EFFICIENT WATER QUALITY ANALYSIS AND PREDICTION USING MACHINE LEARNING

IBM-PROJECT-PNT2022TMID25561

NALAIYATHIRAN PROJECT BASED LEARNING ON PROFESSIONAL READINESS FOR INNOVATION , EMPLOYING AND ENTERPRENEURSHIP

PORJECT REPORT BY:

N GAYATHRI-210919104012

SNR SHAMYUKTHA-210919104040

M MOHANESWARI-210919104025

A S MELITTA-210919104701

BACHELOR OF ENGINEERING

COMPUTER SCIENCE ENGINEERING

LOYOLA INSTITUTE OF TECHNOLOGY

CHENNAI-600123

ANNA UNIVERSITY::CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this project "Efficient water quality analysis and predication using machine learning " is the bonafide work of " Gayathri N, Mohaneswari M, Shamyuktha SNR, Melitta A S,who carried out the project work under my supervision.

SIGNATURE
Dr.G.Bhuvaneswari,
HEAD OF THE DEPARTMENT,
Computer Science Engineering,
Loyola institute of technology
Palanchur,
Chennai-600123

SIGNATURE
Mr.Noor Mohammed,
ASSISTANT PROFESSOR,
Computer Science Engineering,
Loyola institute of technology,
Palanchur,
Chennai-600 123.

CONTENTS

7. CODING & SOLUTIONING

a. Feature1

b. Feature2

c. Database Schema(ifApplicable)

8. TESTING

a. Test Cases

b. User Acceptance Testing

9. RESULTS

a. Performance Metrics

10. ADVANTAGES&DISADVANTAGES

11. CONCLUSION

12. FUTURESCOPE

13. APPENDIX

SourceCode

GitHub&ProjectDemoLink

Efficient Water Quality Analysis & Prediction Using Machine Learning

1. INTRODUCTION

Water quality has a direct impact on public health and the environment. Water is used for various practices,such as drinking ,agriculture, and industry.Recently, development of water sports and entertainment has greatly helped to attract tourists (Jennings 2007). Among various sources of water supply, due to easy access, rivers have been used more frequently for the development of human societies. Using other water resources such as groundwater and seawater sometimes assisted with problems. For example, using ground water without suitable recharge will lead to land subsidence

1.1 Project Overview

With the rapid increase in the volume of data on the aquatic environment, machine learning has become an important tool for data analysis, classification, and prediction. Unlike traditional models used in water-related research, data-driven models based on machine learning can efficiently solve more complex nonlinear problems. In water environment research, models and conclusions derived from machine learning have been applied to the construction, monitoring, simulation, evaluation, and optimization of various water treatment and management systems. Additionally, machine learning can provide solutions for water pollution control,water quality improvement, and watershed ecosystem security management. In this review, we describe the cases in which machine learning algorithms have been applied to evaluate the water quality in different water environments ,such as surface water, groundwater, drinking water, sewage, and seawater. Furthermore, we propose possible future applications of machine learning approaches to water environments

1.2 purpose

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 water borne 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. The purpose of this project is to Predict Water Quality by considering all water quality standard indicators.

2. LITERATURESURVEY

Many work shad been conducted to predict water quality using Machine Learning (ML)approaches .Some researchers used the traditional Machine Learning models, such as DecisionTree, Artificial Neural Network , Support Vector Machine, K-Nearest Neighbors and Naïve Bayes . However, in recent years, some researchers are moving towards more advanced ML ensemble models, such as Gradient Boosting and Random Forest Traditional Machine Learning models, such as the Decision Tree model, are frequently found in the literature and performed well on water quality data. However, decision-tree-based ensemble models, including Random Forest (RF) and Gradient Boosting (GB), always outperform the single decisiontree.Among the reasons for this are its ability to manage both regular attributes and data,not being sensitive to missing values and being highly efficient. Compared to other ML models, decision-tree-based models are more favorable to short-term prediction and may have a quicker calculation speed [6] .Gakii and Jepkoech compared five different decision tree classifiers, which are Logistic Model Tree (LMT),Hoeffding tree, Random Forest and Decision Stump. They found that J48 showed the highest accuracy of94%, while Decision Stump showed the lowest accuracy. Another study by Jeihouni et al. also compared fivedecision-tree-basedmodels,which are RandomTree, RandomForest, Ordinary DecisionTree(ODT) ,Chisquare Automatic Interaction Detector and Iterative Dichotomiser (ID3), to determine high water quality zones. They found that ODT and Random Forest produce higher accuracy compared to the other algorithms and the methods are more suitable for continuous dataset.

Another popular Machine Learning model to predict water quality is Artificial Neural Network(ANN). ANN is a remarkable data-driven model that can cater both linear and non-linear associations amongou tput and input data. It is used to treat the non-linearity of water quality data and the uncertainty of contaminant source. However ,the performance of ANN can be obstructed if the training data are imbalanced and when all initial weights of the parameter have the same value. In India, Aradhana and Singhused ANN algorithms to predict water quality. They found that Lavenberg Marquardt (LM) algorithm has a better performance than the Gradient Descent Adaptive(GDA) algorithm. Abyaneh[5] used ANN and multivariate linear regression models in his research and found that the ANN model outperforms the MLR model. However, the research only assessed the performance of the ANN model using root-mean-squareerror (RMSE), coefficient of correlation (r) and bias values. Although ANN models are the most broadly used,they have a drawback as the prediction power becomes weak if they are used with a small dataset and the testing data are outside the range of the training data.

The ensemble method is a Machine Learning technique that combines several base learners' decisions to produce a more precise prediction than what can be achieved with having each base learner's decision. This method has also gained wide attention among researchers recently. The diversity and accuracy of each base learner are two important features to makethe ensemble learners work properly. The en semble method ensures the two features in several ways based on its working principle. There are two commonly used ensemble families in Machine Learning, which are bagging and boosting. Both the bagging and boosting methods provide a higher stability to the classifiers and are good in reducing variance. Boosting can reduce the bias, while bagging can solve the overfitting problem.
.A famous ensemble model that uses the bagging algorithm is Random Forest.It is a classificationmodelthat uses multiple base models, typically decision trees, on a given subset of data independently and makes decisions based on all models.It uses feature and omness and bagging when building each individual ldecisiontree to produce an independent forest of trees.

Existing problem

the main problem lies here. For testing the water quality we have to conduct lab tests on the water which is costly and time-consuming as well. So, in this paper, we propose an alternative approach using artificial intelligence to predict water quality. This method uses a significant and easily available water quality index which is set by the WHO(World Health Organisation). The data taken in this paper is taken from the PCPB India which includes 3277 examples of the distinct wellspring. In this paper,WQI(WaterQualityIndex)iscalculatedusingAItechniques.Soinfuturework,wecanintegratethiswith IoT based framework to study large datasets and to expand our study to a larger scale. By usingthat it can predict the water quality fast and more accurately than any other IoT framework. That IoTframework system uses some limits for the sensor to check the parameters like ph, Temperature, Turbidity, and so on. And further after reading this parameter pass these readings to the Arduino microcontroller and ZigBee handset for further prediction

2.2 References

Srivastava,G.;Kumar,P.Waterqualityindexwithmissingparameters.Int.J.Res.Eng.Technol.2013,2,609–614.

PCRWR.WaterQualityofFiltrationPlants,MonitoringReport;PCRWR:Islamabad,Pakistan,2010.Availableonline:http://www.pcrwr.gov.pk/Publications/Water%20Quality%20Reports/FILTRTAION%20PLANTS%20REPOT-CDA.pdf (accessedon23August2019).

Sakizadeh,M.Artificialintelligenceforthepredictionofwaterqualityindexingroundwatersystems.Model.EarthSyst.Environ.2016,2,8. [CrossRef]

2.3Problem Statement Definition

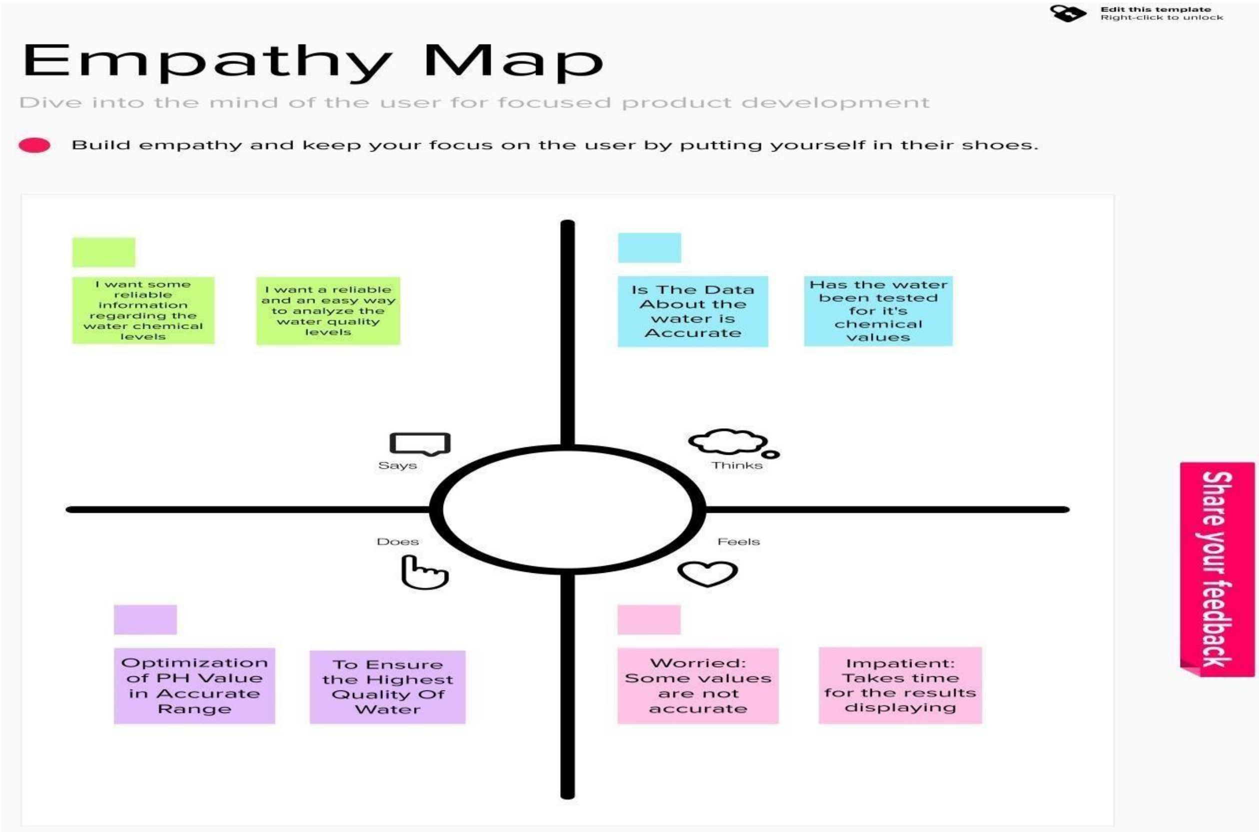
Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs out weigh the costs of under taking the interventions.

3. IDEATION & PROPOSEDSOLUTION

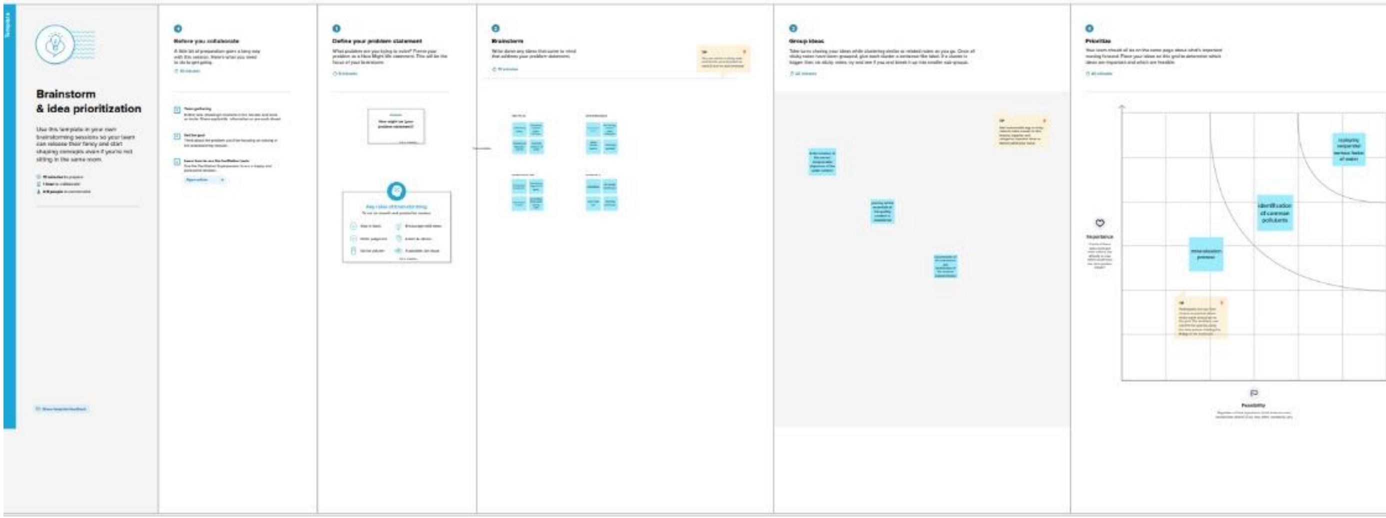
3.1 Empathy Map Canvas

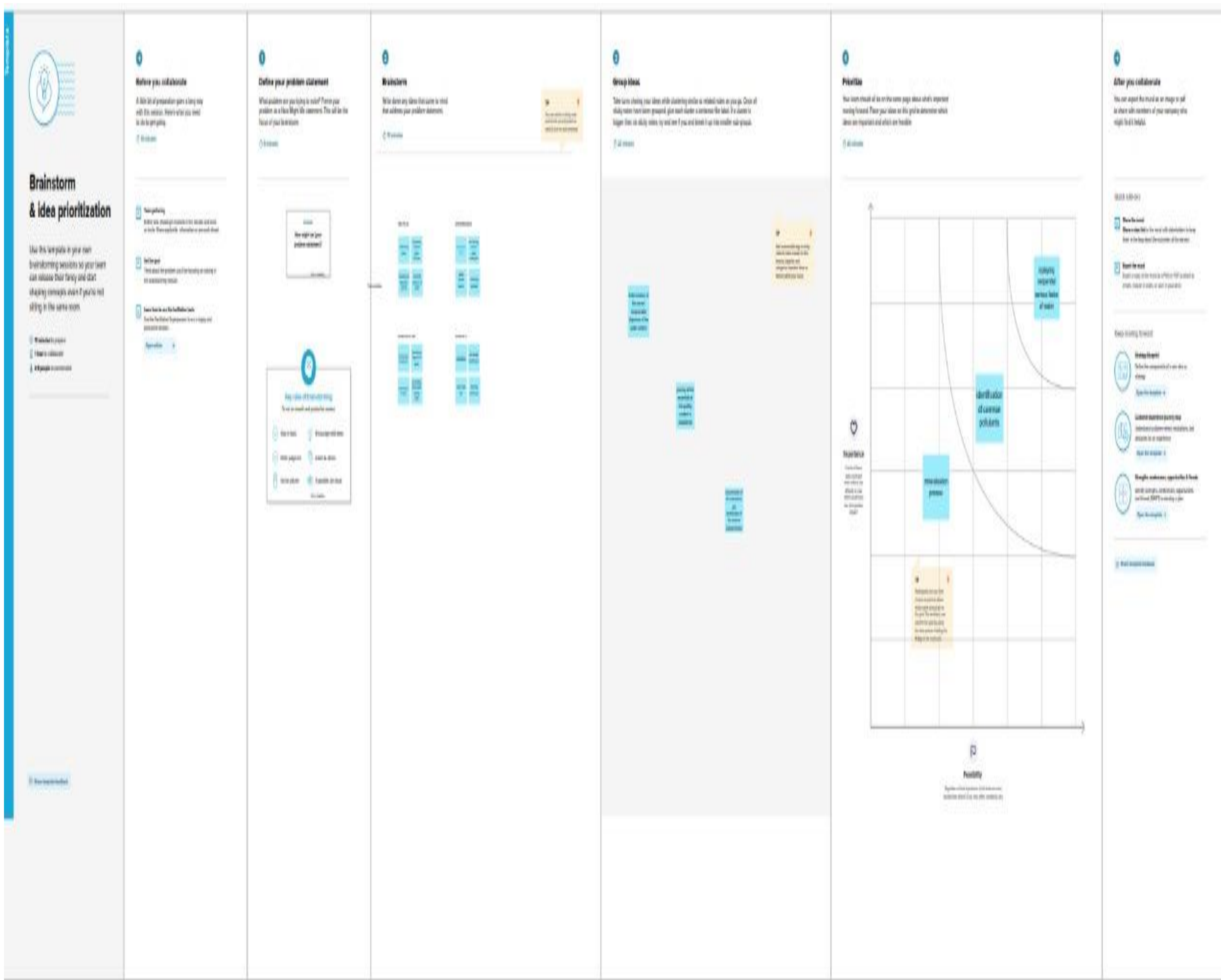
An empathy map canvas serves as a foundation for outstanding user experiences, which focus on providing the experience customers want rather than forcing design teams to rely on guesswork.

Empathy map canvases help identify exactly what it is that users are looking for so brands can deliver. They can be particularly beneficial for getting teams on the same page about who users are and what they want from the brand.



3.2 Ideation & Brainstorming





3.3 Proposed Solution

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 explore saseries of supervised machinelearning 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 ,pHand total dissolved solids.

4. REQUIREMENT ANALYSIS

4.1 Functional requirement

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Registration through Form Registration through Gmail Registration through LinkedIN
FR-2	User Confirmation	Confirmation via Email Confirmation via OTP
FR-3	Executive administration	Regulation of monitoring the water environment status and regulatory compliance like pollution event emergency management, and it includes two different functions: early warning/forecast monitoring.
FR-4	Data handling	File contains water quality metrics for different water bodies.
FR-5	Quality analysis	Analyze with the acquired information of the water across various water quality indicator like (PH, Turbidity, TDS, Temperature) using different models.
FR-6	Model prediction	Confirming based on water quality index and shows the machine learning prediction (Good, Partially Good, Poor) with the percentage of presence of various parameter.
FR-7	Remote Visualization	Visualization through charts based on present and past values of all the parameter for future forecast.
FR-8	Notification services	Confirming through notification of water status prediction with parameter presence along with timestamp.

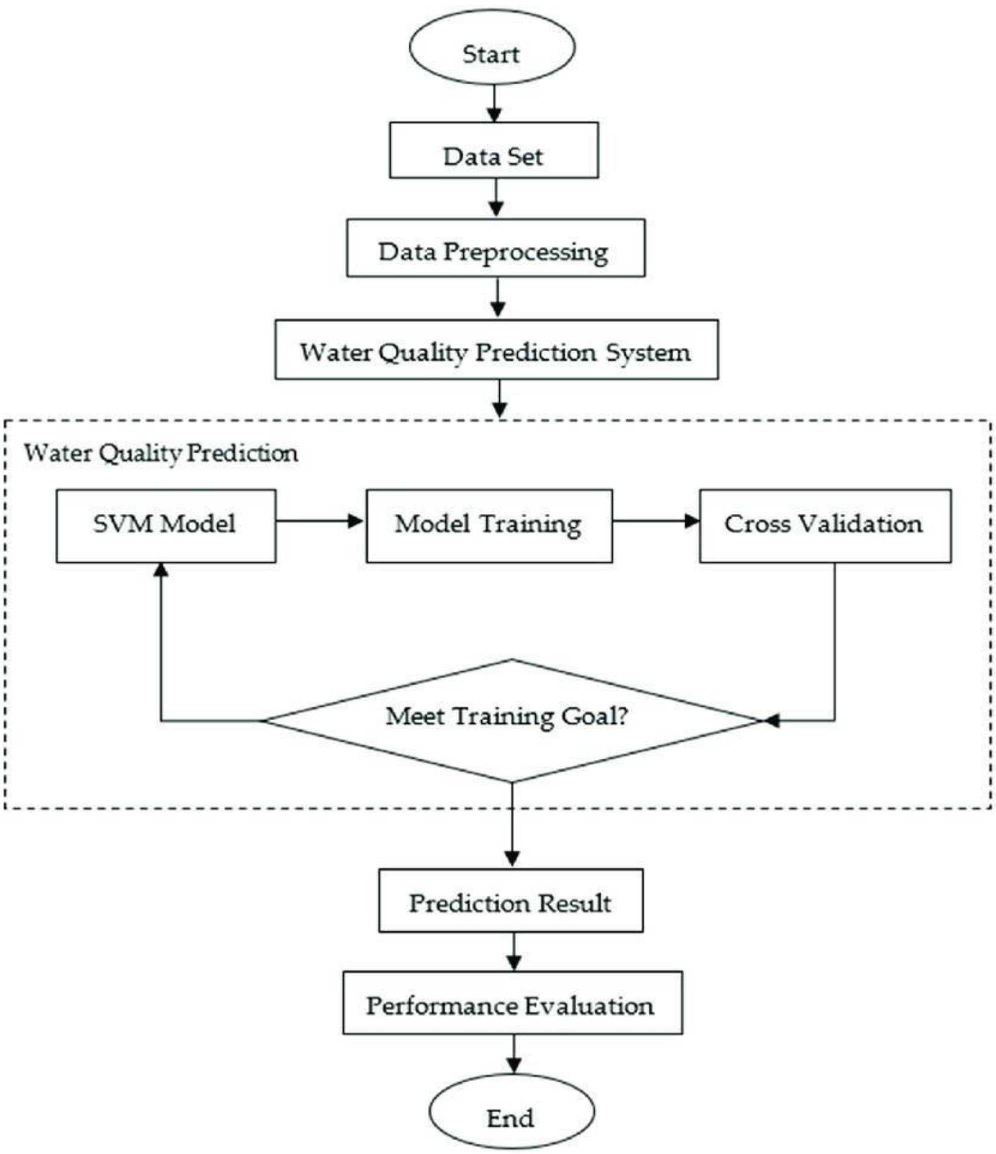
FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Registration through Form Registration through Gmail Registration through LinkedIN
FR-2	User Confirmation	Confirmation via Email Confirmation via OTP
FR-3	Executive administration	Regulation of monitoring the water environment status and regulatory compliance like pollution event emergency management, and it includes two different functions: early warning/forecast monitoring.
FR-4	Data handling	File contains water quality metrics for different water bodies.
FR-5	Quality analysis	Analyze with the acquired information of the water across various water quality indicator like (PH, Turbidity, TDS, Temperature) using different models.
FR-6	Model prediction	Confirming based on water quality index and shows the machine learning prediction (Good, Partially Good, Poor) with the percentage of presence of various parameter.
FR-7	Remote Visualization	Visualization through charts based on present and past values of all the parameter for future forecast.
FR-8	Notification services	Confirming through notification of water status prediction with parameter presence along with timestamp.

4.2 Non-Functional requirements

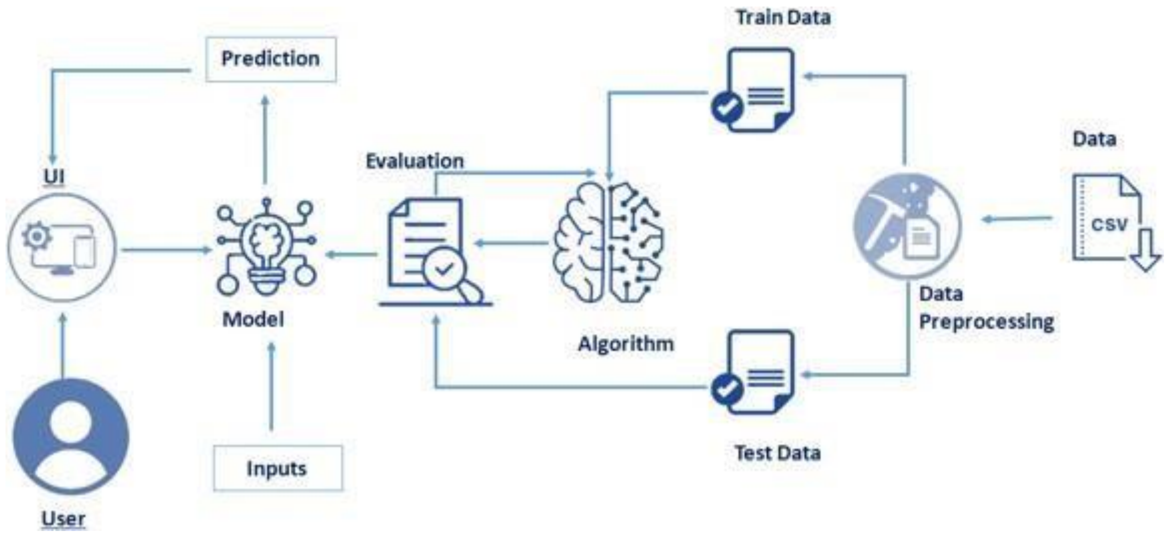
FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	The system provides a natural interaction with the users. Accurate water quality prediction with short time analysis and provide prediction safe to drink or not using some parameters and provide a great significance for water environment protection.
NFR-2	**Security**	The model enables with the high security system as the user's data will not be shared to the other sources. The system is protected with the user name and password throughout the process.
NFR-3	**Reliability**	The system is very reliable as it can last for long period of time when it is well maintained. The model can be extended in large scale by increasing the datasets.
NFR-4	**Performance**	Our system should run on 32 bit (x86) or 64 bit (x64) Dual-core 2.66-GHZ or faster processor. It should not exceed 2 GB RAM.
NFR-5	**Availability**	The system should be available for the duration of the user access the system until the user terminate the access. The system response to request of the user in less time and the recovery is done is less time.
NFR-6	**Scalability**	It provides an efficient outcome and has the ability to increase or decrease the performance of the system based on the datasets.

5. PROJECT DESIGN

5.1 Data Flow Diagrams



5.2 Solution &Technical Architecture



6. PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Data Collection	USN-1	Collecting dataset for pre-processing	10	High	Billu Dilip Budharaju Sumanth Aduri Charansai Bandla Venkata Akash Bhuvanendra Chowdary V
Sprint-1		USN-2	Data pre-processing-Used to transform the data into useful format.	10	Medium	Billu Dilip Budharaju Sumanth Aduri Charansai Bandla Venkata Akash Bhuvanendra Chowdary V
Sprint-2	Model Building	USN-3	Calculate the Water Quality Index (WQI) using Regression algorithm of machine learning.	10	High	Billu Dilip Budharaju Sumanth Aduri Charansai Bandla Venkata Akash Bhuvanendra Chowdary V
Sprint-2		USN-4	Splitting the data into training and testing from the entire dataset.	10	Medium	Billu Dilip Budharaju Sumanth Aduri Charansai Bandla Venkata Akash Bhuvanendra Chowdary V
Sprint-3	Training and Testing	USN-5	Training the model using regression algorithm and testing the performance of the model	20	Medium	Billu Dilip Budharaju Sumanth Aduri Charansai Bandla Venkata Akash Bhuvanendra Chowdary V
Sprint-4	Implementation of Web page	USN-6	Implementing the web page for collecting the data from user	10	High	Billu Dilip Budharaju Sumanth Aduri Charansai Bandla Venkata Akash Bhuvanendra Chowdary V
Sprint-4		USN-6	Deploying the model using IBM Cloud and IBM Watson Studio	10	Medium	Billu Dilip Budharaju Sumanth Aduri Charansai Bandla Venkata Akash Bhuvanendra Chowdary V

6.2 Sprint Delivery Schedule

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022	20	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022		
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022		
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022		

Velocity:

Sprint 1 Average Velocity:
Average Velocity = 20/2 = 10

Sprint 2 Average Velocity:
Average Velocity = 20/2 = 10

Sprint 3 Average Velocity:
Average Velocity = 20/1 = 20

Sprint 4 Average Velocity:
Average Velocity = 20/2 = 10

6.3 Reports from JIRA

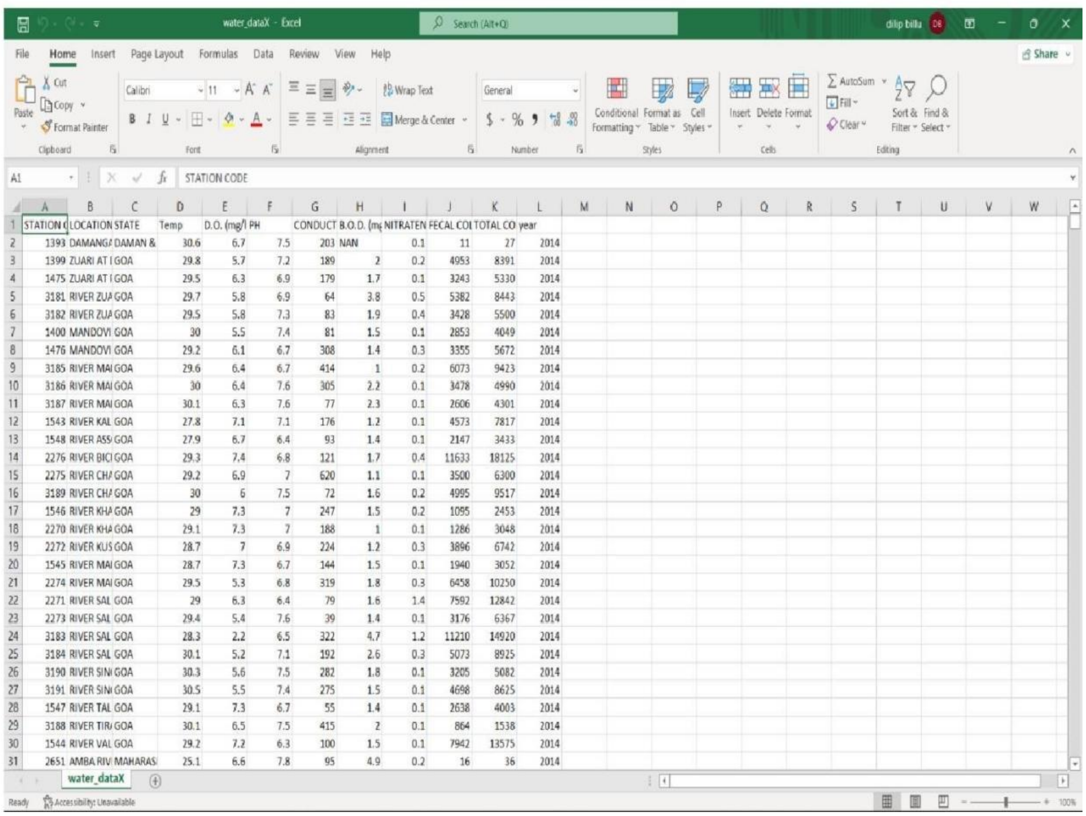
7. CODING & SOLUTIONING

7.1 FEATURE 1

Data collection and creation:

Data mining techniques require domain knowledge in order to generate predictions. For water quality applications, it is vital to understand how various water quality parameters influence water quality.Thisinformationcancomefromadomainexpertorhistoricaldatacollections.Forthe forecasting task, two types of data sets were used: a carefully created huge synthetic data set and an available real dataset

Data Collection



STATION C	LOCATION	STATE	Temp	D.O. (mg/l	PH	CONDUCT	B.O.D. (mg	NITRATEN	FECAL COL	TOTAL CO	year
1393	DAMANGA	DAMAN &	30.6	6.7	7.5	203	NAN	0.1	11	27	2014
1399	ZUARI AT I	GOA	29.8	5.7	7.2	189	2	0.2	4953	8391	2014
1475	ZUARI AT I	GOA	29.5	6.3	6.9	179	1.7	0.1	3243	5330	2014
3181	RIVER ZUA	GOA	29.7	5.8	6.9	64	3.8	0.5	5382	8443	2014
3182	RIVER ZUA	GOA	29.5	5.8	7.3	83	1.9	0.4	3428	5500	2014
1400	MANDOVI	GOA	30	5.5	7.4	81	1.5	0.1	2853	4049	2014
1476	MANDOVI	GOA	29.2	6.1	6.7	308	1.4	0.3	3355	5672	2014
3185	RIVER MAI	GOA	29.6	6.4	6.7	414	1	0.2	6073	9423	2014
3186	RIVER MAI	GOA	30	6.4	7.6	305	2.2	0.1	3478	4990	2014
3187	RIVER MAI	GOA	30.1	6.3	7.6	77	2.3	0.1	2606	4301	2014
1543	RIVER KAL	GOA	27.8	7.1	7.1	176	1.2	0.1	4573	7817	2014
1548	RIVER ASS	GOA	27.9	6.7	6.4	93	1.4	0.1	2147	3433	2014
2276	RIVER BICI	GOA	29.3	7.4	6.8	121	1.7	0.4	11633	18125	2014
2275	RIVER CHA	GOA	29.2	6.9	7	620	1.1	0.1	3500	6300	2014
3189	RIVER CHA	GOA	30	6	7.5	72	1.6	0.2	4995	9517	2014
1546	RIVER KHA	GOA	29	7.3	7	247	1.5	0.2	1095	2453	2014
2270	RIVER KHA	GOA	29.1	7.3	7	188	1	0.1	1286	3048	2014
2272	RIVER KUS	GOA	28.7	7	6.9	224	1.2	0.3	3896	6742	2014
1545	RIVER MAI	GOA	28.7	7.3	6.7	144	1.5	0.1	1940	3052	2014
2274	RIVER MAI	GOA	29.5	5.3	6.8	319	1.8	0.3	6458	10250	2014
2271	RIVER SAL	GOA	29	6.3	6.4	79	1.6	1.4	7592	12842	2014
2273	RIVER SAL	GOA	29.4	5.4	7.6	39	1.4	0.1	3176	6367	2014
3183	RIVER SAL	GOA	28.3	2.2	6.5	322	4.7	1.2	11210	14920	2014
3184	RIVER SAL	GOA	30.1	5.2	7.1	192	2.6	0.3	5073	8925	2014
3190	RIVER SIN	GOA	30.3	5.6	7.5	282	1.8	0.1	3205	5082	2014
3191	RIVER SIN	GOA	30.5	5.5	7.4	275	1.5	0.1	4698	8625	2014
1547	RIVER TAL	GOA	29.1	7.3	6.7	55	1.4	0.1	2638	4003	2014
3188	RIVER TIR	GOA	30.1	6.5	7.5	415	2	0.1	864	1538	2014
1544	RIVER VAL	GOA	29.2	7.2	6.3	100	1.5	0.1	7942	13575	2014
2651	AMBA RIV	MAHARAS	25.1	6.6	7.8	95	4.9	0.2	16	36	2014

7.2 FEATURE 2

Performance Measures Results True Positives (TP) are when the model predicts the positive class properly. True Negatives (TN) is one of the components of a confusion matrix designed to demonstrate how classification algorithms work. Positive outcomes that the model predicted incorrectly are known as False Positives (FP). False Negatives (FN) are negative outcomes that the model predicts negative class. Accuracy is the most basic and intuitive performance metric, consisting of the ratio of successfully predicted observations to total observations. Accuracy=TP+TN/(TP+FP+FN+TN)

ANALYSING THE DATA:

In [5]:
```
data.head()
```
Out[5]:

	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
0	1393	DAMANGANGA AT D/S OF MADHUBAN, DAMAN	DAMAN & DIU	30.6	6.7	7.5	203	NAN	0.1	11	27
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
2	1475	ZUARI AT PANCHAWADI	GOA	29.5	6.3	6.9	179	1.7	0.1	3243	5330
3	3181	RIVER ZUARI AT BORIM BRIDGE	GOA	29.7	5.8	6.9	64	3.8	0.5	5382	8443
4	3182	RIVER ZUARI AT MARCAIM JETTY	GOA	29.5	5.8	7.3	83	1.9	0.4	3428	5500

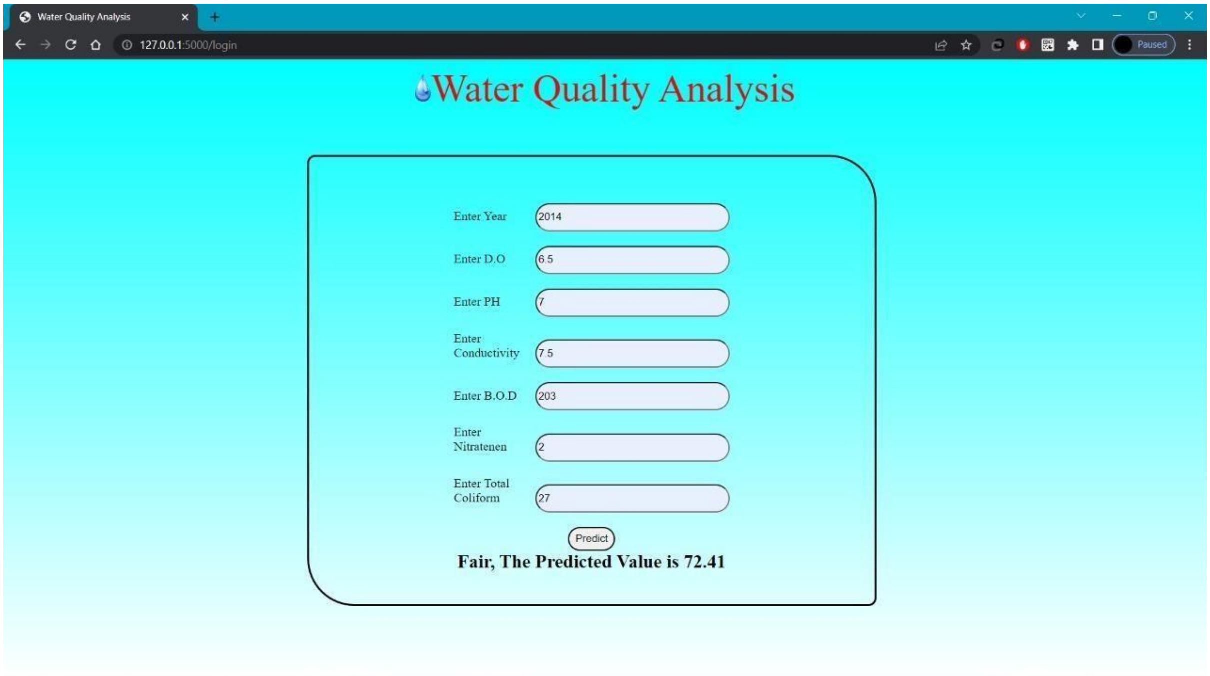
In [6]:
```
data.describe()
```
Out[6]:

	year
count	1991.000000
mean	2010.038172
std	3.057333
min	2003.000000
25%	2008.000000
50%	2011.000000
75%	2013.000000
max	2014.000000

In [7]:
```
data.info()
```
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1991 entries, 0 to 1990
Data columns (total 12 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   STATION CODE                        1991 non-null   object
 1   LOCATIONS                           1991 non-null   object
 2   STATE                               1991 non-null   object
 3   Temp                                1991 non-null   object
 4   D.O. (mg/l)                         1991 non-null   object
 5   PH                                  1991 non-null   object
 6   CONDUCTIVITY (µmhos/cm)             1991 non-null   object
 7   B.O.D. (mg/l)                       1991 non-null   object
 8   NITRATENAN N+ NITRITENANN (mg/l)    1991 non-null   object
 9   FECAL COLIFORM (MPN/100ml)          1991 non-null   object
 10  TOTAL COLIFORM (MPN/100ml)Mean      1991 non-null   object
```

8. TESTING

8.1 Test Case 1



8.2 Test Case 2

8.2User Acceptance Testing

1. Purpose of Document:

The purpose of this report is to briefly explain the test coverage and open issues of the project at the time of the release to User Acceptance Testing(UAT).

2. Defect Analysis:

This report shows the number of resolved or closed bugs at each severity level, and how they were resolved

Resolution	Severity 1	Severity 2	Severity 3	Severity 4	Subtotal
By Design	10	4	2	3	20
Duplicate	1	0	3	0	4
External	2	3	0	1	6
Fixed	11	2	4	20	37
Not Reproduced	0	0	1	0	1
Skipped	0	0	1	1	2
Won't Fix	0	5	2	1	8
Totals	24	14	13	26	77

3. Test Case Analysis:

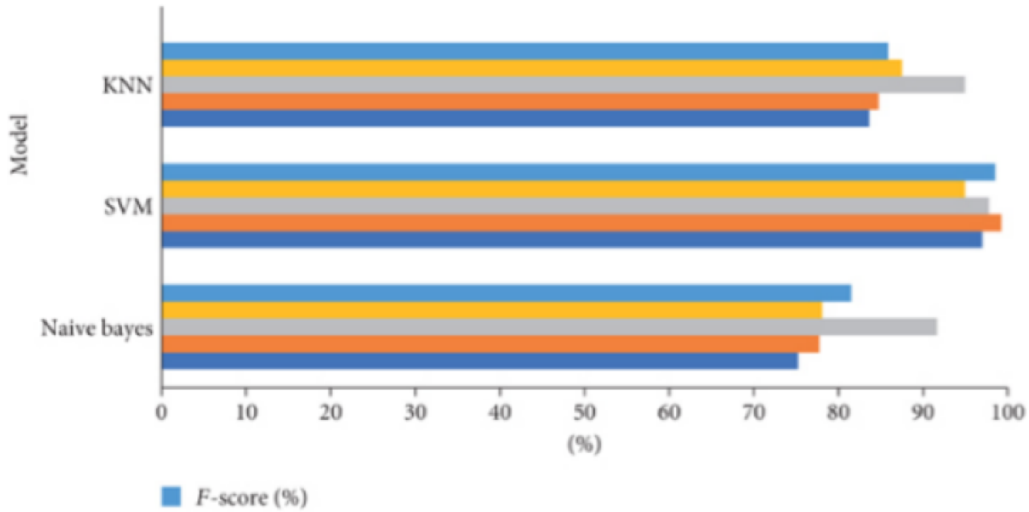
This report shows the number of test cases that have passed ,failed, and untested**.**

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

9.RESULT

9.1 PERFORMANCE METRICS

For validating the developed model, the dataset has been divided into 70% training and 30%testing subsets. While the ANN and LSTM models were used to predict the WQI, the SVM, KNN, and Naive Bayes were utilized for the water quality classification prediction



SO,WE ARE GOING TO USE SVC

Performance Measures Results True Positives (TP) are when the model predicts the positive class properly. True Negatives (TN) is one of the components of a confusion matrix designed to demonstrate how classification algorithms work. Positive outcomes that the model predicted incorrectly are known as False Positives (FP). False Negatives (FN) are negative outcomes that the model predicts negative class. Accuracy is the most basic and intuitive performance metric, consisting of the ratio of successfully predicted observations to tota lobservations .Accuracy=TP+TN/(TP+FP+FN+TN)

10. 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.

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 networks erver required**.**

11. CONCLUSION

Portability determines the quality of water, which is one of the most important resources for existence. Traditionally, testing water quality required an expensive and time-consuming lab analysis. This study looked into an alternative machine learning method for predicting water quality using only a few simple water quality criteria. To estimate, a set of representative supervised machine learning algorithms was used. It would detect water of bad quality before it was released for consumption and notify the appropriate authorities It will hopefully reduce the number of individuals who drink low-quality water, lowering the risk of diseases like typhoid and diarrhea. In this case, using a prescriptive analysis based on projected values would result in future capabilities to assist decision and policy makers.

12. SOURCE CODE

Machinelearninghasbeenwidelyusedasapowerfultooltosolveproblemsin 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 machinelearning 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 machinelearning 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.

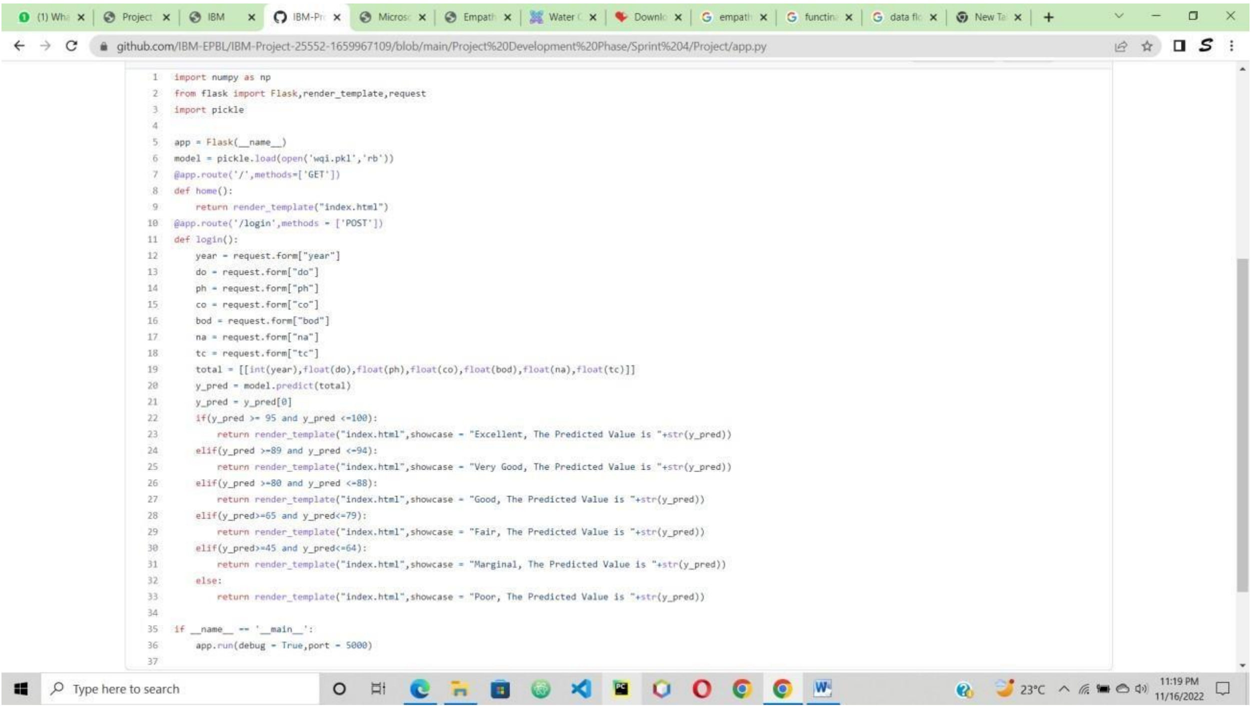
(3) Interdisciplinary talent with knowledge in different fields should be trained to develop more advanced machine learning techniques and apply them in engineering practices.

13. APPENDIX

REQUIREMENT.TXT

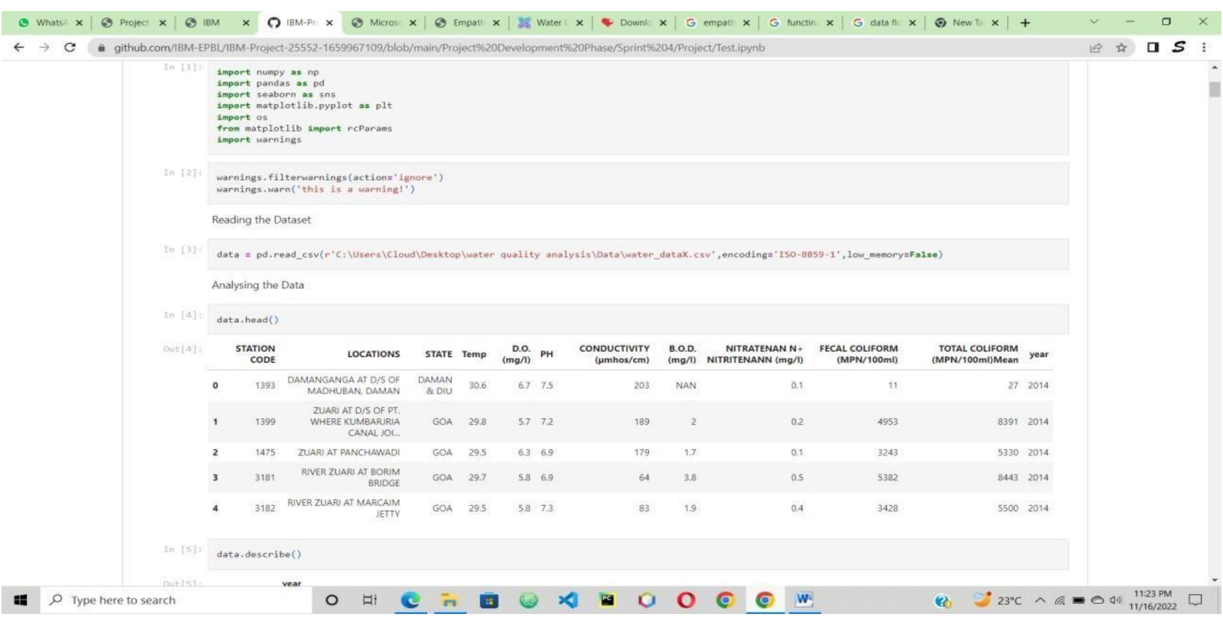
```
Flask == 2.2.2

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

gevent

requests

flask-cors==3.0.10
```

APP.PY




```python
import numpy as np
from flask import Flask,render_template,request
import pickle

app = Flask(__name__)
model = pickle.load(open('wqi.pkl','rb'))
@app.route('/',methods=['GET'])
def home():
    return render_template("index.html")
@app.route('/login',methods = ['POST'])
def login():
    year = request.form["year"]
    do = request.form["do"]
    ph = request.form["ph"]
    co = request.form["co"]
    bod = request.form["bod"]
    na = request.form["na"]
    tc = request.form["tc"]
    total = [[int(year),float(do),float(ph),float(co),float(bod),float(na),float(tc)]]
    y_pred = model.predict(total)
    y_pred = y_pred[0]
    if(y_pred >= 95 and y_pred <=100):
        return render_template("index.html",showcase = "Excellent, The Predicted Value is "+str(y_pred))
    elif(y_pred >=89 and y_pred <=94):
        return render_template("index.html",showcase = "Very Good, The Predicted Value is "+str(y_pred))
    elif(y_pred >=80 and y_pred <=88):
        return render_template("index.html",showcase = "Good, The Predicted Value is "+str(y_pred))
    elif(y_pred>=65 and y_pred<=79):
        return render_template("index.html",showcase = "Fair, The Predicted Value is "+str(y_pred))
    elif(y_pred>=45 and y_pred<=64):
        return render_template("index.html",showcase = "Marginal, The Predicted Value is "+str(y_pred))
    else:
        return render_template("index.html",showcase = "Poor, The Predicted Value is "+str(y_pred))

if __name__ == '__main__':
    app.run(debug = True,port = 5000)
```

TEST.IPYNB



In [1]:
```python
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]:
```python
warnings.filterwarnings(action='ignore')
warnings.warn('this is a warning!')
```

Reading the Dataset

In [3]:
```python
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]:
```python
data.head()
```

Out[4]:

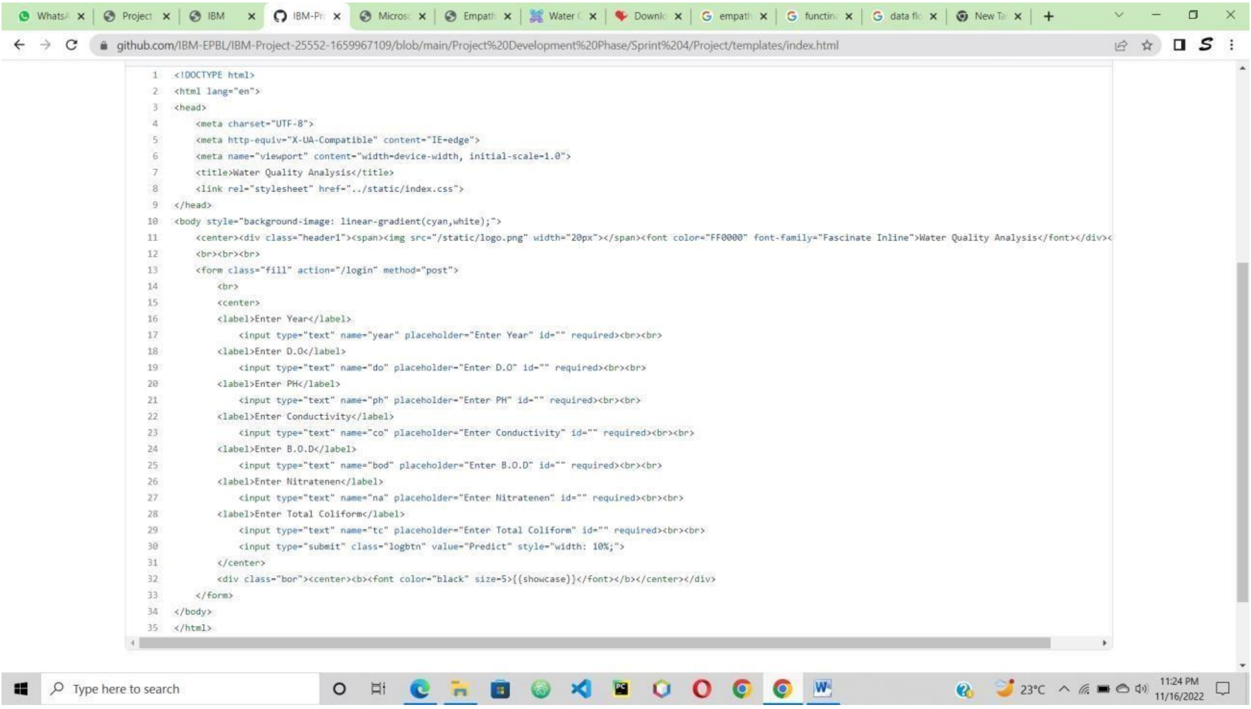
	STATION CODE	LOCATIONS	STATE	Temp	D.O. (mg/l)	PH	CONDUCTIVITY (µmhos/cm)	B.O.D. (mg/l)	NITRATENAN N+ NITRITENANN (mg/l)	FECAL COLIFORM (MPN/100ml)	TOTAL COLIFORM (MPN/100ml)Mean	year
0	1393	DAMANGANGA AT D/S OF MADHUBAN, DAMAN	DAMAN & DIU	30.6	6.7	7.5	203	NAN	0.1	11	27	2014
1	1399	ZUARI AT D/S OF PT. 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]:
```python
data.describe()
```

Out[5]:

year

INDEX.HTML



```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="../static/index.css">
</head>
<body style="background-image: linear-gradient(cyan,white);">
    <center><div class="header1"><span><img src="/static/logo.png" width="20px"></span><font color="FF0000" font-family="Fascinate Inline">Water Quality Analysis</font></div><
    <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: 10%;">
        </center>
        <div class="bor"><center><b><font color="black" size=5>{{showcase}}</font></b></center></div>
    </form>
</body>
</html>
```

LINKS:

GITHUB- https://github.com/IBM-EPBL/IBM-Project-53536-1661414817/blob/main/DESIGN%20AND%20PLANNING/IDEATION/Empathy%20map.pdf

DEMOLINK- https://drive.google.com/file/d/1DBRB6RwlyfBx989eLSieViEQXm2BSQci/view?usp=drivesdk