IRAQI Academic Scientific Journals
العراقية المجلات الأكاديمية العلمية
Al-Kut University College Journal
ISSN (e) : 2616 - 7808 II ISSN (P) : 2414 - 7419
www.kutcollegejournal.alkutcollege.edu.iq
k.u.c.j.sci@alkutcollege.edu.iq
Vol. 9 , Issue 2 , December 2024

Application of Artificial Intelligence for Water Quality Evaluation

Hani Kadhim Al-Dhalemi [1] , Mohamad Abou Taam [2] , Mahmoud Koabaz [3]

Abstract

In response to the deterioration in water quality, several governments have started extensive mitigation programs essential stop control regular programs to protect. As a result, the classification of done using the water quality index (WQI). One newly created method for categorizing water quality is soft computing. Consequently, the goal of this project is to create a machine learning system for categorizing water quality. Both in the field and in the lab, measurements of (NH3-N) were made. The concentrations of these six parameters were of the DOE-WQI (Water Quality Index Department of Environment) technique.

Keywords: Artificial Intelligence, Machine Learning, Water Quality Evaluation, Model Classification

Affiliations of Authors

[1, 2, 3] Science Department, American University of Culture & Education, Lebanon, Beirut, 14/5840 Mazraa Beirut, 11052070

[1] hka032@auceonline.com
[2] maboutaam@auce.edu.lb
[3] mahmoudkoabaz@auce.edu.lb

[3] **Corresponding Author**

Paper Info.
Published: Dec. 2024

تطبيق الذكاء الاصطناعي لتقييم جودة المياه

أ. د. محمد أبو طعام [1] ، أ. د. محمود قعباز [2] ، هاني كاظم ظالمي [3]

المستخلص

استجابةً للتدهور في جودة المياه، بدأت العديد من الحكومات برامج تخفيف واسعة النطاق ضرورية لوقف البرامج المنتظمة للتحكم في الحماية. ونتيجة لذلك، تم تصنيفها باستخدام مؤشر جودة المياه (WQI). إحدى الطرق التي تم إنشاؤها حديثًا لتصنيف جودة المياه هي الحوسبة الناعمة. وبالتالي، فإن الهدف من هذا المشروع هو إنشاء نظام تعلم آلي لتصنيف جودة المياه. تم إجراء قياسات (NH3-N) سواء في الميدان أو في المختبر. كانت تركيزات هذه المعلمات الستة من تقنية DOE-WQI (مؤشر جودة المياه المعتمد في الاقسام البيئة) .

الكلمات المفتاحية: الذكاء الاصطناعي، التعلم الالي، تقييم جودة المياه، نماذج التصنيف

انتساب الباحثين

[1، 2، 3] قسم العلوم ، الجامعة الأمريكية للثقافة والتعليم، لبنان، بيروت، 5840/14 مزرعة بيروت 11052070

[1] hka032@auceonline.com
[2] maboutaam@auce.edu.lb
[3] mahmoudkoabaz@auce.edu.lb

[3] **المؤلف المراسل**

معلومات البحث
تأريخ النشر: كانون الاول 2024

Introduction

Water elements required the survival of known life, including that of humans. Water of a high standard is necessary for all living things to survive. The amount of pollution that aquatic life can tolerate has a limit. The survival of these organisms is in danger if these restrictions are not followed. Water has properties of both a gas and a liquid in its supercritical form, including low density and high viscosity. The following table (table 1) lists a few attributes of water:

Table (1) Some Water Characteristics

	Normal water	Subcritical water		Supercritical water	
Temperature (°C)	25	250	350	400	400

Pressure (MPa)	0.1	5	25	25	50
Density, p (g cm^3)	1	0.8	0.6	0.17	0.58
Dielectric constant, ε (F m^{21})	78.5	27.1	14.07	5.9	10.5
Ionic product, pKw	14	11.2	12	19.4	11.9
Heat capacity, C, (KJ Kg K^1)	4.22	4.86	10.1	13	6.8
Dynamic viscosity, η (mPa s)	0.89	0.11	0.064	0.03	0.07

Most of the rivers, lakes, and streams in the world meet specified quality requirements. Additionally, there are requirements for the caliber of water in certain circumstances. For example, irrigation water shouldn't be toxic to soil microbes or be so salty that it kills plants. Both conditions could lead to ecological collapse. The type of water utilized in various industrial processes has varied requirements. Natural water resources include ground and surface water, two of the most affordable sources of fresh water. However, these supplies can get contaminated through industrial, commercial, and even natural processes.

So, a worrying reduction in water quality has been a result of fast development. Lack of public awareness and unclean infrastructure are other major factors that have an impact on drinking water quality [1]. The implications of contaminated water on ecosystems, infrastructure, and human health are severe enough to require significant consideration. A UN report estimates that 1.5 million people each year pass away from ailments brought on by tainted water. In underdeveloped nations, 80% of health problems are formally attributed to water contamination. There are apparently 2.5 billion incidences of sickness and five million fatalities each year [2]. Such a death toll exceeds the sum of deaths from terrorism, crime, and natural disasters [3].

To analyze and, ideally, predict water quality (WQ), it is critical to provide different approaches. When predicting the patterns of WQ, it is advised to take the temporal dimension into account in order specific combination of models, as opposed to one model alone, produces better outcomes when used to predict the WQ [4-6]. There are numerous methods that can be used to forecast and model the WQ. These methods include data analysis, statistical procedures, graphical models, and prediction algorithms. To ascertain association between, multivariate statistical approaches have been applied [7]. Transition probabilities, multivariate interpolation, and other geostatistical analyses were performed using do regression analysis [5] and.

Massive population growth, all had negative consequences on WQ ecosystems [8, 9]. Having models that can forecast the WQ is essential for monitoring water contamination.

To effectively plan for and control water consumption, high-quality water resource modeling is absolutely necessary. In the past, researchers would frequently collect water samples from monitoring stations to determine the status of the water. The findings are also delayed because this takes so long. Artificial intelligence (AI) techniques like logic have been used by researchers.

To effectively plan for and control water consumption, high-quality water resource modeling is absolutely necessary. In the past,

researchers would frequently collect water samples from monitoring stations to determine the status of the water. The findings are also delayed because this takes so long. Artificial intelligence (AI) techniques like

Prioritizing water quality is necessary to guarantee the long-term success of a diversion strategy. One must anticipate the patterns of variation in that quality system at a specific time. Planning and managing water consumption depend on being able to predict the quality of future water supplies. Developing plausible strategies to prevent and manage water contamination as well as forecasting future changes in water safety at various levels of pollution are both crucial stages toward a cleaner, safer water supply. In water diversion designs, it's crucial to establish the water's general uniformity. Large amounts of water must be moved to meet daily drinking water needs. As a result, techniques for forecasting water quality in contemporary society ought to be investigated [10].

Money must be redirected to remedy problems with water distribution, just like with any infrastructure, which can be troublesome if the water is of low quality. Better water treatment and water quality management have become more necessary as a result to ensure that everyone has access to affordable, potable water. To get over these challenges, it will be necessary to conduct methodical analyses of raw water, disposal systems, and organizational monitoring concerns [11]. The ability to predict changes in water quality with accuracy is crucial for the success of aquaculture. Pre-processing water quality data is a common first step in the method of water quality measure prediction. This composition comprises two phases as a result. First, a correlation study of the various water quality indicators must be done.

This study examines how effectively several AI techniques, including. The development of ANN and SVM involved numerous iterations of transfer functions and kernels. A comparison of the performance of ANN and SVM findings reveals that both models can anticipate the various components of water quality. Tansig and RBF were found to perform well when compared to other transfer and kernel functions while developing ANN and SVM. Despite having respectable outcomes compared to other models, the GMDH model, when attempting to predict the many elements of water quality, only misses the precision of ANN and SVM. SVM was demonstrated to be the most accurate model among all those examined using error indices. The results of the models' analysis showed that they shared a common tendency toward overestimation. When employing the DDR index, it was found that the SVM model performed better than the other models.

Can Decomposition Methods Improve Soft Computing Models in Every Case? Predicting the Florida St. Johns River's Dissolved Oxygen Concentration [12]

In this study, stand-alone and hybrid soft computing models are compared for forecasting dissolved oxygen (DO) concentration using several water qualities measures. First, two different soft computing models—the multilayer perceptron (MLP) neural network and the cascade correlation neural network (CCNN)—were used to forecast the DO concentration in the St. Johns River, Florida, USA. By defining six combinations of input parameters, such as the DO concentration and water quality metrics chloride (Cl), nitrogen oxides (NOx), and total dissolved solids

Research materials:

The parameters employed, the number of contaminants present, and the defined boundary conditions all affect how effective an early warning system is. The system's reliability is most heavily influenced by the model's parameters. Some data suggests incorporating parameter optimization technology can increase the proposed system's accuracy. This frequently leads to better results with less time and effort required.

A water quality model's parameter estimation may be challenging and imprecise, especially if the model has a high degree of dimension. There are two different methods for estimating parameters:

i. Automated calibration and.
ii. using trial and error.

When a water quality model has a lot of parameters, automatic calibration is employed; when a model has few parameters, trial-and-error is utilized; this is largely based on the skill of the modeler and reduces uncertainty in model predictions [13].

Many measures the biological component of this inquiry will be the instead of the physical, it is the chemical processes that cause pollution. A controlled laboratory environment is used to assess some of these features, whereas field meters are used to measure others directly in the field [14].

Commercially accessible sensor electrodes for monitoring devices enable in-field monitoring simply for a few o**f these measures. The objective** of this study is to incorporate these variables that generate. Some of the standards utilized in this study to evaluate the water quality are as follows:

Oxygen Dissolved (DO):

This is how many oxygen molecules there are in every liter of water. The standard unit of measurement for oxygen content in water is milligrams per liter. With this knowledge, it is possible to determine whether retaining all of the formation have an impact on the molecule concentrations in a body of water. numerous factors, including air, water temperature, and wind mixing, the amount of dissolved oxygen in a body of water, as well as photosynthetic activity. Other significant include lake and other surface waters. Salinity, temperature, and other variables all have an impact on the oxygen content of saltwater. The main contributors to oxygen deprivation are the respiration of the hypolimnion accumulation compounds, other processes occur. Fish populations and their capacity for growth are significantly impacted by this. chosen to utilize to more accurately classify water. This makes the correlations between threshold and ambient temperature easier to see. Instead, it makes the defined limit reliant of the environment. as a function of expressed as a percentage (%).

Temperature:

The temperature of an area or object determines how hot or cold it is. This is one of the most crucial things to look at if you want to understand how the thermodynamics of the lake operate. The units of measurement for temperature are degrees Celsius and degrees Fahrenheit. The surface water temperature directly affects molecular mobility, dissolved oxygen saturation, and fluid dynamics. Numerous studies have found that the relationship between temperature and the amount of dissolved oxygen is inverse., which 13 has a decrease in oxygen solubility due to a rise in temperature. By

doing this, the lake's dissolved oxygen content is controlled. The observed alterations may be affected by factors such as sunlight exposure, turbidity, groundwater inputs, and ambient air temperature. Lake dynamics and temperature changes have a significant impact on the biological and chemical processes that take place there. Because of the increased activity of pollutants, more pollution incidents take place in the hotter summer months. It may have a substantial effect on the aquatic life of a lake if the temperature climbs by 10 degrees Celsius. All through the day and year, surface water naturally experiences temperature changes.

Ph:

The pH of a solution shows how concentrated the hydrogen ions are. It indicates whether the water is alkaline or acidic. Water is considered acidic if the number is less than 7, and alkaline if the number is greater than 7. According to the collected data, algae blooms typically begin in the spring when the pH is high. These conditions feature high quantities of dissolved oxygen and a somewhat neutral pH because photosynthesizing algae emit oxygen and utilize carbon dioxide as fuel. There is virtually any vertical variation in the pH of the water. When high pH depth profiles are assessed, in the majority of situations, It's a transient phenomenon. Growing algal blooms result in high pH and supersaturated dissolved oxygen because they consume carbon dioxide and produce oxygen [30]. More dissolved inorganic carbon is taken up by photosynthesis during the summer, which raises pH. The result of water evaporation is comparable. The fact that the locals regularly wash their plates and other kitchenware in detergents could be one reason for the lake's highly alkaline ph.

Phosphate:

Phosphate, also referred to as orthophosphate, is the form of phosphorus that is the easiest to obtain. In terms of concentrations, mg/L is used. Only when it is present in extremely high concentrations in an aquatic environment does it represent a concern to human health can be divided into three major groups: Phosphate endangers the floating aquatic life in surface water. Water quality indicators that we objectively examined in this analysis before using them in the rest of the work. Which parameters to use depends on several factors, including expert judgment

Indicators of water quality and allowable limits:

Effective pollution control of aquatic life needs consideration of basic environmental factors including water's dissolved oxygen content. Various governments have devised several techniques to track and assess pollution indicators and their effects. The Water Framework Directive (WFD) also lists 17 water quality factors that should be observed to determine the state of surface water. The biological, hydro morphological, and physio-chemical features of a system are some of these components. The physio-chemical components will be the main topic of this inquiry. Controlling the great majority of material pollutants is the main objective. Physio-chemical variables such as temperature, dissolved. As Seen in Table (2).

Table (2) WFD quality parameter contamination limits

Water Quality Parameter	Permissible limit for surface waters
PH	6.5-9.0
Temperature	Increase of 10°C affects aquatic life
Dissolved oxygen	Fresh water: 7mg/l 9mg/l Early life fishes: 9.5 mg/l in cold water and 6.0 mg/l in warm water
Ammonium	Greater than 0.1mg/l and less than 1mg/l
Nitrate	50mg NO 3/1
Nitrite	0.10 mg NO 2/1
Turbidity	Based on dissolved solids
Phosphates	Less than 50µg/l at entry point and less than 25µg/l within the lake
Dissolved organic carbon	10mg/l threshold
Conductivity	Fresh water streams 150µS/cm to 500 µS/cm

Model XGBOOSTINGe

The objective technique combines a number classifier. Iterative education involves starting with a weak learner and progressing to an advanced student [15]. Both gradient boosting and XGBoost are based on the same ideas. The two systems differ greatly in the specifics of how they are put into practice. The performance of XGBoost can be improved by regularizing the trees in a variety of ways [16].

Let's say there is a set of inputs and outputs like (x1, y1), (x2, y2),..., (xn, yn). The tree ensemble approach predicts the results using K addative functions, each of which represents a CART. The anticipated result is provided by the Equation (1).

$$\hat{y}i = \sum_{k}^{k} fk(xi), fk \in F \qquad (1)$$

where the CART space is denoted by f in F.

In order to approximate the functions, given a set of parameters, the following regularized objective function [15] must be minimized by Equation (2).

$$Obj(0) = \sum_{i}^{n} l(\hat{y}i, yi) + \sum_{k}^{k} \Omega(fk) \qquad (2)$$

where the training loss function, represented by the first term l(yi, yi), assesses the discrepancy between the output that was anticipated and the output that was actually produced. can be used to measure the training loss function by Equation (3).

$$MSE = \sum_{i}^{n} (yi - \hat{y}i)^2 \qquad (3)$$

and Logistic Loss, which are both expressed in the following equation (4) [5].

$$Logistic\ Loss = \sum_{i}^{n} [yiln(1 + e - \hat{y}i) + (1 - yiln(1 + e\hat{y}i))] \qquad (4)$$

The regularization term, or second term, (fk), penalizes model complexity to prevent overfitting. The regularization term in XGBoost is provided by Equation (5) [17].

$$\hat{y}_i^0 = 0$$
$$\hat{y}_i^1 = f_1(x_i) = \hat{y}_i^0 + f_1(x_i)$$
$$\hat{y}_i^1 = f_1(x_i) + f_2(x_i) = \hat{y}_i^1 + f_2(x_i)$$
$$\hat{y}_i^t = \sum_{k=1}^{k=1} fk(xi) = \hat{y}it - 1 + ft(xi) \qquad (5)$$

It is the tree that is added at each stage that maximizes the objective function. One alternative for the objective function is Equation (6). [18]

$$obj^{(t)} = \sum_i l(Yi, l(Y_i, Y_i) + \sum_k^k \Omega(f_k)$$
$$= \sum_{i=1}^{n} l(Y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) \qquad (6)$$

The objective compressed used by, as thoroughly explained by Chen and Gastrin. The quality of the tree is then assessed using the score function. We must thus take action to stop overfitting even more. The shrinkage variable shrinks the feature weights to attain a specific learning rate, or. Additionally, by providing row and column subsampling, XGBoost aids in the management of bias and variation in Random Forest [15]. The figure (1) described in this manner.

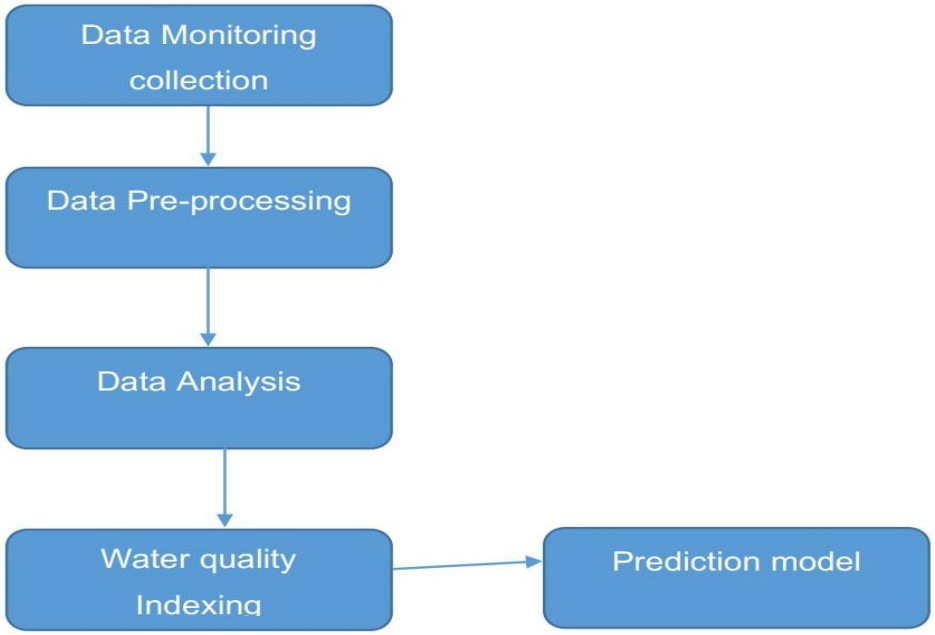

Figure (1): The diagram

From Kaggle has been incorporated into our data-monitoring system. Data preparation techniques include carefully examining to look for patterns and offer support for creating a goal to deliver indisputable be able to send out timely alerts regarding water contamination.

Evaluation metrics

Model testing and refinement are essential in machine learning. There are several measures we may use to evaluate the model's performance, and they will change depending on the challenge. We neglect regression metrics like mean squared error, mean absolute error, etc. in this thesis and instead

pay attention to supervised classification problems in the context of medical imaging.

Accuracy: The most important factor in sorting information is classification accuracy. It is an easy way to gauge how well the examined model classifies its forecasts. Equation (7).

$$\textbf{Accuracy} = \frac{\textbf{Number of correct predictions}}{\textbf{Total number of predictions}}$$

$$\textbf{Binary Accuracy} = \frac{\textbf{TP} + \textbf{TN}}{\textbf{TP} + \textbf{TN} + \textbf{FP} + \textbf{FN}} \qquad \textbf{(7)}$$

Precision is crucial in the realm of medicine since poor diagnoses can have serious effects for patients' lives. Doctors have the option of employing these tools as a second opinion when using data from models with high accuracy scores, but they shouldn't rely completely on them.

Forecasting Values

After establishing the numerous outcomes, a categorization can generate, we define the following metrics by Equation (8).

- Positive Predictive Value

$$\textbf{PPV} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FP}}$$

- Negative Predictive Value

$$\textbf{NPV} = \frac{\textbf{TN}}{\textbf{TN} + \textbf{FN}} \qquad \textbf{(8)}$$

Cross checking

Although the method is effective, it has a serious problem. We only consider combination using this approach. way the split distorted the overall picture. Given the data-splitting method's independence, we suggest using cross-validation to solve this issue. A particular kind of cross-validation that will be the subject of this paper is k-fold cross-validation. As Seen in Figure (2).

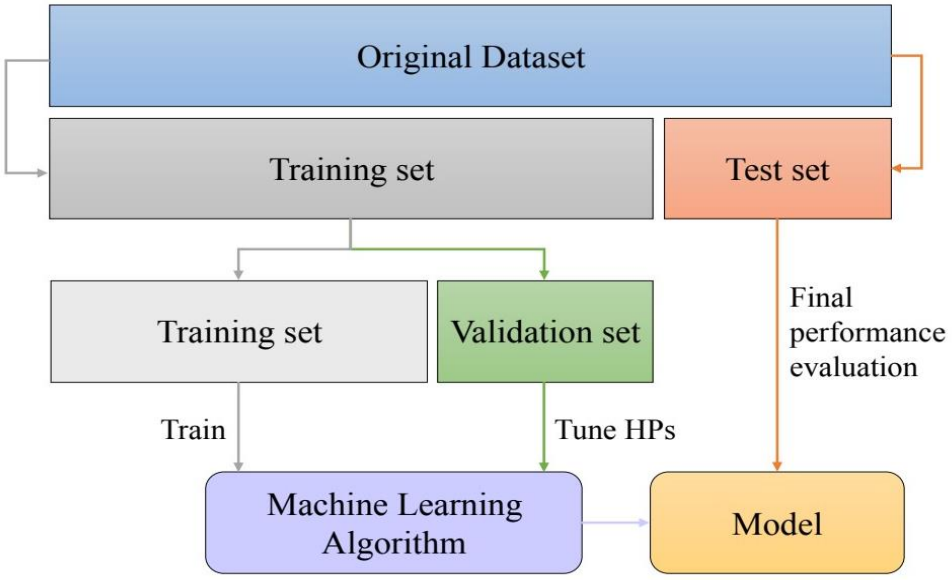

Figure (2): shows an illustration of the subsets' make-up as they are utilized to train and test a model

Our objective is to choose larger sum of .The training data are then divided into k roughly similar sections and given unique labels (F1, F2, etc.). After that, we continue training the model by

applying the hyper parameters as seen in equation (9).

k
(U ***Fi*** **)** ****
Fj **(9)**
i=1

Observe how effectively it functions on Fj. then averaged, for every conceivable combination of hyperparameters, this operation is repeated several times. Last but not least, the model tested on the test set had the best average cross-validation performance. A sample of data partitioning into k folds is shown in Figure (3). In order to do k-fold cross-validation, the training data is typically partitioned into k distinct groups, or "folds," at random. Major issues may arise from this, particularly if the data is wildly unrepresentative. This problem can be resolved by using the so-called stratified sampling approach since it appropriately depicts the distribution of the data within each strata. the use of stratification.

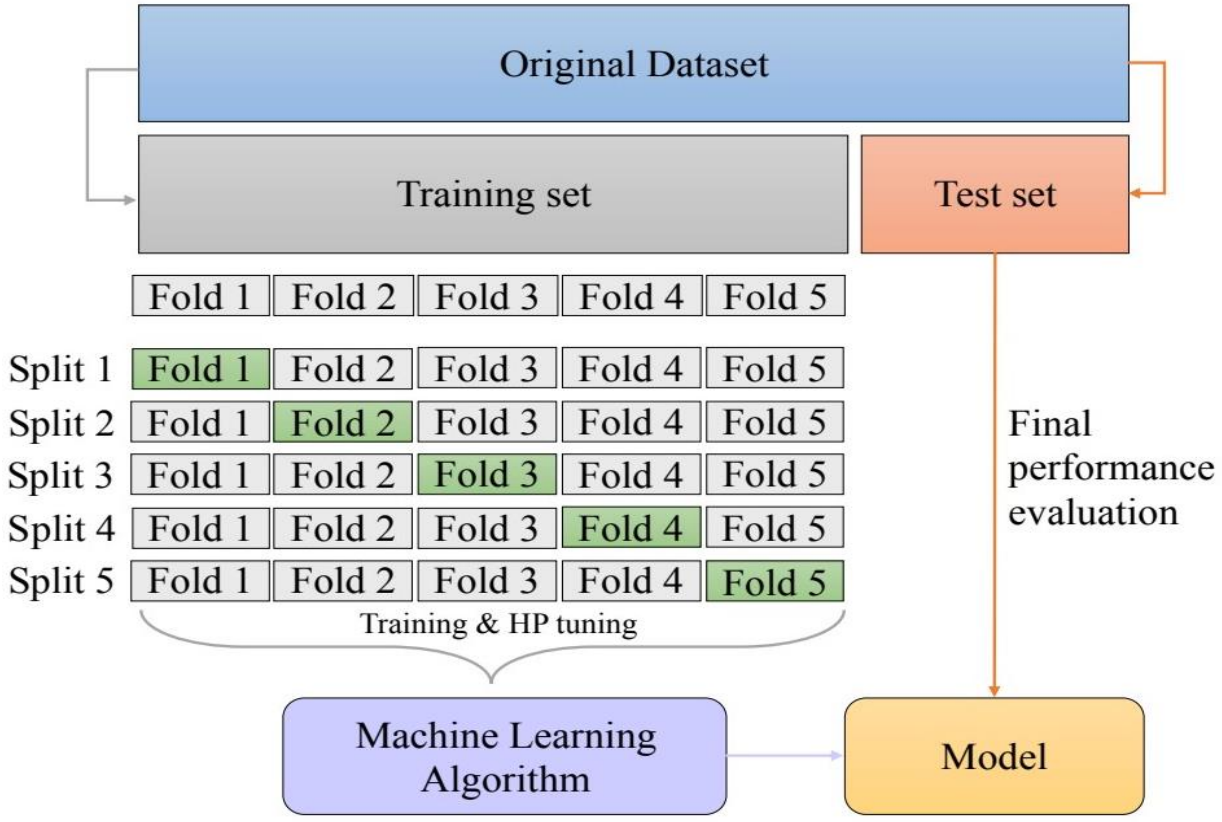

Figure (3): Shows a k-fold cross-validation example

Results

'Potable water' refers to drinking water, which is getting harder to get globally. Freshwater supplies all throughout the world are under stress due to increased use. The water you drink can become aesthetically unappealing or a health threat due to an unending array of impurities. Unless the water is unfiltered, microplastics containing adsorbed mercury species may still be present in treated drinking water.

1. PH value: The PH value is important when determining the acid-base balance of water. It

also indicates whether the water is acidic or alkaline. The WHO states that a pH range between 6.5 and 8.5 is the maximum permitted range. The ranges of the current investigation were between 6.52 to 6.83, which is acceptable by WHO standards.

2. Hardness: Salts high in calcium and magnesium are frequently to blame for hardness. These salts are released by the geological structures through which water travels. The amount of time that water is exposed to the material that causes hardness has an impact on its raw hardness.

3. Solids, or total dissolved solids (TDS): Both the water's color and flavor were affected by these minerals. This is a vital consideration when dealing with water. High TDS readings are a sign of mineral-rich water. The suggested range for TDS is 500 to 1000 mg/l, with 1000 mg/l being the maximum level that is acceptable for human consumption.

4. Chloramines: The two disinfectants most frequently employed in municipal water systems are chlorine and chloramine. Adding ammonia to chlorine to filter drinking water is the most frequent way that chloramines are created. Up to 4 mg/L (4 ppm) of chlorine is considered acceptable in drinking water.

5. Sulfate: Soil and rock contain sulfates, which are prevalent organic molecules in geological materials. They are absorbed into the air, groundwater, vegetation, and food. The main sulfate user is the chemical sector. Sulfate levels in saltwater are around 2,700 mg/L. Although certain regions have substantially greater amounts (up to 1000 mg/L), the typical concentration in freshwater ranges from 3 to 30 mg/L.

6. Conductivity: When it comes to electricity, pure water behaves more like an insulator than a conductor. Ion concentration increases the electrical conductivity of water. The concentration of dissolved solids often affects the electrical conductivity of water.

7. Organic carbon: The Total Organic Carbon (TOC) in source waters is derived from both naturally occurring organic matter (NOM) that has decomposed and manmade sources. The total organic carbon content (TOC) statistic measures the amount of organic carbon in water. The US Environmental Protection Agency's recommendations state that total organic carbon (TOC) levels in treated/drinking water should not exceed 2 mg/L and those in the source water utilized for treatment should not exceed 4 mg/L.

8. Trihalomethanes: Even after being chlorinated, water may still contain trihalomethanes. The amount of organic matter in the water, the amount of chlorine needed to treat the water, and the temperature of the treated water all have an impact on the concentration of THMs in drinking water.

9. The quantity of solid particles suspended in the water will define its turbidity. This test measures the light-emitting characteristics of water and assesses the colloidal matter quality of waste discharge. The Wando Genet Campus's average turbidity level (0.98 NTU) is below the 5.00 NTU WHO limit.

10. If the water has a potability score of 1, it is suitable for human consumption; if it has a score of 0, it is not.

The figure (4) and (5) shows the output and the descriptive statistics of the results.

Results:

	ph	Hardness	Solids	...	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	...	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	...	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	...	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	...	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	...	31.997993	4.075075	0

Figure (4): The outcome

	ph	Hardness	Solids	...	Trihalomethanes	Turbidity	Potability
count	2785.000000	3276.000000	3276.000000	...	3114.000000	3276.000000	3276.000000
mean	7.080795	196.369496	22014.092526	...	66.396293	3.966786	0.390110
std	1.594320	32.879761	8768.570828	...	16.175008	0.780382	0.487849
min	0.000000	47.432000	320.942611	...	0.738000	1.450000	0.000000
25%	6.093092	176.850538	15666.690297	...	55.844536	3.439711	0.000000
50%	7.036752	196.967627	20927.833607	...	66.622485	3.955028	0.000000
75%	8.062066	216.667456	27332.762127	...	77.337473	4.500320	1.000000
max	14.000000	323.124000	61227.196008	...	124.000000	6.739000	1.000000

Figure (5): shows the minimum, maximum, count, and standard deviation

Mean: If potability is 0.39, it signifies that there are more 0s in the data than 1s, indicating that the water is not potable.

The variance is known as the standard deviation (Std).

Min: what is each attribute's minimal value?

Maximum: these are the values for each attribute at their highest possible levels. 25%: this is significant since it shows how the data vary between the ranges of 0 and 25, as well as 50 and 75. The figure (6) shows the results of the potability.

To summarize the distribution of various composites in different types of water (alkaline, seawater, tap, bottled, distilled, and acidic), here are some key points:

Alkaline Water: Typically has higher pH levels (above 7) and contains minerals like calcium, potassium, and magnesium.

Seawater: Rich in salts, primarily sodium chloride, and contains various minerals and trace elements.

Tap Water: Composition varies by location but generally includes chlorine, fluoride, and trace amounts of metals and minerals.

Bottled Water: Can vary widely; some are mineral-rich, while others are purified and may contain added minerals for taste.

Distilled Water: Pure H2O with almost no dissolved minerals or impurities. Acidic Water: Lower pH levels (below 7), often containing higher levels of dissolved carbon dioxide and other acids.

The figure (7) shows the results of the Distribution of all composites. The figure (8) represents the output results of the used models.

The results are shown in a comparative table (Table 3), where three machine learning models—XGBoost, Rough Forest, and Progressive Boosting—are evaluated based on their performance percentages. XGBoost achieves an accuracy of 76.0%, while Rough Forest and Progressive Boosting both show slightly higher accuracies of 77.0%.

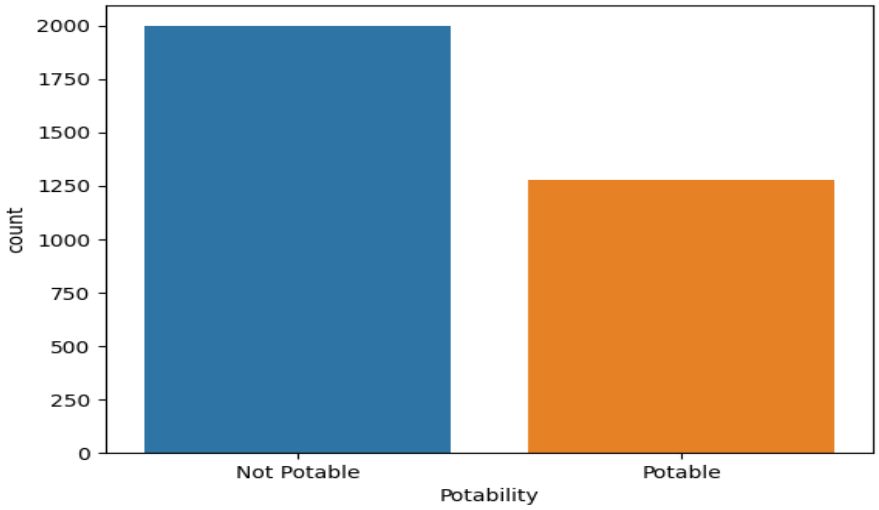

Figure (6): The potability of 2000 are non-potable and 1 250 are potable

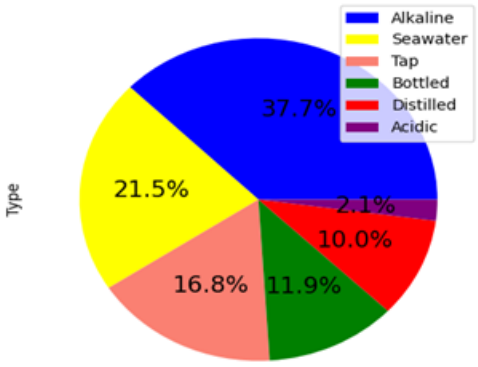

Figure (7): Distribution of all composites

	precision	recall	f1-score	support
0	0.82	0.81	0.81	429
1	0.65	0.66	0.65	227
accuracy			0.76	656
macro avg	0.73	0.73	0.73	656
weighted avg	0.76	0.76	0.76	656

Figure (8): Output result

Table 3. Comparison table,

Model XGBoost	Rough Forest	Progressive Boosting
76.0%	77.0%	77.0%

Conclusions

Overall, the goals of this study were met, and examples of the use were given, covering most aspects of the regular research activity in the field of artificial intelligence for positions in environmental sciences. This work also emphasizes how important it is to collaborate with starting because it frequently happens that data sets are unsuitable for the desired activities. In this thesis, various machine learning models for predicting the classification of water quality (WQC) were examined, which, according to the water quality score, is a unique class. as well as a number of water quality-related input parameters were used in the suggested technique. was shown to predict the WQC the most accurately among the three used methods. It had the best balanced accuracy and the smallest categorization error, accuracy, precision, focus, and f-measure. Classification models provide exceptionally accurate prediction models and are useful for assessing.

Reference

[1] Zeilhofer P., Zeilhofer L. V. A. C., Hardoim E. L., Lima Z. M. ., Oliveira C. S. GIS applications for mapping and spatial modeling of urban-use water quality: a case study in District of Cuiabá, Mato Grosso, Brazil. Cadernos de Saúde Pública. 2007;23(4):875–884. doi: 10.1590/S0102-311X2007000400015.

[2] Kahlown M. A., Tahir M. A., Rasheed H. National Water Quality Monitoring Programme, Fifth Monitoring Report (2005–2006) Islamabad, Pakistan: Pakistan Council of Research in Water Resources Islamabad; 2007. http://www.pcrwr.gov.pk/Publications/Water%20Quality%20Reports/Water%20Quality%20Monitoring%20Report%202005-06.pdf.

[3] UN water. Development; 2010. Clean water for a healthy world. https://www.undp.org/content/undp/en/home/presscenter/articles/2010/03/22/clean-water-for-a-healthy-world.html.

[4] Taskaya-Temizel T., Casey M. C. A comparative study of autoregressive neural network hybrids. Neural Networks. 2005;18(5–6):781–789. doi: 10.1016/j.neunet.2005.06.003.

[5] Babu C. N., Reddy B. E. A moving-average filter based hybrid ARIMA-ANN model for forecasting time series data. Applied Soft Computing. 2014;23:27–38. doi: 10.1016/j.asoc.2014.05.028.

[6] Zhang X., Hu N., Cheng Z., Zhong H. Vibration data recovery based on compressed sensing. Acta Physica Sinica. 2014;63(20):119–128. doi: 10.7498/aps.63.200506.

[7] Farrell-Poe K., Payne W., Emanuel R. Water Quality & Monitoring. University of Arizona Repository; 2000. http://hdl.handle.net/10150/146901.

[8] Cabral Pinto M. M. S., Ordens C. M., Condesso de Melo M. T., et al. An inter-disciplinary approach to evaluate human health risks due to long-term exposure to contaminated groundwater near a chemical complex. Exposure and Health. 2020;12(2):199–214. doi: 10.1007/s12403-019-00305-z.

[9] Cabral Pinto M. M. S., Marinho-Reis A. P., Almeida A., et al. Human predisposition to cognitive impairment and its relation with environmental exposure to potentially toxic elements. Environmental Geochemistry and Health. 2018;40(5):1767–1784. doi: 10.1007/s10653-017-9928-3.

[10] Zhou, J.; Wang, Y.; Xiao, F.; Sun, L. Water Quality Prediction Method Based on IGRA and LSTM. Water 2018, 10, 1148.

[11] Hu, Z.; Zhang, Y.; Zhao, Y.; Xie, M.; Zhong, J.; Tu, Z.; Liu, J. A Water Quality Prediction Method Based on the Deep LSTMNetwork Considering Correlation in Smart Mariculture. Sensors 2019, 19, 1420.

[12] Clark, R.; Hakim, S.; Ostfeld, A. Handbook of Water and Wastewater Systems Protection (Protecting Critical Infrastructure); Springer:New York, NY, USA, 2011.

[13]J. Park, J.H. Park, J.S. Choi, J.C. Joo, K. Park, H.C. Yoon, C.Y. Park, W.H. Lee, T.Y. Heo Ensemble model development for the prediction of a disaster index in water treatment systems Water, 12 (2020), Article 113195, 10.3390/w12113195

[14] Wang, Y., Zhang, W., Engel, B. A., Peng, H., Theller, L., Shi, Y., & Hu, S. (2017). Accurate early warning to water quality pollutant risk by mobile model system with optimization technology. Journal' of' Environment' management.

[15] Jerome H. Friedman. “Greedy Function Approximation: A Gradient Boosting Machine”. In: The Annals of Statistics 29.5 (Oct. 2001), pp. 1189–1232. ISSN: 00905364.

[17] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining. KDD ’16. ACM, Aug. 2016, pp. 785– 794. ISBN: 9781450342322.

[18] Bayram, A., Kankal, M., Tayfur, G., & Onsoy, H. (2014). Predictionof suspended sediment concentration from water quality variables. Neural' Computing' and'Applications,'24(5), 1079F1087.