Predicting the Water Potability Index Using Machine Learning

Article *in* Environment and Ecology Research · July 2023

DOI: 10.13189/eer.2023.110402

3 authors:

Ivan Ivanov
Sofia University "St. Kliment Ohridski"
121 PUBLICATIONS **1,028** CITATIONS
SEE PROFILE

Borislava Toleva
Sofia University "St. Kliment Ohridski"
24 PUBLICATIONS **168** CITATIONS
SEE PROFILE

George Taylor
Horizon Research Publishing(HRPUB)
572 PUBLICATIONS **1,381** CITATIONS
SEE PROFILE

Environment and Ecology Research 11(4): 537-542, 2023
DOI: 10.13189/eer.2023.110402 http://www.hrpub.org

Predicting the Water Potability Index Using Machine Learning

Ivan Ivanov, Borislava Toleva*

Faculty of Economics and Business Administration, Sofia University, Bulgaria

Received March 21, 2023; Revised June 9, 2023; Accepted July 19, 2023

Cite This Paper in the Following Citation Styles

(a): *[1] Ivan Ivanov, Borislava Toleva , "Predicting the Water Potability Index Using Machine Learning," Environment and Ecology Research, Vol. 11, No. 4, pp. 537- 542, 2023. DOI: 10.13189/eer.2023.110402.*

(b): *Ivan Ivanov, Borislava Toleva (2023). Predicting the Water Potability Index Using Machine Learning. Environment and Ecology Research, 11(4), 537 - 542. DOI: 10.13189/eer.2023.110402.*

Abstract Water potability is a key topic in ecology as it defines the areas where life can exist and the quality of health and food. Without potable water, vast regions can be unpopulated. Poor water quality affects the quality and quantity of food and the spread of diseases. There is a tendency that sources of potable water have started to deteriorate in recent years, so the topic of potable water quality has become central in environmental and ecology studies. A central question is prediction of water quality in various areas. This research proposes an improved machine learning algorithm for predicting the potability of water. The proposed algorithm is simple, and it is easier to apply compared to other existing algorithms. It can be applied to various datasets for water quality providing a quick insight into the question whether new water sources in the area are more likely to have potable water. It can also be applied to various datasets about water quality. Therefore, a quick review of the water and its quality in each region can be done using the proposed algorithm.

Keywords Water Potability, Machine Learning, Classification, Unbalanced Data

1. Introduction

Potable water is water that is fit for drinking without representing a hazard to one's health. The topic of potable water has been an extremely popular theme in recent research as people's health, agriculture and wildlife are dependent on the quantity of clean and potable water. As a result of global warming, the global balance in quality and quantity of potable water has become disturbed [1], [2]. Potable water supply is becoming limited in vast areas around the world, so alternative water sources need to be found [3-4]. Increased number of scientists warn against losses of potable water and deterioration of its quality. Therefore, many scientists have started to examine factors that affect the quantity and quality of potable water [5], [32], how loss of potable water can be prevented [6] and how to tackle drought [7].

Machine learning (ML) models have been used in all mentioned areas of water research, particularly in predicting water quality. For instance, Zhu and Wang [8] compare the performance of forty-five machine learning models to assess the quality of potable water. They [8] provide a review of existing academic research about water quality. They classify ML models into five groups – for surface water, groundwater, drinking water, wastewater and marine water. According to their classification, the Partial least squares (PLS) regression, Support vector regression (SVR), and Deep neural network (DNN) work best when applied to surface water quality. However, when surface waters are polluted, more sophisticated models are necessary [8], [9]. Long short-term memory (LSTM) networks [10] and Bootstrapped wavelet neural networks (BWNN) [11] can be used in time series water-quality data that are fluctuating and lack seasonality. Artificial Neural Networks (ANN) [12] and Support vector machines (SVM) [13] are good to model water quality components.

Groundwater is another source of potable water, for which ML models exist. Zhu and Wang [8] review several types of ML models used to assess the quality of groundwater. Their conclusion is that the best results can be achieved using the decision tree classifier (DT) [14].

The principal components analysis (PCA) and multivariate statistical analysis [15] provide the best results in identifying the source of groundwater pollution. ML models can also be applied to examining the quality of drinking water, where the focus is put in many directions, e.g., water loss through pipes, reuse of drinking water, etc. The ANN and SVM models [16] have the advantage of short computing time, which makes them suitable for monitoring drinking water quality in real time. The SVM model is robust to noise, while the ANN is extremely sensitive to noisy data. Wastewater quality can be assessed using various ML models [17-20]. Machine learning models are also used to monitor the pollutants of marine water [21-25]. ML models have become an integral part of water quality research as Zhu and Wang [8] and Zawawi et al. [33] show. Therefore, they are constantly improved [10-25] to predict the quality of water in each region better.

In this context, the aim of this research is to propose a simple but effective ML algorithm that can provide a quick insight into the potability of water in each region.

2. Materials and Methods

2.1. Dataset Description

The underlying data are publicly available in [26]. The dataset contains data about the Water Potability Index. The data have 3276 observations. The dataset consists of 10 variables, 9 independent and 1 target variables. The target variable is called 'Potability'. It describes whether the water is drinking (1) or not (label 0). Table 1 describes the columns of the underlying data for the model.

Table 1. Dataset Description [26]

Variable	Description
1. pH	pH of 1. water (0 to 14).
2. Hardness	Capacity of water to precipitate soap in mg/L.
3. Solids	Total dissolved solids in ppm.
4. Chloramines	Amount of Chloramines in ppm.
5. Sulfate	Amount of Sulfates dissolved in mg/L.
6. Conductivity	Electrical conductivity of water in μS/cm.
7. Organic_carbon	Amount of organic carbon in ppm.
8. Trihalomethanes	Amount of Trihalomethanes in μg/L.
9. Turbidity	Measure of light emitting property of water in NTU.
10. Potability	Indicates if water is safe for human consumption. Potable -1 and Not potable -0

The standard ML algorithm for classification requires steps 1, 3, 4, 5 and 6 [29]. It consists of handling missing data (step 1), standardizing independent variables (step 3), splitting data into training and test sets (step 4), performing classification (step 5) and assessing the model (step 6). However, when the labels in the dependent variable are imbalanced, these steps may not be enough to provide reliable results. We propose a novel step 2 that involves random shuffling of the dataset prior to standardization and splitting into training and test sets. We also modify step 5 of the standard algorithm by notifying each classification model that there is class imbalance. Next subsection clarifies the proposed algorithm.

2.2. The Proposed Algorithm

Step 1: Import the data in Python 3.7 and handle missing data. All missing data are imputed using the mean value of the respective variable. The variables called 'ph', 'Sulfate' and 'Trihalomethanes' have missing values. Therefore, step one is applied to them. Apply the command below to each of the three variables.

```
dataset["ph"].fillna(value=dataset["ph"].mean(),
inplace=True)
```

Step 2 (new): Shuffle the import data randomly to avoid overfitting using:

```
random.seed(24)
random.shuffle(dataset)
```

This step is an additional step to the classic ML algorithm [29], which applies before splitting into train/test sets. Standard ML algorithms require standardization of input independent variables and then train/test splitting [29]. However, this research introduces a new step by randomly shuffling the dataset before standardizing the data and defining train/test splits. This step is introduced to shuffle classes in the water potability dataset as they are imbalanced, thus having the prevalence of class 'nonpotable' over class 'potable' water. Random shuffling would allow for unbiased prediction that is not affected by the prevalence of class 'non-potable'.

Step 3: The next step is to standardize the input variables to remove the differences of the measurement unit that can also affect the prediction. Standardization is performed in Python using the following commands:

```
sc=StandardScaler()
X=sc.fit_transform(X)
```

Step 4: Split the data into training and test sets using X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.28, random_state=124), where the size of the test set is 28%, while the training set contains 72% of the shuffled dataset. The train/test split is done to avoid overfitting [29]. According to [31], the train_test_split function in Python can be used as an alternative to cross validation. Its advantage is that the dataset can be split into training and test sets using a predefined proportion as it is in the presented algorithm.

Step 5: Fit several classification models like SVM, DT and random forest (RF) and set the parameter class_weight to 'balanced' so that the class imbalance in the dataset can be handled. Setting the parameter class_weight to 'balanced' is also another step to remove the bias coming from class imbalance.

Run the three models by the commands below:

SVM: SVM = SVC(C=10, kernel='rbf', gamma='auto', class_weight='balanced')

DT: DT= DecisionTreeClassifier(class_weight='balanced')

RF: RF = RandomForestClassifier(class_weight='balanced')

Step 6: Assess the performance of the algorithms by exploring accuracy and classification metrics like precision, recall and f1-score [29]. Calculating precision, recall and f1-score consists of producing the elements of a confusion matrix [29]. Confusion matrix provides details about the number of correctly and wrongly predicted instances from the model. Together with accuracy, the confusion matrix is a measure of the model quality. The elements of the confusion matrix are true positive (TP), true negative (TN), false positive (FP), and false negative (FN) instances.

The formulas are the following:

Accuracy = (TP+TN)/(TP+TN+FP+FN);

Precision (Class 0) = TP/ (TP+FP) = Precision

Precision (Class 1) = TN/ (TN+FN);

Recall (Class 0) = TP / (TP+FN) = Sensitivity= Recall

Recall (Class 1) = TN / (TN+FP) = Specificity

The Sensitivity parameter provides information on the accuracy of only positive predictions. Specificity represents the accuracy only of negative predictions. Accuracy provides information about the overall performance of the model. More information about these metrics can be found in [29].

In summary, the proposed algorithm contains two novel steps – step 2 – random shuffling of the data before any transformations and step 5 – setting the parameter class_weight to 'balanced' to handle class imbalance in the water potability dataset. Applying these additional steps to the classic ML algorithm [29] would allow for a quick prediction of water quality in various datasets with class imbalance using a simple to follow algorithm. The simplicity of the proposed algorithm is its biggest advantage. Next section demonstrates the output from the proposed algorithm and comments on its prediction quality.

3. Results

3.1. Accuracy

Our modifications result in high accuracy as table 2 shows. The highest accuracy achieved is 0.88 by the DT modification. We also achieve accuracy of 0.83 using the SVC and 0.81 using the random forest. Table 2 also demonstrates the accuracy by another research on the same dataset [30]. In table 2 Mod 1 DT, Mod 2 SVC and Mod 3 RF present our modifications, while the remaining models can be found in [30].

Table 2. Model comparison: this research vs [30]

Model	Accuracy
Random Forest*	0.8
Gradient Boost*	0.76
Decision Tree*	0.73
Support Vector*	0.69
AdaBoost*	0.68
Support Vector*	0.67
KNeighbors*	0.65
BernoulliNB*	0.61
GaussianNB*	0.57
Passive aggressive*	0.54
Nearest centroid*	0.52
Logistic regression*	0.52
Ridge*	0.52
Stochastic gradient descent	0.51
Perceptron*	0.51
Gradient Boosting after parameters tuning*	0.81
Random Forest after parameters tuning*	0.81
Mod 1 DT (this research)	**0.88**
Mod 2 SVC (this research)	**0.83**
Mod 3 RF (this research)	**0.81**

Patel [30] presents an algorithm where independent variables are standardized, and a SMOTE technique is used to balance the classes. Then, they run several classification models without parameter adjustment and with parameter adjustment (table 2). When parameter adjustment is not applied, the accuracy in [30] is low, between 0.51 and 0.69. There are several models, however, that perform well despite lack of parameters adjustment – the Random Forest (0.8), the Gradient Boost (0.76) and the Decision Tree (0.81) as shown in table 2 [30]. In comparison, our models (0.81 – 0.88) perform better than Patel's unadjusted versions (0.51-0.69) [30] and have comparable accuracy (0.81, 0.83) to the unadjusted version of Patel's Random Forest and Gradient Boost (0.8, 0.76, 0.81). Therefore, the SMOTE technique used by Patel to handle class imbalance comes at the price of properly adjusting the model to get accurate predictions. This is both time consuming and can introduce bias through the parameter adjustment process.

Therefore, our modifications are simple and easy to apply as they include Python's built-in functions without additional steps and packages. Moreover, our algorithm achieves high accuracy using various classification models, allowing the focus to be on the data and the outcome of the classification task rather than finding the most appropriate classification model. Using our algorithm, we can gain quick insights into the potability of water in a particular region, we can predict the type of water (in other datasets) and any other indicator related to water and its quality.

Patel's [30] adjusted versions of the gradient boost and random forest models use further tuning of the parameters using cross validation. However, the performance of the random forest remains unchanged (0.80 unadjusted vs 0.81 adjusted version), while the gradient boost performance improved from 0.76 to 0.81. Their key finding is that using SMOTE to balance classes and adjusting the tuning parameters of the classification can improve the prediction of water potability. However, SMOTE is a technique to synthetically derive more observations of the minority class, so that a new balanced dataset can be produced. This makes the algorithm complicated. Our algorithm, on the other hand, handles class imbalance in a simple way using Python's built-in functions and resampling. We first resample data randomly, then split into train and test sets by using the shuffled dataset. Then, we run a classification model by setting the class weight parameter to balance. We use these steps to balance the dataset without introducing synthetical observations to the dataset. Our modifications work with the original water potability data. Therefore, we also avoid the bias that may come from inserting synthetical observations into the dataset. Our algorithm is simple, yet efficient as can be seen in table 2.

Our results also show that our algorithm provides satisfactory performance of the decision tree, random forest and support vector machines without additional tuning of the model. Our algorithm outperforms Patel's [30] SVC and random forest and is competitive to their decision tree. So, our algorithm can be applied as a fast approach to handling imbalanced classes in all water datasets where classification task is present. In the case of water potability, Mod 1 DT provides the best accuracy (0.88) among all experiments shown in table 2.

Unlike Patel, we do not preselect the most important features in the dataset. They [30] do that to remove noisy features and use only the key features that would boost the accuracy of the model. Removing noisy features would remove variables that would increase accuracy artificially [28]. However, we handle this issue by splitting into train and test sets. Also, the number of features is so small that identifying the most important features is not necessary. Table 2 shows that our approach is more effective than Patel's in terms of accuracy. As a result, we can predict better the potability of water.

3.2. Classification Metrics

Table 3 shows the classification metrics for our SVC and Patel's [30].

Table 3. Classification metrics of two SVC models.

Proposed SVC			
	precision	recall	f1-score
Not potable	0.88	0.86	0.87
Potable	0.74	0.77	0.76
	accuracy	0.83	
Patel's SVC [30]			
Not potable	0.74	0.71	0.77
Potable	0.78	0.8	0.73
	accuracy	0.75	

Table 3 demonstrates that the proposed SVC model performs better than Patel's version. Our SVC models have a comparable performance in terms of predicting the 'not potable' water, but it predicts better the label 'potable'. This can be seen from the precision and recall measures in table 3. As a result, the f1-score and accuracy of our SVC are higher than Patel's [30]. Similar findings can be observed in table 4.

Table 4. Comparison of proposed DT and Patel's

Proposed DT			
	precision	recall	f1-score
Not potable	0.92	0.91	0.91
Potable	1	0.80	0.81
	accuracy	0.88	
Patel's DT [30]			
Not potable	0.74	0.71	0.73
Potable	0.71	0.74	0.73
	accuracy	0.73	

The decision tree with SMOTE technique [30] demonstrates lower classification metrics than the proposed DT. As a result, the accuracy is also lower. Although our modification uses the original dataset without artificial observations to balance the classes, the proposed algorithm predicts better both potable and non-potable water. The accuracy of our model is also better – 0.83 compared to 0.73 in Patel's research. In addition, the accuracy of the proposed DT is the highest one compared to all modifications of Patel (table 3). This finding follows from the improved prediction ability of our DT version to classify correctly the two labels. Table 5 shows that the precision, recall and f1-score measures are significantly improved compared to [30].

Table 5. Comparison between proposed RF and Patel's FR

Proposed RF			
	precision	recall	f1-score
0	0.82	0.82	0.82
1	0.81	0.81	0.81
	accuracy	0.81	
Patel's RF [30]			
0	0.83	0.8	0.82
1	0.8	0.81	0.81
	accuracy	0.81	

Table 5 shows that the classification metrics of the proposed RF and Patel's are similar. The accuracy as well. The two algorithms predict correctly about 80% of the two classes. However, the version of RF we propose is much simpler to apply.

4. Discussion

In this research we present a simple but effective algorithm to perform fast and accurate prediction of potable water. We improve the accuracy and classification metrics of the decision tree, support vector machines and the random forest. We propose a novel approach to handling class imbalance and predicting better water potability by using Python's built-in functions. We can summarize the key findings from this research as follows:

1 Using Python's built-in functions can successfully tackle class imbalance for water potability. The parameter in question is 'class_weight'=balance.

2 When having class imbalance in water potability data, resampling the data randomly prior to train/test split can help avoid overfitting coming from class imbalance. This can be used as an alternative to other techniques for class imbalance like SMOTE [30].

3 Applying our proposed algorithm is a simple and fast way to predict labels with class imbalance in the water potability dataset. It does not require tuning of the parameters of the model, nor additional techniques to balance classes. It does not involve preselecting important variables, which saves time and makes the model easy for analysis. This makes it an effective algorithm for initial and further modelling of imbalanced water data of any kind.

A future extension of our research would be to do more experiments with other classification algorithms and water datasets to explore the possibility of presenting a universal fast and effective algorithm for analyzing water potability and other water related datasets. Further experiments can be conducted to explore the question if applying this algorithm to datasets with larger quantity of variables requires feature selection or its advantages remain unchanged in such an environment.

5. Conclusions

Despite the future possibilities for research, the proposed algorithm can be used as a fast alternative to predict water potability. The algorithm can be applied to all types of water datasets that require classification tasks. It can be used as a data explanatory approach to analyzing water potability data quickly at the early stage of the research.

Acknowledgements

This research was financed by Sofia University Research Programme, project number 80-10-11/11.04.2023.

REFERENCES

[1] Dawood T., Elwakil E., Novoa H., Delgado J., Toward urban sustainability and clean potable water: Prediction of water quality via artificial neural networks, Journal of Cleaner Production, Volume 291, pp. 125266, 2021, DOI: DOI: https://doi.org/10.1016/S0015-1882(11)70082-9.

[2] Bennett A., Potable water: New technology enables use of alternative water sources, Filtration + Separation, Volume 48, Issue 2, pp. 24-27, 2011. https://doi.org/10.1016/S0015-1882(11)70082-9.

[3] Lugo A., Chathurika G., Bandara L., Xu X., Almeida J., Abeysiriwardana-Arachchige I., Nirmalakhandan N., Xu P., Life cycle energy use and greenhouse gas emissions for a novel algal-osmosis membrane system versus conventional advanced potable water reuse processes: Part I, Journal of Environmental Management, Volume 331, 117293, 2023, DOI: https://doi.org/10.1016/j.jenvman.2023.117293.

[4] Warsinger D., Chakraborty S., Tow E., Plumlee M., Bellona Ch., Loutatidou S., Karimi L., Mikelonis A., Achilli A., Ghassemi A., Padhye L., Snyder S., Curcio S., Vecitis Ch., Arafat H., Lienhard J., A review of polymeric membranes and processes for potable water reuse, Progress in Polymer Science, Volume 81, pp. 209-237, 2018, DOI: https://doi.org/10.1016/j.progpolymsci.2018.01.004.

[5] Hartley K., Tortajada C., Biswas A., A formal model concerning policy strategies to build public acceptance of potable water reuse, Journal of Environmental Management, Volume 250, 109505, 2019, DOI: https://doi.org/10.1016/j.jenvman.2019.109505.

[6] Yu T., Chen X., Yan W., Xu Z., Ye M., Leak detection in water distribution systems by classifying vibration signals, Mechanical Systems and Signal Processing, Volume 185, pp. 109810, 2023, DOI: https://doi.org/10.1016/j.ymssp.2022.109810.

[7] Kent R., Chapter 2.5 - Flood and Drought Prevention and Disaster Mitigation: Combating Land Degradation with an Integrated Natural Systems Strategy, Editor(s): Ilan Chabay, Martin Frick, Jennifer Helgeson, Land Restoration, Academic Press, pp. 133-161, 2016, DOI: https://doi.org/10.1016/B978-0-12-801231-4.00014-8.

[8] Zhu M., Wang J., Yang X., Zhang Y., Zhang L., Ren H.,

Wu B., Ye L., A review of the application of machine learning in water quality evaluation, Eco-Environment & Health, Volume 1, Issue 2, pp. 107-116, 2022, DOI: https://doi.org/10.1016/j.eehl.2022.06.001.

[9] Wu Y., Zhang X., Xiao Y., Feng J., Attention neural network for water image classification under IoT environment, Appl. Sci. 10, pp. 030909, 2020, DOI: https:// doi.org/10.3390/app10030909.

[10] Wang Y., Zheng T., Zhao Y., Jiang J., Wang Y., Guo L., Wang P., Monthly water quality forecasting and uncertainty assessment via bootstrapped wavelet neural networks under missing data for Harbin, China, Environ. Sci. Pollut. Control Ser. 20, pp. 8909–8923, 2013, DOI: https://doi.org/10.1007/s11356-013-1874-8.

[11] Zhi W., Feng D., Tsai W., Sterle G., Harpold A., Shen C., Li L., From hydrometeorology to river water quality: can a deep learning model predict dissolved oxygen at the continental scale? Environ. Sci. Technol. 55, 2357–2368, 2021, https://doi.org/10.1021/acs.est.0c06783.

[12] Parsaie A., Nasrolahi A., Haghiabi A., Water quality prediction using machine learning methods, Water Qual. Res. J. 53, pp. 3–13, 2018, DOI: https://doi.org/10.2166/ wqrj.2018.025.

[13] Liu M., Lu J., Support vector machine-an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river? Environ. Sci. Pollut. Control Ser. 21, pp. 11036–11053, 2014, DOI: https://doi.org/ 10.1007/s11356-014-3046-x.

[14] Jeihouni M., Toomanian A., Mansourian A., Decision tree-based data mining and rule induction for identifying high quality groundwater zones to water supply management: a novel hybrid use of data mining and GIS, Water Resour. Manag. 34, pp. 139–154, 2019, DOI: https://doi.org/10.1007/s11269-019-02447-w.

[15] Chen T., Zhang H., Sun C., Li H., Gao Y., Multivariate statistical approaches to identify the major factors governing groundwater quality, Appl. Water Sci. 8, 215, 2018, DOI: https://doi.org/10.1007/s13201-018-0837-0.

[16] Bouamar M., Ladjal M., Evaluation of the performances of ANN and SVM techniques used in water quality classification, 14th IEEE International Conference on Electronics, Circuits and Systems, pp. 1047–1050, 2007, DOI: https:// doi.org/10.1109/ICECS.2007.4511173.

[17] Chen H., Chen A., Xu L., Xie H., Qiao H., Lin Q., Cai K., A deep learning CNN architecture applied in smart near-infrared analysis of water pollution for agricultural irrigation resources, Agric. Water Manag. 240, 106303, 2020, DOI: https://doi.org/10.1016/j.agwat.2020.106303.

[18] Rosen C., Lennox J., Multivariate and multiscale monitoring of wastewater treatment operation, Water Res. 35, pp. 3402–3410, 2001, https://doi.org/10.1016/S0043-1354(01)00069-0.

[19] Foschi J., Turolla A., Antonelli M., Soft sensor predictor of E. coli concentration based on conventional monitoring parameters for wastewater disinfection control, Water Res. 191, 116806, 2021, DOI: https://doi.org/10.1016/j.watres.2021.116806.

[20] Cecconi F., Rosso D., Soft sensing for on-line fault detection of ammonium sensors in water resource recovery facilities, Environ. Sci. Technol. 55, pp. 10067-10076, 2021, DOI: https://doi.org/10.1021/acs.est.0c06111.

[21] Sheng L., Zhou J., Li J., Pan Y., Liu L., Water quality prediction method based on preferred classification, IET Cyber-Physical Systems: Theory & Applications 5, pp. 176-180, 2020, DOI: https://doi.org/10.1049/iet-cps.2019.0062.

[22] Zhou J., Wang Y., Xiao F., Wang Y., Sun L., Water quality prediction method based on IGRA and LSTM, Water 10, pp.1148, 2018, DOI: https://doi.org/10.3390/w10091148.

[23] Du Z., Qi J., Wu Z., Zhang F., Liu R., A spatially weighted neural network based water quality assessment method for large-scale coastal areas, Environ. Sci. Technol. 55, pp. 2553–2563, 2021, DOI: https://doi.org/10.1021/acs.est.0c05928.

[24] Liyanaarachchi S., Shu L., Muthukumaran S., Jegatheesan V., Baskaran K., Problems in seawater industrial desalination processes and potential sustainable solutions: a review, Rev. Environ. Sci. Biotechnol. 13, pp. 203–214, 2015, DOI: https://doi.org/10.1007/s11157-013-9326-y.

[25] Chawla P., Cao X., Fu Y., Hu C., Wang M., Wang S., Gao J., Water quality prediction of salton sea using machine learning and big data techniques, Int. J. Environ. Anal. Chem., 1963713, 2021, DOI: https://doi.org/10.1080/03067319.2021.1963713

[26] Drinking water quality index, https://www.kaggle.com/datasets/adityakadiwal/water-potability?resource=download, accessed January 2023

[27] Kouadri S., Elbeltagi A., Islam A., and Kateb S., "Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast)," Applied Water Science, vol. 11, no. 12, pp. 190, 2021.

[28] Hassan M., Hassan M., Akter L., et al., "Efficient prediction of water quality index (WQI) using machine learning algorithms," Human-Centric Intelligent Systems, vol. 1, no. 3-4, pp. 86–97, 2021.

[29] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning: with Applications in R. New York: Springer, 2013.

[30] Patel J., Amipara Ch., Ahanger T., Ladhva T., Gupta R., Alsaab H., Althobaiti Y., Ratna R., A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI, Computational Intelligence and Neuroscience, Volume 2022, Article ID 9283293, pp. 15, 2022, DOI: https://doi.org/10.1155/2022/9283293

[31] Train/test Split in Python https://realpython.com/train-test-split-python-data/, accessed May 2023

[32] Carhuanayocc R., Cisneros N., Condori R., Pérez G., "A Monte Carlo Simulation for the Improvement of Drinking Water and Sewerage Services in a Northern Settlement in Peru," Environment and Ecology Research, Vol. 10, No. 5, pp. 614 - 625, 2022. DOI: 10.13189/eer.2022.100509.

[33] Zawawi I., Haniffah M., Aris H., "Trend Analysis on Water Quality Index Using the Least Squares Regression Models," Environment and Ecology Research, Vol. 10, No. 5, pp. 561-571, 2022. DOI: 10.13189/eer.2022.100504.