Efficient Water Quality Analysis And Prediction Using Machine Learning

1. INTRODUCTION:

1.1 Project Overview:

Water is an essential resource for human existence. In recent years, water pollution has become a serious problem affecting water quality. 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. Using machine learning algorithms to develop a model that is capable of predicting the water quality index and then the water quality class. The method we propose is based on four water parameters: temperature, pH, turbidity and coliforms. The use of the multiple regression algorithms has proven to be important and effective in predicting the water quality index. In addition, the adoption of the artificial neural network provides the most highly efficient way to classify the water quality.

1.2 Purpose:

Water quality modeling helps people understand the eminence of water quality issues and models provide evidence for policy makers to make decisions In order to properly mitigate water.

2. LITERATURE SURVVEY:

2.1 Existing Problem:

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.

2.2 References:

1. Water Quality Prediction Based on Machine Learning Techniques - Zhao Fu, Cheng Mei Yang, Jacimaria Batista and Yingtao Jiang, published in January 2020.

2. Efficient Water Quality Prediction Using Supervised Machine Learning - Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Asad A. Shah, Rabia Irfan and José García-Nieto, published in October 2019.

3. Evaluation of Multivariate Linear Regression And Artificial Neural Networks in Prediction of Quality Parameters - Hamid Zare Abyaneh, published in January 2014.

4. Prediction of Water Quality System for Aquaculture using Machine Learning - Kiran babu T S, Manoj Challa, published in June 2019.

5. Water Quality Prediction Models Based on Machine Learning - Rongli Gai, Jiahui Yang, published in May 2022.

6. Water quality prediction based on Naive Bayes algorithm - M. Ilic, Z.Srdjevic, B.Srdjevic, published in January 2022.

7. Multi-task learning framework for predicting water quality using non-linear machine learning technique - D.Senthilkumar, D.George Washington,

A.K.Reshmy, M. Noornisha, published in April 2022.

8. Water Pollution Prediction Based on Deep Belief Network in Big Data of Water Environment Monitoring - Li Liang, published in December 2021.

9. A study on water quality prediction by a hybrid CNNLSTM model with attention mechanism - Yurong Yang, Qingyu Xiong, Chao Wu, Qinghong Zou1, Yang Yu1, Hualing Yi1, Min Gao, published in June 2021.

10. Smart Urban Water Quality Prediction System Using Machine Learning - Bharath Singh J, Nirmitha S, Kaviya S S, published in August 2021.

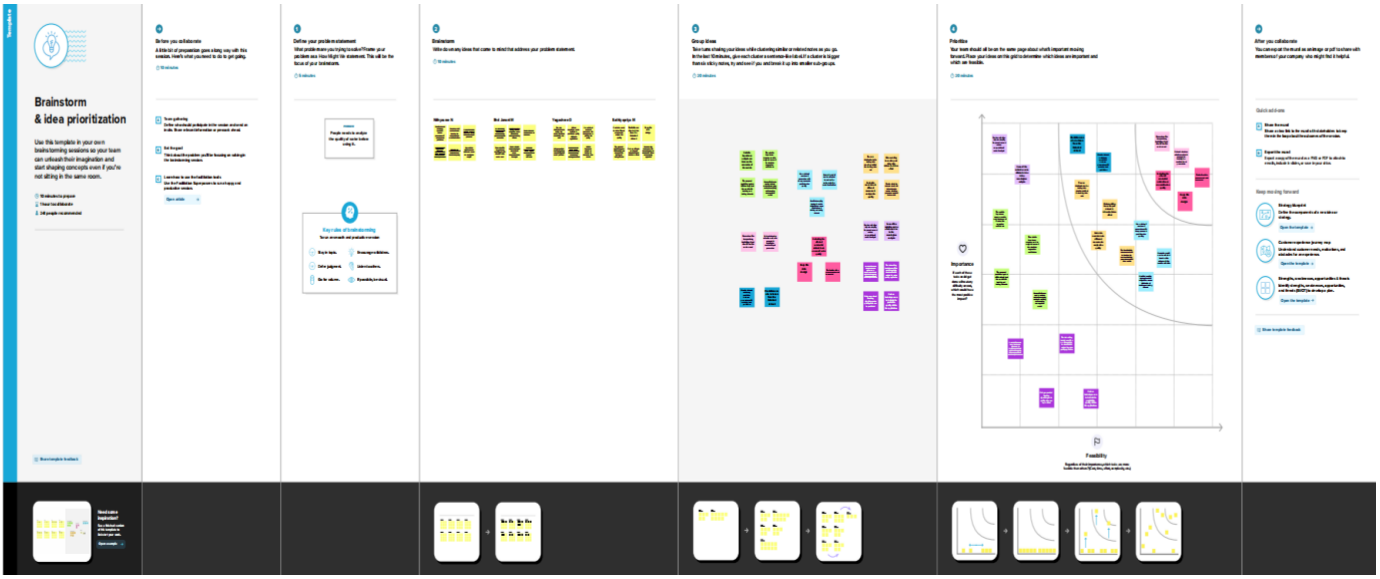
2.3 Problem Statement Definition:

The water level is predicted in an hourly manner to ensure the growth and survival of aquatic life. The web application is built using Flask to alert the user to critical situations. The impact of water parameter changes can be effectively treated if the information is analyzed and water quality is expected ahead of time.

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas:

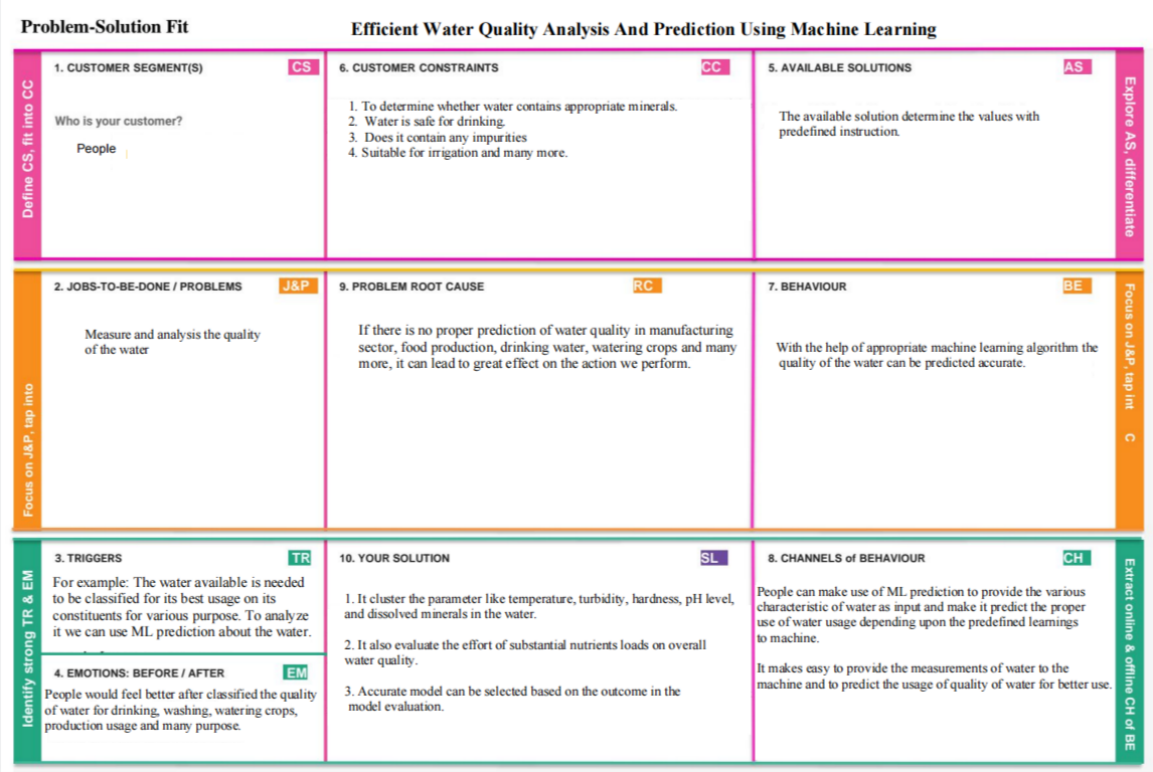
3.2 Ideation & Brainstorming:



3.3 Proposed Solution:

S.No.	Parameter	Description
1.	Problem Statement (Problem to besolved)	People need to analyse the quality of water before using it for variouspurpose.
2.	Idea / Solution description	Massive dataset and strong correlation between parameters will make the best prediction.
3.	Novelty / Uniqueness	Accurate model can be selected based on the outcome in the model evaluation.
4.	Social Impact / Customer Satisfaction	Helps people to better categorise the available water for various usage depending upon the analysis for which water conservation can be practised.
5.	Business Model (Revenue Model)	Machine Learning model can be implemented which is capable of integrating linear and non-linear relationships in dataset.
6.	Scalability of the Solution	Feature selection helps to simplify the procedure and reduce computational cost of analysis.

3.4 Problem Solution fit:



4. REQUIREMENT ANALYSIS

4.1 Functional requirement:

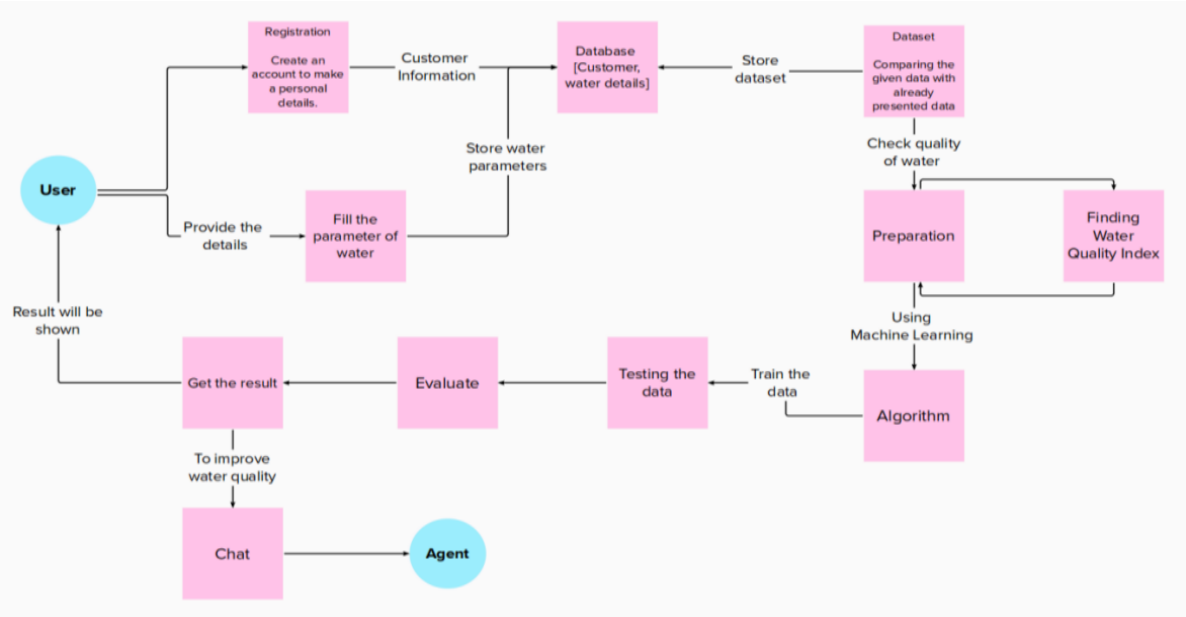
FR No.	Functional Requirement (Epic)	Sub Requirement (Story/ Sub-Task)
FR-1	User Registration	Registration through Gmail and Form.
FR-2	User Confirmation	Confirmation via OTP.
FR-3	User Problem	Efficient water quality analysis and prediction with the list of parameters.
FR-4	Solution by Agent	Issue is solved by agent via email chats.
FR-5	Default solution	Frequent solution to the problem is displayed.

4.2 Non-Functional requirements:

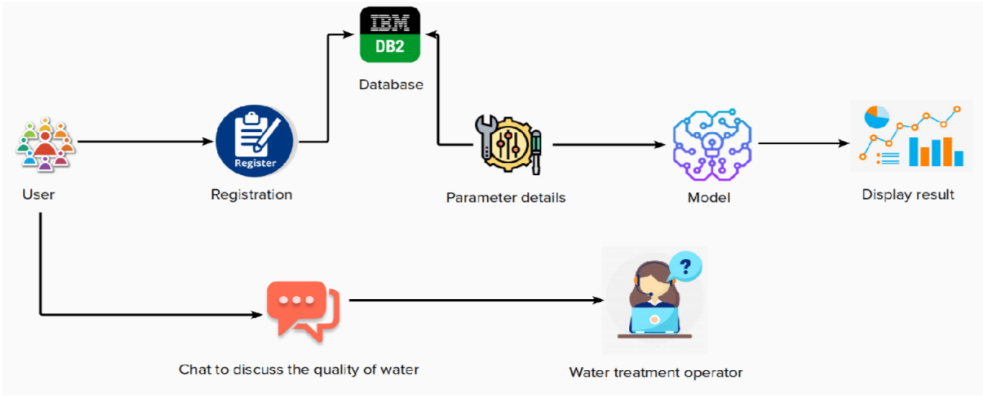
FR No.	Non-Functional Requirement	Description
NFR-1	Usability	User friendly interface to provide the inputs.
NFR-2	Security	It is secured as each process is verified by using mail.
NFR-3	Reliability	Analysis can be used for various purposes.
NFR-4	Performance	The analysis is always accurate to the mark.
NFR-5	Availability	Analysis can be made at any time in need.
NFR-6	Scalability	It is highly scalable due its data backupmaintained.

5. PROJECT DESIGN

5.1 Data Flow Diagrams:



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)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	I can access my account /dashboard	High	
	Confirmation	USN-2	As a user, I will receive confirmation emai lonce I have registered for the application	I can receive confirmationemail & clickconfirm	High	
	Verification	USN-3	As a user, I can register for the applicationthrough OTP message or email	I can receive OTPandprovide for verification	Low	
	Parameter Passing	USN-4	As a user, I can provide values for variousparamete rs of waterquality	I can choose therequired parameter for input	High	
	Predicting	USN-5	Using ML algorithm, predictions are madeusing the parameter provided	Algorithm makes prediction using paramete rs	Medium	
	Result	USN-6	Quality of the water is determined.	Result will be displayed	High	
External Agent	Solution Providing	USN-1	Better waterusage ideas are provided basedon the quality of water	Based on the result,ideasare provided	Medium	

6. PROJECT PLANNING AND SCHEDULING

6.1 Sprint Planning & Estimation:

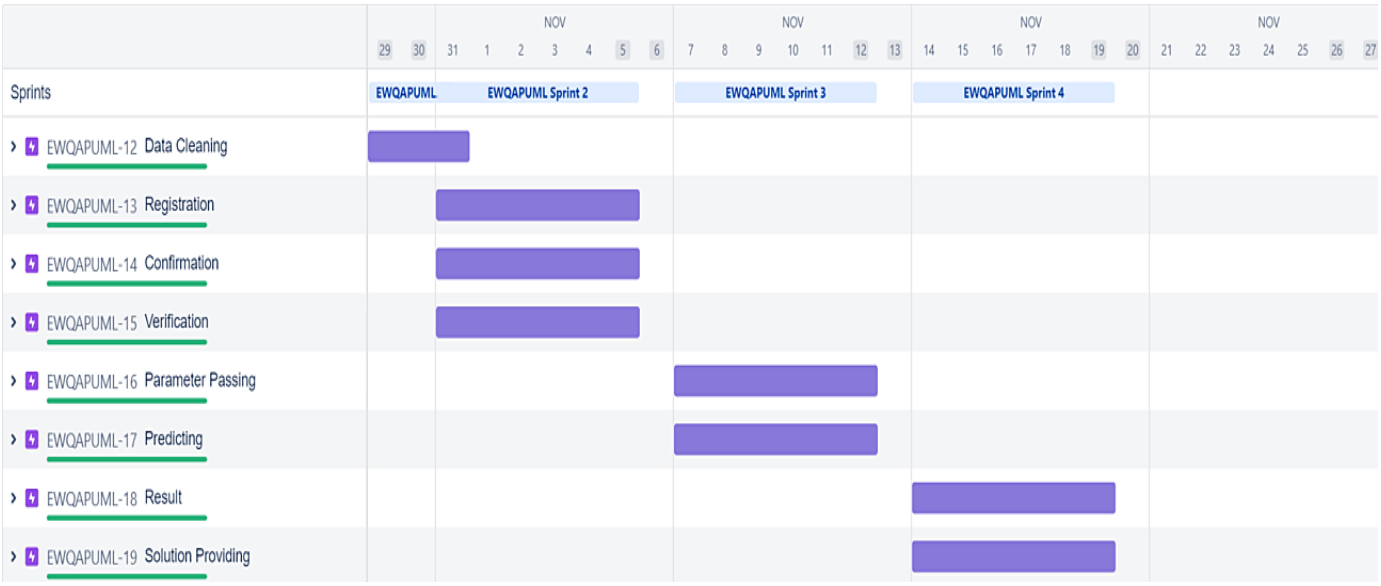
Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-2	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	2	High	Yogashree. D
Sprint-2	Confirmation	USN-2	As a user, I will receive confirmation emailonce I have registered for the application	1	High	Shri Janani. M
Sprint-2	Verification	USN-3	As a user, I can register for the application through OTP message or email	2	Low	Sathiyapriya. M
Sprint-3	Parameter Passing	USN-4	As a user, I can provide values for various parameters of water quality	1	High	Nithyasree. N
Sprint-3	Predicting	USN-5	Using ML algorithm, predictions are madeusing the parameter provided	2	Medium	Sathiyapriya. M
Sprint-4	Result	USN-6	Quality of the water is determined	1	High	Nithyasree. N
Sprint-1	Data Cleaning	USN-5	Removing the null values and outliers from thedata	1	Low	Yogashree. D

Sprint-1	Data Pre-processingAnd Model Building	USN-5	Scaling thedata and training the model withthe data	3	High	Nithyasree. N
Sprint-4	Solution Providing	USN-1	Better waterusage ideas are provided basedon the quality of water	1	Medium	Shri Janani. M

6.2 Sprint Delivery Schedule:

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on PlannedEnd Date)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022	20	4 Nov2022
Sprint-2	20	6 Days	31 Oct 2022	05 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:



7. CODING & SOLUTIONING

7.1 Feature 1: Data Collection:

In this feature the required packages are imported and along with that the data in the dataset is read using pandas.

```
data=pd.read_csv("C:/Users/Jothy Natarajan/Downloads/water_dataX.csv",encoding= 'ISO-8859-1',low_memory=False)
```

```
data.head(10)
```

	STATION CODE	LOCATIONS	STATE	Temp	D.O. (mg/l)	PH	CONDUCTIVITY (µmhos/cm)	B.O.D. (mg/l)	NITRATENAN N+ NITRITENANN (mg/l)	FECAL COLIFORM (MPN/100ml)	TOTAL COLIFORM (MPN/100ml)Mean	year
0	1393	DAMANGANGA AT D/S OF MADHUBAN, DAMAN	DAMAN & DIU	30.6	6.7	7.5	203*	NAN	0.1	11	27	2014
1	1399	ZUARI AT D/S OF PT. WHERE KUMBARJRIA CANAL JOI...	GOA	29.8	5.7	7.2	189	2	0.2	4953	8391	2014

7.2 Feature 2: Handling the Null Values:

Dataset has about 500 null values in it.We have removed the null values by filling their mean values in it,because the variables are continuous.

```
In [10]: data.isnull().sum()
```

```
Out[10]: 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 [11]: 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)
```

7.3 Feature 3: Data Pre-Processing:

```
In [15]: 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 [16]: 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 [17]: data['nco']=data.TOTAL_COLIFORM.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 [18]: 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 our data frame for detecting the water's quality first we need to identify the factor water quality index. For calculating the water quality index, the pre-processing of independent variables need to be. Before that, the predefined scaling for each variable is to be carried out.

```
In [19]: data['nec']=data.CONDUCTIVITY.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 [20]: data['nna']=data.NITRATENAN.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) else 0)))))

In [21]: 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
```

7.4 Feature 4: Model building:

Splitting the data frame into tarin and test dataset. The train dataset is about 80% of total size and the test dataset is about 20% of total size. The train dataset is used for training the model while the test dataset is for evaluation of the model.
Here we have used extreme Gradient Boosting Regressor algorithm**.**

```
In [28]: x_train,x_test,y_train,y_test= train_test_split(x,y,train_size = 0.8, test_size = 0.2,random_state =42)

In [47]: model = XGBRegressor().fit(x_train, y_train)
         y_pred=model.predict(x_test)
         model.score(x_test,y_test)

         D:\anaconda\lib\site-packages\xgboost\data.py:250: FutureWarning: pandas.Int64Index is deprecated and will be removed from pand
         as in a future version. Use pandas.Index with the appropriate dtype instead.
           elif isinstance(data.columns, (pd.Int64Index, pd.RangeIndex)):

Out[47]: 0.9830935225772242
```

7.5 Feature 5: Model Evaluation:

The score of the model is about 98% approx. The Mean Absolute Error is about 0.5638 which very good and Mean Square Error is 2.99. The overall score is very high and it doesn’t show any traces of overfitting.

```
In [48]: 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: 0.5638007431221487
         MSE: 2.994258691391396
         RMSE: 1.7303926408163541
```

7.6 Feature 6: Saving the Model:

```
In [50]: import pickle
         pickle.dump(model, open('wqi.pkl','wb'))
         model = pickle.load(open('wqi.pkl','rb'))
```

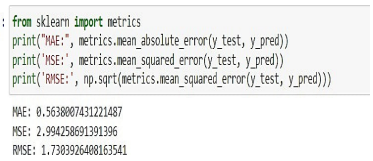
8. TESTING

8.1 User Acceptance Testing:

Section	Total Cases	Not Tested	Fail	Pass
Print Engine	8	0	0	8
Client Application	53	0	0	53
Security	5	0	0	5
Outsource Shipping	4	0	0	4
Exception Reporting	5	0	0	5
Final Report Output	6	0	0	6
Version Control	3	0	0	3

9. RESULTS

9.1 Performance Metrics:

S.No.	Parameter	Values	Screenshot
1.	Metrics	**Regression Model:** MAE - 0.5638007431221487 MSE - 2.994258691391396 RMSE - 1.7303926408163541 R2 score - 0.9830935225772242	
2.	Tune the Model	Hyper parameter Tuning - GridSearchCV	

10. ADVANTAGES &DISADVANTAGES

Advantage:

1. Benefits of predictive modeling
2. Gaining a better understanding of competition.
3. Employing strategies to gain a competitive advantage.
4. Optimizing existing products or services.
5. Understanding consumer needs.
6. Understanding the general consumer base of an industry or company.
7. Reducing time, effort and cost of estimating outcomes

Disadvantage:

1. Inadequate Training Data
2. Poor quality of data.
3. Non-representative training data.
4. Overfitting and Underfitting.
5. Monitoring and maintenance.
6. Getting bad recommendations.
7. Lack of skilled resources

11. CONCLUSION

Through the prediction from ewqa quality of water will be determined and will get various benefits from it.

12. FUTURE SCOPE

- Through water analysis unhygienic water will be predicted, it leads to the prevention of disease.
- The future scope of this project is monitoring environmental conditions, drinking water quality, treatment and disinfection of waste water
- Analysis will make to discuss about the use and process of industrial water and domestically used water.
- Prediction will tell about the availability of drinking water in the world.

13 APPENDIX

13.1 Source Code:

https://github.com/IBM-EPBL/IBM-Project-8500-1658921071/blob/main/Final%20Deliverables/Final%20code/Water%20Quality%20Analysis.ipynb

13.2 GitHub &Project Demo Link:

https://github.com/IBM-EPBL/IBM-Project-8500-1658921071

https://drive.google.com/file/d/1R8S9ecIu6b3ovfCi9B8mDA3pZ3_qDue4/view?usp=share_link