EFFICIENT WATER QUALITY ANALYSIS &PREDICTION USING MACHINE LEARNING

TABLE OF CONTENT:

SERIAL NUMBER	TITLE	PAGE NUMBER
1	Abstract	1
2	Introduction	2
3	Project Description	2
4	Literature Review	3
5	working model	3
6	Context	4
7	Procedure	4
8	Conclusion	5

ABSTRACT:

Water makes up about 70% of the earth's surface and is one of the most important sources vital to sustaining life. Rapid urbanization and industrialization have led to a deterioration of water quality at an alarming rate, resulting in harrowing diseases. Water quality has been conventionally estimated through expensive and time-consuming lab and statistical analyses, which render the contemporary notion of real-time monitoring moot. The alarming consequences of poor water quality necessitate an alternative method, which is quicker and inexpensive. With this motivation, this research explores a series of supervised machine learning algorithms to estimate the water quality index (WQI), which is a singular index to describe the general quality of water, and the water quality class (WQC), which is a distinctive class defined on the basis of the WQI. The proposed methodology employs four input

parameters, namely, temperature, turbidity, pH and total dissolved solids. Of all the employed algorithms, gradient boosting, with a learning rate of 0.1 and polynomial regression, with a degree of 2, predict the WQI most efficiently, having a mean absolute error (MAE) of 1.9642 and 2.7273, respectively. Whereas multi-layer perceptron (MLP), with a configuration of (3, 7), classifies the WQC most efficiently, with an accuracy of 0.8507. The proposed methodology achieves reasonable accuracy using a minimal number of parameters to validate the possibility of its use in real time water quality detection system

INTRODUCTION :

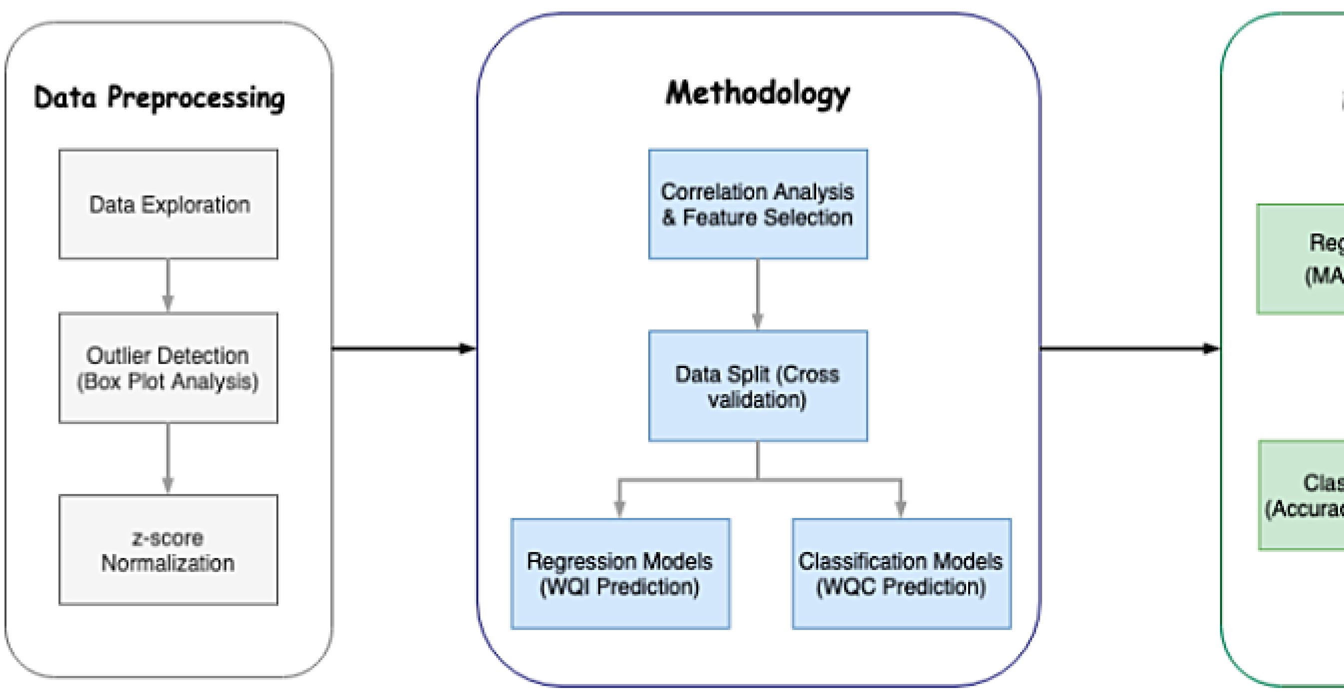
Data science helps in gathering and getting useful information to make sound decisions

Machine learning helps in applying algorithms to get the predicted result

Python language helps in applying the tools and techniques

Statistics helps in analyzing and understanding data and accounting for the relevant uncertains

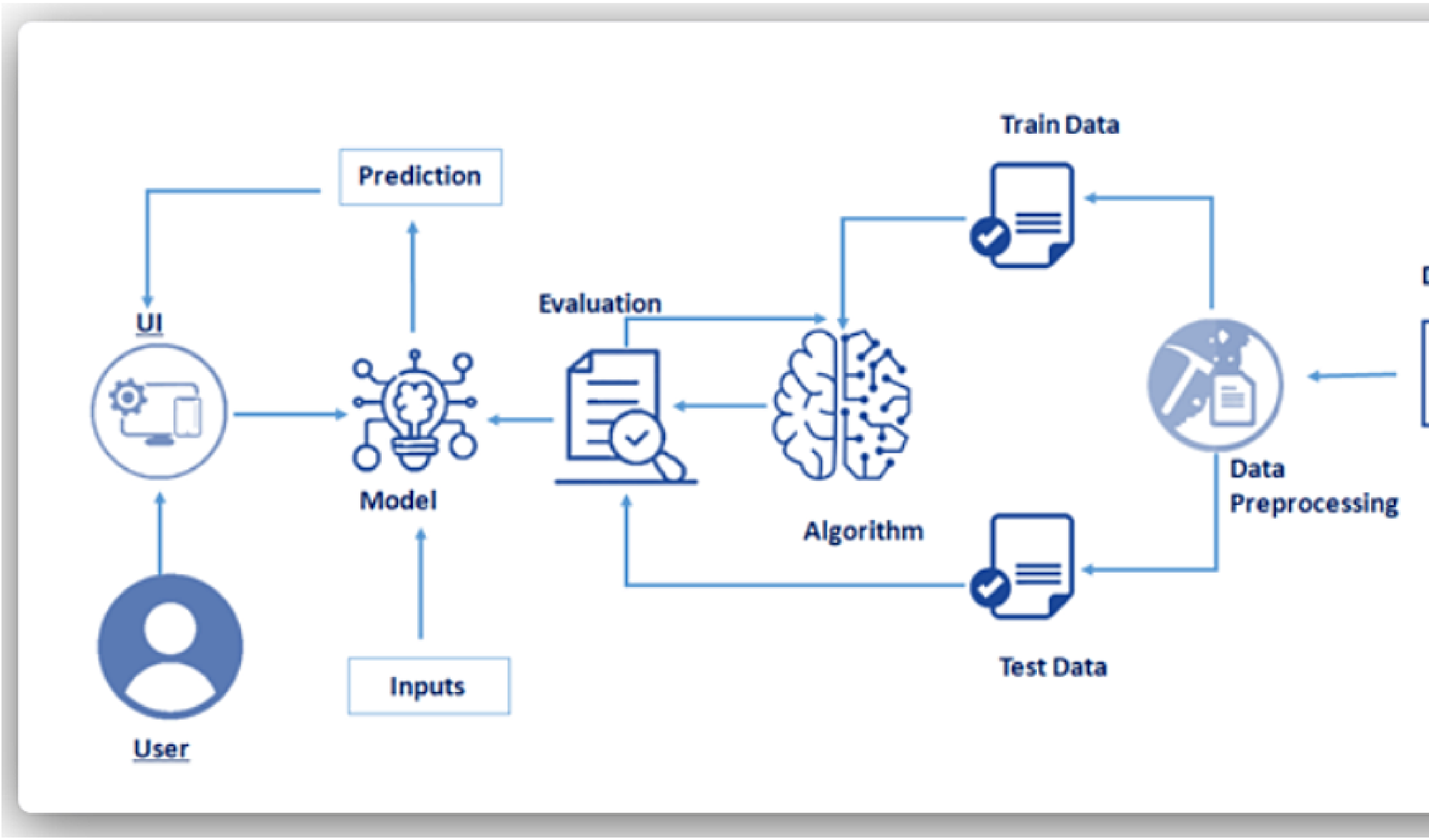
Aim of this project to predict the potability using classification machine learning model



LITERATURE SURVEY :

A literature survey on water quality of indian water bodies

TECHNICAL ARCHITECTURE :



WORKING MODEL :

SCOPE OF WORK :

The final goal is to built the rigid machine learning model that can accurately predict potability of water on some inputs

TOOLS FOR IMPLEMENTATION :

Anaconda

Jupyter Notebook
Python/R
Google collaboratory

LIBRARIES USED :

Numpy - to perform wide variety of mathematical operation on arrays
Pandas - provides various datastructures and ooperation for manipulating numerical data
sklean - tool for predictive data analysis , features various classifications , regressions and clustering algorithms

REQUIREMENTS :

Dataset that can contain relevant information
proper data visualization
selection of correct classification model
calculating the accuracy of the result

PROJECT DESCRIPTION :

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.

SKILLS REQUIRED :
Python
Python Web Frame Works
Python For Data Visualization
Data Preprocessing Techniques

Machine Learning
IBM Cloud
IBM Watson Studio
Python-Flas

MACHINE LEARNING :

Definition:

The data used for this research was obtained from Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its Model

DATA PROCESSING :

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

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

CONTEXT :

Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions.

1. pH value:

6.52–6.83 which are in the range of WHO standards. PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended a maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards

. 2. Hardness:

Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium. EFFICIENT WATER QUALITY ANALYSIS AND PREDICTION USING MACHINE LEARNING 16

3. Solids (Total dissolved solids - TDS):

Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced an unwanted taste and diluted color in the appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. The Desired limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which is prescribed for drinking purposes.

4. Chloramines:

Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

5. Sulphate:

Sulphates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulphate is in the chemical industry. Sulphate concentration in seawater is about 2,700 milligrams per litre (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.

6. Conductivity:

Pure water is not a good conductor of electric current rather it's a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceed 400 μS/cm.

7. Organic carbon :

EFFICIENT WATER QUALITY ANALYSIS AND PREDICTION USING MACHINE LEARNING 17 Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to the US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is used for treatment.

8. Trio halomethanes:

THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.

9. Turbidity:

The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.

10.Potability: Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

Water is not safe to drink
Water is safe to drink

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from matplotlib import pyplot as pt
import warnings
```

```
data = pd.read_csv(r'Documents\dataset.csv',encoding='ISO-8859-1',
data.head()
```

	STATION CODE	LOCATION	STATE	Temp	D.O. (mg/l)	PH	CONDUCTIVITY(µmhos/cm)	B.O.D. (mg/l)	NITRATEN N+ NITRITENANN(mg/l)	FECALCOLIFORM(MPN/100M
0	1393	DAMANGANGA AT D/S OF MADHUBAN, DAMAN	DAMAN & DIU	30.6	6.7	7.5	203	NAN	0.1	
1	1399	ZUARI AT D/S OF PT. WHERE KUMBARJRIA CANAL JOI...	GOA	29.8	5.7	7.2	189	2	0.2	49
2	1475	ZUARI AT PANCHAWADI	GOA	29.5	6.3	6.9	179	1.7	0.1	32
3	3181	RIVER ZUARI AT BORIM BRIDGE	GOA	29.7	5.8	6.9	64	3.8	0.5	53
4	3182	RIVER ZUARI AT MARCAIM JETTY	GOA	29.5	5.8	7.3	83	1.9	0.4	34

```
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>=
                                  else (60 if(8.8>=x>=8.6) or (6.8>=
                                      else(40 if(9>=x>=8.8) or (6.7>
                                          else 0)))))
```

```
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)))))
```

```
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)))))
```

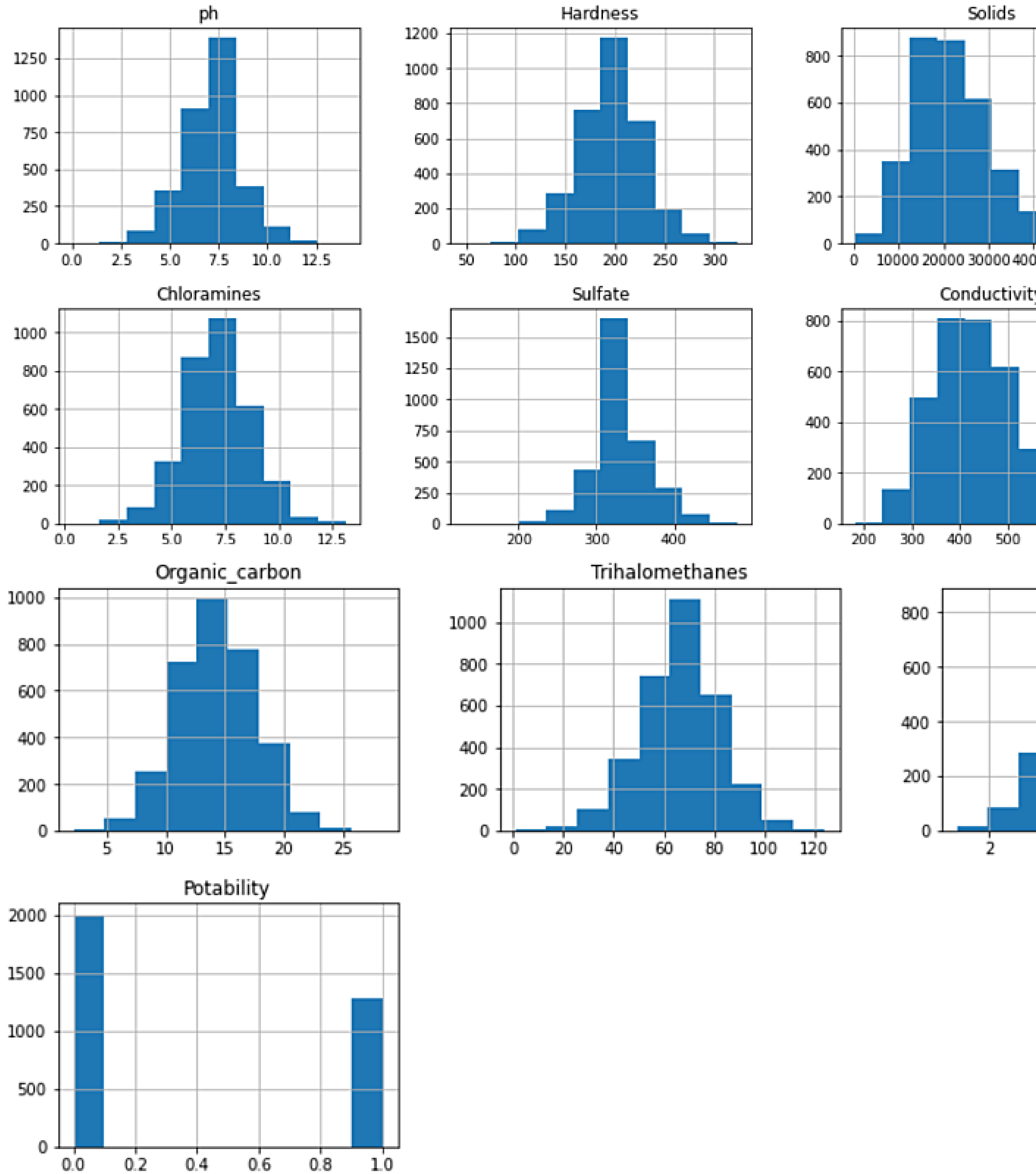
```
data['nco']=data.tc.apply(lambda x: (100 if(5>=x>=0)
                                else(80 if(50>=x>=5)
                                  else (60 if(500>=x>=50)
                                      else(40 if(10000>=x>=500)
                                          else 0)))))
```

```
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)))))
```

```
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.wc
data
```

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes
count	2785.000000	3276.000000	3276.000000	3276.000000	2495.000000	3276.000000	3276.000000	3114.000000
mean	7.080795	196.369496	22014.092526	7.122277	333.775777	426.205111	14.284970	66.396293
std	1.594320	32.879761	8768.570828	1.583085	41.416840	80.824064	3.308162	16.175008
min	0.000000	47.432000	320.942611	0.352000	129.000000	181.483754	2.200000	0.738000
25%	6.093092	176.850538	15666.690300	6.127421	307.699498	365.734414	12.065801	55.844536
50%	7.036752	196.967627	20927.833605	7.130299	333.073546	421.884968	14.218338	66.622485
75%	8.062066	216.667456	27332.762125	8.114887	359.950170	481.792305	16.557652	77.337473
max	14.000000	323.124000	61227.196010	13.127000	481.030642	753.342620	28.300000	124.000000



CONCLUSIONS :

The machine learning algorithms have been used to design the water quality analysis model which predicts whether the water is safe for drinking or not

after testing different algorithms , the support vector machine learningwith radial kernel is found to be be the best performing model among all others in terms of accuracy with the percentage accuracy of 67.5 % followed by k-nearest neighbour with the accuracy of 63.1% followed by the logistics regressions model with accuracy of 62.8%

The quality of water is predicted using the Machine Learning. Water is one of the most essential resources for survival and its quality is determined through Water Quality Index(WQI).

Conventionally, to test water quality, one has to go through expensive and cumbersome lab analysis.

This research explored an alternative method of machine learning to predict water quality using minimal and easily available water quality parameter