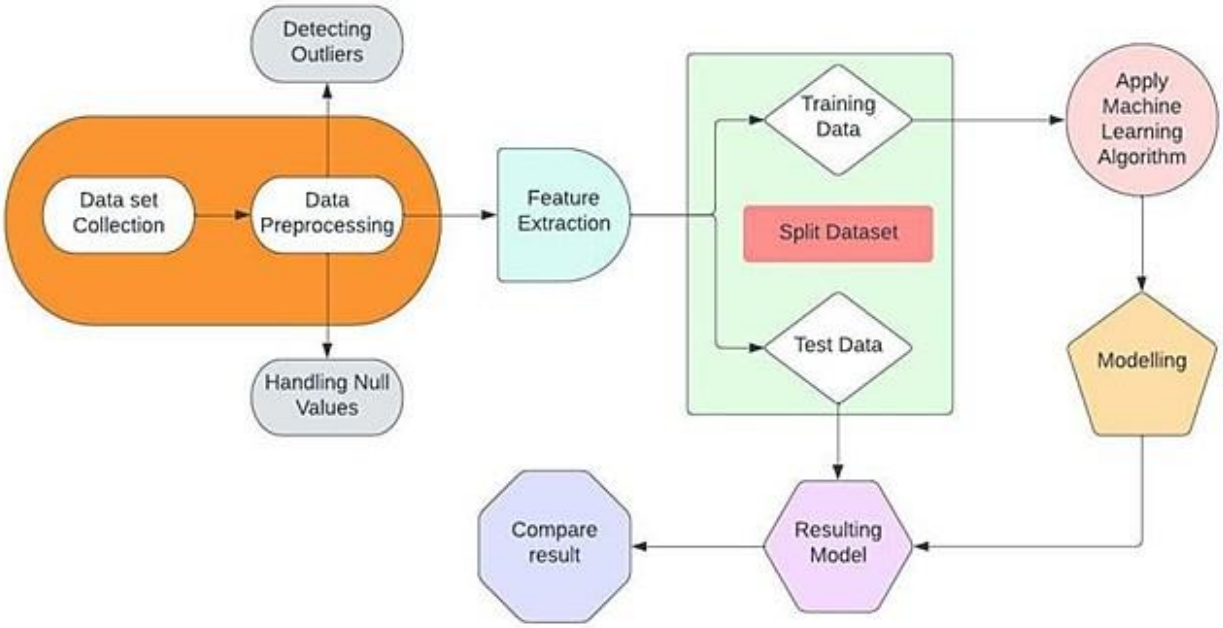
FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	Customers can access the system more efficiently and in a simpler way.The customerscan have the opportunity to view a better interpretation of results. The customers are alsorecommended the purification techniques based on the impurities.
NFR-2	**Security**	All the predicted information is accessed onlyby the authenticated users
NFR-3	**Reliability**	It should be reliable in producing effective andefficient water quality prediction results. It should ensure the trust and belief among people thatthis water quality prediction systemproduces correct results when used.
NFR-4	**Performance**	The system should be consistent in producingthe prediction results of the Water Quality Index (WQI) and also needs to ensure better throughput and response time compared to othersystems.
NFR-5	**Availability**	The system canbe utilized by the customers 24/7 and it shouldbe availed to test anykind ofwatersamples anywhere
NFR-6	**Scalability**	It can be used by a wide variety of users like testingagencies, private and public laboratories,restaurants and hotels, and peoplewho wish to test the quality of water they consume. The system should also be compatible enough to be integrated with futuretechnologies also.

5. PROJECT DESIGN

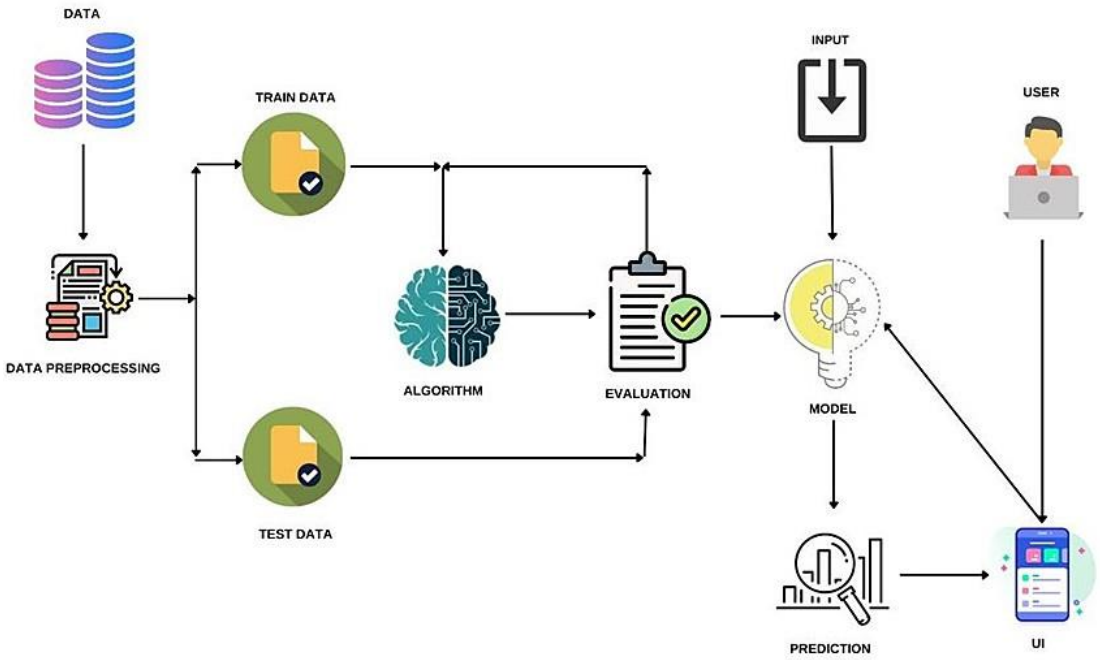
5.1 DATA FLOW DIAGRAMS



5.2 SOLUTION AND TECHNICALARCHITECTURE



Solution Architecture



Technical Architecture

Table-1: Components & Technologies:

S. No	Component	Description	Technology
1.	User Interface	User interacts by using webuserinterface.	HTML, CSS and Python Flask
2.	Application Logic-1(Login)	User can able to login if that personis already registered tothesite.	HTML, CSS and Python Flask
3.	Application Logic-2(Register)	User needsto be registered ifthatpersonis new to the site.	HTML, CSS and Python Flask.
4.	Application Logic-3(Reporting Form)	User needs to click on the reporting formin order to gettheprediction result	Front end-HTML,CSS and Python Flask. Back end – Query Languages, Python.
5.	Database	Data Type-String, Numeralvalues.	Query Languages such as MySQL, NoSQL etc.
6.	Cloud Database	Database Serviceon Cloud.	IBM DB2, IBM Cloud ant etc.
7.	File Storage	File storage requirements.	Local File-system.
8.	External API-1	Anyone can access the detailswithsome restrictions to the personal detailsof other users.	Web API.
9.	External API-2	Accessibility.	Aadhar API.
10.	Machine LearningModel	Predictthe result basedon thetraining and testing dataset.	Data RecognitionModel, etc.
11.	Infrastructure (Server / Cloud)	Application Deployment onLocalSystem.	Local System.

Table-2: Application Characteristics:

S. No	Characteristics	Description	Technology
1.	Open-Source Frameworks	Frameworks are used for predictivedata analysis, providing clear and actionableerror messages.	Tensor flow, Sci-kitlearn, Keras.
2.	Security Implementatio ns	OTP will be sent to the registered email id. Unauthorized users couldnot 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 isneeded.	Load Balancer.
5.	Performance	The modelcould be able to process large number of datasets.	Load Balancer.

5.3 USER STORIES

User Type	Functional Requirement(Epic)	User Story Number	User Story/ Task	Acceptance criteria	Priority	Release
People (web user)		USN-1	As a user, I can understand the detailed description of water quality on the homepage	I can accessthe web page	High	Sprint-1
	Input form	USN-2	As a user,I can enter the details required to analysis the water qualitywithuse of form provided in the web page.	I can giveinputs 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 waterresource organisation) toknow the details of water	I can contact people with Whatsapp, instagram, twitter, mail and also I canmake call	Medium	Sprint-3

6. PROJECT PLANNING AND SCHEDULING

6.1 SPRINT PLANNING AND ESTIMATION

Sprint	Functional Requirement(Epic)	User Story Number	User Story/ Task	Story Points	Priority	Team Members
Sprint-1	Data Preparation	USN-1	Collecting waterdataset andpre-processing it	10	High	DINESH M DINESHKUMAR C
Sprint-1	Model Building	USN-2	Create an ML modelto predictwater quality	5	Medium	DINESH M DINESHKUMAR C GOPINATH T GOKUL KM
Sprint-1	Model Evaluation	USN-3	Calculate the performance, errorrate, and complexityof the ML model and evaluate the dataset based on the parameter that the dataset consists of.	5	Medium	

Sprint-2	Model Deployment	USN-4	As a user, I needto deploy the model andneed to find the results.	20	Medium	
Sprint-3	Web page (Form)	USN-5	As a user, Ican use the application on byentering the water datasetto analyze or predict theresults.	20	Medium	DINESH M DINESHKUMAR C GOPINATH T GOKUL KM
Sprint-4	Dashboard	USN-6	As a user, I can predict the water quality by clicking the submit button and the application will show whether thewater isefficient for use or not.	20	High	DINESH M DINESHKUMAR C

Project Tracker:

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date	Story Points 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

Velocity:

Sprint 1: 1 userstories x 20 story points = 20

Sprint 2: 1 userstories x 20 story points = 20

Sprint 3: 1 userstories x 20 story points = 20

Sprint 4: 1 userstories x 20 story points = 20

Total = 80

The average sprintvelocity is 80 ÷ 4 = **20.**

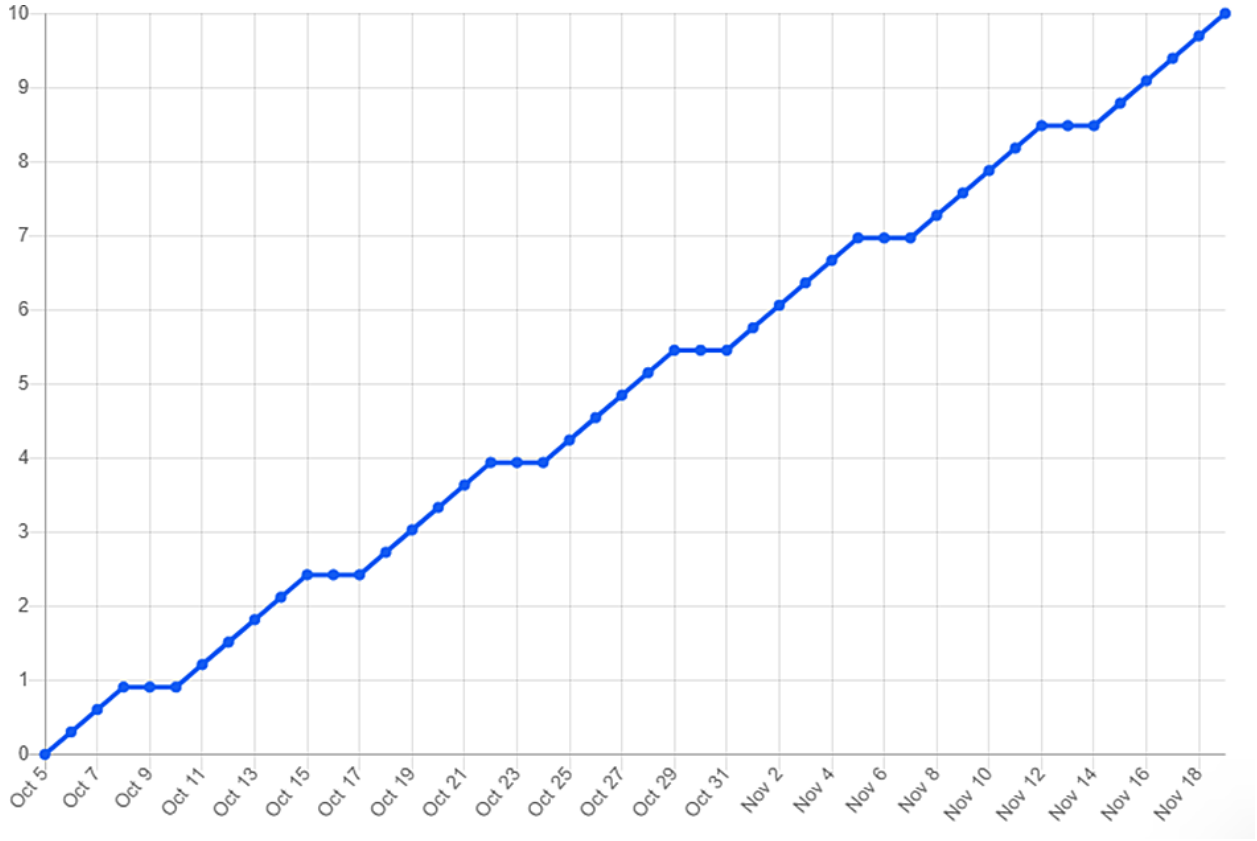
a. **PROJECT DELIVERY SCHEDULE**

TITLE	DESCRIPTION	DATE
Literature Survey& Information Gathering	Literature survey on the selected project & gatheringinformation by referring the technical papers, research publications, journals etc.	1 SEPTEMBER 2022

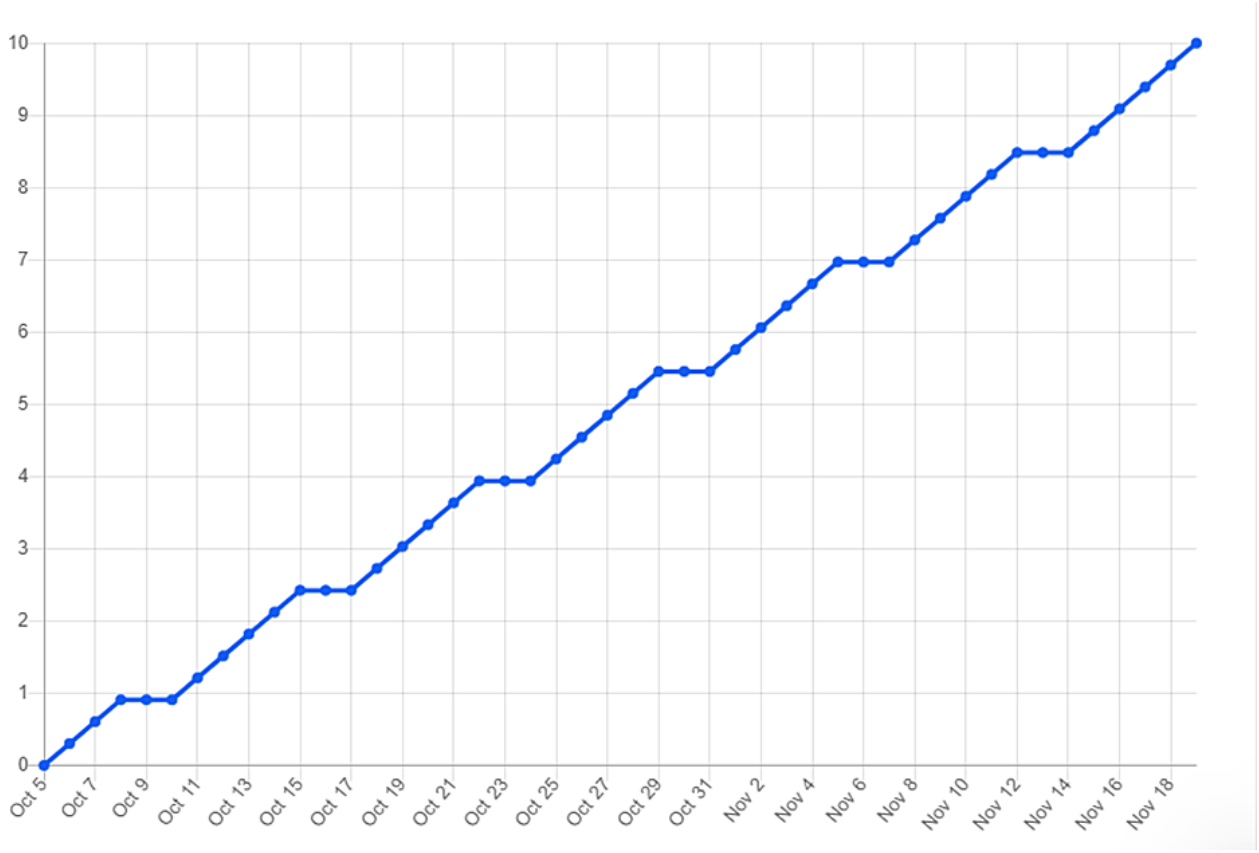
Prepare Empathy Map	Prepare Empathy Map Canvasto capture the user Pains and Gains,prepare list of problem Statements that are to be solved by this project.	7 SEPTEMBER 2022 & 9 SEPTEMBER 2022
Ideation	List the ideas by organizinga brainstorming session and prioritize the top threeideas based on the feasibility and importance.	15 SEPTEMBER 2022
Proposed Solution	Prepare the proposed solutiondocument, which includes novelty, feasibility ofidea, revenue model,social impact,scalability of solution,etc.	22 SEPTEMBER 2022
Problem Solution Fit	Prepare problem - solution fitdocument.	30 SEPTEMBER 2022
Solution Architecture	Prepare solution architecture document.	30 SEPTEMBER 2022
Customer Journey	Prepare the customer journeymaps to understandthe user interactions and experiences with the application (entry to exit).	6 OCTOBER2022
Functional Requirement	Preparethe functional requirementdocument.	11 OCTOBER2022
Data Flow Diagrams andUser_Stories	Prepare the Data flow diagrams andUser Stories for the problem	14 OCTOBER2022
Technology Stack Architecture	Prepare the Technology StackArchitechture	17 OCTOBER2022
Prepare Milestone &ActivityList	Prepare the milestones andactivity list of theproject.	21 OCTOBER 2022

Project Development Phase	Develop Project Development Phase which include Sprint 1, Sprint 2, Sprint 3, Sprint4	15 NOVEMBER 2022

6.3 REPORTS FROM JIRABURNDOWN CHART



BURNUP CHART



7. CODING AND SOLUTIONING

7.1 FEATURE 1 (RANDOMFOREST 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 arraysand matrices, along with a large collection of high- level mathematical functions to operate on these arrays. Moreover, NumPy forms the foundation of the Machine Learningstack. Pandas is an open-source Python package thatis most widely used for data science/data analysis and machine learning tasks. Sea born is a Pythondata visualization librarybased on matplotlib. It provides a high-level interface for drawing attractiveand 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 andhard 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 ofyour data, it is distribution, null values and much more. You can either explore data using graphs or through some python functions. There will betwo 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 measuresof central tendencyand measures of variability. Measures of central tendency include themean, 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 learningalgorithms can then decide in a betterway 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 designedto be highly efficient, flexible,and portable. It implements machinelearning algorithms under the GradientBoosting 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 classesused to connect with the system softwareand handle inputsand outputs. It simplifies the life of a developer while giving them theability to use certain extensions and makes the online applications scalable and maintainable. Flask is a web application frameworkwritten in Python. A Web Application Framework or a simply a Web Framework represents a collection of libraries and modules that enable web application developers to write applications without worrying about low-level detailssuch 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 Poocco. Flask is based on the Werkzeg WSGI toolkit and the Jinja2 template engine. Both are Pocco projects. The Web Server Gateway Interface (Web Server GatewayInterface, WSGI) has been used as a standard for Python web application development. WSGI is the specification of a common interface between web serversand web applications. Flask is often referredto as a micro-framework. It is designed tokeepthecoreoftheapplication simple and scalable. Insteadof 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'thave ahuge learningcurve. HTML stands for Hyper Text Markup Language. HTMLis the standard markup language for creating Web pages. HTML describes the structure of a Web page. HTML consists of a series of elements. HTMLelements tell the browserhow 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 deployedusing Flask is

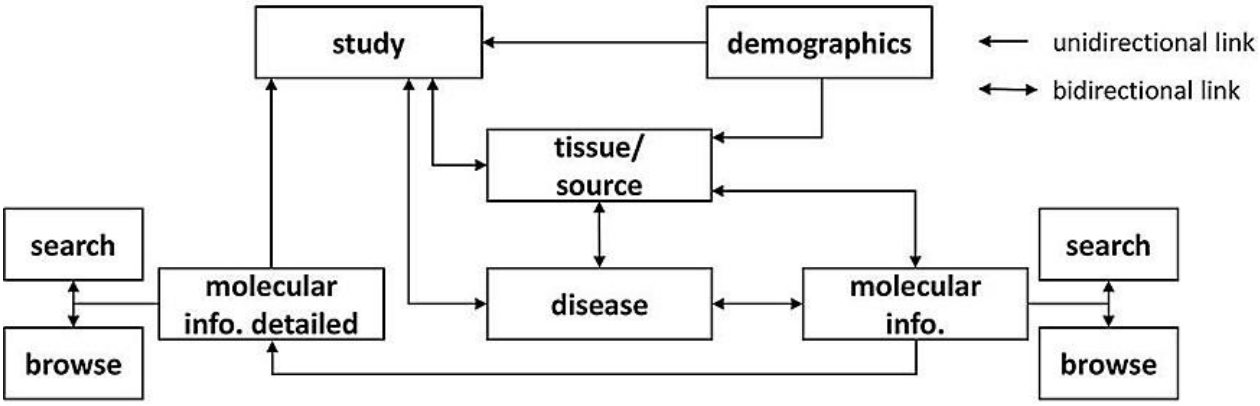
used to predict the Chronic Kidney Disease. Hypertext markup language (HTML) is the basic language used to create documents for the Web and, along with HTTP (hypertext transfer protocol) and URLs (universal

resource locators), is one of the threemain 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 provideadditional information about it, including its content, structureand appearance. This information facilitates automated operations on the text,including formatting it for display, searching it and even modifying it. Some type of markup languageis employed by every word processing program and by nearlyevery other program that displays text, although such languages and their tags are typically hidden from the user.HTML consistsof 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 includeparagraphing, 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 spacesfor 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 rapidlygenerated 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,providinginformation about genes, transcripts, proteins or metabolites. In this review, we presentomics databases availablecurrently, 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 broadinformation scope and kidney- specificdatabases distinctively concentrated on kidney pathologies. In research, databasescan be used as a rich source ofinformation 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 providea 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 miningshould provide powerful insights into unlocking the mysteries of kidney disease, leading to a potential impact on pharmacological intervention andtherapeutic disease management.

8. TESTING

8.1 TEST CASES

Test CaseID		**Test Case Description**	Test the Water qualityPrediction Functionality		
Created By	DINESH M	**Reviewed By**	DINESHKUMAR C		
Tester'sName	GOKUL KM GOPINATH T	**Date Tested**	November 15, 2022	**Test Case (Pass/ Fail /Not Executed)**	Pass

S #	Prerequisites:
1	Access to Chrome Browser
2	Entering the details required
3	checkfor correct values
4	Application to train themodel

S #	Test Data
1	By Clicking the websitelink
2	Detailsshould be in ainteger format
3	Data should be filled
4	Providethe datasets formodel training

Test Scenario Verify whether the deployed project predictsas 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

TESTINGTEST CASE 1:

TEST CASE 2:

9. RESULTS

9.1 PERFORMANCE METRICES

TITLE	DESCRIPTION	DATE
Literature Survey& Information Gathering	Literature survey on the selected project & gatheringinformation by referring the technical papers, research publications, journals etc.	1 SEPTEMBER 2022
Prepare Empathy Map	Prepare Empathy Map Canvastocapture 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 organizinga brainstorming session and prioritize the top three ideas based on thefeasibility and importance.	15 SEPTEMBER 2022

Proposed Solution	Prepare the proposed solutiondocument, which includes novelty, feasibility ofidea, revenue model,social impact,scalability of solution,etc.	22 SEPTEMBER 2022
Problem Solution Fit	Prepare problem - solution fitdocument.	30 SEPTEMBER 2022
Solution Architecture	Prepare solution architecturedocument.	30 SEPTEMBER 2022
Customer Journey	Prepare the customer journeymaps to understandthe user interactions and experiences with the application (entry to exit).	6 OCTOBER2022
Functional Requirement	Preparethe functional requirementdocument.	11 OCTOBER2022
Data Flow Diagrams andUser_Stories	Prepare the Data flow diagrams and User Stories for the problem	14 OCTOBER2022
Technology Stack Architecture	Prepare the Technology StackArchitechture	17 OCTOBER2022
Prepare Milestone &ActivityList	Prepare the milestones andactivity list of theproject.	21 OCTOBER 2022
Project Development Phase	Develop Project Development Phase which include Sprint 1, Sprint 2, Sprint 3, Sprint4	15 NOVEMBER 2022

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 theseasons, 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 fromother nodes, and therefore cannot interact with the Exchange Network Technical expertise and network server required.

11. CONCLUSION

Water is one of the most essential resourcesfor survival and its qualityis 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 availablewater 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 representativesupervised 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 Tireh River (Iran). To this end most dataset related well-known components, such as pH, SO_4, Na, Ca, Cl, Mg, HCO_3 etc., were

collected. Results indicated that the applied models have suitable performance for predicting water quality.

2. FUTURE SCOPE

In future works, we propose integrating the findings of this research in a large- scale IoT-based online monitoring system using only the sensorsof 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 identifypoor quality waterbefore it is released for consumption and alert concernedauthorities. 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 proposedmodels. So far, the water quality datasetfrom the LVW collected by Southern Nevada Water Authorityand 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 datasetsto build a more reliable water qualityprediction model.

3. 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 resourceallocation, 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 learningis usually dependenton large amountsof high- quality data. Obtaining sufficient data with high accuracy in watertreatment 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 hindersthe wide application of machine learningapproaches.

3. The implementation of machine learning algorithms in practical applications requires researchersto have certainprofessional background knowledge.

To overcome the above-mentioned challenges, the following aspectsshould be consideredin 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

learningapproaches.

2. The feasibility and reliability of the algorithms should be improved, and more universalalgorithms and models should be developed accordingto the water treatment and management requirements.

3. Interdisciplinary talentwith knowledge in different fieldsshould be trainedto develop more advanced machinelearning 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

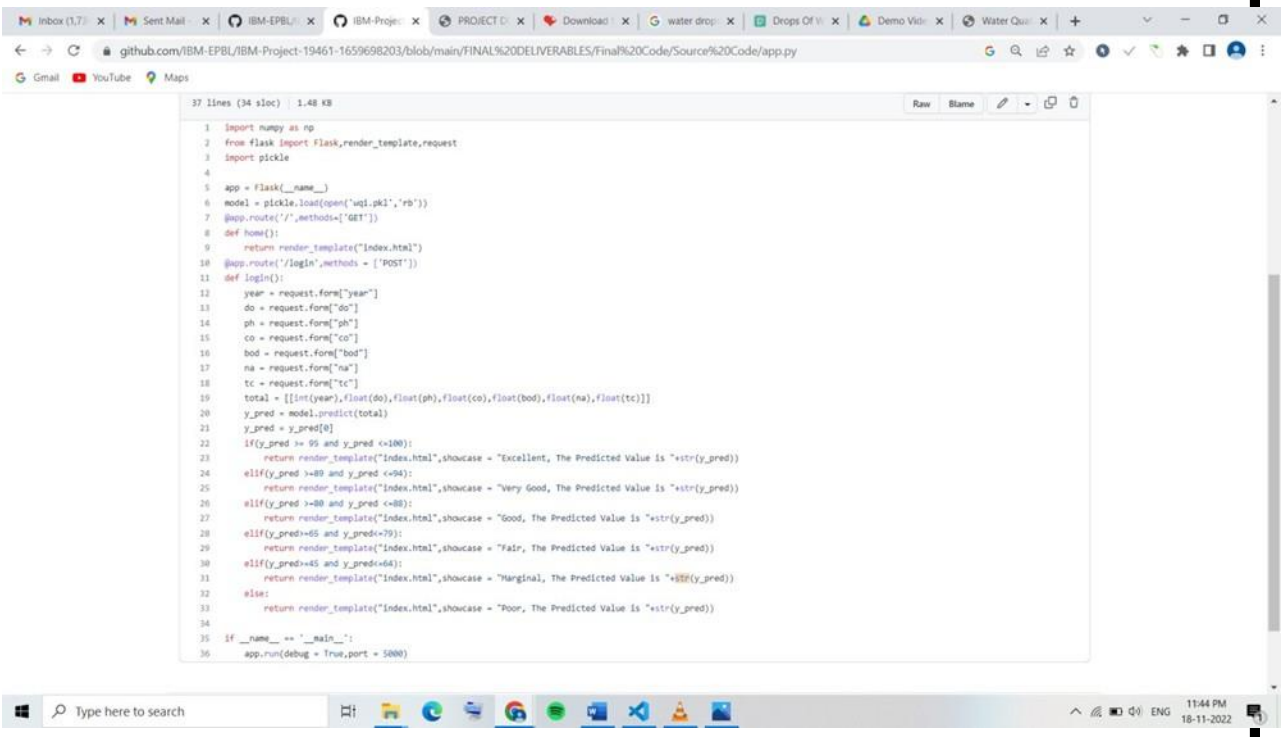
Xgboost = 1.7.1

Gunicorn= 20.1.0

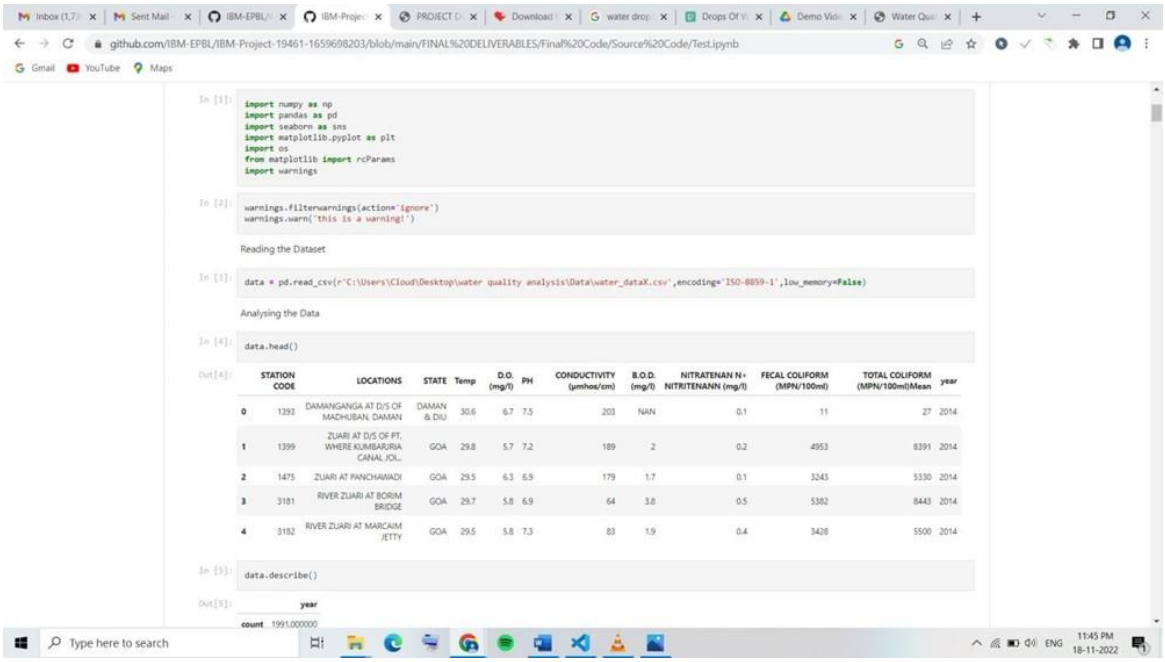
Matplotlib = 3.6.2

Seaborn = 0.12.1

APP.py:



TEST.ipynb:



```
In [1]:
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import os
from matplotlib import rcParams
import warnings

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

Reading the Dataset

```
In [3]:
data = pd.read_csv(r'C:\Users\Cloud\Desktop\water quality analysis\Data\water_dataX.csv',encoding='ISO-8859-1',low_memory=False)
```

Analysing the Data

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

Out[4]:

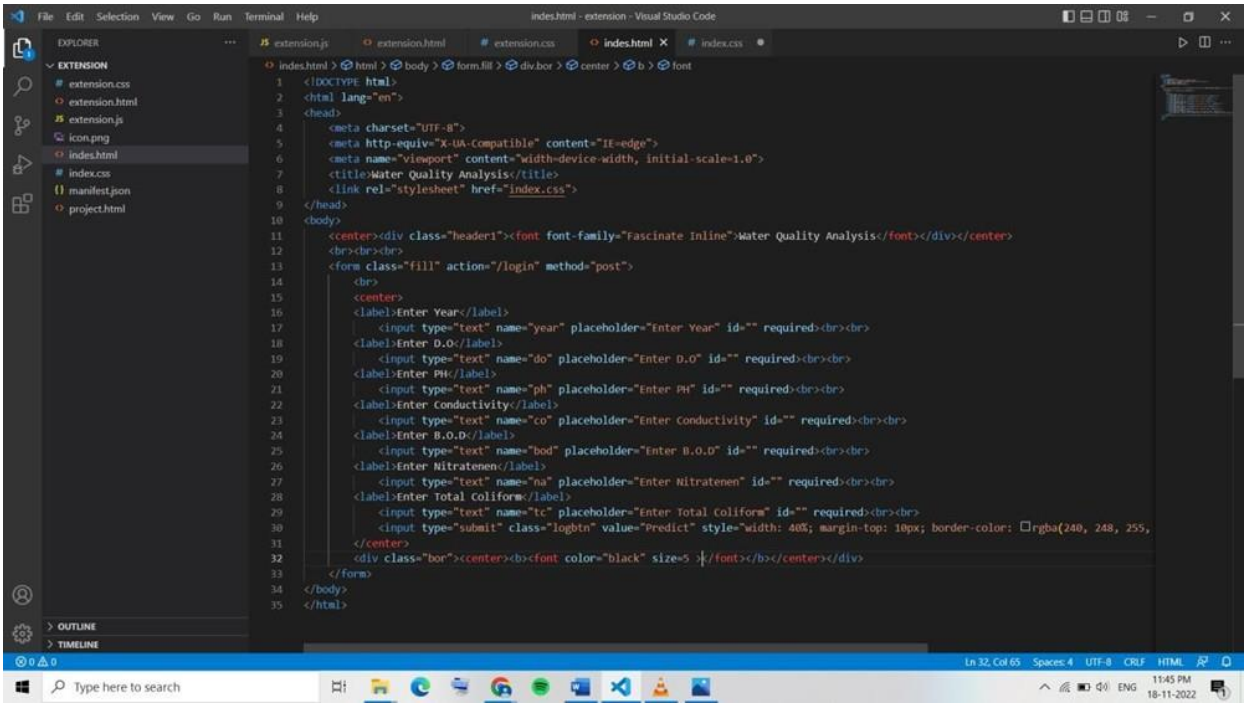
	STATION CODE	LOCATIONS	STATE	Temp	D.O. (mg/l)	PH	CONDUCTIVITY (µmhos/cm)	B.O.D. (mg/l)	NITRATENAN N+ NITRITENANN (mg/l)	FECAL COLIFORM (MPN/100ml)	TOTAL COLIFORM (MPN/100ml)Mean	year
0	1393	DAMANGANGA AT D/S OF MADHUBAN, DAMAN	DAMAN & DIU	30.6	6.7	7.5	203	NAN	0.1	11	27	2014
1	1399	ZUARI AT D/S OF PT. WHERE KUMBARJRIA CANAL JOI...	GOA	29.8	5.7	7.2	189	2	0.2	4953	8391	2014
2	1475	ZUARI AT PANCHAWADI	GOA	29.5	6.3	6.9	179	1.7	0.1	3243	5330	2014
3	3181	RIVER ZUARI AT BORIM BRIDGE	GOA	29.7	5.8	6.9	64	3.8	0.5	5382	8443	2014
4	3182	RIVER ZUARI AT MARCAIM JETTY	GOA	29.5	5.8	7.3	83	1.9	0.4	3428	5500	2014

```
In [5]:
data.describe()
```

Out[5]:

	year
count	1991.000000

INDEX.html:



```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <title>Water Quality Analysis</title>
    <link rel="stylesheet" href="index.css">
</head>
<body>
    <center><div class="header1"><font font-family="Fascinate Inline">Water Quality Analysis</font></div></center>
    <br><br><br>
    <form class="fill" action="/login" method="post">
        <br>
        <center>
        <label>Enter Year</label>
            <input type="text" name="year" placeholder="Enter Year" id="" required><br><br>
        <label>Enter D.O</label>
            <input type="text" name="do" placeholder="Enter D.O" id="" required><br><br>
        <label>Enter PH</label>
            <input type="text" name="ph" placeholder="Enter PH" id="" required><br><br>
        <label>Enter Conductivity</label>
            <input type="text" name="co" placeholder="Enter Conductivity" id="" required><br><br>
        <label>Enter B.O.D</label>
            <input type="text" name="bod" placeholder="Enter B.O.D" id="" required><br><br>
        <label>Enter Nitratenen</label>
            <input type="text" name="na" placeholder="Enter Nitratenen" id="" required><br><br>
        <label>Enter Total Coliform</label>
            <input type="text" name="tc" placeholder="Enter Total Coliform" id="" required><br><br>
            <input type="submit" class="logbtn" value="Predict" style="width: 40%; margin-top: 10px; border-color: rgba(240, 248, 255,
        </center>
        <div class="bor"><center><b><font color="black" size=5 ></font></b></center></div>
    </form>
</body>
</html>
```

LINKS:

GITHUB :

IBM-EPBL/IBM-Project-18686-1659688445:
Efficient Water Quality Analysis & Prediction using Machine Learning (github.com)

IBM-Project-18686-1659688445