www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 15(5), Dec 2024, 228-236. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2024.v15i5.583

Prediction of Water Quality through Dissolved Oxygen Saturation using Data Mining: A Case Study of Puebla Mexico

M. Claudia Denicia Carral[1], G. Jafet Yáñez García[1], Ana L. Ballinas Hernández[1], and Gustavo M. Minquiz Xolo[1]

[1] Benemérita Universidad Autónoma de Puebla, Complejo Regional Centro, México
claudia.denicia@correo.buap.mx, gustavo.yanezg@alumno.buap.mx, analuisa.ballinas@correo.buap.mx, gustavo.minquiz@correo.buap.mx

Abstract. Computational sciences have been highlighted in the application to any area with good results. One of the primary interests of humanity is water quality because it is a vital resource for the existence of living beings. This research applies a data mining model based on CRISP-DM methodology built to classify water contamination in lotic systems in Puebla City, located in the center of Mexico. The classification was carried out through physicochemical parameters using a water quality evaluation model based on the amount of dissolved oxygen percentage (%DO). Results demonstrate that the application of decision trees and K-NN algorithms, using chemical parameters, are effective in determining the presence of contamination in lotic water bodies and represent a novel way to evaluate water quality in the water system along rivers in Puebla, Mexico.

Keywords: Artificial Intelligence, Data Mining, Water Quality, Water Contamination

Article Info
Received May 10,2024.
Accepted Nov 20, 2024.

1 Introduction

Water is considered the natural resource that most impacts in activities of human beings, since industry, livestock, agriculture, and survival of many species, including human beings, depend on it. Water quality (WQ) has decreased drastically over time, for this reason scientific community has made great efforts to measure and predict water quality indices through the study of some parameters that intervene in its deterioration [1].
The National Water Commission (CONAGUA) oversees measuring water quality in Mexico using some quality indicators, which allow water to be classified using a monitoring system through traffic lights that are determined by means of a Water Quality Index (WQI). This monitoring model also considers classification through indicators such as Escherichia coli, total coliforms, biochemical oxygen demand and the saturation percentage of oxygen dissolved in water.
Dissolved oxygen (DO) refers to free and uncompounded oxygen in water or other liquids, which participates in various biochemical and physiological activities. Measurement of dissolved oxygen (milligrams per liter mg/L) is of great importance to know water quality and thus determine if there is contamination in water. In recent years, different methods based on computational models have been explored to predict DO in water, obtaining encouraging results in the study of water quality and its contaminants worldwide [2].
This work presents a data mining model through the CRISP-DM methodology that allows analyzing water quality and contamination through the prediction of oxygen saturation percentage in lotic water bodies in the state of Puebla. The methodology applied for the analysis is data mining, nevertheless, there exists a difference

through the selection of the techniques and pre-processing of the attributes, and the effect on the algorithms and metrics of the evaluation selected. This research used methodologies where techniques of attributes are based on the quality modeled.
Rest of the paper is organized as follows. Section 2 provides a review of works that include the use of artificial intelligence models for determining water quality and parameters that cause contamination in water bodies. Section 3 describes methods and materials, which include the study area, the description of data sets preparation, as well as the algorithms used and the evaluation method. Section 4 shows experimental results. Finally, section 5 shows conclusions and future work.

2 Related Work

Scientific community in different areas has dedicated efforts to improve water quality. Among research in computational area, the use of artificial intelligence to calculate, predict, and classify water quality has stood out [3]. Most studies focus on the prediction of some quality parameter or on the calculation of some metrics that combines these parameters; main techniques use neural networks [4], decision trees, and regression models [5]. One of the most important parameters is dissolved oxygen, for this reason several research focuses on its prediction through neural network models and data mining [6]. Main parameters used include Biochemical Oxygen Demand (BOD), Toxicity (TOX), Chemical Oxygen Demand (COD), Water Temperature (WATER_TEMP), Hydrogen Potential (PH_FIELD), Electrical Conductivity (CONDUC_FIELD) and Total Suspended Solids (SST) [2], [7]. Metrics used to evaluate results are very diverse, including mean squared error, Nash Sutcliffe efficiency, sum of squared errors, among others. Other authors evaluate their results through classification models, determining classes related to oxygen percentage and its relationship with water quality [5]. Other research proposed a predictive mechanism to provide an acceptable prediction of quality water-use datasheets. The study was developed on the Support Vector Regression (SVR) technique set with Pearson VII Universal Kernel (PUK) and the evolutionary algorithm of Particle Swarm Optimization (PSO) [9]. Artificial intelligence is another technique for better water quality prediction, to analyze the data to make a prediction modeling. The models are Adaptive Neuro-Fuzzy Inference System (ANFIS) and Multy-layer Perceptron Neural Networks (MLP-ANN) with a R^2 equal to or higher than 0.9 [10]. The combination of wavelet transforms with the BP neural network to predict the water quality through the model showed a mean absolute percentage error of 3.822%, with a high learning speed to enhance the accuracy prediction [11]. The principal component regression and gradient boosting classifier approach are other alternatives used to predict the quality and classification of water with a 95% prediction accuracy in quality and 100% in classification [12]. One of the key points to determine if a body of water is contaminated is the selection of parameters to use. [8] shows a classification of parameters that intervene in different international regulations that evaluate water quality. The parameter selection during algorithm analysis is one of the most determinant steps for classification. To get reliable results, the analysis process begins with data selection, preparation, and modeling. The parameter selection is based on human health and the available data. There is a tendency in the literature to choose only one set of parameters, and this tendency was tested in the study [3]. In this study, the parameters selected for analyzing water quality are shown in a comparative table. The research and mix of some parameters led to relevant results in the automatic study prediction of water quality.

However, many of these investigations are carried out in environments other than Mexico [13] and very few focus on analysis for water bodies from the state of Puebla, which in addition to being the fifth most populated state in Mexico, concentrates diverse rivers that they connect with other states. Works such as Salcedo Sánchez et al. [14] evaluate water quality of Puebla Valley and provide a mathematical framework to evaluate based on water chemical parameters. Another work focused on Puebla is described in [15], where the correlation of urbanization levels with water quality is analyzed. However, none of the reviewed works explore water quality of Puebla through a data mining perspective, or from prediction models to classify water quality and its contamination.

3 Methods and Materials

To carry out this research, CRISP-DM (CRoss Industry Standard Process for Data Mining) was used, a data mining model that covers six stages of development through a hierarchy that goes from general to specific. Stages included in this methodology are [16]: understanding business, which consists of understanding the

problem to be solved from business approach; understanding data, which includes data collection, as well as activities to understand and analyze data; data preparation, which encompasses data selection, cleaning, and transformation; modeling, where data mining algorithms are applied; evaluation, which is the process of applying metrics; and finally deployment, which is the final stage in which the knowledge obtained is used to improve processes.

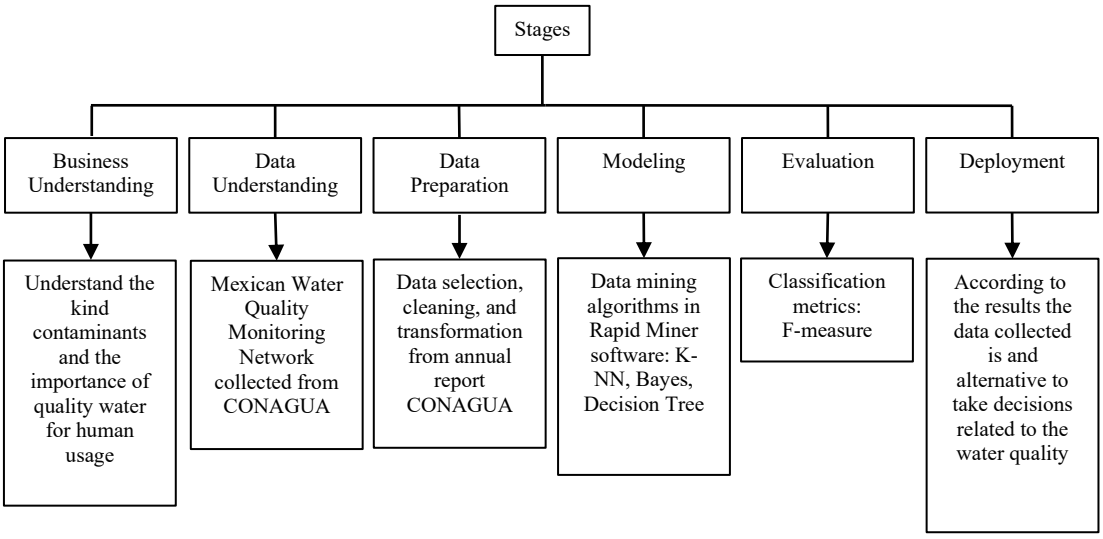

Figure 1. Map of monitoring stations in Puebla. [18]

Selection of this methodology has two foundations, on one hand, it is a methodology that includes stages of understanding problem and data analysis, necessary in applied data mining research, and on other hand, it has been proven that its use is in increase [17].

3.1 Study Area

The state of Puebla, in the center of Mexico, was selected as the study area (see Figure 2). The state of Puebla has a vast number of streams and water bodies, many of which have high levels of pollution caused by illegal discharges from industry, human activities, and impact of climate change.

In this research, data obtained from 75 monitoring stations distributed in 35 municipalities in the state of Puebla were analyzed. Data were downloaded from the CONAGUA repository [19], each station monitors different variables that intervene in water quality. This work used lotic surface water systems, that is, systems with water that flows rapidly in a single direction.

3.2 Datasets

Data set comes from the Mexican Water Quality Monitoring Network collected from 2012 to 2021 from 75 monitoring stations located in Puebla. Initial set consists of 2,394 records for surface waters of lotic water bodies, each record is described through 254 attributes, of which 250 attributes are parameters related to water quality, and the remaining 4 contain information regarding the monitoring station: site_key, monitoring_key, site_name, monitoring_date. Measurements are monthly; however, some stations do not report measurements with this continuity. Table 1 shows distribution of water bodies and number of monitoring stations located along them.

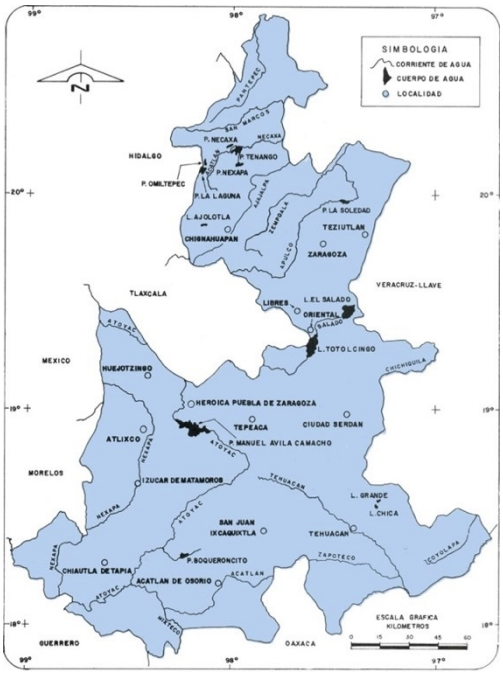

Figure 2. Map of lotic surface water in Puebla. [18]

Table 1. Lotic water bodies and monitoring stations in the state of Puebla

Water body Subtype	Watershed area	# monitoring stations
Stream	Alto Atoyac	8
	Cazones	1
	Salado	2
Ravine	Nexapa	1
Water discharge	Alto Atoyac	2
	Libres Oriental	2
	Nexapa	2
River	Alto Atoyac	16
	Bajo Atoyac	5
	Nexapa	25
	Cazones	5
	Grande de Tulancingo	1
	Mixteco	1
	Tecolutla	4
	Total	75

3.3 Data Preparation

Initial data were pre-processed because they presented several problems such as missing data, formatting errors, and inconsistent data. It was observed that data set contained a relevant number of missing data, since of the 254 attributes that described measurements, only 42 attributes contain sufficient data to calculate water quality.

Data were divided according to parameters used to measure water quality, described by [8], which are: physical parameters that describe noticeable features through senses, such as odor, taste, or turbidity; chemical parameters, which analyze distribution of individual substances that constitute water structure; biological parameters, which allow to identify ability to transport different substances, dissolve others and maintain original water structure. Other important parameters are anion parameters, which are used to analyze neutral energy or conductivity in water; and heavy metal parameters, which are used to analyze metallic elements whose density is at least five times greater than that of water. Table 2 shows selected parameters, maximum and minimum values, average and standard deviation.

Table 2. Attributes selected for the study

Water quality parameters	Parameters/Symbol/Unit	Min	Max	Mean	Std. Dev.
Physical Parameters	Electrical conductivity(EC) (µS/cm)	15	35865	980.24	1003.84
	Temperature (Temp)(°C)	5.5	48.2	20.59	4.12
	Color (Color)(TCU)	5	2000	101.94	151.21
	Turbidity (Turb) (NTU)	2.50	16780	255.84	828.37
	Total Dissolved Solids (TDS)(mg/L)	<10	18900	330.21	912.32
Chemical Parameters	pH value (ph)(pH unit)	4.59	13.66	7.81	0.46
	Hardess (Hads) (mg/L)	<20	2780	319	251.90
	Bio-chemical Oxygen Demand (BOD) (mg/L)	2	3400	41.42	125.43
	Chemical Oxygen Demand (COD) (mg/L)	<10	48806.40	135.60	1041.02
	Total Organic Carbon (TOC) (mg/L)	0.50	1502.05	36.38	75.16
	Dissolved Organic Carbon (DOC) (mg/L)	0.50	782.99	24.09	51.93
Biological Parameters	Total coliform (TC) (NMP/100 mL)	3	2419600	23094.63	114502.23
	Fecal coliform (FC) (NMP/100 mL)	3	9300000	22102.97	203959.85
	Escherichia coli (E-coli) (NMP/100 mL)	1	2419600	12332.33	73976.87
Anions	Nitrate-nitrogen (NO3-N) (mg/L)	0.02	43.33	1.02	2.18
	Ammonia Nitrogen (NH3-N) (mg/L)	0.02	154.59	6.71	14.34
Heavy Metals	Arsenic (As)(mg/L)	<10	139.17	45.11	26.50
	Chromium (Cr) (mg/L)	0.005	0.706	0.021	0.033
	Cadmium (Cd) (mg/L)	0.003	0.160	0.004	0.004
	Lead (Pb) (mg/L)	0.005	0.326	0.013	0.030
	Nickel (Ni)(mg/L)	0.001	0.0480	0.017	0.028
	Mercury (Hg)(mg/L)	0	0.546	0.001	0.013

Evaluation of water quality was determined through the dissolved oxygen saturation percentage parameter (%DO), which measures amount of dissolved oxygen in a water sample compared to the maximum amount that could be present at the same temperature and pressure. 2,103 records were maintained and classified according to contamination thresholds reported in [19], see Table 3. For the class with contamination, 1,100 records were categorized and for the class without dangerous contamination, 1,003 records were maintained, therefore the set is balanced.

Table 3. Classification of water quality by %DO

% OD	Water quality by %DO	Samples	Contaminated Water	Samples
$70 < OD \leq 110$	Excellent	334	NO	
$50 < OD \leq 70$ $110 < OD \leq 120$	Good	669	NO	1003
$30 < OD \leq 50$ $120 < OD \leq 130$	Low pollution	403	SI	
$10 < OD \leq 30$ $130 < OD \leq 150$	Polluted	245	SI	1100
$OD \leq 10$ $OD > 150$	Highly polluted	452	SI	

To find the correlation between variables, a correlation matrix was generated, see Figure 2. It could be observed that the relationship between the dissolved oxygen percentage in lotic water bodies of Puebla is negative with respect to EC, Color, BOD, COD, TOC, DOC, NH3-N while with Temp, ph, NO3-N, and AS it maintains a positive relationship. A positive correlation was also observed between most of the physical and chemical parameters, and very weak relationships with biological parameters and the rest of parameters.

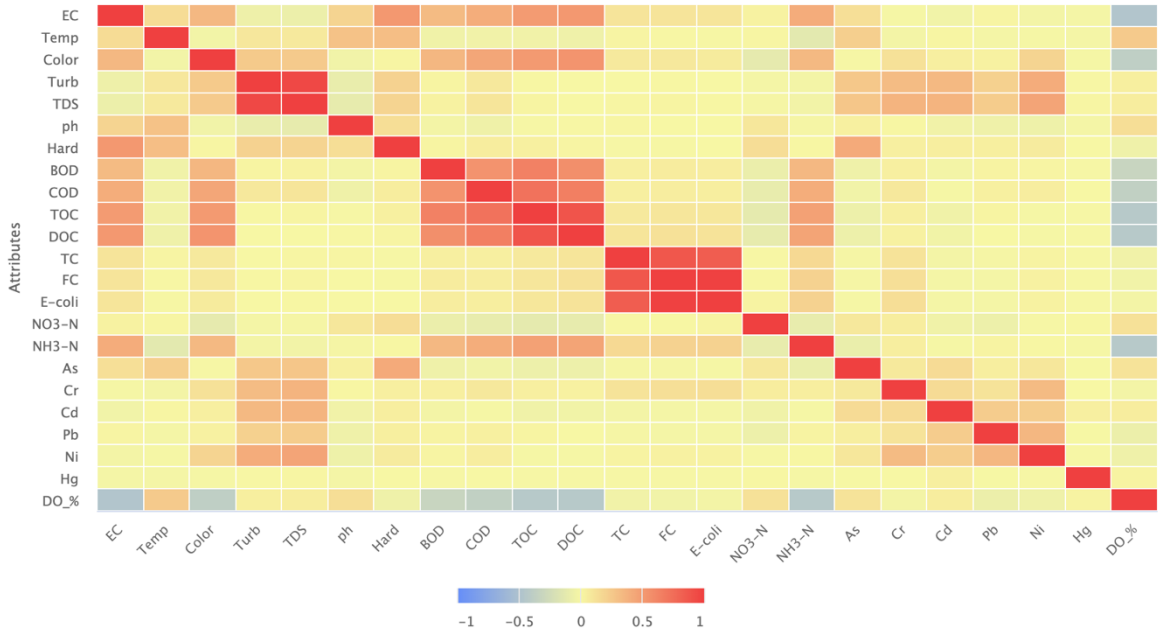

Figure 3. Correlation matrix between pollutants

According to the correlation matrix (Fig 3), it could be observed that the greatest correlation is between parameters of the same type (marked in red), so data sets were generated based on this observation, as well as a review of the state of the art, quality metrics, and an attribute selection model. In the end, 22 different data sets were obtained, which differ in the number of attributes used to represent each water body. Each set explored the attributes by category and later the attributes were mixed.

3.4 Prediction Algorithms

Tests were carried out using three classification algorithms, selection was made based on the state of the art. The decision tree algorithm has shown good results for analyzing hydrological parameters with high precision. Decision tree models create tree diagrams from a logical sequence of simple tests, each test compares a numerical attribute with a threshold value or a nominal attribute with a set of possible values [20]. The k-NN or nearest neighbor algorithm is based on comparing an unknown example with the k training examples that are nearest neighbors of unknown example and it has been used with good results in water quality classification [5]. The Naïve-Bayes algorithm [21], based on Bayes' Theorem, calculates the probability of a subsequent event occurring given probability of previous events and is a good option for classify parameters that are independent of each other. Experiments were run using Rapid Miner software.

3.5 Evaluation

Evaluation was performed using the cross-validation method [22]. To execute this method, set of examples must be divided into k parts of same size (k-folds). Then, one of parts is used as a test subset and (k-1) are used as training subsets. Process is repeated k times to ensure all examples are used for training and testing. For this research, 10-fold cross validation was used. The metric used was F-measure, which evaluates instances percentage classified correctly, based on precision and recall metrics [23].

4 Experimental Results

Results obtained with each of test sets are shown in following table. All trees were generated with a depth of 5 and using information gain, as a quality criterion, the K-NN algorithm used Euclidean distance and 10 neighbors. Following tables show results obtained, Table 4 shows results of classification using a single type

of parameters, best results are obtained using chemical parameters and decision tree algorithm. While the lowest values are using heavy metals according to Naive-Bayes with a value of 21.81.

Table 4. Results obtained by classifying with 1 type of parameter

Experiment	F-measure		
	K-NN	Decision-Tree	Naive-Bayes
Physical Parameters	74.04	*75.67*	67.2
Chemical Parameters	**77.93**	***80.32***	**77.82**
Biological Parameters	54.99	61.06	*64.44*
Anions	*77.1*	76.77	72.14
Heavy Metals	56.05	63.9	21.81

Table 5 shows the combination of two types of parameters. The best results were combining chemical parameters with biological ones using the decision tree algorithm with 81.69. But 63.83 is the lowest value. Additionally, physical and chemical combinations were the best result using K-NN with a value of 80.82. As can be seen from Table 5, there are different combinations of parameters with each of the three algorithms.

Table 5. Results obtained by classifying with 2 types of parameters

Experiment	F-measure		
	K-NN	Decision-Tree	Naive-Bayes
Physical and Chemical	**80.82**	80.61	76.87
Physical and Biological	70.27	73.49	67.91
Physical and Anions	75.66	79.31	74.48
Physical and Heavy Metals	76.84	72.15	24.80
Chemical and Biological	71.76	**81.69**	75.24
Chemical and Anions	77.92	79.62	77.45
Chemical and Heavy Metals	79.27	81.31	**78.84**
Biological and Anions	70.77	76.23	71.02
Biological and Heavy Metals	62.61	63.83	65.46
Anions and Heavy Metals	76.85	79.11	30.50

Table 6 shows the combination of three parameters, and the better results were with physical, chemical, and anions. This behavior demonstrates that the best decision is the tree option. This alternative has a value of 81.51, while 80.32 is the value with the same combination using the K-NN. On the other hand, applying Naive-Bayes is a 77.67 value with the same combination. Besides, chemicals, biology, and anions are the second highest values in the decision tree.

Table 6. Results obtained by classifying with 3 types of parameters

Experiment	F-measure		
	K-NN	Decision-Tree	Naive-Bayes
Physical, Biology and Chemical	74.45	79.27	75.7
Physical, Biology and Anions	69	79.38	73.21
Physical, Biology and Heavy Metals	69.81	72.70	72.93
Physical, Chemical and Anions	80.32	**81.51**	77.67
Physical, Chemical and Heavy Metals	**80.78**	79.64	79.26
Chemical, Biology and Heavy Metals	71.76	80.6	77.67
Chemical, Biology and Anions	71.42	81.16	76.68
Biology, Anions and Heavy Metals	71.82	78.22	72.29
Anions, Heavy Metals and Physical	75.35	79.83	33.73
Chemical, Anions and Heavy Metals	79.03	80.60	**79.61**

Table 7 shows results of combinations of 4 parameters, best results are obtained again with decision tree and parameters physical, chemical, anions, and heavy metals. This classification displayed that the three alternatives have higher values with this combination. Finally, all parameters were combined, and best results were obtained with the decision tree algorithm, obtaining an F-measure value that exceeds 90%, see Table 8.

Table 7. Results obtained by classifying with 4 types of parameter

Experiment	F-measure		
	K-NN	Decision-Tree	Naive-Bayes
Physical, Chemical, Biology and Anions	73.26	81.12	77.86
Physical, Chemical, Biology and Heavy Metals	73.26	79.81	76.92
Chemical, Biology, Anions and Heavy Metals	72.69	81.35	77.57
Physical, Biology, Anions and Heavy Metals	66.76	79.21	74.43
Physical, Chemical, Anions and Heavy Metals	**81.26**	**81.66**	**79.56**

Table 8. Results obtained by classifying with 5 types of parameters.

Experiment	F-measure		
	K-NN	Decision-Tree	Naive-Bayes
Physical, Chemical, Biology, Anions and Heavy Metals Parameters	72.9	**90.84**	79.16

5 Conclusions and Directions for further Research

In this work, quality of surface water in lotic systems in the state of Puebla was analyzed through a data mining process that followed the CRISP-DM guidelines, using the dissolved oxygen saturation percentage as a quality metric to predict the presence of contamination in water through classification algorithms.

Results showed a F-measure greater than 90% in predicting water quality. We attribute these results to the appropriate selection of quality parameters and pre-processing tasks in which careful selection and cleaning were made of data. Combination of different physicochemical parameters resulted in a correct classification, even when only one type of parameter was used, results are encouraging, exceeding 70% of F-measure in most cases.

Decision trees gave better results, they also have the advantage of generating understandable and easy to interpret models compared to numerical steps of other models. In context of water quality classification, quality values can be clearly observed for different parameters that intervene as water contaminants and an additional advantage is decision trees are not affected by missing values.

It is worth mentioning that results obtained are relevant in research of data mining applications and are also useful in analysis and decision-making regarding water contamination in the state of Puebla. Future work focuses on analyzing other case studies to verify results obtained with lotic water bodies of Puebla, and to observe if the behavior is maintained. Another line of research will consist of selection and testing of other sets of parameters, as well as some attribute selection techniques and new classification algorithms.

6 Acknowledgements

Authors appreciate the support granted through the VIEP project of the Benemérita Universidad de Puebla for elaboration of this work.

References

1. Rocha, F. C., Andrade, E. M., Lopez, F. B. Water quality index calculated from biological, physical and chemical attributes. Environ. Monit. Assess. 187, 4163 (2015) doi: 10.1007/s10661-014-4163-1

2. Ziyad Sami, B. F., Latif, S. D., Ahmed, A. N., Chow, M. F., Murti, M. A., Suhendi, A., Ziyad Sami, B. H., Wong, J. K., Birima, A. H., El-Shafie, A. Machine learning algorithm as a sustainable tool for dissolved oxygen prediction: a case study of Feitsui Reservoir, Taiwan. Sci. Rep. 12(1), 1-12 (2022) doi: 10.1038/s41598-022-06969-z
3. 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-Environ. Heal. 1(2) 107–116 (2022) doi: 10.1016/j.eehl.2022.06.001
4. Zhang, Y. F., Fitch, P., Thorburn, P. J. Predicting the trend of dissolved oxygen based on the kPCA-RNN model. Wat. 12(2), 585 (2020) doi: 10.3390/w12020585
5. Tung, T. M., Yaseen, Z. M. A survey on river water quality modelling using artificial intelligence models: 2000–2020. J. Hydrol. 858, 124670 (2020) doi: 10.1016/j.jhydrol.2020.124670
6. Bolick, M. M., Post, C. J., Naser, M. Z., Mikhailova, E. A. Comparison of machine learning algorithms to predict dissolved oxygen in an urban stream. Environ. Sci. Pollut. Res. 30(32), 78075–78096 (2023) doi: 10.1007/s11356-023-27481-5
7. Sánchez, E., Colmenarejo, M. F., Vicente, J., Rubio, A., García, M. G., Travieso, L., Borja, R. Use of the water quality index and dissolved oxygen deficit as simple indicators of watersheds pollution. Ecol. Indic. 7(2), 315–328 (2007) doi: 10.1016/j.ecolind.2006.02.005
8. Akhtar, N. Ishak, M.I.S. Ahmad, M.I. Umar, K. Md Yusuff, M.S. Anees, M.T. Qadir, A. Ali Almanasir, Y.K. Modification of the Water Quality Index (WQI) Process for Simple Calculation Using the Multi-Criteria Decision-Making (MCDM) Method: A Review. Wat. 13(7), 905 (2021) doi 10.3390/w13070905
9. López, I. D., Figueroa, A., & Corrales, J. C. Adaptive prediction of water quality using computational intelligence techniques. In International Conference on Computational Science and Its Applications (pp. 45-59). Cham: Springer International Publishing (2017). https://doi.org/10.1007/978-3-319-62395-5_4
10. Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., ... & Elshafie, A. Machine learning methods for better water quality prediction. Journal of Hydrology, 578, 124084 (2019).. https://doi.org/10.1016/j.jhydrol.2019.124084
11. Xu, L., & Liu, S. Study of short-term water quality prediction model based on wavelet neural network. Mathematical and Computer Modelling, 58(3-4), 807-813 (2013). https://doi.org/10.1016/j.jhydrol.2019.124084
Khan, M. S. I., Islam, N., Uddin, J., Islam, S., & Nasir, M. K. (2022). Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. Journal of King Saud University-Computer and Information Sciences, 34(8), 4773-4781(2022). https://doi.org/10.1016/j.jksuci.2021.06.003
12. Areerachakul, S., Sanguansintukul, S. Water quality classification using neural networks: Case study of canals in Bangkok, Thailand. Int. Conf. Internet Technol. Secur. Trans. ICITST 2009, 1-5 (2009) doi: 10.1109/icitst.2009.5402577
13. Salcedo-Sánchez, E. R., Garrido Hoyos, S. E., Esteller Alberich, M. V., Martínez Morales, M. Application of water quality index to evaluate groundwater quality (temporal and spatial variation) of an intensively exploited aquifer (Puebla valley, Mexico). Environ. Monit. Assess. 188(10), 573 (2016) doi: 10.1007/s10661-016-5515-9
14. Estrada-Rivera, A., Díaz Fonseca, A., Treviño Mora, S., García Suastegui, W. A., Chávez Bravo, E., Castelán Vega, R., Morán Perales, J. L., Handal-Silva, A. The Impact of Urbanization on Water Quality: Case Study on the Alto Atoyac Basin in Puebla, Mexico. Sustain. 14(2), 1-15 (2022) doi: 10.3390/su14020667
15. López, C. P. Data Mining. The CRISP-DM Methodology. The CLEM language and IBM SPSS Modeler. Lulu Press, Inc. (2021)
16. Schröer, C., Kruse, F., Gómez, J. M. A systematic literature review on applying CRISP-DM process model. Procedia Comput. Sci. 181, 526–534 (2019) doi: 10.1016/j.procs.2021.01.199
17. CEIGEP, Anuario Estadístico y Geográfico de Puebla 2017, 2021. http://ceigep.puebla.gob.mx/# (accessed Jul. 26, 2023)
18. C. N. del A. CONAGUA, Calidad del agua en México | Comisión Nacional del Agua | Gobierno | gob.mx, Gobierno de Máxico, 2023. https://www.gob.mx/conagua/articulos/calidad-del-agua (accessed Jul. 25, 2023).
19. Kotsiantis, S. B. Decision trees: A recent overview. Artif. Intell. Rev. 39(4), 261–283 (2013) doi: 10.1007/s10462-011-9272-4
20. Rish, I. An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence 3(22), 41-46 (2001)
21. Browne, M. W. Cross-Validation Methods. J. Math. Psychol. 44, 108-132 (2000) doi: 10.1006/jmps.1999.1279
22. Sasaki, Y. The truth of the F-measure. Teach tutor mater, 1(5), 1-5 (2007).