CHRIST THE KING
ENGINEERING COLLEGE
Karamadai, Coimbatore – 641 104
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

Date	20 November2022
Team ID	PNT2022TMID42412
Project Name	Efficient Water QualityAnalysis and PredictionUsing Machine Learning

1. INTRODUCTION

1.1 Project Overview

Water is considered as a vital resource that affects various aspects of human health and lives. The quality of water is a major concern for people living in urban areas. The quality of water serves as a powerful environmental determinant and a foundation for the prevention and control of waterborne diseases. However predicting the urban water quality is a challenging task since the water quality varies in urban spaces non-linearly and depends on multiple factors, such as meteorology, water usage patterns, and land uses, so this project aims at building a Machine Learning (ML) model to Predict Water Quality by considering all water quality standard indicators.

1.2 Purpose

- To evaluate the quality of water and determine the safety of water.
- To monitor changes in water quality.
- To determine whether water is suitable for the health of the natural environment.
- To determine whether water is suitable for human consumption and other uses

2. LITERATURE SURVEY

a. **References**

1. **https://scikit-learn.org/stable/**
2. **https://stackoverflow.com/**
3. **https://www.geeksforgeeks.org/**
4. **https://youtu.be/T3PsRW6wZSY**

2.2 Problem Statement Definition

1. Water is one of the most essential for the existence of life. The safety and accessibility if drinking-water are major concerns throughout the globe

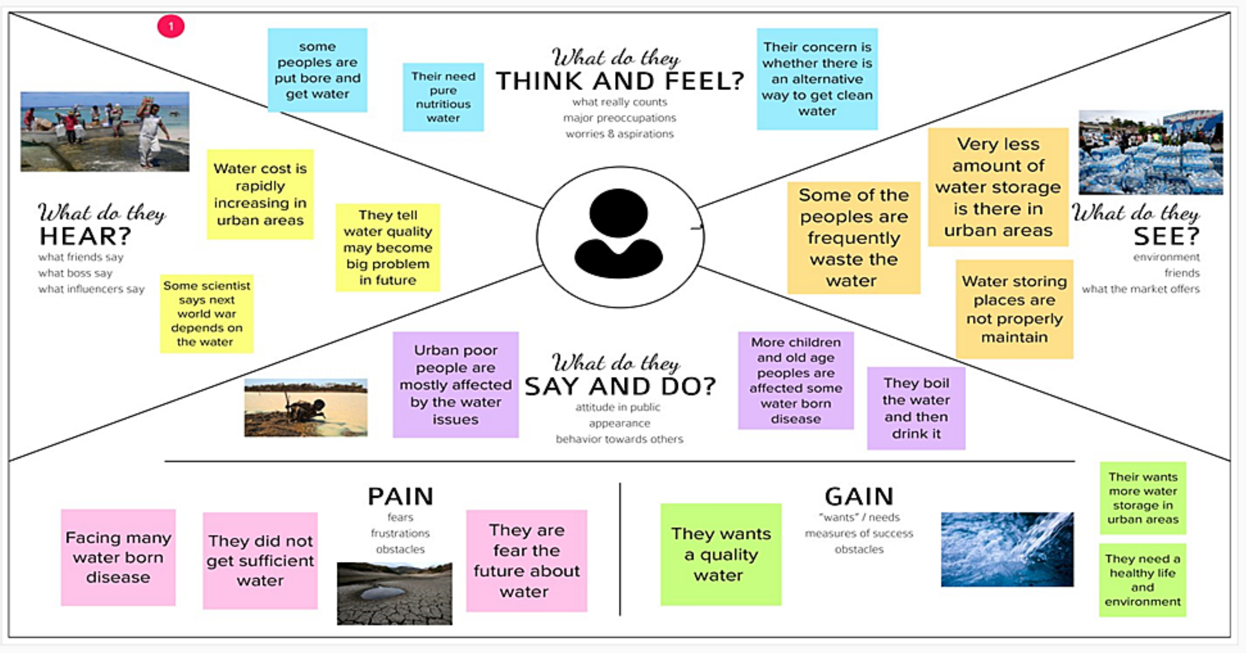
2. Water makes up about 70% of the surface and is one of the most important sources vital to sustaining life.

3. Water quality has been conventionally estimated through expensive and time consuming lab and statically analysis.

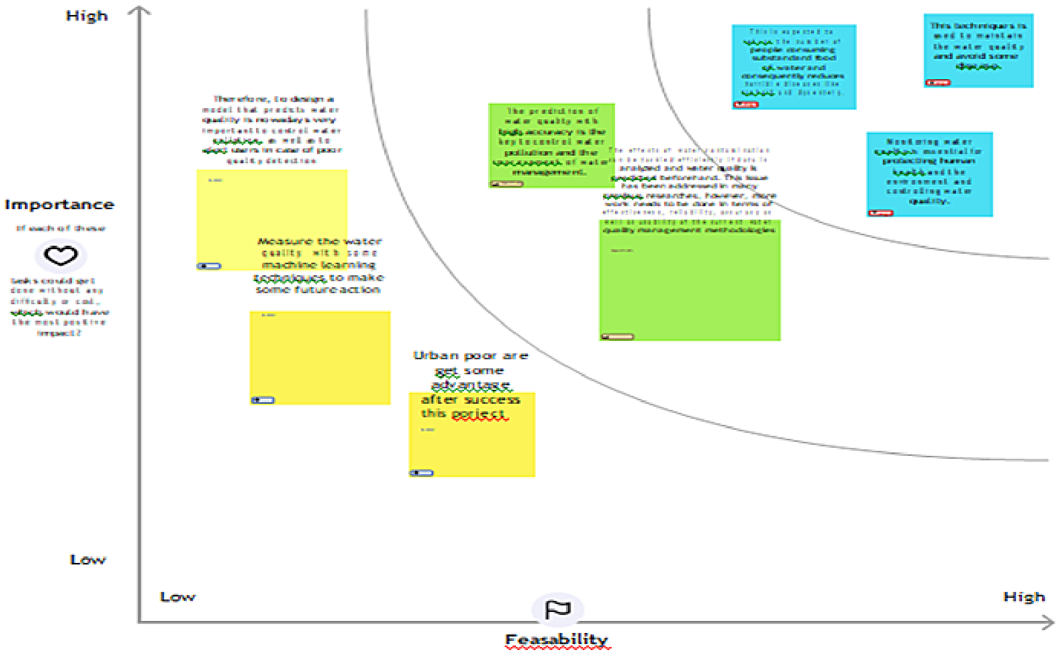
4. This system is proposed to check the water quality and warm the user before water gets contaminated using Machine Learning.

3. IDEATION & PROPOSED SOLUTION

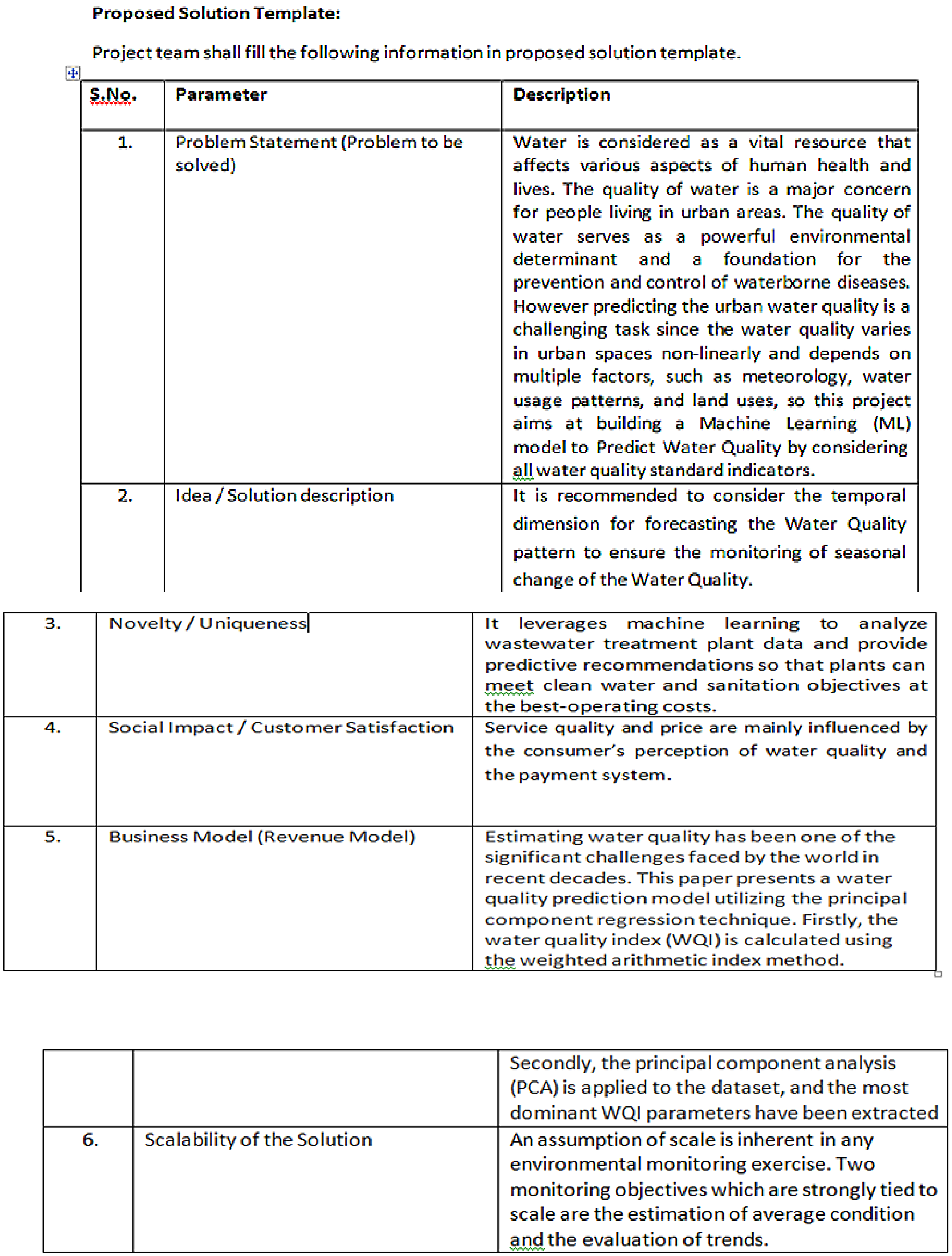
3.1 Empathy Map Canvas


1
some peoples are put bore and get water
Their need pure nutritious water
What do they THINK AND FEEL?
what really counts
major preoccupations
worries & aspirations
Their concern is whether there is an alternative way to get clean water
Water cost is rapidly increasing in urban areas
They tell water quality may become big problem in future
What do they HEAR?
what friends say
what boss say
what influencers say
Some scientist says next world war depends on the water
Some of the peoples are frequently waste the water
Very less amount of water storage is there in urban areas
Water storing places are not properly maintain
What do they SEE?
environment
friends
what the market offers
Urban poor people are mostly affected by the water issues
What do they SAY AND DO?
attitude in public
appearance
behavior towards others
More children and old age peoples are affected some water born disease
They boil the water and then drink it
PAIN
fears
frustrations
obstacles
Facing many water born disease
They did not get sufficient water
They are fear the future about water
GAIN
"wants" / needs
measures of success
obstacles
They wants a quality water
Their wants more water storage in urban areas
They need a healthy life and environment

3.2 Ideation & Brainstorming


High
Importance
If each of these tasks could get done without any difficulty or cost, which would have the most positive impact?
Low
Low
Feasability
High
Therefore, to design a model that predicts water quality is nowadays very important to control water pollution, as well as to alert users in case of poor quality detection
Measure the water quality with some machine learning techniques to make some future action
Urban poor are get some advantage after success this project
The prediction of water quality with high accuracy is the key to control water pollution and the improvement of water management.
The effects of water contamination can be tackled efficiently if data is analyzed and water quality is predicted beforehand. This issue has been addressed in many previous researches, however, more work needs to be done in terms of effectiveness, reliability, accuracy as well as usability of the current water quality management methodologies.
This is expected to reduce the number of people consuming substandard food or water and consequently reduces terrible diseases like typhoid and dysentery.
This techniques is used to maintain the water quality and avoid some disease.
Monitoring water quality is essential for protecting human health and the environment and controlling water quality.

3.3 Proposed Solution

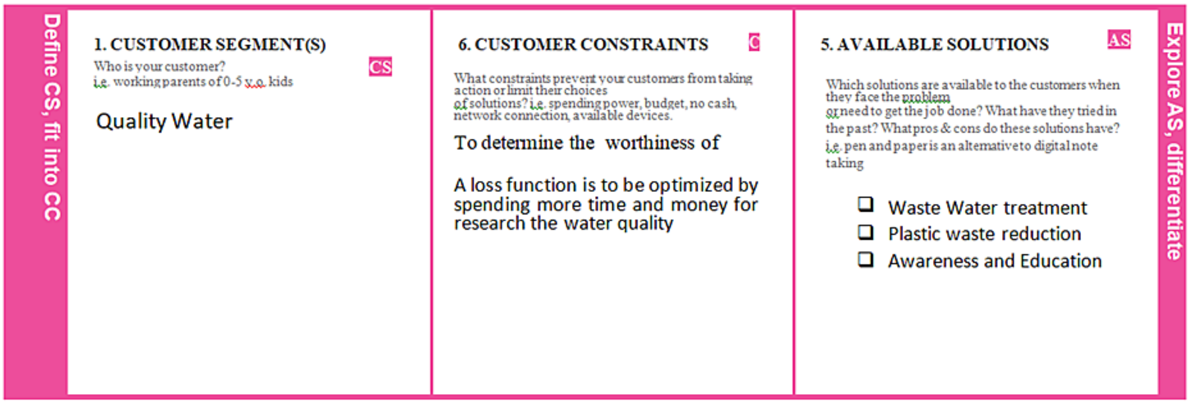


Proposed Solution Template:

Project team shall fill the following information in proposed solution template.

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 description	It is recommended to consider the temporal dimension for forecasting the Water Quality pattern to ensure the monitoring of seasonal change of the Water Quality.
3.	Novelty / Uniqueness	It leverages machine learning to analyze wastewater treatment plant data and provide predictive recommendations so that plants can meet clean water and sanitation objectives at the best-operating costs.
4.	Social Impact / Customer Satisfaction	Service quality and price are mainly influenced by the consumer's perception of water quality and the payment system.
5.	Business Model (Revenue Model)	Estimating water quality has been one of the significant challenges faced by the world in recent decades. This paper presents a water quality prediction model utilizing the principal component regression technique. Firstly, the water quality index (WQI) is calculated using the weighted arithmetic index method. Secondly, the principal component analysis (PCA) is applied to the dataset, and the most dominant WQI parameters have been extracted
6.	Scalability of the Solution	An assumption of scale is inherent in any environmental monitoring exercise. Two monitoring objectives which are strongly tied to scale are the estimation of average condition and the evaluation of trends.

3.4 Problem Solution fit



Define CS, fit into CC

1. CUSTOMER SEGMENT(S) CS

Who is your customer?
i.e. working parents of 0-5 y.o. kids

Quality Water

6. CUSTOMER CONSTRAINTS CC

What constraints prevent your customers from taking action or limit their choices of solutions? i.e. spending power, budget, no cash, network connection, available devices.

To determine the worthiness of

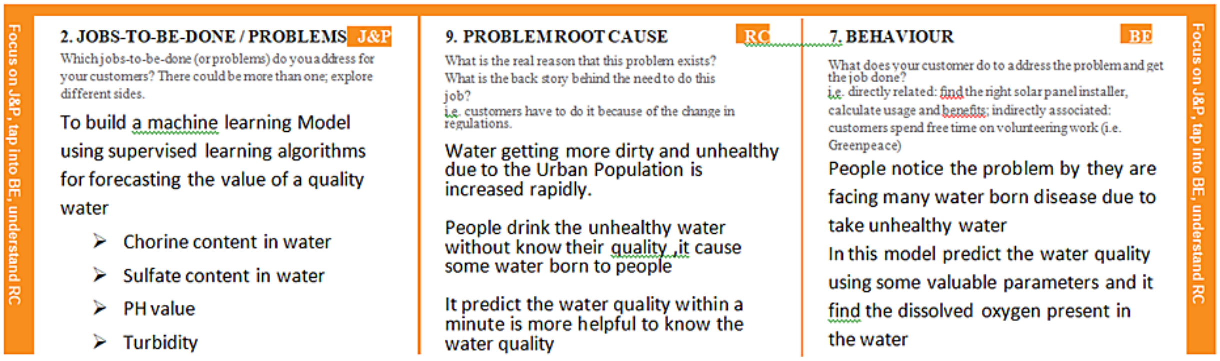
A loss function is to be optimized by spending more time and money for research the water quality

5. AVAILABLE SOLUTIONS AS

Which solutions are available to the customers when they face the problem or need to get the job done? What have they tried in the past? What pros & cons do these solutions have? i.e. pen and paper is an alternative to digital note taking

- ❑ Waste Water treatment
- ❑ Plastic waste reduction
- ❑ Awareness and Education

Explore AS, differentiate



Focus on J&P, tap into BE, understand RC

2. JOBS-TO-BE-DONE / PROBLEMS J&P

Which jobs-to-be-done (or problems) do you address for your customers? There could be more than one; explore different sides.

To build a machine learning Model using supervised learning algorithms for forecasting the value of a quality water

- ➢ Chorine content in water
- ➢ Sulfate content in water
- ➢ PH value
- ➢ Turbidity

9. PROBLEM ROOT CAUSE RC

What is the real reason that this problem exists? What is the back story behind the need to do this job?
i.e. customers have to do it because of the change in regulations.

Water getting more dirty and unhealthy due to the Urban Population is increased rapidly.

People drink the unhealthy water without know their quality ,it cause some water born to people

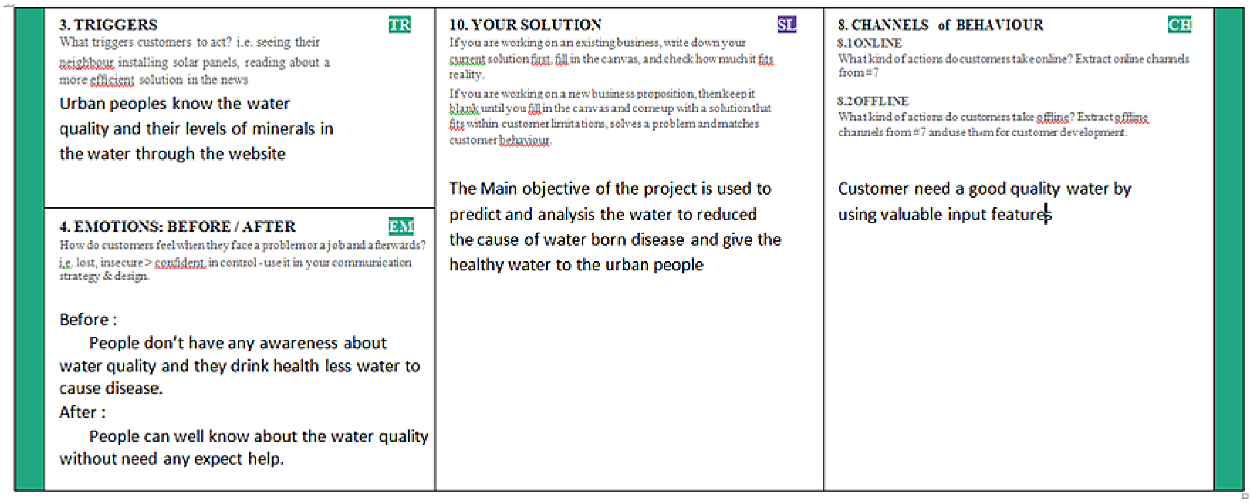
It predict the water quality within a minute is more helpful to know the water quality

7. BEHAVIOUR BE

What does your customer do to address the problem and get the job done?
i.e. directly related: find the right solar panel installer, calculate usage and benefits; indirectly associated: customers spend free time on volunteering work (i.e. Greenpeace)

People notice the problem by they are facing many water born disease due to take unhealthy water
In this model predict the water quality using some valuable parameters and it find the dissolved oxygen present in the water

Focus on J&P, tap into BE, understand RC



3. TRIGGERS TR

What triggers customers to act? i.e. seeing their neighbour installing solar panels, reading about a more efficient solution in the news

Urban peoples know the water quality and their levels of minerals in the water through the website

4. EMOTIONS: BEFORE / AFTER EM

How do customers feel when they face a problem or a job and afterwards?
i.e. lost, insecure > confident, in control - use it in your communication strategy & design.

Before :
People don't have any awareness about water quality and they drink health less water to cause disease.
After :
People can well know about the water quality without need any expect help.

10. YOUR SOLUTION SL

If you are working on an existing business, write down your current solution first, fill in the canvas, and check how much it fits reality.

If you are working on a new business proposition, then keep it blank until you fill in the canvas and come up with a solution that fits within customer limitations, solves a problem and matches customer behaviour.

The Main objective of the project is used to predict and analysis the water to reduced the cause of water born disease and give the healthy water to the urban people

8. CHANNELS of BEHAVIOUR CH

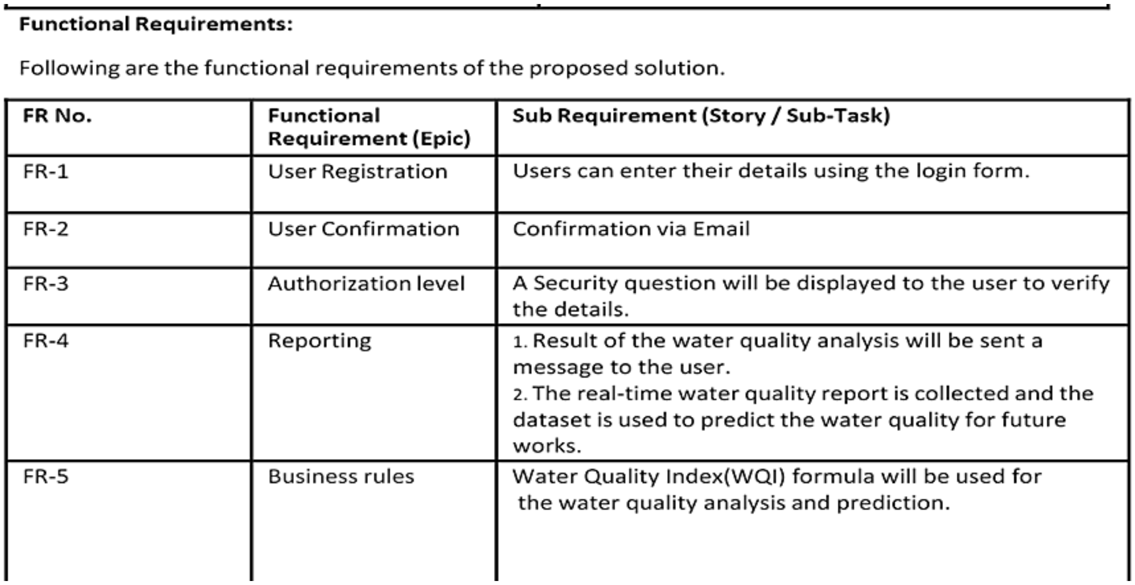
8.1 ONLINE
What kind of actions do customers take online? Extract online channels from #7

8.2 OFFLINE
What kind of actions do customers take offline? Extract offline channels from #7 and use them for customer development.

Customer need a good quality water by using valuable input features

4. REQUIREMENT ANALYSIS

4.1 Functional requirement

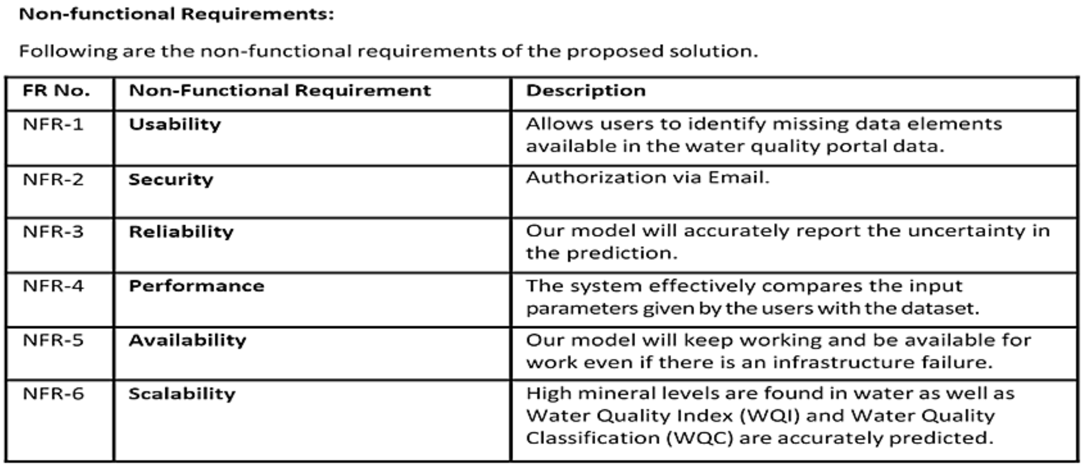


Functional Requirements:

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Users can enter their details using the login form.
FR-2	User Confirmation	Confirmation via Email
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



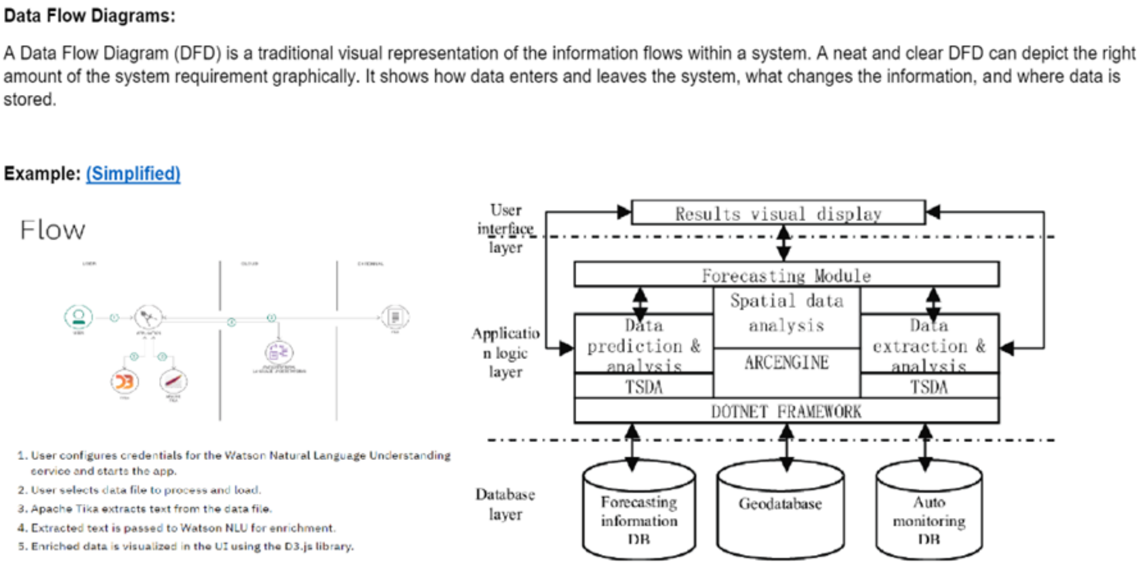
Non-functional Requirements:

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

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 & User Stories



Data Flow Diagrams:

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.

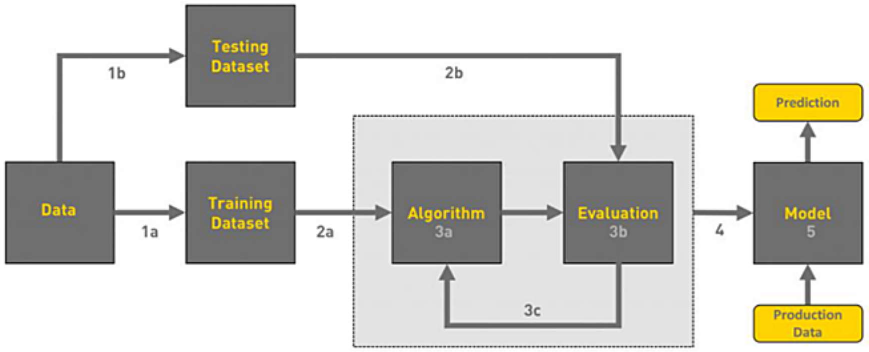
Example: (Simplified)

Flow

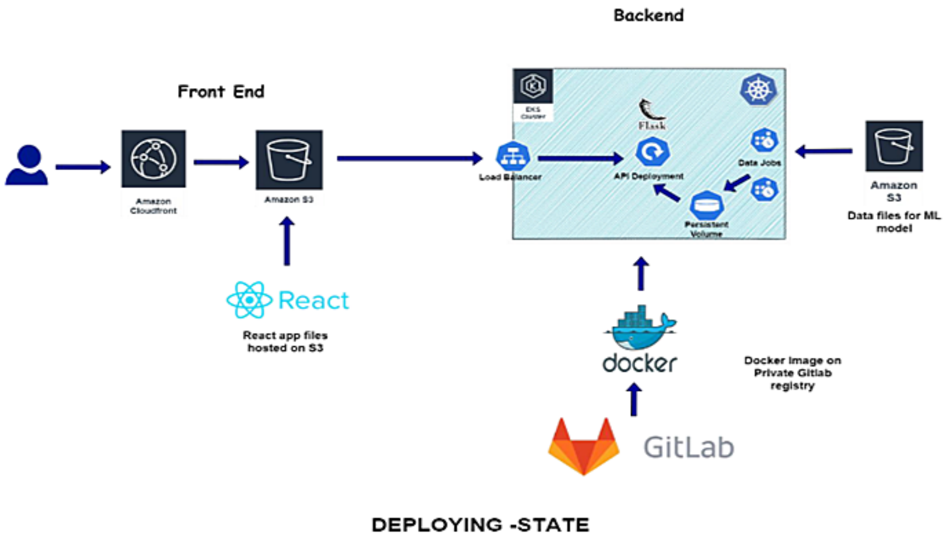
1. User configures credentials for the Watson Natural Language Understanding service and starts the app.
2. User selects data file to process and load.
3. Apache Tika extracts text from the data file.
4. Extracted text is passed to Watson NLU for enrichment.
5. Enriched data is visualized in the UI using the D3.js library.

5.2 Solution & Technical Architecture

Technical Architecture:



DEVELOPING -STATE



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	Registration	USN-1	As a user, I can register for the application by enteringmy email, password, and confirming my password.	1 0	High	Selva Kumar. E, Arul Christober.T
Sprint-1	Confirmation Mail	USN-2	As a user, I will receive confirmation email once I have registered for the application.	1 0	High	Selva Kumar. E,Arul Christober.T
Sprint-2	Data collection	USN-1	As a user, I can Collect the data from Kaggle.	1 0	Low	Selva Kumar. E,Arul Christober.T
Sprint-2	Model Building	USN-2	As a user, I can create the machine learning model.	1 0	Medium	Selva Kumar. E,Arul Christober.T
Sprint-3	Connect to IBM-Watson	USN-1	As a user, I can connect my model into ibm cloud.	2 0	High	Selva Kumar. E,Shilpa Merlin.P

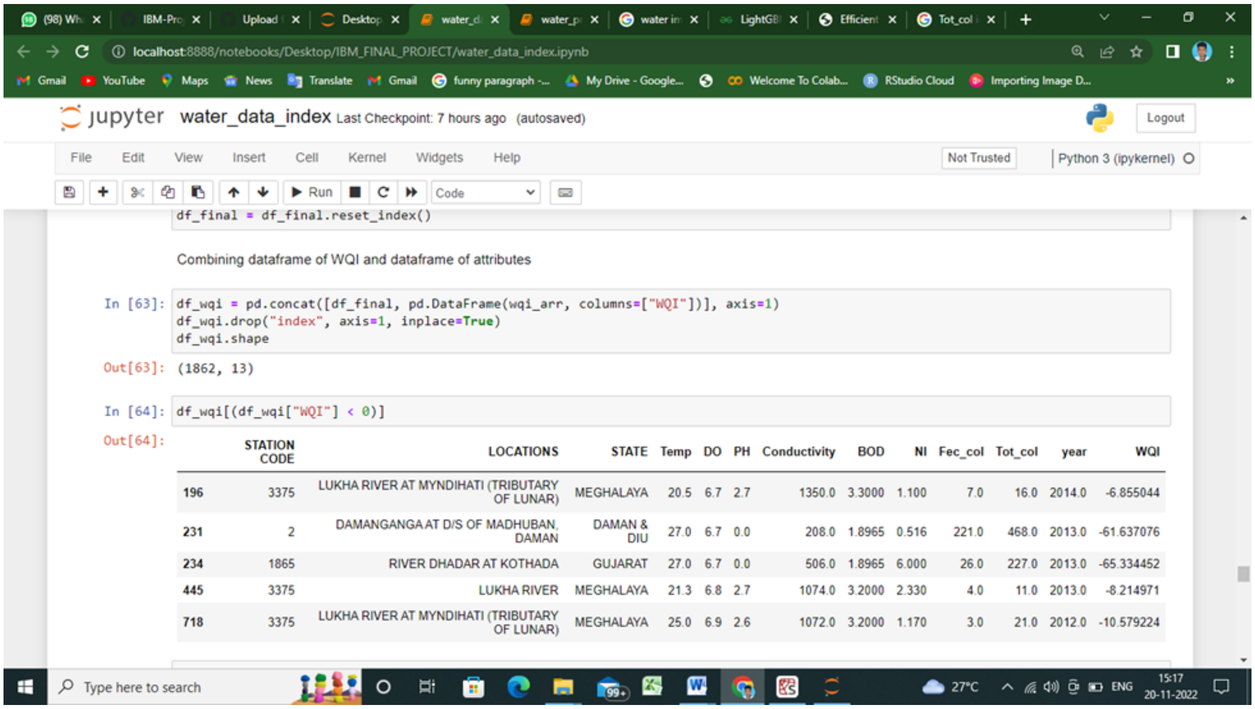
Sprint-4	Prediction	USN-1	As a user,I can get a outputof my inputparameter inthe website window.	2 0	High	Selva Kumar. E,Sri karthiga.V

6.2 Sprint Delivery Schedule

Sprint	Total Story Points	Duration	Sprint StartDate	Sprint End Date(Planned)	Story PointsCompleted (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	20	05 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

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

7.1 Water Quality index



localhost:8888/notebooks/Desktop/IBM_FINAL_PROJECT/water_data_index.ipynb
jupyter water_data_index Last Checkpoint: 7 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Type here to search
27°C
ENG
15:17
20-11-2022

```
df_final = df_final.reset_index()
```

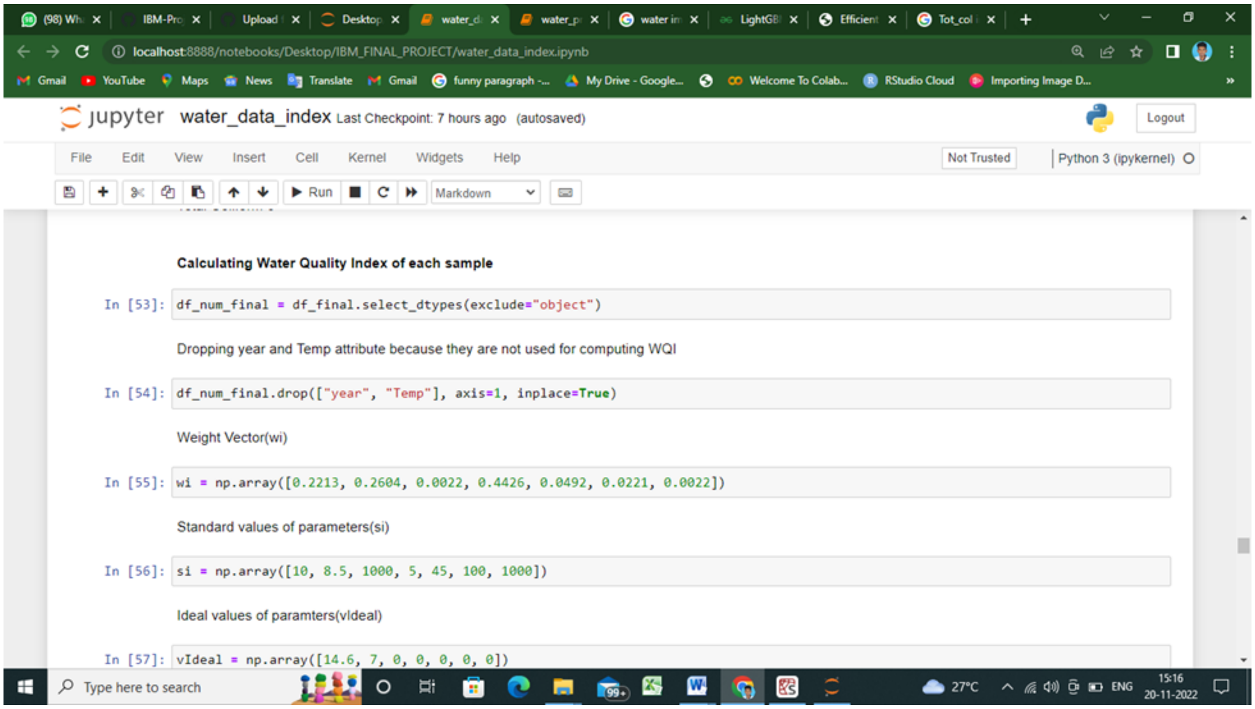
Combining dataframe of WQI and dataframe of attributes

```
In [63]: df_wqi = pd.concat([df_final, pd.DataFrame(wqi_arr, columns=["WQI"])], axis=1)
         df_wqi.drop("index", axis=1, inplace=True)
         df_wqi.shape
Out[63]: (1862, 13)
```

```
In [64]: df_wqi[(df_wqi["WQI"] < 0)]
```

Out[64]:

	STATION CODE	LOCATIONS	STATE	Temp	DO	PH	Conductivity	BOD	NI	Fec_col	Tot_col	year	WQI
196	3375	LUKHA RIVER AT MYNDIHATI (TRIBUTARY OF LUNAR)	MEGHALAYA	20.5	6.7	2.7	1350.0	3.3000	1.100	7.0	16.0	2014.0	-6.855044
231	2	DAMANGANGA AT D/S OF MADHUBAN, DAMAN	DAMAN & DIU	27.0	6.7	0.0	208.0	1.8965	0.516	221.0	468.0	2013.0	-61.637076
234	1865	RIVER DHADAR AT KOTHADA	GUJARAT	27.0	6.7	0.0	506.0	1.8965	6.000	26.0	227.0	2013.0	-65.334452
445	3375	LUKHA RIVER	MEGHALAYA	21.3	6.8	2.7	1074.0	3.2000	2.330	4.0	11.0	2013.0	-8.214971
718	3375	LUKHA RIVER AT MYNDIHATI (TRIBUTARY OF LUNAR)	MEGHALAYA	25.0	6.9	2.6	1072.0	3.2000	1.170	3.0	21.0	2012.0	-10.579224



jupyter water_data_index Last Checkpoint: 7 hours ago (autosaved)

Calculating Water Quality Index of each sample

```
In [53]: df_num_final = df_final.select_dtypes(exclude="object")
```

Dropping year and Temp attribute because they are not used for computing WQI

```
In [54]: df_num_final.drop(["year", "Temp"], axis=1, inplace=True)
```

Weight Vector(wi)

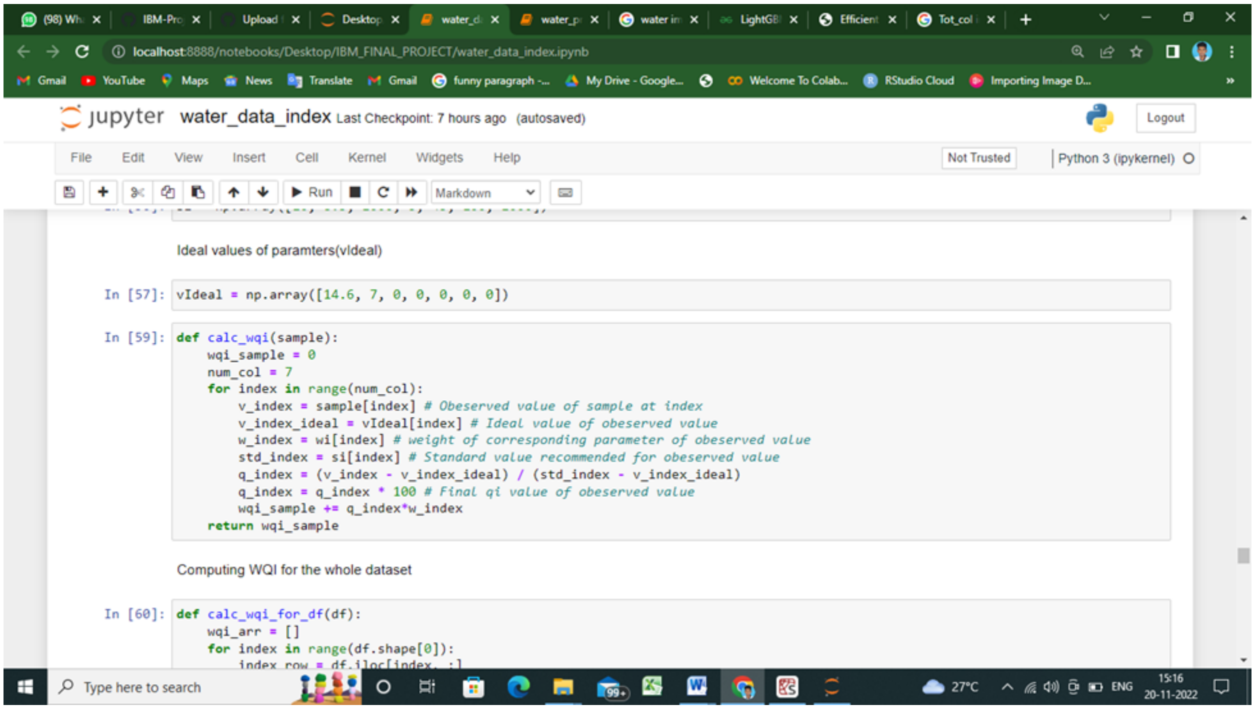
```
In [55]: wi = np.array([0.2213, 0.2604, 0.0022, 0.4426, 0.0492, 0.0221, 0.0022])
```

Standard values of parameters(si)

```
In [56]: si = np.array([10, 8.5, 1000, 5, 45, 100, 1000])
```

Ideal values of paramters(vIdeal)

```
In [57]: vIdeal = np.array([14.6, 7, 0, 0, 0, 0, 0])
```



jupyter water_data_index Last Checkpoint: 7 hours ago (autosaved)

Ideal values of paramters(vIdeal)

```
In [57]: vIdeal = np.array([14.6, 7, 0, 0, 0, 0, 0])
```

```
In [59]: def calc_wqi(sample):
             wqi_sample = 0
             num_col = 7
             for index in range(num_col):
                 v_index = sample[index] # Obeserved value of sample at index
                 v_index_ideal = vIdeal[index] # Ideal value of obeserved value
                 w_index = wi[index] # weight of corresponding parameter of obeserved value
                 std_index = si[index] # Standard value recommended for obeserved value
                 q_index = (v_index - v_index_ideal) / (std_index - v_index_ideal)
                 q_index = q_index * 100 # Final qi value of obeserved value
                 wqi_sample += q_index*w_index
             return wqi_sample
```

Computing WQI for the whole dataset

```
In [60]: def calc_wqi_for_df(df):
             wqi_arr = []
             for index in range(df.shape[0]):
                 index_row = df.iloc[index, :]
```

7.2 Flask app

import pickle

```
model=pickle.load(open('model.pkl','rb'))
@app.route('/home') # slash is the url binding to render this page
def page():
   return render_template("home.html")
@app.route('/y_predict',methods=["GET/POST"])
def login():
        a=request.form('Temp')
        b=request.form('DO')
        c=request.form('PH')
        d=request.form('Conductivity')
        e=request.form('BOD')
       f=request.form('NI')
        g=request.form('Fec_col')
        h=request.form('Tot_col')
      t=[[float(a),float(b),float(c),float(d),float(e),float(f),float(g),float(h)]]
    Prediction = model.predict(t)
    if (Prediction==1):
        return render_template("home.html",y="The Water Quality is good and say for
drink"+str(Prediction))
    else:
        return render_template("home.html",y="The Water Quality is bad and don't
drink"+str(Prediction))
if __name__ =="__main__":
 app.run(debug= False)
```

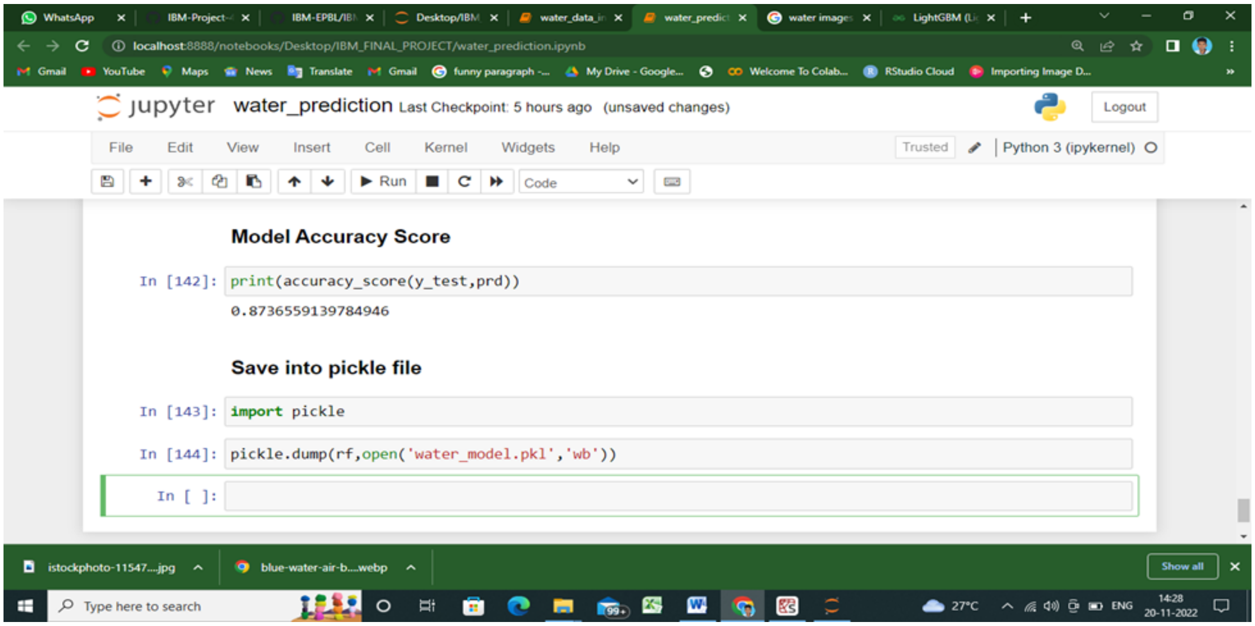
8. TESTING

8.1 Test Cases

8.2 User Acceptance Testing

9. RESULTS

9.1 Performance Metrics



Model Accuracy Score

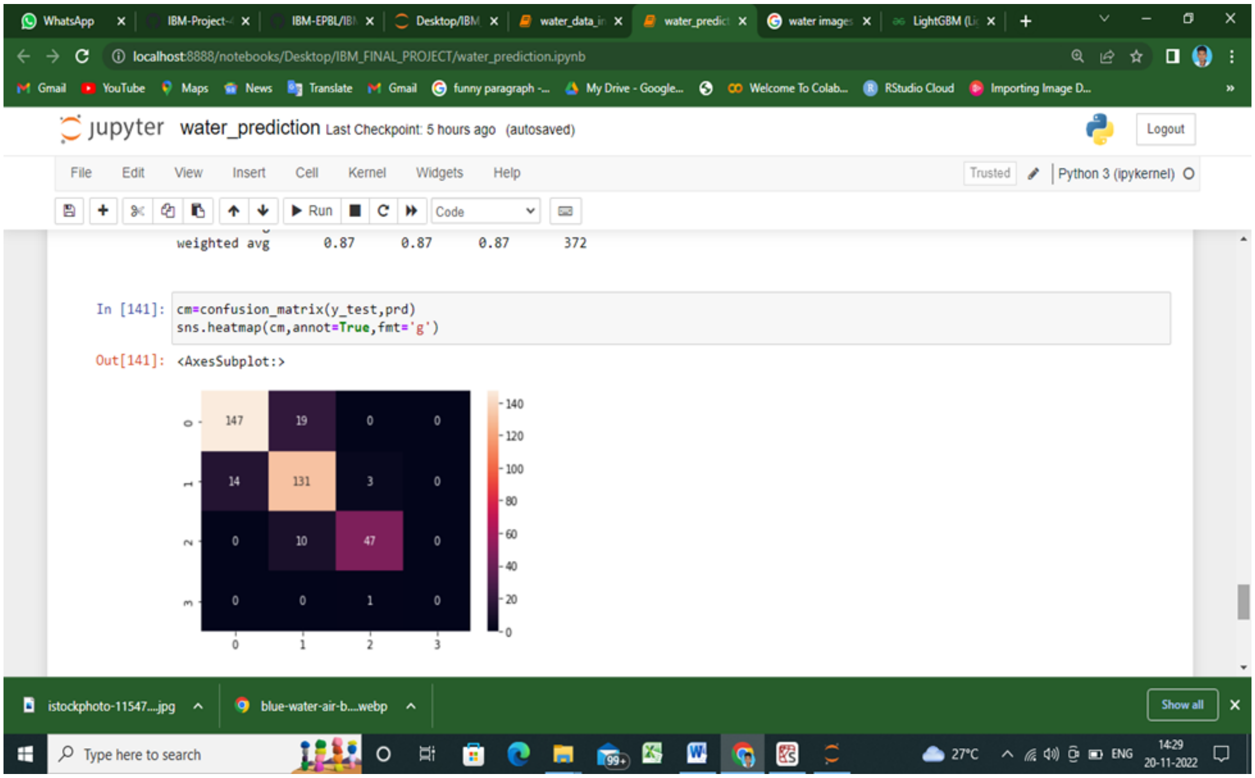
```
In [142]: print(accuracy_score(y_test,prd))
          0.8736559139784946
```

Save into pickle file

```
In [143]: import pickle

In [144]: pickle.dump(rf,open('water_model.pkl','wb'))

In [ ]:
```



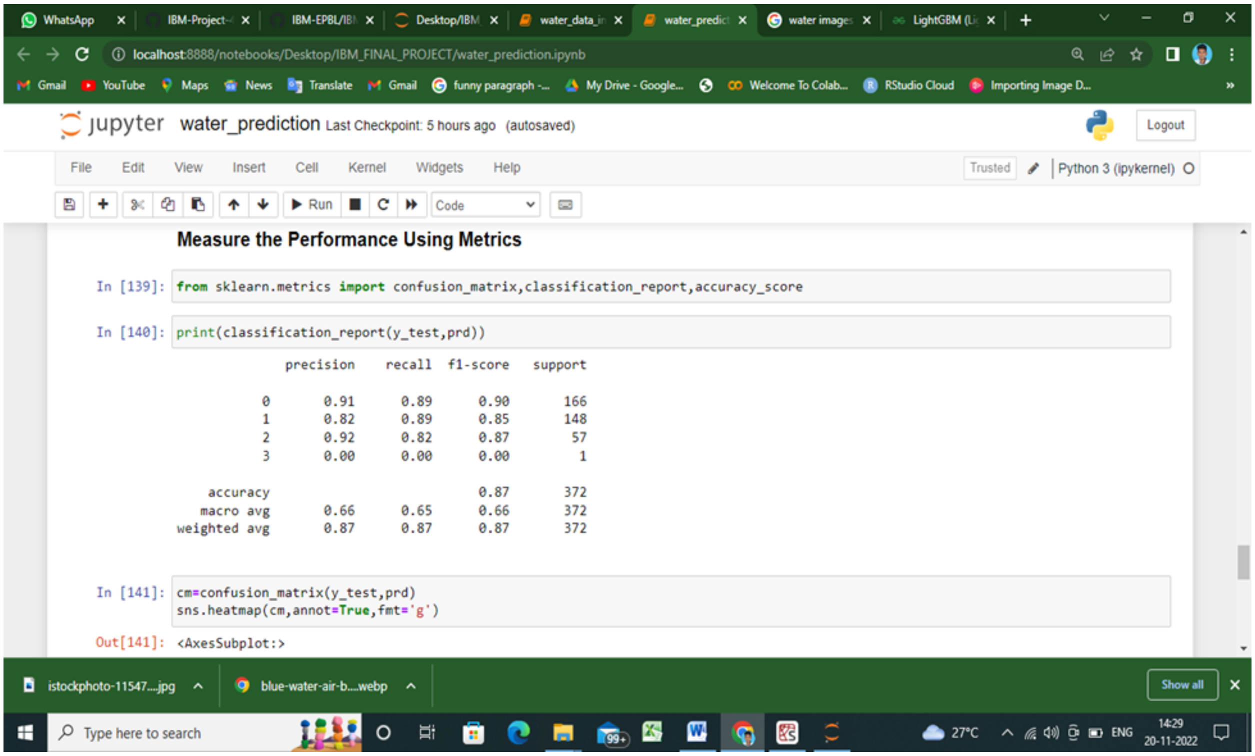
jupyter water_prediction Last Checkpoint: 5 hours ago (autosaved)
localhost:8888/notebooks/Desktop/IBM_FINAL_PROJECT/water_prediction.ipynb
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Logout
Run
Code

```
weighted avg       0.87      0.87      0.87       372
```

```
In [141]: cm=confusion_matrix(y_test,prd)
          sns.heatmap(cm,annot=True,fmt='g')

Out[141]: <AxesSubplot:>
```

147
19
0
0
14
131
3
0
0
10
47
0
0
0
1
0
0
1
2
3
140
120
100
80
60
40
20
0
istockphoto-11547....jpg
blue-water-air-b....webp
Show all
Type here to search
27°C
ENG
14:29
20-11-2022



jupyter water_prediction Last Checkpoint: 5 hours ago (autosaved)
localhost:8888/notebooks/Desktop/IBM_FINAL_PROJECT/water_prediction.ipynb
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Logout
Run
Code

Measure the Performance Using Metrics

```
In [139]: from sklearn.metrics import confusion_matrix,classification_report,accuracy_score

In [140]: print(classification_report(y_test,prd))

              precision    recall  f1-score   support

           0       0.91      0.89      0.90       166
           1       0.82      0.89      0.85       148
           2       0.92      0.82      0.87        57
           3       0.00      0.00      0.00         1

    accuracy                           0.87       372
   macro avg       0.66      0.65      0.66       372
weighted avg       0.87      0.87      0.87       372

In [141]: cm=confusion_matrix(y_test,prd)
          sns.heatmap(cm,annot=True,fmt='g')

Out[141]: <AxesSubplot:>
```

10. ADVANTAGES & DISADVANTAGES

10.1 ADVANTAGES

- Water is an essential element of our lives, giving life to all living creatures on Earth.
- The importance of ensuring that this water is of sound quality is extremely important, particularly if that water is intended for consumption.
- Water quality testing can provide valuable data on the condition of a particular body of water, and whether it may need special treatment before use.

10.2 DISADVANTAGES

- Quality water monitoring system requires a lot of man hours for its installation and operation. This is because water quality monitoring system consists of multiple instruments and they all are very time consuming.
- Quality water monitoring system has a high initial cost and it is very difficult for businesses to afford the total expense of quality water monitoring system. Usually a good quality water monitoring system will be for thousands of dollars or lakhs.
- Quality water monitoring system is very time consuming and the whole procedure is not reliable.
-

11. CONCLUSION

In this work, we presented approaches for event detection on water quality time series data. This work represents a case study and its aim is to find the best model on anomaly detection on water quality systems. Time series are analysed using statistical algorithms for a long time. Today machine learning algorithms are very popular on performing very well on time series data sets. Our experiment shows the weakness of machine learning algorithms when applied to a highly imbalanced data set.

12. FUTURE SCOPE

Given the dynamics of chemical innovation, production, consumption, use, disposal, and consequent emission into the aquatic environment, the challenge for a successful amendment and implementation of the European Water Framework Directive [29] is to define more specific strategies for protecting and enhancing the status of aquatic ecosystems. In particular, strategies for identifying river basin specific pollutants, improvements in the diagnostics of ecological impacts and more powerful approaches for establishing causal links between chemical and ecological assessments are required.

13.GitHub & Project Demo Link

Github Link:

https://github.com/IBM-EPBL/IBM-Project-7104-1658847437/tree/main/Final%20Deliverabiles

Project Demo Link:

https://drive.google.com/file/d/1mbhyh6E5TW85vldxrsxImyvJ-v9V8GSU/view?usp=sharing