CORPORATE EMPLOYE ATTRITION ANALYTICS

NALAITHIRAN PROJECT BASED LEARNING

ON

EFFICIENT WATER QUALITY ANALYTICS AND PREDICTION USING MACHINE LEARNING

A PROJECT REPORT
BY

JINO MON EM - 19110040

MATHAVAN G - 19110054

ABINESH - 19110003

MOHAMMAD IQBAL - 19110055

BACHELOR OF TECHNOLOGY

IN

INFORMATION TECHNOLOGY

HINDUSTHAN COLLEGE OF ENGINEERING AND TECHOLOGY
Approved by AICTE, New Delhi, Accredited with ‘A’ Grade by NAAC
(An Autonomous Institution, Affiliated to Anna University, Chennai)
COIMBATORE – 641032

November 2022

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBERVIATIONS

WQI	Water Quality Index
WQC	Water Quality Class
MLR	Mutlivariate Linear Regression
RMSE	Root Mean Square Error

ABSTRACT

Water is considered a vital resource that affects various aspects of human health and life. 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.

CHAPTER 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. 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 and 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.

Water quality has a direct impact on public health and the environment. Water is used for various practices, such as drinking, agriculture, and industry. Recently, development of water sports and entertainment has greatly helped to attract tourists. Water quality monitoring is a must to keep a reliable and safe water supply. Water contamination has become increasingly significant as the economy has grown and urbanization has expanded. A survey has done withthis in mind, that investigates some supervised machine learning (ML) algorithms for estimating the water quality index (WQI), which is a single index that describes and general quality ofwater, which is a different class based on the WQI. pH, Temperature, total dissolved solids, and turbidity are four input parameters used in the suggested methodology.

1.2 Project Overview

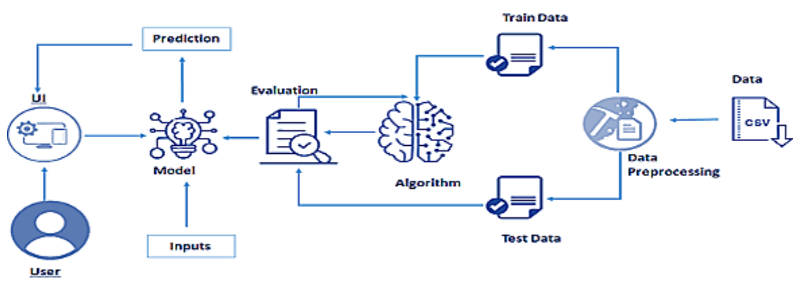

Fig 1.1 Project Overview

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. A representative set of supervised machine learning algorithms were employed on the said dataset for predicting the water quality index. 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 best WQI.

1.2 Purpose

Water quality is normally determined by a set of physical and chemical parameters that are closely related to the water's intended usage. The acceptable and unacceptable values for each variable must then be established. Water that meets the predetermined parameters for a specific application is considered appropriate for that application. Data on training pH and hardness testing data Solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes, turbidity, and potability are all terms that can be used to describe something. The definition of theclass labels for each data sample present in the data are all factors that go into selecting the water quality data set, which is a prerequisite to model construction. The more challenging method is tocombine the values of a group of physical and chemical variables into a single value. A quality value function (usually linear) represented the equivalence between the variable and its quality level was included in the index for each variable. These functions were created using direct measurements of a substance's concentration or the value of a physical variable derived from water sample studies. The major goal of this research is to examine how machine learning

algorithms may be used to predict water quality. The proposed methodology employs four input parameters, namely, temperature, turbidity, Ph and total dissolved solids. 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.

CHAPTER 2

LITERATURE SURVEY

2.1 Existing Problem

The basic idea of this research is to devise a comprehensive methodology that analyzes and predicts the water quality of particular regions with the help of certain water quality parameters. These parameters include physical, biological, or chemical factors which influence water quality. There are certain quality standards set up by international organizations like the World Health Organization (WHO) and the Environmental Protection Agency (EPA), which serve as a benchmark for determining the quality of water. In its document “Efficient Water Quality Analysis and Prediction using Machine Learning”, EPA mentions a total of 101 parameters that affect water quality in one way or another. However, some parameters have a greater and more visible effect on water quality than others.

TITILE: Improving The Robustness Of Beach Water Quality Modeling Using An Ensemble Machine Learning Approach

AUTHOR: Wang et al (2021)

This study demonstrates the utility of using a model stacking approach for predictive modeling of beach water quality. Since model stacking averages out noise from its base models, it is theoretically more promising than individual models in generating predictions with greateraccuracy and robustness. The results from this study suggest that the model stacking algorithm has promise for improving the reliability of predictive modeling for beach microbial water quality of other sites with similar hydrogeological and environmental conditions such as other beaches along the Great Lakes. A comprehensive test needs to be done to understand the strength and weaknesses of individual base models and the stacking approach.This study indicatedthat the model stacking approachmay improve the robustness of beach water quality modeling.

TITILE: Application of Artificial Neural Network In Water Quality Index Prediction
AUTHOR: M Yilma, Z Kiflie, A Windsperger, and N Gessese (2018)

The Little Akaki River is one of the most polluted Rivers in Ethiopia as reported in many studies. These studies, however, mainly used concentration measurements of certain constituents and compared them against local and international standards. The multitude of objectives and differences in selected constituents in the various reports require scientific knowledge to understand the River water quality status. The water quality index is a useful method to get a summary of water bodies' pollution extent. To this end, the Canadian Council of Ministers of the Environment-water quality index approach was used. Furthermore, modeling of the index was performed using a trained and validated artificial neural network. Twelve water quality parameters from 27 sampling sites in the Dry season (January/February 2017) and Wet season (October/November 2015) were used for index determination. Results show that all samplingsites except one site upstream were under the poor water quality category.

TITILE: Improving Prediction of Water Quality IndicesUsing Novel Hybrid Machine-LearningAlgorithms
AUTHOR: DT Bui, K Khosravi, J Tiefenbacher, H Nguyen, N Kazakis (2020)

River water quality assessment is one of the most important tasks to enhance water resources management plans. A water quality index (WQI) considers several water quality variables simultaneously. Traditionally WQI calculations consume time and are often fraught with errors during derivations of sub-indices. In this study, 4 standalone (random forest (RF), M5P, random tree (RT), and reduced error pruning tree (REPT)) and 12 hybrid data-mining algorithms (combinations of standalone with bagging (BA), CV parameter selection (CVPS) and randomizable filtered classification (RFC)) were used to create Iran WQI (IRWQIsc) predictions. Six years (2012 to 2018) of monthly data from two water quality monitoring stations within the Talar catchment were compiled. The models were evaluated using several statistical and visual evaluation metrics. The result shows that fecal coliform (FC) and total solids (TS) had the greatest and least effect on the prediction of IRWQIsc.

TITILE: Prediction of Water QualityParameters Using Anfis Optimized By Intelligence Algorithms

AUTHOR: An Azad, H Karami, S Farzin, A Saeedian, HKashi, F Sayyahi (2018)

Water quality management and control have high importance in the planning and development of water resources. This study investigated the application of Genetic Algorithm (GA), Ant Colony Optimization for Continuous Domains (ACOR), and Differential Evolution (DE) in improving the performance of adaptive neuro-fuzzy inference system (ANFIS), for evaluating the quality parameters of Gorganroud River water, such as Electrical Conductivity (EC), Sodium Absorption Ratio (SAR) and Total Hardness (TH). Accordingly, initially most suitable inputs were estimated for every model using sensitivity analysis and then all of the quality parameters were predicted using mentioned models. Also, ANFIS-DE and ANFIS-GA models had the best performance in prediction of SAR in test stage. It is noteworthy that ANFIS showed the best performance in the prediction of all mentioned water quality parameters in training stage. The results indicated the ability of mentioned algorithms in improving the accuracy of ANFIS for predicting the quality parameters of river water.

TITILE: Integrating Water Quality and Operation into A Prediction of Water Production in Drinking Water Treatment Plants by GeneticAlgorithm Enhanced

AUTHOR: Y Zhang, X Gao, K Smith, G Inial, S Liu, LBConil, et al (2019)

Stringent regulations and deteriorating source water quality could greatly influence the water production capacity of drinking water treatment plants (DWTPs). Using models to predict the performance of DWTPs under stress provides valuable information for decision making and future planning. A hybrid statistic model named HANN was established by combining an artificial neural network (ANN) with a genetic algorithm (GA) aiming at forecasting the overall performance of DWTPs nationwide in China. Water quality parameters like temperature and chemical oxygen demand (COD) and operational parameters like electricity consumption and chemical consumption were selected as input variables while drinking water production was employed. The scenario analysis showed that the HANN model was capable of predicting water production variation based on the parameter variations, indicating that the HANN model could bea general management tool for decision-makers and DWTP managers to make plans in advance of regulatory changes, source water quality variations and market demands.

2.1.1 COMPARATIVE ANALYSIS OF LITERATURE SURVEY

S.No.	ARTICLE NAME	AUTHOR NAME	PUBLISHED YEAR	DRAWBACKS
1	Improving the robustness of beach water quality modeling using an ensemble machine learning approach	L Wang, Z Zhu, L Sassoubre, G Yu, C Liao, Q Hu, et al.	2021	The model is built only to predict beach water quality
2	Application of artificial neural network in water quality index prediction: a case study in Little Akaki River, Addis Ababa, Ethiopia.	M Yilma, Z Kiflie, A Windsperger, and N Gessese	2018	Usage of ANN model which can be replaced by simple ML model with the same level of accuracy.
3	Improving prediction of water quality indices using novel hybrid machine learning algorithms.	DT Bui, K Khosravi, J Tiefenbacher, HNguyen, and N Kazakis	2020	The model is built with high complexity but gives the same accuracy level as simple models can.
4	Prediction of Water Quality Parameters Using ANFIS Optimized by Intelligence Algorithms (Case Study: Gorganrood River)	An Azad, H Karami, S Farzin, A Saeedian, H Kashi, and F Sayyahi	2018	It is specific to Gorganroud River water
5	Integrating water quality and operation into a prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network	Y Zhang, X Gao, K Smith, G Inial, S Liu, LB Conil, etal.	2019	Usage of ANN model which can be replaced by simple ML model with the same level of accuracy

Table 2.1 Comparative Analysis Of Literature Survey

2.2 References

1. L Wang, Z Zhu, L Sassoubre, G Yu, C Liao, Q Hu, et al., Improving the robustness of beach water quality modeling using an ensemble machine learning approach, Science of The Total Environment, Vol. 765, 2021, pp. 142760.
2. M Yilma, Z Kiflie, A Windsperger, and N Gessese, Application of artificial neural network in water quality index prediction: a case study in little Akaki River, Addis Ababa, Ethiopia, Modeling Earth Systemsand Environment, Vol. 4, 2018, pp.175-187.
3. DT Bui, K Khosravi, J Tiefenbacher, H Nguyen, and N Kazakis, Improving prediction of water qualityindices using novel hybrid machine-learning algorithms, Science ofTheTotal Environment, Vol. 721, 2020, pp. 137612.
4. An Azad,H Karami, S Farzin, A Saeedian, H Kashi, and F Sayyahi,Prediction of waterquality parameters using ANFIS optimized by intelligence algorithms (case study: Gorganrood river), KSCE Journal of Civil Engineering, Vol. 22, 2018,pp. 2206-2213.
5. Y Zhang, X Gao, K Smith, G Inial, S Liu, LB Conil, et al., Integrating water quality and operation into a prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network, Water Research, Vol. 164, 2019, pp. 114888.

2.3 Problem Statement Definition

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. Rapid industrialization has consequently led to the 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. Water quality is currentlyestimated through expensive and time-consuming lab which require sample collection, transport these samples to the lab collected from one of the water sources and it takes a considerable amount of time for the calculation of results, which is quite ineffective if the water is polluted with waste that causes diseases. The main motivation is to propose and evaluate an alternative method based on Machine Learning for the efficient analysis and prediction of water quality in real time.

CHAPTER 3

IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas

An **empathy map** is a collaborative visualization used to articulate what we know about a particular type of user. It externalizes knowledge about users in order to

1) Create a shared understanding of user needs

2) Aid in decision making.

Fig 3.1 Empathy Map

Empathy maps are split into 4 quadrants (*Says*, *Thinks*, *Does*, and *Feels*), with the user or persona in the middle. Empathy maps provide a glance into who a user is as a whole and are **not** chronological or sequential.

a. The **Says/Hear** quadrant contains what the user says out loud or hear or some other usability study related to ideology.

- Water is the domestic need we must acquire it properly.
- Quality of water based on their surroundings.
- Human body can accept only pure water of ph 7.35-7.4

b. The **Feels/See** quadrant is the user's emotional state
Ask yourself: what worries the user? What does the user get excited about? How does the user feel about the experience?

- To aware the element present in the water.
- To have a clean drinking water.
- To have a healthy life.
- Know the quality of water.

c. The **Thinks** quadrant captures what the user is thinking about water quality efficient.

- How accuracy the quality if water?
- What was the pH quality of the water?

d. The **Does** quadrant encloses the actions the user takes,

What does the user physically do?, How does the user go about doing it?

- Find the elements present inside the water.
- Find the range for the quality of drinking water.

e. **Goal** quadrant

- All water can be tested and its load helathy life.
- Quality of water can be determined
- Analyse the perfect water for domestic usage

3.2 Ideation & Brainstorming

Brainstorm and idea prioritization involves three steps:

- Team gathering, collaboration, select the problem statement
- Brainstorm, idea listing, grouping
- Idea prioritization

i) Team gathering, collaboration, select the problem statement

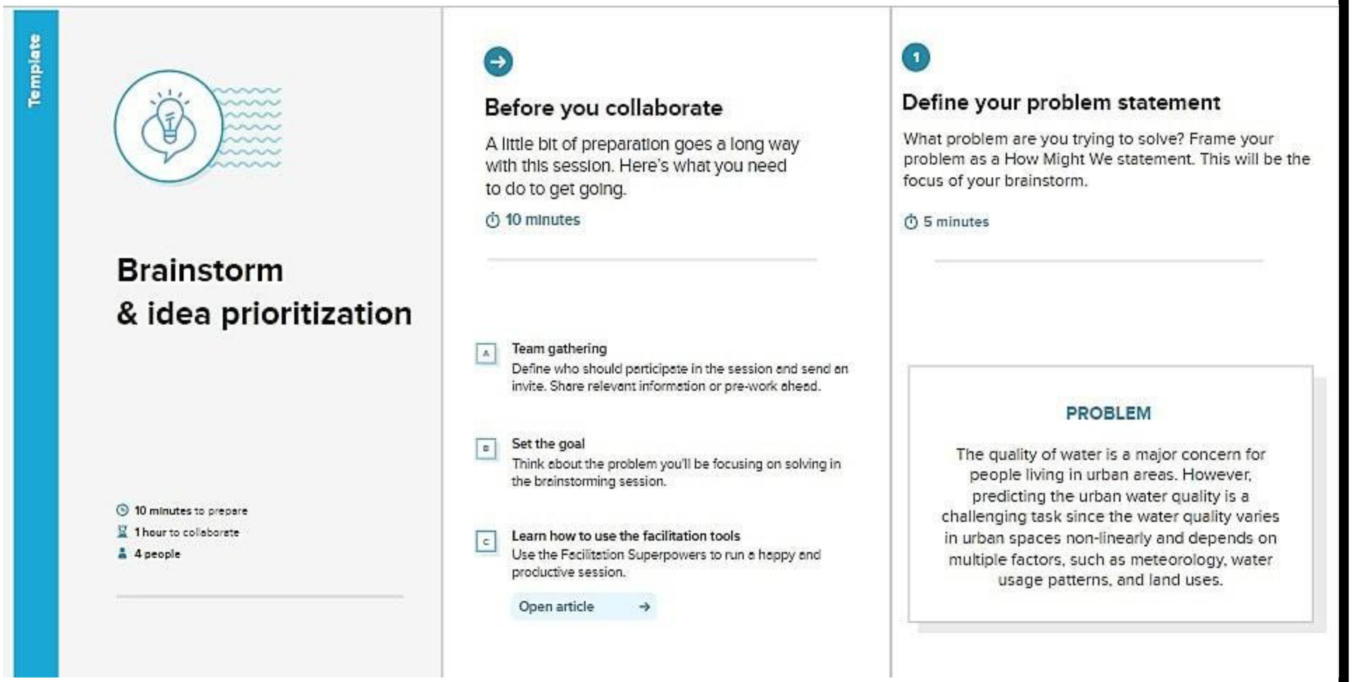

Fig 3.2 Team Gathering, Problem Statement

The problem statement is that the quality of water is a major concern for people in urban areas. However, predicting the urban areas of 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 use.

ii) Brainstorm

Fig 3.2.1 Brainstorm

Brainstorming is a group problem-solving method that involves the spontaneous contribution of creative ideas and solutions. This technique requires intensive, freewheeling discussion in which every member of the group is encouraged to think aloud and suggest as many ideas as possible based on their diverse knowledge.

iii) Group Ideas

Take turns sharing your Iideas while clustering similiar or related notes as you go. Once all sticky not have been grouped, give each cluster a sentence-like kabel. If a cluster is bigger than six sticky notes, try and see if you and break it up into similar sub-groups

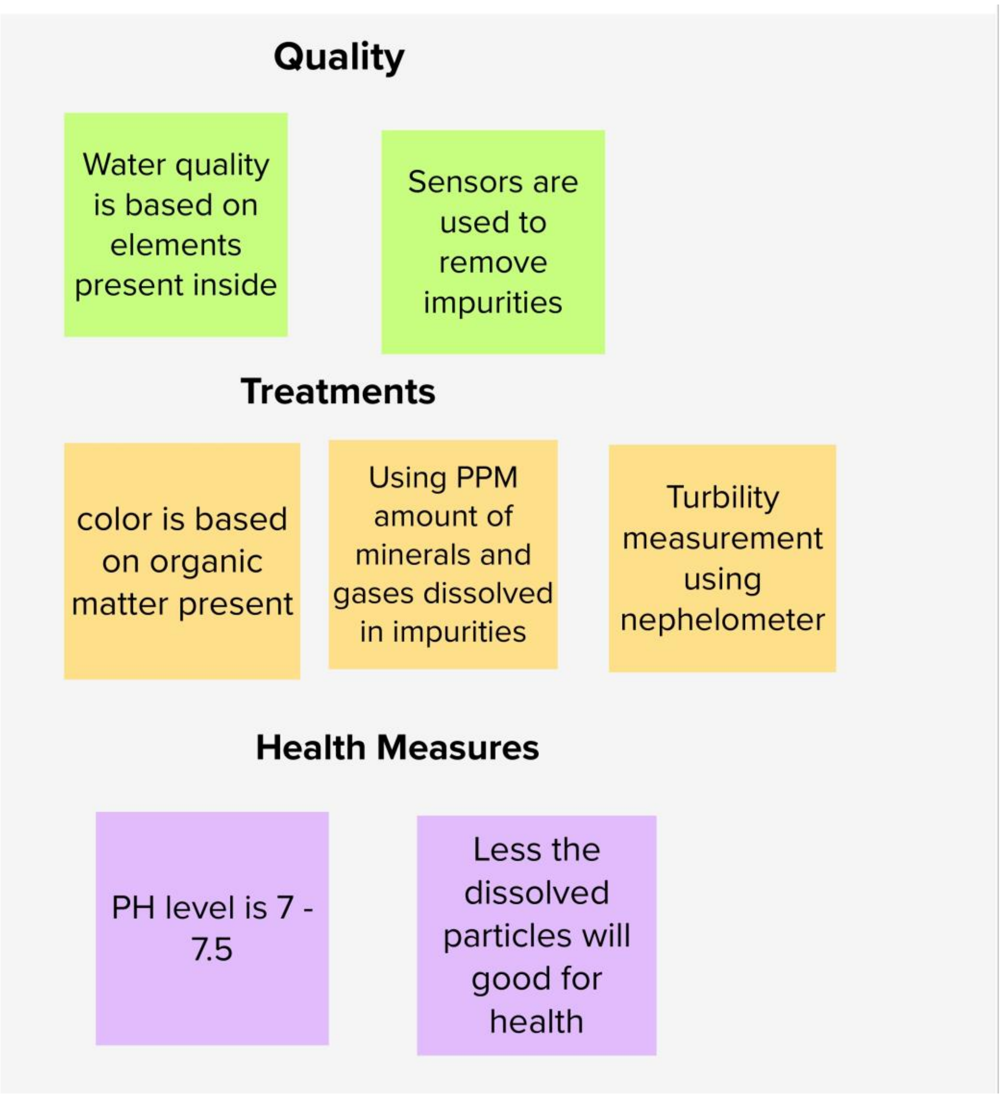

Fig 3.2.2 Group Ideas

iv) Idea Prioritization

Fig 3.3.3 Idea prioritization

Fig 3.2.3 Idea Prioritization

Idea prioritization is just a part of the idea management process. Having a structured idea management process and a systematic way of gathering, evaluating and prioritizing new ideas takes time. To make it work, the entire idea management process should be integrated to the everyday ways of working.

3.3 Proposed Solution

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	The quality of water is a major concern for people living in urban areas. 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. Water quality is currently estimated through expensive and time-consuming labs which require sample collection, and transport of these samples to the lab collected from one of the water sources and it takes a considerable amount of time for the calculation of results which is quite ineffective if the water polluted withwaste that causes diseases.
2.	Idea / Solution description	To solve the above problem, we may use machine learning algorithms to predict the quality of water by considering all water quality standard indicators, and to increase accuracy we may use ensemble techniques and PCA to reduce complexity in learning
3.	Novelty / Uniqueness	The model can be used to determine whether the water possess the quality of drinking water. Thus helpful in ensuring healthy lives.
4.	Social Impact / Customer Satisfaction	The quality of water serves as a powerful environmental determinant and a foundation for the prevention and control of waterborne diseases. So this project aims at building a Machine Learning (ML) model to Predict Water Quality by considering all water quality standard indicators. Thus, it helps people to ensure the standard of health concerning drinking water.
5.	Business Model (Revenue Model)	Our model is capable to help branded water bottle companies ensure water quality and decides whether further purification of water is needed or not.
6.	Scalability of the Solution	It is expected that our model helps in getting all require aspects of water

Table 3.1 Proposed Solution

3.4 Problem Solution Fit

Customer Segments People,Residential,Commerical, Lab testing	**Customer Constraints** Water is essential for every one to sustain. If the water is not in good quality it may cause diseases. The Disease caused by the impure water is avoided with this application	**Available Solutions** The available solution is finding water quality index and water quality class.With the help of WQI and WQC the pH level of water can be calculated.
Jobs-to-be-done / Problems ● Check the quality of water ● Check whether the water is usable or not ● Gives the reason for un-usability ● Customer can check the water quality by themselves without expert's support	**Problem Root Cause** ● Identity appropriate Solution ● Collect sufficient amount of data ● Identity the associated casual factor	**Behaviour** The study attemptsto assess the users water use behaviour using available resoureces, prevailing socio-economics conditions and personal aspects of users. This research work suggests the need for ensuring water quality is important before use.
Triggers With the help of this application users can avoid the fesr of water quality. Since the user knows the quality of water they are goning to use, they can avoid most of the health issues that are caused by poor quality water	**Your Solution** The data from different sources are taken and with help of a water quality analyst we will be getting an idea about the constraints where we start to get the data and preprocess it. By using some ML algorithms and some analysis methods the hardness, conductivity and turbiditiy are identified and results are provided based on this application	**Channels of Behaviour** **Online :** Through advertising in social media, news platform makes cusmoter to know and realized the importance of monitoring the level of water quality that we consume for our needs and to provide awareness about the need for measuring the water quality level
Emotions: Before / After Before there is no technology to analysis the quality of water, so there is some fear with the qaulity of water.	**After** It is easy to calculate the water quality with the help of this application	**Offline :** ● By attaining the standard quality of satisfy all parameters is consider as pure water ● Words of mouth among customers

Fig 3.4 Problem Solution Fit

CHAPTER 4

REQUIREMENT ANALYSIS

4.1 Functional Requirements

FR No.	Functional Requirements (Epic)	Sub Requirements (Story / Sub-Task)
FR-1	User Interface	The detailed description about waterquality should be provided.
FR-2	User Form	Values and measures require to predict the water quality should be given as input in the form.
FR-3	Machine Learning Model employment	Develop the Machine Learning Regression Model to predict the Water Quality Index (WQI). Develop the Machine Learning Classification Model to predict the Water Quality Classification (WQC).
FR-4	Testing The WaterSamples	Provides an option to test any kind of water samples with required parameters and to calculate the Water Quality Index and impurities present
FR-5	Reporting	If anyissues are faced by the customer or user it will be directly notified to the developer.

Fig 4.1 Functional Requirements

4.2 Non-Functional Requirements

NFR No.	Non-Functional Requirement	Description
NFR-1	Usability	Customers can access the system more efficiently and in a simpler way. The customers can have the opportunity to view a better interpretation of results. The customers arealso recommended with the purification techniquesbasedon the impurities.
NFR-2	Security	All the predicted information is accessed only by the authenticated users
NFR-3	Reliability	It should be reliable in producing effective and efficient water quality prediction results. It should ensure the trust and belief among people thatthis water quality prediction system producescorrect results whenused.
NFR-4	Performance	The system should be consistent in producing the prediction results of Water Quality Index (WQI) and also needs to ensure better throughput and response time compared to othersystems.
NFR-5	Availability	The system can be utilised by the customers 24/7 and it shouldbe availed to test anykind of water samples anywhere
NFR-6	Scalability	It can be used by wide variety of users like testing agencies, private and publiclaboratories, restaurants and hotels and people who wish to test the quality of water they consume. The system should also be compatible enough so as to be integrated with the future technologies also.

Fig 4.2 Non - Funactional Requirements

CHAPTER 5

PROJECT DESIGN

5.1 Data Flow Diagram

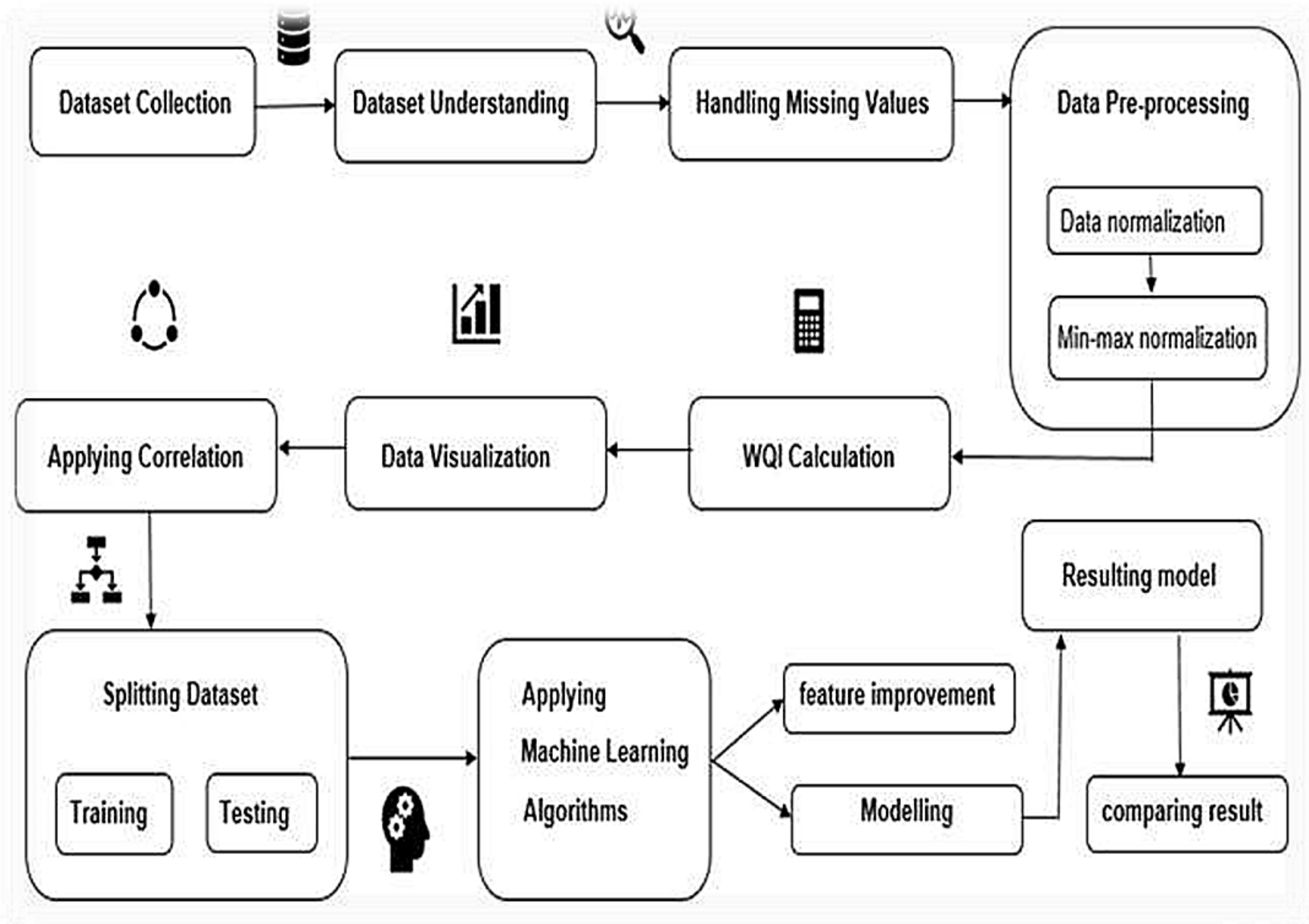

Fig 5.1 Data Flow Diagram

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 & Technical Architecture

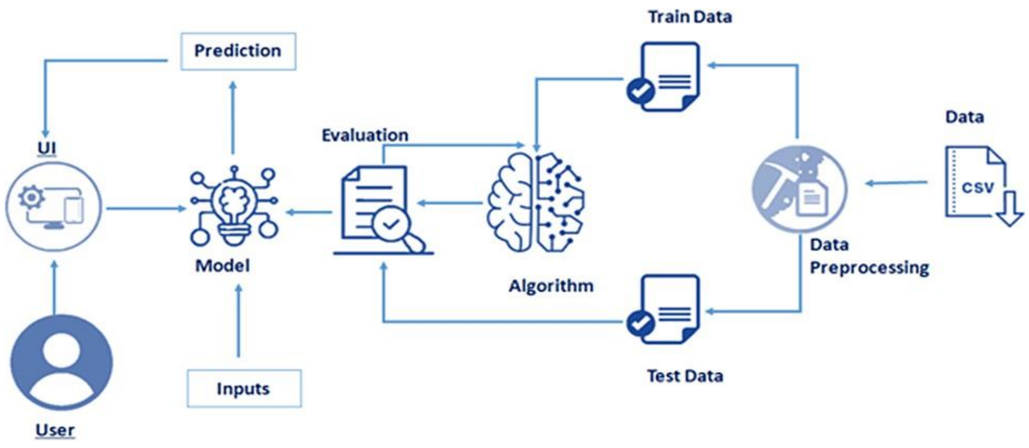

Fig 5.2 Solution Architecture

5.3 User Stories

User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
People (web user)		USN-1	As a user, I can understand the detailed description of water quality on the home page	I can access the web page	High	Sprint-1
	Input form	USN-2	As a user, I can enter the details required to analysis the water quality with use of form provided in the web page.	I can give inputs in the form and it is processed and visualize the water quality.	High	Sprint-2
		USN-3	As a user, I can contact the Customer care (people at the water resource organisation) to know the details of water	I can contact people with Whatsapp, instagram, twitter, mail and also I can make call	Medium	Sprint-3

Fig 5.3 User Stories

CHAPTER 6

PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

Sprint	Functional Requiremet (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	Dharaneesh S K
Sprint-1	Data Preprocessing	USN-2	Used to transform the data into useful format.	7	Medium	Dhanish S
Sprint-2	Model Building	USN-3	Calculate the WaterQuality Index (WQI).	10	High	Kaviya K K
Sprint-2	Model Building	USN-4	Splitting the Modelinto Training and Testing from the overall dataset.	7	Medium	Gowri J
Sprint-3	Trainingand Testing	USN-5	Train the Modelusing Regression algorithm and Testing thePerformance of themodel.	10	High	Gowri J
Sprint-3	Application Building	USN-6	Build the HTML and Python code	7	Medium	Dhanish S, Gowri J
Sprint-4	Implementation of the Application	USN-7	Run Flask App	10	High	Dharaneesh S K
Sprint-4	Implementation of the Application	USN-8	Deploy the Modelon IBM Cloud.	7	Medium	Kaviya K K

Fig 6.1 Sprint Planning

Project Tracker,Velocity

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 Oct2022	29 Oct2022	8	29 Oct2022
Sprint-2	10	6 Days	31 Oct2022	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

Fig 6.2 Project Tracker

VELOCITY

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

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

6.2 Reports From Jira

Fig 6.2 Burndown Chart

The priority level conveys the severity of an issue so that agents can react accordingly, it identifies the relative importance of an incident and is usually based on the impact and urgency of an issue. It help your agents prioritise issues, and identifies the required time for actions to be taken to resolve it. It shows work that has been completed in an sprint, and the total work remaining.

6.3 Sprint Delivery Schedule

S.No.	MILESTONES	ACTIVITIES	TEAM	DATE
1.	Data Collection	Download the Dataset	Dharaneesh S K	24 OCT 2022
2.	Data Preprocessing	Importing the Libraries	Dhanish S	24 OCT 2022
3.	Data Preprocessing	Reading the Dataset	Dharaneesh S K	24 OCT 2022
4.	Data Preprocessing	Analyse the Data	Dharaneesh S K	25 OCT 2022
5.	Data Preprocessing	Handling the Missing Values	Dhanish S	26 OCT 2022
6.	Data Preprocessing	Water Quality Index (WQI) Calculation	Kaviya K K	27 OCT 2022
7.	Data Preprocessing	Data Visualisation	Gowri J	28 OCT 2022
8.	Data Preprocessing	Splitting dependent and Independent Columns.	Gowri J	29 OCT 2022
9.	Data Preprocessing	Splitting thedata into Training and Testing	Gowri J	29 OCT 2022
10.	Model evaluation	Model Evaluation	Kaviya K K	31 OCT 2022
11.	Model evaluation	Save the Model	Dhanish S	05 NOV 2022
12.	Application building	Build the HTML code.	Gowri J	07 NOV 2022
13.	Application building	Build the Python code	Dhanish S	10 NOV 2022
14.	Application building	Run Flask App	Dharaneesh S K	12 NOV 2022
15.	Train the modelon IBM	Register ForIBM Cloud	Kaviya K K	14 NOV 2022
16.	Train the modelon IBM	Train theML Model onIBM	Kaviya K K	17 NOV 2022
17.	Train the modelon IBM	Integrate FlaskwithScoring Endpoint.	Gowri J	19 NOV 2022

Fig 6.3 Sprint DEvliverable Schedule

CHAPTER 7

CODING & SOLUTION

7.1 Feature (Random Forest Classifier)

The first feature of the deployement is the process of Random Forest Classifier is used to train and test the model for detecting the Efficent Wter Quality Analysis And Prediction with the help of collected and pre-processed dataset collections.

Train Test Split:

from sklearnn_test_split X_train,X.model_selection

import train_test,y_train,y_test = train_test_split(X_smote,y_smote,test_size=0.3, random_state=33)

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

RandomForestClassifier:

from sklearn.ensemble import RandomForestClassifier

RandomForest = RandomForestClassifier()

RandomForest = RandomForest.fit(X_train,y_train)

Predictions:

y_pred = RandomForest.predict(X_test)

Performance:

print('Accuracy:', accuracy_score(y_test,y_pred))

print(confusion_matrix(y_test,y_pred))

print(classification_report(y_test,y_pred))

Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting.

GradientBoostingClassifier:

from sklearn.ensemble import GradientBoostingClassifier

GradientBoost = GradientBoostingClassifier()

GradientBoost = GradientBoost.fit(X_train,y_train)

Predictions:

y_pred = GradientBoost.predict(X_test)

Performance:

print('Accuracy:', accuracy_score(y_test,y_pred))

print(confusion_matrix(y_test,y_pred))

print(classification_report(y_test,y_pred))

AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners. These are models that achieve accuracy just above random chance on a classification problem. The most suited and therefore most common algorithm used with AdaBoost are decision trees with one level.

AdaBoostClassifier:

from sklearn.ensemble import AdaBoostClassifier

AdaBoost = AdaBoostClassifier()

AdaBoost = AdaBoost.fit(X_train,y_train)

Predictions:

y_pred = AdaBoost.predict(X_test)

Performance:

print('Accuracy:', accuracy_score(y_test,y_pred))

print(confusion_matrix(y_test,y_pred))

print(classification_report(y_test,y_pred))

7.2 Feature 2 (Flask Connectivity)

Python flask is the first feature that helps to complete this project. It allows the user to create local server and host the website in a local machine.

from flask import Flask,

render_template,

request import numpy as np

import pickle import requests

import json

Here we import all the necessary features of this project involving in Python flask.

```
header = {'Content-Type': 'application/json', 'Authorization': 'Bearer ' + mltoken}
app = Flask(name) model = pickle.load(open('water.pkl', 'rb'))
@app.route('/',methods=['GET'])
def Home():
    return render_template('index.html')
```

Here we created a local client’s own server which serves the .html pages to the users

```
@app.route("/predict", methods=['POST'])
def predict():
    if request.method == 'POST':
        Age = int(request.form['Age']) Gender = int(request.form['Gender'])
        Total_Bilirubin = float(request.form['Total_Bilirubin'])
        Alkaline_Phosphotase = int(request.form['Alkaline_Phosphotase'])
        Alamine_Aminotransferase = int(request.form['Alamine_Aminotransferase'])
int(request.form['Aspartate_Aminotransferase'])
```

Total_Protiens = float(request.form['Total_Protiens'])

Albumin = float(request.form['Albumin'])

Albumin_and_Globulin_Ratio = float(request.form['Albumin_and_Globulin_Ratio'])

Values = np.array([[Age,Gender,Total_Bilirubin,Alkaline_Phosphotase,Alamine_Aminotransferase,Aspartate_Aminotransferase,Total_Protiens,Albumin,Albumin_and_Globul in_Ratio]])

prediction = model.predict(values) return render_template('result.html', prediction=prediction)

if name == "main":

app.run(debug=True)

Here we use the inputs from the html pages which has to be get by using request method in Python Flask. By validating the values from the database, we allow the user to access the home page. render_template: Used for rendering html pages on browser. url_for: Passing the control of the program to another function. session: Creates a separate session for the individual user

CHAPTER 8

TESTING

8.1 Test Case

Test Case Id	15407	Test Case Description	Efficient Water Quality Analysis and Prediction Using Machine Learning

S.No.	PREREQUISITES	TEST DATA
1.	Access to Chrome Browser	By clicking the website link
2.	Entering the details required	Details should be in a integer format
3.	Check for correct values	Data sholud be filled
4.	Application to train the model	Provide the datasets for model training

Table 8.1 Test Case

Test Scenario : Verify whether the deployed project predicts as per expected

Step	Step Details	Expected Results	Acutal Results	Pass/Fail/Not/ Executed/ Suspended
1.	Navigate to website link	Site should open	As Expected	Pass
2.	Enter the details	Details should be entered	As Expected	Pass
3.	Click Submit	Details should be entered	As Expected	Pass
4.	Output results	Result are generated	As Expected	Pass

Table 8.2 Test Scenario

8.2 User Acceptance

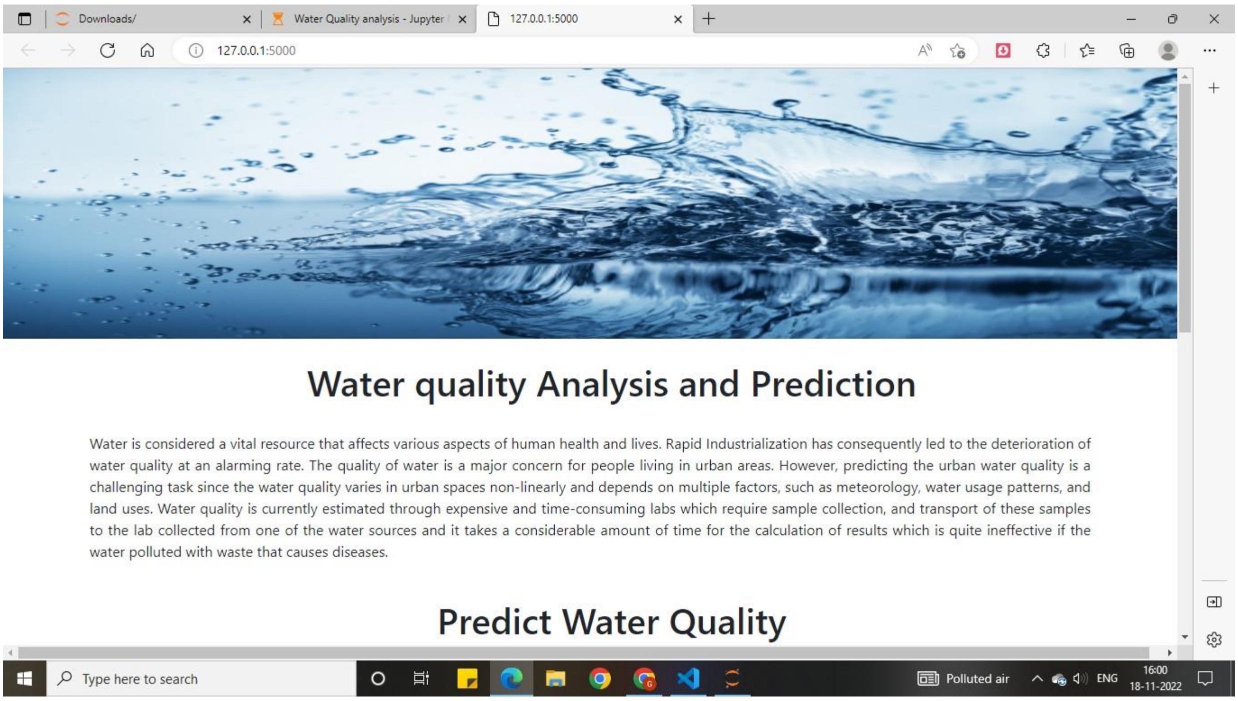

Fig 8.1 User Web Page

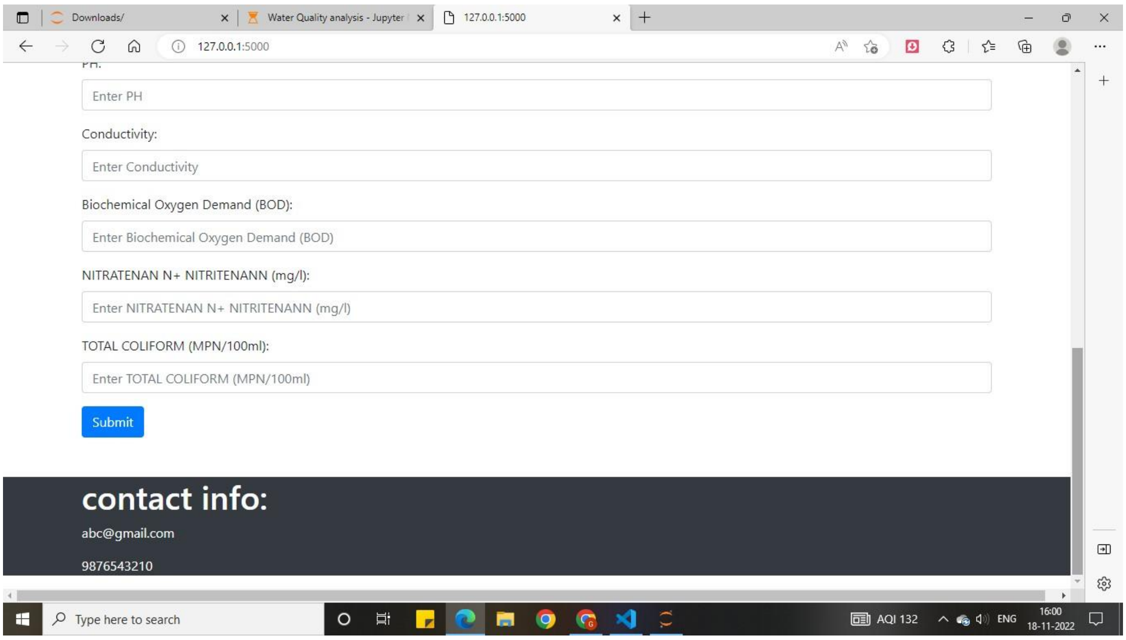

Fig 8.2 Water Quality Deatils

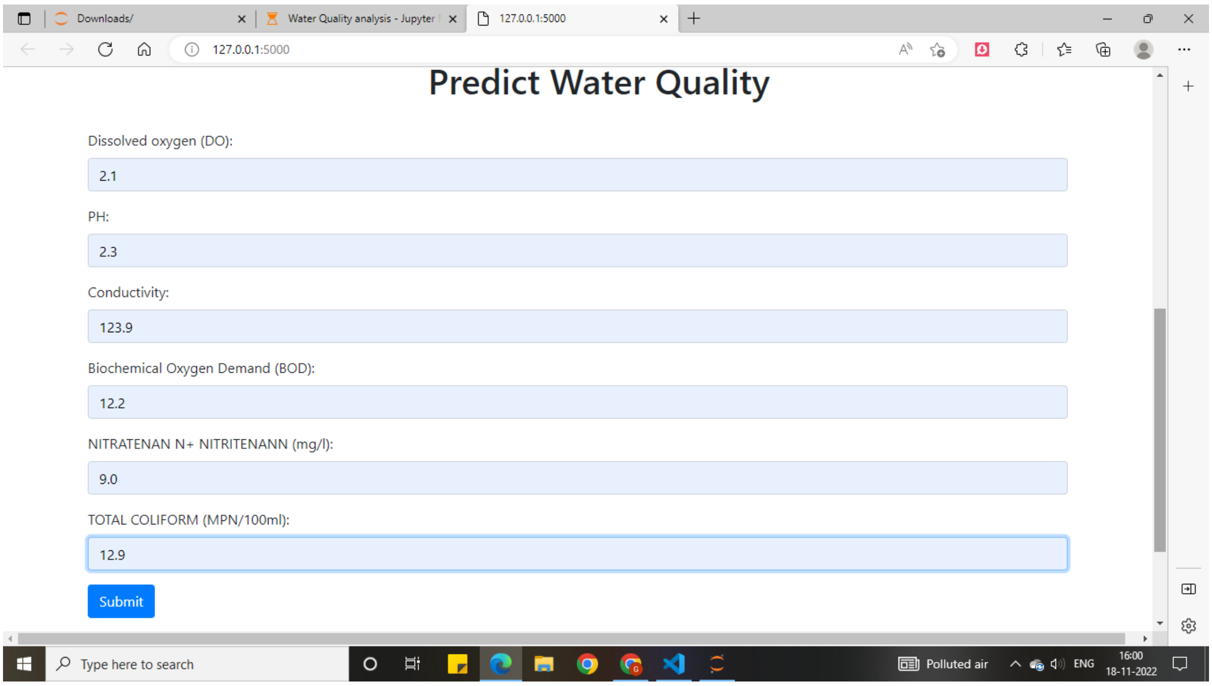

Fig 8.3 Entering the values

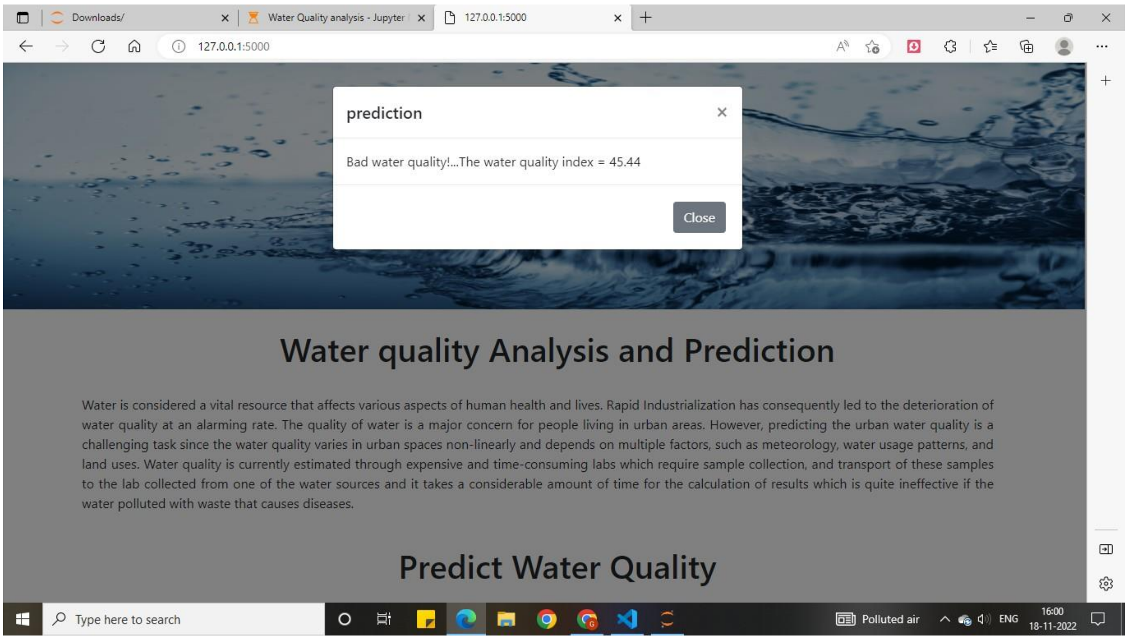

Fig 8.4 Final Prediction

In this, the data which was entered by the user will be analyzed and predicted. This above figure shows the water quality whether it is bad or good and also show the water quality index level.

CHAPTER 9

RESULTS

9.1 Performnace Matrix

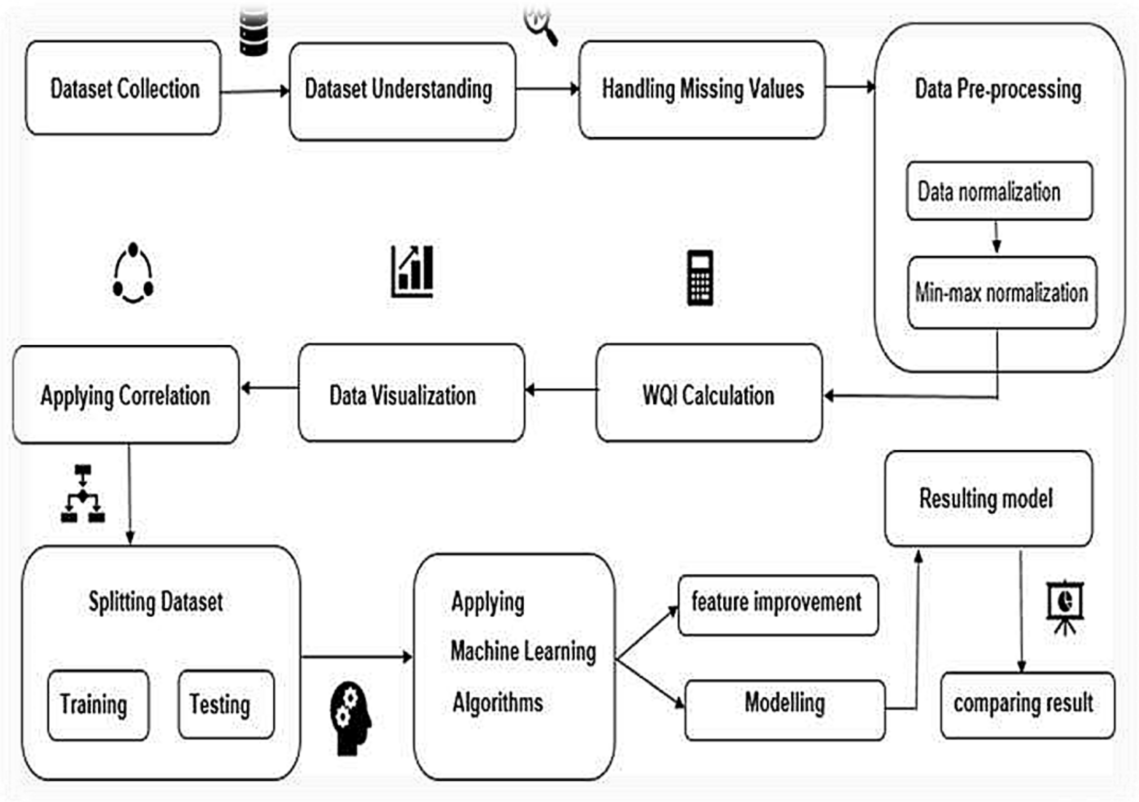

Fig 9.1 Performance Matrix

Root mean square error or root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance. To compute RMSE, calculate the residual (difference between prediction and truth) for each data point, compute the norm of residual for each data point, compute the mean of residuals and take the square root of that mean. RMSE is commonly used in supervised learning applications, as RMSE uses and needs true measurements at each predicted data point.

RMSE has a double purpose:

- To serve as a heuristic for training models
- To evaluate trained models for usefulness / accuracy

For training models, it doesn’t really matter what units we are using, since all we care about during training is having a heuristic to help us decrease the error with each iteration. We care only about relative size of the error from one step to the next, not the absolute size of the error. But in evaluating trained models in data science for usefulness / accuracy , we do care aboutunits, because we aren’t just trying to see if we’re doing better than last time: we want to know if our model can actually help us solve a practical problem. The subtlety here is that evaluating whether RMSE is sufficiently small or not will depend on how accurate we need our model to be for our given application. In this case RMSE isn’t really telling us anything about the accuracy ofour underlying model: we were guaranteed to be able to tweak parameters to get RMSE = 0 as measured measured on our existing data points regardless of whether there is any relationship between the two real quantities at all.

CHAPTER 10

ADVANTAGES & DISADVANTAGES

10.1 Advantages

- 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 water quality patterns to ensure the monitoring of the seasonal change of the water quality.

- There are several methodologies proposed for the prediction and modeling of the water quality. These methodologies include statistical approaches, visual modeling, analyzing algorithms, and predictive algorithms. For the sake of the determination of the correlation andrelationship 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 water quality environments. Thus, having models for the prediction of the water quality is of great help for monitoring water contamination.

10.2 Disadvantages

- Referring to water quality studies, an error can be defined as a value that does not represent the true concentration of a variable such as turbidity. These may arise from both human and technical error during sample collection, preparation, analysis and recording of results.

- Erroneous values can be recorded even where an organization has a clearly defined monitoring protocol. If invalid values are subsequently combined with valid data, the integrity of the latter is also impaired. Incorporating erroneous values into a management toollike a WQI or model, could result in wrong conclusions that might be costly to the environment or humans.

CHAPTER 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 representativesupervised 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 anddiarrhea. In this case, using a prescriptive analysis based on projected values would result in future capabilities to assist decision and policy makers. 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

CHAPTER 12

FUTURE SCOPE

Water is considered a vital resource that affects various aspects of human health and life. 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. Water is one of the most important natural resources for all living organisms on earth. The monitoring of treated wastewater discharge quality is vitally important for the stability and protection of the ecosystem. Collecting and analyzing water samples in the laboratory consumes much time and resources. In the last decade, many machine learning techniques, like multivariate linear regression (MLR) and Machine Learning model, have been proposed to address thesolution. The goal of an ambient water quality management program is to establish appropriate standards for water quality in water bodies receiving pollutant loads and then to ensure that these standards are met. Realistic standard setting takes into account the basin's hydrologic, ecological, and land use conditions, the potential uses of the receiving water bodies, and the institutional capacity to set and enforce water quality standards. Those models can achieve good performance in the water quality prediction. Water resources management involves the monitoring and management of water quality as much as the monitoring and management of water quantity. Various models can assist in predicting the water quality impacts of alternative land and water management policies and practices.

CHAPTER 13

APPENDIX

```
#importing libraries

import numpy as np

import pandas as pd

from sklearn import datasets

from sklearn import neighbors

#reading datset

data=pd.read_csv(r"C:\Users\gowri\Downloads\Data Collection\Dataset.csv",encoding="ISO-8859-1")

#visualizing first 5 rows of the dataset

data.head()

data.info()

data.describe()

#checking null values

data.isna().sum()

data.dtypes

data['Temp']=pd.to_numeric(data['Temp'],errors='coerce')

data['D.O. (mg/l)']=pd.to_numeric(data['D.O. (mg/l)'],errors='coerce')

data['PH']=pd.to_numeric(data['PH'],errors='coerce')

data['B.O.D. (mg/l)']=pd.to_numeric(data['B.O.D. (mg/l)'],errors='coerce')

data['CONDUCTIVITY (µmhos/cm)']=pd.to_numeric(data['CONDUCTIVITY (µmhos/cm)'],errors='coerce')

data['NITRATENAN N+ NITRITENANN (mg/l)']=pd.to_numeric(data['NITRATENAN N+ NITRITENANN (mg/l)'],errors='coerce')

data['TOTAL COLIFORM (MPN/100ml)Mean']=pd.to_numeric(data['TOTAL COLIFORM
```

```
(MPN/100ml)Mean'],errors='coerce')
data.dtypes
#initialization
start=2
end=1779
station=data.iloc [start:end ,0]
location=data.iloc [start:end ,1]
state=data.iloc [start:end ,2]
do= data.iloc [start:end ,4].astype(np.float64)
value=0
ph = data.iloc[ start:end,5]
co = data.iloc [start:end ,6].astype(np.float64)
year=data.iloc[start:end,11]
tc=data.iloc [2:end ,10].astype(np.float64)
bod = data.iloc [start:end ,7].astype(np.float64)
na= data.iloc [start:end ,8].astype(np.float64)
na.dtype
data=pd.concat([station,location,state,do,ph,co,bod,na,tc,year],axis=1)
data. columns = ['station','location','state','do','ph','co','bod','na','tc','year']
data.head()
#calulation of Ph
data['npH']=data.ph.apply(lambda x: (100 if (8.5>=x>=7) else(80 if  (8.6>=x>=8.5) or
(6.9>=x>=6.8) else(60 if (8.8>=x>=8.6) or (6.8>=x>=6.7) else(40 if (9>=x>=8.8) or
(6.7>=x>=6.5)else 0)))))
#calculation of dissolved oxygen
data['ndo']=data.do.apply(lambda x:(100 if (x>=6) else(80 if  (6>=x>=5.1) else(60 if (5>=x>=4.1)
```

```
else(40 if (4>=x>=3) else 0)))))

#calculation of total coliform

data['nco']=data.tc.apply(lambda x:(100 if (5>=x>=0)  else(80 if  (50>=x>=5) else(60 if
(500>=x>=50) else(40 if (10000>=x>=500) else 0)))))

#calc of B.D.O

data['nbdo']=data.bod.apply(lambda x:(100 if (3>=x>=0)  else(80 if  (6>=x>=3) else(60 if
(80>=x>=6) else(40 if (125>=x>=80) else 0)))))

#calculation of electrical conductivity

data['nec']=data.co.apply(lambda x:(100 if (75>=x>=0)  else(80 if  (150>=x>=75) else(60 if
(225>=x>=150) else(40 if (300>=x>=225)  else 0)))))

#Calulation of nitrate

data['nna']=data.na.apply(lambda x:(100 if (20>=x>=0) else(80 if  (50>=x>=20) else(60 if
(100>=x>=50) else(40 if (200>=x>=100) else 0)))))

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

#calculation overall wqi for each year

overall_wqi=data.groupby('year')['wqi'].mean()

overall_wqi.head()

data1=overall_wqi.reset_index(level=0,inplace=False)

data1
```

```
#visualizing the filttered data
year=data1['year'].values
AQI=data1['wqi'].values
data1['wqi']=pd.to_numeric(data1['wqi'],errors='coerce')
data1['year']=pd.to_numeric(data1['year'],errors='coerce')
import matplotlib.pyplot as plt
plt.rcParams['figure.figsize'] = (20.0, 10.0)
from mpl_toolkits.mplot3d import Axes3D
fig = plt.figure()
ax = Axes3D(fig)
ax.scatter(year,AQI, color='green')
plt.show()
data1 = data1[np.isfinite(data1['wqi'])]
data1.head()
#scatter plot of data points
cols =['year']
x=data1[cols]
y = data1['wqi']
plt.scatter(x,y)
plt.show()
import matplotlib.pyplot as plt
data1=data1.set_index('year')
data1.plot(figsize=(15,6),color="red")
```

```
plt.show()

df=data[["do","ph","co","bod","na","tc","wqi"]]

df

df.isnull().sum()

df.dropna(axis=0,inplace=True)

df.isnull().sum()

x=df.iloc[:,:-1]

x

y=df['wqi']

from sklearn.model_selection import train_test_split

x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=4)

x_train.shape

x_test.shape

from sklearn.linear_model import LinearRegression

model=LinearRegression()

model.fit(x_train,y_train)

ypred=model.predict(x_test)

ypred[[4]]

y_test

from sklearn import metrics

print(np.sqrt(metrics.mean_squared_error(y_test,ypred)))

from sklearn.tree import DecisionTreeRegressor

model = DecisionTreeRegressor(random_state = 0)
```

```
model.fit(x_train, y_train)

ypred=model.predict(x_test)

from sklearn import metrics

print(np.sqrt(metrics.mean_squared_error(y_test,ypred)))

from sklearn.ensemble import RandomForestRegressor

model = RandomForestRegressor(n_estimators = 100, random_state = 0)

model.fit(x_train, y_train)

ypred=model.predict(x_test)

from sklearn import metrics

print(np.sqrt(metrics.mean_squared_error(y_test,ypred)))

def predict(do,ph,co,bod,ni,tc):

   a=model.predict([[do,ph,co,bod,ni,tc]])

    for i in a:

      wqi=float(i)

      i=int(i)

   if i>=91 and i<=100:

      return "Excellent water quality!..."+"The water quality index = {:.2f}".format(wqi)

   elif i>=71 and i<=90:

      return "Good water quality!..."+"The water quality index = {:.2f}".format(wqi)

   elif i>=51 and i<=70:

      return "Medium water quality!..."+"The water quality index = {:.2f}".format(wqi)

   elif i>=26 and i<=50:

      return "Bad water quality!..."+"The water quality index = {:.2f}".format(wqi)
```

```
    elif i>=0 and i<=25:
        return "Very bad water quality!..."+"The water quality index = {:.2f}".format(wqi)
    else:
 return "invalid data!..."
 from flask import Flask, request, render_template, render_template_string
app = Flask(_name_)
@app.route('/', methods =["GET", "POST"])
def gfg():
    if request.method == "POST":
        do = request.form.get("do")
        ph = request.form.get("ph")
        co = request.form.get("co")
        bod = request.form.get("bod")
        ni = request.form.get("ni")
        tc = request.form.get("tc")
          key = predict(do,ph,co,bod,ni,tc)
        return  render_template("form1.html",key=key)
    return render_template("form.html")
if__name__=='__main__':
            app.run()
```

Form 1

```
<!DOCTYPE html>

<html lang="en">

<head>

  <meta charset="utf-8">

  <meta name="viewport" content="width=device-width, initial-scale=1">

  <link rel="stylesheet"
href="https://cdn.jsdelivr.net/npm/bootstrap@4.6.1/dist/css/bootstrap.min.css">

  <script src="https://cdn.jsdelivr.net/npm/jquery@3.6.0/dist/jquery.slim.min.js"></script>

  <script src="https://cdn.jsdelivr.net/npm/popper.js@1.16.1/dist/umd/popper.min.js"></script>

  <script src="https://cdn.jsdelivr.net/npm/bootstrap@4.6.1/dist/js/bootstrap.bundle.min.js"></script>

  <style>

    .size{

      width:100%;

      height:300px;

    }

    .b{

      font-size: medium;

      font-weight: 700;

    }
```

```
    body{
      max-width: 100%;
    }
  </style>
</head>
<body>
<div class="fluid-container">
  <div class="row">
   <img
src="https://th.bing.com/th/id/R.1dc2fd1484d247f496b0b32030532460?rik=pD2YAgw5rIB7kA&riu
=http%3a%2f%2fwww.hdwallpaperspulse.com%2fwp-
content%2fuploads%2f2019%2f09%2f19%2fwater-art-hd-
image.jpg&ehk=4OSDL04JOwrrK8ehwy1L0gJ8%2bnBI0MEdnMer1paUuIY%3d&risl=1&pid=Im
gRaw&r=0" class="size">
        </div>
         <br>
        <center><h1 class="ml-5">Water quality Analysis and Prediction</h1></center>
        <br>
         <div class="container text-justify"><p>Water is considered a vital resource that affects various
aspects of human health and lives.
```

Rapid Industrialization has consequently led to the deterioration of water quality at an alarming rate. The quality of water is a major concern for people living in urban areas. 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. Water quality is currently estimated through expensive and time-consuming labs which require sample collection, and transport of these samples to the lab collected from one of the water sources and it takes a considerable amount of time for the calculation of results which is quite ineffective if the water polluted with waste that causes diseases.</p>

</div>

<center><h1 class="ml-5">Predict Water Quality</h1></center>

<div class="container">

<form action="{{ url_for('gfg')}}" method="POST">

<div class="form-group">

<label>Dissolved oxygen (DO):</label>

<input type="text" class="form-control" id="do" placeholder="Enter Dissolved oxygen (DO)" name="do">

</div>

<div class="form-group">

<label>PH:</label>

<input type="text" class="form-control" id="ph" placeholder="Enter PH" name="ph">

```
</div>

<div class="form-group">

  <label>Conductivity:</label>

<input type="text" class="form-control" id="co" placeholder="Enter Conductivity" name="co">

</div>

div class="form-group">

 <label>Biochemical Oxygen Demand (BOD):</label>

    input type="text" class="form-control" id="bod" placeholder="Enter Biochemical Oxygen Demand (BOD)" name="bod">

</div>

<div class="form-group">

  <label>NITRATENAN N+ NITRITENANN (mg/l):</label>

  <input type="text" class="form-control" id="ni" placeholder="Enter NITRATENAN N+ NITRITENANN (mg/l)" name="ni">

</div>

<div class="form-group">

  <label>TOTAL COLIFORM (MPN/100ml):</label>

    <input type="text" class="form-control" id="tc" placeholder="Enter TOTAL COLIFORM (MPN/100ml)" name="tc">

</div>

<button type="submit" class="btn btn-primary">Submit</button>
```

```
        </form>
        <br>
         <br>
        </div>
        </div>
       <footer>
      <div class="fluid-container bg-dark">
        <div class="container text-white">
        <h1>
          contact info:
        </h1>
        <p>abc@gmail.com</p>
      <p>
        9876543210
      </p>        </div></div>
    </footer>
</body>
</html>
```

Github and Project Video Demo Link

Github Link:

https://github.com/IBM-EPBL/IBM-Project-21196-1659774863

Project Video Demo Link:

https://drive.google.com/file/d/1iM6hJ0L7sKRhAZmQ38re3mrhbkIv-b8D/view?usp=sharing