EFFICIENT WATER QUALITY ANALYSIS AND PREDICTION USING MACHINE LEARNING

DONE BY

TEAM ID: PNT2022TMID23758

RAJARANGANAYAKI R
RAKSHAMBIKA S
SHANMUGA VALLI S
VIDHYA P

In partial fulfillment for the award of the degree of

In

BACHELOR OF ENGINEERING

In

COMPUTER SCIENCE AND ENGINEERING

VIVEKANANDHA COLLEGE OF ENGINEERING FOR WOMEN

NOV 2022

CONTENTS

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 groundwater 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

Hence, rapid industrial development has prompted the decay of water quality at a disturbing rate. Furthermore, infrastructures, with the absence of public awareness, and less hygienic qualities, significantly affect the quality of drinking water . In fact, the consequences of polluted drinking water are so dangerous and can badly affect health, the environment, and infrastructures. As per the United Nations (UN) report, about 1.5 million people die each year because of contaminated water-driven diseases. In developing countries, it is announced that 80% of health problems are caused by contaminated water. Five million deaths and 2.5 billion illnesses are reported annually .Such a mortality rate is higher than deaths resulting from accidents, crimes, and terrorist attacks . Therefore, it is very important to suggest new approaches to analyze and, if possible, to predict the water quality (WQ). It is recommended to consider the temporal dimension for forecasting the WQ patterns to ensure the monitoring of the seasonal change of the WQ . However, using a special variation of models together to predict the WQ grants better results than using a single model . There are several methodologies proposed for the prediction and modeling of the WQ. These methodologies include statistical approaches, visual modeling, analyzing algorithms, and predictive algorithms. For the sake of the determination of the correlation and relationship among different water quality parameters, multivariate statistical techniques have been employed . The geostatistical approaches were used for transitional probability, multivariate interpolation, and regression analysis . Massive increases in population, the industrial revolution, and the use of fertilizers and pesticides have led to serious effects on the WQ environments . Thus, having models for the prediction of the WQ is of great help for monitoring water contamination.

2. LITERATURE SURVEY

Many works had been conducted to predict water quality using Machine Learning (ML) approaches. Some researchers used the traditional Machine Learning models, such as Decision Tree , 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 decision tree. 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 of 94%, while Decision Stump showed the lowest accuracy. Another study by Jeihouni et al. also compared five decision-tree-based models, which are Random Tree, Random Forest, Ordinary Decision Tree (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 among output 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 Singh used 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-square error (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 make the ensemble learners work properly . The ensemble 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 classification model that uses multiple base models, typically decision trees, on a given subset of data independently and makes decisions based on all models
. It uses feature randomness and bagging when building each individual decision tree 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(Water Quality Index) is calculated using AI techniques. So in future work, we can integrate this with IoT based framework to study large datasets and to expand our study to a larger scale. By using that it can predict the water quality fast and more accurately than any other IoT framework. That IoT framework 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. Water quality index with missing parameters. Int. J. Res. Eng. Technol. 2013,2,609–614.

PCRWR. Water Quality of Filtration Plants, Monitoring Report; PCRWR: Islamabad, Pakistan, 2010. Availableonline:http://www.pcrwr.gov.pk/Publications/Water%20Quality%20Reports/FILTRTAION%20PLANTS% 20REPOT-CDA.pdf (accessed on 23 August 2019).

Sakizadeh, M. Artificial intelligence for the prediction of water quality index in groundwater systems.Model. Earth Syst. Environ. 2016, 2, 8. [CrossRef]

2.3Problem Statement Definition

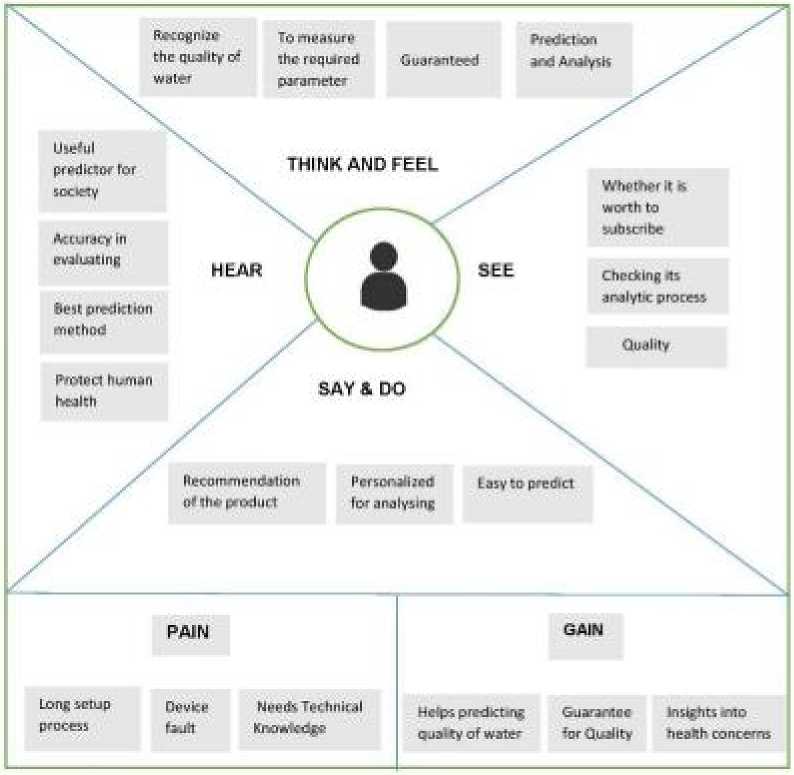
Access to safe drinking-water is essential to health, a basic human right and a component of effectivepolicy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation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

3. IDEATION & PROPOSED SOLUTION

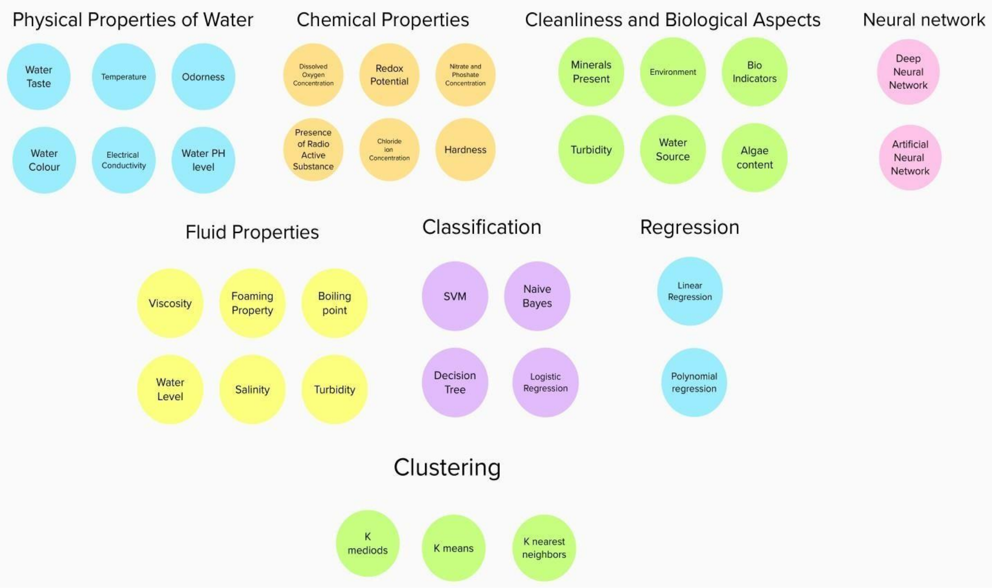
3.1 Empathy Map Canvas

An empathy map canvas serves as a foundation for outstanding user experiences, which focuson 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 usersare and what they want from the brand.


Recognize the quality of water
To measure the required parameter
Guaranteed
Prediction and Analysis
THINK AND FEEL
Useful predictor for society
Accuracy in evaluating
Best prediction method
Protect human health
HEAR
SEE
Whether it is worth to subscribe
Checking its analytic process
Quality
SAY & DO
Recommendation of the product
Personalized for analysing
Easy to predict
PAIN
Long setup process
Device fault
Needs Technical Knowledge
GAIN
Helps predicting quality of water
Guarantee for Quality
Insights into health concerns

3.2 Ideation & Brainstorming

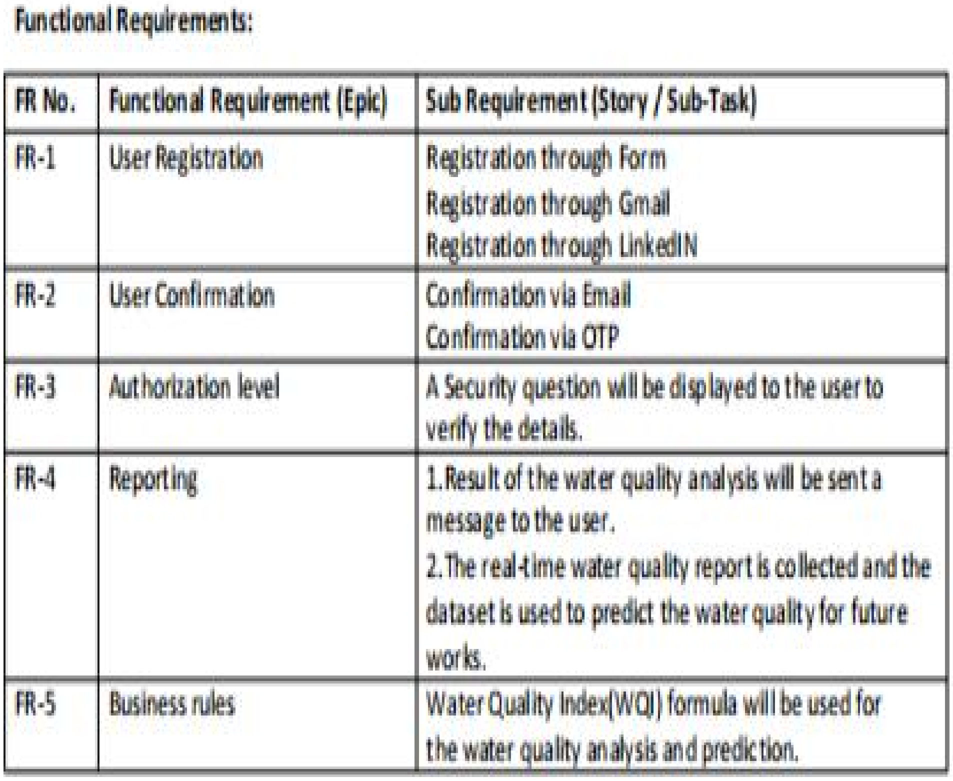


3.3 Proposed Solution

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 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.

4. REQUIREMENT ANALYSIS

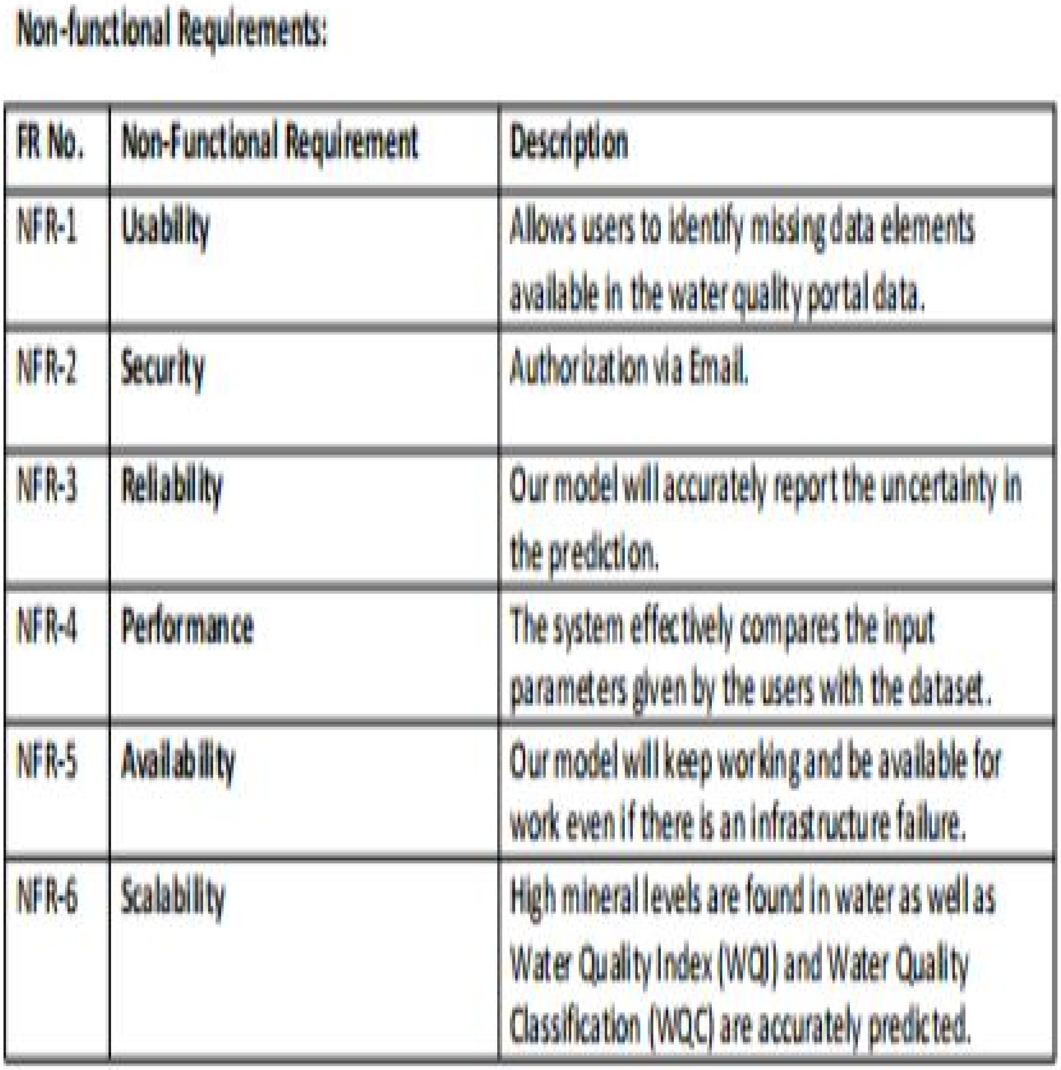
4.1 Functional requirement



Functional Requirements:

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	Authorization level	A Security question will be displayed to the user to verify the details.
FR-4	Reporting	1.Result of the water quality analysis will be sent a message to the user. 2.The real-time water quality report is collected and the dataset is used to predict the water quality for future works.
FR-5	Business rules	Water Quality Index(WQI) formula will be used for the water quality analysis and prediction.

4.2 Non-Functional requirements

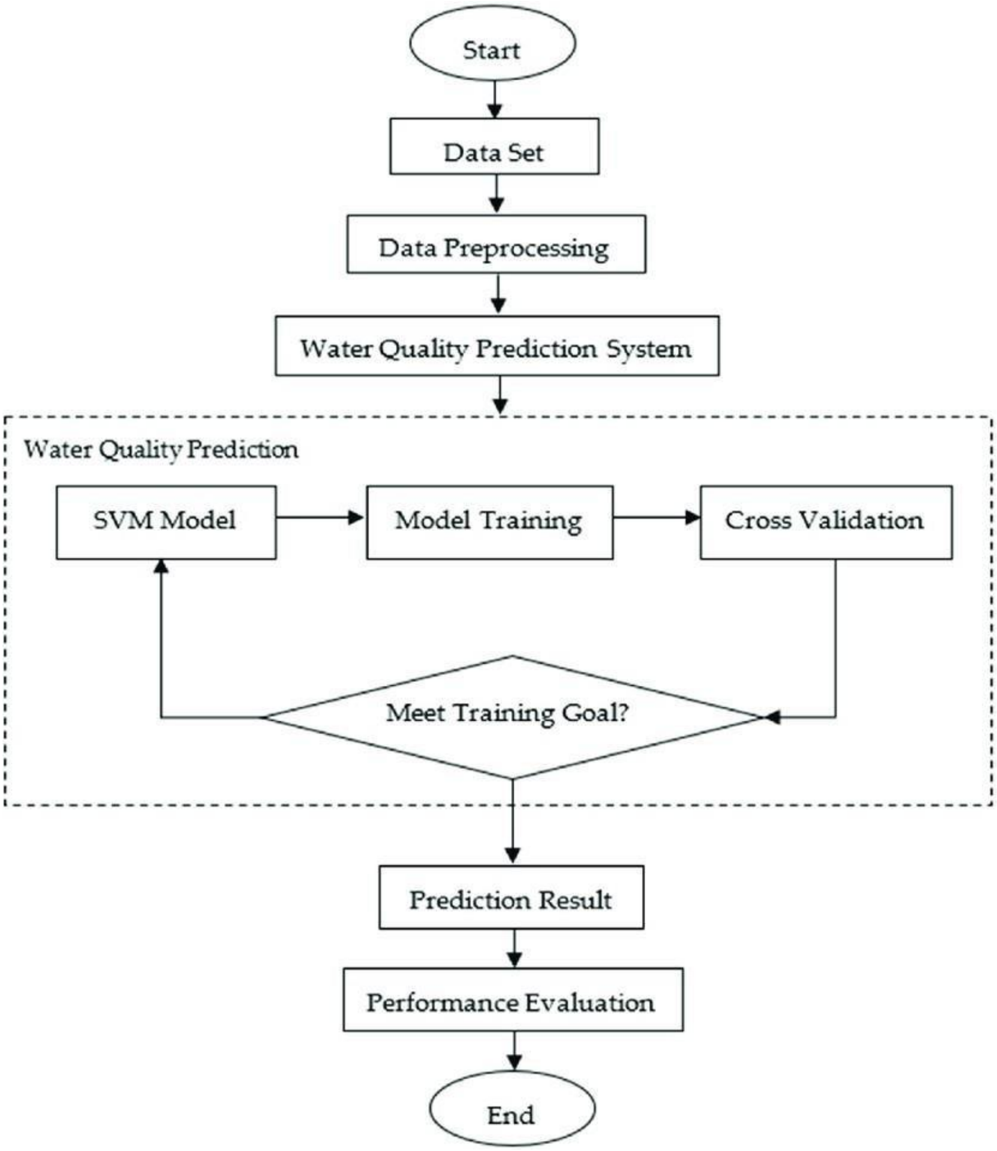


Non-functional Requirements:

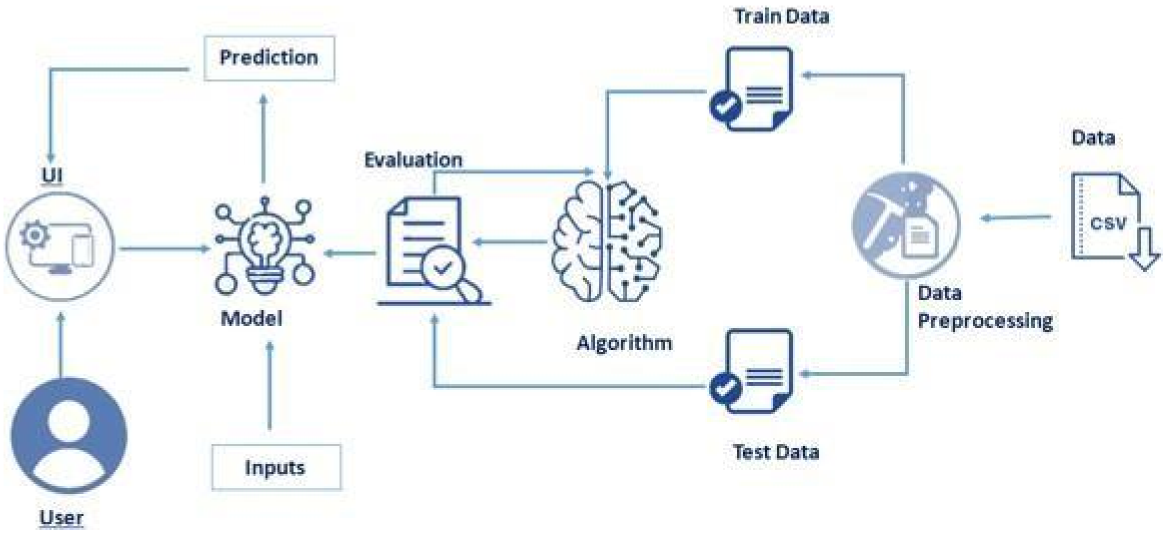
FR No.	Non-Functional Requirement	Description
NFR-1	Usability	Allows users to identify missing data elements available in the water quality portal data.
NFR-2	Security	Authorization via Email.
NFR-3	Reliability	Our model will accurately report the uncertainty in the prediction.
NFR-4	Performance	The system effectively compares the input parameters given by the users with the dataset.
NFR-5	Availability	Our model will keep working and be available for work even if there is an infrastructure failure.
NFR-6	Scalability	High mineral levels are found in water as well as Water Quality Index (WQI) and Water Quality Classification (WQC) are accurately predicted.

5. PROJECT DESIGN

5.1 Data Flow Diagrams



5.2 Solution & Technical Architecture

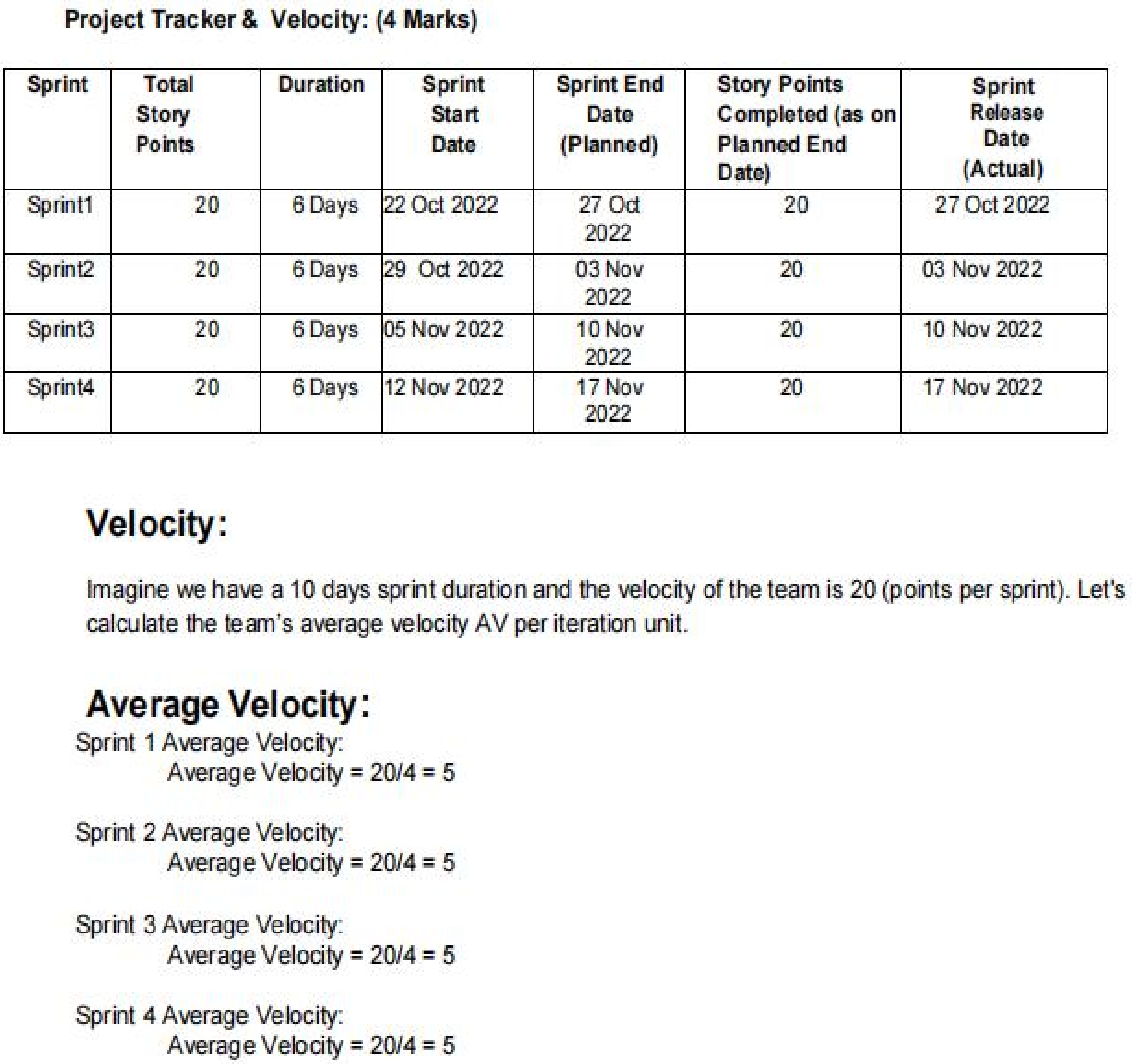


6. PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint1	Data Collection	USN-1,2	Collecting/ downloading dataset for pre- processing.	12	High	Rajaranganayaki R Shanmugavalli S Rakshambika S Vidhya P
Sprint1	Data Pre processing	USN-1,2	formats the data and handles the missing data in the dataset.	8	Medium	Rajaranganayaki R Shanmugavalli S Rakshambika S Vidhya P
Sprint2	Model Building	USN-1,2	Calculate the Water Quality Index (WQI) using specified formula for every parameter.	10	High	Shanmugavalli S Rajaranganayaki R Vidhya P Rakshambika S
Sprint2	Accessing datasets	USN-1,2	Splitting the data into training and testing dataset from the entire dataset.	10	High	Rakshambika S Rajaranganayaki R Shanmugavalli S Vidhya P
Sprint3	Training and Testing	USN-1,2	Training the model using Random Forest Regression algorithm and testing the performance of the model (accuracy rate)	20	High	Vidhya P Rajaranganayaki R Shanmugavalli S Rakshambika S
Sprint4	Implementation of Web page and user login	USN-1,2	Implementing the web page for collecting the data from user	12	High	Rajaranganayaki R Vidhya P Rakshambika S Shanmugavalli S
Sprint4	Web application	USN-1,2	It will display the current information of the water quality.	8	Medium	Shanmugavalli S Rajaranganayaki R Rakshambika S Vidhya P

6.2 Sprint Delivery Schedule



Project Tracker & Velocity: (4 Marks)

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint1	20	6 Days	22 Oct 2022	27 Oct 2022	20	27 Oct 2022
Sprint2	20	6 Days	29 Oct 2022	03 Nov 2022	20	03 Nov 2022
Sprint3	20	6 Days	05 Nov 2022	10 Nov 2022	20	10 Nov 2022
Sprint4	20	6 Days	12 Nov 2022	17 Nov 2022	20	17 Nov 2022

Velocity:

Imagine we have a 10 days sprint duration and the velocity of the team is 20 (points per sprint). Let's calculate the team's average velocity AV per iteration unit.

Average Velocity:

Sprint 1 Average Velocity:
Average Velocity = 20/4 = 5

Sprint 2 Average Velocity:
Average Velocity = 20/4 = 5

Sprint 3 Average Velocity:
Average Velocity = 20/4 = 5

Sprint 4 Average Velocity:
Average Velocity = 20/4 = 5

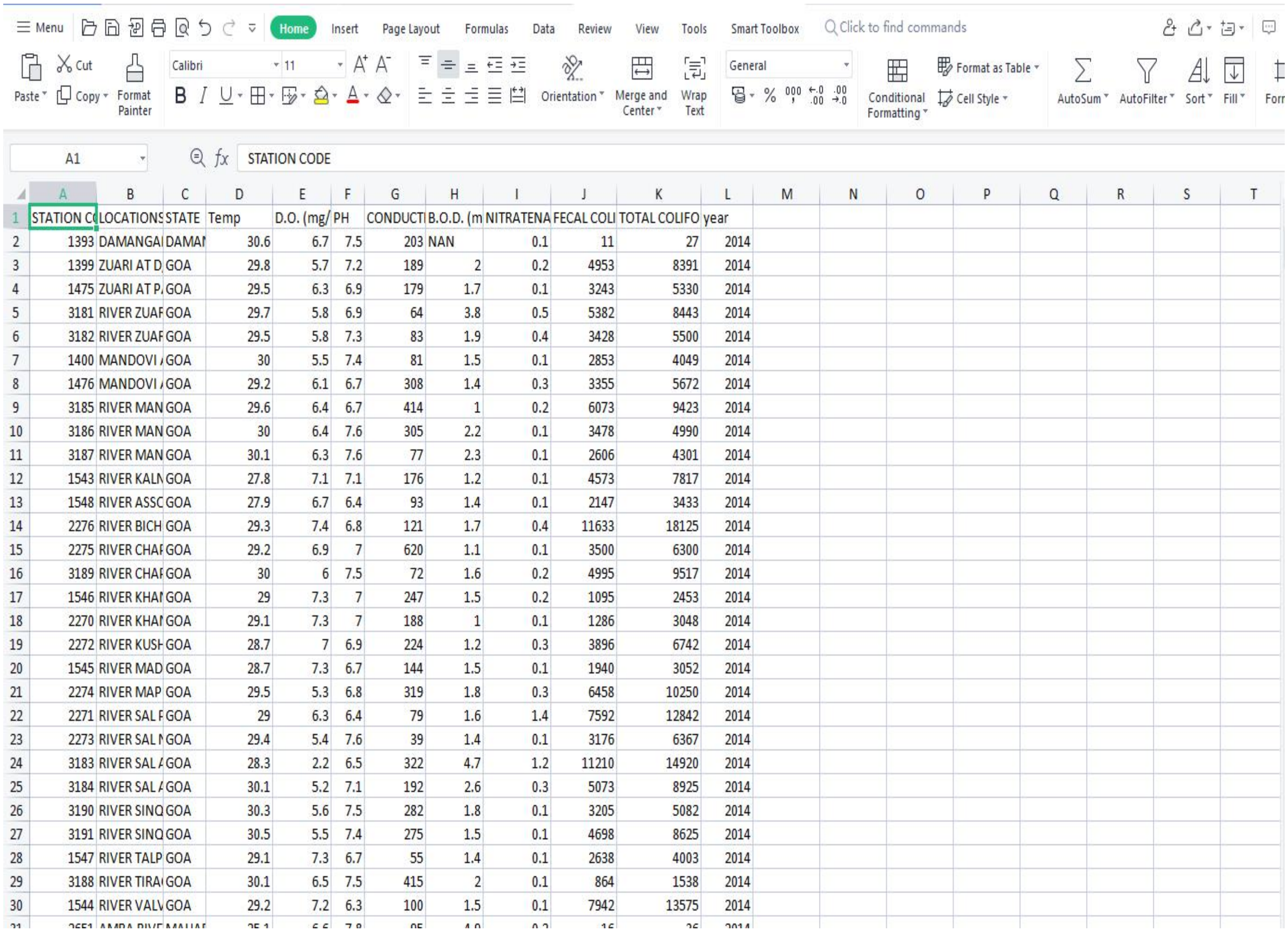
7. CODING & SOLUTIONING

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. This information can come from a domain expert or historical data collections. For the forecasting task, two types of data sets were used: a carefully created huge synthetic data set and an available real data set

Data Collection

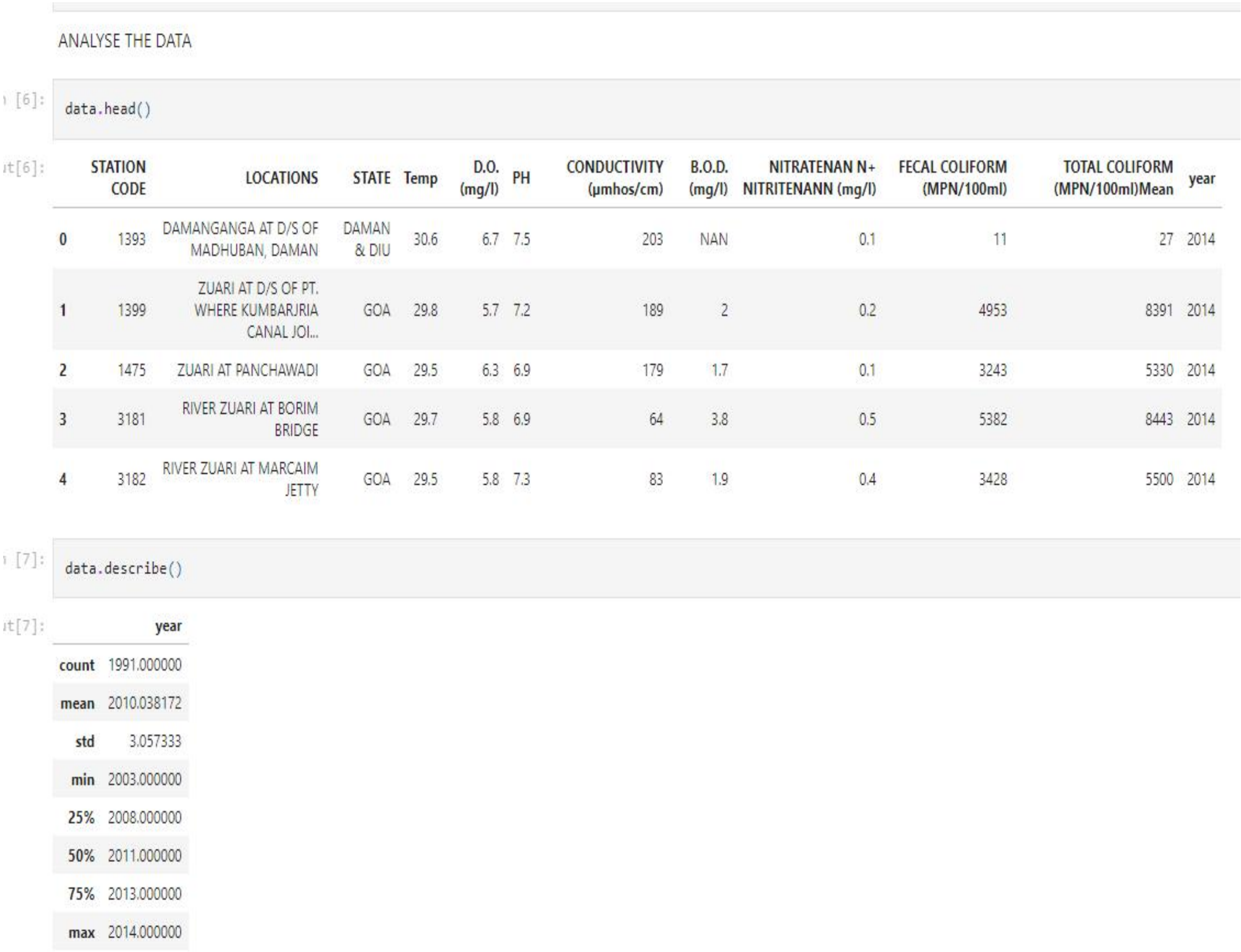


A1 fx STATION CODE

	A	B	C	D	E	F	G	H	I	J	K	L
1	STATION C	LOCATIONS	STATE	Temp	D.O. (mg/	PH	CONDUCTI	B.O.D. (m	NITRATENA	FECAL COLI	TOTAL COLIFO	year
2	1393	DAMANGAI	DAMAI	30.6	6.7	7.5	203	NAN	0.1	11	27	2014
3	1399	ZUARI AT D	GOA	29.8	5.7	7.2	189	2	0.2	4953	8391	2014
4	1475	ZUARI AT P.	GOA	29.5	6.3	6.9	179	1.7	0.1	3243	5330	2014
5	3181	RIVER ZUAF	GOA	29.7	5.8	6.9	64	3.8	0.5	5382	8443	2014
6	3182	RIVER ZUAF	GOA	29.5	5.8	7.3	83	1.9	0.4	3428	5500	2014
7	1400	MANDOVI /	GOA	30	5.5	7.4	81	1.5	0.1	2853	4049	2014
8	1476	MANDOVI /	GOA	29.2	6.1	6.7	308	1.4	0.3	3355	5672	2014
9	3185	RIVER MAN	GOA	29.6	6.4	6.7	414	1	0.2	6073	9423	2014
10	3186	RIVER MAN	GOA	30	6.4	7.6	305	2.2	0.1	3478	4990	2014
11	3187	RIVER MAN	GOA	30.1	6.3	7.6	77	2.3	0.1	2606	4301	2014
12	1543	RIVER KALN	GOA	27.8	7.1	7.1	176	1.2	0.1	4573	7817	2014
13	1548	RIVER ASSC	GOA	27.9	6.7	6.4	93	1.4	0.1	2147	3433	2014
14	2276	RIVER BICH	GOA	29.3	7.4	6.8	121	1.7	0.4	11633	18125	2014
15	2275	RIVER CHAF	GOA	29.2	6.9	7	620	1.1	0.1	3500	6300	2014
16	3189	RIVER CHAF	GOA	30	6	7.5	72	1.6	0.2	4995	9517	2014
17	1546	RIVER KHAI	GOA	29	7.3	7	247	1.5	0.2	1095	2453	2014
18	2270	RIVER KHAI	GOA	29.1	7.3	7	188	1	0.1	1286	3048	2014
19	2272	RIVER KUSH	GOA	28.7	7	6.9	224	1.2	0.3	3896	6742	2014
20	1545	RIVER MAD	GOA	28.7	7.3	6.7	144	1.5	0.1	1940	3052	2014
21	2274	RIVER MAP	GOA	29.5	5.3	6.8	319	1.8	0.3	6458	10250	2014
22	2271	RIVER SAL F	GOA	29	6.3	6.4	79	1.6	1.4	7592	12842	2014
23	2273	RIVER SAL N	GOA	29.4	5.4	7.6	39	1.4	0.1	3176	6367	2014
24	3183	RIVER SAL /	GOA	28.3	2.2	6.5	322	4.7	1.2	11210	14920	2014
25	3184	RIVER SAL /	GOA	30.1	5.2	7.1	192	2.6	0.3	5073	8925	2014
26	3190	RIVER SINQ	GOA	30.3	5.6	7.5	282	1.8	0.1	3205	5082	2014
27	3191	RIVER SINQ	GOA	30.5	5.5	7.4	275	1.5	0.1	4698	8625	2014
28	1547	RIVER TALP	GOA	29.1	7.3	6.7	55	1.4	0.1	2638	4003	2014
29	3188	RIVER TIRA	GOA	30.1	6.5	7.5	415	2	0.1	864	1538	2014
30	1544	RIVER VALV	GOA	29.2	7.2	6.3	100	1.5	0.1	7942	13575	2014
31	2651	AMBA RIVE	MAHAF	25.1	6.6	7.8	95	4.0	0.2	16	26	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, consistingof the ratio of successfully predicted observations to total observations. Accuracy = TP+TN/(TP+FP+FN+TN)



ANALYSE THE DATA

```
ı [6]: data.head()
```

ıt[6]:

	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

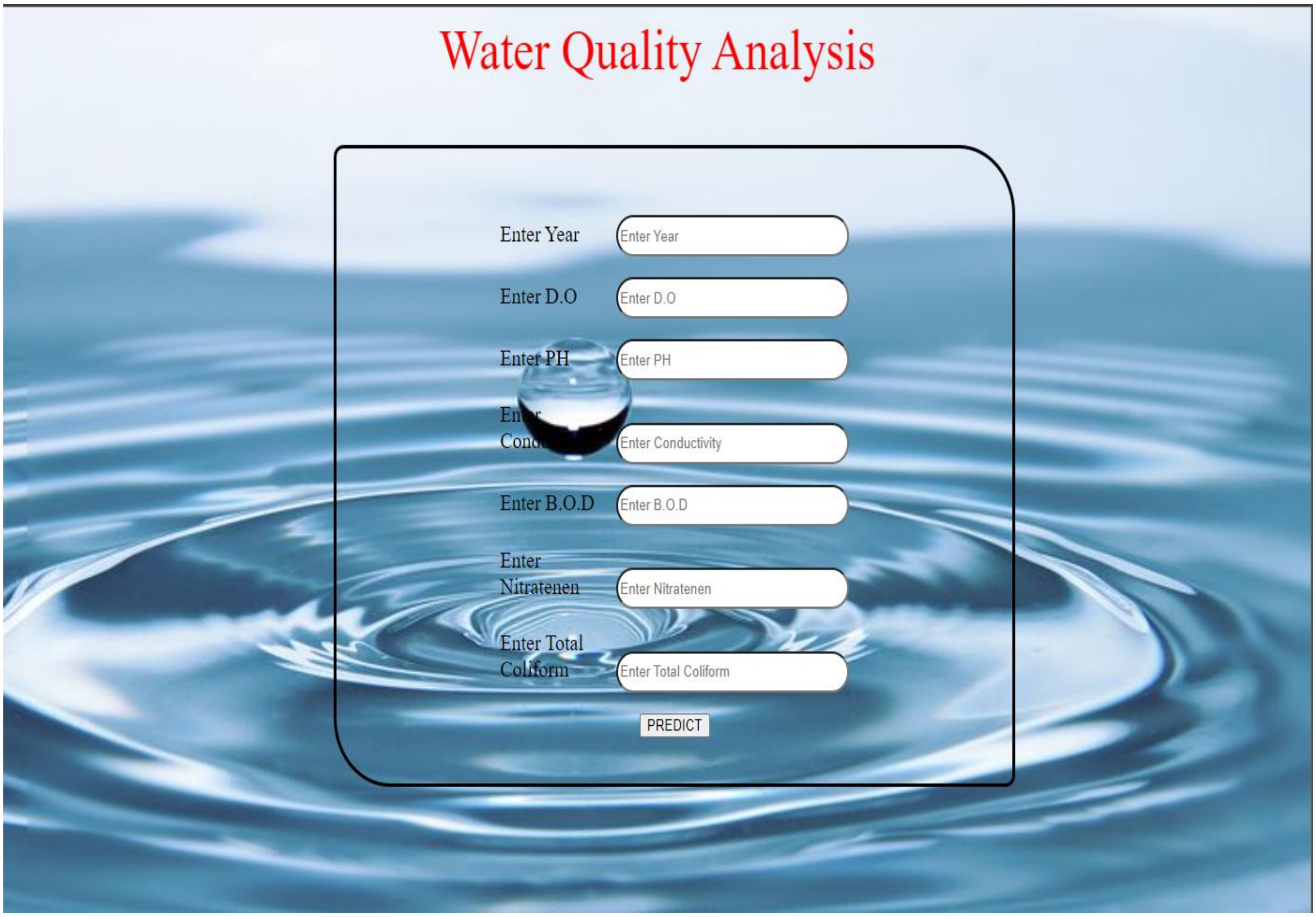
```
ı [7]: data.describe()
```

ıt[7]:

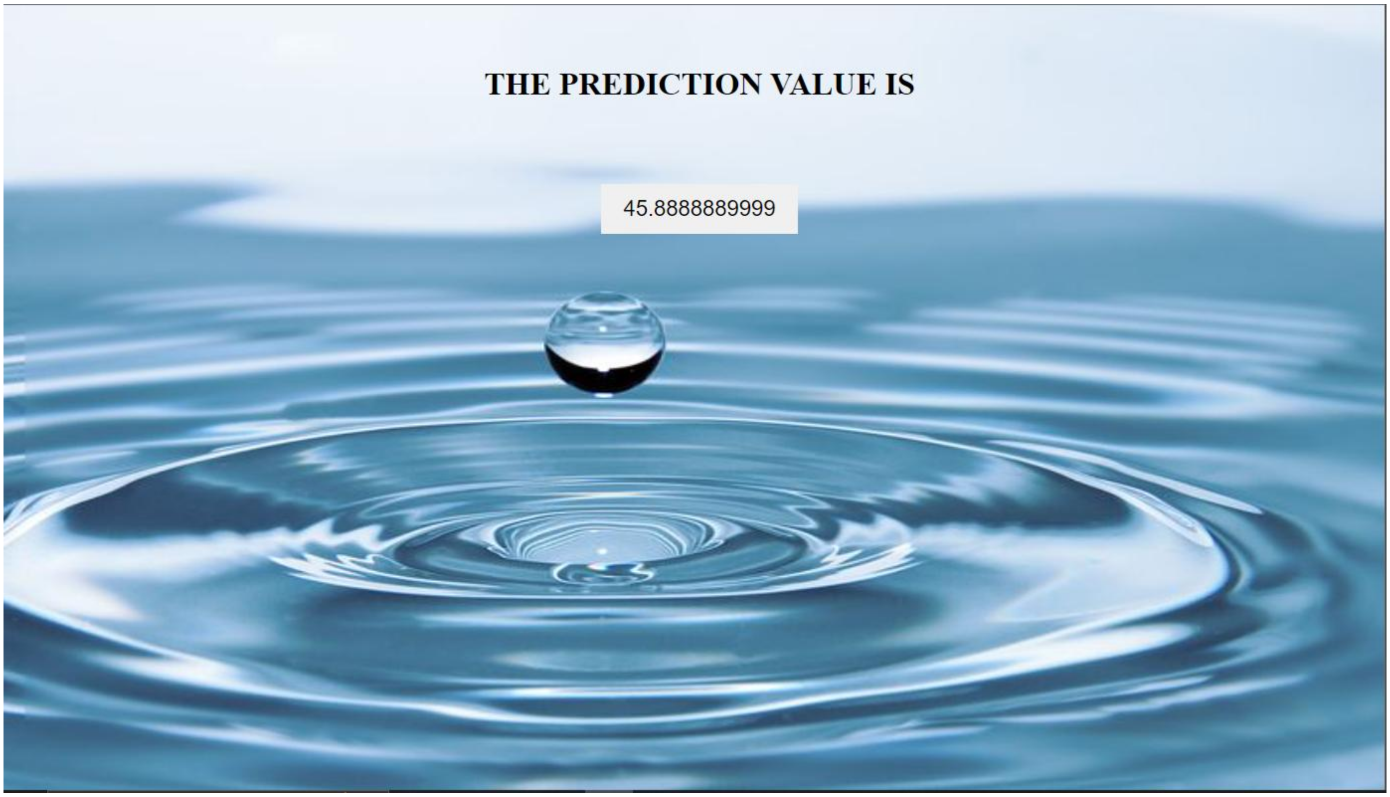
	year
count	1991.000000
mean	2010.038172
std	3.057333
min	2003.000000
25%	2008.000000
50%	2011.000000
75%	2013.000000
max	2014.000000

8. TESTING

8.1 Test Case1



8.2 Test Case2



THE PREDICTION VALUE IS

45.8888889999

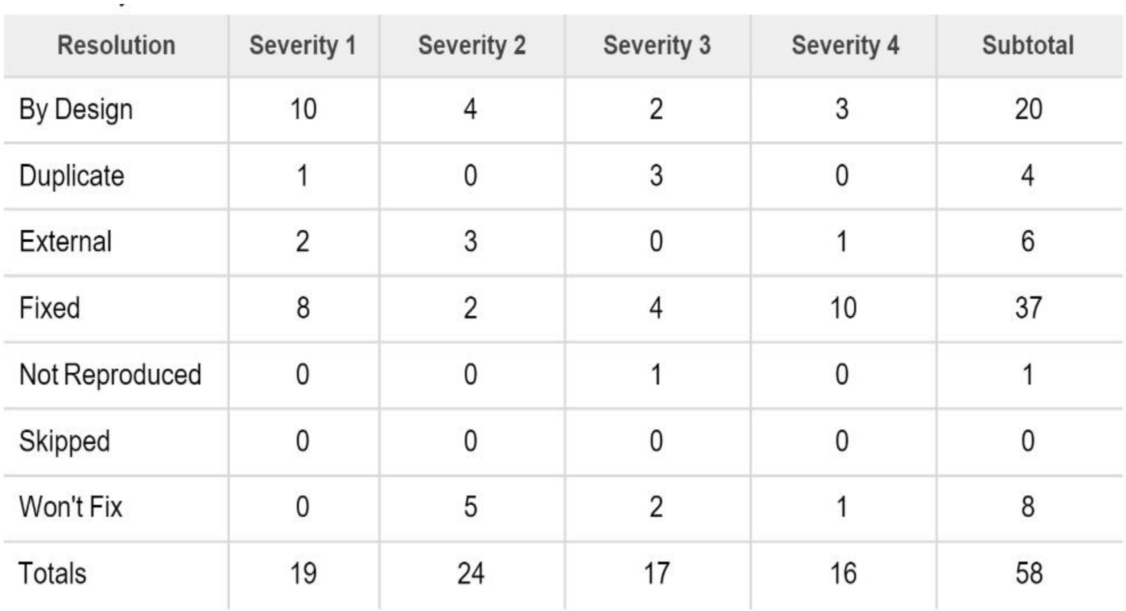
8.2 User Acceptance Testing

1. Purpose of Document :

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

2. Defect Analysis:

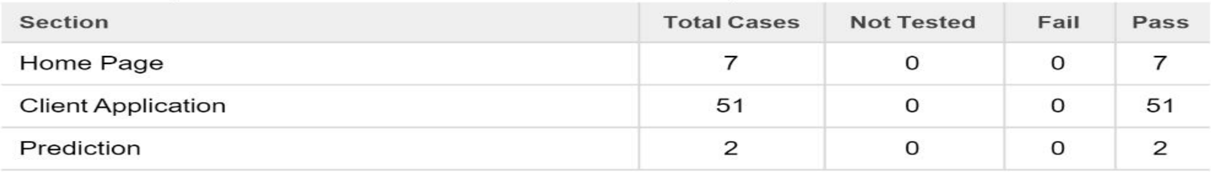
This report shows the number of resolved or closed bugs at each severity level, and how they wereresolved

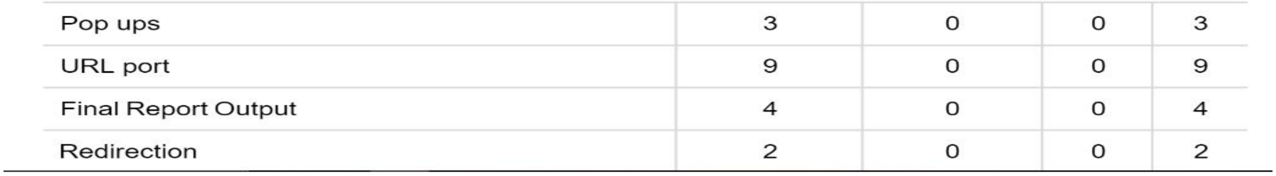


Resolution	Severity 1	Severity 2	Severity 3	Severity 4	Subtotal
By Design	10	4	2	3	20
Duplicate	1	0	3	0	4
External	2	3	0	1	6
Fixed	8	2	4	10	37
Not Reproduced	0	0	1	0	1
Skipped	0	0	0	0	0
Won't Fix	0	5	2	1	8
Totals	19	24	17	16	58

3. Test Case Analysis:

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



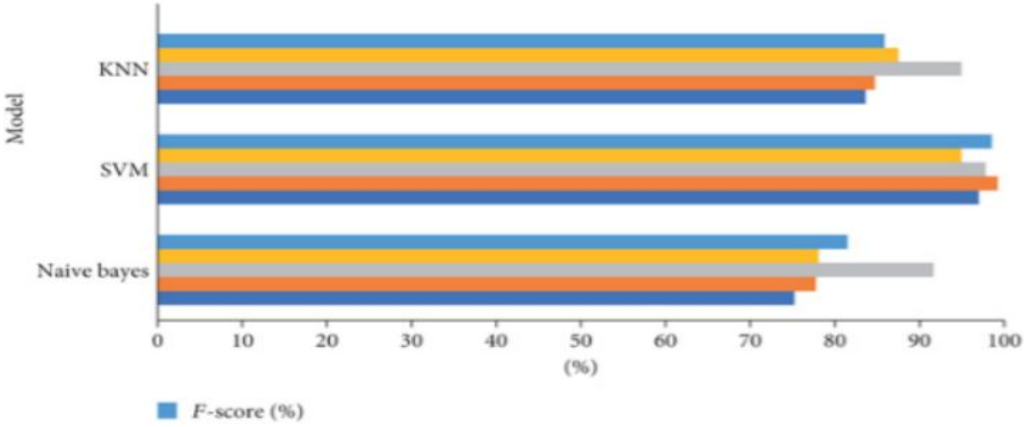


Section	Total Cases	Not Tested	Fail	Pass
Home Page	7	0	0	7
Client Application	51	0	0	51
Prediction	2	0	0	2
Pop ups	3	0	0	3
URL port	9	0	0	9
Final Report Output	4	0	0	4
Redirection	2	0	0	2

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, consistingof the ratio of successfully predicted observations to total observations. 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 network server 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 onlya 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

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 researchersto have certain professional background knowledge.

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

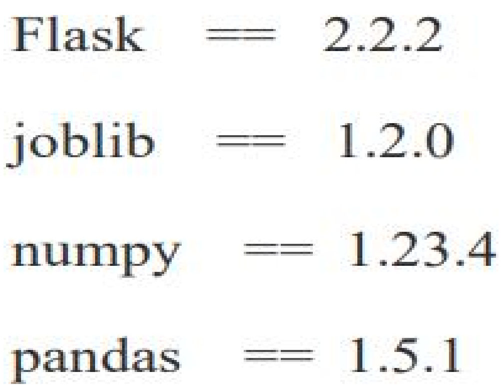
(1) More advanced sensors, including soft sensors, should be developed and applied in water quality monitoring to collect sufficiently accurate data to facilitate the application of machine learning approaches.

(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.

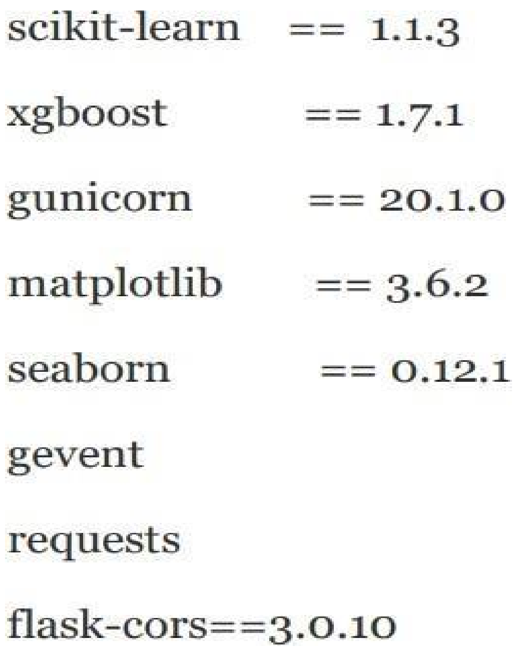
(3) Interdisciplinary talent with knowledge in different fields should be trained to develop more advanced machine learning techniques and apply them in engineering practices.

13. APPENDIX

REQUIREMENT.TXT

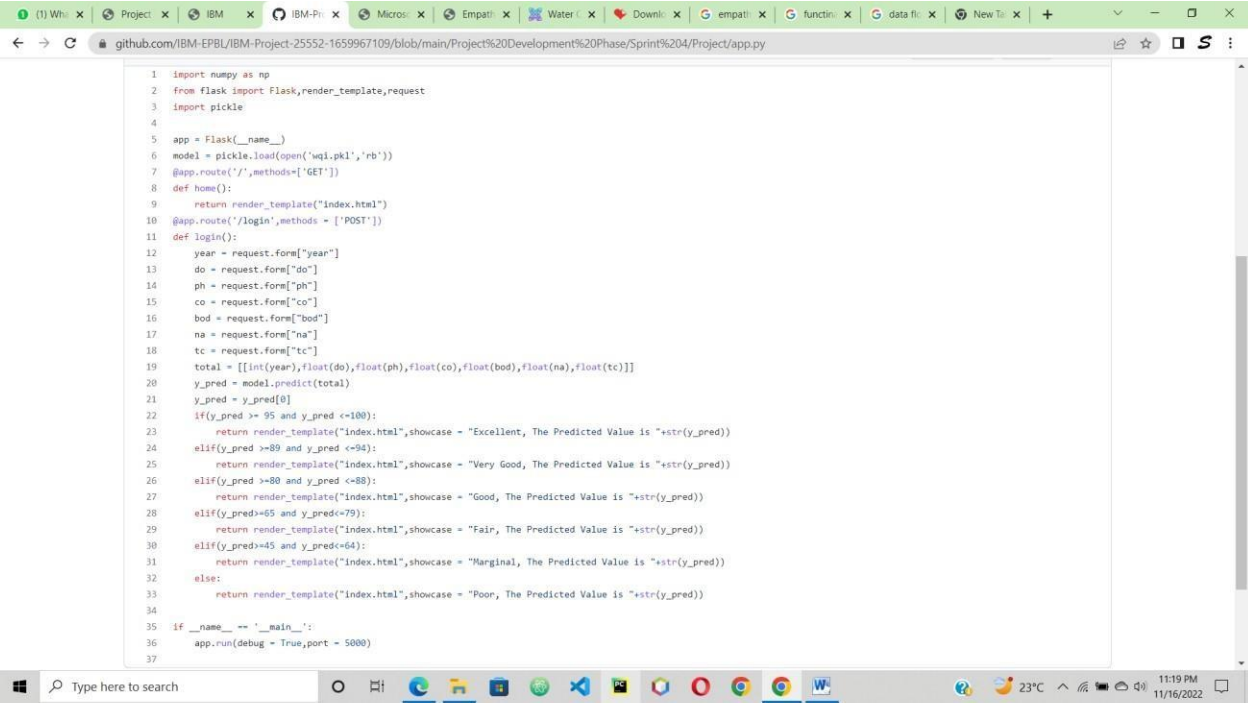


```
Flask   ==  2.2.2

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
```

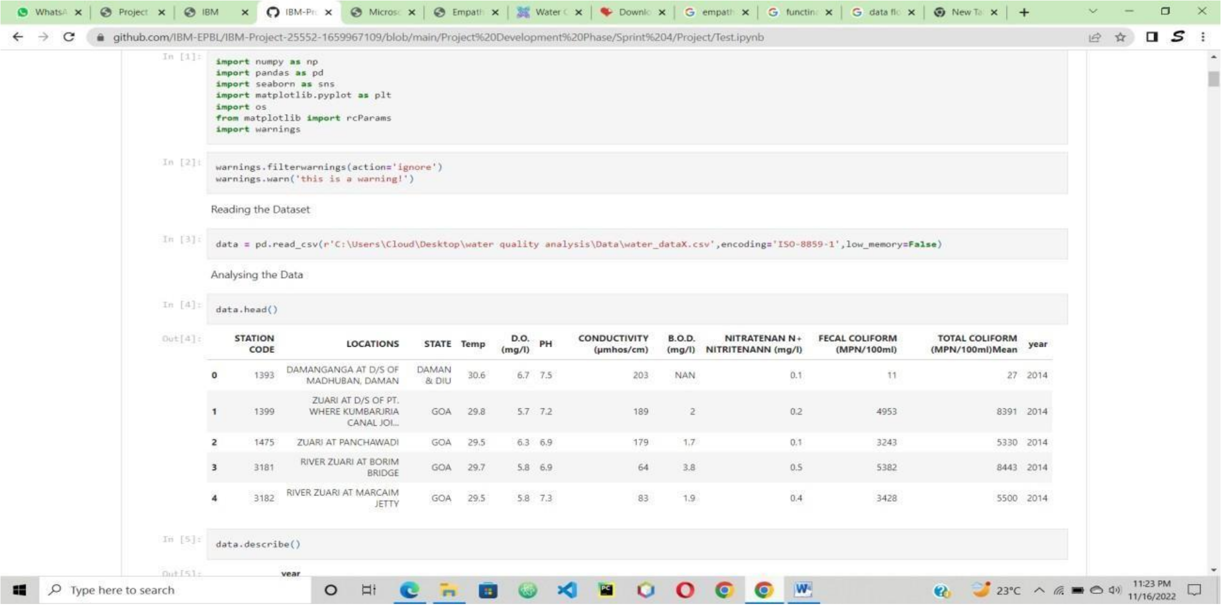
APP.PY



github.com/IBM-EPBL/IBM-Project-25552-1659967109/blob/main/Project%20Development%20Phase/Sprint%204/Project/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



github.com/IBM-EPBL/IBM-Project-25552-1659967109/blob/main/Project%20Development%20Phase/Sprint%204/Project/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()
```

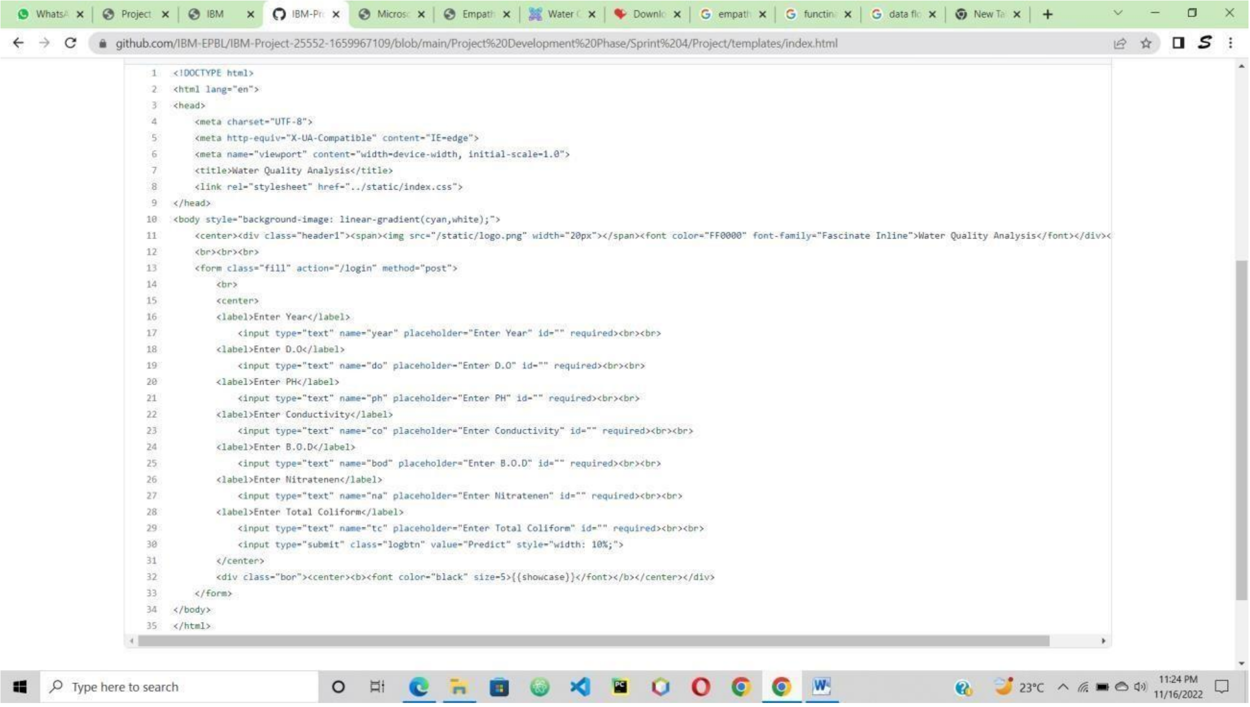
Out[4]:

	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

INDEX.HTML



github.com/IBM-EPBL/IBM-Project-25552-1659967109/blob/main/Project%20Development%20Phase/Sprint%204/Project/templates/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-34779-1660276564/blob/main/Project%20Development%20Phase/User%20Acceptance%20Testing/PNT2022TMID2375%20UAT%20Report.pdf