Improve the Purity and Quality of the Water Using the Correlations Among the Feature Variables

L Roshini and C.R.K. Reddy

March 14, 2023

Improve The Purity and Quality of the Water Using the Correlations Among the feature Variables

Roshini L[1] [0009-0001-9709-0289] and C.R.K.Reddy[2] [0000-0003-2246-3930]

[1] Research Scholar, Osmania University, Hyderabad, India-500044
[1] Assistant Professor, CSIT Department, CVR College of Engineering, Hyderabad, India-501510
[2] Department of Computer Science & Engineering, MGIT, Hyderabad, India-500044,
roshini1604@gmail.com,crkreddy_cse@mgit.ac.in

Abstract: The assessment of water purity and the position of sources for safe potable water has been one of the major problems the world has encountered recently. The primary challenge is preserving water and maintaining the purity of the water from extreme environmental pollution because the mode of water poisoning is unpredictable, making it difficult to assess and maintain. As a result, maximizing resource management is very efficient for improving water purity. There are some particular water-related factors required for the determination of the water quality indicator (WQI). However, the traditional way of computing the WQI requires a lot of effort, and occasionally mistakes are discovered during the computation procedure. In this research, we evaluate the cleanliness of the water and forecast its grade using machine learning techniques. If the water is out of purity our algorithms will identify the chemical compositions of the water and proposes with a required combination in order to make the water balance with pH and stable for drinking.Therefore, effective administration of water supplies is crucial to enhancing water purity, both conventionally and through the use of machine learning algorithms. This study's primary objectives are the collection of water data, analysis of chemical combinations, correlations between combinations, water samples were analyzed for their physico-chemical characteristics, the WQI was calculated using this information, and machine learning methods were used to forecast water purity. Several statistical metrics were used to evaluate the success of the forecast algorithms.

Keywords: Correlations, Machine Learning, Water purity, WQI, chemical compositions

1. INTRODUCTION

No animal or plant on the earth can survive without water. However, due to the industry's yearly expansion in reaction to increasing demand, and because these industries routinely discharge toxic refuse into waterways and lakes, water is frequently polluted. Water contamination causes millions of deaths, vast financial losses, and farming area degradation every year. According to numerous studies, the condition of groundwater in the majority of countries has drastically declined recently. As a result, the condition of the groundwater is declining daily [1]. A simple, effective, and sensible way to evaluate water quality for different reasons is to determine the "quality determining factors" of water, which contribute to the identification of water pollution.

Monitoring and evaluating water purity is an essential job for the best effective administration of different water supplies. The rise in demand for pure, unpolluted water is either directly or tangentially related to the ongoing expansion of human activity. In order to create fresh, cutting-edge models for water quality forecasts, academics are motivated by the need for pure water to study the physical characteristics of groundwater in the local region. Lately, the availability of potable water supplies has become a major concern on a global scale[2]. As a result, it is now essential to evaluate water purity and forecast its future state.

The water quality indicator (WQI), a single, special measurement made by merging a number of physical and chemical variables, is used to gauge the level of water pollution. The WQI is calculated using carefully selected water quality factors, and the results are remarkable. There are several different formulas for WQIs that have been approved by various scholars to evaluate the purity and applicability of both groundwater and public water. However, a number of factors, including poor sanitation, extensive industrialization, and mining, and the addition of excessive fertilizers and insecticides to agricultural

lands, are to blame for the decline in water quality, which is alarming for both human civilization and the environment as a whole [3].
The laboratory analysis of the collected water samples included the following calculations: pH, Total Hardness, Alkalinity, Turbidity, Iron (Fe), Chloride (Cl), Fluoride (F), Dissolved Solids (TDS), Sulphate (SO4), Nitrate (NO3), Calcium (Ca), Magnesium (Mg), Arsenic (As), Conductivity Total Coliform, Fecal Coliform, and Total Residual Chlorine.
The assessment of the water purity is considered with the chemicals compositions their index levels ranges, and quantity available in the water. The combinations change as the change of composition, this can show the impact on the next available composition in the water, which might make the water more pollutant. To solve this issue our model Purity regulator based composition model (PRCM) will describe the solvents to be added within the range and at the same time maintain the water purity pH values for healthy drinking. Hence the study took different parameters.

2. RELATED WORKS

The groundwater quality assessment is very important and imperative, so regularly this is being done, all over periodically. Since the groundwater is used as drinking water hence a complete physico-chemical analysis become necessary as it is related to health. The various agencies (UNESCO, WHO, BSI, etc.) have set physico-chemical & other parameters to assess the quality of water. The reports of such studies carried out by various workers in the wider area are worth mentioning.

Prasad et al[4] have analyzed findings of physico-chemical variables like pH, EC, TDS, total hardness, nitrate, and sulfate. According to them, EC and TDS showed a strong favorable association. According to Garg et al. [5], some metals, such as iron, copper, zinc, and nickel, are necessary for the correct division of living systems, whereas non-essential metals, such as poisonous Lead, Cr, Cd, and Al, are not. They discovered that the water in the drains closest to the disposal site has the greatest concentrations of metal ions, which may be the result of seeping from the solid debris.

3. METHODOLOGIES

The usual approach is applied to evaluate the different physiochemical parameters [6]. Adopting the recommended method[7], the gathered water samples are kept for analysis of hazardous heavy metals by adding 5 milliliters of pure HNO3 to a liter of water samples to keep a pH lower than 4.0. The samples are then precisely filtered using Whitman 40-filter paper, and measurement can be performed using an atomic absorption spectrometer (AAS).

Operational tracking must be precise and dependent on observing pertinent biological, hydro-morphological, and physico-chemical parameters in order to determine the parameters to be determined for water evaluation. These global environmental surveillance tools offer three types of criteria for measuring water quality:

Water quality

A collection of physical, molecular, biological, and bacteriological traits serve as indicators of water purity. These traits are also referred to as markers or factors. Water can be categorized according to physical, molecular, biological, and bacteriological factors, which enables it to be used for a particular purpose.
The standards for water purity vary depending on the intended uses. Therefore, drinkable water cannot contain any toxins or microorganisms that could harm a person's health. Large quantities of sodium ions, high nitrate concentrations, or high concentrations of other pollutants are prohibited in water used in irrigation. Water usage regulations for businesses are less stringent than those for potable water [3].
The sort of water supply is another factor that affects water purity, which also varies based on geological, climatic, and land-use factors [3]. A legislative structure for the preservation of water has also been created by the European Union [EU] [4]. All nations have set water purity standards in compliance with the recommendations of the WHO [3]. The EU's core regulations are strictly adhered to in European nations [4].
Numerous water quality factors or signs need to be evaluated and watched in order to comprehend the general health of an environment and the state of the water. According to [6], tracking of surface water

quality is required to evaluate temporal and geographic area differences. Environmental evaluations and the creation of water guidelines are both results of the process of tracking the purity of atmospheric water. For the study of water quality, a surveillance program and the criteria to be evaluated should be carefully selected for each location and each variety of water. Despite the fact that many water factors are crucial for maintaining ecology or human health, it is not practical to analyze every parameter.[6, 7].
While biological factors offer a comprehensive and complicated study of the ecosystem, the rapid evaluation of water purity relies heavily on molecular and physical factors [8].

4. DATA SET

The data has 21 parameters with salinity indicators, Heavy metals, other relevant indicators and the target parameter.

Table 1: Parameters of the data

Aluminum	Ammonia	Arsenic
Barium	Cadmium	Chloramine
Chromium	Copper	Fluoride
Bacteria	Viruses	Lead
Nitrates	Nitrites	Mercury
Perchlorate	Radium	Selenium
Silver	Uranium	is safe

aluminium	ammonia	arsenic	barium	cadmium	chloramine	chromium	copper	flouride	bacteria	viruses	lead	nitrates	nitrites	mercury	perchlorate	radium	selenium
1.65	9.08	0.04	2.85	0.007	0.35	0.83	0.17	0.05	0.2	0	0.054	16.08	1.13	0.007	37.75	6.78	0.
2.32	21.16	0.01	3.31	0.002	5.28	0.68	0.66	0.9	0.65	0.65	0.1	2.01	1.93	0.003	32.26	3.21	0.
1.01	14.02	0.04	0.58	0.008	4.24	0.53	0.02	0.99	0.05	0.003	0.078	14.16	1.11	0.006	50.28	7.07	0.
1.36	11.33	0.04	2.96	0.001	7.23	0.03	1.66	1.08	0.71	0.71	0.016	1.41	1.29	0.004	9.12	1.72	0.
0.92	24.33	0.03	0.2	0.006	2.67	0.69	0.57	0.61	0.13	0.001	0.117	6.74	1.11	0.003	16.9	2.41	0.
0.94	14.47	0.03	2.88	0.003	0.8	0.43	1.38	0.11	0.67	0.67	0.135	9.75	1.89	0.006	27.17	5.42	0.
2.36	5.6	0.01	1.35	0.004	1.28	0.62	1.88	0.33	0.13	0.007	0.021	18.6	1.78	0.007	45.34	2.84	0
3.93	19.87	0.04	0.66	0.001	6.22	0.1	1.86	0.86	0.16	0.005	0.197	13.65	1.81	0.001	53.35	7.24	0.
0.6	24.58	0.01	0.71	0.005	3.14	0.77	1.45	0.98	0.35	0.002	0.167	14.66	1.84	0.004	23.43	4.99	0.
0.22	16.76	0.02	1.37	0.007	6.4	0.49	0.82	1.24	0.83	0.83	0.109	4.79	1.46	0.01	30.42	0.08	0.
3.27	3.6	0.001	2.69	0.005	5.75	0.15	0.6	1.29	0.04	0.008	0.145	8.47	1.25	0.006	55.4	7.8	0.
1.35	21.96	0.04	0.84	0.002	0.1	0.76	0.17	0.58	0.52	0.52	0.011	18.4	1.49	0.009	21.52	1.3	0.
1.88	19.26	0.02	2.78	0.008	0.05	0.42	1	0.09	0.91	0.91	0.103	4.37	1.95	0.006	22.12	1.97	0.
4.93	23.98	0.04	3.05	0.008	0.7	0.51	1.35	1.07	0.7	0.7	0.101	1.16	1.11	0.008	26.8	5.58	0.
2.89	18.82	0.05	3.77	0.008	5.99	0.54	0.79	0.54	0.2	0.009	0.126	17.56	1.82	0.009	17.54	4.33	0
0.61	2.41	0.03	0.59	0.002	1.94	0.77	1.54	0.62	0.23	0.001	0.017	1.99	1.08	0.007	11.16	0.98	0.
3.47	15.84	0.02	0.06	0.001	5.29	0.47	1.08	1.43	0.89	0.89	0.08	1.91	1.2	0.008	0.18	6.89	0.
2.11	17.03	0.02	0.88	0.009	7.78	0.88	1.15	0.34	0.85	0.85	0.065	17.86	1.53	0.003	19.4	1.14	0
4.88	26.94	0.02	0.36	0.001	1.21	0.68	0.71	0.99	0.75	0.75	0.071	0.31	1.22	0.002	56.7	1	
4.12	17.99	0.02	3.43	0.006	0.01	0.41	1.82	0.22	0.99	0.99	0.108	8.06	1.76	0.005	24.29	0.88	0
0.68	18.99	0.001	0.04	0.006	4.57	0.2	1.18	1	0.92	0.92	0.086	9.46	1.41	0.007	21.79	3.05	0.
1.15	8.12	0.02	0.97	0.007	3.47	0.65	1.51	1.46	0.58	0.58	0.061	8.96	1.5	0.004	14.6	1.74	0.
0.27	10.67	0.02	0.55	0.001	3.74	0.12	1.77	0.43	0.8	0.8	0.114	12.69	1.18	0.008	34.64	0.9	0.

Figure 1: Sample data set with parameters

4.1 Parameter Correlation:

A connection between two factors is indicated by correlation. The correlation coefficient refers to the measurement of the relationship between two such factors. It is also referred to as the strength of the relationship between the two factors.

Covariance

The correlation of two variables, x, and y, in a set of data measures their linear connection. A positive covariance would suggest that the factors have a positive linear connection, whereas a negative covariance would suggest the reverse.

The instance when specified in terms of the group averages, covariance is:

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

A data set's correlation value for two variables is calculated by dividing their covariance by the sum of their respective standard deviations. It measures the straight relationship between the two in a standardized way.
The sample correlation coefficient is defined loosely by the following formula, where s_x and s_y are the sample standard deviations and s_{xy} is the sample covariance.

$$r_{xy} = \frac{s_{xy}}{s_x \quad s_y}$$

The scatter plot will almost definitely follow a straight line with a positive slope if the correlation value is close to 1, showing that the variables are positively linearly related. When the number is 1, the scatter plot almost perfectly traces a straight line with a downward inclination, demonstrating a negative linear relationship between the variables. Additionally, a number of zero denotes a weak linear relationship between the variables.
The correlation coefficient is calculated using the following algorithm, where x stands for the values of the independent variable (in this case, height) and y stands for the values of the dependent variable (in this case anatomical dead space). The computation looks like this:

$$r = \frac{\sum(x - \overline{x})(y - \overline{y}\ )}{\sqrt{[\sum(x - x^{-2})(y - y^{-2}\ )]}}$$

This can be shown to be equal to:

$$r = \frac{\sum xy \ - \ \bar{n}\ \bar{x}\bar{y}}{(n-1)SD(x)SD(y)}$$

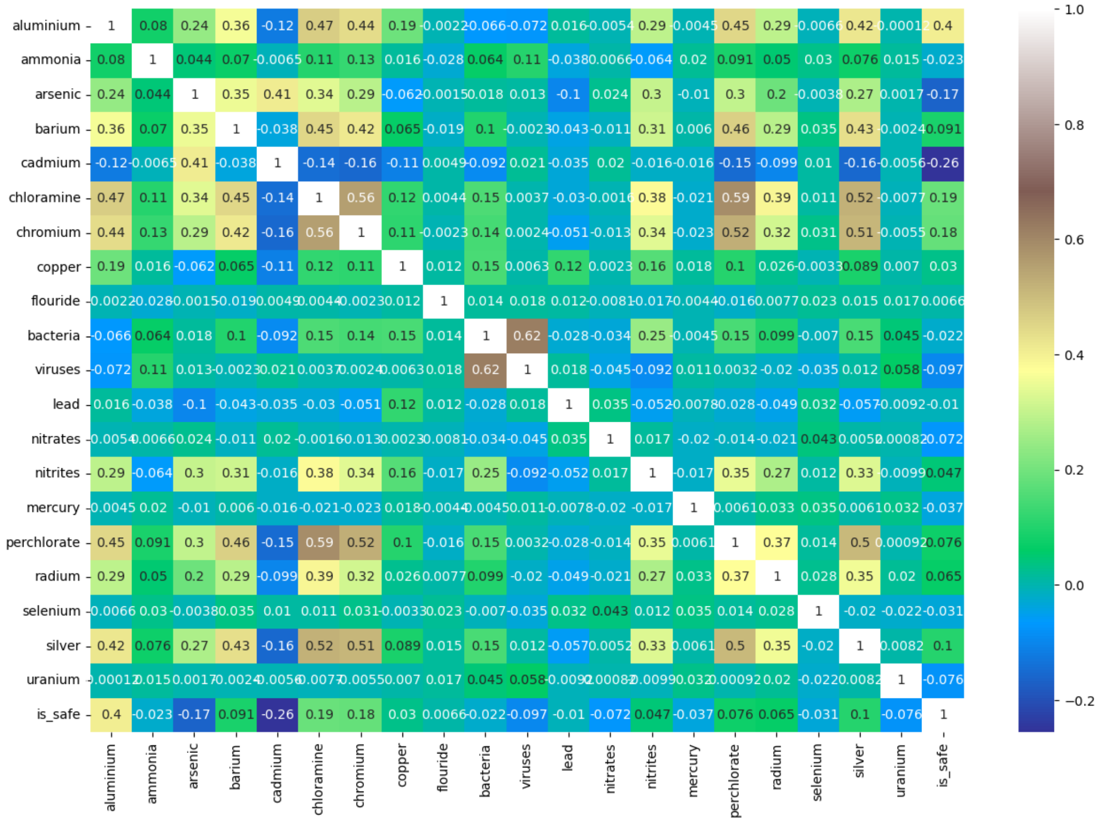

Figure 2: The correlation between every parameter used in the data set

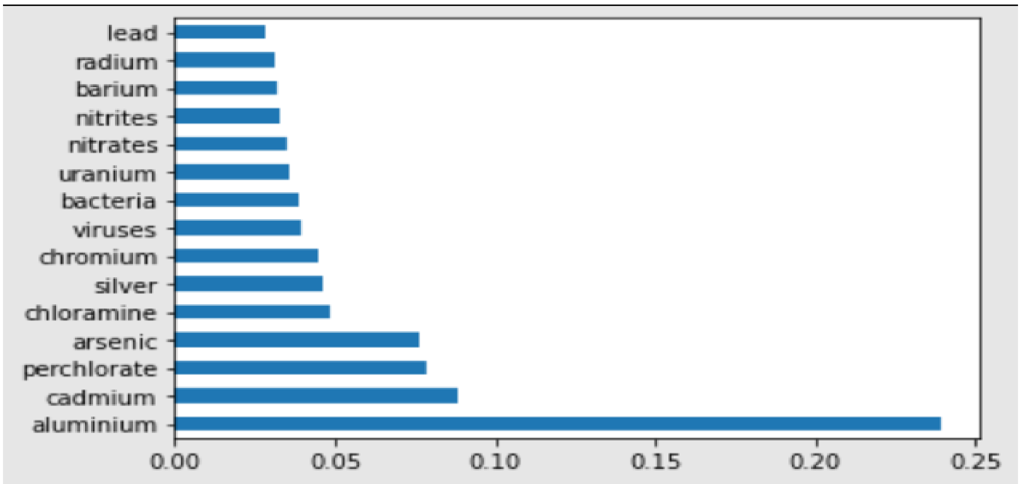

Figure 3: Extra Tree classier for identifying the best parameters from the data set

Table 2: Parameters with high relation to the other parameters from the data set

Parameters	Scores
Aluminum	2139.32
Perchlorate	868.43
Nitrates	840.12
Chloramine	121.83
Radium	78.54
Barium	63.28
Chromium	61.23
Virus	48.43
Cadmium	32.56

Table 3: Parameters with range values:

Parameter	Range
Ammonia	dangerous-if-greater-than-32.5
Arsenic	dangerous-if-greater-than-0.01
Barium	dangerous-if-greater-than-2
cadmium	dangerous-if-greater-than-0.005
Chloramine	dangerous-if-greater-than-4
Chromium	dangerous-if-greater-than-0.1
Copper	dangerous-if-greater-than-1.3
fluoride	dangerous-if-greater-than-1.5
Bacteria	dangerous-if-greater-than-0
Viruses	dangerous-if-greater-than-0
Lead	dangerous-if-greater-than-0.015
Nitrates	dangerous-if-greater-than-10
Nitrites	dangerous-if-greater-than-1
Mercury	dangerous-if-greater-than-0.002
Perchlorate	dangerous-if-greater-than-56
Radium	dangerous-if-greater-than-5
Selenium	dangerous-if-greater-than-0.5
Silver	dangerous-if-greater-than-0.1
Uranium	dangerous-if-greater-than-0.3
is safe	class-attribute-{0---not-safe,-1---safe}

5. RESULTS AND EVALUATION

All presently understood living forms are dependent on water, which has a wide range of uses. Therefore, everyone should care about the purity of the water. A group of physical, molecular, biological, and bacteriological traits can be used to describe the condition of water. These criteria enable water to be categorized into various groups, enabling its use for a particular reason.

The relationships between outside factors and how they affect various aspects of water purity are established in this chapter. There are various Water Quality Index (WQI) algorithms that can be used to provide data on water purity. Several correlation factors have been created in order to research the relationships between the various water quality metrics. We have used Pearson and Spearman relationships to create a cohesive statistical strategy. We used a set of data from our prior study, comprised of 13 factors recorded for water samples, to illustrate how WQI can be computed and analyzed.

The acidity and alkalinity of groundwater, as well as the value produced by the acidic-basic interplay of a number of its mineral and biological components, are significant indicators of pH. The existence of iron and manganese natural metallic ions, soil and organic materials, algae, plants, and industrial pollutants can all give water its color.

The connection between two factors is known as correlation. When the value of one measure rises or falls in direct proportion to the value of the other, this is known as a direct connection. When one parameter increases in response to another, the correlation is said to be positive, and when one parameter increases in response to another, the correlation is said to be negative. The number of the association coefficient (r) ranges from +1 to -1. The factors' association is described as strong when it is between 0.8 and 1.0, average when it is between 0.5 and 0.8, and feeble when it is between 0.0 and 0.5 [14]. A association coefficient (r) was computed between a number of water quality metrics, and the results are displayed in Figure 2.

The experiment is conduction the water sample collected and the data parameters has been evaluated and stored as a csv file. The data now has 21 columns with 8000 rows. The process has been evaluated using the python latest version 10.0 and used required libraries for the process.

The data is balanced with the target column, independent features from the data set has been extracted and the target feature has been divided and stored into different variables. The chi Square model is used to identify the efficient features, Extra tree classifier for then correlation rate of the features with other features. The training and testing is applied on the features and target columns with XGBoost algorithm which evaluates the water quality and recommends on adding the extra minerals to improve the quality of the water. This might has an issues if the added component might can increase or react with the other features of the water as recommend component should be with the permissible range.

The feature correlation is important as the solvents added will impact with other features, so correlation model will be used in order to handle the bottle necks of the solvents.

The model:

XGBClassifier(base-score=None,-booster=None,-callbacks=None,-colsample-bylevel=None,-colsample-bynode=None,-colsample-bytree=None,-early-stopping-rounds=None,-enable-categorical=False,-eval-metric=None,-feature-types=None,-gamma=0,-gpu-id=None,-grow-policy=None,-importance-type=None,-interaction-constraints=None,-learning-rate=None,-max-bin=None,-threshold=None,-max-cat-to-onehot=None-max-delta-step=None,-max-depth=None,-max-leaves=None,-min-child-weight=None, missing=nan,-monotone-constraints=None,-n-estimators=100,-n-jobs=None,-num-parallel-tree=None,-predictor=None, random-state=None)

The experiment is conduct with sample data

```
[ ]  actual = [2.8,0.01,2,0.005,4,0.1,1.3,1.5,0,0,0.015,10,1,0.002,56,5,0.5,0.1,0.3]
```

Figure 4: The sample data for testing the water purity

```
water is unsafe
Please decrease value of aluminium from : 3.0 to : 0.20000000000000018
Please decrease value of arsenic from : 0.05 to : 0.04
Please decrease value of barium from : 4 to : 2.0
Please decrease value of cadmium from : 0.01 to : 0.005
Please decrease value of chloramine from : 6 to : 2.0
Please decrease value of chromium from : 0.5 to : 0.4
Please decrease value of copper from : 2.0 to : 0.7
Please decrease value of flouride from : 2.0 to : 0.5
Please decrease value of bacteria from : 0.01 to : 0.01
Please decrease value of viruses from : 0.01 to : 0.01
Please decrease value of lead from : 0.02 to : 0.005000000000000001
Please decrease value of nitrates from : 15 to : 5.0
Please decrease value of nitrites from : 2 to : 1.0
Please decrease value of mercury from : 0.005 to : 0.003
Please decrease value of perchlorate from : 60 to : 4.0
Please decrease value of radium from : 10 to : 5.0
Please increase value of selenium from : 0.1 to : -0.4
Please decrease value of silver from : 0.5 to : 0.4
Please decrease value of uranium from : 0.5 to : 0.2
```

Figure 5: The recommended features to be managed to make the water pure.

Table 3 : Parameters with range values:

Parameter	Range
Aluminium	3.0 to: 0.20000000000000018
Arsenic	0.05 to: 0.04
Barium	4 to: 2.0
cadmium	0.01 to: 0.005
Chloramine	6 to: 2.0
Chromium	0.5 to: 0.4
Copper	2.0 to: 0.7
fluoride	2.0 to: 0.5
Bacteria	0.01 to: 0.01
Viruses	0.01 to: 0.01
Lead	0.02 to : 0.005000000000000001
Nitrates	15 to : 5.0
Nitrites	2 to : 1.0
Mercury	0.005 to : 0.003
Perchlorate	60 to : 4.0
Radium	10 to : 5.0
Selenium	0.1 to : -0.4
Silver	0.5 to : 0.4
Uranium	0.5 to : 0.2

The Parameters are described as increasing or decreasing in order to make the water pure enough to drink. The whole model works on the principle of correlation of the features of the data set. The model seems to be accurate and efficient in recommending the values for the features.

6 . CONCLUSION

The mean, standard deviation, and coefficient of variance were obtained from the statistical analysis of the empirically determined water quality metrics on water samples. Correlation values were computed because they reveal how different characteristics relate to one another. The findings of the correlation study demonstrate a significant and favorable association between the characteristics of nitrites, chloramine, chromium, and copper. The nitrites, chloramine, chromium, and copper of water are essential drinking water parameters because they are strongly associated with eight out of seventeen parameters in the research region, it can be inferred from the above talk.

References

1 Krishnan, K.S.D. Multiple Linear Regression-Based Water Quality Parameter Modeling to Detect Hexavalent Chromium in Drinking Water. In Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 22–24 March 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 2434–2439.

2 Ighalo JO, Adeniyi AG (2020) A comprehensive review of water quality monitoring and assessment in Nigeria. Chemosphere 260:127569

3 Bhoi SK, Mallick C, Mohanty CR (2022) Estimating the Water Quality Class of a Major Irrigation Canal in Odisha, India: ASupervised Machine Learning Approach. Nat Environ Pollut. Technol 21(2):433–446

4 Prasad B.C. and Narayana T.S., Nat. Environ. Poll. Tech., 3, 10 (2004)

5 Garg A., Kripalani C. and Brighu U., Nat. Environ., Poll. Tech., 3, 1(2004) 73.

6 Ahmed AN, Othman FB, Afan HA, Ibrahim RK, Fai CM, Hossain MS, Elshafie A (2019) Machine learning methods for better water quality prediction. J Hydrol 578:1–18

7 Pant, R.R.Spatio temporal variations of hydrogeo chemistry and its controlling factors in the Gandakiriver basin, Central Himalaya Nepal. Sci. Total. Environ. 2018, 622–623, 770–782. [CrossRef]

8. Khatunm, M. Phytoplankton assemblage with relation to water quality in Turag River of Bangladesh. Casp. J. Environ. Sci. 2020, 18, 31–45.

9. Rose-Rodrígue, R. Water and fertilizers use efficiency in two hydroponic systems for tomato production. Hortic. Bras. 2020, 38, 47–52. [CrossRef]

10. Trombadore, O. Effective Data Convergence, Mapping, and Pollution Categorization of Ghats at Ganga River Front in Varanasi; Springer: Berlin/Heidelberg, Germany, 2020; Volume 27, pp. 15912–15924.

11. King, K.W.; Williams, M.R.; Macrae, M.L.; Fausey, N.R.; Frankenberger, J.; Smith, D.R.; Kleinman, P.J.A.; Brown, L.C. Phosphorus Transport in Agricultural Subsurface Drainage: A Review. J. Environ. Qual. 2015, 44, 467–485. [CrossRef]

12. Hupfer, M.; Lewandowski, J. Oxygen Controls the Phosphorus Release from Lake Sediments—A Long-Lasting Paradigm in Limnology. Int. Rev. Hydrobiol. 2008, 93, 415–432. [CrossRef]

13. Tammeorg, O.; Möls, T.; Niemistö, J.; Holmroos, H.; Horppila, J. The actual role of oxygen deficit in the linkage of the water quality and benthic phosphorus release: Potential implications for lake restoration. Sci. Total Environ. 2017, 599–600, 732–738. [CrossRef]

14. Søndergaard, M.; Jensen, J.P.; Jeppesen, E. Role of sediment and internal loading of phosphorus in shallow lakes. Hydrobiologia 2003, 506–509, 135–145. [CrossRef]

15. Cooke, G.D.; Welch, E.B.; Peterson, S.A.; Nichols, S.A. Restoration and Management of Lakes and Reservoirs, 3rd ed.; Taylor & Francis/CRC Press: Boca Raton, FL, USA, 2005; p. 591.

16. Klapper, H. Control of Eutrophication in Inland Waters; Ellis Hornwood: New York, NY, USA, 1991; pp. 116–266.

17. Lossow, K.; Gawro ´nska, H. Lakes—Review of restoration methods. Munic. Rev. 2000, 9, 91–106.

18. Kuha, J.K.; Palomäki, A.H.; Keskinen, J.T.; Karjalainen, J.S. Negligible effect of hypolimnetic oxygenation on the trophic state of lake Jyväsjärvi, Finland. Limnologica 2016, 58, 1–6. [CrossRef]