Efficient Water Quality Analysis & Prediction using Machine Learning

1. INTRODUCTION

Water is considered as a vital resource that affects various aspects of human health and lives. The quality of water is a major concern for people living in urban areas. The quality of water serves as a powerful environmental determinant and a foundation for the prevention and control of waterborne diseases. However predicting the urban water quality is a challenging task since the water quality varies in urban spaces non-linearly and depends on multiple factors, such as meteorology, water usage patterns, and land uses, so this project aims at building a Machine Learning (ML) model to Predict Water Quality by considering all water quality standard indicators. 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 (Jennings 2007). Among various sources of water supply, due to easy access, rivers have been used more frequently for the development of human societies. Using other water resources such as groundwater and seawater sometimes assisted with problems. For example, using groundwater without suitable recharge will lead to land subsidence (Motagh et al. 2017) and using seawater is usually associated with pollution transmission (El-Kowrany et al. 2016). Therefore, the use of rivers has attracted attention. Several investigations related to rivers around the world have been conducted and a field of engineering named river engineering has been proposed. n river engineering, studies on morphological changes, sediment transport, water quality, and pollution transmission mechanisms are very important (Julien 2002; Dey 2014). Flow structure, sediment transport and morphology of rivers are investigated in the hydraulics of rivers in river engineering (Wu 2007). The study of water quality of rivers is a common theme in earth sciences. To evaluate the quality of rivers two approaches are considered, including measuring the water quality components and defining the mechanism of pollution transmission (Kashefipour 2002; Kashefipour & Falconer 2002; Naseri Maleki & Kashefipour 2012; Qishlaqi et al. 2016). Among water quality components, measuring the dissolved oxygen (DO), chemical oxygen demand (COD), biochemical oxygen demand (BOD), electrical conductivity (EC), pH, temperature, K, Na, Mg, etc. have been proposed (Şener et al. 2017). To this end, governments have constructed hydrometry stations along rivers that cross from urban areas, agro-industrial projects, industrial estates, and rivers that join dams' reservoirs (Herschy 1993; Kejiang 1993). In hydrometry stations, the water quality components are measured and the stage-discharge relation is defined. Obtained values from hydrometry stations contain basic information for feasibility studies and development of water conservation projects. 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 (Jennings). Among various sources of water supply, due to easy access, rivers have been used more frequently for the development of human societies. Using other water resources such as groundwater and seawater sometimes assisted with problems. For example, using groundwater without suitable recharge will lead to land subsidence (Motagh et al.) and using seawater is usually associated with pollution transmission (El-Kowrany et al.). Therefore, the use of rivers has attracted attention. Several investigations related to rivers around the world have been conducted and a field of engineering named river engineering has been proposed

1.1 Project Overview

The water quality prediction system is built based on the water quality dataset collected by the SNWA. Figure 4 illustrates the workflow of the proposed wastewater quality prediction system. First, the network structure of the ANFIS model needs to be defined based on the size of input dataset. The second step is selecting proper input parameters to predict the target parameter. In this study, four input parameters are selected from each sample to predict the target parameter. Meanwhile, all collected water quality data samples are partitioned into training and testing sets based on the stratified sampling method. To find the optimum ANFIS model to solve this problem, the network structure and membership functions need to be adjusted in each loop. The root mean square error (RMSE) is utilized to evaluate the simulation results.the RMSE of the current prediction is smaller than the previous simulation error, the new model will be stored; if it is even smaller than the target threshold, the system will automatically terminate the whole process. The function of each part of the whole system is detailed in the following subsections.

1.2 Purpose

water quality testing, frequently referred to as environmental water quality testing, is a common component of an environmental impact assessment and subsequent continued environmental monitoring. Water quality testing is a key activity that must take place to characterize the quality of the environment. Water quality testing can be done to establish a benchmark before the possibility of disturbing the environment, or after a hazardous event that could have potentially threatened an ecosystem.

2. LITERATURE SURVEY

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 URL https://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 toestimate 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 realtime 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 Figure 1. Water 2019, 11, x FOR PEER REVIEW 3 of 14 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. Zhang et al. [12] have improved a hybrid artificial neural network (HANN) model by the genetic algorithm (GA) for the prediction of drinking water treatment plants in china. The model has trained, validated, and has been continually validated using monthly data from 45 DWTPs across China that comprises eleven input variables for water quality and operational performance. The HANN model has shown better ability and consistency in forecasting the total water output of DWTPs in combination with the water quality and operational factors. Their prediction shows that the HANN model has improved its performance from 0.71 to 0.93 (R2) by increasing the training data provided, as shown by the fact that the model has the ability to grow to the greatest level of performance.

2.1 Existing problem

Water quality is one of the main challenges that societies will face during the 21st century, threatening human health, limiting food production, reducing ecosystem functions, and hindering economic growth. Water quality degradation translates directly into environmental, social and economic problems. The availability of the world's scarce water resources is increasingly limited due to the worsening pollution of freshwater resources caused by the disposal of large quantities of insufficiently treated, or untreated, wastewater into rivers, lakes, aquifers and coastal waters. Furthermore, newly emerging pollutants like personal care products and pharmaceuticals, pesticides, and industrial and household chemicals, and changing climate patterns represent a new water quality challenge, with still unknown long-term impacts on human health and ecosystems.

2.2 References

https://github.com/mohammed840/Water-Quality-Prediction-machine-learning-python#readme

https://github.com/Printutcarsh/Water-Quality-Prediction

https://github.com/pydeveloperashish/Water-Quality-Prediction-using-Machine-Learning

https://www.researchgate.net/publication/336808732_Efficient_Water_Quality_Prediction_Using_Supervised_Machine_Learning

https://iwaponline.com/wqrj/article/53/1/3/38171/Water-quality-prediction-using-machine-learning

https://www.datascience2000.in/2021/10/water-quality-prediction-using-machine.html

https://github.com/siddharth271101/Water-Quality-Analysis

https://github.com/katreparitosh/Water-Quality-Analysis

https://github.com/Oggy23/Analysis-and-prediction-of-water-quality-and-water-quality-index

https://github.com/Anbulenin/Indian-Water-pollution-prediction-using-Logistic-Regression

https://www.kaggle.com/code/anbarivan/indian-water-quality-analysis-and-prediction

https://www.irjmets.com/uploadedfiles/paper/issue_5_may_2022/22391/final/fin_irjmets1651989957.pdf

https://www.sciencedirect.com/science/article/abs/pii/S0048969720362896?via%3Dihub

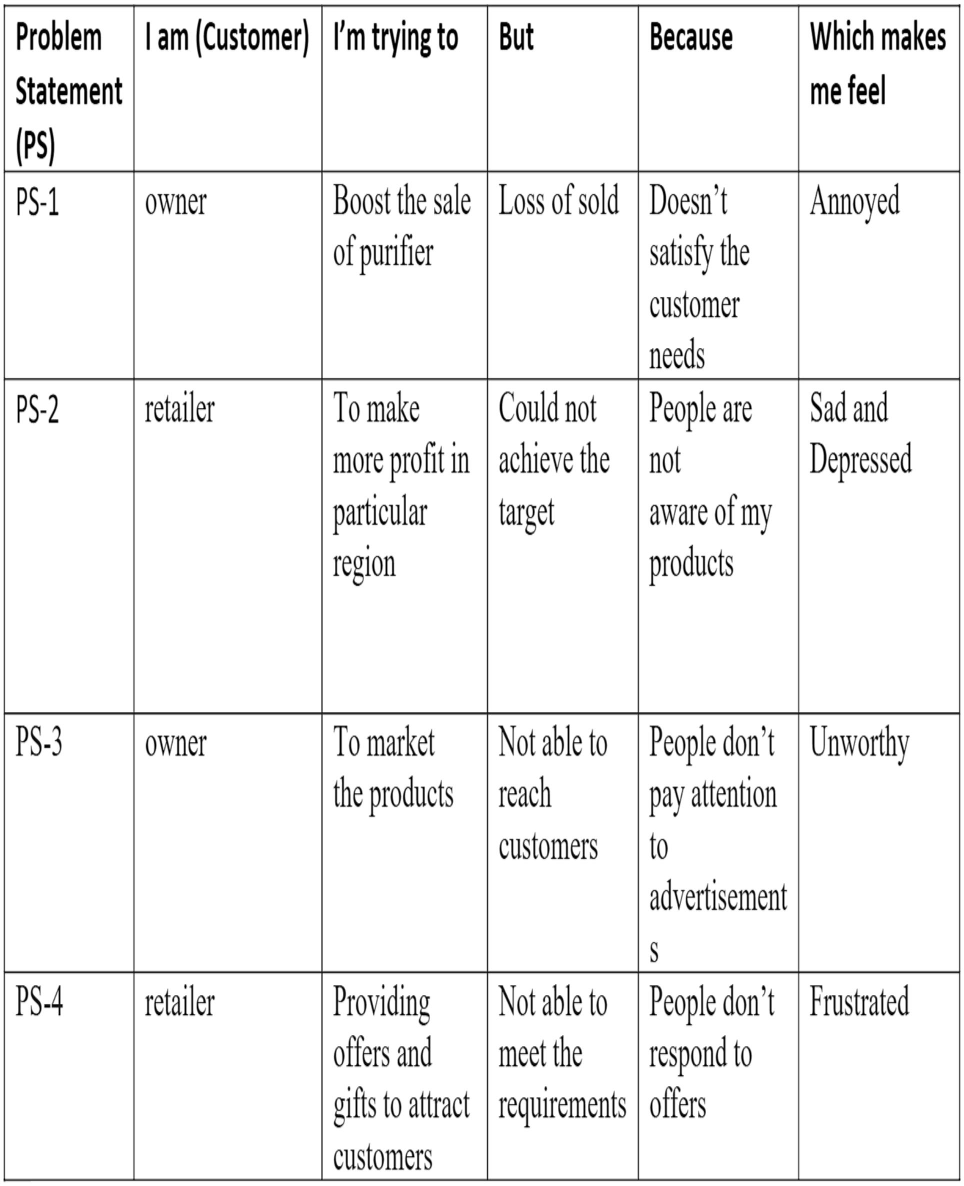
https://en.wikipedia.org/wiki/Water_quality

https://okeanus.com/news/the-purpose-of-performing-water-quality-testing

https://floridakeys.noaa.gov/ocean/waterquality.html

https://www.sciencedirect.com/science/article/abs/pii/S0048969708008231

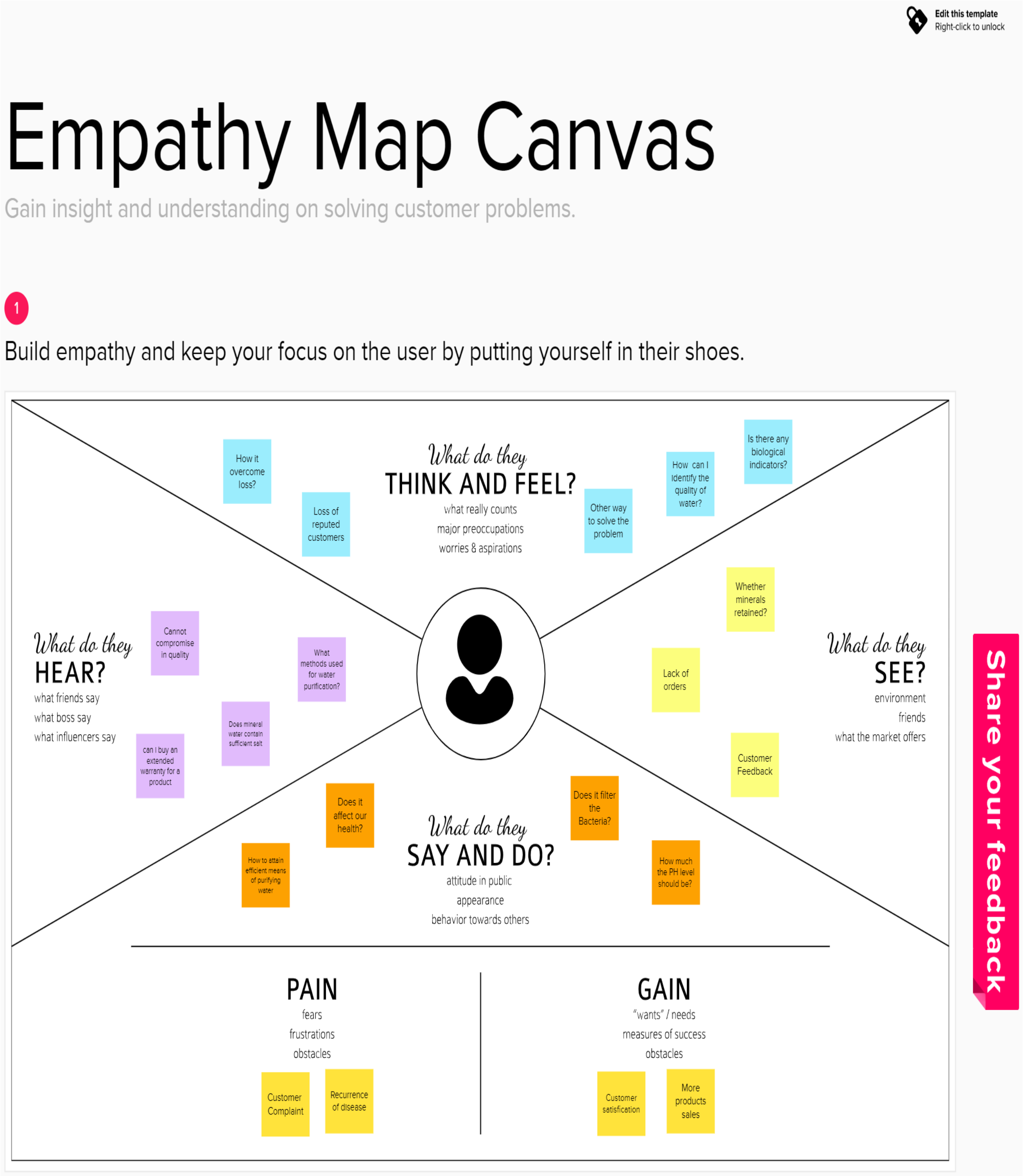
2.3 Problem Statement Definition



Problem Statement (PS)	I am (Customer)	I'm trying to	But	Because	Which makes me feel
PS-1	owner	Boost the sale of purifier	Loss of sold	Doesn't satisfy the customer needs	Annoyed
PS-2	retailer	To make more profit in particular region	Could not achieve the target	People are not aware of my products	Sad and Depressed
PS-3	owner	To market the products	Not able to reach customers	People don't pay attention to advertisement s	Unworthy
PS-4	retailer	Providing offers and gifts to attract customers	Not able to meet the requirements	People don't respond to offers	Frustrated

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas



Edit this template
Right-click to unlock

Empathy Map Canvas

Gain insight and understanding on solving customer problems.

1

Build empathy and keep your focus on the user by putting yourself in their shoes.

3.2 Ideation & Brainstorming

Problem statement:

Brainstorm for PONTUS a connected micro device for workspace improvement.

BRAINSTORM:

Lack of orders.	Improve the customer satisfication.
More product sales.	More offer increase the customer.

Key perfromances.	To think different ways to increase customer.
Encourage the employees.	To increase the purifier sales.

Make more profit.	Boost sales.
Boost profit.	Positive attitude.

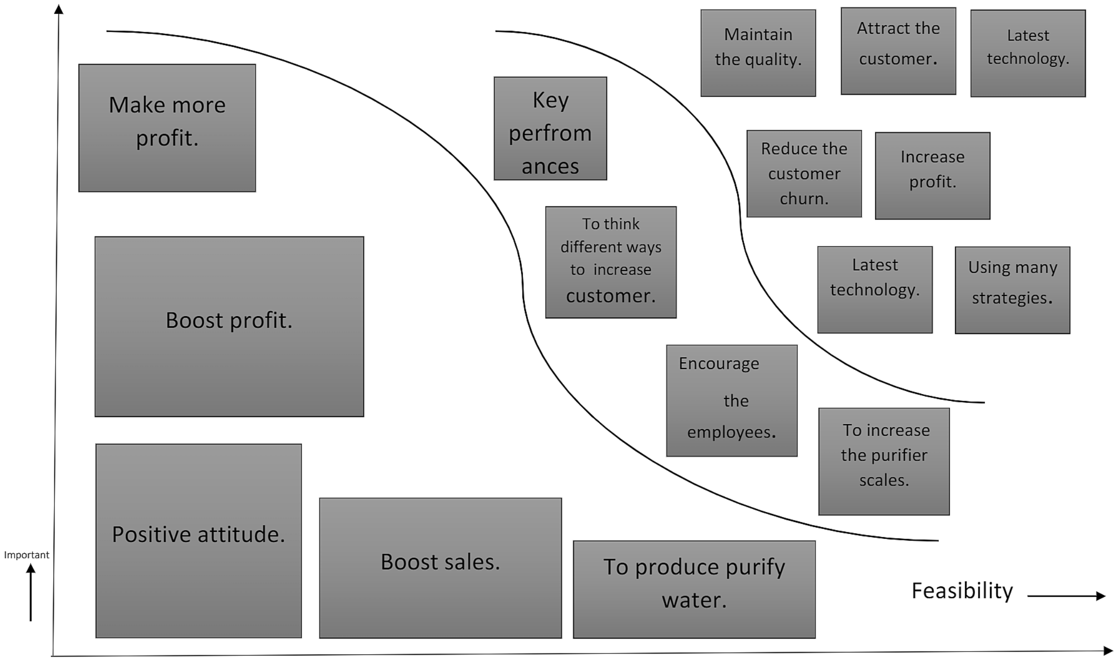
To produce the quality protect.	To produce purify water.
Water treatment and hygiene.	Implement new ideas.

Group ideas:

Maintain the quality.	Attract the customer.
Reduce the customer churn.	Increase profit.

Latest technology.	Marketing through twitter groups.
Using many strategies.	Latest technology

Prioritize:



3.3 Proposed Solution

S. No	Parameter	Description
1.	Problem Statement (Problem to be solved)	Due to the fast growing urbanization supply of safe drinking water is a challenge for the every city authority. Water can be polluted any time. So the water we reserved in the water tank at our roof top or basement in our society or apartment may not be safe. Still in India most of the people use simple water purifier that is not enough to get surety of pure water. Sometimes the water has dangerous particles or chemical mixed and general purpose water purifier cannot purify that. And it's impossible to check the quality of water manually in every time. So an automatic real-time monitoring system is required to monitor the health of the water reserved in our water tank of the society or apartment. So it can warn us automatically if there is any problem with the reserved water.

2.	Idea / Solution description	In this technique, our model predicts that the water is safe to drink or not using some parameters like Ph value, conductivity, hardness, etc. Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection.
3.	Novelty / Uniqueness	Warning when to change the filter. Detecting salt value.
4.	Social Impact / Customer Satisfaction	Customer satisfaction is an important goal in total quality management. In order to meet this goal, it is necessary to use on evaluation model for measuring the customer satisfaction level in a water supply domain. Some important criteria such as water quality, responsibility of the company, etc. are distinguished and used in the proposed model. To integrate all of these criteria in a unit index, the analytic hierarchy process technique is used.
5.	Business Model (Revenue Model)	Water is one of the essential component for human living. Water quality has a direct impact o public health and the environment. Water quality models have different information, but generally have the same purpose, which is to provide evidentiary support of water issues. Understand the material needs. Apply for carbon finance.
6.	Scalability of the Solution	The most common treatment for reducing scale formation is to **“soften” the water**. “Softening” is a process where calcium and magnesium in the water are exchanged with sodium. Commercial softeners are available either through a plumbing Equipment supplier or a water treatment professional.

3.4 Problem Solution fit

1 CUSTOMER SEGMENTS

- Loyal customer
- People who wants to avoid water borne diseases

6 CUSTOMER LIMITATIONS

- Difficult to reach the people

5 AVAILABLE SOLUTIONS

- Smart operation reminders and self diagnosis
- Alarm function

2 PROBLEMS/PAINS

- Leakage of water
- The faucet on the water purifier is making strange noise

9 PROBLEM ROOT/CAUSE

- Limit productions disrupt supply chain lead to conflict with other water users and harm corporate reputation

7 BEHAVIOUR

- Habitual buying behaviour
- Variety seeking behaviour
- Complex buying behavior

3 TRIGGER TO ACT

- To increase the purifier sales

10 YOUR SOLUTION

- Photocatalytic water purification technology
- Water treatment and hygiene

8 CHANNELS OF BEHAVIOR

- Advertising through social medias

4 EMOTIONS

- Cost for product services
- Changes in water taste

4. REQUIREMENT ANALYSIS

4.1 Functional requirement

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	User Purchasing	Protect purchased
FR-4	User payment	Conform payment.

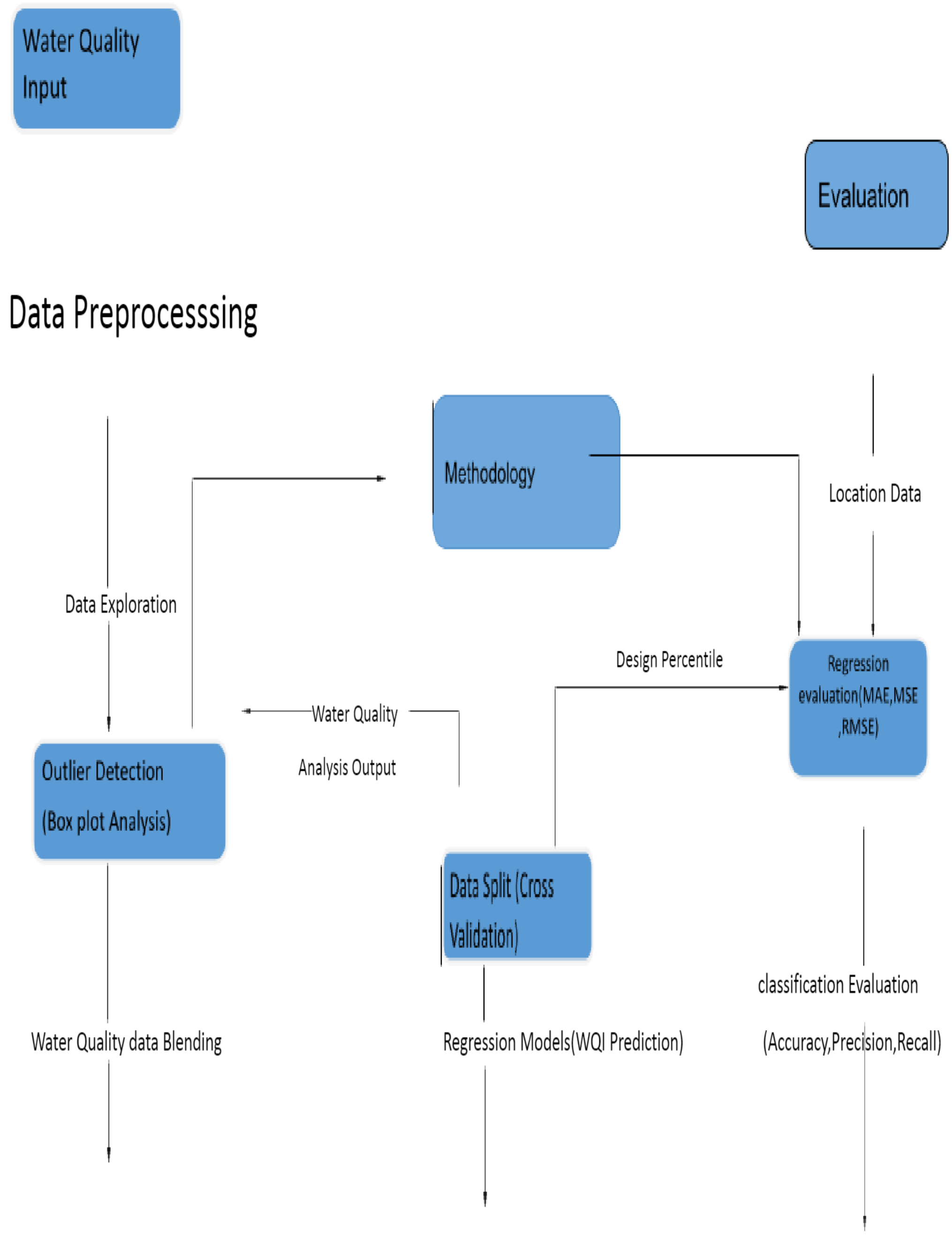
4.2 Non-Functional requirements

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

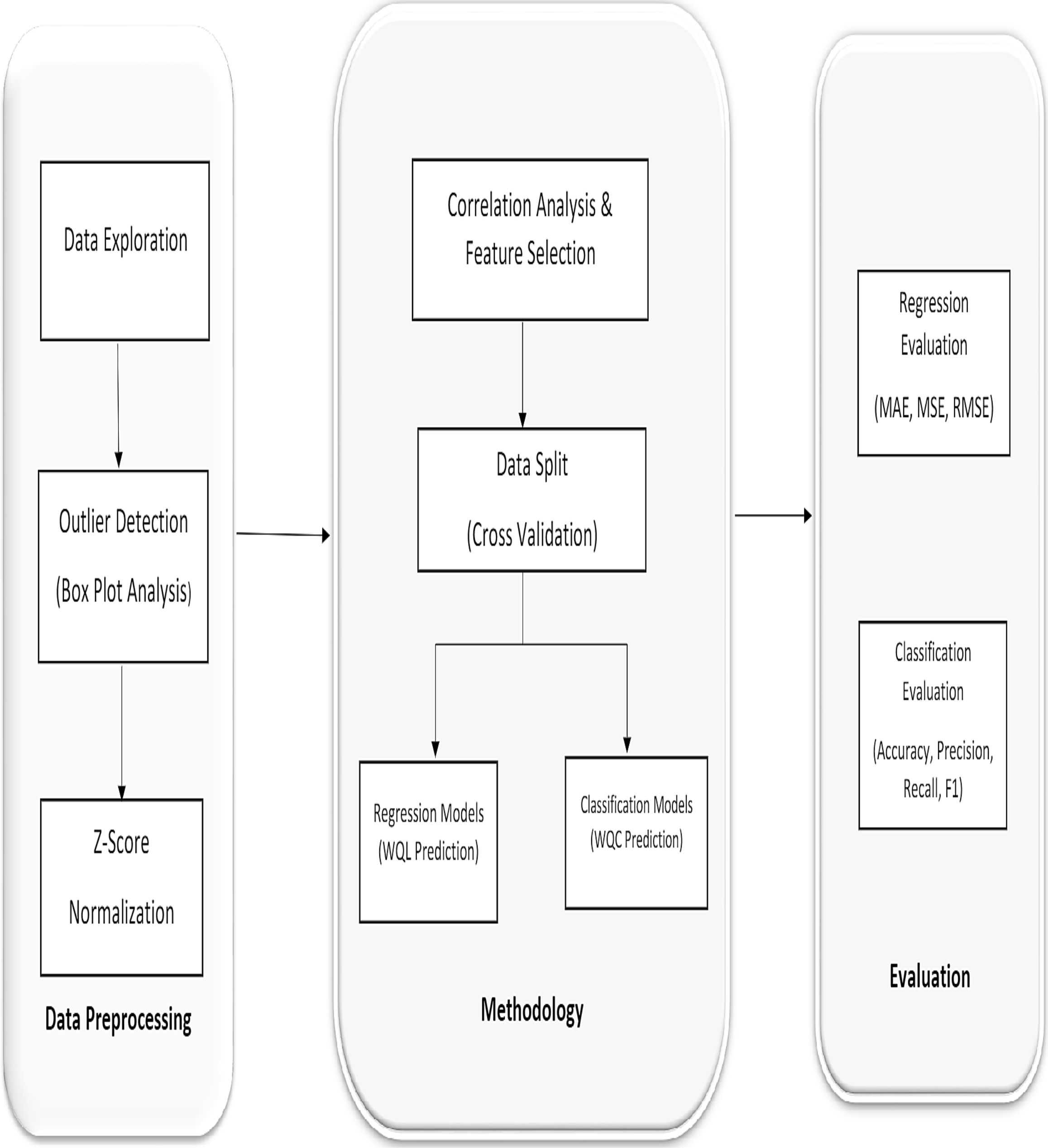
FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	System has been made user friendly by developing a web application, so it's easy to use.
NFR-2	**Security**	Indicate the whether water can be used to drink or not.
NFR-3	**Reliability**	System should give reliable predicted results.
NFR-4	**Performance**	Our LSTM model will have improved performance because of the use of datasets with lowest time intervals and has high precession. For checking the accuracy we have shown the performance metrics using RMSE.
NFR-5	**Availability**	Activated carbon filter source, Bio-sand filter source, Domestic reverse osmosis filter system source.
NFR-6	**Scalability**	If more parameters required, it can be added easily. Number of visualizations can be increased. Currently the system predicts for hourly manner this interval can be changed accordingly.

5. PROJECT DESIGN

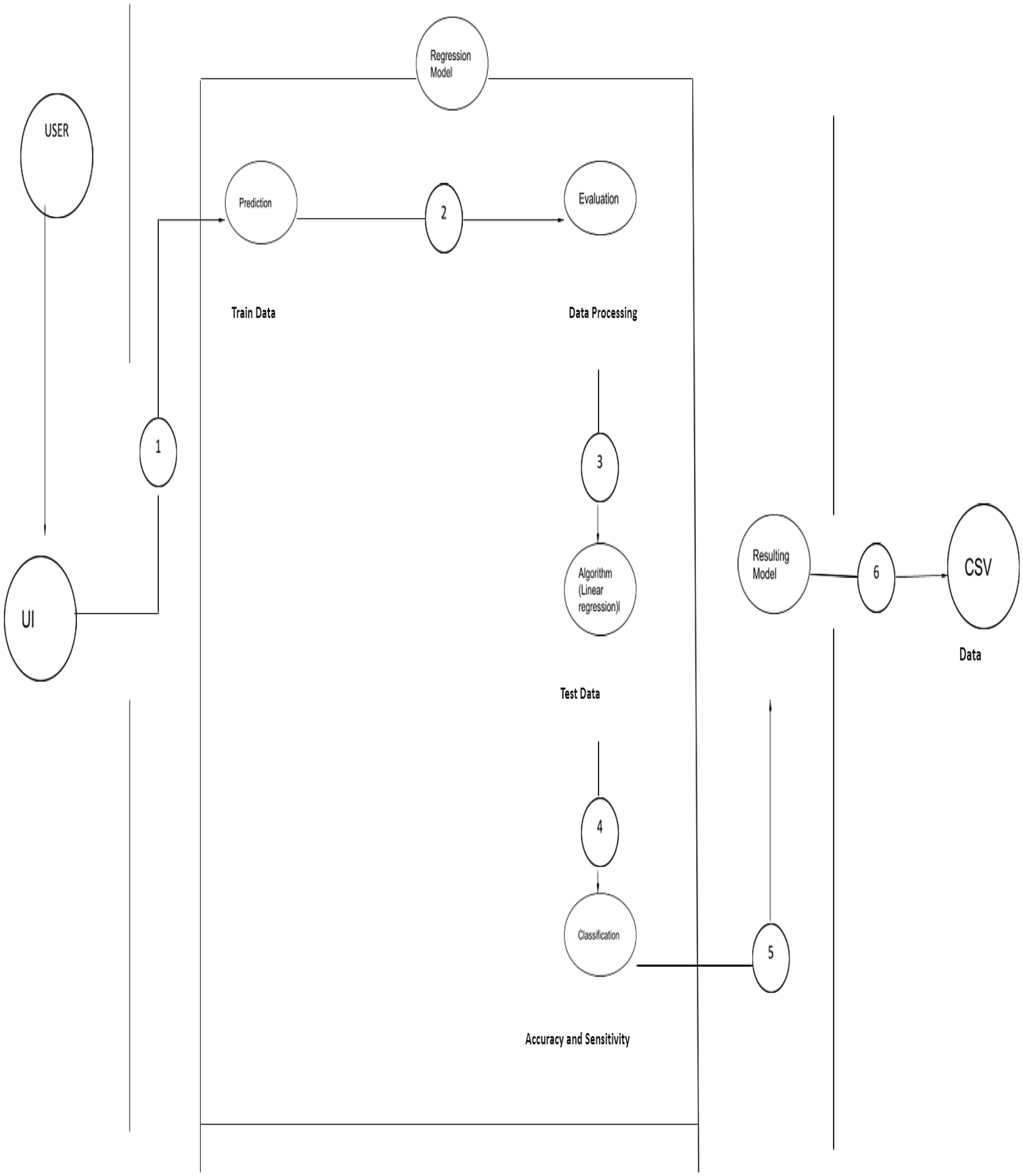
5.1 Data Flow Diagrams


Water Quality Input
Evaluation
Data Preprocesssing
Methodology
Location Data
Data Exploration
Design Percentile
Regression evaluation(MAE,MSE ,RMSE)
Water Quality
Analysis Output
Outlier Detection
(Box plot Analysis)
Data Split (Cross Validation)
classification Evaluation
Water Quality data Blending
Regression Models(WQI Prediction)
(Accuracy,Precision,Recall)

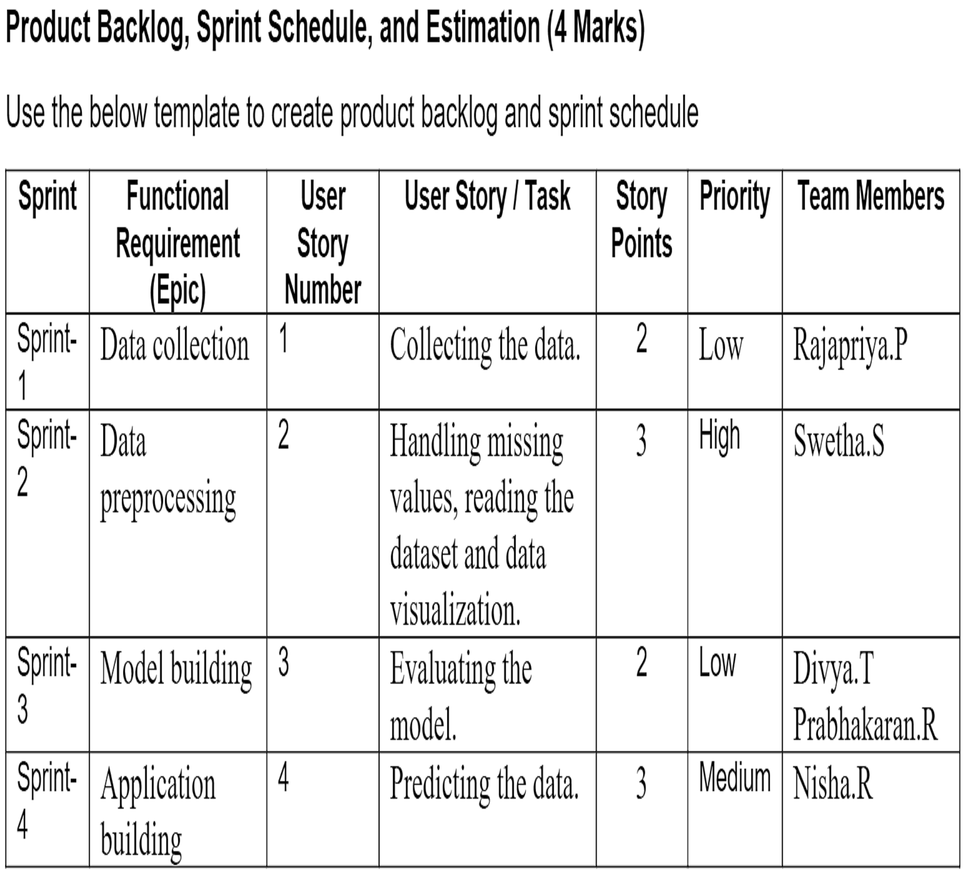
5.2 Solution & Technical Architecture



Technical Architecture



5.3 User Stories



Product Backlog, Sprint Schedule, and Estimation (4 Marks)

Use the below template to create product backlog and sprint schedule

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Data collection	1	Collecting the data.	2	Low	Rajapriya.P
Sprint-2	Data preprocessing	2	Handling missing values, reading the dataset and data visualization.	3	High	Swetha.S
Sprint-3	Model building	3	Evaluating the model.	2	Low	Divya.T Prabhakaran.R
Sprint-4	Application building	4	Predicting the data.	3	Medium	Nisha.R

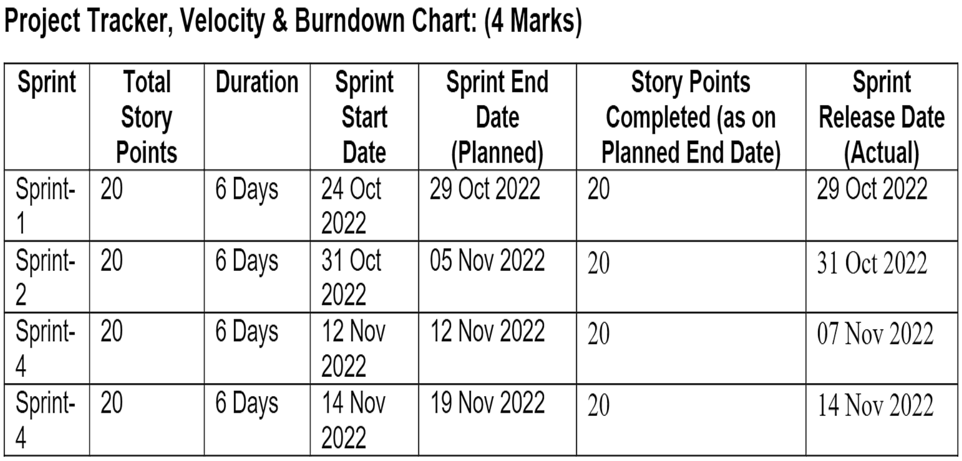
6. PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

Use the below template to create product backlog and sprint schedule

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Data collection	1	Collecting the data.	2	Low	Rajapriya.P
Sprint-2	Data preprocessing	2	Handling missing values, reading the dataset and data visualization.	3	High	Swetha.S
Sprint-3	Model building	3	Evaluating the model.	2	Low	Divya.T Prabhakaran.R
Sprint-4	Application building	4	Predicting the data.	3	Medium	Nisha.R

6.2 Sprint Delivery Schedule



Project Tracker, Velocity & Burndown Chart: (4 Marks)

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	20	6 Days	24 Oct 2022	29 Oct 2022	20	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	20	31 Oct 2022
Sprint-4	20	6 Days	12 Nov 2022	12 Nov 2022	20	07 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	20	14 Nov 2022

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

MODEL BUILDING

```
pip install findspark

pip install geopandas

pip install pyspark

import os
import pandas as pd
import numpy as np
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
import geopandas as gpd
import warnings
warnings.filterwarnings("ignore")

from pylab import *
from pyspark.sql.functions import udf, concat, col, lit
from pyspark import SparkConf, SparkContext
from pyspark.sql import SparkSession, SQLContext

from pyspark.sql.types import *
import pyspark.sql.functions as F
sc = SparkContext.getOrCreate(SparkConf().setMaster("local[*]"))
from pyspark.sql import SparkSession
spark = SparkSession \
    .builder \
    .getOrCreate()
sqlContext = SQLContext(sc)

from google.colab import files
uploaded = files.upload()

df = spark.read.format("csv").option("header", "true").load('water_data.csv')
```

```
df.show(5)

df.dtypes

from pyspark.sql.types import FloatType

df = df.withColumn("TEMP",df["TEMP"].cast(FloatType()))
df = df.withColumn("pH",df["pH"].cast(FloatType()))
df = df.withColumn("DO",df["DO"].cast(FloatType()))
df = df.withColumn("CONDUCTIVITY",df["CONDUCTIVITY"].cast(FloatType()))
df = df.withColumn("BOD",df["BOD"].cast(FloatType()))
df = df.withColumn("NITRATE_N_NITRITE_N",df["NITRATE_N_NITRITE_N"].cast(FloatType()))
df = df.withColumn("FECAL_COLIFORM",df["FECAL_COLIFORM"].cast(FloatType()))
df.dtypes

df=df.drop('TOTAL_COLIFORM')

df.createOrReplaceTempView("df_sql")

df_clean = spark.sql('''Select * from df_sql where TEMP is not null and DO is not null
                 and pH is not null and BOD is not null and CONDUCTIVITY is not null
                 and NITRATE_N_NITRITE_N is not null and FECAL_COLIFORM is not null''')

df_clean.createOrReplaceTempView("df_sql")

do = spark.sql("Select DO from df_sql")
do = do.rdd.map(lambda row : row.DO).collect()
ph = spark.sql("Select pH from df_sql")
ph = ph.rdd.map(lambda row : row.pH).collect()
bod = spark.sql("Select BOD from df_sql")
bod = bod.rdd.map(lambda row : row.BOD).collect()
nn = spark.sql("Select NITRATE_N_NITRITE_N from df_sql")
nn = nn.rdd.map(lambda row : row.NITRATE_N_NITRITE_N).collect()

fig,ax = plt.subplots(num=None,figsize=(14,6), dpi=80, facecolor='w', edgecolor='k')
size=len(do)
ax.plot(range(0,size), do, color='blue', animated=True, linewidth=1, label='Dissolved Oxygen')
ax.plot(range(0,size), ph, color='red', animated=True, linewidth=1, label='pH')
fig,ax2 = plt.subplots(num=None,figsize=(14,6), dpi=80, facecolor='w', edgecolor='k')
ax2.plot(range(0,size), bod, color='orange', animated=True, linewidth=1, label='BOD')
ax2.plot(range(0,size), nn, color='green', animated=True, linewidth=1, label='NN')
```

```
legend=ax.legend()
legend=ax2.legend()

con = spark.sql("Select CONDUCTIVITY from df_sql")
con = con.rdd.map(lambda row : row.CONDUCTIVITY).collect()
fec = spark.sql("Select FECAL_COLIFORM from df_sql")
fec = fec.rdd.map(lambda row : row.FECAL_COLIFORM).collect()

fig,ax = plt.subplots(num=None,figsize=(14,6), dpi=80, facecolor='w', edgecolor='k')
ax.plot(range(0,size), con, color='blue', animated=True, linewidth=1)
fig,ax2 = plt.subplots(num=None,figsize=(14,6), dpi=80, facecolor='w', edgecolor='k')
ax2.plot(range(0,size), fec, color='red', animated=True, linewidth=1)

df=df_clean.toPandas()
df.dtypes

start=0
end=448
station=df.iloc [start:end ,0]
location=df.iloc [start:end ,1]
state=df.iloc [start:end ,2]
do= df.iloc [start:end ,4].astype(np.float64)
value=0
ph = df.iloc[ start:end,5]
co = df.iloc [start:end ,6].astype(np.float64)
bod = df.iloc [start:end ,7].astype(np.float64)
na= df.iloc [start:end ,8].astype(np.float64)
fc=df.iloc [2:end ,9].astype(np.float64)

df=pd.concat([station,location,state,do,ph,co,bod,na,fc],axis=1)
df. columns = ['station','location','state','do','ph','co','bod','na','fc']

df['npH']=df.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)))))

df['ndo']=df.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)))))

df['nco']=df.fc.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)))))

df['nbdo']=df.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)))))

df['nec']=df.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)))))

df['nna']=df.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)))))

df.head()
df.dtypes

df['wph']=df.npH * 0.165
df['wdo']=df.ndo * 0.281
df['wbdo']=df.nbdo * 0.234
df['wec']=df.nec* 0.009
df['wna']=df.nna * 0.028
df['wco']=df.nco * 0.281
df['wqi']=df.wph+df.wdo+df.wbdo+df.wec+df.wna+df.wco
df

df['quality']=df.wqi.apply(lambda x:('Excellent' if (25>=x>=0)
                  else('Good' if  (50>=x>=26)
                            else('Poor' if (75>=x>=51)
                               else('Very Poor' if (100>=x>=76)
                                  else 'Unsuitable')))))

sns.lineplot(df["state"])

spark_df = sqlContext.createDataFrame(df)

spark_df.show()

spark_df.createOrReplaceTempView("df_sql")

State = spark.sql("Select state from df_sql")
State = State.rdd.map(lambda row : row.state).collect()

Wqi = spark.sql("Select wqi from df_sql")
Wqi = Wqi.rdd.map(lambda row : row.wqi).collect()

plt.barh(State,Wqi)

plt.xlabel("WQI")
plt.ylabel("STATES")


plt.show()

spark_df = sqlContext.createDataFrame(df)

from pyspark.ml.feature import StringIndexer

from pyspark.ml.feature import VectorAssembler
from pyspark.ml.feature import Normalizer

vectorAssembler = VectorAssembler(inputCols=["npH","ndo","nbdo","nec","nna","nco"],
outputCol="features")
normalizer = Normalizer(inputCol="features",outputCol="features_norm")

indexer = StringIndexer(inputCol="quality",outputCol="label")
vectorAssembler2 = VectorAssembler(inputCols=["npH","ndo","nbdo","nec","nna","nco","wqi"],
outputCol="features2")
normalizer2 = Normalizer(inputCol="features2",outputCol="features_norm2")
```

```
from pyspark.ml.classification import LogisticRegression

lor = LogisticRegression(featuresCol="features_norm2",labelCol="label",maxIter=10)

from pyspark.ml import Pipeline

pipeline2 = Pipeline(stages=[indexer,vectorAssembler2,normalizer2,lor])

train_data,test_data=spark_df.randomSplit([0.8,0.2])

model3 = pipeline2.fit(train_data)

predictions2 = model3.transform(train_data)

predictions2 = model3.transform(train_data)

from pyspark.ml.evaluation import MulticlassClassificationEvaluator
eval =
MulticlassClassificationEvaluator().setMetricName('accuracy').setLabelCol('label').setPredictionCol('prediction')
eval.evaluate(predictions2)

names = ["Very Poor","Poor","Good","Unsuitable","Excellent"]

predictions2.createOrReplaceTempView("predictions2_sql")

pred = spark.sql("Select prediction from predictions2_sql")
pred = pred.rdd.map(lambda row : int(row.prediction)).collect()
qua = spark.sql("Select quality from predictions2_sql")
qua = qua.rdd.map(lambda row : row.quality).collect()

for x in range(100):
   print("Predicted:", names[pred[x]], "Actual:", qua[x])
```

8. TESTING

The EPA sets standards and regulations for the presence and levels of over 90 contaminants in public drinking water, including *E.coli*, *Salmonella*, *Cryptosporidium*, metals such as lead, and disinfection byproducts. Learn more about these germs in the Diseases and Contaminants page.

9. RESULTS

9.1 Performance Metrics

In this part of the paper, the results of prediction of the internal relations between the water quality components are presented. To develop an optimal model, an approach that was introduced by Parsaie & Haghiabi ([illegible]c) was considered. They stated that for developing the ANN, some steps should be considered to reduce the trial and error process. They stated that for the initial design of ANN model, after dataset division, in the first step one hidden layer consisting of numbers of neurons equal to input features is considered. At this stage, the performance of different transfer functions is evaluated and the best ones are chosen. In the next step, the size of the network is modified to improve the precision of the developed model. To this end, the numbers of neurons or number of hidden layers would increase. The last two stages of this approach are also applicable to the design of SVM. The main point relating to the design of the SVM is defining the kernel function. In this study, 80% of the dataset was used for training and the remaining 20% for testing. These two groups of datasets were used for developing the ANN, SVM and GMDH. For developing the GMDH model as stated in the Material and methods section, the designer only controls the threshold values and develops the network structure. The RMSE index was chosen for threshold values. The RMSE values of GMDH given in Table 2 in the testing stage were chosen as the threshold criteria. A summary of results of each applied model for predicting the water quality components are given in Table 2. For example, as presented in this table, for predicting the Ca, other water quality components including Cl, EC, HCO3, Mg, Na, SO4, TDS, pH were considered as inputs. Table 2 shows that for predicting the Ca, the SVM is most accurate in comparison with others (GMDH and ANN). As presented in Table 2, the best performance of ANN with coefficient of determination (0.92 and 0.84) and root means square error (0.238 and 0.295) in training and testing stages is related to the tansig function as best transfer function. It is notable that the structure of ANN, after the trial and error process, consisted of two hidden layers where its first and second hidden layers included eight and three neurons, respectively. A comparison of three models in terms of predicting the Cl declared that the SVM model has the best performance. For predicting the Ec, all three models have suitable performance. For predicting the HCO3, the performances of SVM and ANN are close together and their precision is greater than the GMDH. This result was repeated for Mg. For predicting the Na, the SVM has the best performance. For estimation of SO4, the best performance was related SVM. For estimation of TDS, all three models have suitable performance and their accuracy is close together. For estimation of pH, the best performance is related to SVM. The accuracy of this model in the prediction of pH indicates that its accuracy has been reduced by a small amount. The structure of SVM and ANN models for predicting the Ca are shown in Figures 4 and 5, respectively. Reviewing Table 2 indicates that the RBF and tansing functions have the best performance in comparison with other tested kernel and transfer functions. The results of three applied AI models for estimation of Ca and Cl in training and testing stages are shown in Figure 6. To present further information related to the performance of applied models throughout the dataset, the DDR index, introduced by Noori et al. (▯▯▯▯), was calculated. This index is calculated using Equation (8). This index shows the performances of applied models related to properties of lower and over estimation. Results of DDR for testing stages of applied models for all water quality components are shown in Figure 7. As shown in this figure, the most amount of data dispersion related DDR index is related to the ANN model and the lowest data dispersion is related to the SVM. Although the accuracy of GMDH was less than SVM and somewhere less than the ANN, the data dispersion related DDR values are close to SVM. Reviewing Figure 7 shows that all three models have a slight over-estimation property: DDR ¼ Predicted Value Observed Value ▯ ▯ 1

10. ADVANTAGES & DISADVANTAGES

ADVANTAGES

Leads to Better Health:

Water quality monitoring system will help us to know the most healthy water in the plant, and it can lead to better health too. Quality water helps prevent waterborne illness. Although the Environmental Protection Agency regulates water treatment facilities for chemicals, microorganisms, and other contaminants, it doesn’t test for a variety of contaminants.

Leads to Better Water Treatment:

In general, water treatment is one of the most important and sometimes the only thing that needs be done in a business because if there is any problem with water, it will affect the productivity of a company. By noticing those problem through water quality monitoring system, then you can change your procedures and answer those problem easily.

High Efficiency:

With water quality monitoring systems you can boost your efficiency of your plant and this will lead to less energy consumption. This is because you can withhold production if there is any problem with water quality testing center. With such systems there are no more surprises when you go for a check up from plant to plant or from production unit to the store room.

Cost Effective:

Quality water monitoring system can save a lot of money for you. This is because if you have a big company and it takes more than 5,000 liters of water in your plant per day, then your monthly bill will be more than 1 crore if you are using untreated water. With water quality monitoring system, it will cost less and you can save a lot from your company.

Safety:

Quality water monitoring system can help you to ensure that your water is safe for human consumption. This is because water quality testing methods can show when chemicals have accumulated in your water as it is being produced and not added/used. This can help you to avoid any chemical or bacterial contamination of your drinking water.

Performance:

Quality water monitoring system can help you to check your performance with water quality testing methods. This is because you can use those testing results and compare them with the performance of your plant and adjust that points where you are lacking. This will boost the performance of your entire company, and it will help you to get more profits in comparison to losses.

DISADVANTAGES

Labor Intensive For Installation And Operation:

Quality water monitoring system requires a lot of man hours for its installation and operation. This is because water quality monitoring system consists of multiple instruments and they all are very time consuming.

High Initial Costs:

Quality water monitoring system has a high initial cost and it is very difficult for businesses to afford the total expense of quality water monitoring system. Usually a good quality water monitoring system will be for thousands of dollars or lakhs.

Maintenance Costs:

Because of high labor intensive and high initial costs, quality water monitoring system has a high maintenance cost which is why after some time you will have to replace your entire system. This is because your system will not be able to provide the same level of performing water quality testing methods after some time.

Time Consuming:

Quality water monitoring system is very time consuming and the whole procedure is not reliable. This is because with water quality monitoring system you will have to change your test report every month, but some companies will ask for test reports every week or even daily.

11. CONCLUSION

The performance of machine learning techniques such as RF, NN, MLR, SVM, and BTM to predict the water quality components of an Indian water quality dataset was evaluated in this work. The most well-known dataset variables, such as BOD, DO, TC, Nitrate, pH, and Temp, were obtained for this purpose. The findings revealed that the applied models performed well in forecasting water quality parameters; however, the greatest performance was linked with the MLR with Accuracy Upper. Further research will be done to build models that combine the proposed method with other techniques and deep learning approaches to improve the efficacy of the selection process.AI solutions such as machine learning ease the task of WQI prediction. The AI-based WQI prediction system supports efforts to provide timely and efficient water pollution prevention and response systems by forecasting the change in the WQI based on historical data. In this paper, eight standalone machine learning regression algorithms (DT, LR, Ridge, Lasso, SVR, RF, ET and ANN) were compared for their predictions of the WQI using three sets of water parameter features. An open dataset based on data from Indian rivers collected between 2003 to 2014 was used. The WQI was measured using six water quality features, including the pH, DO, CO, BOD, NA, and FC. Two sets of derivative features were derived, namely the water quality rating scale and water quality weight score. The original water quality features and the two sets of derivative features were then used in the WQI prediction. The results show that LR and Ridge trained using the water quality rating scale are able to predict the WQI accurately, with MSE=0 and r=1. The results outperformed the performances of existing models. Overall, it was observed that the regression algorithm and set of features used are the main factors affecting the performance of an WQI prediction model. Future research directions and challenges were also addressed in this work.

12. GitHub & Project Demo Link

GitHub Link

https://github.com/IBM-EPBL/IBM-Project-49421-1660818717

Project Demo Link

https://www.mediafire.com/file/9dudjy0qs8ose4r/PNT2022TMID49117.mp4/file