Table of content:

ABSTRACT:

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

INTRODUCTION:

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

Water quality is currently estimated through expensive and time-consuming lab and statistical analyses, which require sample collection, transport to labs, and a considerable amount of time 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 [4]. The horrific consequences of water pollution necessitate a quicker and cheaper alternative. In this regard, the main motivation in this study is to propose and evaluate an alternative method based on supervised machine learning for the efficient prediction of water quality in real-time. This research is conducted on the dataset of Rawal water shed, situated in Pakistan, acquired by The Pakistan Council of Research in Water Resources (PCRWR) (Available online at URL http://www.pcrwr.gov.pk/). A representative set of supervised machine learning algorithms were employed on the said dataset for predicting the water quality index (WQI) and water quality class (WQC). The main contributions of this study are summarized as follows: • A first analysis was conducted on the available data to clean, normalize and perform feature selection on the water quality measures, and therefore, to obtain the minimum relevant subset that allows high precision with low cost. In this way, expensive and cumbersome lab analysis with specific sensors can be avoided in further similar analyses. • A series of representative supervised prediction (classification and regression) algorithms were tested on the dataset worked here. The complete methodology is proposed in the context of water quality numerical analysis. • After much experimentation, the results reflect that gradient boosting and polynomial regression predict the WQI best with a mean absolute error (MAE) of 1.9642 and 2.7273, respectively, whereas multi-layer perceptron (MLP) classifies the WQC best, with an accuracy of 0.8507. The remainder of this paper is organized as follows: Section 2 provides a literature review in this domain. In Section 3, we explore the dataset and perform preprocessing. In Section 4, we employ various machine learning methodologies to predict water quality using minimal parameters and discuss the results of regression and classification algorithms, in terms of error rates and classification precision. In Section 5, we discuss the implications and novelty of our study and finally in Section 6, we conclude the paper and provide future lines of work. 2. Literature Review This research explores the methodologies that have been employed to help solve problems related to

water quality. Typically, conventional lab analysis and statistical analysis are used in research to aid in determining water quality, while some analyses employ machine learning methodologies to assist in finding an optimized solution for the water quality problem. Local research employing lab analysis helped us gain a greater insight into the water quality problem in Pakistan. In one such research study, Daud et al. [5] gathered water samples from different areas of Pakistan and tested them against different parameters using a manual lab analysis and found a high presence of E. coli and fecal coliform due to industrial and sewerage waste. Alamgir et al. [6] tested 46 different samples from Orangi town, Karachi, using manual lab analysis and found them to be high in sulphates and total fecal coliform count. After getting familiar with the water quality research concerning Pakistan, we explored research employing machine learning methodologies in the realm of water quality. When it comes to estimating water quality using machine learning, Shafi et al. [7] estimated water quality using classical machine learning algorithms namely, Support Vector Machines (SVM), Neural Networks (NN), Deep Neural Networks (Deep NN) and k Nearest Neighbours (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.

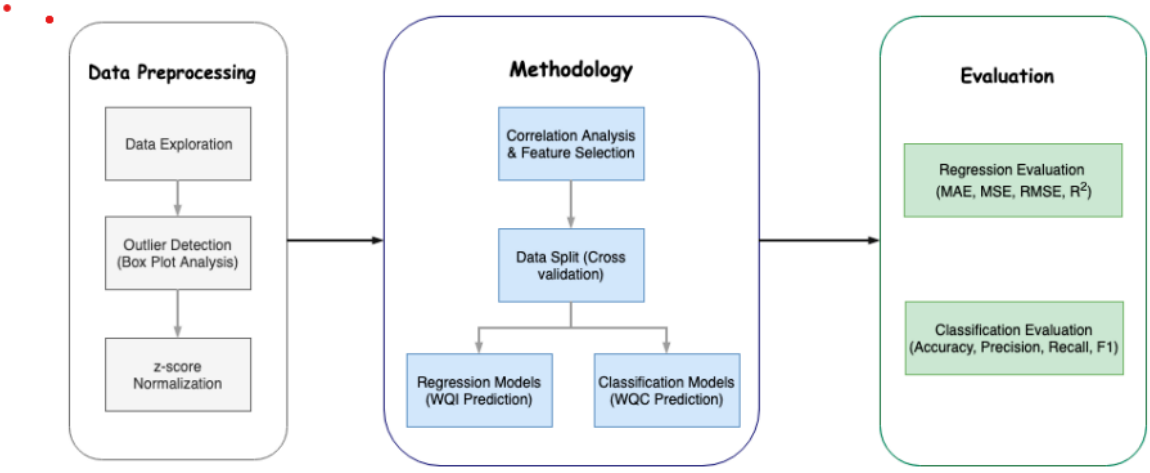


Fig. Methodology flow

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.

Category: Machine Learning

Skills Required
Python, Python Web Frame Works, Python For Data Visualization, Data Preprocessing Techniques, Machine Learning, IBM Cloud, IBM Watson Studio, Python-Flask

Importance:

● Water is one of the essential component for human living. Water quality has a direct impact on public health and the environment.

● Among various sources of water supply, 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.

● It is important to develop a solution to analyse the quality of water for benefits of humankind.

Social Impact:

● Can drive to the vision of healthy nation.

Business Model/ Impact:

● Can sell our service/product to water purifier companies.

● Can collaborate with governments in analysing and providing the water quality solutions.

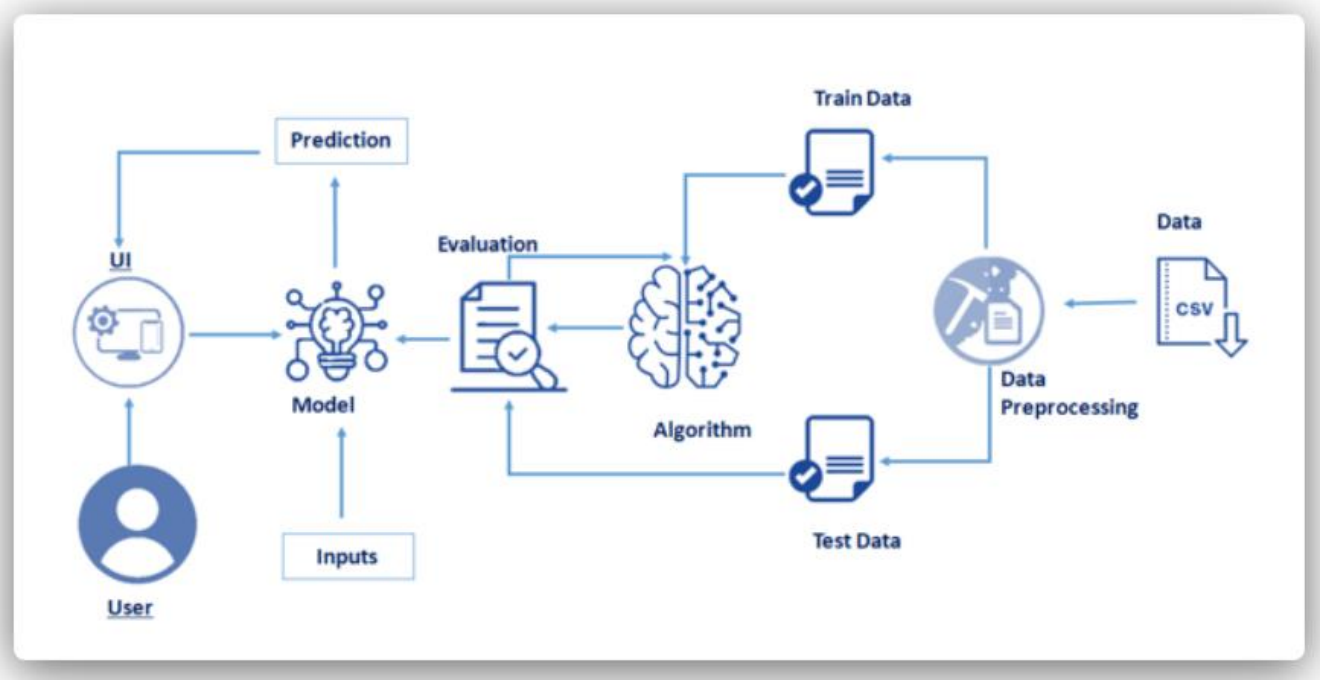
Existing Solutions:

● https://aperainst.com/zentestr-pc60-z-smart-multi-parameter-pocket-tester-kit

Recommended Technology Stack:

● Data Science, Data Analytics, Artificial intelligence, Machine Learning, IBM watson etc

Technical Architecture:



Literature Review:

1. Assessment of ground water quality and its Impact in around Mangalam near Tirupati ; G.DILLI RANI, M.SUMAN, C.NARASIMAHA RAO, P.REDDIRANI, R.PRATHIBA V.G.PRASHANTH, P. VENKATESWARLU Ground water quality and its impact on human health in and around

Mangalam, near Tirupathi, India was assessed. Water samples were collected from 8 different areas in and around Mangalam and analyzed for physicochemical parameters such as pH, electrical conductivity, total dissolved solids, total hardness, calcium, chlorides, sulphates, nitrates and dissolved oxygen. The found values of physicochemical parameters were compared with the World Health Organisation water quality standards. Based on the analysis, it was found that ground water of some of the areas was polluted and not suitable for drinking purpose. Thus the ground water of the area needs purification before drinking. 2. Analysis of Drinking Water Quality and its Impact on human health in Chandragiri, near Tirupati, India. S.V.DORAIRAJU, C. NARASIMHARAO, M. BUJAGENDRA RAJU,AND P.V.CHALAPATHI. Drinking water samples were collected from different locations of Chandragiri, near Tirupati, Andhra Pradesh, India and analyzed to assess physicochemical parameters and suitability of water for drinking purpose. Physicochemical parameters such as pH, hardness, alkalinity, calcium, magnesium, iron, nitrates, chlorides, sulphates, electrical conductivity, total solids (TS), total dissolved solids (TDS), total suspended solids (TSS), dissolved oxygen (DO), chemical oxygen demand (COD) and bio chemical oxygen demand (BOD) were determined. The found values were compared with the World Health Organisation water quality standards. Interpretation of data shows that drinking water of some of the areas was polluted and not suitable for drinking purpose. Thus the ground water of these areas needs purification before drinking. 3. Hydro chemical characterization of ground water in around tirupati areaE. Balaji, A. Nagaraju, Y. Sreedhar, A. Thejaswini, Zahed Sharifi In the management of water resources, quality of water is just as important as its quantity. The main aim of this study has been to assess the variability of groundwater parameters to develop water quality of Tirupati area and its suitability for domestic and irrigation purpose. Further, the samples were analyzed for pH, EC, TDS, carbonates, bicarbonates, alkalinity, chlorides, sulfates, hardness, fluoride, calcium, magnesium, sodium, and potassium. Based on the analytical results, chemical indices like percent sodium, sodium absorption ratio (SAR), adjusted SAR, percent sodium (Na %), residual sodium carbonate (RSC) and permeability index (PI) have been

calculated. Chadha rectangular diagram for geochemical classification and hydro chemical processes of groundwater indicated that most of waters are Ca–Mg–HCO3 and Ca– Mg–Cl types. Assessment of water samples from various methods indicated that majority of the water samples are suitable for domestic and irrigation purpose. 4. Statistical and Analytical Evaluation of ground water quality of Tirupati Area, A.Naraju, Z. Sharifi, E. Balaji. The multivariate statistical analysis, hydro geochemical modelling using visual MINTEQ software, indices of base exchange and Gibbs ratio were simultaneously applied to groundwater hydro chemical data of the Tirupati area. These techniques were applied to know the principal processes controlling the water chemistry. Fiftygroundwater samples were analysed for pH, electrical conductivity (EC), Ca, Mg, Na, K, HCO3, CO3, Cl, and SO4. The results showed that the abundance of the major ions in the water samples is in following order: Na > Ca > Mg > K and HCO3 > Cl > SO4 > CO3 > F. 5. The Physico-Chemical And Bacteriological Analysis Of Ground Water In Around Tirupati- R. Usha, A. Vasavi, Spoorthi And P.M.Swamy. In the present study, an attempt has been made to investigate the quality of ground water in and Around Tirupati, Chittoor District, Andhra Pradesh. The various parameters monitored include pH, Temperature, Total Suspended Solids, Total Dissolved Solids, Total Solids, Dissolved Oxygen, Biochemical Oxygen Demand, Alkalinity, Chlorides, Hardness and Colony Count. The results showed that all water samples have neutral pH,. 6. Groundwater Quality Assessment using Correlation and Regression Model in Tirupati, Ambiga Kannapiran . Groundwater is the most important natural source required for drinking to the public's around the world, particularly in rural areas. An attempt has been made in order to determine the spatial distribution of groundwater quality parameters and to study the correlation and regression method. The physical and chemical analysis results were compared to the standard guideline values as recommended by the Bureau of Indian standards for drinking and public health in order to have an indication of the present groundwater quality. 7. Analysis of Physico-Chemical Characteristics of Industrial Effluents in Tirupati- Putaka Ramesh, K. Abraham, B. suresh, T. Damodharam . Physico-chemical characteristics of industrial effluents

were collected from three industrial sites in and around Tirupati. Industrial effluents were studied in two years month by month from Jan 2014 - Dec 2015. The present research work deals with the study of some of the important physicochemical parameters of industrial waste water collected from Tirupati industrial region. Results indicated that pH values of effluent samples. 8. Determination of heavy metals in surface water and ground water in and around Tirupati-V. Hanuman Reddy, P.M.N. Prasad, A.V. Ramana Reddy and Y.V. Rami Reddy Water Quality is one of the most important concerns. The heavy metals levels up to ppb levels in drinking water quality may cause seviour health problems and also cause cancer. In this study we made an attempt to know the concentration of eight heavy metals in ground water and surface water in different locations of Tirupati, Chittoor District, Andhra Pradesh up to ppb levels. 9. Statistical Analysis of the Hydro geochemical Evolution of Ground water in Rangampeta area of Tirupati – A. Nagaraju, K. Sunil Kumar, A. Thejaswini, Z. Sharisi Multivariate statistical techniques involving factor analysis (FA) and R-mode hierarchical cluster analysis (HCA) were performed on 30 groundwater samples from Rangampeta, Chittoor District, Andhra Pradesh, South India to extract principal processes controlling the water chemistry. The groundwater samples were analyzed for distribution of chemical elements Ca, Mg, Na, K, Si, HCO3, CO3, Cl, and SO4. It also includes pH, and electrical conductivity (EC). 10. Testing of Ground Water Quality For drinking purpose in Tirupati- Dr. R. Bhavani and smt. S. Sharada Testing of groundwater quality is very important before using it for drinking purpose. In the present study ten samples at various locations of Tirupati were collected and the values of various chemical parameters like total dissolved solids (TDS), pH, chlorides, hardness, sulphates and fluorides were determined. These values were compared with Drinking water standards of IS: 10500-1991 to assess the suitability for drinking purpose. 11. Water Quality monitoring on Tirumala and Tirupati- K. Raju, T. Damodharam. An attempt has been made to evaluate the water quality of supplemented and ground water in Tirumala and Tirupati, Chittoor District, Andhra Pradesh, India. The Tirumala and Tirupati are the most popular pilgrimage and education areas in Andhra Pradesh. Twelve areas of

Tirumala and Tirupati have been selected, where the peoples are used supplemented and groundwater for drinking purpose, and the water samples were subjected to systematic analysis with a view to understand the potability of drinking water sources. 12. Seasonal assessment of water quality in Tirupati- Tho Damodharam, S. Suresh. Water quality studies were conducted for surface and sub-surface water in Tirupati. Water was monitored during summer, winter and rainy seasons to assess the status and identify the impacts due to domestic, industrial and agricultural activities. Water quality assessments showed that the treated surface water supplied to Tirupati and Tirumula are within the limits of Indian standards for drinking water whereas sub surface water at certain locations near agricultural and drainage areas varied seasonally and showed higher values of hardness and chloride in April and December exceeding the limits. Seasonal assessment of water quality in Tirupati, A.P. India.

Summary:

Machine Learning:

Definition:

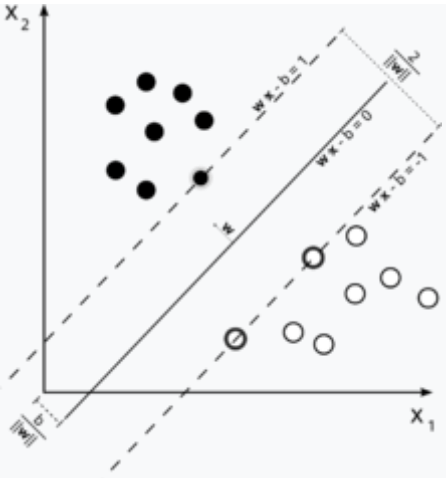
The data used for this research was obtained from Machine learning is **a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy**.

Machine learning approaches are traditionally divided into three broad categories, which correspond to learning paradigms, depending on the nature of the "signal" or "feedback" available to the learning system:

- Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.
- Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

- Reinforcement learning: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or playing a game against an opponent). As it navigates its problem space, the program is provided feedback that's analogous to rewards, which it tries to maximize.

Supervised learning



A support-vector machine is a supervised learning model that divides the data into regions separated by a linear boundary. Here, the linear boundary divides the black circles from the white.

Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs.[36] The data is known as training data, and consists of a set of training examples. Each training example has one or more inputs and the desired output, also known as a supervisory signal. In the mathematical model, each training example is represented by an array or vector, sometimes called a feature vector, and the training data is represented by a matrix. Through iterative optimization of an objective function, supervised learning algorithms learn a function that can be used to predict the output associated with new inputs.[37] An optimal function will allow the algorithm to correctly determine the output for inputs that were not a part of the training data. An algorithm that improves the accuracy of its outputs or predictions over time is said to have learned to perform that task.[20]

Types of supervised-learning algorithms include active learning, classification and regression.[28] Classification algorithms are used when the outputs are restricted to a limited set of values, and regression algorithms are used when the outputs may have any numerical value within a

range. As an example, for a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email.

Similarity learning is an area of supervised machine learning closely related to regression and classification, but the goal is to learn from examples using a similarity function that measures how similar or related two objects are. It has applications in ranking, recommendation systems, visual identity tracking, face verification, and speaker verification.

Unsupervised learning[edit]

Main article: Unsupervised learning

See also: Cluster analysis

Unsupervised learning algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points. The algorithms, therefore, learn from test data that has not been labeled, classified or categorized. Instead of responding to feedback, unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data. A central application of unsupervised learning is in the field of density estimation in statistics, such as finding the probability density function.[38] Though unsupervised learning encompasses other domains involving summarizing and explaining data features.

Cluster analysis is the assignment of a set of observations into subsets (called *clusters*) so that observations within the same cluster are similar according to one or more predesignated criteria, while observations drawn from different clusters are dissimilar. Different clustering techniques make different assumptions on the structure of the data, often defined by some *similarity metric* and evaluated, for example, by *internal compactness*, or the similarity between members of the same cluster, and *separation*, the difference between clusters. Other methods are based on *estimated density* and *graph connectivity*.

Semi-supervised learning

Main article: Semi-supervised learning

Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled

training data). Some of the training examples are missing training labels, yet many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce a considerable improvement in learning accuracy.

In weakly supervised learning, the training labels are noisy, limited, or imprecise; however, these labels are often cheaper to obtain, resulting in larger effective training sets.

Reinforcement learning

Main article: Reinforcement learning

Reinforcement learning is an area of machine learning concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward. Due to its generality, the field is studied in many other disciplines, such as game theory, control theory, operations research, information theory, simulation-based optimization, multi-agent systems, swarm intelligence, statistics and genetic algorithms. In machine learning, the environment is typically represented as a Markov decision process (MDP). Many reinforcement learning algorithms use dynamic programming techniques.[40] Reinforcement learning algorithms do not assume knowledge of an exact mathematical model of the MDP, and are used when exact models are infeasible. Reinforcement learning algorithms are used in autonomous vehicles or in learning to play a game against a human opponent.

Dimensionality reduction

Dimensionality reduction is a process of reducing the number of random variables under consideration by obtaining a set of principal variables.[41] In other words, it is a process of reducing the dimension of the feature set, also called the "number of features". Most of the dimensionality reduction techniques can be considered as either feature elimination or extraction. One of the popular methods of dimensionality reduction is principal component analysis (PCA). PCA involves changing higher-dimensional data (e.g., 3D) to a smaller space (e.g., 2D). This results in a smaller dimension of data (2D instead of 3D), while keeping all original variables in the model without changing the data.[42] The manifold hypothesis proposes that high-dimensional data sets lie along low-dimensional manifolds, and many dimensionality reduction

techniques make this assumption, leading to the area of manifold learning and manifold regularization.

Data processing:

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

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

Water Quality Index (WQI):

Water quality index (WQI) is the singular measure that indicates the quality of water and it is calculated using various parameters that are truly reflective of the water's quality. To conventionally calculate the WQI, nine water quality parameters are used, but if we did not have all of them, we could still estimate the water quality index with at least six defined parameters. We had five parameters, namely fecal coliform, pH, temperature, turbidity and total dissolved solids in our dataset. We also considered nitrites as the sixth parameter as the weight and relative importance of nitrites in the WQI calculation is stated to be equal to that of nitrates in multiple WQI studies [13–15]. Using these parameters and their assigned weightages, we calculated the

WQI of each sample as reflected in Equation (1), where q reflects the value of a parameter in the range of 0–100 and w_factor represents the weight of a particular parameter as listed in Table 2. WQI is fundamentally Figure 2. Outlier detection using box plot analysis. 3.3. Water Quality Index (WQI) Water quality index (WQI) is the singular measure that indicates the quality of water and it is calculated using various parameters that are truly reflective of the water's quality. To conventionally calculate the WQI, nine water quality parameters are used, but if we did not have all of them, we could still estimate the water quality index with at least six defined parameters. We had five parameters, namely fecal coliform, pH, temperature, turbidity and total dissolved solids in our dataset. We also considered nitrites as the sixth parameter as the weight and relative importance of nitrites in the WQI calculation is stated to be equal to that of nitrates in multiple WQI studies [13–15].

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.

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:

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. Trihalomethanes:

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.

(1) Water is not safe to drink and
(2) Water is safe to drink.

PROCEDURE:

Importing the necessary libraries for the program.

```
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
```

Read the dataset which have the necessary data.

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

Now the small part of dataset is

	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/100ML)	TOTAL COLIFORM(MPN/100ml)mean
0	1393	DAMANGANGA AT D/S OF MADHUBAN, DAMAN	DAMAN & DIU	30.6	6.7	7.5	203	NAN	0.1	11	27
1	1399	ZUARI AT D/S OF PT. WHERE KUMBARJRIA CANAL JOI...	GOA	29.8	5.7	7.2	189	2	0.2	4953	8391
2	1475	ZUARI AT PANCHAWADI	GOA	29.5	6.3	6.9	179	1.7	0.1	3243	5330
3	3181	RIVER ZUARI AT BORIM BRIDGE	GOA	29.7	5.8	6.9	64	3.8	0.5	5382	8443
4	3182	RIVER ZUARI AT MARCAIM JETTY	GOA	29.5	5.8	7.3	83	1.9	0.4	3428	5500

calculating the pH by the following

```
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)))))
```

calculating the B.D.O by the following

```
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)))))
```

Calculating the dissolved oxygen by the following

```
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)))))
```

Calculating the total coliform by the following

```
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)))))
```

Calculating the electric conductivity by the following

```
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)))))
```

Calculating the nitrate by the following

```
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)))))
```

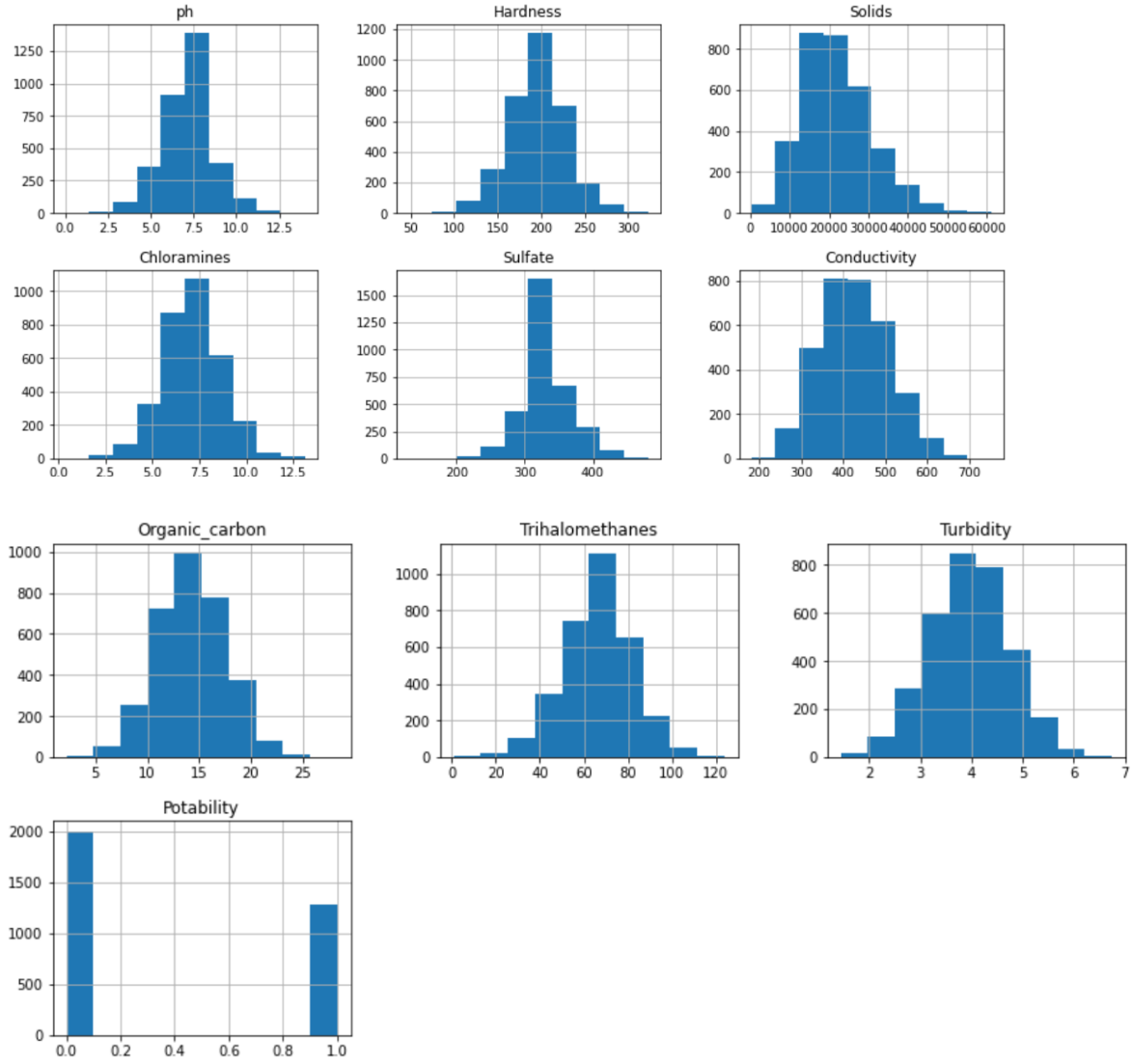
Calculation of water quality index by the following

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

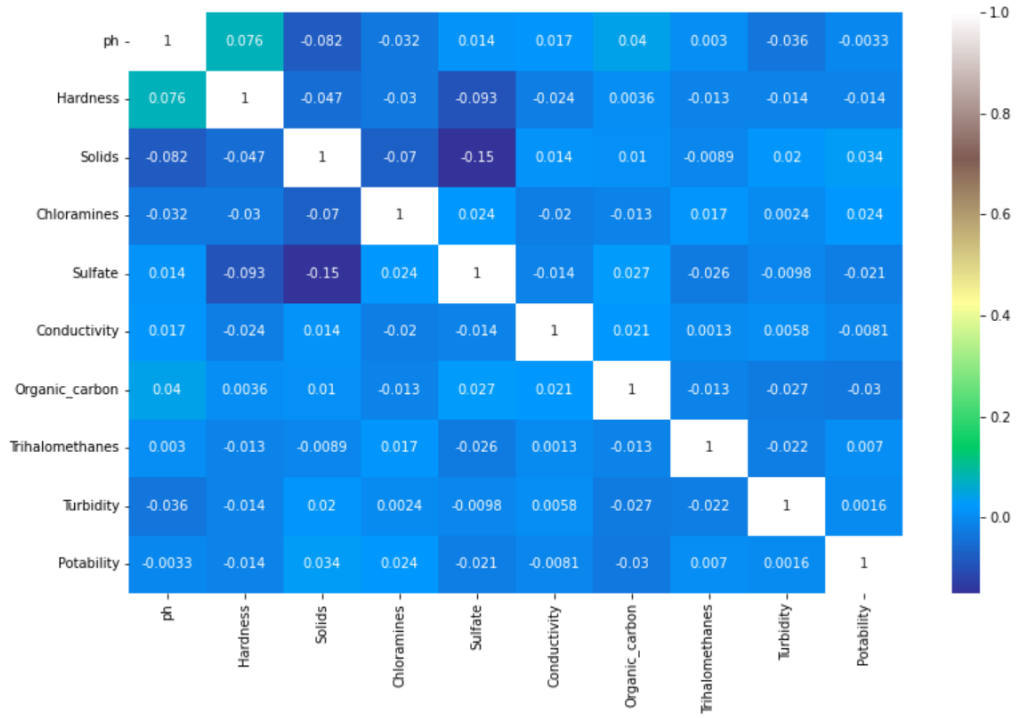
The dataset for the potability is following

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	2785.000000	3276.000000	3276.000000	3276.000000	2495.000000	3276.000000	3276.000000	3114.000000	3276.000000	3276.000000
mean	7.080795	196.369496	22014.092526	7.122277	333.775777	426.205111	14.284970	66.396293	3.966786	0.390110
std	1.594320	32.879761	8768.570828	1.583085	41.416840	80.824064	3.308162	16.175008	0.780382	0.487849
min	0.000000	47.432000	320.942611	0.352000	129.000000	181.483754	2.200000	0.738000	1.450000	0.000000
25%	6.093092	176.850538	15666.690300	6.127421	307.699498	365.734414	12.065801	55.844536	3.439711	0.000000
50%	7.036752	196.967627	20927.833605	7.130299	333.073546	421.884968	14.218338	66.622485	3.955028	0.000000
75%	8.062066	216.667456	27332.762125	8.114887	359.950170	481.792305	16.557652	77.337473	4.500320	1.000000
max	14.000000	323.124000	61227.196010	13.127000	481.030642	753.342620	28.300000	124.000000	6.739000	1.000000

The plot for the potability dataset is



The map for the above dataset is



Conclusion:

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 parameters.

In future works, we propose integrating the findings of this research in a large-scale IoT-based online monitoring system using only the sensors of the required parameters. The tested algorithms would predict the water quality immediately based on the real-time data fed from the IoT system. The proposed IoT system would employ the parameter sensors of pH, turbidity, temperature and TDS for parameter readings and communicate those readings using an Arduino microcontroller and ZigBee transceiver. It would identify poor quality water before it is released for consumption and alert concerned authorities. It will hopefully result in curtailment of people consuming poor quality water and consequently de-escalate harrowing diseases like typhoid and diarrhea. In this regard, the application of a prescriptive analysis from the

expected values would lead to future facilities to support decision and policy makers.