PROJECT BASED EXPERIMENTAL LEARNING PROGRAM - NALAIYA THIRAN

PROJECT REPORT

EFFICIENT WATER QUALITY ANALYSIS & PREDICTION USING MACHINE LEARNING

TEAM ID: PNT2022TMID04079

TEAM LEAD :
1) NIRMAL RAJA K L - 412519106087

TEAM MEMBERS:
2) JAGAN G - 412519106050
3) NISVANTH S - 412519106090
4) HARSHAVARDAN M - 412519106043

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

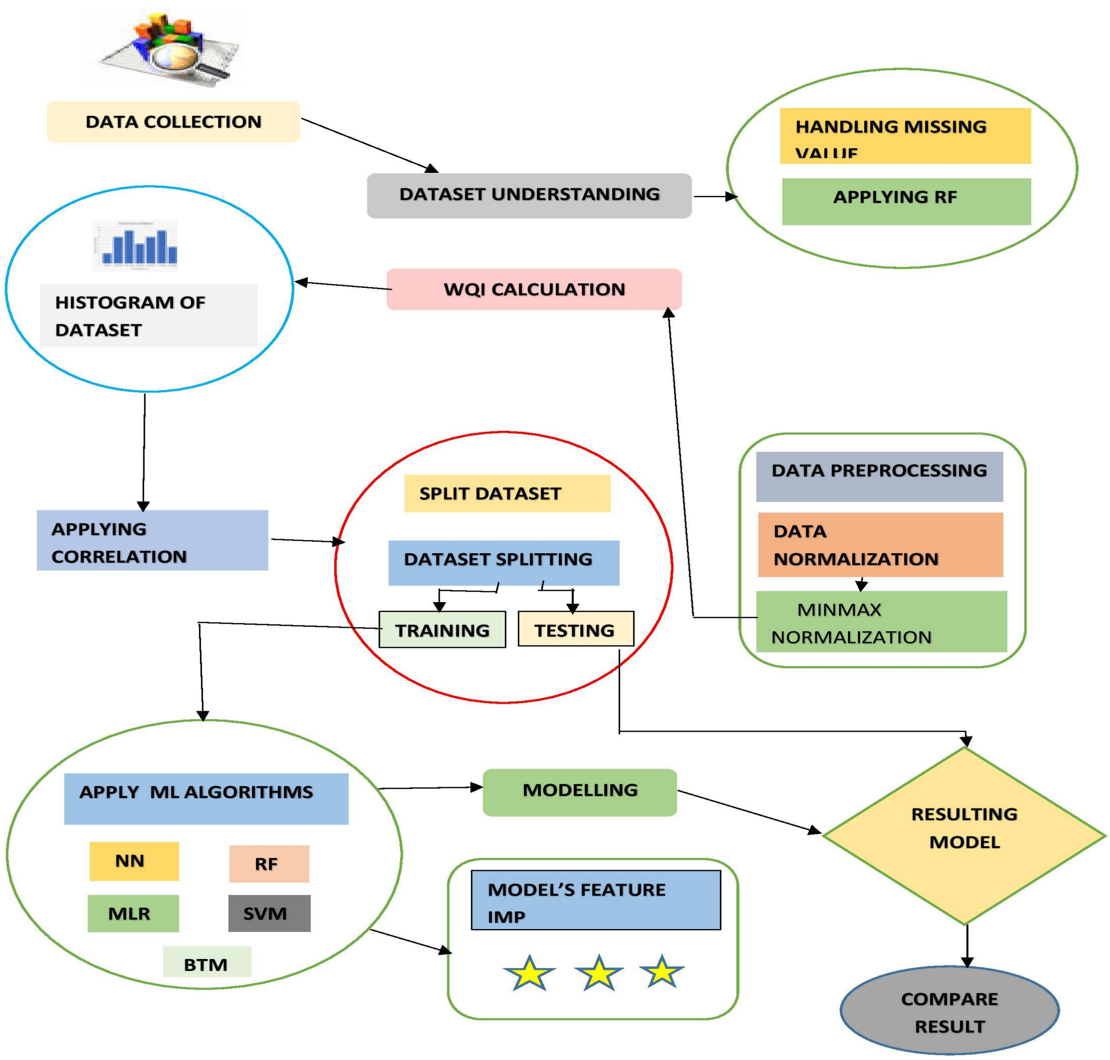
SRI SAIRAM ENGINEERING COLLEGE, CHENNAI

CONTENTS

Project Design Phase-I
Solution Architecture

Date	20 October 2022
Team ID	PNT2022TMID04079
Project Name	Project - Efficient Water Quality Analysis and Prediction using Machine Learning
Maximum Marks	4 Marks

Solution Architecture:



Project Design Phase-I
Proposed Solution Template

Date	19 October 2022
Team ID	PNT2022TMID04079
Project Name	Project – Efficient Water Quality Analysis and Prediction using Machine Learning
Maximum Marks	2 Marks

Proposed Solution Template:

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	To establish safe drinking water sources in the future, it is imperative to understand the quality and pollution level of existing groundwater. The prediction of water quality with high accuracy is the key to controlling water pollution and the improvement of water management. In this study, a Machine learning (ML) based model is proposed for predicting groundwater quality
2.	Idea / Solution description	In this study, a Machine learning (ML) based model is proposed for predicting groundwater quality and compared with three other machine learning (ML) models, namely, random forest (RF), eXtreme gradient boosting (XGBoost), and artificial neural network (ANN). A total of 226 groundwater samples are collected fan agriculturally intensive areas of India, and various physicochemical parameters are measured to compute the entropy weight-based groundwater quality index (EWQI).
3.	Novelty / Uniqueness	•The main reason behind the success of this ML technique is that it ignores the requirements of selective features that are most representative compared to that of traditional ML algorithms. •The ML technique is a self-deterministic approach that learns features to discover the correct representation required for the given task the ML technique is a self-deterministic approach that learns features to discover the correct representation required for the given task. •ML methods can progressively construct high-level attributes from the given dataset

4.	Social Impact / Customer Satisfaction	Generation of unprocessed effluents, municipal refuse, factory wastes, and junking of compostable and non-compostable effluents has hugely contaminated nature-provided water bodies like rivers, lakes, and ponds. Therefore, there is a necessity to look into the water standards before usage. This is a problem that can greatly benefit from Artificial Intelligence (AI). Traditional methods require human inspection and are time-consuming. Automatic Machine Learning (AutoML) facilities supply machine learning with the push of a button, or, on a minimum level, ensure to retain algorithm execution, data pipelines, and code, generally, are kept from sight and are anticipated to be the stepping stone for normalizing AI.
5.	Business Model (Revenue Model)	Assessment of water quality using conventional methods causes losses in eco-nomic value, which in turn affects the decision-making capacity for water quality management programs. Therefore, to tackle these issues, it is essential to adopt a potential and cost-efficient approach for quick and accurate assessment of water quality. In our project, the application of machine learning (ML) techniques can be an effective and reliable approach for the evaluation of water quality.
6.	Scalability of the Solution	•Objective weighting system-based approaches are more reliable because they consider local variations in a dataset during the computation process. •This ML technique is an extension of the artificial neural network method; it has additional complex architectures that make this approach suitable for managing multi-dimensional inputs because of its high model configuration flexibility, greater generalization power, and robust learning capacity.

Project Title: Efficient Water Quality Analysis and Prediction using machine learning

Project Design Phase-I : Solution Fit Template

Team Id: PNT2022TMID04079

Define CS, fit into CL

1. CUSTOMER SEGMENT(S) — CS

- Urban people's
- Stakeholder's of RO based companies.
- Manufacturing companies.

6. CUSTOMER LIMITATIONS EG. BUDGET, DEVICES — CL

- Spending power
- Budget
- Lack of efficient computer system
- Untrained customers

5. AVAILABLE SOLUTIONS PROS & CONS — AS

- Chemical methods
- AI techniques

Explore AS, differentiate

Focus on PR, tap into BE, understand RC

2. PROBLEMS / PAINS + ITS FREQUENCY — PR

- Urban people are mostly self-employed their livelihood are not stable. So, this method will be a cost efficient method for them.
- To check whether the water quality is in compliance with the standards, and hence, suitable or not for the designated use.

9. PROBLEM ROOT / CAUSE — RC

People think that testing the water quality for normal usage are bad investment right now because their too expensive , and possible changes to law might influence the return of investment significantly and diminish the benefits .

7. BEHAVIOR + ITS INTENSITY — BE

- Choosing of efficient providers .
- When their expected standard of water is achieved we can expect this behaviour

Focus on PR, tap into BE, understand RC

Identify strong TR & EM

3. TRIGGERS TO ACT — TR

- Seeing their neighbours using efficient water quality analysis method for their individual purpose.
- Reading about innovative and efficient solutions

4. EMOTIONS BEFORE / AFTER — EM

- Before the implementation of this system people were infuriated about their water needs.
- After accomplishing this system they will be reimbursed .

10. YOUR SOLUTION — SL

This ML technique is an extension of the artificial neural network method; it has additional complex architectures that make this approach suitable for managing multi-dimensional inputs because of its high model configuration flexibility, greater generalization power, and robust learning capacity.

8. CHANNELS of BEHAVIOR — CH

ONLINE

Extract channels from behaviour block

OFFLINE

Extract channels from behaviour block and use for customer development

Extract online & offline CH of BE

Ideation Phase

Literature Survey On The Selected Project

Information Gathering

Date	19 October 2022
Team ID	PNT2022TMID04079
Project Name	Efficient Water Quality Analysis and Prediction using Machine Learning
Maximum Marks	4 Marks

Efficient Water Quality Analysis and Prediction Using Machine Learning

Abstract:

Water makes up about 70% of the earth's surface and is one of the most important sources vital to sustaining life. Rapid urbanization and industrialization have led to a deterioration of water quality at an alarming rate, resulting in harrowing diseases. Water quality has been conventionally estimated through expensive and time-consuming lab and statistical analyses, which render the contemporary notion of real-time monitoring moot. The alarming consequences of poor water quality necessitate an alternative method, which is quicker and inexpensive. With this motivation, this research explores a series of supervised machine learning algorithms to estimate the water quality index (WQI), which is a singular index to describe the general quality of water, and the water quality class (WQC), which is a distinctive class defined on the basis of the WQI. The proposed methodology employs four input parameters, namely, temperature, turbidity, pH and total dissolved solids. Of all the employed algorithms, gradient boosting, with a learning rate of 0.1 and polynomial regression, with a degree of 2, predict the WQI most efficiently, having a mean absolute error (MAE) of 1.9642 and 2.7273, respectively. Whereas multi-layer perceptron (MLP), with a configuration of (3, 7), classifies the WQC most efficiently, with an accuracy of 0.8507. The proposed methodology achieves reasonable accuracy using a minimal number of parameters to validate the possibility of its use in real time water quality detection systems.

Keywords: water quality prediction; supervised machine learning; smart city; gradient boosting; multi-layer perceptron

---

1. Introduction

Water is the most important of sources, vital for sustaining all kinds of life; however, it is in constant threat of pollution by life itself. Water is one of the most communicable mediums with a far reach. Rapid industrialization has consequently led to deterioration of water quality at an alarming rate. Poor water quality results have been known to be one of the major factors of escalation of harrowing diseases. As reported, in developing countries, 80% of the diseases are water borne diseases, which have led to 5 million deaths and 2.5 billion illnesses [1]. The most common of these diseases in Pakistan are diarrhea, typhoid, gastroenteritis, cryptosporidium infections, some forms of hepatitis and giardiasis intestinal worms [2]. In Pakistan, water borne diseases, cause a GDP loss of 0.6–1.44% every year [3]. This makes it a pressing problem, particularly in a developing country like Pakistan.

Water quality is currently estimated through expensive and time-consuming lab and statistical analyses, which require sample collection, transport to labs, and a considerable amount of time an calculation, which is quite ineffective given water is quite a communicable medium and time is of the essence if water is polluted with disease-inducing waste [4]. The horrific consequences of water pollution necessitate a quicker and cheaper alternative.

In this regard, the main motivation in this study is to propose and evaluate an alternative method based on supervised machine learning for the efficient prediction of water quality in real-time.

This research is conducted on the dataset of Rawal water shed, situated in Pakistan, acquired by The Pakistan Council of Research in Water Resources (PCRWR) (Available online at URL http://www.pcrwr.gov.pk/). A representative set of supervised machine learning algorithms were employed on the said dataset for predicting the water quality index (WQI) and water quality class (WQC) .

The main contributions of this study are summarized as follows:

- A first analysis was conducted on the available data to clean, normalize and perform feature selection on the water quality measures, and therefore, to obtain the minimum relevant subset that allows high precision with low cost. In this way, expensive and cumbersome lab analysis with specific sensors can be avoided in further similar analyses.
- A series of representative supervised prediction (classification and regression) algorithms were tested on the dataset worked here. The complete methodology is proposed in the context of water quality numerical analysis.
- After much experimentation, the results reflect that gradient boosting and polynomial regression predict the WQI best with a mean absolute error (MAE) of 1.9642 and 2.7273, respectively, whereas multi-layer perceptron (MLP) classifies the WQC best, with an accuracy of 0.8507.

The remainder of this paper is organized as follows: Section2provides a literature review in this domain. In Section3, we explore the dataset and perform preprocessing. In Section4, we employ various machine learning methodologies to predict water quality using minimal parameters and discuss the results of regression and classification algorithms, in terms of error rates and classification precision. In Section5, we discuss the implications and novelty of our study and finally in Section6, we conclude the paper and provide future lines of work.

1.Literature Review

This research explores the methodologies that have been employed to help solve problems related to water quality. Typically, conventional lab analysis and statistical analysis are used in research to aid in determining water quality, while some analyses employ machine learning methodologies to assist in finding an optimized solution for the water quality problem.

Local research employing lab analysis helped us gain a greater insight into the water quality problem in Pakistan. In one such research study, Daud et al. [5] gathered water samples from different areas of Pakistan and tested them against different parameters using a manual lab analysis and found a high presence of *E. coli* and fecal coliform due to industrial and sewerage waste. Alamgir et al. [6] tested 46 different samples from Orangi town, Karachi, using manual lab analysis and found them to be high in sulphates and total fecal coliform count.

After getting familiar with the water quality research concerning Pakistan, we explored research employing machine learning methodologies in the realm of water quality. When it comes to estimating water quality using machine learning, Shafi et al. [7] estimated water quality using classical machine learning algorithms namely, Support Vector Machines (SVM), Neural Networks (NN), Deep Neural Networks (Deep NN) and k Nearest Neighbors (kNN), with the highest accuracy of 93% with Deep NN. The estimated water quality in their work is based on only three parameters: turbidity, temperature and pH, which are tested according to World Health Organization (WHO) standards (Available online at URLhttps://www.who.int/airpollution/guidelines/en/). Using only three parameters and comparing them to standardized values is quite a limitation when predicting water quality. Ahmad et al. [8] employed single feed forward neural networks and a combination of multiple neural networks to estimate the WQI. They used 25 water quality parameters as the input. Using a combination of backward elimination and forward selection selective combination methods, they achieved an R2 and MSE of 0.9270, 0.9390 and 0.1200, 0.1158, respectively. The use of 25 parameters makes their solution a little immoderate in terms of an inexpensive real time system, given the price of the parameter sensors. Sakizadeh [9] predicted the WQI using 16 water quality parameters and ANN with Bayesian regularization. His study yielded correlation coefficients between the observed and predicted values of 0.94 and 0.77, respectively. Abyaneh [10] predicted the chemical oxygen demand (COD) and the biochemical oxygen demand (BOD) using two conventional machine learning methodologies namely, ANN and multivariate linear regression. They used four parameters, namely pH, temperature, total suspended solids (TSS) and total suspended (TS) to predict the COD and BOD. Ali and Qamar [11] used the unsupervised technique of the average linkage (within groups) method of hierarchical clustering to classify samples into water quality classes. However, they ignored the major parameters associated with WQI during the learning process and they did not use any standardized water quality index to evaluate their predictions. Gazzaz et al. [4] used ANN to predict the WQI with a model explaining almost 99.5% of variation in the data. They used 23 parameters to predict the WQI, which turns out to be quite expensive if one is to use it for an IoT system, given the prices of the sensors. Rankovic et al. [12] predicted the dissolved oxygen (DO) using a feedforward neural network (FNN). They used 10 parameters to predict the DO, which again defeats the purpose if it has to be used for a real-time WQI estimation with an IoT system.

Most of the research either employed manual lab analysis, not estimating the water quality index standard, or used too many parameters to be efficient enough. The proposed methodology improves on these notions and the methodology being followed is depicted in Figure1.

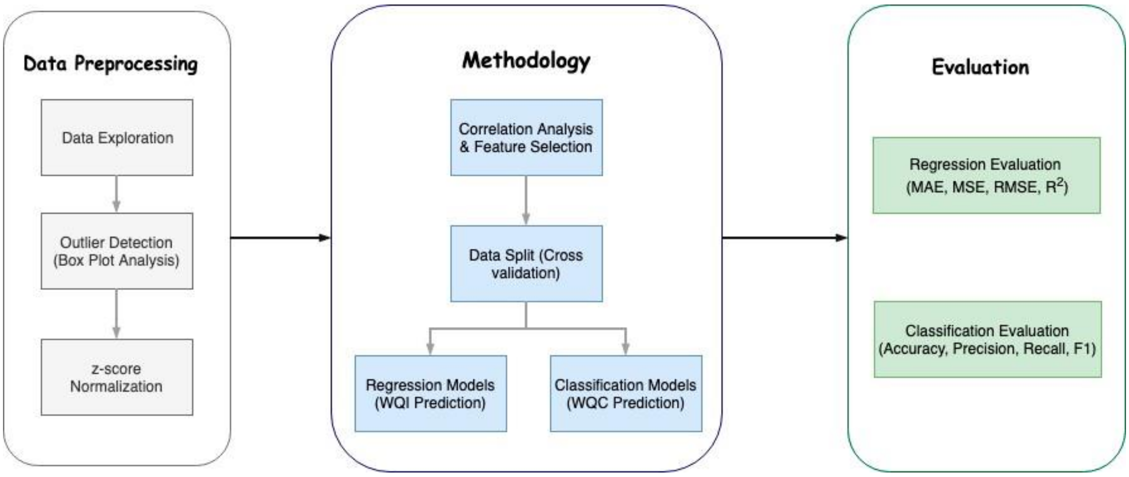


Figure 1. Methodology flow.

1. Data Preprocessig

The data used for this research was obtained from PCRWR and it was cleaned by performing a box plot analysis, discussed in this section. After the data were cleaned, they were normalized using q-value normalization to convert them to the range of 0–100 to calculate the WQI using six available parameters. Once the WQI was calculated, all original values were normalized using z-score, so they were on the same scale. The complete procedure is detailed next.

1.1. Data Collection

The dataset collected from PCRWR contained 663 samples from 13 different sources of Rawal Water Lake collected throughout 2009 to 2012. It contained 51 samples from each source and the 12 parameters listed in Table1.

Table 1. Parameters along with their "WHO" standard limits [11].

Parameter	WHO Limits
Alkalinity	500 mg/L
Appearance	Clear
Calcium	200 mg/L
Chlorides	200 mg/L
Conductance	2000 µS/cm
Fecal Coliforms	Nil Colonies/100 mL
Hardness as $CaCO_3$	500 mg/L
Nitrite as NO_2^-	$<$1 mg/L
pH	6.5–8.5
Temperature	$^{\circ}$C
Total Dissolved Solids	1000 mg/L
Turbidity	5 NTU

1.2. Boxplot Analysis and Outlier Detection

We chose boxplot analysis for outlier detection because most of the parameters varied enough and were on the higher end of the values, and a boxplot provides insightful visualization to decide outlier detection threshold values depending upon the problem domain. Boxplot analysis showed that most parameters lied outside the box, deeming outliers normal, so we adapted an upper cap strategy to filter out outliers We recognized the parameter values that were very different from other values and replaced them with the max threshold value. We set the max threshold value as the parameter value that was just below the outlier values. For example, , for turbidity, as

(1)

reflected in Figure2, we set the threshold value as the sample value, which was 753, and applied it to all values above 753, so that all the values that lied above 753 were assigned the value of 753. We repeated the same process with all the parameters and manually removed the outliers such as to not risk any data loss at all, given our limited dataset [4]. In addition, we were extremely lenient while choosing the upper threshold of parameters so as not to bias the dataset and just to loosely penalize the values that seemed way out of limits and unlikely to occur.

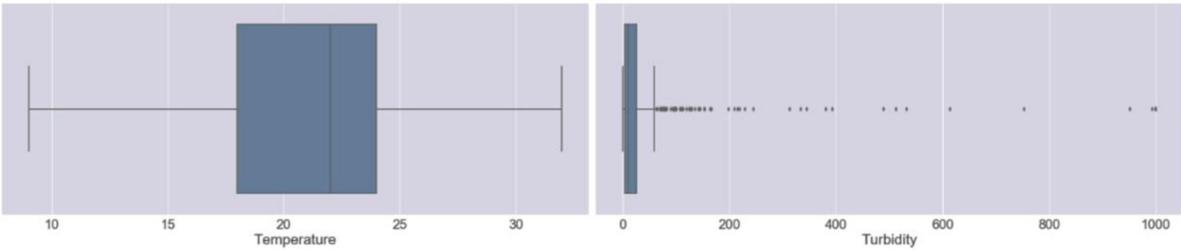


Figure 2. Outlier detection using box plot analysis.

1.3. Water Qualiity Index (WQI)

Water quality index (WQI) is the singular measure that indicates the quality of water and it is calculated using various parameters that are truly reflective of the water's quality. To conventionally calculate the WQI, nine water quality parameters are used, but if we did not have all of them, we could still estimate the water quality index with at least six defined parameters. We had five parameters, namely fecal coliform, pH, temperature, turbidity and total dissolved solids in our dataset. We also considered nitrites as the sixth parameter as the weight and relative importance of nitrites in the WQI calculation is stated to be equal to that of nitrates in multiple WQI studies [13–15]. Using these parameters and their assigned weightages, we calculated the WQI of each sample as reflected in Equation (1), where q_{value} reflects the value of a parameter in the range of 0–100 and *w_factor* represents the weight of a particular parameter as listed in Table2. WQI is fundamentally calculated by initially multiplying the q value of each parameter by its corresponding weight, adding them all up and then dividing the result by the sum of weights of the employed parameters [14,15].

$$\mathrm{WQI} = \frac{q_{value} \times w\_factor}{w\_fact} \qquad (1)$$

Table 2. Parameters weights for the WQI calculation [14,15].

Weighing Factor	Weight
pH	0.11
Temperature	0.10
Turbidity	0.08
Total Dissolved Values	0.07
Nitrates	0.10
Fecal Coliform	0.16

1.4. Water Qulaity Class (WQC)

Once we had estimated the WQI, we defined the water quality class (WQC) of each sample using the WQI in classification algorithms [14,15] as shown in Table3.

Table 3. Ranges [14,15].

Water Quality Index Range	Class
0–25	Very bad
25–50	Bad
50–70	Medium
70–90	Good
90–100	Excellent

1.5. Q-Value Normalization

Q-value normalization was used to normalize the parameters, particularly the water quality parameters to fit them in the range of 0 to 100 for easier index calculation. Figure3shows the q-value charts for six of the water quality parameters. We used them to convert five of these parameters within the range of 0 to 100 [14,15]. For the sixth parameter, namely nitrites, due to unavailability of its q-value ranges, we used the WHO standards to distinctly convert them to the 0–100 range by means of a set of thresholds as follows: assigning 100 if its below 1, 80 if its below 2, 50 if its below 3 and 0 if its greater than 3, reflecting strict penalization. Once the values were q-normalized and were in the range of 0–100, they were used for calculating the WQI of the dataset using (1).

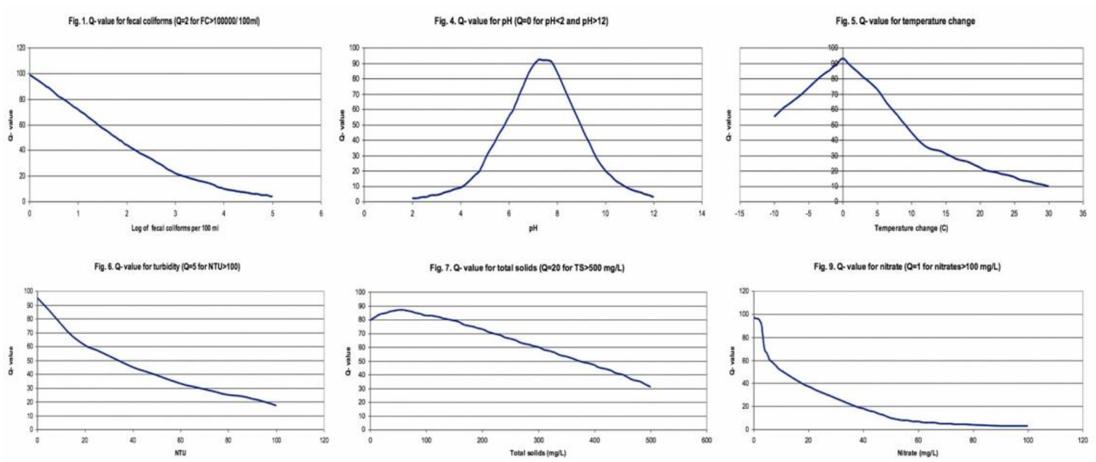


Figure 3. Q-value normalization range charts.

1.6. Z-Score Normalization

The z-score is a conventional standardization and normalization method that represents the number of standard deviations; a raw data point is above or below the population mean. It ideally lies between −3 and +3. It normalizes the dataset to the aforementioned scale to convert all the data with varying scales to the default scale.

To normalize the data using the z-score, we subtracted the mean of the population from a raw data point and divided it by the standard deviation, which gives a score ideally varying between −3 and +3; hence, reflecting how many standard deviations a point is above or below the mean as computed by Equation (2), where *x* represents the value of a particular sample, µ represents the mean and σ represents the standard deviation [16].

$$\text{z\_score} = \frac{(x - \mu)}{\sigma} \quad (2)$$

1.7. Data Analysis

After all the data processing, for data analysis, several machine learning algorithms were employed to predict the WQI and WQC using the minimal number of parameters. Before applying a machine learning algorithm, there are some preliminary steps, like correlation analysis and data splitting, to prepare the data to be given as input to the actual machine learning algorithms.

1.7.1. Correlation Analysis

To find the dependent variables and to predict hard-to-estimate variables through easily attainable parameters, we performed correlation analysis to extract the possible relationships between the parameters. We used the most commonly used and effective correlation method, known as the Pearson correlation. We applied the Pearson correlation on the raw values of the parameters listed in Table 4 and applied it after normalizing the values through q-value normalization as explained in the subsequent section.

As the correlation chart in Table 4 indicates:

- Alkalinity (Alk) is highly correlated with hardness ($CaCO_3$) and calcium (Ca).
- Hardness is highly correlated with alkalinity and calcium, and loosely correlated with pH.
- Conductance is highly correlated with total dissolved solids, chlorides and fecal coliform count, and loosely correlated with calcium and temperature.
- Calcium is highly correlated with alkalinity and hardness, while loosely correlated with TDS, chlorides, conductance and pH.
- TDS is highly correlated with conductance, chlorides and fecal coliform, and loosely correlated with calcium and temperature.
- Chlorides are highly correlated with conductance and TDS, and loosely correlated with temperature, calcium and fecal coliform.
- Fecal coliform is correlated with conductance and TDS, and loosely correlated with chlorides.

Now that we have listed the correlation analysis observations, we find that our predicting parameter WQI is correlated with seven parameters, namely temperature, turbidity, pH, hardness as $CaCO_3$, conductance, total dissolved solids and fecal coliform count. We have to choose the minimal number of parameters to predict the WQI, in order to lower the cost of the system. The three parameters whose sensors are easily available, cost the lowest and contribute distinctly to the WQI are temperature, turbidity and pH, which deems them naturally selected. The other convenient parameter is total dissolved solids, whose sensor is also easily available and is correlated with conductance and fecal coliform count, which means selecting TDS would allow us to discard the other two parameters. We leave the remaining inconvenient parameter, hardness as $CaCO_3$, out because it is not highly correlated comparatively and is not easy to acquire.

To conclude the correlation analysis, we selected four parameters for the prediction of WQI, namely, temperature, turbidity, pH and total dissolved solids. We initially just considered the first three parameters, given their low cost, and if needed, TDS will be included later to analyze its contribution to the accuracy.

Table 4. Correlation Analysis Chart.

	Temp	**Turb**	**pH**	**Alk**	**$CaCO_3$**	**Cond**	**Ca**	**TDS**	**Cl**	**NO_2**	**FC**	**WQI**
Temp	1.000	0.103	0.005	−0.193	−0.288	0.266	−0.150	0.274	0.293	−0.154	0.194	**−0.467**
Turb	0.103	1.000	−0.0886	−0.093	−0.146	0.048	−0.122	0.042	0.037	0.0002	0.037	**−0.354**
pH	0.005	−0.088	1.000	−0.177	−0.278	−0.065	−0.236	−0.060	−0.149	0.167	0.054	**−0.431**
Alk	−0.193	−0.092	−0.177	1.000	0.462	0.011	0.444	0.012	0.061	0.046	0.013	0.223
$CaCO_3$	−0.288	−0.146	−0.278	0.462	1.000	0.068	0.637	0.060	0.135	0.078	0.016	0.360
Cond	0.266	0.048	−0.064	0.011	0.068	1.000	0.225	0.973	0.780	0.100	0.456	−0.370
Ca	−0.150	−0.122	−0.236	0.444	0.637	0.225	1.000	0.219	0.262	0.124	0.113	0.188
TDS	0.273	0.041	−0.060	0.012	0.060	0.974	0.219	1.000	0.765	0.095	0.454	**−0.381**
Cl	0.292	0.037	−0.149	0.061	0.135	0.780	0.262	0.765	1.000	0.036	0.353	−0.274
NO_2	−0.154	0.0002	0.167	0.046	0.078	0.100	0.124	0.095	0.036	1.000	0.193	−0.209
FC	0.194	0.037	0.053	0.012	0.016	0.456	0.113	0.454	0.353	0.193	1.000	−0.421
WQI	−0.467	−0.354	−0.431	0.223	0.360	−0.370	0.188	−0.381	−0.274	−0.209	−0.421	1.000

1.7.2. Data Splitting–Cross Validation

The last step prior to applying the machine learning model is splitting the provided data in order to train the model, test it with a certain part of the data and compute the accuracy measures to establish the model's performance. This research explores the cross validation data splitting technique.

Cross validation splits the data into k subsets and iterates over all the subsets, considering k-1 subsets as the training dataset and 1 subset as the testing dataset. This ensures an efficient split and use of proper and definitive data for training and testing. This is generally computationally expensive, given the iterations, but our research uses a small dataset, which is mostly the case with water quality datasets, making cross validation more suited for this problem. We split the data into k = 6 subsets and ran cross validation. Therefore, as the complete training set consists of 663 samples, we ensured at least 100 samples for each fold subset, including the test set.

1.7.3. Machine Learning Algorithms

We used both regression and classification algorithms. We used the regression algorithms to estimate the WQI and the classification algorithms to classify samples into the previously defined WQC. We used eight regression algorithms and 10 classification algorithms. The following algorithms were employed in our study:

(1) Multiple Linear Regression

Multiple linear regression is a form of linear regression used when there is more than one predicting variable at play. When there are multiple input variables, we use multiple linear regression to assess the input of each variable that affects the output, as reflected in Equation (3), where y is the output for which machine learning has been applied to predict the value, x is the observed value, β is the slope on the observed value, and ∈ is the error term [17].

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \in \quad (3)$$

(2) Polynomial Regression

Polynomial regression is used when the relation between input and output variables is not linear and a little complex. We used a higher order of variables to capture the relation of input and output variables, which is not as linear. We used the order of two. Using a higher order of variables does carry the risk of overfitting, as reflected in Equation (4), where y is the output for which machine learning has been applied to predict the value, x is the observed value, β is the fitting value, i is the number of parameters considered, k is the order of the polynomial equation, and $\in_i$ is the error term or residuals of the ith predictor [18]. We used it with 2-degree polynomials with an order of C.

$$y = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_k x_i^k + \in_i, \ for\ i = 1, 2, \ldots, n \quad (4)$$

(3) Random Forest

Random forest is a model that uses multiple base models on subsets of the given data and makes decisions based on all the models. In random forest, the base model is a decision tree, carrying all the pros of a decision tree with the additional efficiency of using multiple models [19].

(4) Gradient Boosting Algorithm

This is the most contemporary algorithm used in most competitions. It uses an additive model that allows for optimization of differentiable loss function. We used it with a loss function of 'ls', a min_samples_split of 2 and a learning rate of 0.1 [20].

(5) Support Vector Machines

Support vector machines (SVMs) are mostly used for classification but they can be used for regression as well. Visualizing data points plotted on a plane, SVMs define a hyperplane between the classes and extend the margin in order to maximize the distinction between two classes, which results in fewer close miscalculations [21].

(6) Ridge Regression

Ridge regression works on the same principles as linear regression, it just adds a certain bias to negate the effect of large variances and to void the requirement of unbiased estimators. It penalizes the coefficients that are far from zero and minimizes the sum of squared residuals [22,23].

(7) Lasso Regression

Lasso regression works on the same principles as ridge regression, the only difference is how they penalize their coefficients that are off. Lasso penalizes the sum of absolute errors instead of the sum of squared coefficients [24].

(8) Elastic Net Regression

Elastic net regression combines the best of both ridge and lasso regression. It combines the method of penalties of both methods and minimizes the loss function [25].

(9) Neural Net/Multi-Layer Perceptrons (MLP)

Neural nets are loosely based on the structure of neurons. They contain multiple layers with interconnected nodes. They contain an input layer and output layer, and hidden layers in between these two mandatory layers. The input layer takes in the predicting parameters and the output layer shows the prediction based on the input. They iterate through each training data point and generalize the model by giving and updating the weight on each node of each layer. The trained model then uses those weights to decide what units to activate based on the input. Multi-layer perceptron (MLP) is a conventional model of neural net, which is mostly used for classification, but it can be used for regression as well [26]. We used it for classification with the configuration of (3, 7) running for a maximum of 200 epochs using 'lbfgs' solver.

(10) Gaussian Naïve Bayes

Naïve Bayes is a simple and a fast algorithm that works on the principle of Bayes theorem with the assumption that the probability of the presence of one feature is unrelated to the probability of the presence of the other feature [27].

(11) Logistic Regression

Logistic regression is a classification algorithm. It is based on the logistic function or the sigmoid function, hence the name. It is the most common algorithm used in the case of binary classification, but in our case we used multinomial logistic regression because there was more than two classes [28]. We used it with 'warn' solver and l2 penalty.

(12) Stochastic gradient descent

This iterative optimization algorithm minimizes the loss function iteratively to find the global optimum. In stochastic gradient descent, the sample selection is random [29].

(13) K Nearest Neighbor

The K nearest neighbor algorithm classifies by finding the given points nearest N neighbors and assigns the class of majority of n neighbors to it. In the case of a draw, one could employ different techniques to resolve it, e.g., increase n or add bias towards one class. K nearest neighbor is not recommended for large datasets because all the processing takes place while testing, and it iterates through the whole training data and computes nearest neighbors each time [30]. We used a n = 5 configuration for our model.

(14) Decision Tree

A decision tree is a simple self-explanatory algorithm, which can be used for both classification and regression. The decision tree, after training, makes decisions based on values of all the relevant input parameters. It uses entropy to select the root variable, and, based on this, it looks towards the other parameters' values. It has all the parameter decisions arranged in a top-to-down tree and projects the decision based on different values of different parameters [31].

(15) Bagging Classifier

A bagging classifier fits multiple base classifiers on random subsets of data and then averages out their predictions to form the final prediction. It greatly helps out with the variance [32].

We used default values for the algorithms, except MLP, which uses a (3, 7) configuration.

2.Results

In this section, prior to discussing the results, we will describe different measures used to assess the accuracy of the applied machine learning algorithms.

1.8. Accuracy Measures

As mentioned earlier, this research employed two types of supervised machine learning algorithms, i.e., regression and classification. The results yielded by both types of algorithms were evaluated differently.

For regression, we used the following measures:

(1) Mean Absolute Error (MAE)

Mean absolute error (MAE) is a measure of accuracy for regression. It sums up absolute values of errors and divides them by the total number of values. It gives equal weight to each error value.

The formula for calculating MAE is shown in Equation (5), where x_{obs} refers to the actual value, x_{pred} refers to the predicted value, and n refers to the total number of samples considered [33].

$$\text{MAE} = \frac{\sum(\| x_{obs} - x_{pred} \|)}{n} \quad (5)$$

(2) Mean Square Error (MSE)

Mean square error (MSE) is the sum of squares of errors divided by the total number of predicted values. This attributes greater weight to larger errors. This is particularly useful in problems where there needs to be a larger weight for larger errors. It is measured by Equation (6), where x_{obs} is the actual value, x_{pred} is the predicted value, and n is the total number of samples considered [33].

$$\text{MSE} = \frac{\sum (x_{obs} - x_{pred})^2}{n} \quad (6)$$

(3) Root Mean Squared Error (RMSE)

Root mean squared error (RMSE) is just the square root of MSE and scales the values of MSE near to the ranges of observed values. It is estimated from Equation (7), where x_{obs} points to the actual value, x_{pred} points to the predicted value, and n points to the total number of samples considered [33].

$$\text{RMSE} = \sqrt{\frac{\sum (x_{obs} - x_{pred})^2}{n}} \quad (7)$$

(4) R Squared Error (RSE)

R squared error (RSE), also known as the coefficient of determination, and often denoted as R^2, determines the goodness of fit of the model. It particularly explains the amount of variance of the dependent variable that is explainable through the independent variable, as shown in Equation (8). Higher RSE values mean that the independent variables largely explain the variance of the dependent variable [34].

$$\text{RSE or } R^2 = 1 - \frac{\text{Explained variation}}{\text{Total variation}} \quad (8)$$

For classification, we used the following measures:

(1) Accuracy

Accuracy is the correct number of predictions made by the model over all the observed values. Accuracy is measured by Equation (9), where TP refers to true positive, TN refers to true negative, FP refers to false positive and FN refers to false negative [7,35].

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (9)$$

(2) Precision

Precision is the proportion of correctly classified instances of a particular positive class out of the total classified instances of that class. Precision is calculated with the formula shown in Equation (10), where TP refers to true positive and FP refers to false positive [7,35,36].

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

(3) Recall

Recall is the proportion of instances of a particular positive class that were actually classified correctly. Recall is calculated with the formula shown in Equation (11), where *TP* refers to true positive and *FN* refers to false negative [7,35,36].

$$\text{Recall} = \frac{TP}{TP + FN} \quad (11)$$

(4) F1 Score

As precision and recall, individually, do not cover all aspects of the accuracy, we took their harmonic mean to reflect the F1 score, as shown in Equation (12), which covers both aspects and reflects the overall accuracy measure better. It ranges between 0 and 1. The higher the score, the better the accuracy [7,35,36].

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

1.9. Results for Regression Algorithms

As water quality parameter sensors are expensive, this study aimed to use a minimal number of parameters with cheap sensors to predict water quality. Initially, we used four parameters, namely temperature, turbidity, pH and total dissolved solids. While employing the regression algorithms, we found gradient boosting, having a MAE of 1.9642, MSE of 7.2011, RMSE of 2.6835 and RSE of 0.7485, to be the most efficient algorithm, as shown in Table5.

Table 5. Regression results using four parameters.

Algorithm	MAE	MSE	RMSE	R Squared
Linear Regression	2.6312	11.7550	3.4286	0.6573
Polynomial Regression	2.0037	7.9467	2.8190	0.7134
Random Forest	2.3053	9.5669	3.0930	0.6705
Gradient Boosting	**1.9642**	**7.2011**	**2.6835**	**0.7485**
SVM	2.4373	10.6333	3.2609	0.3458
Ridge Regression	2.6323	11.7500	3.4278	0.4971
Lasso Regression	3.5850	20.1185	4.4854	−2.9327
Elastic Net Regression	3.6595	20.9698	4.5793	−4.0050

Following that, we tried to reduce more parameters, and so we decided to drop total dissolved solids, which is a little harder to acquire than the others. We found that the polynomial regression, with a MAE of 2.7273, MSE of 12.7307, and RMSE of 3.5680, to be the most efficient algorithm, although linear regression and gradient boosting had the best RSE values, with 0.5384 and 0.5051, respectively, as shown in Table6. There was an increase in the overall error rate, but the increase was not alarming and still performed well within the limits, given the cost.

Table 6. Regression results using three parameters.

Algorithm	MAE	MSE	RMSE	R Squared
Linear Regression	3.1375	15.8321	3.9790	**0.5384**
Polynomial Regression	**2.7273**	**12.7307**	**3.5680**	0.4851
Random Forest	3.0404	15.2473	3.9048	0.4107
Gradient Boosting	2.8060	13.2710	3.6429	0.5051
SVM	2.8252	13.8546	3.7222	0.1546
Ridge Regression	3.1386	15.8327	3.9790	0.2031
Lasso Regression	3.8800	22.9966	4.7955	−3.6636
Elastic Net Regression	3.9697	24.0678	4.9059	−5.5210

1.9.1. Results for Classification Algorithms

Using classification algorithms, we predicted water quality class (WQC), which was assigned to samples based on their pre calculated WQI. The same parameters, as in previous section, were used for classification as well. Initially, the same four parameters were considered. We found that MLP, in such a setting, performed better than the other algorithms, with an accuracy of 0.8507, precision of 0.5659, recall of 0.5640, and F1 score of 0.5649, as shown in Table7.

Table 7. Classification results using four parameters.

Algorithm	Accuracy	Precision	Recall	F1 Score
MLP	**0.8507**	**0.5659**	**0.5640**	**0.5649**
Guassian Naïve Bayes	0.7843	0.4964	0.5491	0.5025
Logistic Regression	0.8401	0.5520	0.5594	0.5548
Stochastic Gradient Descent	0.8205	0.5473	0.5424	0.5443
KNN	0.7270	0.4734	0.4783	0.4750
Decision Tree	0.7949	0.5298	0.5250	0.5268
Random Forest	0.7587	0.5063	0.5011	0.5027
SVM	0.7979	0.5187	0.5327	0.5228
Gradient Boosting Classifier	0.8130	0.5375	0.5376	0.5376
Bagging Classifier	0.8100	0.5410	0.5354	0.5374

In this section, we iterated through our study's results and established that gradient boosting and polynomial regression performed better in predicting WQI, whereas MLP performed better in predicting WQC.

3.Discussion

Water Quality is conventionally calculated using water quality parameters, which are acquired through time consuming lab analysis. We explored alternative methods of machine learning to estimate it and found several studies employing them. These studies used more than 10 water quality parameters to predict WQI. Ahmad et al. [8] used 25 input parameters, Sakizadeh [9] used 16 parameters, Gazzaz et al. [4] used 23 input parameters in their methodology, and Rankovic et al. [12] used 10 input parameters, which is unsuitable for inexpensive real time systems. Whereas, our methodology employs only four water quality parameters to predict WQI, with a MAE of 1.96, and to predict water quality class with an accuracy of 85%. Our results make a base for an inexpensive real time water quality detection system, while other studies, although they use machine learning, use too many parameters to be incorporated in real time systems.

4.Conclusions and Future Work

Water is one of the most essential resources for survival and its quality is determined through WQI.

Conventionally, to test water quality, one has to go through expensive and cumbersome lab analysis. Thisresearch explored an alternative method of machine learning to predict water quality using minimal and easily available water quality parameters. The data used to conduct the study were acquired from PCRWR and contained 663 samples from 12 different sources of Rawal Lake, Pakistan. A set of representative supervised machine learning algorithms were employed to estimate WQI. This showed that polynomial regression with a degree of 2, and gradient boosting, with a learning rate of 0.1, outperformed other regression algorithms by predicting WQI most efficiently, while MLP with a configuration of (3, 7) outperformed other classification algorithms by classifying WQC most efficiently. In future works, we propose integrating the findings of this research in a large-scale IoT-based online monitoring system using only the sensors of the required parameters. The tested algorithms would predict the water quality immediately based on the real-time data fed from the IoT system. The proposed IoT system would employ the parameter sensors of pH, turbidity, temperature and

TDS for parameter readings and communicate those readings using an Arduino microcontroller and ZigBee transceiver. It would identify poor quality water before it is released for consumption and alert concerned authorities. It will hopefully result in curtailment of people consuming poor quality water and consequently de-escalate harrowing diseases like typhoid and diarrhea. In this regard, the application of a prescriptive analysis from the expected values would lead to future facilities to support decision and policy makers.

Author Contributions: Conceptualization, R.M.; Data curation, H.A.; Formal analysis, U.A. and R.I.; Investigation, A.A.S.; Methodology, U.A.; Resources, R.M.; Supervision, R.M.; Validation, U.A. and R.M.; Writing—original draft, U.A.; Writing—review & editing, R.M., H.A., R.I. and J.G.-N.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. PCRWR. *National Water Quality Monitoring Programme, Fifth Monitoring Report (2005–2006)*; Pakistan Council of Research in Water Resources Islamabad: Islamabad, Pakistan, 2007. Available online: http://www.pcrwr.gov.pk/Publications/Water%20Quality%20Reports/Water%20Quality%20Monitoring%20Report%202005-06.pdf (accessed on 23 August 2019).
2. Mehmood, S.; Ahmad, A.; Ahmed, A.; Khalid, N.; Javed, T. Drinking Water Quality in Capital City of Pakistan. *Open Access Sci. Rep.* **2013**, *2*. [CrossRef]
3. PCRWR. *Water Quality of Filtration Plants, Monitoring Report*; PCRWR: Islamabad, Pakistan, 2010. Available online: http://www.pcrwr.gov.pk/Publications/Water%20Quality%20Reports/FILTRTAION%20PLANTS%20REPOT-CDA.pdf (accessed on 23 August 2019).
4. Gazzaz, N.M.; Yusoff, M.K.; Aris, A.Z.; Juahir, H.; Ramli, M.F. Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. *Mar. Pollut. Bull.* **2012**, *64*, 2409–2420. [CrossRef]
5. Daud, M.K.; Nafees, M.; Ali, S.; Rizwan, M.; Bajwa, R.A.; Shakoor, M.B.; Arshad, M.U.; Chatha, S.A.S.; Deeba, F.; Murad, W.; et al. Drinking water quality status and contamination in Pakistan. *BioMed Res. Int.* **2017**, *2017*, 7908183. [CrossRef]
6. Alamgir, A.; Khan, M.N.A.; Hany; Shaukat, S.S.; Mehmood, K.; Ahmed, A.; Ali, S.J.; Ahmed, S. Public health quality of drinking water supply in Orangi town, Karachi, Pakistan. *Bull. Environ. Pharmacol. Life Sci.* **2015**, *4*, 88–94.
7. Shafi, U.; Mumtaz, R.; Anwar, H.; Qamar, A.M.; Khurshid, H. Surface Water Pollution Detection using Internet of Things. In Proceedings of the 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT), Islamabad, Pakistan, 8–10 October 2018; pp. 92–96.
8. Ahmad, Z.; Rahim, N.; Bahadori, A.; Zhang, J. Improving water quality index prediction in Perak River basin Malaysia through a combination of multiple neural networks. *Int. J. River Basin Manag.* **2017**, *15*, 79–87. [CrossRef]
9. Sakizadeh, M. Artificial intelligence for the prediction of water quality index in groundwater systems. *Model. Earth Syst. Environ.* **2016**, *2*, 8. [CrossRef]
10. 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. [CrossRef]
11. Ali, M.; Qamar, A.M. Data analysis, quality indexing and prediction of water quality for the management of rawal watershed in Pakistan. In Proceedings of the Eighth International Conference on Digital Information Management (ICDIM 2013), Islamabad, Pakistan, 10–12 September 2013; pp. 108–113.
12. Ranković, V.; Radulović, J.; Radojević, I.; Ostojić, A.; Čomić, L. Neural network modeling of dissolved oxygen in the Gruža reservoir, Serbia. *Ecol. Model.* **2010**, *221*, 1239–1244. [CrossRef]
13. Kangabam, R.D.; Bhoominathan, S.D.; Kanagaraj, S.; Govindaraju, M. Development of a water quality index (WQI) for the Loktak Lake in India. *Appl. Water Sci.* **2017**, *7*, 2907–2918. [CrossRef]
14. Thukral, A.; Bhardwaj, R.; Kaur, R. Water quality indices. *Sat* **2005**, *1*, 99.
15. Srivastava, G.; Kumar, P. Water quality index with missing parameters. *Int. J. Res. Eng. Technol.* **2013**, *2*, 609–614.

16. Jayalakshmi, T.; Santhakumaran, A. Statistical normalization and back propagation for classification. *Int. J. Comput. Theory Eng.* **2011**, *3*, 1793–8201.
17. Amral, N.; Ozveren, C.; King, D. Short term load forecasting using multiple linear regression. In Proceedings of the 2007 42 nd International Universities Power Engineering Conference, Brighton, UK, 4–6 September 2007; pp. 1192–1198.
18. Ostertagov á, E. Modelling using polynomial regression. *Procedia Eng.* **2012**, *48*, 500–506. [CrossRef]
19. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
20. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [CrossRef]
21. Tong, S.; Koller, D. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.* **2001**, *2*, 45–66.
22. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
23. Zhang, Y.; Duchi, J.; Wainwright, M. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.* **2015**, *16*, 3299–3340.
24. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
25. Zou, H.; Hastie, T. Regression shrinkage and selection via the elastic net, with applications to microarrays. *J. R. Stat. Soc. Ser. B* **2003**, *67*, 301–320. [CrossRef]
26. Günther, F.; Fritsch, S. Neuralnet: Training of neural networks. *R J.* **2010**, *2*, 30–38. [CrossRef]
27. Zhang, H. The optimality of naive Bayes. *AA* **2004**, *1*, 3.
28. Hosmer, D.W., Jr.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; John Wiley Sons: Hoboken, NJ, USA, 2013.
29. Bottou, L. Large-scale machine learning with stochastic gradient descent. In Proceedings of the COMPSTAT'2010, Paris, France, 22–27August 2010; pp. 177–186.
30. Beyer, K.; Goldstein, J.; Ramakrishnan, R.; Shaft, U. When is "nearest neighbor" meaningful? In Proceedings of the International Conference on Database Theory, Jerusalem, Israel, 10–12 January 1999; pp. 217–235.
31. Quinlan, J.R. Decision trees and decision-making. *IEEE Trans. Syst. Man Cybern.* **1990**, *20*, 339–346. [CrossRef]
32. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
33. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **2005**, *30*, 79–82. [CrossRef]
34. Menard, S. Coefficients of determination for multiple logistic regression analysis. *Am. Stat.* **2000**, *54*, 17–24.
35. Sokolova, M.; Japkowicz, N.; Szpakowicz, S. Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In Proceedings of the Australasian Joint Conference on Artificial Intelligence, Hobart, Australia, 4–8 December 2006; pp. 1015–1021.
36. Goutte, C.; Gaussier, E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In Proceedings of the European Conference on Information Retrieval, Santiago de Compostela, Spain, 21–23 March 2005; pp. 345–359.

Ideation Phase

Date	19 Octoberr 2022
Team ID	PNT2022TMID04079
Project Name	Efficient Water Quality Analysis and Prediction using Machine Learning
Maximum Marks	4 Marks

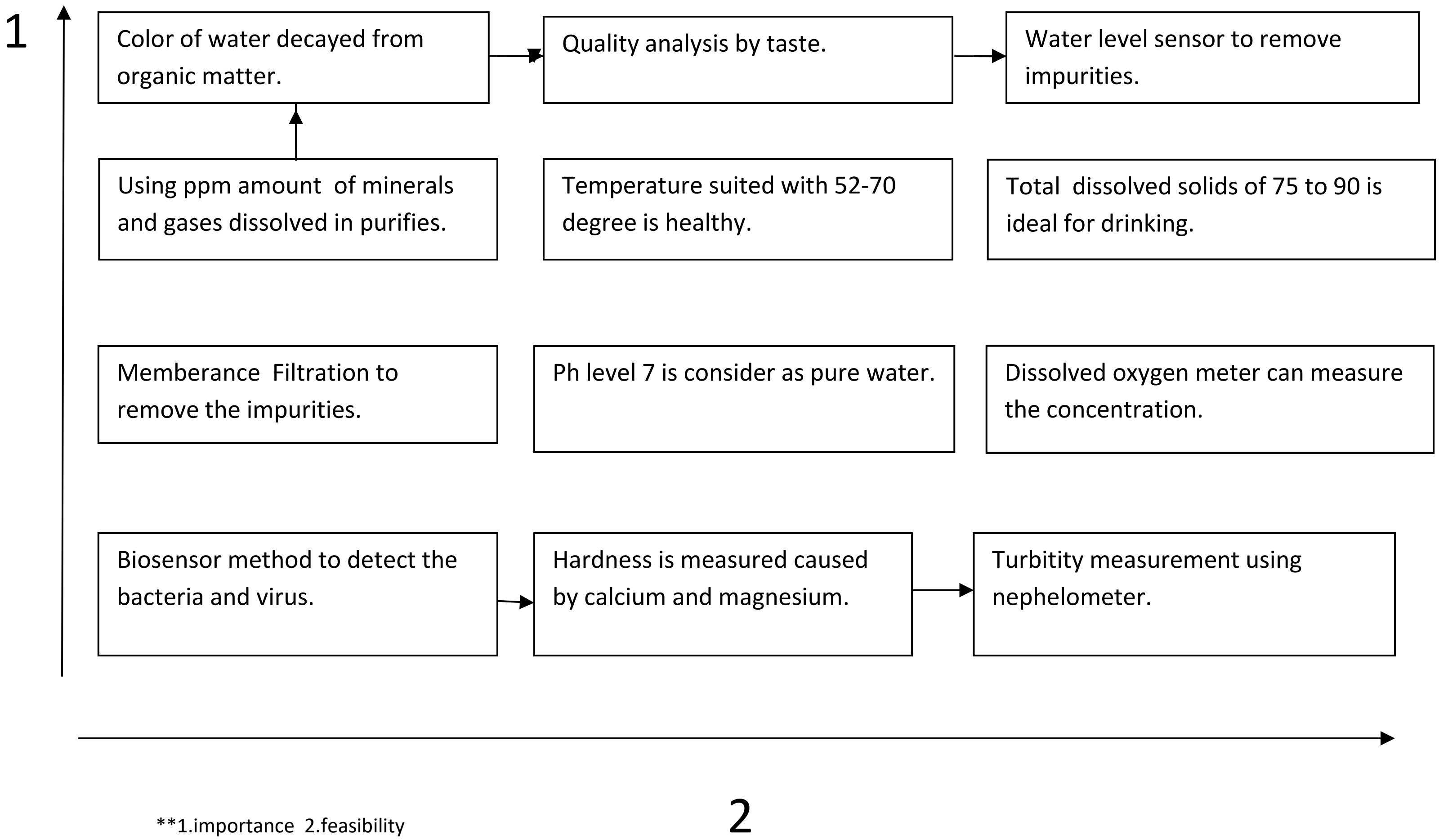
Efficient water quality analysis & prediction using machine learning

1. Water makes up about 70% of the surface and is one of the most important sources vital to sustaining life.
2. Water quality has been conventionally estimated through expensive and time consuming lab and statical analysis.
3. With this motivation ,we explore a series of supervised machine learning algorithm to estimate the water quality.

Big Idea:

Temperature suited with 52-70 degree is healthy.	Biosensor method to detect the bacteria and virus.	Hardness is measured caused by calcium and magnesium.
Ph level 7 is consider as pure water.	Memberance Filtration to remove the impurities.	Dissolved oxygen meter can measure the concentration.
Total dissolved solids of 75 to 90 is ideal for drinking.	Color of water decayed from organic matter.	Using ppm amount of minerals and gases dissolved in purifies.
Quality analysis by taste.	Turbitity measurement using nephelometer.	Water level sensor to remove impurities.

Idea prioritization:



**1.importance 2.feasibility

Ideation Phase
Empathize & Discover

Date	19 October 2022
Team ID	PNT2022TMID04079
Project Name	Project - Efficient Water Quality Analysis & Prediction using Machine Learning
Maximum Marks	4 Marks

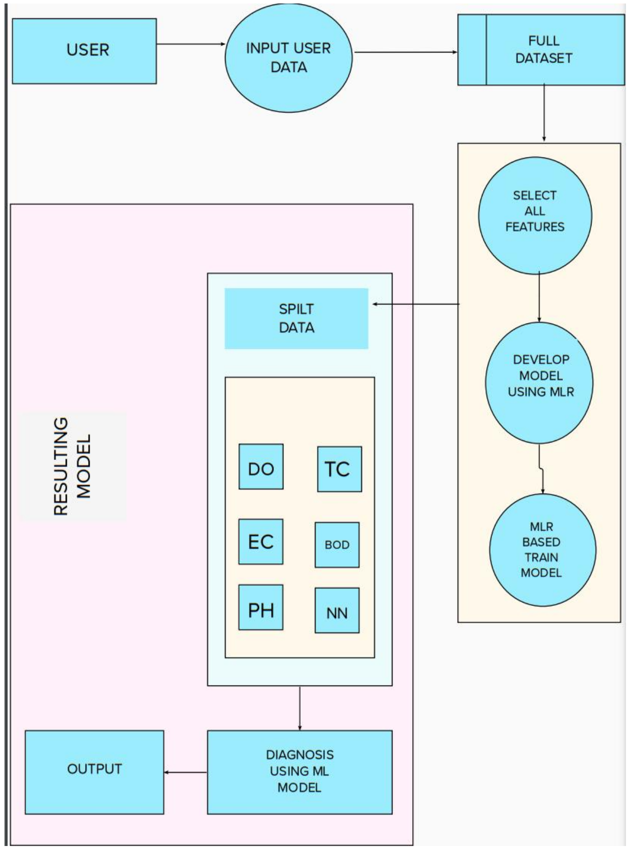
EMPATHY MAP

TOPIC: EFFICIENT WATER QUALITY ANALYSIS AND PREDICTION USING MACHINE LEARNING

Project Design Phase-II
Data Flow Diagram
&
User Stories

Date	19 October 2022
Team ID	PNT2022TMID04079
Project Name	Efficient Water Quality Analysis and Prediction Using Machine Learning
Maximum Marks	4 Marks

DATAFLOW DIAGRAM:



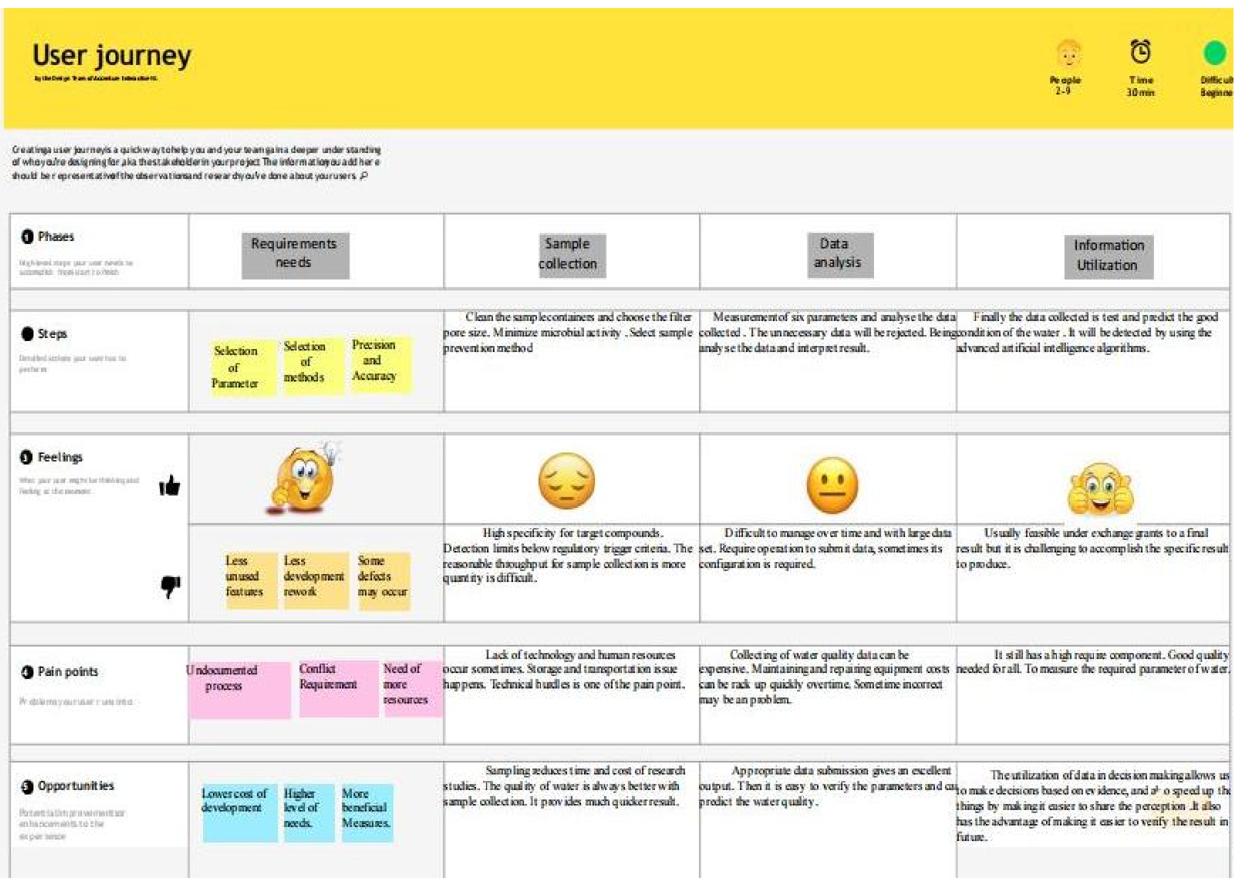
USER STORIES

USER TYPE	FUNCTIONAL REQUIREMENT (EPIC)	USER STORY NUMBER	USER STORY / TASK	ACCEPTANCE CRITERIA	PRIORITY	RELEASE
Customer (Mobile 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	Sprint-1
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint-1
		USN-3	As a user, I can register through website	I can register and access the account with website	High	Sprint-1
		USN-4	As a user, I can register for the application through Gmail	I can register and access the gmail	Medium	Sprint-1
	Login	USN-5	As a user, I can log into the application by entering email & password	I can successfully login into application	High	Sprint-1
	Dashboard	USN-6	As a user,I can access the dashboard	I can referred dashboard for certainty	Medium	Sprint-1
Customer (Ordinary people,Industry)	Analysis the water quality	USN-7	As a user,I can access the water quality analysis in all over india	I can predict the water quality earlier	High	Sprint-1
Customer Care Executive	Customer queries	USN-8	As a user ,I can register the complaint in website	I can get immediate solution	High	Sprint-1
Administrator	Getting value	USN-9	when there is a issuses in getting analysed value	through administrator getting predicted value	Low	Sprint-2

Project Design Phase-II

Customer Journey

Date	20 October 2022
Team ID	**PNT2022TMID04079**
Project Name	Efficient Water Quality Analysis and Prediction using Machine Learning
Team Leader	Rajaranganayaki R
Team Members	Rakshambika S, Vidhya P , Shanmuga valli S
Maximum Marks	2 Marks



Project Design Phase-II
Solution Requirements (Functional & Non-functional)

Date	19 October 2022
Team ID	**PNT2022TMID04079**
Project Name	Project – Efficient Water Quality Analysis and Prediction using Machine Learning
Maximum Marks	4 Marks

Functional Requirements:

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Registration through Form Registration through Gmail Registration through LinkedIN
FR-2	User Confirmation	Confirmation via Email Confirmation via OTP
FR-3	Authorization level	A Security question will be displayed to the user to verify the details.
FR-4	Reporting	1.Result of the water quality analysis will be sent a message to the user. 2.The real-time water quality report is collected and the dataset is used to predict the water quality for future works.
FR-5	Business rules	Water Quality Index(WQI) formula will be used for the water quality analysis and prediction.

Non-functional Requirements:

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

Project Design Phase-II
Technology Stack (Architecture & Stack)

Date	20 October 2022
Team ID	PNT2022TMID04079
Project Name	Efficient Water Quality Analysis and Prediction using Machine Learning
Maximum Marks	4 Marks

Technical Architecture:

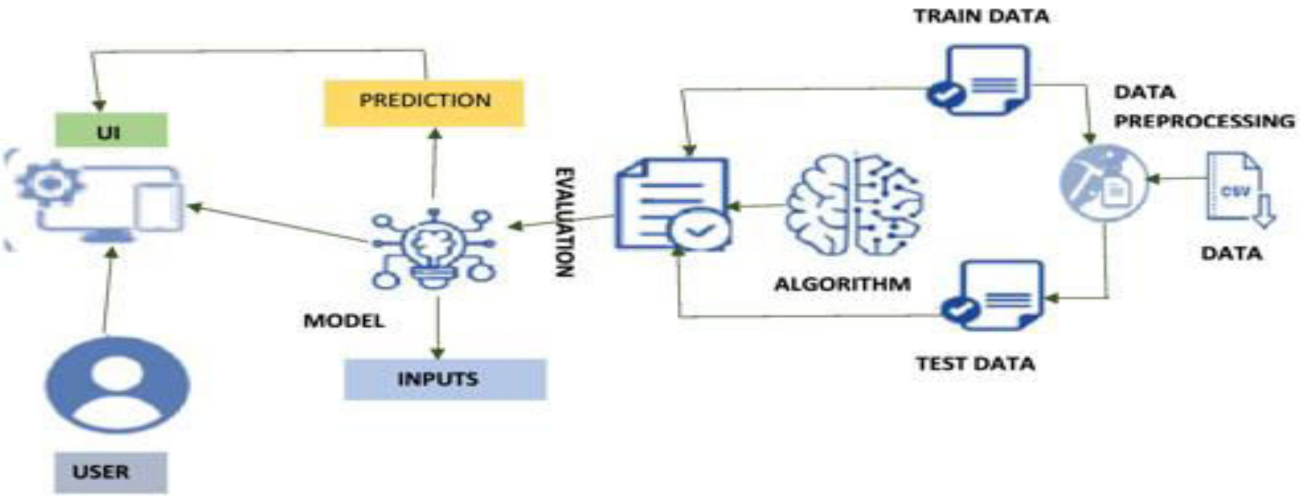


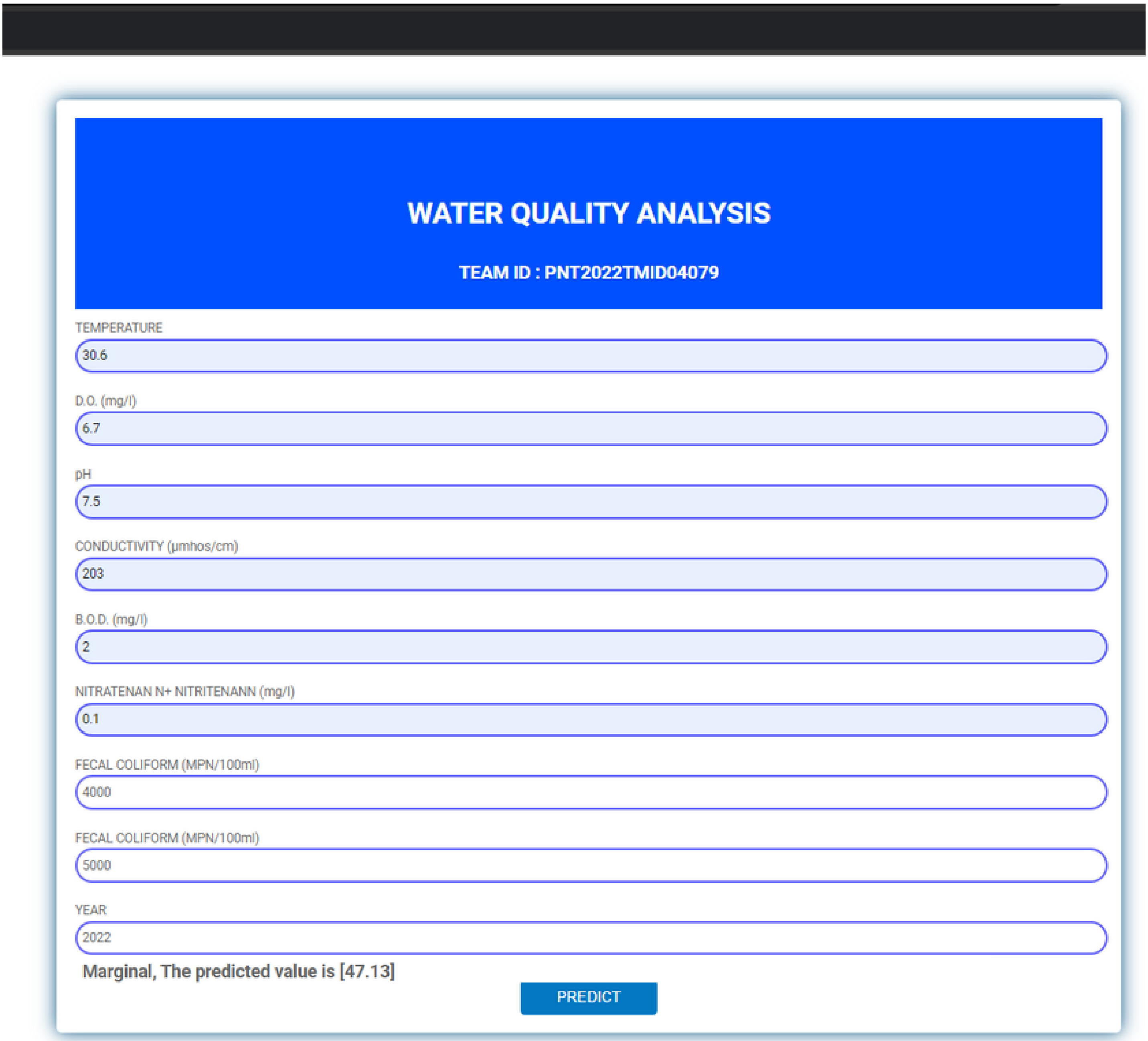
Table-1 : Components & Technologies:

S.NO	Component	Description	Technology
1.	User Interface	How user interacts with application e.g. Web UI, Mobile App, Chatbot etc.	HTML, CSS, Python
2.	Application Logic-1	Logic for a process in the application	ML Algorithms.
3.	Application Logic-2	Logic for a process in the application	IBM Watson STT service
4.	Dataset	Data Type, Configurations etc.	Dataset used for this project is downloaded from Kaggle.
5.	Cloud Database	Database Service on Cloud	IBM DB2, IBM Cloudnet etc.
6.	File Storage	File storage requirements	IBM Block Storage or Other Storage Service or Local Filesystem
7.	Machine Learning Model	Purpose of Machine Learning Model	Classification and Regression model
8.	Infrastructure (Server / Cloud)	Application Deployment on Local System / Cloud Local Server Configuration: Cloud Server Configuration :	Local, Cloud Foundry, Kubernetes, etc.

Table-2: Application Characteristics:

S.NO	Characteristics	Description	Technology
1.	Scalable Architecture	Water quality index (WQI) and water quality Classification (WQC) are accurately predicted.	Surface water quality assessment tool will be used here
2.	Availability	Our model will keep working and be available for work even if there is infrastructure failure.	Machine learning
3.	Performance	The system effectively compares the input parameters given by the users with the dataset	Digital twin technology

RESULT



Thus the real-time measured values are entered and the Water Quality index and its quality are obtained as output.

ADVANTANGE & DISADVANTAGE

- **ADVANTAGES**

Reliable one with the prediction accuracy.
The future of water quality modeling seems to be very bright and remarkable
Elective technique for AI to foresee water quality utilizing negligible and effectively accessible water quality boundaries.
This project can be used in urban areas to predict the quality of the drinking water thereby preventing the spread of diseases such as dysentery, typhoid and cholera due to consumption of contaminated water.
System is low cost and efficient.

DISADVANTAGES

There needs to be a more user-centric approach towards tackling the water quality issues, by using user friendly tools and an interactive environment so that the solution actually benefits in tackling water quality issues.
Not all models have been able to numerically predict the magnesium absorption ratio (MAR) and the permeability index (PI), so classification models may be able to improve the accuracy of predictions.
Internet Connectivity and times may be a problem, since data won't be updated.

CONCLUSION

- Water is one of the most important resources for survival, and its quality is measured using WQI (Water quality Index). To that purpose, most dataset-related well-known components, such as temperature, PH, and so on, are used.
-
- DO (dissolved oxygen), conductivity, BOD (biochemical oxygen demand), nitratenan, faecal coliform, and other parameters were measured.
-
- We can analyse drinking water by designing and implementing robust hardware and software.
- We must train the dataset using the Random Forest Regression technique to notice the gradual increase in prediction rate. Unneeded data will be rejected. Analyze the information and interpret the results. Finally, we test the acquired data and forecast the water's good state.
- We assess the models' accuracy using the test dataset created in the previous

FUTURE SCOPE

- **In future work, we suggest incorporating the findings of this study into a large-scale IoT-based online monitoring system employing only the essential parameter sensors. The water would be predicted by the tested algorithms.**
-
- **Quality is promptly determined based on real-time data from the IoT system.**
- **It would detect low-quality water before it was released for human consumption and notify the appropriate authorities. It will presumably reduce the number of individuals who drink contaminated water and, as a result, deescalate dreadful illnesses like typhoid and diarrhoea. In this sense, the use of a prescriptive analysis based on projected values would result in future capabilities to assist decision and policy makers.**

APPENDIX

CODE:
https://github.com/IBM-EPBL/IBM-Project-21994-1659800695

DEMO LINK:
https://drive.google.com/file/d/1TdY2sitNSBw9-agtZCjgRrY6Lm26efqt/view?usp=share_link