CLASSIFICATION OF WATER QUALITY BY DIFFERENT ARTIFICIAL INTELLIGENCE ALGORITHMS

Mario Elias Carvalho do Nascimento [1], Ralpho Rinaldo dos Reis [1]

[1] Postgraduate Program in Environmental Engineering and Technology, Western Paraná State University, Cascavel, Paraná, Brazil.
marioelias.carvalho@gmail.com, ralpho.reis@unioeste.br

Corresponding author: marioelias.carvalho@gmail.com

ABSTRACT

The water and the monitoring of its quality is an area of great importance for modern society. Indexes for assessing water quality (WQI) have been developed for many years, serving as easy-to-interpret tools. However, with the increase in data collection and consequently an increase in the complexity of systems, there is a need to use automated and more modern techniques. With the increasing use of machine learning algorithms to optimize processes, this work proposed to evaluate 10 classic machine learning models (LDA – Linear Discriminant Analysis, QDA – Quadratic Discriminant Analysis, LR – Logistic Regression, Perceptron, RC – Ridge Classifier, GNB – Gaussian Naive Bayes, KNN – Knearest Neighbors, SVM – Support Vector Machine, DT – Decision Tree and MLP – Multilayer Perceptron) and 5 ensemble models (ADB - AdaBoost, BAG - Bagging, ET - Extra Tree, GDB - Gradient Boosting and RF – Random Forest). The statistical evaluation of the obtained models was made using the following metrics: balanced accuracy, precision, recall and f1, confusion matrix. Then, the number of variables was reduced to ensure an acceptable classification and the limit of 4 features was reached (tc – thermotolerant coliforms, bod – biochemical oxygen demand, do – dissolved oxygen and tp – total phosphorus). The SVM and GDB techniques stood out as they obtained the best metrics. This work demonstrated the feasibility of using artificial intelligence models to classify water quality, as well as the possibility of predicting this classification with a smaller number of measured variables.

KEYWORDS: *index water quality; machine learning; classical models; ensemble models; classification multi-class*.

1. INTRODUCTION

Water is universally recognized as a vital resource essential for sustaining life on Earth, and its physical, chemical, and biological properties require continuous monitoring to ensure its quality and availability. This awareness has led to the development of methodologies to systematically assess water quality, with indices emerging as the most widely adopted approach. Water quality indices (WQIs) provide a simplified, user-friendly means of translating complex environmental data into interpretable metrics, which can be easily understood by the general public and policymakers. Since the late 1960s, numerous WQIs have been proposed, each characterized by distinct methodologies, parameters, and applications.

One of the earliest contributions to the field was made by [1], who developed an index comprising eight parameters scaled from 0 to 100. Each parameter was assigned weights ranging from 1 to 4, reflecting its relative impact on water quality. This index set the foundation for future advancements by providing a quantifiable measure of water quality. Subsequently, [2] introduced a color-coded map for classifying water quality, integrating chemical and biological parameters to facilitate visual interpretation.

Building on these early developments, [3, 4] utilized the Delphi method to construct an index represented on a scale from 0 to 100, with colors ranging from dark red (poor quality) to dark blue (excellent quality). This index incorporated nine physical, chemical, and biological variables, each assigned weights based on their significance. Initially, the index employed a summation method for calculation, but this was later replaced with a product-based approach in 1973 to enhance its sensitivity.

The evolution of water quality indices continued with numerous studies that expanded their scope and refined their methodologies. For example, [5] focused on a surface water pollution index tailored to assess contamination levels. Meanwhile, [6] introduced a composite WQI derived from the integration of two other indices: the industrial and municipal effluent index and the ambient water quality index. This approach allowed for a more holistic evaluation of water quality across diverse sources.

In a comparative study, [7] evaluated the statistical performance of five different WQIs, providing valuable insights into their reliability and applicability. Similarly, [8] proposed a classification scheme categorizing water quality into five levels, ranging from very bad to very good, thereby standardizing quality assessments. An innovative index designed for specific applications was presented by [9], who applied it across four distinct water use classes: bathing, water supply, fish spawning, and general use.

Comprehensive reviews by [10] and [11] further enriched the field by documenting a wide range of indices developed globally. These reviews underscored the diversity in validation methodologies, parameter selection, and contextual applicability, highlighting the adaptability of WQIs to address varied environmental challenges.

Artificial intelligence is a powerful mathematical tool that can be used to calculate WQIs, as it has already been used in several other areas. [12] demonstrated that AI has been researched and applied in the area of higher education for at least 30 years. [13] presented the use of algorithms in processing X-ray and computed tomography images to optimize the diagnosis of COVID-19. [14] analyzed the techniques applied in marketing and pointed out strategies for use by companies that wish to increase their market share. [15] discussed the changes and challenges that algorithms have in optimizing Industry 4.0. [16] pointed out the main AI methodologies used in dentistry. [17] commented on several machine learning and data mining methods for analyzing and validating food quality. [18] reported ensemble machine learning paradigms in hydrology, emphasizing simulation and prediction.

WQIs and AI algorithms have been improved to optimize processes. On the other hand, despite the efforts of several researchers to integrate these distinct areas, there are still gaps to be explored: in the simulation, prediction, classification, identification, and modeling of the parameters involved in WQIs. There is little integration between computing professionals and those in the areas of engineering and environmental sciences [19]. However, AI techniques prove to be effective in analyzing and managing water quality [20]. According to [21], river pollution promotes social, environmental, economic, human health and water scarcity problems, with monitoring being an important factor in decision-making, especially for relevant rivers, such as the Tietê.

The Tietê River is an important river that runs through the state of São Paulo, Brazil. This Brazilian state has the highest population density and the highest GDP in the country. It is relevant in the generation of electrical energy, in the food industry, in transport and navigation, and suffers from the degradation of the quality of its waters resulting from several factors: urban and industrial effluents, solid material carried by rain, deforestation of riparian forests, among others [22]. In 2022, environmental reports indicate that in metropolitan regions water quality has been improving, this is due to sanitation works. On the other hand, regions that have experienced intense economic changes, where there is intensive agriculture, low vegetation cover, excessive use of fertilizers and pesticides and increased urbanization, showed a drop in WQIs. Therefore, there is a need for constant monitoring of these waters [23].

The objective of this work was to classify the WQI of the Tietê River, using 15 machine learning models, to analyze and statistically compare their performance. In this way, we will present which of them have the best performance and can be used. Furthermore, the relevance of the physical,

chemical, and biological parameters measured was determined to determine the possibility of reducing the number of parameters to obtain an efficient classification.

This article was divided into four sections. The first is an introduction to the topic and the problem, the second is the methodologies, how the datasets were set up and the main mathematical and data manipulation methods used to develop the work, the third is the results obtained in the process and a brief comparation with results from other articles found in the bibliography that had as their object the water quality index and/or use machine learning techniques in classification and the fourth the conclusion of the work.

2. MATERIALS AND METHODS

2.1 Characterization of the study area

Tietê River crosses the state of São Paulo, running from East to West, with a length of 1,100 km, as shown in figure 1. Its source is located in the city of Salesópolis and its mouth is in the municipality of Itapura, making its flow into the Paraná River. It comprises an area of 9,172,066 hectares, with 265 municipalities and a population of 25 million inhabitants. It is an intrastate river, having 6 sub-basins: Tiete/Batalha, Baixo Tietê, Alto Tietê, Piracicaba, Sorocaba/ Médio Tietê, Tietê/Jacaré.

Figure 1. Tietê River

2.2 Construction of Datasets

Data from WQI (response variable) and 9 physical, chemical, and biological parameters (independent variables), namely: hydrogen potential (pH), dissolved oxygen (do), biochemical oxygen demand (bod), thermotolerant coliforms (tc), total nitrogen (tn), total phosphorus (tp), total solids (ts), turbidity (turb), temperature (temp) were obtained from reports issued by the Environmental Company of the State of São Paulo (CETESB). The WQI used presents a range between 0 and 100, divided into ranges of values and by color: very bad (0 and ≤ 19, purple); bad (> 19 and ≤ 36, red); regular (> 36 and ≤ 51, yellow); good (> 51 and ≤ 79, green) and excellent (> 79, blue). This index is calculated from the 9 parameters mentioned above. Thus, if one of the parameters is not measured, the calculation of the WQI is not feasible [4].

The present work developed 2 WQI datasets. The first used as its object of study data obtained from 78 measurement points located on the Tiete River and its tributaries, from 1994 to 2019 [24]. After cleaning up missing or inconsistent data, it totaled 7101 samples, this database was called “tiete”. The second dataset was assembled using all measurement points from the year 2019 in the state of São

Paulo, except those that are part of the "tiete" dataset. After using cleaning techniques in pre-processing, it ended up with 2332 samples, and it was named "sp2019".

2.3 Simulations, outlier detection, transformation, and relative importance of features

2.3.1 Simulation

The simulations were developed utilizing a range of sophisticated tools and programming libraries to ensure precision and efficiency in both machine learning (ML) algorithm training and data analysis. The Anaconda Navigator 2.3.2 [25] graphical user interface (GUI) was employed as the primary environment for managing dependencies and packages. Computational tasks were carried out within the Jupyter Notebook 6.4.8 [26] interactive environment, facilitating seamless integration and execution of code.

The simulations were programmed using the Python 3.9.12 [27] language, which provided robust support for machine learning and data manipulation tasks. To train and validate ML algorithms, the Scikit-learn 1.2.2 [28] library was utilized, offering a comprehensive suite of tools for predictive modeling and data processing. The visualization and diagnostic capabilities of machine learning models were enhanced using the Yellowbrick 1.5 [29] library, which provided a wide array of visual tools to interpret and evaluate model performance. Additionally, the Pandas 2.0.1 [30] library was employed for efficient and flexible manipulation of datasets, enabling structured data analysis and preprocessing.

2.3.2 Outlier detection

The ensemble technique, Isolation Forest, was used to detect outliers in the "tiete" dataset. The "tiete" dataset with 7101 samples, which was used as a trainset and testset of the machine learning models, after using the outlier detection technique, presented 6545 samples, and it was divided into a scale of 80% (5236 samples) for the set of training and 20% (1309 samples) for the test set. Table 1 summarizes by class and presents the percentage of division of the dataset.

Table 1 Demonstration of the "tiete" dataset without outliers.

Classes	Trainset (80%)	Testset (20%)
Very bad	920 (17.57%)	211 (16.11%;)
Bad	1216 (23.22%)	348 (26.58%)
Regular	644 (12.29%)	165 (10.61%.)
Good	1777 (33.93%)	446 (34.07%)
Excellent	679 (12.96%)	139 (12.60%)
Total	5236	1309

2.3.3 Data transformation

In the area of machine learning there is a need to transform data, when the magnitude scales of the variables are very different. In this case, standardization is needed to ensure better performance of the algorithms. There are some common techniques for transforming data, min-max and z-score are the most used. However, we opted for the Yeo-Johnson transformation technique presents a good response to negative data or zeros [31].

2.3.4 Relative importance of features and variable reduction

The *features_importance_* attribute, present in the Random Forest algorithm, was used to order the relative importance of each feature present in the database, "tiete". The hyperparameters *class_weight* was modified from *none* to *balanced* to assign greater weights to less frequent classes.

2.4 Artificial intelligence techniques

2.4.1 Classic machine learning techniques

The LDA, QDA, LR, Perceptron, RC, GNB, KNN, SVM, DT, and MLP. The hyperparameters details are presented in table 2.

Table 2 Hyperparameters of classic models.

Model	Solver	Penalty	Class	N_neighbors	W	Activation
LDA	Svd	-	-	-	-	-
QDA	-	-	-	-	-	-
LR	Newtoncg	L2	Balanced	-	1.0	-
Perceptron	-	-	Balanced	-	-	-
RC	Self	-	Balanced	-	-	-
GNB	-	-	-	-	-	-
KNN	-	-	-	5	-	-
SVM	Rbf	-	-	-	1.0	-
DT	Gini	-	Balanced	-	-	-
MLP	Adam	-	-	-	-	Relu

2.4.2 Ensemble machine learning techniques

The Bagging, Random Forest, Extremely Randomized Tree, AdaBoost, and Gradient Boosting algorithms are described in the supporting material. Table 3 is presented here with the hyperparameters of the models.

Table 3 Hyperparameters of ensemble models.

Model	Estimator	Lear. Rate	Class	Max_depth	Criterion	Loss
ADA	100	1.15	Balanced	3	-	-
BAG	100	-	-	-	-	-
ET	100		Balanced	3	Gini	-
GDB	100	0.1	-	3	Friedman_mse	Log_loss
RF	100	-	Balanced	3	Gini	-

2.5 Model evaluation

To evaluate the machine learning models, graphical techniques were used: rocauc, precision–recall curve and confusion matrix. Mathematical techniques were also used: precision, recall, f1, balanced accuracy. In tuning the hyperparameters, the validation curve technique was used. In this way, evaluating how changing a parameter of the model can influence its performance. It is necessary to choose a specific metric (accuracy, balanced accuracy, precision, recall or f1) to evaluate the impact of changing the parameter. In this work, the metric chosen was f1.

The confusion matrix technique relates the accuracy between the predicted classes and the real classes. Precision-recall curve presents the tradeoff between precision and recall at different thresholds. ROCAUC curve combines the ROC (Receiver Operating Characteristic) assessment that compares the predictive quality of the model in several tradeoffs relating sensitivity and specificity and AUC (Area Under Curve) calculates the interaction between false positives and true positives, normally the ROC curve is used for binary classifications of the model output, but the yellowbricks library can draw the graph by doing a binary classification by class when the problem is multi-class.

Precision, equation 1, is defined as the ratio between the number of true positives (T_p) and the sum of false positives (F_p) and true positives (T_p).

$$P = \frac{T_p}{T_p + F_p} \quad (1)$$

Recall, equation 2, is the ratio of true positives (T_p) to the sum of false negatives (F_n) and true positives (T_p).

$$R = \frac{T_p}{T_p + F_n} \tag{2}$$

F1, equation 3, is the harmonic mean between precision and recall.

$$F1 = 2\ x\ \frac{P\ x\ R}{P + R} \tag{3}$$

Balanced accuracy, equation 4, is a relationship between sensitivity, rate of true positives (T_p), and specificity, rate of true negatives (T_n). The metric has the advantage of not being influenced by unbalanced classes [32].

$$BA = \frac{1}{2}\left(\frac{T_p}{T_p + F_n} + \frac{T_n}{T_n + F_p}\right) \tag{4}$$

3 RESULTS AND DISCUSSION

3.3 Datasets

The "tiete" dataset presents water quality stratified into 5 levels: "very bad", "bad", "regular", "good", "excellent". Table 4 summarizes the data from this dataset before and after using the outlier detection and removal technique.

Table 4 Demonstration of the "tiete" dataset with and without outliers.

Classes	With outliers	Without outliers
Very bad	1511 (21.27%)	1131 (17.28%)
Bad	1694 (23.85%)	1564 (23.89%)
Regular	834 (11.74%)	818 (12.49)
Good	2243 (31.58%)	2223 (33.96%)
Excellent	819 (11.53%)	809 (12.36%)
Total	7101	6545

The descriptive statistics after removing the outliers in the "tiete" dataset is presented in table 5.

Table 5 Descriptive statistics of the "tiete" dataset without outliers.

Variable	Mean	Sd	Min	Max
pH	7,170	0.468	4.5	9.6
do (mg/L)	4,155	2,870	0	16.80
bod (mg/L)	15,443	20,410	0	210
tc (NMP/100mL)	655263	1636620	0	25000000
tn (mg/L)	7,738	7,819	0.130	69,270
tp (mg/L)	0.606	0.789	0.002	13
ts (mg/L)	215.20	122,428	1	978
turb (UNT)	29,086	36,568	0	330
temp (ºC)	23,019	3,292	13	35

The "sp2019" dataset has 2332 stratified samples: "very bad" – 93 (3.98%), "bad" – 204 (8.74%), "regular" – 374 (16.03%), "good" – 1392 (59.69%) and "excellent" – 269 (11.53%). Table 6 summarizes the descriptive statistics of the variables.

Table 6 Descriptive statistics of the "sp2019" dataset.

Variable	Mean	Sd	Min	Max
pH	7,106	0.531	3.40	10
do (mg/L)	6,197	2,272	0.10	17.20
bod (mg/L)	7,641	18.24	two	332
tc (NMP/100mL)	371620	2444414	two	8166667
tn (mg/L)	3,865	6,590	0.360	78.21
tp (mg/L)	0.371	0.876	0.007	9.21
ts (mg/L)	149.28	124.53	14.80	1520
turb (UNT)	31.74	63.14	0	1200
temp (ºC)	22.95	3.42	11	31.20

3.4 Evaluation of models with 9 variables

3.4.1 Training and cross-validation

Details of the metrics and values of the confusion matrix containing the maximum accuracy per water quality class can be seen in table 7.

Table 7 Performance of ML models for the training set

Model	Ba[a]	P[b]	R[c]	F1	Confusion matrix				
					0	1	2	3	4
LDA	0.871	0.878	0.877	0.876	889	993	502	1603	603
QDA	0.878	0.891	0.891	0.891	862	1050	480	1649	623
LR	0.912	0.911	0.904	0.905	893	1045	568	1572	654
Perceptron	0.735	0.746	0.753	0.744	856	1156	26	1700	118
RC	0.643	0.698	0.610	0.556	870	534	20	1518	371
NBG	0.828	0.850	0.852	0.850	847	1049	360	1592	613
KNN	0.903	0.915	0.916	0.915	875	1098	492	1686	643
SVM	0.965	0.961	0.959	0.960	901	1143	621	1685	673
DT	0.772	0.808	0.789	0.795	725	920	424	1518	545
MLP	0.918	0.928	0.928	0.928	888	1096	537	1713	627
ADA	0.894	0.899	0.896	0.897	866	1025	520	1628	655
BAG	1.0	1.0	1.0	1.0	920	1216	644	1777	679
ET	0.844	0.845	0.816	0.817	912	860	509	1320	672
GDB	0.978	0.982	0.982	0.982	914	1192	609	1766	663
RF	0.847	0.851	0.835	0.837	856	908	497	1451	658
Max	-	-	-	-	920	1216	644	1777	679

[a] balanced accuracy, [b] precision, [c] recall, 0 – "very bad", 1 – "bad", 2 – "regular", 3 – "good" and 4 – "excellent"

The algorithm that presented the best training performance was Bagging, obtaining all metrics equal to 1. Consequently, correctly predicting all water quality levels in the trainset. On the other hand, this adjustment meant that the model could be in overfitting, an unwanted state. Corroborating this deduction, the values of the metrics using the cross-validation technique were very low when compared to the previous metrics.

Table 8 shows the values of the metrics when using the 10-fold cross-validation.

Table 8 Cross-validation performance (10 folds) of ML models

Model	Ba[a]	P[b]	R[c]	F1
LDA	0.871 ± 0.017	0.879 ± 0.013	0.876 ± 0.013	0.876 ± 0.014
QDA	0.871 ± 0.014	0.885 ± 0.013	0.885 ± 0.013	0.885 ± 0.013
LR	0.908 ± 0.012	0.908 ± 0.011	0.900 ± 0.012	0.901 ± 0.012
Perceptron	0.623 ± 0.057	0.693 ± 0.069	0.662 ± 0.104	0.652 ± 0.092
RC	0.639 ± 0.014	0.699 ± 0.028	0.602 ± 0.015	0.547 ± 0.017
NBG	0.826 ± 0.020	0.848 ± 0.016	0.849 ± 0.015	0.848 ± 0.016
KNN	0.857 ± 0.017	0.875 ± 0.014	0.876 ± 0.013	0.874 ± 0.013
SVM	0.928 ± 0.012	0.930 ± 0.011	0.927 ± 0.010	0.928 ± 0.010

DT	0.763 ± 0.018	0.802 ± 0.016	0.782 ± 0.016	0.788 ± 0.015
MLP	0.908 ± 0.018	0.921 ± 0.014	0.921 ± 0.013	0.921 ± 0.014
ADA	0.881 ± 0.017	0.888 ± 0.009	0.877 ± 0.011	0.879 ± 0.011
BAG	0.884 ± 0.017	0.912 ± 0.010	0.911 ± 0.011	0.909 ± 0.011
ET	0.840 ± 0.012	0.841 ± 0.011	0.815 ± 0.012	0.815 ± 0.012
GDB	0.916 ± 0.010	0.930 ± 0.008	0.930 ± 0.008	0.929 ± 0.008
RF	0.836 ± 0.014	0.842 ± 0.014	0.824 ± 0.015	0.827 ± 0.015

[a] balanced accuracy, [b] precision, [c] recall.

GradientBoosting was the second best in performance, with balanced accuracy, precision, recall and f1 values of 0.978, 0.982, 0.982, and 0.982, respectively. Regarding cross-validation, the following values were obtained in order of the same metrics: 0.916 ± 0.010, 0.930 ± 0.008, 0.930 ± 0.008, and 0.929 ± 0.008. The third best algorithm was the SVM with: 0.965, 0.961, 0.959, and 0.960; and for cross-validation: 0.928 ± 0.012, 0.930 ± 0.011, 0.927 ± 0.010 and 0.928 ± 0.010.

Other models that showed values above 0.90 in the 4 metrics for both training and cross-validation were: LR and MLP. Perceptron, Ridge, and DT were the only models that performed below 0.80 in all evaluated metrics, with the Ridge model having the worst performance of all.

Regarding prediction by level of water quality, no model stood out in all classes. The models that performed well in the classification at a specific level were: GDB for “very bad”, “bad” and “good” water and the SVM for “regular” and “excellent” – see table 7.

3.4.2 External validation

Two external validations were carried out, one using the testset from the “tiete” dataset and the other using the “sp2019” dataset. Tables 9 and 10 respectively summarize the performance of the models using testset and sp2019.

SVM, MLP, and GradientBoosting were the best performing models with the testset. All presented metrics above 0.90, with SVM being the best performing model with 0.922, 0.923, 0.921, and 0.921 for balanced accuracy, precision, recall and f1, respectively. The algorithms with the lowest performance were Perceptron and Ridge, with Ridge being the only model with metrics below 0.70.

SVM obtained the highest metrics, but despite this it did not obtain the greatest number of classification hits at a specific quality level; on the other hand, it was always among the 4 algorithms with the highest number of hits. For the "very bad" level, Ridge and ET stood out. For "bad," the algorithm with the highest accuracy was Bagging. For "regular" it was LR. "Good" went back to Bagging. And for "excellent" it was ET.

Table 9 Performance of ML models for the testset.

Model	**Ba[a]**	**P[b]**	**R[c]**	**F1**	**Confusion matrix**				
					0	1	2	3	4
LDA	0.857	0.861	0.855	0.855	207	276	122	389	125
QDA	0.861	0.872	0.870	0.870	198	291	120	406	124
LR	0.898	0.894	0.884	0.886	210	301	139	376	131
Perceptron	0.666	0.711	0.728	0.714	202	285	21	358	87
RC	0.627	0.686	0.557	0.508	211	31	73	278	136
NBG	0.819	0.834	0.836	0.834	199	292	88	389	126
KNN	0.867	0.880	0.881	0.879	208	295	110	414	126
SVM	0.922	0.923	0.921	0.921	210	310	139	413	133
DT	0.760	0.797	0.774	0.781	176	252	112	372	101
MLP	0.905	0.913	0.912	0.912	210	304	132	421	127
ADA	0.868	0.876	0.862	0.865	199	277	140	390	122
BAG	0.872	0.905	0.904	0.900	207	321	107	431	117
ET	0.841	0.839	0.801	0.802	211	240	128	330	139
GDB	0.904	0.918	0.919	0.918	209	319	127	424	124
RF	0.833	0.834	0.815	0.818	203	260	117	355	132
Max	-	-	-	-	211	348	165	446	139

[a] balanced accuracy, [b] precision, [c] recall, 0 – “very bad”, 1 – “bad”, 2 – “regular”, 3 – “good” and 4 – “excellent”

Table 10 Performance of ML models for the sp2019 dataset.

Model	Ba[a]	P[b]	R[c]	F1	Confusion matrix				
					0	1	2	3	4
LDA	0.435	0.536	0.454	0.451	32	51	36	671	269
QDA	0.373	0.589	0.616	0.585	23	52	80	1204	76
LR	0.407	0.507	0.397	0.397	30	41	46	539	269
Perceptron	0.499	0.615	0.573	0.562	26	75	77	889	269
RC	0.539	0.665	0.541	0.538	82	6	96	828	248
NBG	0.815	0.865	0.861	0.862	77	162	274	1282	212
KNN	0.809	0.857	0.860	0.857	82	166	244	1311	201
SVM	0.900	0.918	0.915	0.916	85	183	334	1301	229
DT	0.739	0.799	0.751	0.767	74	139	257	1081	200
MLP	0.868	0.908	0.909	0.908	84	174	303	1342	215
ADA	0.828	0.882	0.868	0.872	76	137	326	1250	235
BAG	0.848	0.893	0.891	0.888	89	175	261	1353	200
ET	0.853	0.868	0.853	0.855	90	136	301	1207	255
GDB	0.887	0.914	0.913	0.913	85	186	313	1327	218
RF	0.826	0.856	0.843	0.845	90	124	293	1221	237
Max	-	-	-	-	92	204	374	1392	269

[a] balanced accuracy, [b] precision, [c] recall, 0 – "very bad", 1 – "bad", 2 – "regular", 3 – "good" and 4 – "excellent".

For the "sp2019" dataset, only the SVM algorithm obtained metrics above 0.90. The LDA, QDA, Perceptron, and Ridge models obtained the weakest performances with values ranging between 0.373 and 0.665. The ET and RF Algorithm were highlighted in the classification of "very bad" water, with 90 correct answers from 92 samples. GradientBossting scored 186 out of 204 for the "bad" level. SVM correctly classified 334 out of 374 for "regular" water. Bagging managed 1353 hits out of a total of 1392 for the "good" level. RF scored 237 out of 269 classifications for "excellent" water. It is noted for this level of quality 100% hits for the LDA, LR, and Perc algorithms, but this is not due to its performance, but to the bias of classifying the samples at one level above, and as there was no level above "excellent", all of them were classified as "excellent".

The most recent work demonstrates the efficiency of machine learning models in the classification process. The work of [33] stood out. They worked with the under and over-sampling techniques for generating synthetic data with the objective of making a binary classification. They obtained for the SVM, KNN, and MLP models an accuracy of 0.73, 0.75, and 0.75 in the under-sampling technique and 0.73, 0.98, and 0.97 for the over-sample respectively. With the aim of binary classification, [34] proposed a Stacking model based on 3 linear models (MLR, PLS, and SPLS) and 2 non-linear models (RF and BN) to classify the excess value of *E.coli* on 3 beaches in New York, achieving accuracy of 0.78, 0.81, and 0.823 for each location. In this way, we observed that the balanced accuracy values obtained by most of our models with 9 variables – except for Perceptron, Ridge, NBG, and DT – are better than those presented in the literature.

3.5 Evaluation of models after variable reduction

Figure 3 is the representation of the variables present in the "tiete" dataset and their relative importance in relation to the WQI classification. It was observed in the change of *max_depth*, an alternation in the position of importance of the variables, and the 4 most important ones were always (tc, do, bod, and tp). In contrast, the 3 less important variables did not alternate their position, only their value relative importance, obtaining ts (3rd), ph (2nd) and temp (1st) as less important. *Max_depth* equal to 2 was chosen because the response is less biased.

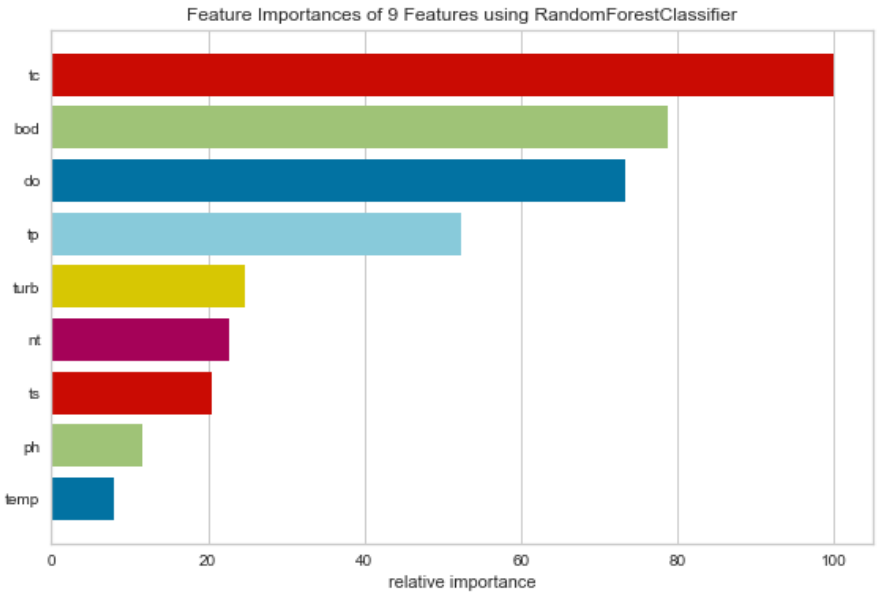

Figure 3 Feature importance: feature and relative importance.

3.5.1 8 (tc, bod, do, tp, turb, tn, ts and ph), 6 (tc, bod, do, tp, turb and tn) and 4 (tc, bod, do and tp) features

For artificial intelligence models built from 8 variables (excluding the temperature variable, which is the least important), 6 variables (excluding temperature, ph, and ts, which are the 3 least important) and 4 variables (tc, bod, do, tp, which are the 4 most important) the metrics values are presented in table 11 for the trainset. The cross-validation data are presented in table 12.

Table 11 Performance of the ML models with 8, 6, and 4 features for the trainset.

Model	Ba[a]	P[b]	R[c]	F1
SVM8	0.937	0.946	0.946	0.946
SVM6	0.933	0.941	0.941	0.941
SVM4	0.906	0.919	0.920	0.920
GDB8	0.979	0.983	0.983	0.983
GDB6	0.973	0.977	0.977	0.977
GDB4	0.949	0.955	0.955	0.955
MLP8	0.918	0.927	0.928	0.927
MLP6	0.909	0.918	0.919	0.918
MLP4	0.884	0.899	0.900	0.900
KNN8	0.916	0.928	0.929	0.928
KNN6	0.931	0.939	0.940	0. 939
KNN4	0.924	0.933	0.934	0.933

[a] balanced accuracy, [b] precision, [c] recall. *8, 6, and 4 – number of features.

Table 12 Performance of the ML models with 8, 6, and 4 features for cross-validation.

Model	Ba[a]	P[b]	R[c]	F1
SVM8	0.923 ± 0.015	0.935 ± 0.011	0.935 ± 0.011	0.934 ± 0.011
SVM6	0.922 ± 0.012	0.933 ± 0.010	0.933 ± 0.010	0.932 ± 0.010
SVM4	0.900 ± 0.016	0.914 ± 0.013	0.914 ± 0.013	0.914 ± 0.013
GDB8	0.916 ± 0.010	0.929 ± 0.008	0.929 ± 0.008	0.928 ± 0.008
GDB6	0.915 ± 0.009	0.927 ± 0.008	0.927 ± 0.008	0.927 ± 0.008
GDB4	0.893 ± 0.009	0.908 ± 0.009	0.907 ± 0.010	0.907 ± 0.010
MLP8	0.903 ± 0.022	0.918 ± 0.017	0.918 ± 0.017	0.917 ± 0.018
MLP6	0.901 ± 0.014	0.915 ± 0.014	0.914 ± 0.014	0.914 ± 0.014
MLP4	0.882 ± 0.009	0.898 ± 0.008	0.898 ± 0.008	0.897 ± 0.008

KNN8	0.870 ± 0.018	0.888 ± 0.017	0.888 ± 0.016	0.887 ± 0.016
KNN6	0.888 ± 0.014	0.904 ± 0.010	0.904 ± 0.010	0.903 ± 0.010
KNN4	0.890 ± 0.012	0.903 ± 0.011	0.903 ± 0.011	0.902 ± 0.011

[a] balanced accuracy, [b] precision, [c] recall. *8, 6, and 4 – number of features.

It is noted that there were no significant variations in the metrics despite the reduction in the number of variables used in the construction of these models.

The metrics calculated for the testset and sp2019 are summarized in table 13.

Table 13 Performance of the ML models with 8, 6, and 4 features for testset and sp2019.

	Testset				**Sp2019**			
Model	**Ba[a]**	**P[b]**	**R[c]**	**F1**	**Ba[a]**	**P[b]**	**R[c]**	**F1**
SVM8	0.916	0.929	0.929	0.928	0.886	0.921	0.920	0.919
SVM6	0.902	0.918	0.918	0.917	0.898	0.925	0.924	0.924
SVM4	0.877	0.896	0.896	0.895	0.857	0.893	0.891	0.891
GDB8	0.908	0.923	0.924	0.923	0.883	0.916	0.914	0.914
GDB6	0.899	0.916	0.917	0.916	0.885	0.913	0.912	0.912
GDB4	0.888	0.901	0.901	0.900	0.865	0.895	0.894	0.893
MLP8	0.903	0.913	0.912	0.912	0.872	0.911	0.911	0.910
MLP6	0.903	0.912	0.912	0.911	0.879	0.911	0.911	0.910
MLP4	0.852	0.875	0.876	0.874	0.839	0.883	0.882	0.880
KNN8	0.878	0.887	0.888	0.886	0.810	0.864	0.866	0.862
KNN6	0.890	0.898	0.899	0.898	0.866	0.903	0.903	0.902
KNN4	0.876	0.894	0.895	0.893	0.857	0.887	0.885	0.885

[a] balanced accuracy, [b] precision, [c] recall. *8,6 and 4 – feature numbers.

These results also show that there were small variations in these metrics in relation to those obtained with the 9-variable models. With the reduction in the number of variables the KNN algorithm demonstrated a slight improvement in the metric values, while the SVM, GDB presented slightly lower values. In this way, the reduction in the number of variables did not significantly affect the new models obtained.

Regarding the metrics obtained, the results of our models (table 13) do not differ from other models reported in the literature built with 6 variables. [35] worked with the Klang River basin in Malaysia, presenting 2 scenarios: a dataset with 5 monitoring points and a dataset with 19 points. It presented an accuracy of 0.832 for the DT model and 0.891 for the RF, when working with the 5-point dataset; and an accuracy of 0.776 for the DT and 0.840 for the RF, when using the dataset for the 19-point dataset. [36] worked on a case of multi-class classification for the water quality of the Langat River, located in Malaysia, obtained values: 0.894 for DT, 0.912 for MLP, and 0.927 for SVM in the f1 metric.

Another study that reports the use of models trained with 4 variables is the one from [35], who developed an artificial ecosystem optimized by machine learning models used for classification and prediction of water quality, which indicated the following f values: KNN - 0.893, LR - 0.870, DT - 0.859, MLP - 0.848, GNB - 0.896, and SGD - 0.843. With 4 variables, our models were compatible or had a slightly better performance than the results obtained with models in the literature.

Table 14 shows which models have the highest number of hits considering the different levels of water quality.

Table 14 CM performance of the 4 best ML models with 8, 6 and 4 features.

	Testset			**Sp2019**		
Class	***8**	***6**	***4**	***8**	***6**	***4**
0	SVM/KNN	SVM	KNN/MLP	SVM	SVM	GDB
1	GDB	GDB	SVM/GDB	SVM	SVM	KNN
2	SVM	SVM	GDB	SVM/GDB	SVM	SVM/MLP
3	SVM	SVM	SVM	SVM	SVM	SVM
4	KNN	KNN/MLP	GDB	GDB	MLP	KNN

0 – "very bad", 1 – "bad", 2 – "regular", 3 – "good" and 4 – "excellent", *8,6 and 4 – number of features.

Some models are consistent in classifying water, even when the number of parameters has changed. Regarding the testset, it was observed that the SVM model obtained the highest number of correct classifications on 8 occasions; the GDB and KNN models, 5 and 4 times, respectively; and the MLP obtained the highest number of correct classifications on 2 levels. In relation to sp2019, there was a dominance of SVM with 10 appearances, GDB with 3, and KNN and MLP with 2 each.

4 CONCLUSIONS

The proposed objectives of this work were achieved, since, among the 15 trained and evaluated models, it was found that the best models are: SVM, GDB, MLP, and KNN, as they obtained the best metrics.

The variables had their relative importance ordered and the minimum limit of 4 features (tc, bod, do, and tp) was obtained for the WQI classification, in which the loss of quality of information on the classification boundaries is acceptable. This statement is corroborated by the values of the balanced accuracy, precision, recall, and f1 metrics that were observed using the sp2019 dataset.

Thus, with the use of machine learning techniques in modeling water quality classification, it was demonstrated that there was no need to: i) have the 9 variables measured to obtain an WQI; ii) use of the WQI equation or calculation of linearization of variables by range of values and iii) use of average quality variation curves.

The models studied in this work presented a performance equal to or superior to the same machine learning models presented in other works by the various authors mentioned here, when observing the absolute numbers of the statistical metrics. However, it should be noted that the objectives, methods, and database of the works listed here are of a different nature, serving only to corroborate the usefulness of machine learning techniques and not for a direct comparison of results.

Finally, given the increase in data generation in water quality monitoring, the complexity of the system, and the increasing use of machine learning techniques, the present work demonstrated technical feasibility for the use of artificial intelligence in the WQI classification process.

ACKNOWLEDGEMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

AUTHOR CONTRIBUTIONS

Conceptualization, Validation, Formal analysis, Investigation, Writing – Original draft preparation were performed by **Mario Elias Carvalho do Nascimento**. Supervision, Project administration, Funding acquisition were performed by **Ralpho Rinaldo dos Reis**. All authors read and approved the final manuscript.

DATA AVAILABLE

The datasets generated during the current study are available in the Zenodo repository, https://doi.org/10.5281/zenodo.10357787.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ETHICAL APPROVAL

This article does not contain any studies with human participants or animals performed by any of the authors.

REFERENCES

[1]. R. K. Horton, "An Index Number System for Rating Water Quality," *Journal of the Water Pollution Control Federation*, Vol. 37, No. 3, 1965, pp. 300-306.

[2]. Liebman H (1969) Atlas of water quality, methods and practical conditions. Oldenbourg, Munich.

[3]. Brown RM, McClelland NI, Deininger RA, Tozer RG (1970). A water quality index—Do we dare? *Water Sew Works* 117(10):339–343

[4]. Brown RM, McClelland NI, Deininger RA, Landwehr JM (1973). Validating the WQI. The paper presented at national meeting of American society of civil engineers on water resources engineering, Washington, DC.

[5]. L. Prati, R. Pavanello', F. Pesarin, "Assessment of surface water quality by a single index of pollution," (1971).

[6]. Inhaber H (1974) An approach to a water quality index for Canada. Water Res 9:821–833.

[7]. Landwehr JM, Deininger RA (1974) An objective of water quality index. Environ Monit Assess, J *Water Pollut Control Fed* 46(7):1804–1807.

[8]. Landwehr JM (1974) Water quality indices: construction and analysis. PhD thesis, University of Michigan, Ann Arbor, Michigan, USA.

[9]. Smith, D. G. (1990). A better Water Quality Indexing System for Rivers and Streams. In *Wat. Res* (Vol. 24, Issue 10).

[10]. Lumb, A., Sharma, T. C., & Bibeault, J.-F. (2011). A Review of Genesis and Evolution of Water Quality Index (WQI) and Some Future Directions. *Water Quality, Exposure and Health*, *3*(1), 11–24. https://doi.org/10.1007/s12403-011-0040-0.

[11]. Gupta, S., & Gupta, S. K. (2021). A critical review on water quality index tool: Genesis, evolution and future directions. *Ecological Informatics*, *63*. https://doi.org/10.1016/j.ecoinf.2021.101299.

[12]. Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? In *International Journal of Educational Technology in Higher Education* (Vol. 16, Issue 1). Springer Netherlands. https://doi.org/10.1186/s41239-019-0171-0

[13]. Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y., & Shen, D. (2021). Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19. In *IEEE Reviews in Biomedical Engineering* (Vol. 14, pp. 4–15). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/RBME.2020.2987975.

[14]. Mehta, P., Jebarajakirthy, C., Maseeh, H. I., Anubha, A., Saha, R., & Dhanda, K. (2022). Artificial intelligence in marketing: A meta-analytic review. *Psychology and Marketing*, *39*(11), 2013–2038. https://doi.org/10.1002/mar.21716.

[15]. R. S. Peres, X. Jia, J. Lee, K. Sun, A. W. Colombo, and J. Barata, "Industrial Artificial Intelligence in Industry 4.0 -Systematic Review, Challenges and Outlook," *IEEE Access*, 2020, https://doi: 10.1109/ACCESS.2020.3042874.

[16]. Ahmed, N., Abbasi, M. S., Zuberi, F., Qamar, W., Halim, M. S. Bin, Maqsood, A., & Alam, M. K. (2021). Artificial Intelligence Techniques: Analysis, Application, and Outcome in Dentistry - A Systematic Review. In *BioMed Research International* (Vol. 2021). Hindawi Limited. https://doi.org/10.1155/2021/9751564.

[17]. Jiménez-Carvelo, A. M., González-Casado, A., Bagur-González, M. G., & Cuadros-Rodríguez, L. (2019). Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review. In *Food Research International* (Vol. 122, pp. 25–39). Elsevier Ltd. https://doi.org/10.1016/j.foodres.2019.03.063.

[18]. Zounemat-Kermani, M., Batelaan, O., Fadaee, M., & Hinkelmann, R. (2021). Ensemble machine learning paradigms in hydrology: A review. In *Journal of Hydrology* (Vol. 598). Elsevier B.V. https://doi.org/10.1016/j.jhydrol.2021.126266.

[19]. Garcia, H. P. (2012). Desenvolvimento de Estratégias Para Utilização de Sistemas Inteligentes No Monitoramento da Qualidade da Água. Tese (Doutorado) – Universidade Federal de Pernambuco,

Programa de Pós-Graduação em Engenharia Química. https://repositorio.ufpe.br/handle/123456789/11826.

[20]. Marques, L. P. (2018). Modelagem Matemática Para Previsão Parâmetros de Qualidade de Água em Corpos Hídricos. Tese (Doutorado) – Universidade Federal de Pernambuco, Programa de Pós-Graduação em Engenharia Química. https://repositorio.ufpe.br/handle/123456789/32663.

[21]. Morais, C. P., Tadini, A. M., Bento, L. R., Oursel, B., Guimaraes, F. E. G., Martin-Neto, L., Mounier, S., & Milori, D. M. B. P. (2021). Assessing extracted organic matter quality from river sediments by elemental and molecular characterization: Application to the Tietê and Piracicaba Rivers (São Paulo, Brazil). *Applied Geochemistry*, *131*. https://doi.org/10.1016/j.apgeochem.2021.105049.

[22]. Mazzilli, B. P., Lavieri, L. G. S., Soares, J. S., Rocha, F. R., Angelini, M., & Favaro, D. I. T. (2022). Trace and major elements, natural and artificial radionuclides assessment in bottom sediments from Tietê River basin, São Paulo State, Brazil: part III. *Journal of Radioanalytical and Nuclear Chemistry*, *331*(1), 129–144. https://doi.org/10.1007/s10967-021-08094-z.

[23]. S.O.S Mata Atlantica. (2022). *Observando o Tietê 2022 O retrato da qualidade da água e a evolução dos indicadores de impacto do Projeto Tietê*. https://cms.sosma.org.br/wp-content/uploads/2022/09/SOSMAObservando-Tiete_22-1.pdf.

[24]. do Nascimento, M. E., & dos Reis, R. R. (2023). Water Quality Index from Tiete River. *Zenodo*. https://doi.org/10.5281/zenodo.10357787.

[25]. Anaconda, *Anaconda Navigator 2.3.2*. [Online]. Available: https://anaconda.org/anaconda/anaconda-navigator. [Accessed: 2023].

[26]. Jupyter, *Jupyter Notebook 6.4.8*. [Online]. Available: https://jupyter.org/. [Accessed: 2023].

[27]. Python Software Foundation, *Python 3.9.12*. [Online]. Available: https://www.python.org/. [Accessed: 2023].

[28]. Scikit-learn, *Scikit-learn 1.2.2 Documentation*. [Online]. Available: https://scikit-learn.org/stable/index.html. [Accessed: 2023].

[29]. Yellowbrick, *Yellowbrick 1.5 Documentation*. [Online]. Available: https://www.scikit-yb.org/en/latest/index.html. [Accessed: 2023].

[30]. Pandas, *Pandas 2.0.1 Documentation*. [Online]. Available: https://pandas.pydata.org/. [Accessed: 2023].

[30]. Yeo, I.-K., & Johnson, R. A. (2000). A New Family of Power Transformations to Improve Normality or Symmetry (Vol. 87, Issue 4). https://about.jstor.org/terms.

[31]. Se'kou, L., Mosley, D., Gilbert, S., Hofmann, H., & Wang, L. (2013). A balanced approach to the multi-class imbalance problem.

[32]. Xu, T., Coco, G., & Neale, M. (2020). A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning. *Water Research*, *177*. https://doi.org/10.1016/j.watres.2020.115788.

[33]. Wang, L., Zhu, Z., Sassoubre, L., Yu, G., Liao, C., Hu, Q., & Wang, Y. (2021). Improving the robustness of beach water quality modeling using an ensemble machine learning approach. *Science of the Total Environment*, *765*. https://doi.org/10.1016/j.scitotenv.2020.142760.

[34]. Tiyasha, Tung, T. M., & Yaseen, Z. M. (2021). Deep Learning for Prediction of Water Quality Index Classification: Tropical Catchment Environmental Assessment. *Natural Resources Research*, *30*(6), 4235–4254. https://doi.org/10.1007/s11053-021-09922-5.

[35]. Shamsuddin, I. I. S., Othman, Z., & Sani, N. S. (2022). Water Quality Index Classification Based on Machine Learning: A Case from the Langat River Basin Model. *Water (Switzerland)*, *14*(19). https://doi.org/10.3390/w14192939.

[36]. Islam, N., & Irshad, K. (2022). Artificial ecosystem optimization with Deep Learning Enabled Water Quality Prediction and Classification model. *Chemosphere*, *309*. https://doi.org/10.1016/j.chemosphere.2022.136615.

Authors

Mario Elias Carvalho do Nascimento PhD in Environmental Engineering and Technology - PPGETA (2020 -) at the State University of Western Paraná and Federal University of Paraná - UNIOESTE / UFPR. Graduated in Control and Automation Engineering from Centro Universitário Assis Gurgacz (2014). Degree in Electrical Engineering from Centro Universitário Assis Gurgacz (2015). Master's degree in Energy Engineering in Agriculture from the State University of Western Paraná - UNIOESTE - Cascavel (2017 - 2019). https://orcid.org/0000-0003-0961-5207

Ralpho Rinaldo dos Reis holds degree in Chemistry from the State University of Campinas (1987), master's degree in Chemistry from State University of Campinas (1993) and PhD in Agricultural Engineering from the State University of West of Paraná (2013). Currently an Associate Professor at the State University of Western Paraná. Permanent professor of the Postgraduate Program in Agricultural Engineering (Masters and Doctorate) and collaborator of the Postgraduate Program Conservation and Management of Natural Resources (Master's) at State University of Western Paraná. He also works as a permanent professor in the Postgraduate Program in Environmental Engineering and Technology (Masters and Doctorate) at the Federal University of Paraná in association with the State University of Western Paraná. https://orcid.org/0000-0002-2476-2136