EFFICIENT WATER QUALITY ANALYSIS & PREDICTION USING MACHINE LEARNING

A PROJECT SUBMISSION BY

DHARSHINI K

DIVYA DHARSHINI P

HARISHMA K

HESINI PRIYA R

ABSTRACT:

Water is one of the vital needs of all living beings. Humans need water in many daily activities like drinking, washing, bathing, cooking etc. If the quality of water is not good then it becomes unfit for drinking and other activities. The quality of water usually described according to its physical, chemical and biological characteristics. Hence it becomes necessary to find the suitability of water for drinking, irrigation and Industry purpose. The groundwater quality based on Sodium percent, Sodium Absorption Ratio and Residual Sodium Carbonate will help to identify the suitability of water for irrigation purpose. Rapid industrialization and use of chemical fertilizers and pesticides in agriculture are causing deterioration of water quality and depletion of aquatic biota. Due to use of contaminated water, human population suffers from water borne diseases. Parameters that may be tested include temperature, pH, turbidity, salinity, nitrates, TDS, Cations, Anions and phosphates.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Project Overview

Ground water, surface water (rivers, streams and ponds), atmospheric water (rain-water, snow and hail) and springs are the main source of water available to the people in general. The qualities of these water bodies vary widely depending on the location and environmental factors. The major source of ground water is precipitation that infiltrates the ground and moves through the soil and pore spaces of rocks. Other sources include water infiltrating from lakes and streams, recharge ponds and waste-water treatment system. As ground water moves through soil, sediment and rocks, many impurities such as disease-causing microorganisms are filtered out. Many water resources in developing countries are unhealthy because they contain harmful physical, chemical and biological agents. To maintain a good health however, water should be safe to drink and meet the local standards and international standards to taste, odour and appearance. To monitor the water resource and ensure sustainability, national and international criteria and guidelines established for water quality standards are being used.(WHO-1993; 2005).The chemistry of water is very dynamic, largely controlled and modified by its medium of contact. Since the chemistry of water directly hints the quality of water for various purposes, its monitoring and assessment gained substantial importance in the present century.

A tremendous increase in the population increased the stress on both surface and the groundwater. It is believed at the beginning of the human civilization itself, groundwater was the most trusted form of drinking water because of the filtering effect of the aquifer. However, in the present world drinking the water directly from the source without proper treatment is a tough task. The groundwater analysis for physical and chemical properties is very important for public health studies. These studies are also main part of pollution studies in the environment. The groundwater contains dissolved solids possesses physical characteristics such as odour, taste and temperature. The natural quality of groundwater depends upon the physical environment, the origin, and the movement of water. As the water moves through the hydrological cycle, various chemical, physical and biological processes change its original quality through reactions with soil, rock and organic matter. Natural processes and human activities cause the changes in groundwater quality, directly or indirectly. According to WHO organization, about 80% of all the diseases in human beings are caused by water.

1.2 Purpose

As ground water moves through soil, sediment and rocks, many impurities such as disease-causing microorganisms are filtered out. Many water resources in developing countries are unhealthy because they contain harmful physical, chemical and biological

agents. To maintain a good health however, water should be safe to drink and meet the local standards and international standards to taste, odour and appearance. To monitor the water resource and ensure sustainability, national and international criteria and guidelines established for water quality standards are being used.(WHO-1993; 2005).The chemistry of water is very dynamic, largely controlled and modified by its medium of contact. Since the chemistry of water directly hints the quality of water for various purposes, its monitoring and assessment gained substantial importance in the present century.

2 Literature Survey

2.1 Existing Problem

The extensive literature review was carried out by referring standard journals, reference books and conference proceedings. The major work carried out by different researchers is summarized below. Dinesh Kumar Tank et.al [01] study focused on the hydrochemistry of groundwater in the Jaipur city to assess the quality of groundwater for determining its suitability for drinking and agricultural purposes. Groundwater samples were collected from eleven stations of Jaipur city during monsoon season and were analyzed for physico-chemical parameters such as pH, EC, TDS, sodium, potassium, calcium, magnesium, chloride, sulphate, carbonate, bicarbonate, nitrate and fluoride. Comparison of the concentration of the chemical constituents with WHO (world health organization) drinking water standards of 1983, the status of groundwater is better for drinking purposes. The parameters like pH, sodium, potassium, carbonate, bicarbonate, chloride are within permissible limit as per WHO but calcium, magnesium and nitrate values exceeding the limit. The calculated values of SAR, RSC and percentage sodium indicate that the water for irrigation uses is excellent to good quality. US Salinity diagram was used for evaluating the water quality for irrigation which suggests that the majority of the groundwater samples were good for irrigation.

Vikas Tomar et.al [02] collected water samples from 67 locations during pre and post-monsoon seasons of the year 2011 from Karnal district, Haryana and were subjected to analysis for chemical characteristics. The type of water that predominates in the study area was of sodium-calcium bicarbonate and magnesium bicarbonate type during pre and post-monsoon seasons of the year 2011 respectively and based on hydro-chemical facies.Based on chemical analysis, the pre and post monsoon water samples were classified as per different standard irrigation criteria to study the chemical changes resulting due to rain and natural recharge. It indicates that Na-Ca-HCO3 type water dominates during pre monsoon and Mg-HCO3 during post monsoon seasons of the year 2011.

Chidanand Patil et.al [03] carried out Physical, chemical, bacteriological analysis of water samples from seven bore wells located around landfill site at Turmuri, Belgaum to ascertain the magnitude of dumpsite pollution on groundwater quality. During the study period,7 bore wells were selected around the landfill area at a distance of 500, 750 and 1000m.The parameters analyzed during the study period were pH, Total dissolved

solids(TDS),Total Hardness, Nitrate, Most Probable Number (MPN) and heavy metal such as Lead using standard laboratory procedures. The pH ranged from 6.01 to 7.3 indicating acidic in nature in the month of Feb and March, but in the month of April and may all the wells within the levels. The pHs of water in wells within 500-700m are contaminated by the leachate of landfill. Concentrations of Hardness, TDS, Nitrate ranged from 0 to 80 mg/L, 49 to 190 mg/L, 4 to 79.89 mg/L respectively. The analysis was done for four months from Feb to May. The results showed that within 500 m bore wells were contaminated by E-Coli bacteria, also nitrate concentration is above the permissible level described by WHO and Bureau of Indian Standards for drinking water and pH were acidic in nature. The polluted water requires certain levels of treatment before use. Public enlightenment on waste sorting, adoption of clean technology, using climate change mitigation strategies and the use of sanitary landfill to prevent further contamination of ground water flow are recommended.

Sarala C.et.al [04] studied the groundwater quality parameters in the surrounding wells of Jawaharnagar, in upper Musi catchment

area of Ranga Reddy district in Andhra Pradesh. The bore wells data was collected from the study area for two seasons i.e., post monsoon in December 2007and pre monsoon in June 2008. The groundwater is acidic in nature and very hard. It is done by using Arc GIS software. The study reveals that the concentrations of major constituents are well within the permissible limits of IS10500-1994, except in few cases where total hardness and fluoride concentrations are high. The fluoride conc. exceeded the permissible limit. From the analysis it was observed that the groundwater is polluted in the entire study area. During last few years, the utilization of surface and groundwater for drinking, industrial and agricultural purposes has increased manifolds but consequently it is observed that the water is polluted and affecting the human health, soil nutrients,livestock,biomass and environment in certain areas.

Priti Singh et.al [05] he assess and map the spatial distribution of ground water quality of the Dhankawadi ward,pune by using GIS.APHA's standard laboratory procedure has been adopted to assess the quality of ground water. The spatial distribution map of pH, Chlorides, Magnesium and Sulphate shows that, these parameters are within range as per standard.TDS and Nitrate concentrations in ground water of the study area exceed the permissible limit at central location at katraj dairy near katraj,Pune.

People can use the ground water for drinking and domestic

purpose in study area except in upper katraj nagar, Pune. Priyanka Pandey et.al [06] he analysis the physiochemical properties of ground water near munciple solid waste dumping sites in Jabalpur. All the samples were collected from bore well and hand pump near the MSW dumping sites and stored at 40C. The temp. of ground water sample ranged from 25.11 to 27.310C.The study is carried out on parameters which are selected for testing are pH,TSS,TDS,COD,Nitrate,Cl,PO4 - ,Fetc.The parameters for both type water are within permissible limit for the use except TDS,TSS,TS. Sarala C.et.al [04] studied the groundwater quality parameters in the surrounding wells of Jawaharnagar, in upper Musi catchment area of Ranga Reddy district in Andhra Pradesh. The bore wells data was collected from the study area for two seasons i.e., post monsoon in December 2007and pre monsoon in June 2008. The groundwater is acidic in nature and very hard. It is done by using Arc GIS software. The study reveals that the concentrations of major constituents are well within the permissible limits of IS10500-1994, except in few cases where total hardness and fluoride concentrations are high. The fluoride conc. exceeded the permissible limit. From the analysis it was observed that the groundwater is polluted in the entire study area. During last few years, the utilization of surface and groundwater for drinking, industrial and agricultural purposes has increased manifolds but consequently it is observed that the water is polluted and affecting the human health, soil nutrients,livestock,biomass and environment in certain areas.

Priti Singh et.al [05] he assess and map the spatial distribution of

ground water quality of the Dhankawadi ward,pune by using GIS.APHA's standard laboratory procedure has been adopted to assess the quality of ground water. The spatial distribution map of pH, Chlorides, Magnesium and Sulphate shows that, these parameters are within range as per standard.TDS and Nitrate concentrations in ground water of the study area exceed the permissible limit at central location at katraj dairy near katraj,Pune.

People can use the ground water for drinking and domestic purpose in study area except in upper katraj nagar, Pune. Priyanka Pandey et.al [06] he analysis the physiochemical properties of ground water near munciple solid waste dumping sites in Jabalpur. All the samples were collected from bore well and hand pump near the MSW dumping sites and stored at 40C. The temp. of ground water sample ranged from 25.11 to 27.310C.The study is carried out on parameters which are selected for testing are pH,TSS,TDS,COD,Nitrate,Cl,PO4 - ,Fetc.

2.2 References

[1] C.Sadashivaia1,C.R.Ramakrishnaiah and G. Ranganna, "Hydrochemical Analysis and Evaluation of Groundwater Quality in TumkurTaluk, Karnataka State, India, International Journal ofEnvironmental Research and Public Health, 2008, 5(3) 158-164.

[2] Adetunde L.A, Glover R.L.K &Oguntola G.O, “assessment of the ground water quality in ogbomoso township of oyo state of nigeria”, IJRRAS8 (1) july 2011, 115-122.

[3] Shimaa M. Ghoraba&A.D.Khan, “ hydrochemistry and groundwater quality assessment in balochistan province, Pakistan”, IJRRAS 17 (2) November 2013, 185-199

[4] M. R. G. Sayyed1, G. S. Wagh2, A. Supekar3, “Assessment of impact on the groundwater quality due to urbanization by hydrogeochemicalfacies analysis in SE part of Pune city, India”, Proceedings of the International Academy of Ecology and Environmental Sciences, 2013, 3(2): 148-15.

[5] Dinesh kumar tank and c. p. Singh chandel, “Analysis of the major ion constituents in groundwater of Jaipur city”, Nature and Science, 2010;8(10), 1-7

[6] VikasTomar, Kamra S.K, Kumar S, Kumar Ajay and Vishal Khajuria, “Hydro-chemical analysis and evaluation of groundwater quality for irrigation in Karnal district ofHaryana state, India”, International Journal of Environmental Sciences, Volume 3, No 2, 2012, pp.756-766.

[7] Mona A. Hagras Assistant Professor, Irrigation &Hydraulics Department, Faculty of Engineering, Ain Shams University, Cairo, Egypt

[8] S. Prabaharan, R. Manonmani, M. Ramalingam and T. Subramani, “Groundwater Contamination due to Municipal Solid Waste Disposal in Salem City using GIS”, International Journal of Earth Science and Engineering, Volume 05, No. 04, pp- 696-702.

[10]ManeT.T.andHinganeHemalataN.“Existing Situation of Solid Waste Management in Pune City, India”, Research Journal of Recent Sciences , Vol. 1, 2012, pp.348-351

[11]S.S. Castaneda , R.J. Sucgang , R.V. Almoneda , N.D.S. Mendoza and C.P.C. David, “Environmental isotopes and major ions for tracing leachate contamination from a municipal landfill in Metro Manila, Philippines”, Journal of Environmental Radioactivity 110 (2012), pp.30-37.

[12]GunjanBhalla, Swamee, P.K, Arvind Kumar, Ajay Bansal , “Assessment of groundwater quality near municipal solid waste landfill byan Aggregate Index Method”, International Journal of Environmental Sciences Volume 2, No 2, 2012, pp. 1492- 1503.

[13]P.I. Agber, A. Ali and N. A. Tsaku, “Assessment of Ground Water Quality, Soil Propertiesand Nutrient Content of Soil in Areas Close to MunicipalRefuse Dump Sites in Makurdi, Nigeria”, J. Biol. Chem. Research. Vol. 30, No. 1, 2013, pp.88-97.

[14]Sarala C, Ravi Babu P, “ Assessment of Groundwater Quality

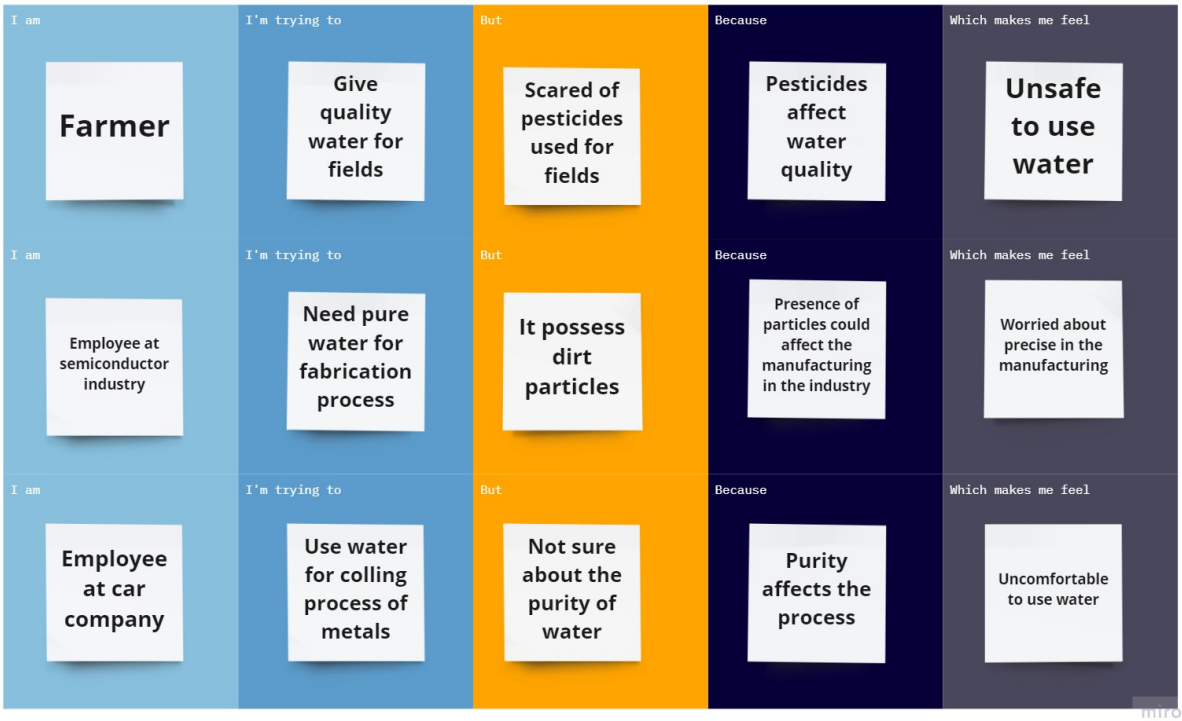
Parameters in and around Jawaharnagar, Hyderabad, International Journal of Scientific and Research Publications, Volume 2, Issue 10, 2012, pp.1- 6. [15]N Shaheda Niloufer, A.V. V. S. Swamy and M K. Syamala Devi, “Impact of Municipal Solid Waste on the Ground Water Quality in Vijayawada City, Andhra Pradesh”, Indian Journal of Applied research, Volume 3, April 2013, pp.62-642.

[16]Chidanand Patil, Shreekant Narayanakar and ArjunVirupakshi, “Assessment of Groundwater Quality Around Solid Waste Landfill Area - A Case Study, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 2, Issue 7, July 2013,pp. 3131-3136.

2.3 Problem Statement Definition

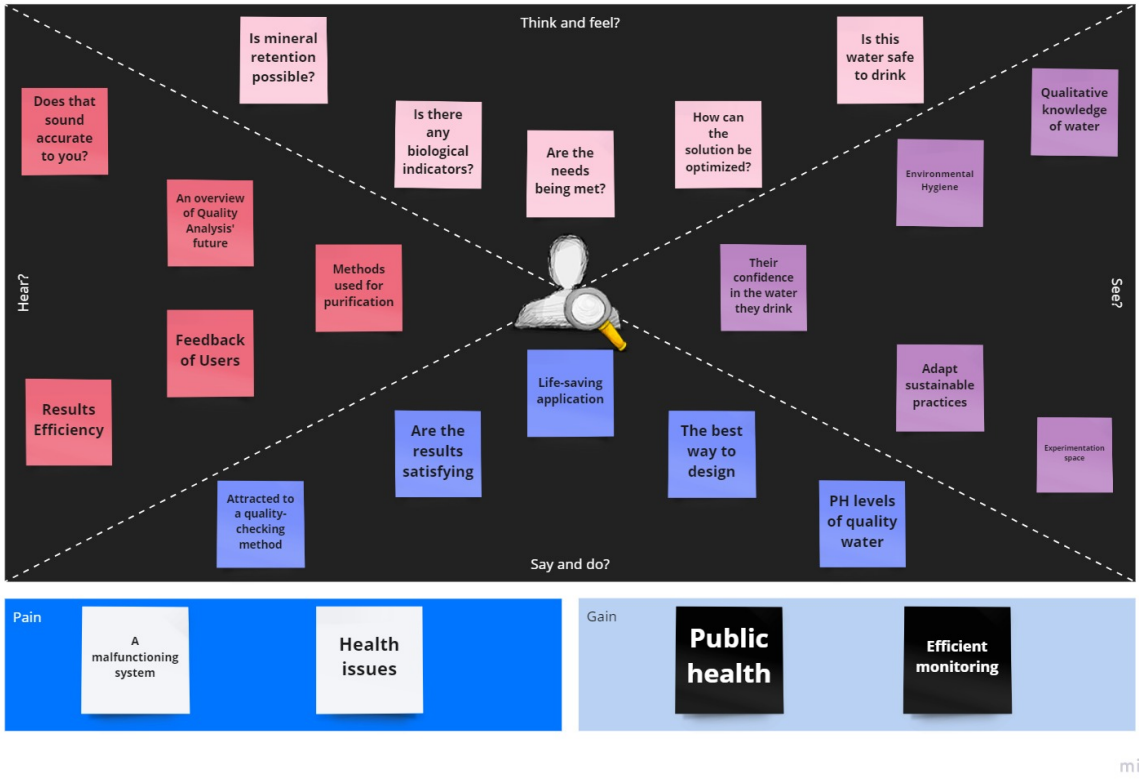
Water is considered as a vital resource that affects various aspects of human health and lives. The quality of water is a major concern for people living in urban areas. The quality of water serves as a powerful environmental determinant and a foundation for the prevention and control of waterborne diseases. 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.

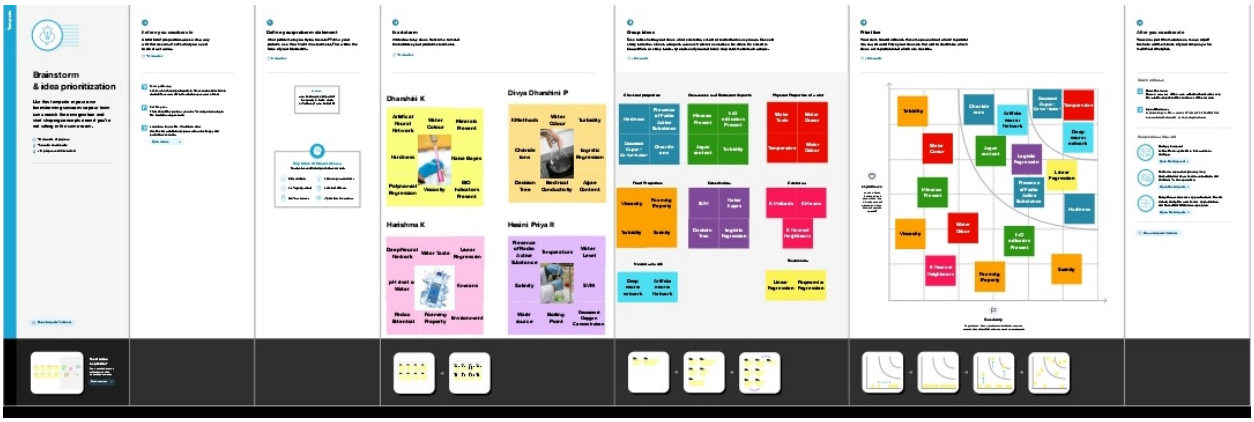


3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas



3.2 Ideation & Brainstorming



3.3 Proposed Solution

S.NO	Parameter	Description
1.	Problem Statement (Problem to be solved)	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.
2.	Idea / Solution	The solution is

	description	derived from the data sets by comparing the accuracy rate with previous data set and the current data set.
3.	Novelty / Uniqueness	Can be determined whether the water can be recycled or reused. User Friendly and Eco friendly.
4.	Social Impact / Customer Satisfaction	Beneficial for people health. By analysing the quality of water, good and healthy water is provided.
5.	Business Model (Revenue Model)	First the application is tested with few people.Later on it comes into the picture where everyone can see by networking. By conducting various activities regarding

		the importance of quality of water.
6.	Scalability of the Solution	Helps in getting all required aspects of water.

3.4 Problem Solution fit

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.

4.2 Non-Functional requirements

NFR 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

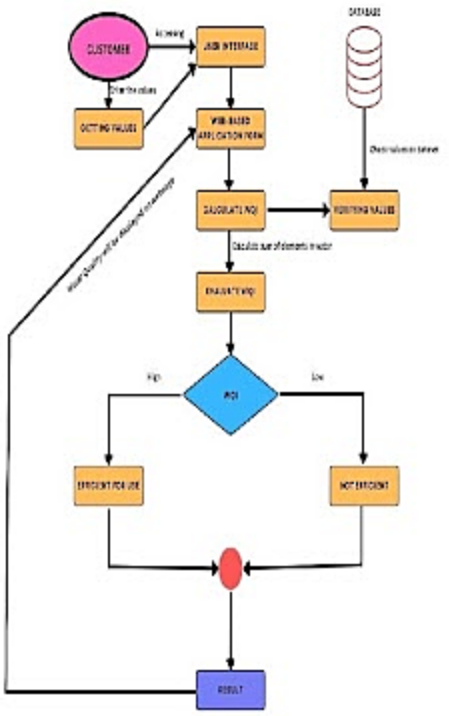
		and has the ability to increase or decrease the performance of the system based on the datasets.

5. PROJECT DESIGN

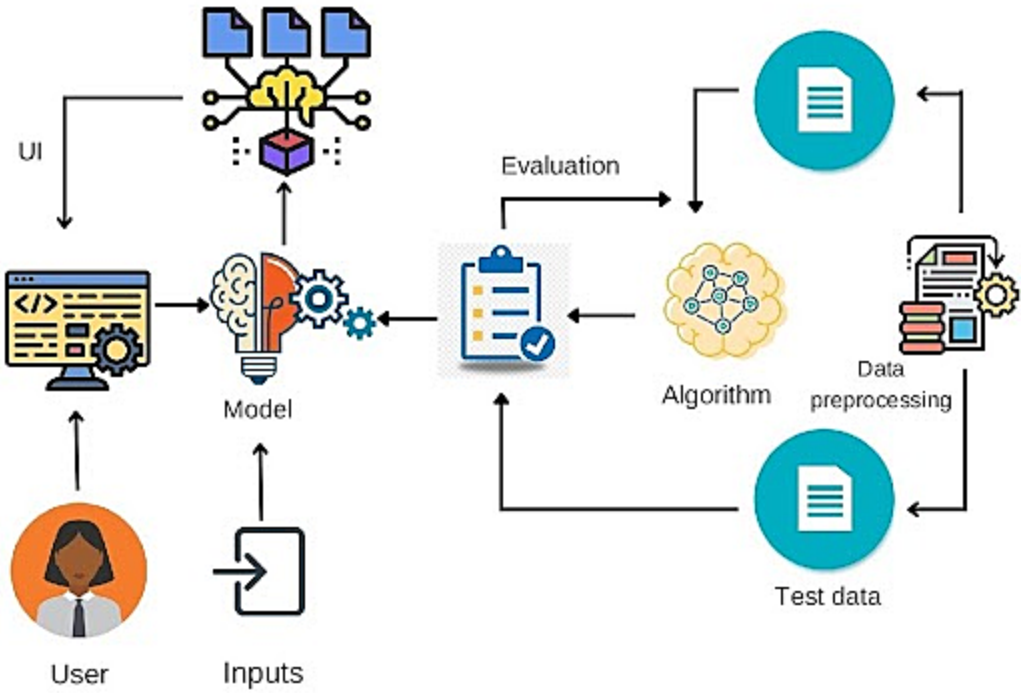
5.1 Data Flow Diagrams

DFD LEVEL 0

DFD LEVEL 1



5.2 Solution & Technical Architecture



5.3 User Stories

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Web user)	Access Web page	USN-1	As a user, anyone can access the web page to check the water quality.	I can access my webpage through online at any time.	High	Sprint-1
Customer	Usage of	USN-2	As per the	Prediction	High	Sprint-2

(Web user)	water		usage of user, the quality of water should be predicted in easy way.	can be done in easy way.		
Customer (Web user)	Accuracy of water	USN-3	By using the prediction model the user will know the quality of water on a daily basis.	The quality analysis of water will be accurate.	High	Sprint-3
Administrator	Manage the web page	USN-4	As an admin, he/she can manage user details and update parameters essential for prediction.	Make changes on User Interface (UI).	High	Sprint-4
Administrator	Calculation of WQI	USN-5	As an admin, he/she can update the calculations for water quality index.	Improves the accuracy of the calculation.	High	Sprint-5

6. PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority
Sprint-1	Data Collection	USN-1	Collecting dataset for pre-processing	2	High
Sprint-1		USN-2	Data pre-processing used to transform the data into useful format.	1	High
Sprint-2	Model Building	USN-3	Calculate the Water Quality Index (WQI) using Regression algorithm of machine learning	2	Low
Sprint-2		USN-4	Splitting the data into training and testing from	2	Medium

			the entire dataset.		
Sprint-3	Application Building	USN-5	Implementing the web page for collecting the data from user	1	High
Sprint-4	Training and Testing	USN-6	Training the model using regression algorithm and testing the performance of the model	2	High
Sprint-4		USN-7	Deploying the model using IBM Cloud and IBM Watson Studio	2	Medium

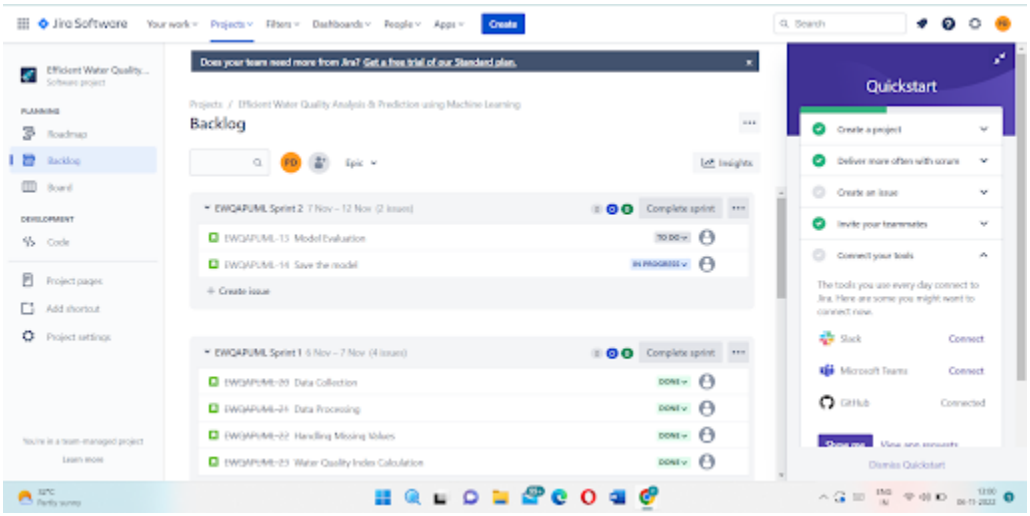
6.2 Sprint Delivery Schedule

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on	Sprint Release Date (Actual)

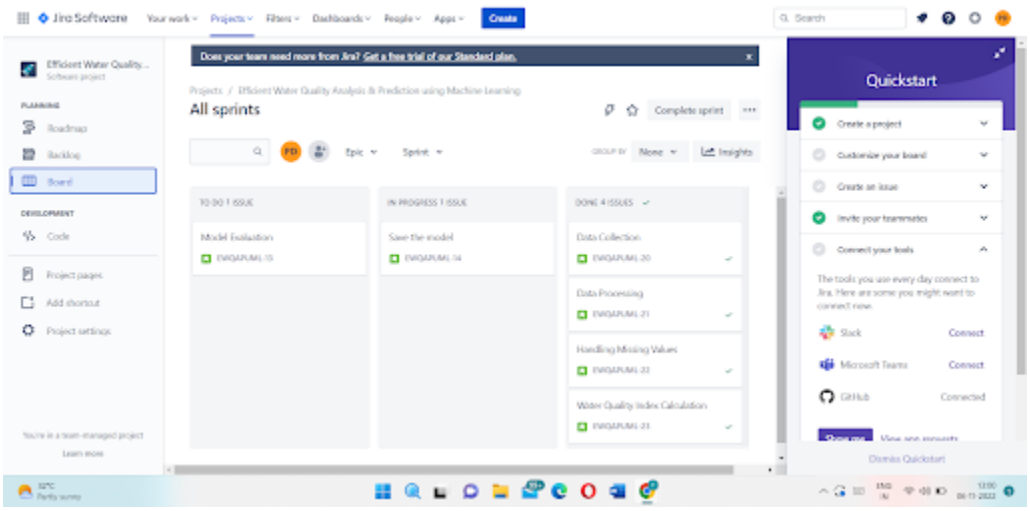
					Planned End Date)	
Sprint-1	20	6 days	24 Oct 2022	29 Oct 2022	20	29 Oct 2022
Sprint-2	20	6 days	31 Oct 2022	5 Nov 2022	20	5 Nov 2022
Sprint-3	20	6 days	07 Nov 2022	12 Nov 2022	20	12 Nov 2022
Sprint-4	20	6 days	14 Nov 2022	19 Nov 2022	20	19 Nov 2022

6.3 Reports from JIRA

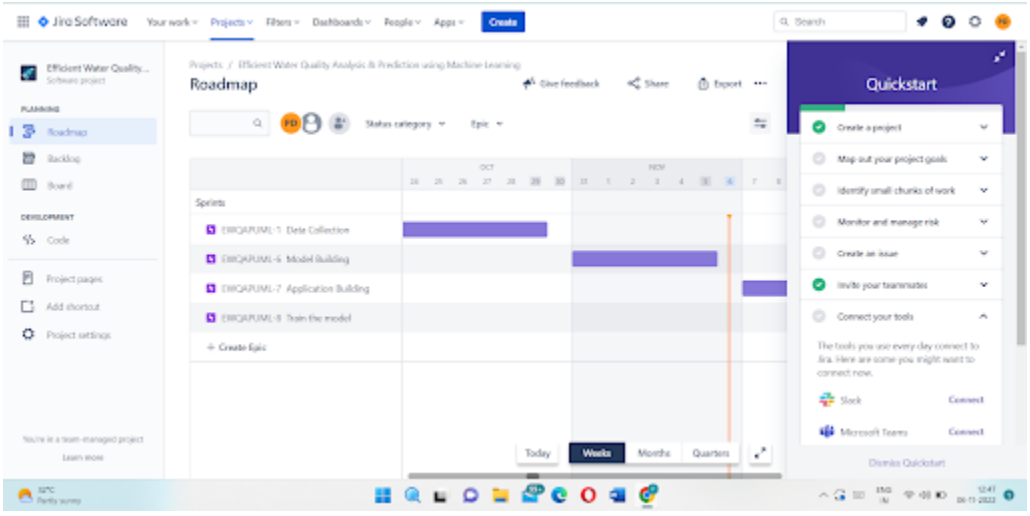
Backlog



Board


Jira Software
All sprints
Quickstart

Roadmap


Jira Software
Roadmap
Quickstart

7. CODING & SOLUTIONING

7.1 Front-end

web.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/css/style.css">
<style>
body {
   background: url('../static/css/water.jpg');
   background-repeat: no-repeat;
   background-size:cover;
}
</style>
</head>

<body>
   <header>
      <nav>
         <div class="row">
            <div class="row1">
               <img src="../static/css/img.png" alt="logo">
            </div>
            <div class="row2">
               <h1 style="color:black"> Water Quality Prediction</h1>
            </div>
         </div>
      </nav>
   </header>
```

```
  <main>
    <div class="column">
      <form action="/login" method="post">
        <label for=""></label>
        <input type="text" name="year" id="" placeholder="Enter Year">
        <label for=""></label>
        <input type="text" name="do" id="" placeholder="Enter D.O">
        <label for=""></label>
        <input type="text" name="ph" id="" placeholder="Enter PH">
        <label for=""></label>
        <input type="text" name="co" id="" placeholder="Enter Conductivity">
        <label for=""></label>
        <input type="text" name="bod" id="" placeholder="Enter B.O.D">
        <label for=""></label>
        <input type="text" name="na" id="" placeholder="Enter Nitratenen">
        <label for=""></label>
        <input type="text" name="tc" id="" placeholder="Enter Total Coliform">
        <label for=""></label>
        <div class="last">
          <input type="submit" value="Predict">
        </div>
        <div class="bor" style="color:black"><b>
          {{showcase}}</b>
        </div>

      </form>
    </div>
  </main>
  </div>
</body>

</html>
```

style.css

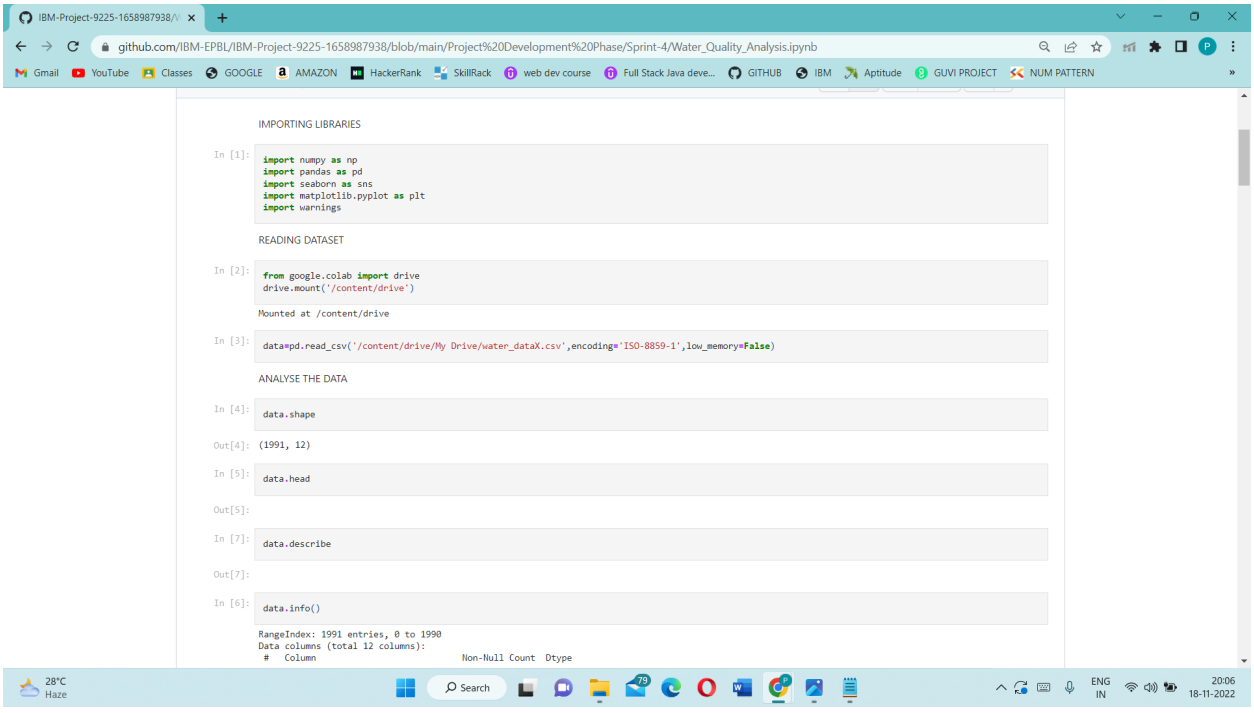
```
* {
  margin: 0;
  padding: 0;
  box-sizing: border-box;
```

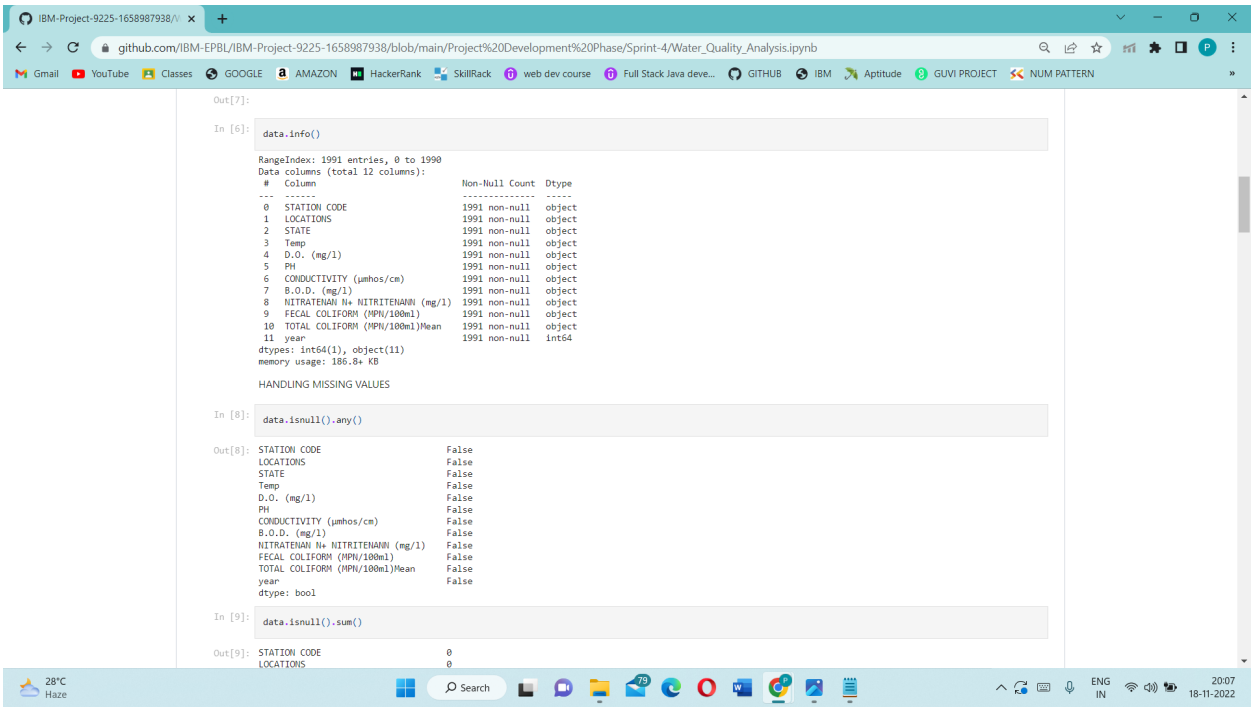
```
}

/* styling the header */

.row1 img {
    height: 70px;
    position: relative;
    left: 54vw;
}

.row2 h1 {
    position: absolute;
    left: 60vw;
    color: #ffff;
    top: 30px;
}

main div.column input {
    display: block;
    position: relative;
    margin: 24px 26px;
    left: 60vw;
    border-radius: 35px;
    width: 250px;
    height: 30px;
}

input[type="text"] {
    text-align: center;
    font-family: 'Courier New', Courier, monospace;
}

main div.last input {
    width: 256px;;
    background: rgb(5, 3, 68);
     background: linear-gradient(90deg, rgba(5, 3, 68, 1) 0%, rgba(0, 12, 36, 1) 0%, rgba(39, 9, 121,
1) 0%, rgba(38, 8, 114, 1) 0%, rgba(94, 0, 255, 1) 45%, rgba(188, 0, 255, 1) 84%);
    font-weight: 600;
    font-family: 'Courier New', Courier, monospace;
}
```

```
main div.last input:hover {
    background: rgb(188, 0, 255);
     background: linear-gradient(90deg, rgba(188, 0, 255, 1) 23%, rgba(94, 0, 255, 1) 63%, rgba(22, 9, 121, 1) 100%, rgba(5, 3, 68, 1) 100%, rgba(0, 12, 36, 1) 100%, rgba(38, 8, 114, 1) 100%);
}

.bor {
    text-align: center;
    margin-left: 60vw;
    color: white;
    font-size: 21px;
    border: 2px solid rgb(251, 253, 255);
    width: 358px;
    padding-left: 4px;
}
```

7.2 Back-end

Water_Quality_Analysis.ipynb





Out[7]:

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

```
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
 11  year                            1991 non-null   int64 
dtypes: int64(1), object(11)
memory usage: 186.8+ KB
```

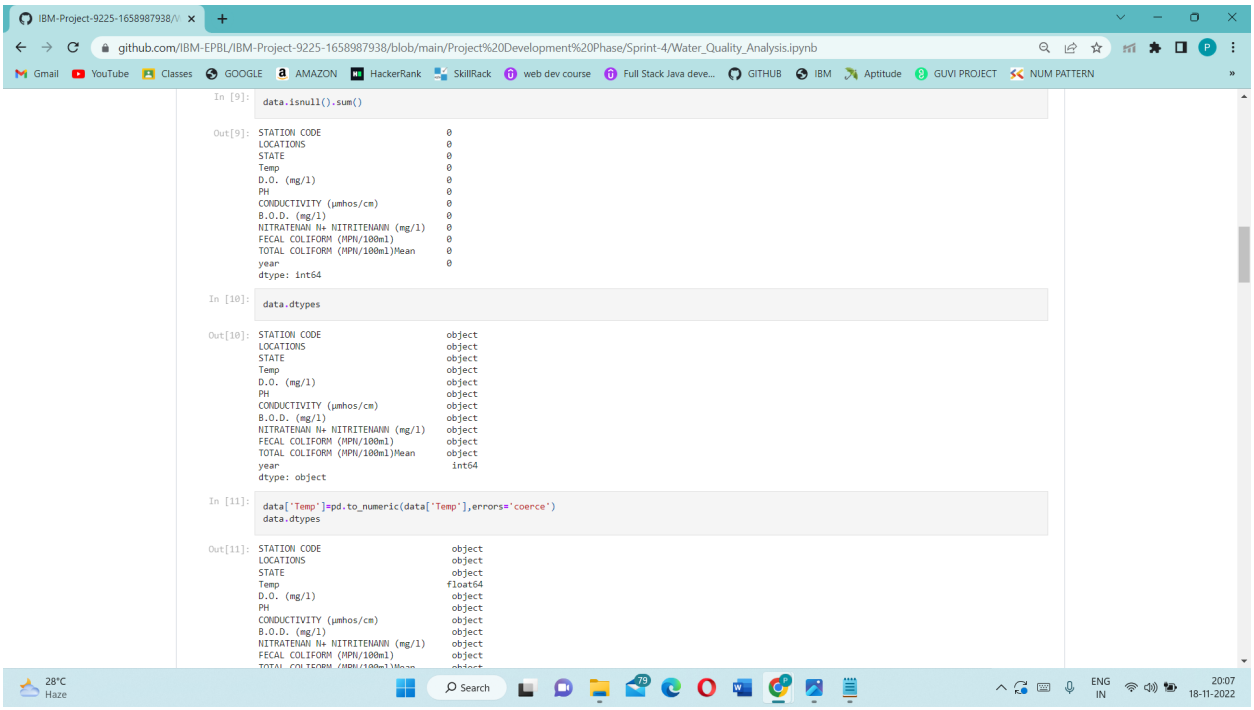
HANDLING MISSING VALUES

In [8]:
```
data.isnull().any()
```

Out[8]:
```
STATION CODE                    False
LOCATIONS                       False
STATE                           False
Temp                            False
D.O. (mg/l)                     False
PH                              False
CONDUCTIVITY (µmhos/cm)         False
B.O.D. (mg/l)                   False
NITRATENAN N+ NITRITENANN (mg/l)  False
FECAL COLIFORM (MPN/100ml)      False
TOTAL COLIFORM (MPN/100ml)Mean  False
year                            False
dtype: bool
```

In [9]:
```
data.isnull().sum()
```

Out[9]:
```
STATION CODE                    0
LOCATIONS                       0
```



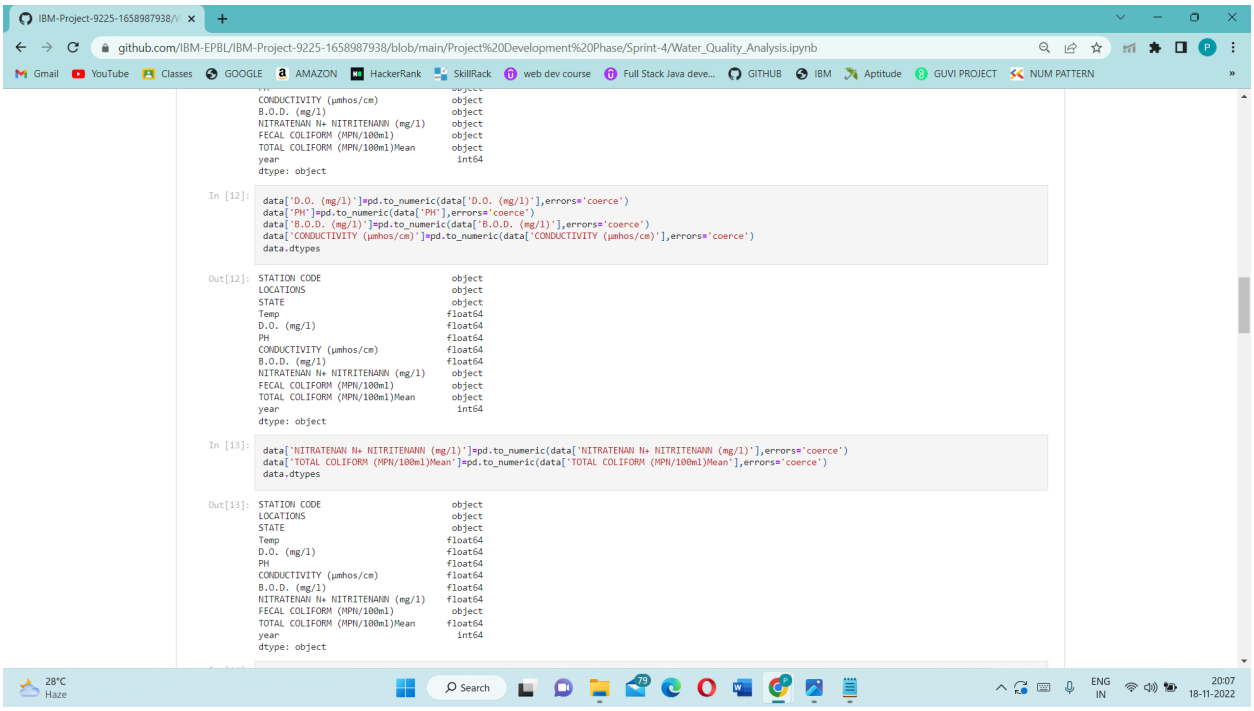
Out[9]:
```
STATION CODE                    0
LOCATIONS                       0
STATE                           0
Temp                            0
D.O. (mg/l)                     0
PH                              0
CONDUCTIVITY (µmhos/cm)         0
B.O.D. (mg/l)                   0
NITRATENAN N+ NITRITENANN (mg/l)  0
FECAL COLIFORM (MPN/100ml)      0
TOTAL COLIFORM (MPN/100ml)Mean  0
year                            0
dtype: int64
```

In [10]:
```
data.dtypes
```

Out[10]:
```
STATION CODE                    object
LOCATIONS                       object
STATE                           object
Temp                            object
D.O. (mg/l)                     object
PH                              object
CONDUCTIVITY (µmhos/cm)         object
B.O.D. (mg/l)                   object
NITRATENAN N+ NITRITENANN (mg/l)  object
FECAL COLIFORM (MPN/100ml)      object
TOTAL COLIFORM (MPN/100ml)Mean  object
year                             int64
dtype: object
```

In [11]:
```
data['Temp']=pd.to_numeric(data['Temp'],errors='coerce')
data.dtypes
```

Out[11]:
```
STATION CODE                     object
LOCATIONS                        object
STATE                            object
Temp                            float64
D.O. (mg/l)                      object
PH                               object
CONDUCTIVITY (µmhos/cm)          object
B.O.D. (mg/l)                    object
NITRATENAN N+ NITRITENANN (mg/l)  object
FECAL COLIFORM (MPN/100ml)       object
```



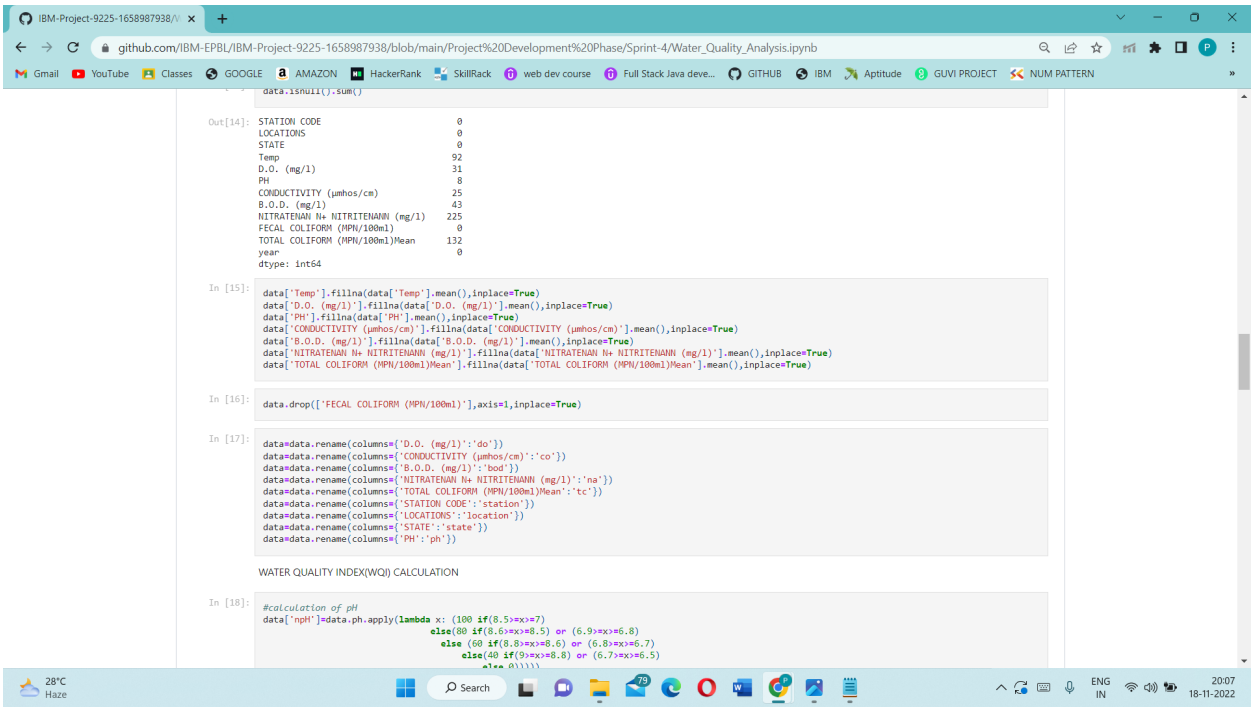
```
CONDUCTIVITY (µmhos/cm)          object
B.O.D. (mg/l)                    object
NITRATENAN N+ NITRITENANN (mg/l) object
FECAL COLIFORM (MPN/100ml)       object
TOTAL COLIFORM (MPN/100ml)Mean   object
year                              int64
dtype: object
```

In [12]:
```
data['D.O. (mg/l)']=pd.to_numeric(data['D.O. (mg/l)'],errors='coerce')
data['PH']=pd.to_numeric(data['PH'],errors='coerce')
data['B.O.D. (mg/l)']=pd.to_numeric(data['B.O.D. (mg/l)'],errors='coerce')
data['CONDUCTIVITY (µmhos/cm)']=pd.to_numeric(data['CONDUCTIVITY (µmhos/cm)'],errors='coerce')
data.dtypes
```

Out[12]:
```
STATION CODE                     object
LOCATIONS                        object
STATE                            object
Temp                            float64
D.O. (mg/l)                     float64
PH                              float64
CONDUCTIVITY (µmhos/cm)         float64
B.O.D. (mg/l)                   float64
NITRATENAN N+ NITRITENANN (mg/l) object
FECAL COLIFORM (MPN/100ml)       object
TOTAL COLIFORM (MPN/100ml)Mean   object
year                              int64
dtype: object
```

In [13]:
```
data['NITRATENAN N+ NITRITENANN (mg/l)']=pd.to_numeric(data['NITRATENAN N+ NITRITENANN (mg/l)'],errors='coerce')
data['TOTAL COLIFORM (MPN/100ml)Mean']=pd.to_numeric(data['TOTAL COLIFORM (MPN/100ml)Mean'],errors='coerce')
data.dtypes
```

Out[13]:
```
STATION CODE                     object
LOCATIONS                        object
STATE                            object
Temp                            float64
D.O. (mg/l)                     float64
PH                              float64
CONDUCTIVITY (µmhos/cm)         float64
B.O.D. (mg/l)                   float64
NITRATENAN N+ NITRITENANN (mg/l) float64
FECAL COLIFORM (MPN/100ml)       object
TOTAL COLIFORM (MPN/100ml)Mean   float64
year                              int64
dtype: object
```



```
data.isnull().sum()
```

Out[14]:
```
STATION CODE                      0
LOCATIONS                         0
STATE                             0
Temp                             92
D.O. (mg/l)                      31
PH                                8
CONDUCTIVITY (µmhos/cm)          25
B.O.D. (mg/l)                    43
NITRATENAN N+ NITRITENANN (mg/l) 225
FECAL COLIFORM (MPN/100ml)        0
TOTAL COLIFORM (MPN/100ml)Mean   132
year                              0
dtype: int64
```

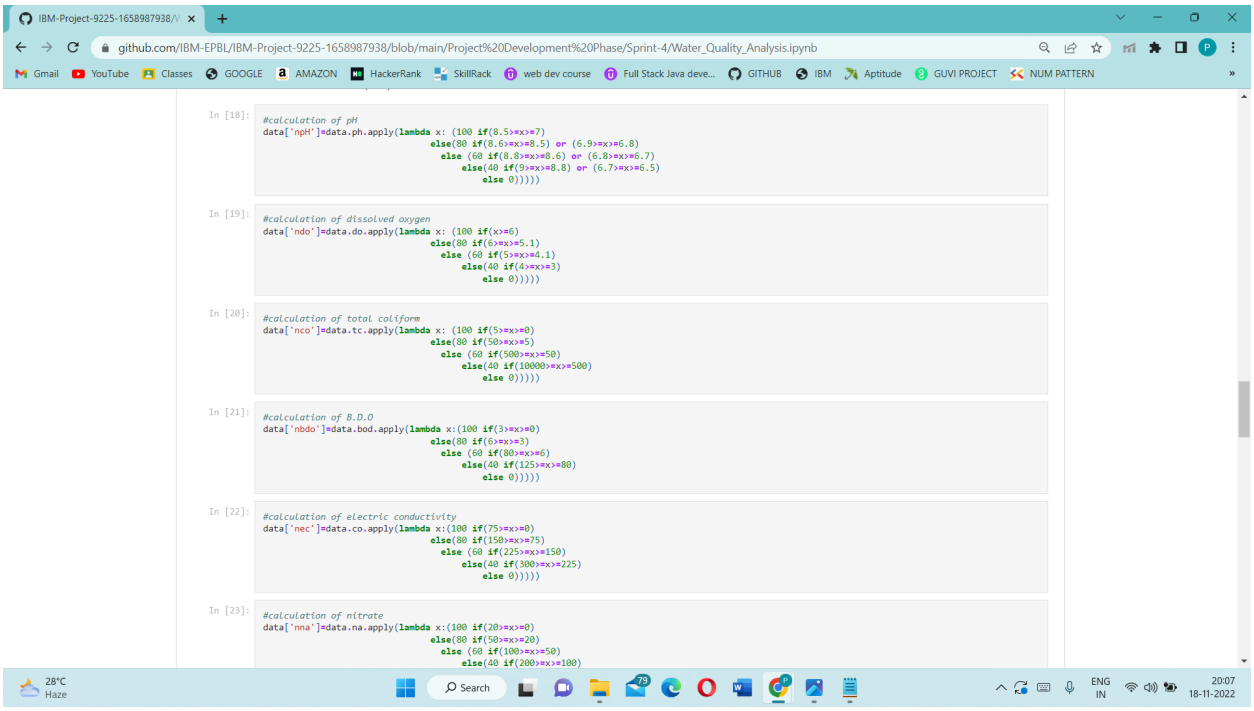
In [15]:
```
data['Temp'].fillna(data['Temp'].mean(),inplace=True)
data['D.O. (mg/l)'].fillna(data['D.O. (mg/l)'].mean(),inplace=True)
data['PH'].fillna(data['PH'].mean(),inplace=True)
data['CONDUCTIVITY (µmhos/cm)'].fillna(data['CONDUCTIVITY (µmhos/cm)'].mean(),inplace=True)
data['B.O.D. (mg/l)'].fillna(data['B.O.D. (mg/l)'].mean(),inplace=True)
data['NITRATENAN N+ NITRITENANN (mg/l)'].fillna(data['NITRATENAN N+ NITRITENANN (mg/l)'].mean(),inplace=True)
data['TOTAL COLIFORM (MPN/100ml)Mean'].fillna(data['TOTAL COLIFORM (MPN/100ml)Mean'].mean(),inplace=True)
```

In [16]:
```
data.drop(['FECAL COLIFORM (MPN/100ml)'],axis=1,inplace=True)
```

In [17]:
```
data=data.rename(columns={'D.O. (mg/l)':'do'})
data=data.rename(columns={'CONDUCTIVITY (µmhos/cm)':'co'})
data=data.rename(columns={'B.O.D. (mg/l)':'bod'})
data=data.rename(columns={'NITRATENAN N+ NITRITENANN (mg/l)':'na'})
data=data.rename(columns={'TOTAL COLIFORM (MPN/100ml)Mean':'tc'})
data=data.rename(columns={'STATION CODE':'station'})
data=data.rename(columns={'LOCATIONS':'location'})
data=data.rename(columns={'STATE':'state'})
data=data.rename(columns={'PH':'ph'})
```

WATER QUALITY INDEX(WQI) CALCULATION

In [18]:
```
#calculation of pH
data['npH']=data.ph.apply(lambda x: (100 if(8.5>=x>=7)
                                  else(80 if(8.6>=x>=8.5) or (6.9>=x>=6.8)
                                    else (60 if(8.8>=x>=8.6) or (6.8>=x>=6.7)
                                       else(40 if(9>=x>=8.8) or (6.7>=x>=6.5)
```



In [18]:
```
#calculation of pH
data['npH']=data.ph.apply(lambda x: (100 if(8.5>=x>=7)
                                  else(80 if(8.6>=x>=8.5) or (6.9>=x>=6.8)
                                    else (60 if(8.8>=x>=8.6) or (6.8>=x>=6.7)
                                        else(40 if(9>=x>=8.8) or (6.7>=x>=6.5)
                                            else 0)))))
```

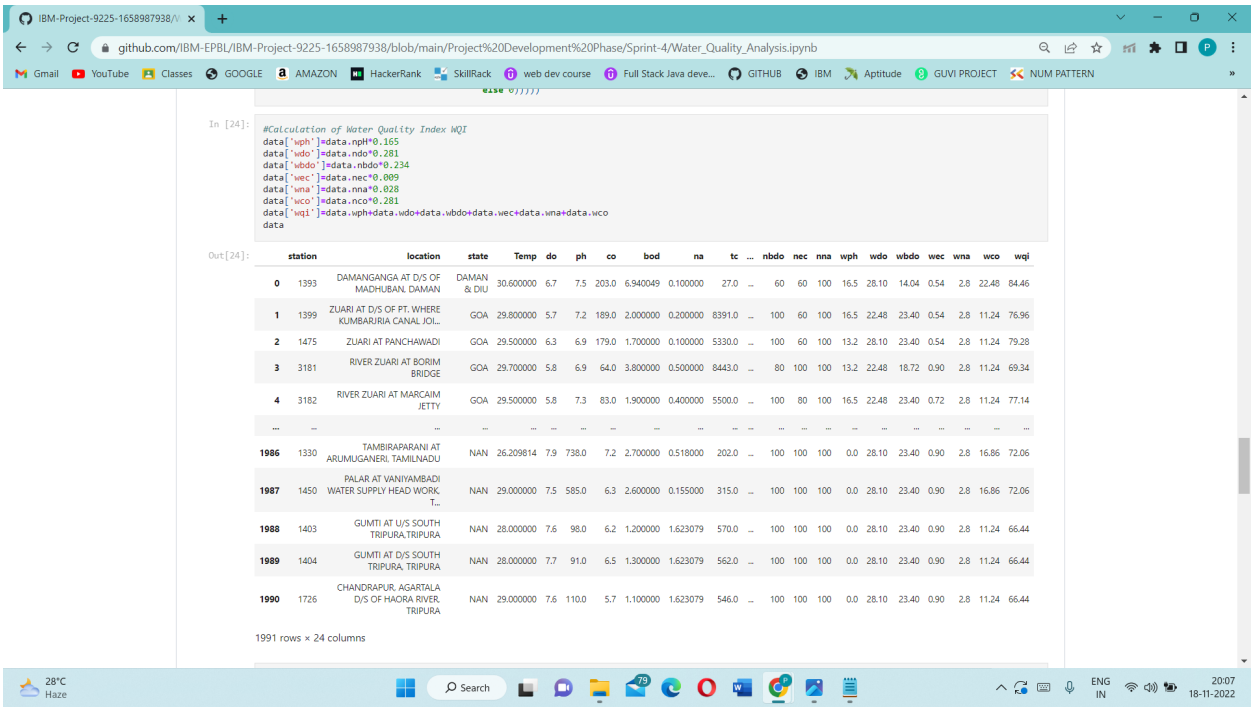
In [19]:
```
#calculation of dissolved oxygen
data['ndo']=data.do.apply(lambda x: (100 if(x>=6)
                                  else(80 if(6>=x>=5.1)
                                    else (60 if(5>=x>=4.1)
                                        else(40 if(4>=x>=3)
                                            else 0)))))
```

In [20]:
```
#calculation of total coliform
data['nco']=data.tc.apply(lambda x: (100 if(5>=x>=0)
                                  else(80 if(50>=x>=5)
                                    else (60 if(500>=x>=50)
                                        else(40 if(10000>=x>=500)
                                            else 0)))))
```

In [21]:
```
#calculation of B.D.O
data['nbdo']=data.bod.apply(lambda x:(100 if(3>=x>=0)
                                  else(80 if(6>=x>=3)
                                    else (60 if(80>=x>=6)
                                        else(40 if(125>=x>=80)
                                            else 0)))))
```

In [22]:
```
#calculation of electric conductivity
data['nec']=data.co.apply(lambda x:(100 if(75>=x>=0)
                                  else(80 if(150>=x>=75)
                                    else (60 if(225>=x>=150)
                                        else(40 if(300>=x>=225)
                                            else 0)))))
```

In [23]:
```
#calculation of nitrate
data['nna']=data.na.apply(lambda x:(100 if(20>=x>=0)
                                  else(80 if(50>=x>=20)
                                    else (60 if(100>=x>=50)
                                        else(40 if(200>=x>=100)
```



In [24]:
```
#Calculation of Water Quality Index WQI
data['wph']=data.npH*0.165
data['wdo']=data.ndo*0.281
data['wbdo']=data.nbdo*0.234
data['wec']=data.nec*0.009
data['wna']=data.nna*0.028
data['wco']=data.nco*0.281
data['wqi']=data.wph+data.wdo+data.wbdo+data.wec+data.wna+data.wco
data
```

Out[24]:

	station	location	state	Temp	do	ph	co	bod	na	tc	...	nbdo	nec	nna	wph	wdo	wbdo	wec	wna	wco	wqi
0	1393	DAMANGANGA AT D/S OF MADHUBAN, DAMAN	DAMAN & DIU	30.600000	6.7	7.5	203.0	6.940049	0.100000	27.0	...	60	60	100	16.5	28.10	14.04	0.54	2.8	22.48	84.46
1	1399	ZUARI AT D/S OF PT. WHERE KUMBARJRIA CANAL JOI...	GOA	29.800000	5.7	7.2	189.0	2.000000	0.200000	8391.0	...	100	60	100	16.5	22.48	23.40	0.54	2.8	11.24	76.96
2	1475	ZUARI AT PANCHAWADI	GOA	29.500000	6.3	6.9	179.0	1.700000	0.100000	5330.0	...	100	60	100	13.2	28.10	23.40	0.54	2.8	11.24	79.28
3	3181	RIVER ZUARI AT BORIM BRIDGE	GOA	29.700000	5.8	6.9	64.0	3.800000	0.500000	8443.0	...	80	100	100	13.2	22.48	18.72	0.90	2.8	11.24	69.34
4	3182	RIVER ZUARI AT MARCAIM JETTY	GOA	29.500000	5.8	7.3	83.0	1.900000	0.400000	5500.0	...	100	80	100	16.5	22.48	23.40	0.72	2.8	11.24	77.14
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1986	1330	TAMBIRAPARANI AT ARUMUGANERI, TAMILNADU	NAN	26.209814	7.9	738.0	7.2	2.700000	0.518000	202.0	...	100	100	100	0.0	28.10	23.40	0.90	2.8	16.86	72.06
1987	1450	PALAR AT VANIYAMBADI WATER SUPPLY HEAD WORK, T...	NAN	29.000000	7.5	585.0	6.3	2.600000	0.155000	315.0	...	100	100	100	0.0	28.10	23.40	0.90	2.8	16.86	72.06
1988	1403	GUMTI AT U/S SOUTH TRIPURA,TRIPURA	NAN	28.000000	7.6	98.0	6.2	1.200000	1.623079	570.0	...	100	100	100	0.0	28.10	23.40	0.90	2.8	11.24	66.44
1989	1404	GUMTI AT D/S SOUTH TRIPURA, TRIPURA	NAN	28.000000	7.7	91.0	6.5	1.300000	1.623079	562.0	...	100	100	100	0.0	28.10	23.40	0.90	2.8	11.24	66.44
1990	1726	CHANDRAPUR, AGARTALA D/S OF HAORA RIVER, TRIPURA	NAN	29.000000	7.6	110.0	5.7	1.100000	1.623079	546.0	...	100	100	100	0.0	28.10	23.40	0.90	2.8	11.24	66.44

1991 rows × 24 columns

IBM-Project-9225-1658987938/W

github.com/IBM-EPBL/IBM-Project-9225-1658987938/blob/main/Project%20Development%20Phase/Sprint-4/Water_Quality_Analysis.ipynb

Gmail YouTube Classes GOOGLE AMAZON HackerRank SkillRack web dev course Full Stack Java deve... GITHUB IBM Aptitude GUVI PROJECT NUM PATTERN

```
#Calculation of overall WQI for each year
average = data.groupby('year')['wqi'].mean()
average.head()
```

Out[25]:
```
year
2003    66.239545
2004    61.290000
2005    73.762689
2006    72.909714
2007    74.233000
Name: wqi, dtype: float64
```

SPLITTING DEPENDENT AND INDEPENDENT COLUMNS

In [26]:
```
data.head()
data.drop(['location','station','state'],axis =1,inplace=True)
```

In [27]:
```
data.head()
```

Out[27]:

	Temp	do	ph	co	bod	na	tc	year	npH	ndo	...	nbdo	nec	nna	wph	wdo	wbdo	wec	wna	wco	wqi
0	30.6	6.7	7.5	203.0	6.940049	0.1	27.0	2014	100	100	...	60	60	100	16.5	28.10	14.04	0.54	2.8	22.48	84.46
1	29.8	5.7	7.2	189.0	2.000000	0.2	8391.0	2014	100	80	...	100	60	100	16.5	22.48	23.40	0.54	2.8	11.24	76.96
2	29.5	6.3	6.9	179.0	1.700000	0.1	5330.0	2014	80	100	...	100	60	100	13.2	28.10	23.40	0.54	2.8	11.24	79.28
3	29.7	5.8	6.9	64.0	3.800000	0.5	8443.0	2014	80	80	...	80	100	100	13.2	22.48	18.72	0.90	2.8	11.24	69.34
4	29.5	5.8	7.3	83.0	1.900000	0.4	5500.0	2014	100	80	...	100	80	100	16.5	22.48	23.40	0.72	2.8	11.24	77.14

5 rows × 21 columns

In [28]:
```
x=data.iloc[:,0:7].values
x.shape
```

Out[28]: (1991, 7)

In [29]:
```
y=data.iloc[:,-1:].values
y.shape
```

Out[29]: (1991, 1)

28°C Haze Search ENG IN 20:07 18-11-2022

IBM-Project-9225-1658987938/W

github.com/IBM-EPBL/IBM-Project-9225-1658987938/blob/main/Project%20Development%20Phase/Sprint-4/Water_Quality_Analysis.ipynb

Gmail YouTube Classes GOOGLE AMAZON HackerRank SkillRack web dev course Full Stack Java deve... GITHUB IBM Aptitude GUVI PROJECT NUM PATTERN

In [30]:
```
print(x)
```

```
[[3.06000000e+01 6.70000000e+00 7.50000000e+00 ... 6.94004877e+00
  1.00000000e-01 2.70000000e+01]
 [2.98000000e+01 5.70000000e+00 7.20000000e+00 ... 2.00000000e+00
  2.00000000e-01 8.39100000e+03]
 [2.95000000e+01 6.30000000e+00 6.90000000e+00 ... 1.70000000e+00
  1.00000000e-01 5.33000000e+03]
 ...
 [2.80000000e+01 7.60000000e+00 9.80000000e+01 ... 1.20000000e+00
  1.62307871e+00 5.70000000e+02]
 [2.80000000e+01 7.70000000e+00 9.10000000e+01 ... 1.30000000e+00
  1.62307871e+00 5.62000000e+02]
 [2.90000000e+01 7.60000000e+00 1.10000000e+02 ... 1.10000000e+00
  1.62307871e+00 5.46000000e+02]]
```

In [31]:
```
print(y)
```

```
[[84.46]
 [76.96]
 [79.28]
 ...
 [66.44]
 [66.44]
 [66.44]]
```

SPLITTING THE DATA INTO TRAIN AND TEST

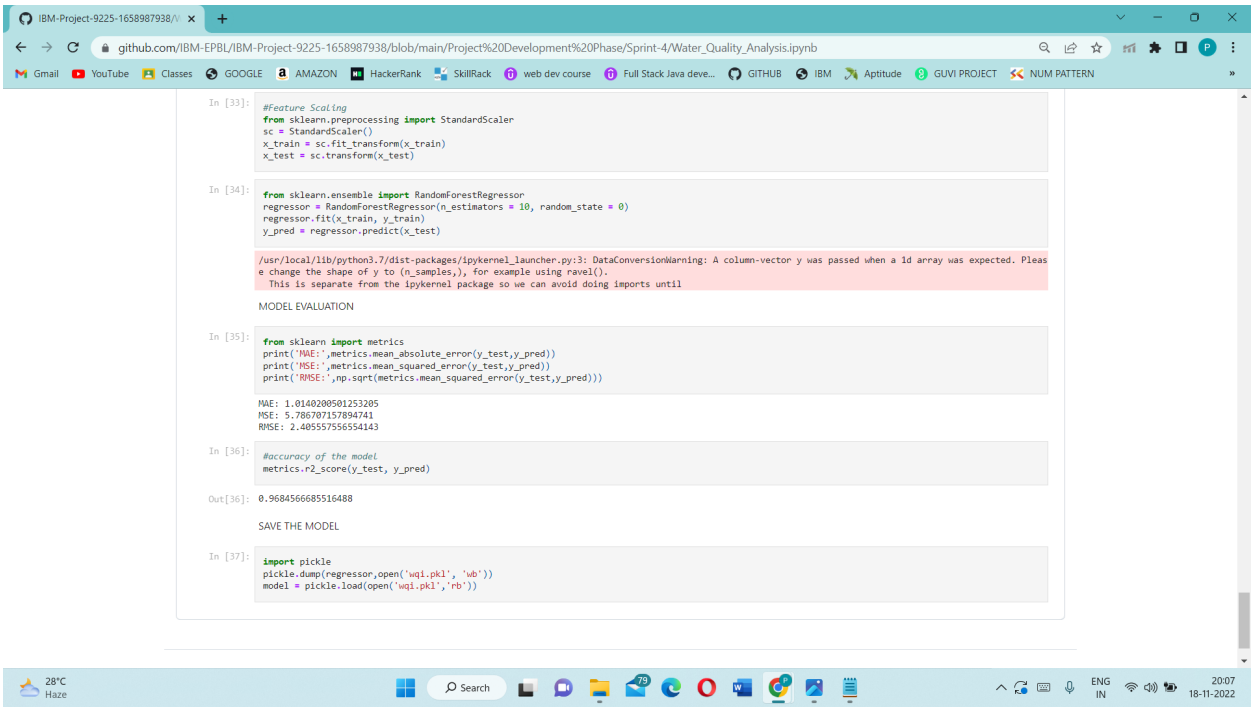
In [32]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size = 0.2,random_state=10)
```

RANDOM FOREST REGRESSION

In [33]:
```
#Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
x_train = sc.fit_transform(x_train)
x_test = sc.transform(x_test)
```

In [34]:
```
from sklearn.ensemble import RandomForestRegressor
regressor = RandomForestRegressor(n_estimators = 10, random_state = 0)
regressor.fit(x_train, y_train)
y_pred = regressor.predict(x_test)
```

28°C Haze Search ENG IN 20:07 18-11-2022



```
In [33]:
#Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
x_train = sc.fit_transform(x_train)
x_test = sc.transform(x_test)

In [34]:
from sklearn.ensemble import RandomForestRegressor
regressor = RandomForestRegressor(n_estimators = 10, random_state = 0)
regressor.fit(x_train, y_train)
y_pred = regressor.predict(x_test)

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:3: DataConversionWarning: A column-vector y was passed when a 1d array was expected. Pleas
e change the shape of y to (n_samples,), for example using ravel().
  This is separate from the ipykernel package so we can avoid doing imports until

MODEL EVALUATION

In [35]:
from sklearn import metrics
print('MAE:',metrics.mean_absolute_error(y_test,y_pred))
print('MSE:',metrics.mean_squared_error(y_test,y_pred))
print('RMSE:',np.sqrt(metrics.mean_squared_error(y_test,y_pred)))

MAE: 1.0140200501253205
MSE: 5.786707157894741
RMSE: 2.405557556554143

In [36]:
#accuracy of the model
metrics.r2_score(y_test, y_pred)

Out[36]:
0.9684566685516488

SAVE THE MODEL

In [37]:
import pickle
pickle.dump(regressor,open('wqi.pkl', 'wb'))
model = pickle.load(open('wqi.pkl','rb'))
```

app.py

```
import numpy as np
from flask import Flask,render_template,request
import pickle
app= Flask(__name__)
@app.route('/')
def home() :
  return render_template("web.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"]
  tc = request.form["tc"]
  na = request.form["na"]

  #ibm start
  import requests
```

```
import json
 # NOTE: you must manually set API_KEY below using information retrieved from your IBM Cloud account.
API_KEY = "H3ZSwhKcbaPPAmxup7UnpF9FBPEBXJHbaaa-yUeIJpJU"
token_response = requests.post('https://iam.cloud.ibm.com/identity/token', data={"apikey":
API_KEY, "grant_type": 'urn:ibm:params:oauth:grant-type:apikey'})
mltoken = token_response.json()["access_token"]

header = {'Content-Type': 'application/json', 'Authorization': 'Bearer ' + mltoken}

# NOTE: manually define and pass the array(s) of values to be scored in the next line
 payload_scoring = {"input_data": [{"field": [["do","ph","co","bod","tc","na"]], "values": [[do, ph, co, bod, tc, na]]}]}


response_scoring = requests.post('https://us-south.ml.cloud.ibm.com/ml/v4/deployments/b6831db5-17e8-4f64-87a2-56aa167b55a9/predictions?version=2022-11-16', json=payload_scoring,
headers={'Authorization': 'Bearer ' + mltoken})
print("Scoring response")
predict = response_scoring.json()
pred = (predict['predictions'][0]['values'][0][0])
y_pred=round(pred,3)
#ibm end

if(y_pred >= 95 and y_pred<=100):
      return render_template("web.html",showcase = 'Excellent, The Predicted Value Is '+ str(y_pred))
elif(y_pred >= 89 and y_pred<=94):
     return render_template("web.html",showcase = 'Very Good, The Predicted Value Is '+ str(y_pred))
elif(y_pred >= 80 and y_pred<=88):
  return render_template("web.html",showcase = 'Good, The Predicted Value Is '+ str(y_pred))
elif(y_pred >= 65 and y_pred<=79):
  return render_template("web.html",showcase = 'Fair, The Predicted Value Is '+ str(y_pred))
elif(y_pred >= 45 and y_pred<=64):
      return render_template("web.html",showcase = 'Marginal, The Predicted Value Is '+ str(y_pred))
else:
  return render_template("web.html",showcase = 'Poor, The Predicted Value Is '+ str(y_pred))
```

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

wqi.pkl

RandomForestRegressor(n_estimators=10, random_state=0)

8. TESTING

8.1 Test Cases

Testcase ID	Feature Type	Component	Test Scenario	Steps to execute	Expected Result	Actual Result
TC_1	Functional	HomePage	Verify user is able to predict quality of water.	Run 'app.py' file in command prompt and then open the webpage. Give values	User should get the predicted value.	User gets predicted value for their given data.

				in the respective text boxes and click PREDICT.		
TC_2	Functional	HomePage	Verify user is able to predict quality of water.	Run 'app.py' file in command prompt and then open the webpage. Give values in the respective text boxes and don't click PREDICT.	User simply gets error page.	Finally gets the error page.

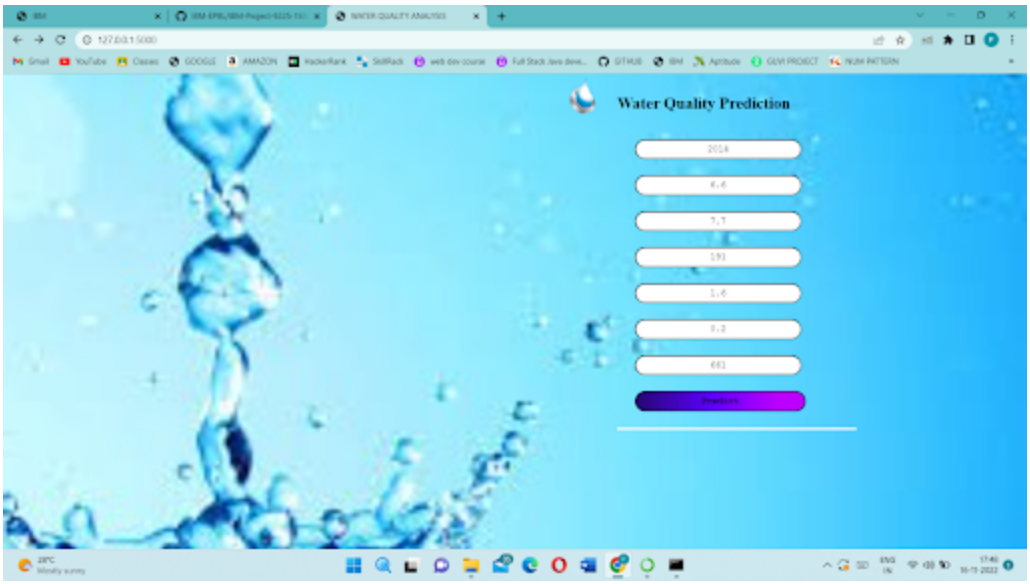
8.2 User Acceptance Testing

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

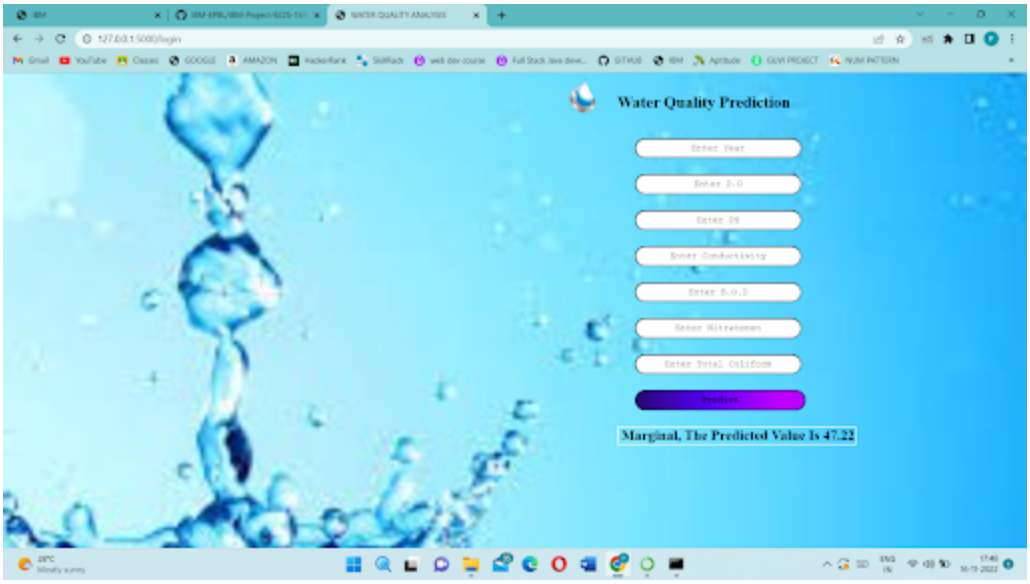
9. RESULTS

Output:

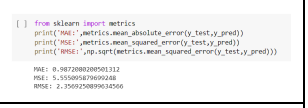
Webpage with data entered



Webpage with predicted value



9.1 Performance Metrics

S. No.	Parameter	Values	Screenshot
1.	Metrics	Regression model-MAE,MSE,RMSE	
2.	Tune the model	Accuracy score	

			[ ] #accuracy of the model
			metrics.r2_score(y_test, y_pred)
			
			0.9697191812580 9

10. ADVANTAGES & DISADVANTAGES

Advantages

- Water quality standards protect human health and avoid the cost related to medical care and productivity loss.

- It can give accurate measurement on pH, total dissolved salt and electrical conductivity.

Disadvantages

- Water pollutants may cause disease or act as poison.
- Health Risk: Open storage structures can be contaminated by animals and can provide a breeding ground for disease carrying insects.

11. CONCLUSION

Potability 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. FUTURE SCOPE

Each study-unit investigation will consist of a high-intensity phase, followed by a low-intensity phase. The high-intensity phase will cover the first 6 years of the study. During this phase, existing

water-quality data and basin characteristics will be evaluated; data for evaluation of surface-water and groundwater quality, river-bed sediments, and aquatic organisms will be collected; and technical and nontechnical reports and papers describing study results will be prepared. During the 4-year low-intensity phase, water-quality sampling will be less frequent than it will be during the highintensity phase. Surface water will be sampled at about 20 sites at weekly to monthly intervals during the high-intensity phase of the study. These sites will represent environmental settings characteristic of the study unit. Major water-quality constituents to be analyzed at these sites include suspended sediment, nutrients, major ions, and trace elements.

13 Appendix

GitHub Link - Github

Demo Link - Demo