International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

Analysis and Prediction of Water Quality Data using Machine Learning Approaches and Exploratory Data Analysis

Ravindra Changala[1], A Tharun[2], A Sai Akshith[3], K Karthik[4]
[1]Assistant Professor, IT Department, Guru Nanak Institutions Technical Campus, Hyderabad, India
[2,3,4] IT Department, Guru Nanak Institutions Technical Campus, Hyderabad, India

Article Info

Publication Issue :
Volume 8, Issue 6
November-December-2022

Page Number : 188-193

Article History
Accepted: 07 Nov 2022
Published: 19 Nov 2022

ABSTRACT

Drinking Water Supply (DWS) is one of the most critical and sensitive systems to maintain city operations globally. In Europe, the contradiction between the fast growth of population and obsolete water supply infrastructure is even more prominent. The high standard water quality requirement not only provides convenience for people's daily life but also challenges the risk response time in the systems. Prevalent water quality regulations are relying on periodic parameter tests. This brings the danger in bacteria broadcast within the testing process which can last for 24-48 hours. In order to cope with these problems, we propose a EDA (Exploratory Data Analysis) model for water quality assessment. This model consists of two dimensions, including water quality parameters and score. Furthermore, we applied this model to predict water quality changes in the DWS system using a Random Forest algorithm using Pycaret. For a case study, we select an industrial water supply system. The preliminary results show that this model can provide high predictions & accuracy i.e., 73.76% for water quality understanding.

Keywords: Water Quality Monitoring, Water Quality Assessment, Water Quality Analysis, Chain of Custody.

I. INTRODUCTION

Water plays a vital role in everyone's life and is observed everywhere and in every form [1]. In Today's world, due to climatic changes and pollution the water quality is been affected in areas and various experiments are done to test the quality of water [2]. Due to poor water quality, risk occurs in the industrial areas which damage the whole environment and causes an economical loss [3].The root cause for many diseases such as typhoid, diarrhea, cholera is due to usage of contaminated water caused by increased industrialization and urbanization in India. [4]. According to reports form WHO, it is estimated that about 77 million people affected by contaminated water in India and 21% of diseases are caused due to it.[5] Due to insufficient rainfall and drying up of main reservoirs that supplies water, India faces water crisis frequently, hence making water one of the most precious and limited land resources. Many Organizations including WHO and BIS has framed standards for water parameters that can be used to efficiently analyze the quality of water. For checking the quality of water, conventionally it required to collect water samples and send it to the lab for testing which is a tedious process. [6] With IoT and Machine Learning algorithms it is easy to obtain the sensor values form a water sample, monitor and predict the quality of water at the comfort of our home. IOT is a buzzing technology that allows sensors to transfer data between them or to the cloud without the intervention of humans. [5]. Water quality index of the water, which helps in determining the quality of water, can be predicted by the extensive use of machine learning regression algorithm.

According to the report from the World Health Organization (WHO) [1], there are still over 681 million people on earth struggling to receive sufficient clean water. In DWS systems, water quality is a key factor across the whole process, from the water source, treatment, and distributed pipelines. Prevalent water quality is controlled using a series of parameters. They are different from countries or regions based on geographical and development conditions. Typical water quality parameters can be divided into three groups, as physical, chemical and biological parameters. To test the parameters in practice can take from several minutes to 24 hours. The outbreaks of contagious bacteria can be much faster than the testing time and therefore cause serious threats to people's health.

In this paper, we address this problem using the data analysis method. This becomes feasible thanks to the significant improvements in sensing and data analysis technologies. To predict the water quality in DWS systems. We need to build a model to provide the changes in water quality parameters. Because the time consumption from computation is much shorter than regular tests, we can, therefore, provide early warnings to risk detection. There are some trial works before in this field. In 2015, Yagur Kroll et al [2] introduced some sensors to monitor biological parameters. For water quality prediction, Holger et al [3] designed an artificial neural network for salinity level in an Australian river named Murray. In Iran, Orouji et al provided a series of algorithms for chemical parameter predictions, such as in [4] [5] [6]. Chang et al [7] proposed a framework to predict NH3-H for Dahan River. However, their works are focusing on single water quality parameter prediction, without considering the relationship between parameters. In addition, geographical and time factors are ignored. We propose a EDA model for water quality analysis,taking into account parameter, and time domains. This model provides a comprehensive understanding of DWS systems for water quality control. Furthermore, we use a Random Forest algorithm to predict these parameters in order to provide early warnings in the industrial process and decision-making support for corresponding actions.

II. RELATED WORKS

This model is used to predict the quality of the water. Data are been collected through sensors and saved in excel sheet. The method called Data preprocessing is to clean the given raw data. Cleaning provides the accurate date. Import and export of these data can be done anytime and anywhere. The following steps are to be followed. First the necessary libraries such as pandas, SCI-KIT learn are imported. And the important second step will be, the dataset are been converted to .CSV format and been uploaded. The dataset is checked for hidden values and if present it is filled with mean of the column.

Quality of surface water is a serious factor affecting human health and ecological systems. Accurate prediction of water quality parameters plays an important role in the management of rivers. Thus, different methods such as (support vector regression) SVR have been employed to predict water quality parameters. This paper applies SVR to predict eight water quality parameters including (sodium (Na+), potassium (K+), magnesium (Mg+2), sulfates (SO4–2), chloride (Cl–), power of hydrogen (pH), electrical conductivity (EC), and total dissolved solids (TDS)) at the Astane station in Sefidrood River, Iran.

To achieve an efficient SVR model, the SVR parameters should be selected carefully. Commonly, various techniques such as trial and error, grid search and metaheuristic algorithms have been applied to estimate these parameters.

Observations:

- Less Improved predictions when compared to new proposed system.
- Slow convergence, easy to fall into the local optimal solution and premature convergence.
- It is not suitable for large data sets

III. LITERATURE SURVEY

Y. Amit, D. Geman, and K. Wilder, Modeling water-quality parameters using genetic algorithm–least squares support vector regression and genetic programming, the modeling and monitoring of water-quality parameters is necessary because of the ever increasing use of water resources and contamination caused by sewage disposal. This study employs two data-driven methods for modeling water-quality parameters. The methods are the least-squares support vector regression (LSSVR) and genetic programming (GP). Model inputs to the LSSVR algorithm and GP were determined using principal component analysis (PCA). The coefficients of the LSSVR were selected by sensitivity analysis employing statistical criteria. The results of the sensitivity analysis of the LSSVR showed that its accuracy depends strongly on the values of its coefficients. The value of the Nash-Sutcliffe (NS) statistic was negative for 60% of the combinations of coefficients applied in the sensitivity analysis. That is, using the mean of a time series would produce a more accurate estimate of water-quality parameters than the LSSVR method in 60% of the combinations of parameters tried. The genetic algorithm (GA) was combined with LSSVR to produce the GA-LSSVR algorithm with which to achieve improved accuracy in modeling water-quality parameters. The GA-LSSVR algorithm and the GP method were employed in

modeling Na+Na+, K+K+, Mg2+Mg2+, SO2–4SO42–, Cl–Cl-, pH, electric conductivity (EC), and total dissolved solids (TDS) in the Sefidrood River, Iran. The results indicate that the GA-LSSVR algorithm has better accuracy for modeling water-quality parameters than GP judged by the coefficient of determination (R2R2) and the NS criterion. The NS static established, however, that the GA-LSSVR and GP methods have the capacity to model water-quality parameters accurately.

N. Mahmoudi, H. Orouji, and E. Fallah-Mehdipour , Integration of shuffled frog leaping algorithm and support vector regression for prediction of water quality parameters, Quality of surface water is a serious factor affecting human health and ecological systems. Accurate prediction of water quality parameters plays an important role in the management of rivers. Thus, different methods such as (support vector regression) SVR have been employed to predict water quality parameters. This paper applies SVR to predict eight water quality parameters including (sodium (Na^+), potassium (K^+), magnesium (Mg^{+2}), sulfates (SO_4^{-2}), chloride (Cl^-), power of hydrogen (pH), electrical conductivity (EC), and total dissolved solids (TDS)) at the Astane station in Sefidrood River, Iran. To achieve an efficient SVR model, the SVR parameters should be selected carefully. Commonly, various techniques such as trial and error, grid search and metaheuristic algorithms have been applied to estimate these parameters. This study presents a novel tool for estimation of quality parameters by coupling SVR and shuffled frog leaping algorithm (SFLA) . Results of SFLA-SVR compared with genetic programming (GP) as a capable method in water quality prediction. Using SFLA-SVR, average of RMSE for training and testing of six combinations of data sets for all of the water quality parameters improved 57.4 % relative to GP. These results indicate that the new proposed SFLA-SVR tool is more efficient and powerful than GP for determining water quality parameters.

F.-J. Chang, Y.-H. Tsai, P.-A. Chen, A. Coynel, and G. Vachaud, "Modeling water quality in an urban river using hydrological factors–data driven approaches", Contrasting seasonal variations occur in river flow and water quality as a result of short duration, severe intensity storms and typhoons in Taiwan. Sudden changes in river flow caused by impending extreme events may impose serious degradation on river water quality and fateful impacts on ecosystems. Water quality is measured in a monthly/quarterly scale, and therefore an estimation of water quality in a daily scale would be of good help for timely river pollution management. This study proposes a systematic analysis scheme (SAS) to assess the spatio-temporal interrelation of water quality in an urban river and construct water quality estimation models using two static and one dynamic artificial neural networks (ANNs) coupled with the Gamma test (GT) based on water quality, hydrological and economic data. The Dahan River basin in Taiwan is the study area. Ammonia nitrogen (NH3–N) is considered as the representative parameter, a correlative indicator in judging the contamination level over the study. Key factors the most closely related to the representative parameter (NH3–N) are extracted by the Gamma test for modeling NH3–N concentration, and as a result, four hydrological factors (discharge, days w/o discharge, water temperature and rainfall) are identified as model inputs. The modeling results demonstrate that the nonlinear autoregressive with exogenous input (NARX) network furnished with recurrent connections can accurately estimate NH3–N concentration with a very high coefficient of efficiency value (0.926) and a low RMSE value (0.386 mg/l). Besides, the NARX network can suitably catch peak values that mainly occur in dry periods (September–April in the study area), which is particularly important to water pollution treatment. The proposed SAS suggests a promising approach to reliably modeling the spatio-temporal NH3–N concentration based solely on hydrological data, without using water quality sampling data. It is worth noticing that such estimation can be made in a much shorter time interval of interest (span from a monthly scale to a daily scale) because hydrological data are long-term collected in a daily scale. The proposed SAS favorably makes NH3–N concentration estimation much easier (with only hydrological field sampling) and more efficient (in shorter time intervals), which can substantially help river managers interpret and estimate water quality responses to natural and/or manmade pollution in a more effective and timely way for river pollution management.

H. Rowley, S. Baluja, and T. Kanade, "Random decision forests", Decision trees are attractive classifiers due to their high execution speed. But trees derived with traditional methods often cannot be grown to arbitrary complexity for possible loss of generalization accuracy on unseen data. The limitation on complexity usually means suboptimal accuracy on training data. Following the principles of stochastic modeling, we propose a method to construct tree-based classifiers whose capacity can be arbitrarily expanded for increases in accuracy for both training and unseen data. The essence of the method is to build multiple trees in randomly selected subspaces of the feature space. Trees in, different subspaces generalize their classification in complementary ways, and their combined classification can be monotonically improved. The validity of the method is demonstrated through experiments on the recognition of handwritten digits.

F A Aziz , M Sarosa and E Rohadi, "Monitoring system water pH rate, turbidity, and temperature of river water ", Water is a substance that is important in life after air if it has been polluted by chemicals it is very dangerous for living things. The development of the industry uses chemicals in a production that causes hazardous waste, even though it has been reprocessed in the filtering process, if it is less than perfect, the water will be polluted if disposed of in the river flow. Identification and monitoring measures are needed that require media to use a device that includes a sensor containing the pH of water that comes from the content of acidic substances in water, turbidity or the content of solid objects in the water, and water temperature. The data collection method uses a node MCU microcontroller. Then the data is sent via a wireless connection to be stored on the database server. Testing to see the results of monitoring using an Android application that is connected to the server. From the results, it can be concluded that the data displayed on the application can run smoothly, the value displayed can be seen in real-time.

IV. PROPOSED SYSTEM

Water is prime natural resource. In this modern environment, fresh water is not available due to increase of population, agricultural and industries. The quality of water is determined by the physico-chemical properties of water and micro biological characteristics.

The physico-chemical analysis of water samples was done my several researchers using by standard methods. Water quality prediction is done using various data mining techniques. In this research data classification techniques are used to study various classifiers and to find out most accurate classifier.

Observations:

It speeds up the experiment cycle exponentially and makes you more productive.

It will get a working model in no time with very little effort.

It can produce good predictions that can be understood easily.

It can handle large datasets efficiently & provides a higher level of accuracy in predicting outcomes over the decision tree algorithm

This model is easily extended on these domains. Secondly, we use a random forest algorithm to predict the most dangerous biological water quality parameters based on easily tested physical and chemical parameters. For a case study, we applied our method in an industrial DWS system in. The results show our method is feasible and fulfills the requirements from the domain.

Dataset Description:

The dataset consists of individual data in that there are 3277 rows & 10 columns in the dataset, which are described below.

1. Ph
2. Hardness
3. Solids
4. Chloramines
5. Sulfate
6. Conductivity
7. Organic_carbon
8. Trihalomethanes
9. Turbidity
10. Potability
11. Score

Shuffled Frog Leaping Algorithm and Support Vector Regression Integration

This study presents a novel tool for estimation of quality parameters by coupling SVR and shuffled frog leaping algorithm (SFLA) . Results of SFLA-SVR compared with genetic programming (GP) as a capable method in water quality prediction. Using SFLA-SVR, average of RMSE for training and testing of six combinations of data sets for all of the water quality parameters improved 57.4 % relative to GP. These results indicate that the new proposed SFLA-SVR tool is more efficient and powerful than GP for determining water quality parameters.

Random Forest method using Pycaret

PyCaret is an open-source, low-code machine learning library in Python that automates machine learning workflows. It is an end-to-end machine learning and model management tool that exponentially speeds up the experiment cycle and makes you more productive which uses sklearn under the hood, lets you create and test regression models with a few lines of code. It includes a variety of algorithms, as well as the ability to plot and do hyper parameter tuning which Random forest trees generate the classification tree based on the predictor variables. First, an analysis on physic chemical properties of ground water quality assessment is done in the laboratory for all the water samples collected from the regions of. Water Quality Index (WQI) values were calculated. After, water quality classification is determined by using WQI values. By using this trained data, Random Forest approach is applied to test the water quality whether they are suitable for drinking or not.

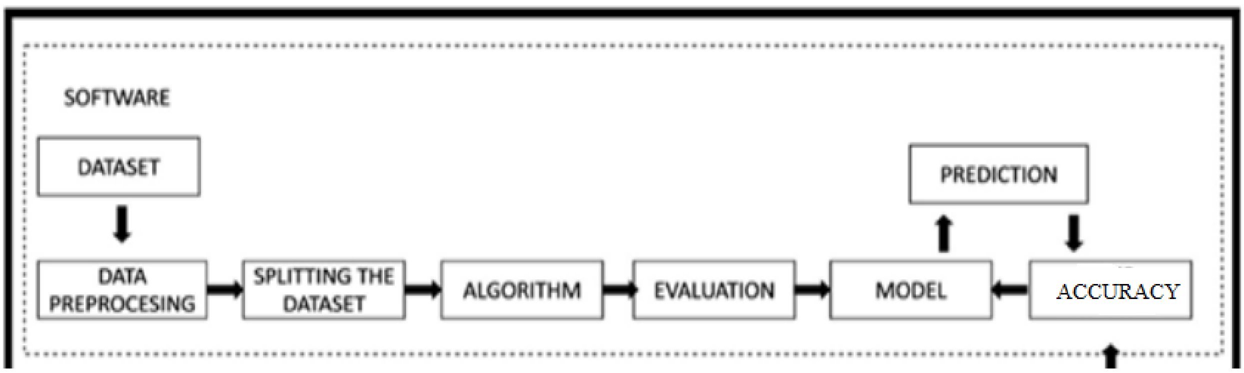


Fig 1. Architecture of the work

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

DEVELOPMENT TOOLS

Sample Code segment

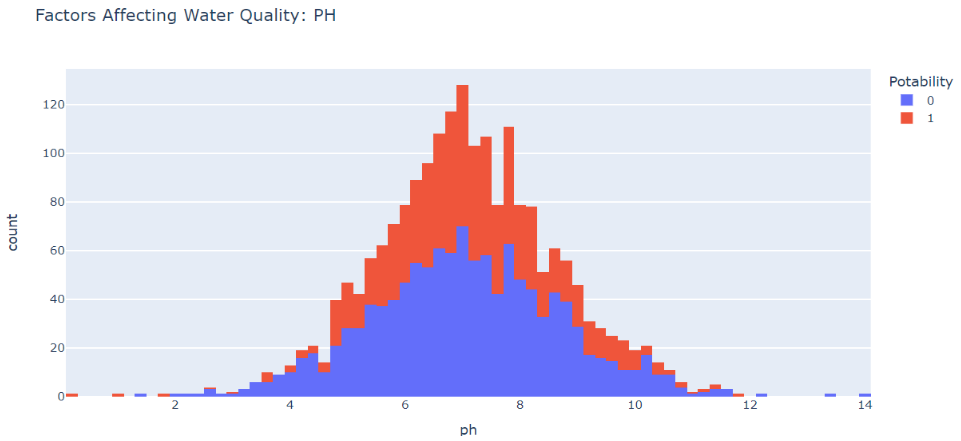
```
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import numpy as np
import plotly.express as px
from pycaret.classification import *
```

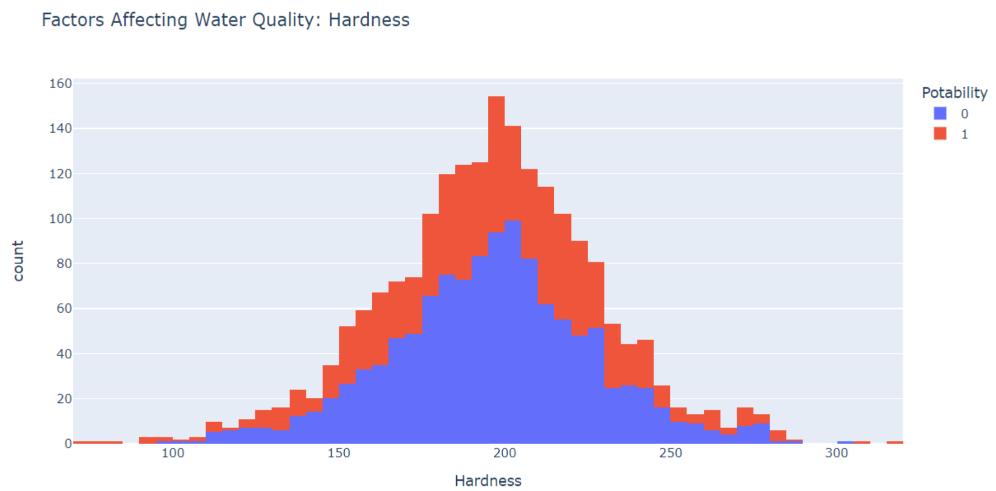
```
data = pd.read_csv("water_potability.csv")
data.head()
data = data.dropna()
data.isnull().sum()
plt.figure(figsize=(15, 10))
sns.countplot(data.Potability)
plt.title("Distribution of Unsafe and Safe Water")
plt.show()
data = data
figure = px.histogram(data, x = "ph",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: PH")
figure.show()
figure = px.histogram(data, x = "Hardness",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Hardness")
figure.show()
figure = px.histogram(data, x = "Solids",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Solids")
figure.show()
figure = px.histogram(data, x = "Chloramines",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Chloramines")
figure.show()
figure = px.histogram(data, x = "Sulfate",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Sulfate")
figure.show()
figure = px.histogram(data, x = "Conductivity",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Conductivity")
figure.show()
figure = px.histogram(data, x = "Organic_carbon",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Organic Carbon")
figure.show()
figure = px.histogram(data, x = "Trihalomethanes",
                      color = "Potability",
                      title= "Factors Affecting Water Quality: Trihalomethanes")
figure.show()
figure = px.histogram(data, x = "Turbidity",
```

```
                color = "Potability",
                title= "Factors Affecting Water Quality: Turbidity")
figure.show()

correlation = data.corr()
correlation["ph"].sort_values(ascending=False)
clf = setup(data, target = "Potability", silent = True, session_id = 786)
compare_models()

model = create_model("rf")
predict = predict_model(model, data=data)
predict.head()
```

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

Table 1. Description of Water data set





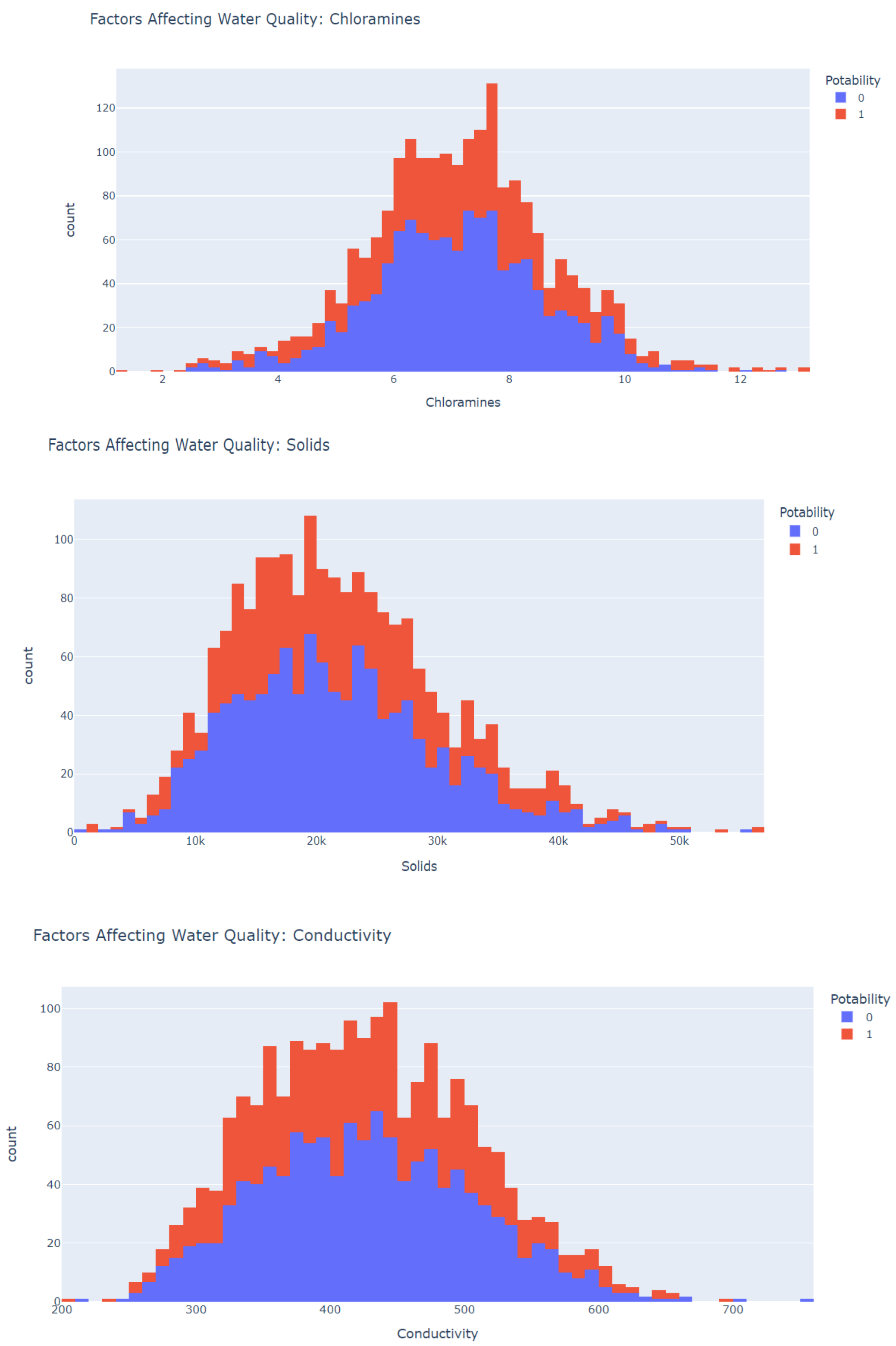
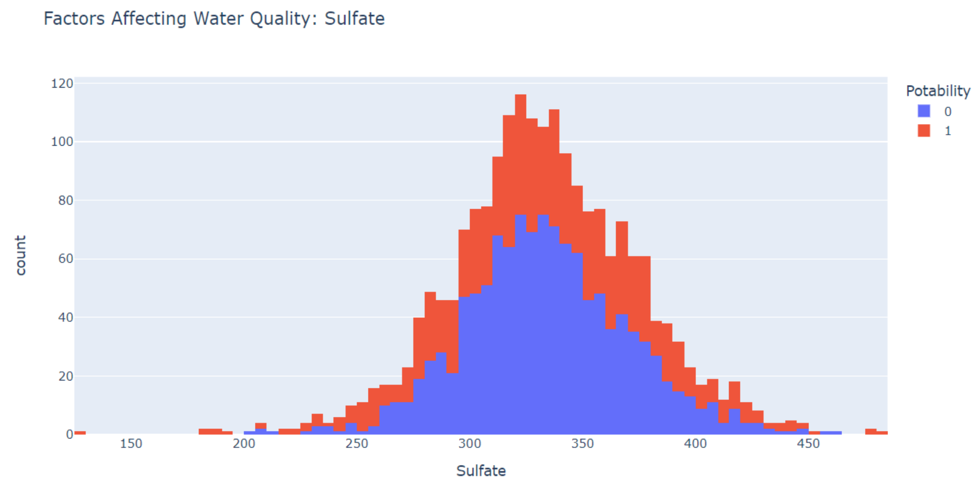


Fig 2. Show various factors affecting water quality



	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.6830	0.7005	0.4197	0.6744	0.5133	0.2976	0.3182	0.724
qda	Quadratic Discriminant Analysis	0.6823	0.7192	0.3985	0.6883	0.5013	0.2917	0.3174	0.022
et	Extra Trees Classifier	0.6816	0.6941	0.3861	0.6858	0.4916	0.2863	0.3123	0.557
lightgbm	Light Gradient Boosting Machine	0.6652	0.6916	0.4762	0.6078	0.5324	0.2781	0.2840	0.172
gbc	Gradient Boosting Classifier	0.6602	0.6738	0.3718	0.6306	0.4667	0.2419	0.2603	0.339
nb	Naive Bayes	0.6184	0.6078	0.2478	0.5545	0.3412	0.1261	0.1462	0.019
dt	Decision Tree Classifier	0.6034	0.5895	0.5186	0.5049	0.5097	0.1775	0.1784	0.027
lr	Logistic Regression	0.5984	0.5199	0.0071	0.1900	0.0134	0.0028	0.0127	0.355
ridge	Ridge Classifier	0.5984	0.0000	0.0089	0.1583	0.0168	0.0035	0.0056	0.021
lda	Linear Discriminant Analysis	0.5977	0.4903	0.0089	0.1500	0.0167	0.0021	0.0024	0.022
ada	Ada Boost Classifier	0.5956	0.5671	0.2919	0.4896	0.3644	0.0972	0.1034	0.173
knn	K Neighbors Classifier	0.5743	0.5423	0.3644	0.4642	0.4070	0.0826	0.0846	0.121
svm	SVM - Linear Kernel	0.5194	0.0000	0.3982	0.1604	0.2287	-0.0014	-0.0104	0.027

Table 2. Various F1 Score values of data set

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability	Label	Score
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0	0	0.87
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0	0	0.91
5	5.584087	188.313324	28748.687739	7.544869	326.678363	280.467916	8.399735	54.917862	2.559708	0	0	0.83
6	10.223862	248.071735	28749.716544	7.513408	393.663396	283.651634	13.789695	84.603556	2.672989	0	0	0.89
7	8.635849	203.361523	13672.091764	4.563009	303.309771	474.607645	12.363817	62.798309	4.401425	0	0	0.94

Table 3. Accuracy of the various factors

V. CONCLUSION & FUTURE ENHANCEMENTS

Our Solution that addresses water quality problems are been discussed in our paper. We have used low cost microcontroller and affordable sensors to build the system. The Webpage integrated with Thing Speak and Machine Learning model was built which will help the user to monitor the values as well as predicts the quality of the sample water.

Our further research is been done to improve the system that will alert the concerned official in case the water quality is poor, so that they would take necessary actions. The system will also fetch real time values from sensors and automatically predict the quality without requiring the user to manually enter the value.

So this is how you can analyze the quality of water and train a machine learning model to classify safe and unsafe water for drinking. Access to safe drinking water is one of the essential needs of all human beings. From a legal point of view, access to drinking water is one of the fundamental human rights. Many factors affect water quality, it is also one of the major research areas in machine learning.

VI. REFERENCES

1. Shamsul Haq, Shoukat Ara, Syed Maqbool Geelani and Asma Absar Bhatt, 2016, Assessment of surface and ground water for irrigational purposes, International Journal of Applied And Pure Science and Agriculture, vol.2, no. 2.
2. Manish Kumar, Gurmeet Singh, Tushara Chaminda, Pham Van Quan and Keisuke Kuroda, 2014, Emerging Water Quality Problems in Developing Countries, The Scientific World Journal . vol.2014, no.158796.
3. Roy and Ritabrata, 2018, An Introduction to water quality analysis, International Journal for Environmental Rehabilitation and Conservation, vol.2018, no. 3, pp. 94-100.
4. Subodh Kumar, Hari Mohan Meena and Kavita Verma, 2017, Water Pollution in India: Its Impact on the Human Health: Causes and Remedies, International Journal of Applied Environmental Sciences, vol.12, no. 2, pp. 275-279.
5. K Gopavanitha and S Nagaraju, 2017, A low cost system for real time water quality monitoring and controlling using IOT, International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), DOI.10.1109/ICECDS.2017.8390054.
6. Brinda Das and P C Jain, 2017, Real-time water quality monitoring system using Internet of Things, 2017 International Conference on Computer, Communications and Electronics (Comptelix), DOI 10.1109/COMPTELIX.2017.8003942.
7. Vrushali Y Kulkarni and Pradeep K Sinha, 2014, Effective Learning and Classification using Random Forest Algorithm, International Journal of Engineering and Innovative Technology (IJEIT), vol.3, no. 11, pp. 45-53.
8. Anna Bosch, Andrew Zisserman and Xavier Munoz , 2007, Image Classification using Random Forests and Ferns, IEEE International Conference on Computer Vision, DOI 10.1109/ICCV.2007.4409066 .
9. Hossam M Zawbaa, Maryam Hazman, Mona Abbass and Aboul Ella Hassanien, 2014, Automatic fruit classification using random forest algorithm, International Conference on Hybrid Intelligent Systems, DOI.10.1109/HIS.2014.7086191 .
10. Gerard Biau, 2012, Analysis of a Random Forests Model, Journal of Machine Learning Research, vol.13, pp. 1063-1095.
11. Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Asad A Shah and Rabia Irfan, 2019, Efficient Water Quality Prediction Using Supervised Machine Learning, MDPI, vol.11, no. 2210, pp. 1-14.

12. Martin Breuss, Python Web Applications: Deploy Your Script as a Flask App, https://realpython.com/python-web-applications/ .
13. Abdelhadi Dyouri, How to Make a Web Application Using Flask in Python 3. https://www.digitalocean.com/community/tutorials/how-to-make-a-web-application-using-flask-in-python-3 .
14. https://in.mathworks.com/help/thingspeak/
15. Kancheti Spandana and V R Seshagiri Rao, 2018, Internet of Things (IOT) Based Smart Water Quality Monitoring System, International Journal of Engineering & Technology , vol.7, no. 3, pp. 259-262.
16. V D Vaishnavi and M A Gaikwad, 2017, Water Quality Monitoring System Based on IOT, Advances in Wireless and Mobile Communications, vol.10, no. 5, pp. 1107-1116. 2455-3506.

Cite this article as :

Ravindra Changala, A Tharun, A Sai Akshith, K Karthik, "Analysis and Prediction of Water Quality Data using Machine Learning Approaches and Exploratory Data Analysis", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 6, pp. 188-193, November-December 2022.
Journal URL : https://ijsrcseit.com/CSEIT228625