

PROJECT REPORT

**EFFICIENT WATER QUALITY ANALYSIS
AND PREDICTION**

By team – PNT2022TMID53347
Batch no - B2-2M4E

**SRI VENKATESWARA COLLEGE OF
ENGINEERING**

KEERTHIRAJAN S (2127190501058)
MANOJKUMAR M (2127190501072)
MINU S (2127190501073)
NIHIL RENGASAMY T(2127190501078)

Under the guidance of,

Dr. Revathi N (Mentor)
Lalitha Gayathri (Industry Mentor)

TABLE OF CONTENTS

S.no	Title	Pg no
1.	Introduction	3
1.a	Project overview	5
1.b	Purpose	7
2.	Literature survey	9
2.1	Existing problems	11
2.2	Reference	13
2.3	Problem statement and definition	15
3.	IDEATION & PROPOSED SOLUTION	17
3.1	Empathy Map Canvas	18
3.2	Ideation & Brainstorming	19
3.3	Proposed Solution	21
3.4	Problem Solution fit	23
4.	REQUIREMENT ANALYSIS	25
4.1	Functional requirement	26
4.2	Non-Functional requirements	27
5.	PROJECT DESIGN	28
5.1	Data Flow Diagrams	39
5.2	Solution & Technical Architecture	30
5.3	User Stories	31
6.	PROJECT PLANNING & SCHEDULING	32
6.1	Sprint Planning & Estimation	34
6.2	Sprint Delivery Schedule	35
6.3	Reports from JIRA	36
7.	CODING & SOLUTIONING (Explain the features added in the project along with code)	37
7.1	Feature 1	38
7.2	Feature 2	39
7.3	Database Schema (if Applicable)	40
8.	TESTING	41
8.1	Test Cases	42

8.	User Acceptance Testing	43
9.	RESULTS	44
9.1.	Performance Metrics	44
10.	ADVANTAGES & DISADVANTAGES	45
11.	CONCLUSION	46
12.	FUTURE SCOPE	47
13.	APPENDIX	48

1.INTRODUCTION

water, a substance composed of the chemical elements hydrogen and oxygen and existing in gaseous, liquid, and solid states. It is one of the most plentiful and essential of compounds. 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

Yinguo Qiu , Hui Xie (2019) “A Novel Spatiotemporal Data Model for River Water Quality Visualization and Analysis” River water quality (RWQ) data has obvious characteristics of spatial and temporal distribution, and tables are conventionally exploited for storage of multi-period monitoring data of RWQ; however, neither effective visualization nor accurate analysis of the obtained data can be realized due to its dispersion character. In this paper, a novel spatiotemporal data model is proposed for RWQ data to realize conveniently data representation and spatiotemporal analysis. In this model, a spatial point, containing both location and dynamic water quality information, is considered as the basic element of river spaces, and methods of expanding a point to a line segment, a flat surface and a cube are designed respectively so as to make this model be applicable to different generalizations of river spaces. Moreover, a temporal data storage structure is designed so that efficient inquiry and advanced analysis of RWQ data can be guaranteed and the occupied memory space can be reduced. Finally, case studies are conducted by performing 3D visualization, trend analysis and anomaly identification on RWQ data, the result of which showing that tridimensional representation of RWQ data can be realized efficiently, the computational complexity is reduced significantly and the occupied memory space of monitoring data is effectively economized..

NohaE.Attar,HeshamR.Lotfy(2022)”PerformanceofArtificialIntelligenceMn Analysis and Prediction of Water Potability”

”Water is a prime necessity for the survival and sustenance of all living beings. Thus, it is very important to maintain a water quality balance. Otherwise, it would seriously damage the health of humans and severely affect the ecological balance among other species. Water quality is an important factor to consider, whether for ecosystem needs or contamination levels that directly impact health, hygiene, food, and the economy. In this study, we have utilized Artificial Intelligence as an efficient technique to predict water quality without resorting to the traditional analysis methods of water monitoring. We have developed a convolutional neural network as a deep learning algorithm and applied five kinds of machine learning algorithms: Backpropagation Neural Network, Random Forest, Decision Jungle, Naive Bayes, Logistic Regression, and Support Vector Machine. The reported results of the classification process showed

that the CNN has superior to the other machine learning algorithms in predicting the water potability based on the eight adopted parameters of water properties. CNN has achieved nearly 97% accuracy in the prediction process, while RF, the best one in the utilized machine learning algorithms, has recorded 81%. In addition, CNN has also proven its capability to reduce classification processing time.

2.1 EXISITNG 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

1. Ling, J.K.B. Water Quality Study and Its Relationship with High Tide and Low Tide at Kuantan River. Bachelor's Thesis, Universiti Malaysia Pahang, Gambang, Malaysia, 2010. Available online: http://umpir.ump.edu.my/id/eprint/2449/1/JACKY_LING_KUO_BAO.PDF (accessed on 22 February 2022).
2. Xu, J.; Gao, X.; Yang, Z.; Xu, T. Trend and Attribution Analysis of Runoff Changes in the Weihe River Basin in the Last 50 Years. *Water* 2022, 14, 47.
3. Wahab, M.A.A.; Jamadon, N.K.; Mohmood, A.; Syahir, A. River Pollution Relationship to the National Health Indicated by Under-Five Child Mortality Rate: A Case Study in Malaysia. *Bioremediat. Sci. Technol. Res.* 2015, 3, 20–25.
4. Abbasi, T.; Abbasi, S.A. *Water Quality Indices*; Elsevier: Amsterdam, The Netherlands, 2012.
5. Abyaneh, H.Z. Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. *J. Environ. Health Sci. Eng.* 2014, 12, 40.
6. Alias, S.W.A.N. Ecosystem Health Assessment of Sungai Pengkalan Chepa Basin: Water Quality and Heavy Metal Analysis. *Sains Malays.* 2020, 49, 1787–1798.

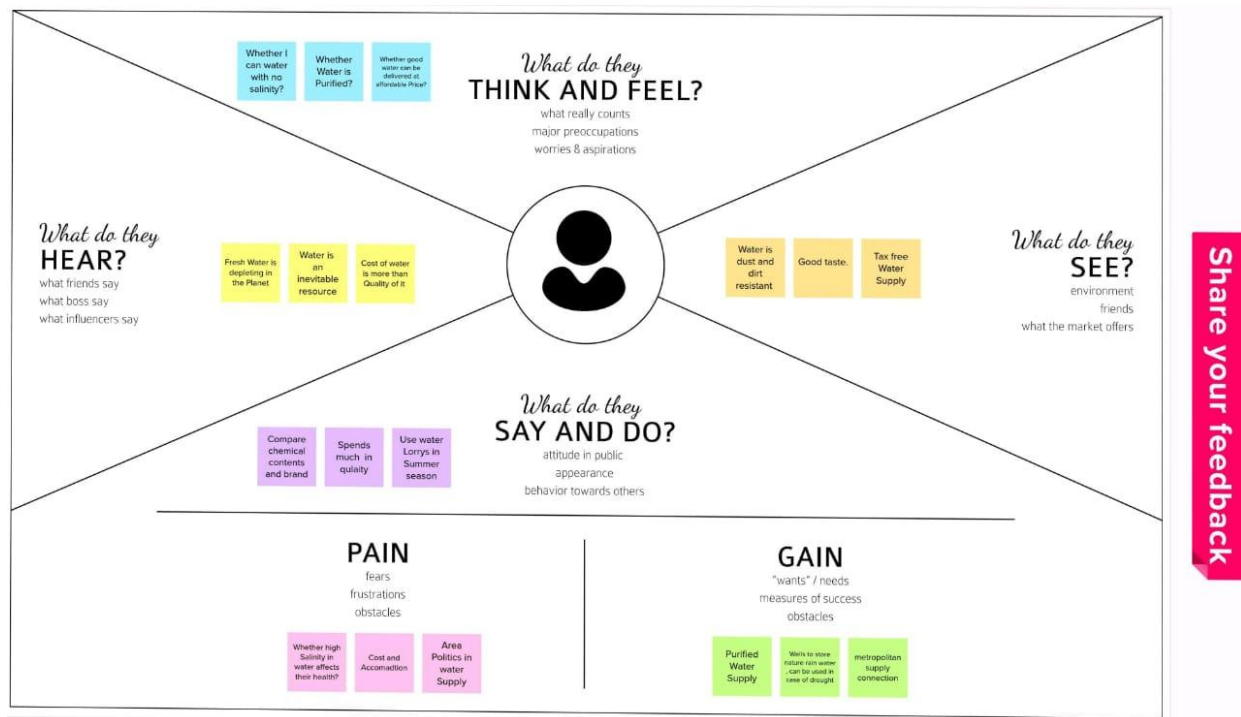
7. Al-Badaii, F.; Shuhaimi-Othman, M.; Gasim, M.B. Water quality assessment of the Semenyih river, Selangor, Malaysia. *J. Chem.* 2013, 2013, 871056.
8. Asadollah, S.B.H.S.; Sharafati, A.; Motta, D.; Yaseen, Z.M. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. *J. Environ. Chem. Eng.* 2021, 9, 104599.
9. Chen, K.; Chen, H.; Zhou, C.; Huang, Y.; Qi, X.; Shen, R.; Liu, F.; Zuo, M.; Zou, X.; Wang, J. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Res.* 2020, 171, 115454.
10. Leros, J.L.; Villarica, M.V. Pattern Extraction of Water Quality Prediction Using Machine Learning Algorithms of Water Reservoir. *Int. J. Mech. Eng. Robot. Res.* 2019, 8, 992–997.
11. Sengorur, B.; Koklu, R.; Ates, A. Water quality assessment using artificial intelligence techniques: SOM and ANN—A case study of Melen River Turkey. *Water Qual. Expo. Health* 2015, 7, 469–490.
12. Aradhana, G.; Singh, N.B. Comparison of Artificial Neural Network algorithm for water quality prediction of River Ganga. *Environ. Res. J.* 2014, 8, 55–63.

2.3 PROBLEM STATEMENT DEFINITION

To predict the water safe or not for Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions.

3.IDEATION AND PROPOSED SOLUTION

3.1 EMPATHY MAP CANVAS



3.2 IDEATION AND BRAINSTORMING

Brainstorm & idea prioritization

Use this template in your own brainstorming sessions so your team can unleash their imagination and start shaping concepts even if you're not sitting in the same room.

⌚ 10 minutes to prepare
🕒 1 hour to collaborate
👥 2-8 people recommended

[Share template feedback](#)

Before you collaborate

A little bit of preparation goes a long way with this session. Here's what you need to do to get going.

⌚ 10 minutes

- A Team gathering**
Define who should participate in the session and send an invite. Share relevant information or pre-work ahead.
- B Set the goal**
Think about the problem you'll be focusing on solving in the brainstorming session.
- C Learn how to use the facilitation tools**
Use the Facilitation Superpowers to run a happy and productive session.

[Open article](#) →

1 Define your problem statement

What problem are you trying to solve? Frame your problem as a How Might We statement. This will be the focus of your brainstorm.

⌚ 5 minutes

PROBLEM

Water is considered as a vital resource that affects various aspects of human health in urban areas. It depends on the factors like Water Usage patterns, land uses and meteorology parameters. So measuring the water Quality using technology helps urban people to use Oblique Water for their daily routine.

2 Key rules of brainstorming
To run an smooth and productive session

- Stay in topic.
- Encourage wild ideas.
- Defer judgment.
- Listen to others.
- Go for volume.
- If possible, be visual.

Need some inspiration?

See a finished version of this template to kickstart your work.

[Open example](#) →

2

Brainstorm

Write down any ideas that come to mind that address your problem statement.

10 minutes

TIP

You can select a sticky note and hit the pencil (switch to sketch) icon to start drawing!



3

Group ideas

Take turns sharing your ideas while clustering similar or related notes as you go. Once all sticky notes have been grouped, give each cluster a sentence-like label. If a cluster is bigger than six sticky notes, try and see if you can break it up into smaller sub-groups.

20 minutes

Studying Dataset



Preventing Chemicals



Satisfying Standards



Conserving nature



About Locations

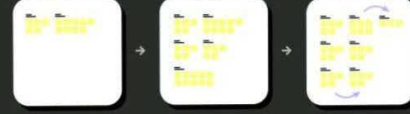


Pipe Supply



TIP

Add customizable tags to sticky notes to make it easier to find, browse, organize, and categorize important ideas as themes within your board.



4

Prioritize

Your team should all be on the same page about what's important moving forward. Place your ideas on this grid to determine which ideas are important and which are feasible.

🕒 20 minutes



→

After you collaborate

You can export the mural as an image or pdf to share with members of your company who might find it helpful.

Quick add-ons

- A Share the mural**
Share a view link to the mural with stakeholders to keep them in the loop about the outcomes of the session.
- B Export the mural**
Export a copy of the mural as a PNG or PDF to attach to emails, include in slides, or save in your drive.

Keep moving forward

- Strategy blueprint**
Define the components of a new idea or strategy.
[Open the template →](#)
- Customer experience journey map**
Understand customer needs, motivations, and obstacles for an experience.
[Open the template →](#)
- Strengths, weaknesses, opportunities & threats**
Identify strengths, weaknesses, opportunities, and threats (SWOT) to develop a plan.
[Open the template →](#)

[Share template feedback](#)



3.3 PROPOSED SOLUTION

S. No	Parameter	Description
1.	Problem Statement (Problem to be solved)	With the rapidly growing urbanization proposal, safe drinking water is a challenge for everyone. Water is being contaminated by several factors. So analysis of water quality and real time monitoring of water is essential.
2.	Idea / Solution description	For Water Quality Index (WQI) prediction several machine learning algorithms have been developed. Using these techniques, our model analyzes the water quality parameters like Alkalinity, pH level, temperature, turbidity, dissolved oxygen, minerals & nutrients (nitrogen, phosphorous). Suitability for the usage of the water for various entities will be deduced based on the WQI calibrated.
3.	Novelty / Uniqueness	In addition to just determining the analysis of water sample, we may perform more processing to determine the water's level of usability and its usage for the appropriate reasons with the help of machine learning.
4.	Social Impact / Customer Satisfaction	Customer satisfaction is an important goal in total quality management. In the recent years water quality level has declined by various pollutants.. Therefore, predicting the water quality is very important in controlling water pollution and providing safe water to the consumers. In order to meet this goal it is necessary to use an evaluation model for measuring the customer satisfaction and used by all

		groups of people in both rural and urban areas.
5.	Business Model (Revenue Model)	The technology and production is improved in business side. It increases the profit and also the logistic way. The revenue model enables the users to find out the harmful effects that can be caused by the water body and also categorises the nearby water bodies for different usage capabilities.
6.	Scalability of the Solution	Scalability of this solution can handle the amount of data collected from water sources to big water bodies and analyze thoroughly in an effective way to instantly serve millions of users.

3.3 PROBLEM SOLUTION FIT

1. CUSTOMER SEGMENT(S)	6. CUSTOMER CONSTRAINTS	5. AVAILABLE SOLUTIONS
<ul style="list-style-type: none">• Private and public laboratories.• Residential and industrial places.• Hotels,restaurants, factories.• Household purposes.	<ul style="list-style-type: none">• Customer has to depend on the testing agencies in order to test the water quality.• The interpretation of result of water quality analysis done by the testing agencies may be trustable or not.• Customers on using a web application to analyze the water quality may require some fundamental prerequisites such as network connection, a system or a mobile.	<ul style="list-style-type: none">• The solution is to have information on water quality parameters like pH level,Temperature, Turbidity,Minerals etc, to analyze the quality of water.• It is possible to find the Water quality index(WQI) and Water quality class(WQC)

2. PROBLEMS / PAINS	9. ROOT/CAUSE	7. BEHAVIOUR
<ul style="list-style-type: none"> • Check the quality of water by gathering information based on many features and qualities in the chemical and physical composition of nature. • Customer can check the water quality without expert's support. • Check the usability of water. 	<ul style="list-style-type: none"> • Poor quality water is one of the major factors of escalation of many diseases. • Rapid urbanization has led to the deterioration of water quality at an increasing rate. • All living things are harmed by improper maintenance of rainwater and surface water contaminated by industrial waste. 	<ul style="list-style-type: none"> • The study attempts to assess the users water behavior using available resources,prevailing socio economic conditions and personal aspects of users.The research work suggests the need for ensuring the water quality. • Customers must have knowledge about the water quality in order for machine learning models to accurately anticipate the water quality.

<p>3.TRIGGERS TO ACT</p> <p>To enhance the standard of living in terms of health aspects by providing good quality water in order to reduce the water borne diseases .</p>	<p>10. YOUR SOLUTION</p> <p>To build an effective and efficient water quality prediction system for all kinds of water samples using Classification algorithms of Machine Learning provides a better and easy interpretation of water samples by using the past historical data of water for prediction and analysis so that the people with no prior knowledge can understand the results of analysis process and can be made available at anytime and at anyplace.</p>	<p>8. CHANNELS OF BEHAVIOUR</p> <p>ONLINE:</p> <ul style="list-style-type: none"> Through Advertising in social media, news platform makes customer to know and realize the importance of monitoring the level of water quality. Customers can make use of web applications to process the data.
<p>4.EMOTIONS</p> <p>BEFORE: Without appropriate technology to analyze the water quality, lead to various diseases.</p> <p>AFTER: Now it is easy to evaluate the quality of water with the help of this application.</p>		<p>OFFLINE:</p> <ul style="list-style-type: none"> To attain standard quality of water by analyzing the water parameters

4. REQUIREMENT ANALYSIS

4.1 FUNCTIONAL REQUIREMENT

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User registration	Registration through Gmail Create an account Follow the instructions
FR-2	User Confirmation	Confirmation via Email and it is predicted by water level sensor
FR-3	Interface sensor	Interface sensor and Water level sensor produces the detection of clean drinking water
FR-4	Accessing datasets	Datasets are collected by data preprocessing method.
FR-5	Mobile application	The efficient of water quality is analyzed, the mobile application is not used .

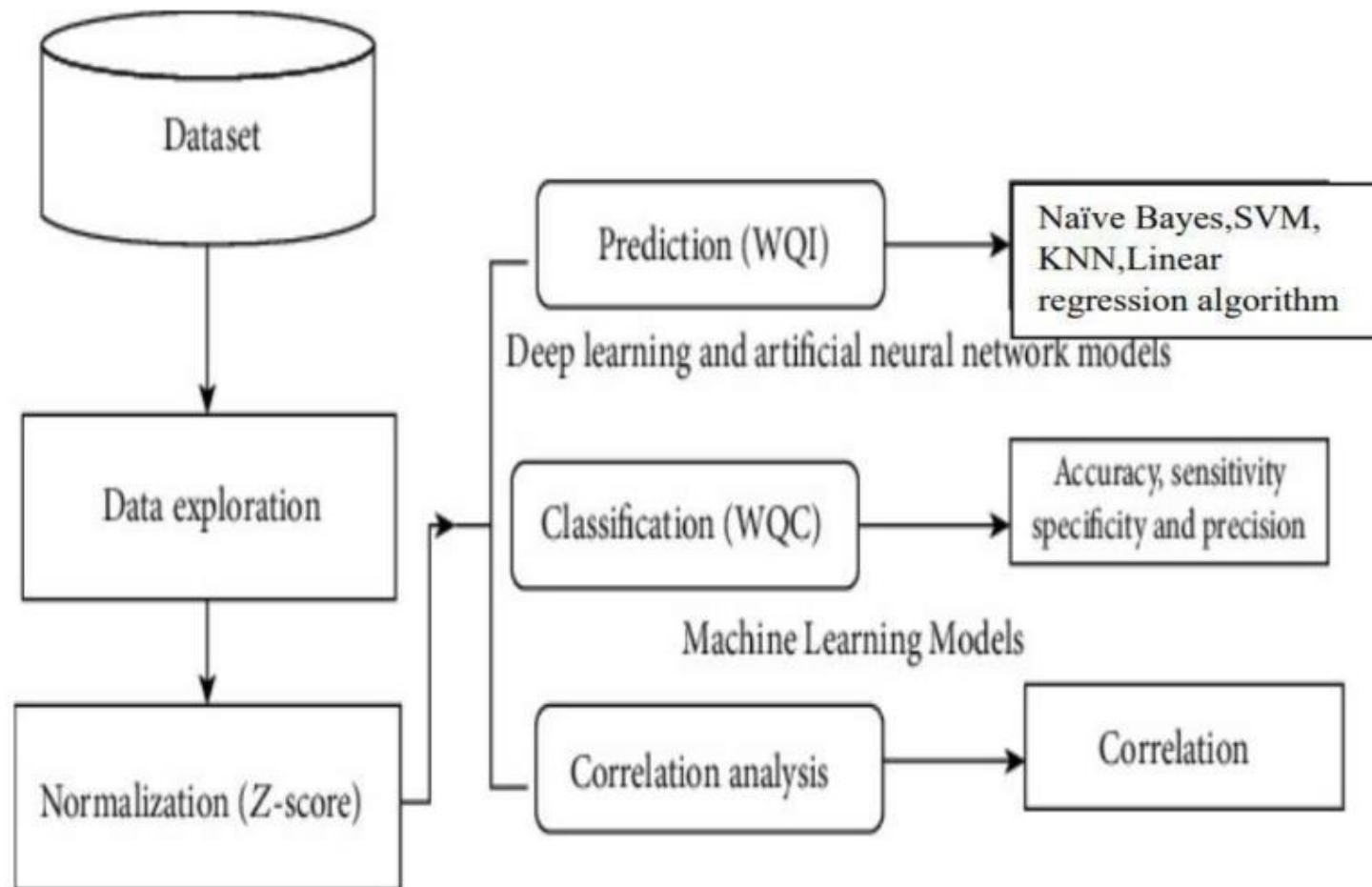
4.2 NON-FUNCTIONAL REQUIREMENT

FR No.	Non-Functional Requirement	Description
NFR-1	Usability	This project is useful for all human being by predicting a purified water.
NFR-2	Security	We have designed this project to secure the people from drinking the impurity water.
NFR-3	Reliability	This project will help everyone in protecting their health. Accurate water quality prediction is the basis of water environment management and is of great significance for water environment protection.
NFR-4	Performance	This system uses different sensors for monitoring the water quality by determine pH,Turbidity,conductivity and temperature. The data preprocessing access the dataset. With the use of this we predict the quality water.
NFR-5	Availability	By developing and deploying resilient hardware and software we can analyze the drinking water .
NFR-6	Scalability	This project used to measure and determine the quality of water. This provide pollution free and purified water.

5.PROJECT DESIGN

5.1 DATA FLOW DIAGRAM

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.



5.2 SOLUTION AND TECHNICAL ARCHITECTURE

There are basically 10 steps for making our model predict the water quality of the water samples. Those steps are:-

A. Problem Identification

In this step, we identify the problem which is solved by our model. So the problem to be solved by our model is water quality prediction using a dataset.

B. Data Extraction:-

In this, we extract the data from the internet to train our data and predict the water quality. So for that, we take the CPCB(Central Pollution Control Board India) dataset which contains 3277 instances of 13 different wellsprings which are collected between 2014 to 2020.

C. Data Exploration:-

In this step, we analyze the data visually by comparing some parameters of water with the WHO standards of water. It gives a slight overview of the data.

D. Data Cleaning

In this step, we clean that data like if there are some missing values in it so we replace them with mean and remove noise from the data..

F. Data Selection

In this step, we select the data types and source of the data. The essential goal of data selection is deciding fitting data type, source, and instrument that permit agents to respond to explore questions sufficiently

G. Data Splitting

In this step, we divide the dataset into smaller subsets for easing the complexity. Normally, with a two-section split, one section is utilized to assess or test the information and the other to prepare the model.

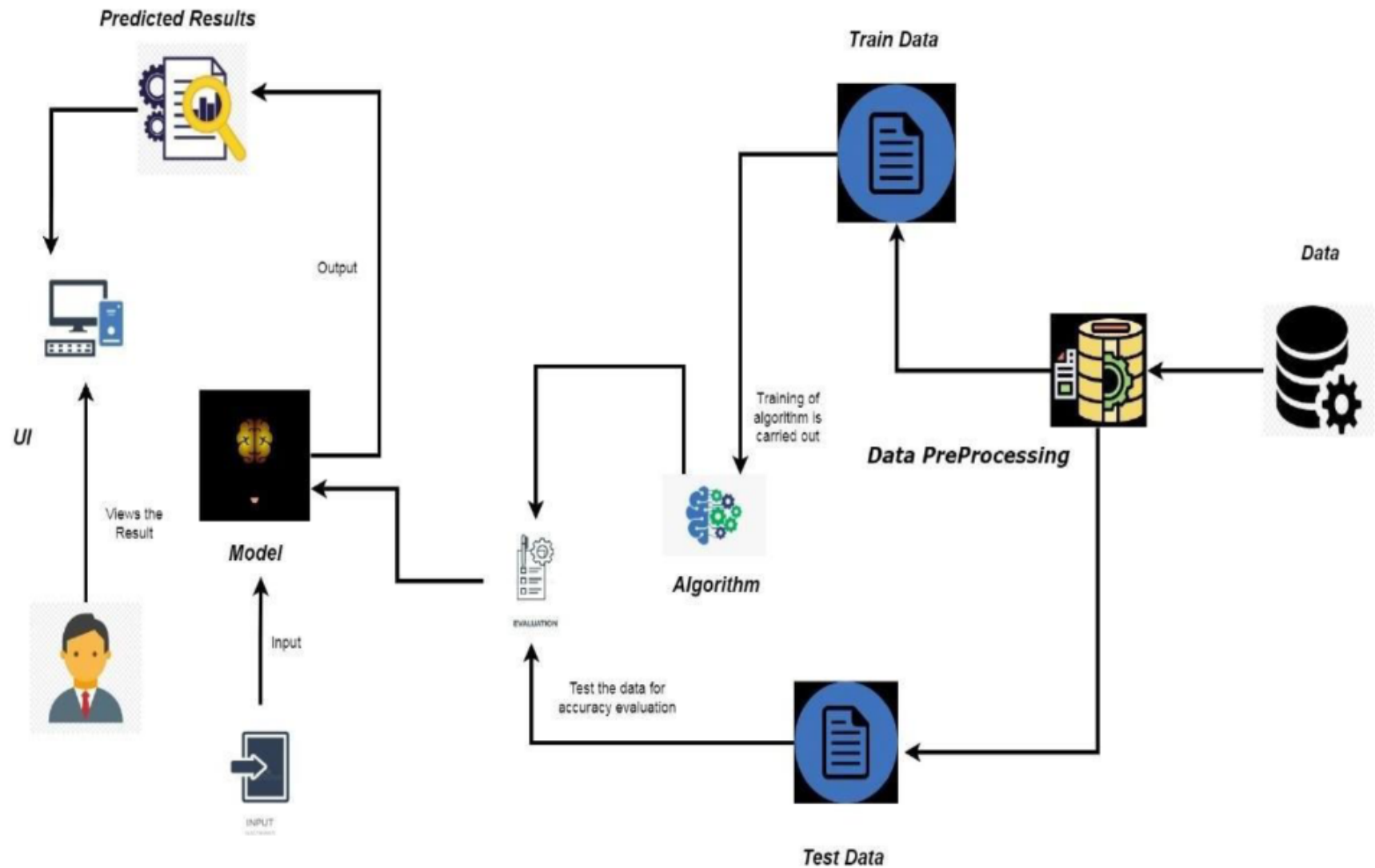
H. Data Modeling

In this step, we create a graph of the dataset for visual representation of data for better understanding. A Data Model is this theoretical model that permits the further structure of conceptual models and to set connections between data.

I. Model Evaluation

Model Evaluation is a fundamental piece of the model improvement process. In this step, we evaluate our model and check how well our model do in the future.

SOLUTION ARCHITECTURE



5.3 USER STORIES

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Data Collection	USN-1	Collect the appropriate dataset for predicting the water quality.	10	High	Nihil Rengamasy T, Manoj Kumar M
Sprint-1		USN-2	Data Preprocessing – Used to transform the data into useful format.	7	Medium	Keerthi Rajan S, Minu S
Sprint-2	Model Building	USN-3	Calculate the Water Quality Index (WQI) using Regression algorithm of Machine Learning.	10	High	Nihil Rengasamy T, Minu S

Sprint-2		USN-4	Splitting the Model into Training and Testing from the overall dataset.	7	Medium	Manoj Kumar M, Keerthi Rajan S
Sprint-3	Training and Testing	USN-5	Train the Model using Regression algorithm and Testing the Performance of the model.	10	High	Minu S, Keerthi Rajan S
Sprint-4	Implementation of the Application	USN-6	Predict the Water Quality Index (WQI) and recommend the appropriate purification technique.	10	High	Nihil Rengasamy T, Manoj Kumar M
Sprint-4		USN-7	Deploy the Model on IBM Cloud.	7	Medium	Keerthi Rajan S, Nihil Rengasamy T

6. PROJECT PLANNING AND SCDULING

6.1 SPRINT PLANNING AND ESTIMATION

Product Backlog, Sprint Schedule, and Estimation:

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority
Sprint-1	Data Preparation	USN-1	Collecting water dataset and pre-processing it	20	High
Sprint-2	Model Building	USN-2	Create an ML model to predict water quality	5	Medium
Sprint-2	Model Evaluation	USN-3	Calculate the performance, error rate, and complexity of the ML model and evaluate the dataset based on the parameter that the dataset consists of.	5	Medium
Sprint-2	Model Deployment	USN-4	As a user, I need to deploy the model and need to find the results.	10	Medium
Sprint-3	Web page (Form)	USN-5	As a user, I can use the application by entering the water dataset to analyze or predict the results.	20	Medium
Sprint-4	Dashboard	USN-6	As a user, I can predict the water quality by clicking the submit button and the application will show whether the water is efficient for use or not.	20	High

6.2 SPRINT SCHEDULE

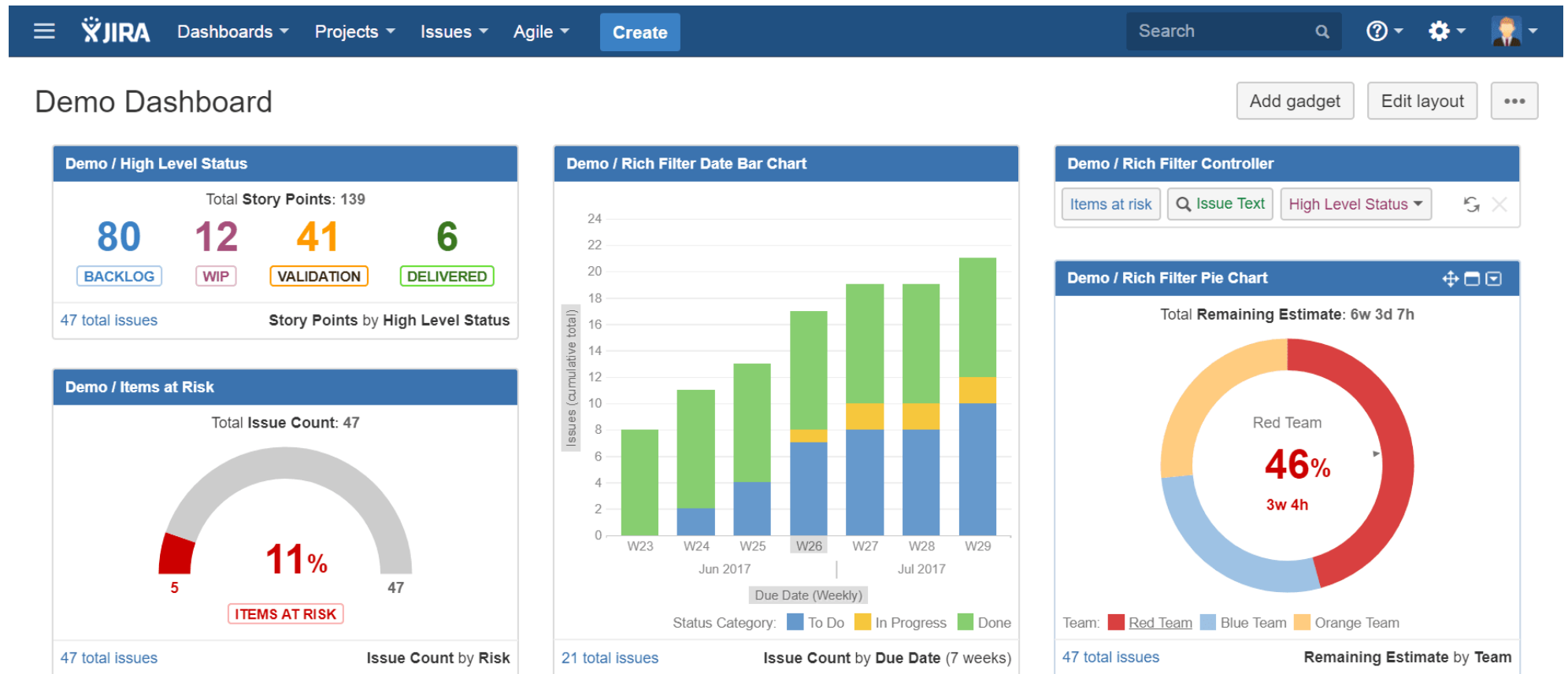
Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	10	6 Days	24 Oct 2022	29 Oct 2022	8	29 Oct 2022
Sprint-2	10	6 Days	31 Oct 2022	05 Nov 2022	7	05 Nov 2022
Sprint-3	10	6 Days	07 Nov 2022	12 Nov 2022	8	12 Nov 2022
Sprint-4	10	6 Days	14 Nov 2022	19 Nov 2022	7	19 Nov 2022

Velocity:

Imagine we have a 6 -day sprint duration, and the velocity of the team is 10 (points per sprint). Let's calculate the team's average velocity (AV) per iteration unit (story points per day).

$$AV = \frac{\text{sprint duration}}{\text{velocity}} = 6/10=0.6$$

6.3 REPORTS FROM JIRA



7. CODING AND SOLUTIONS

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

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

Data Preprocessing

The processing phase is very important in data analysis to improve the data quality. In this phase, the WQI has been calculated from the most significant parameters of the dataset. Then, water samples have been classified on the basis of the WQI values. For obtaining superior accuracy, the -score method has been used as a data normalization technique.

Null Values are checked and replaced with median value

```
df.isnull().sum()
```

```
ph          491
Hardness     0
Solids       0
Chloramines  0
Sulfate     781
Conductivity 0
Organic_carbon 0
Trihalomethanes 162
Turbidity    0
Potability   0
dtype: int64
```

```
for feature in df.columns:
    if df[feature].isnull().sum()>0:
        print(f"{feature} : {round(df[feature].isnull().mean(),4)*100}%")
```

```
ph : 14.99%
Sulfate : 23.84%
Trihalomethanes : 4.95%
```

```
## Fill missing values with median
for feature in df.columns:
    df[feature].fillna(df[feature].median() , inplace = True)
```

```
## find duplicate rows in dataset
duplicate = df[df.duplicated()]
duplicate
```

ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
----	----------	--------	-------------	---------	--------------	----------------	-----------------	-----------	------------

Feature Engineering:

Removing Outliers using outlier Technique:

```
# removing outliers
Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)
IQR = Q3 - Q1
print(IQR)
```

```
ph                1.592377
Hardness          39.816918
Solids            11666.071830
Chloramines        1.987466
Sulfate           33.291119
Conductivity      116.057890
Organic_carbon     4.491850
Trihalomethanes    20.018954
Turbidity          1.060609
Potability         1.000000
dtype: float64
```

```
df = df[~((df < (Q1 - 1.5 * IQR)) | (df > (Q3 + 1.5 * IQR))).any(axis=1)]
df.shape
```

```
(2666, 10)
```

Water Quality Index Calculation

To measure water quality, WQI is used to be calculated using various parameters that significantly affect WQ [40–42]. In this study, a published dataset is considered to test the proposed model, and seven significant water quality parameters are included. The WQI has been calculated using the following formula:

$$WQI = \frac{\sum_{i=1}^N q_i \times w_i}{\sum_{i=1}^N w_i},$$

where: N is the total number of parameters included in the WQI calculations, q_i is the quality rating scale for each parameter calculated by equation (2) below, and w_i is the unit weight for each parameter calculated by equation (3).

$$q_i = 100 \times \left(\frac{V_i - V_{Ideal}}{S_i - V_{Ideal}} \right),$$

where: V_i is the measured value of parameter in the tested water samples, V_{Ideal} is the ideal value of parameter in pure water (0 for all parameters except pH), and S_i is the recommended standard value of parameter (as shown in Table 1)

$$w_i = \frac{K}{S_i},$$

7.2 FEATURE 2

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

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

The screenshot shows a Jupyter Notebook in Visual Studio Code. The notebook is titled "Water_quality.ipynb" and is part of "Sprint 4". The code in the notebook is as follows:

```
[31] # Support vector classifier
from sklearn.svm import SVC
svc_classifier = SVC(class_weight = "balanced" )
svc_classifier.fit(X_train_final, y_train)
y_pred_scv = svc_classifier.predict(X_test_final)
accuracy_score(y_test, y_pred_scv)
```

The output of the code is 0.6225.

```
[32] print(classification_report(y_test, y_pred_scv))
```

The output of the classification report is as follows:

	precision	recall	f1-score	support
0	0.70	0.69	0.70	497
1	0.50	0.50	0.50	303
accuracy			0.62	800
macro avg	0.60	0.60	0.60	800
weighted avg	0.62	0.62	0.62	800

The SVM model was developed in 1995 by Corinna Cortes and Vapnik. It has several unique benefits in solving small samples, and nonlinear and high-dimensional pattern recognition. It can be extended to function in the simulation of other machine learning problems. It uses the hyperplane to separate the points of the input vectors and finds the needed coefficients. The best hyperplane is the line with the largest margin, which is meant the distance between the hyperplane and the nearest input objects. The input points defined in the hyperplane are called *support vectors*. In this work, the linear SVM model along with the Gaussian radial basis function (equation (17)) is used to classify the tested water samples based on their quality.

Hyper Parameter tuning with Support Vector Machines(SVM):

```
# defining parameter range
param_grid = {'C': [0.1, 1, 10, 100, 200, 400, 600, 800],
              'gamma': [1, 0.1, 0.01, 0.001, 0.0001],
              'kernel': ['rbf']}
```

```
from sklearn.model_selection import GridSearchCV
```

```
grid = GridSearchCV(SVC(), param_grid, refit = True, verbose = 3)
```

```
# fitting the model for grid search
grid.fit(X_train_final, y_train)
```

Output exceeds the [size limit](#). Open the full output data [in a text editor](#)

Fitting 5 folds for each of 40 candidates, totalling 200 fits

```
[CV 1/5] END .....C=0.1, gamma=1, kernel=rbf;; score=0.628 total time= 0.5s
[CV 2/5] END .....C=0.1, gamma=1, kernel=rbf;; score=0.630 total time= 0.3s
[CV 3/5] END .....C=0.1, gamma=1, kernel=rbf;; score=0.630 total time= 0.3s
[CV 4/5] END .....C=0.1, gamma=1, kernel=rbf;; score=0.630 total time= 0.3s
[CV 5/5] END .....C=0.1, gamma=1, kernel=rbf;; score=0.627 total time= 0.3s
[CV 1/5] END .....C=0.1, gamma=0.1, kernel=rbf;; score=0.628 total time= 0.2s
[CV 2/5] END .....C=0.1, gamma=0.1, kernel=rbf;; score=0.630 total time= 0.2s
```



```
▸ GridSearchCV
▸ estimator: SVC
  ▸ SVC
```

```
# print best parameter after tuning
print(grid.best_params_)

# print how our model looks after hyper-parameter tuning
print(grid.best_estimator_)
```

```
{'C': 100, 'gamma': 0.01, 'kernel': 'rbf'}
SVC(C=100, gamma=0.01)
```

```
# Support vector classifier
from sklearn.svm import SVC
svc_classifier = SVC(class_weight = "balanced" , C=100, gamma=0.01)
svc_classifier.fit(X_train_final, y_train)
y_pred_scv = svc_classifier.predict(X_test_final)
accuracy_score(y_test, y_pred_scv)
```

```
0.6325
```

```
print(classification_report(y_test, y_pred_xgb))
```

	precision	recall	f1-score	support
0	0.67	0.77	0.72	497
1	0.50	0.38	0.43	303
accuracy			0.62	800
macro avg	0.59	0.57	0.57	800
weighted avg	0.61	0.62	0.61	800

8.TESTING

8.1 Home Page

← → ↻ 127.0.0.1:5000 ☆

Water Quality Prediction

Enter pH value	Enter Hardness
Enter Solids	Enter Chloramines
Enter Sulfate	Enter Conductivity
Enter Organic_carbon	Enter Trihalomethanes
Enter Turbidity	

Predict

8.1.1 TEST CASE - 1

← → ↻ 127.0.0.1:5000/predict ☆

Water Quality Prediction

1	323
2	8
5	753
4	100
4	

Predict

water is safe for human consumption

8.1.2 TEST CASE - 2

← → ↻ 127.0.0.1:5000/predict ☆

Water Quality Prediction

14	290
61227	8
481	4
18	124
6	

Predict

water is not safe for human consumption

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 [ProductName] project at the time of the release to User Acceptance Testing (UAT).

2. Defect Analysis

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

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

3. Test Case Analysis

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

Section	Total Cases	Not Tested	Fail	Pass
Print Engine	7	0	0	7

Client Application	51	0	0	51
Security	2	0	0	2
Outsource Shipping	3	0	0	3
Exception Reporting	9	0	0	9
Final Report Output	4	0	0	4
Version Control	2	0	0	2

9.RESULT

9.1 PERFORMANCE METRICS

For validating the developed model, the dataset has been divided into 70% training and 30% testing subsets. The SVM, Xgboost, and Random Forest were utilized for the water quality classification prediction

SO, WE ARE GOING TO USE SVC

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

Accuracy = $\frac{TP+TN}{TP+FP+FN+TN}$

SN.	Algorithm	Type	ACCURACY	Precision	Recall f1-Score
1	RANDOM FOREST	58.5	0.42	0.38	0.40
2	XGBOOST	61.7	0.43	0.12	0.18

Table 1. Comparison of algorithms SN.

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

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.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 researchers to have certain professional background knowledge.

To overcome the above-mentioned challenges, the following aspects should be considered in future research and engineering practices: (1) More advanced sensors, including soft sensors, should be developed and applied in water quality monitoring to collect sufficiently accurate data to facilitate the application of machine learning approaches. (2) The feasibility and reliability of the algorithms should be improved, and more universal algorithms and models should be developed according to the water treatment and management requirements. (3) Interdisciplinary talent with knowledge in different fields should be trained to develop more advanced machine learning techniques and apply them in engineering practices.

13. APPENDIX REQUIREMENT.TXT

```
Flask == 2.2.2
joblib == 1.2.0
numpy == 1.23.4
pandas == 1.5.1
scikit-learn == 1.1.3
xgboost == 1.7.1
gunicorn == 20.1.0
matplotlib == 3.6.2
seaborn == 0.12.1
gevent
requests
flask-cors==3.0.10
```

APP.PY

```
app.py > ...
1  from flask import Flask, request, render_template
2  import pickle
3  import pandas as pd
4  import numpy as np
5  import joblib
6  scaler = joblib.load("my_scaler.save")
7  app = Flask(__name__)
8  model=pickle.load(open('model.pkl','rb'))
9
10 @app.route("/home")
11 @app.route("/")
12 def hello():
13     return render_template("predict.html")
14
15 @app.route("/predict", methods = ["GET", "POST"])
16 def predict():
17     if request.method == "POST":
18         input_features = [float(x) for x in request.form.values()]
19         features_value = [np.array(input_features)]
20
21         feature_names = ["ph", "Hardness" , "Solids", "Chloramines", "Sulfate",
22                         "Conductivity", "Organic_carbon","Trihalomethanes", "Turbidity"]
23
24         df = pd.DataFrame(features_value, columns = feature_names)
25         df = scaler.transform(df)
26         output = model.predict(df)
27
28         if output[0] == 1:
29             prediction = "safe"
30         else:
31             prediction = "not safe"
32
33
```

Ln 14, Col 1 Spaces: 4 UTF-8 CRLF Python 3.10.8 64-bit (microsoft store)

WATER QUALITY.IPYNB

File Edit Selection View Go Run Terminal Help

Water_quality.ipynb - Sprint 4 - Visual Studio Code

app.py predict.html Water_quality.ipynb style.css

Water_quality.ipynb > Problem Statement > EDA and Feature Engineering > for feature in df.columns:

+ Code + Markdown | Run All | Clear Outputs of All Cells | Restart | Variables | Outline | Python 3.10.8

Support vector Machine

```
[31] # Support vector classifier
from sklearn.svm import SVC
svc_classifier = SVC(class_weight = "balanced" )
svc_classifier.fit(X_train_final, y_train)
y_pred_scv = svc_classifier.predict(X_test_final)
accuracy_score(y_test, y_pred_scv)
```

0.6225

```
[32] print(classification_report(y_test, y_pred_scv))
```

		precision	recall	f1-score	support
	0	0.70	0.69	0.70	497
	1	0.50	0.50	0.50	303
	accuracy			0.62	800
	macro avg	0.60	0.60	0.60	800
	weighted avg	0.62	0.62	0.62	800

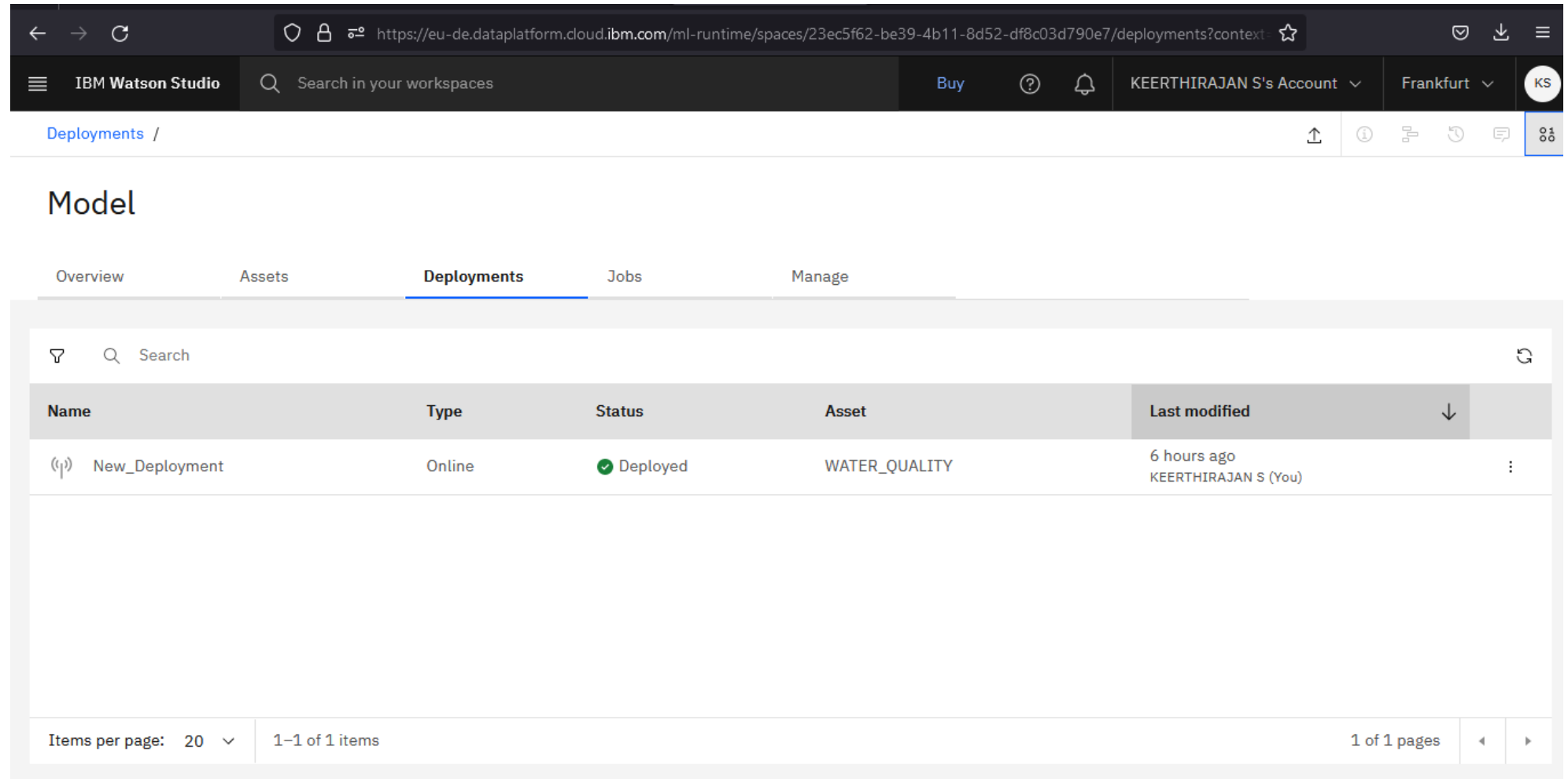
> OUTLINE
> TIMELINE

0 0 0

Jupyter Server Local Cell 19 of 52

The image shows a Visual Studio Code editor window with a dark theme. The top bar displays the menu (File, Edit, Selection, View, Go, Run, Terminal, Help) and the active file (predict.html - Sprint 4 - Visual Studio Code). The Explorer sidebar on the left shows a project structure for 'SPRINT 4' with files like 'style.css', 'predict.html', 'app.py', and 'Water_quality.ipynb'. The main editor area displays the 'predict.html' file, which contains HTML code for a 'water quality prediction' form. The form includes input fields for pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, and Turbidity, along with a 'Predict' button. The code uses Bootstrap classes for styling and includes links to external CSS and JavaScript files. The status bar at the bottom indicates the current line is 28, column 8, with 4 spaces, using UTF-8 encoding and CRLF line endings.

CLOUD DEPLOYMENT STATUS





The screenshot shows the IBM Watson Studio interface. The top navigation bar includes the IBM Watson Studio logo, a search bar, and user account information for KEERTHIRAJAN S. The main content area is titled 'Model' and has tabs for Overview, Assets, Deployments (selected), Jobs, and Manage. Below the tabs is a table of deployments. The table has columns for Name, Type, Status, Asset, and Last modified. One deployment is listed: 'New_Deployment' with Type 'Online', Status 'Deployed', Asset 'WATER_QUALITY', and Last modified '6 hours ago' by 'KEERTHIRAJAN S (You)'. The bottom of the page shows pagination controls: 'Items per page: 20', '1-1 of 1 items', and '1 of 1 pages'.

Deployments /

Model

Overview Assets **Deployments** Jobs Manage

Name	Type	Status	Asset	Last modified
 New_Deployment	Online	 Deployed	WATER_QUALITY	6 hours ago KEERTHIRAJAN S (You)

Items per page: 20 1-1 of 1 items 1 of 1 pages

LINKS:

GITHUB - <https://github.com/IBM-EPBL/IBM-Project-19850-1659707961>

IBM CLOUD - <https://eu-de.ml.cloud.ibm.com/ml/v4/deployments/496f9619-052e-4340-897b-ef55442970f2/predictions?version=2022-11-19>

VIDEO LINK - <https://youtu.be/y8M4dcFnFqs>