PREDICTION OF WATER QUALITY INDEX (WQI) VALUES BY MACHINE LEARNING 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.
mario.nascimento@unioeste.br, ralpho.reis@unioeste.br

Corresponding author: mario.nascimento@unioeste.br

ABSTRACT

This study aimed to explore the use of machine learning algorithms for predicting the Water Quality Index (WQI) values. To achieve this, 11 machine learning algorithms were employed: k nearest neighbors, elastinetCV, linear support vector machine, support vector machine, multilayer perceptron, decision tree, adaboost, bagging, extra trees, gradient boosting, and random forest. The models were statistically evaluated: mean squared error, root mean squared error, and mean absolute error, balanced accuracy, precision, recall, f1, and confusion matrix. A reduction in the number of independent variables from 9 to 4 was also performed. For this reduction, the Spearman correlation technique was used, demonstrating that the most significant variables for predicting WQI were: thermotolerant coliforms, biochemical oxygen demand, dissolved oxygen, and total phosphorus. Thus, this study showed that WQI can be predicted using machine learning techniques trained with only 4 independent variables, without significant differences from traditionally determined index values.

KEYWORDS: *Index Water Quality; Machine Learning; Classical Models; Ensemble Models; Prediction.*

I. INTRODUCTION

Regression techniques are widely used and important in the study of natural phenomena. However, for their use, certain prerequisites must be met linearity, homoscedasticity, independence of errors, and non-multicollinearity. Conventional prediction modeling methods present significant limitations, with dependencies on the sampling and databases used [1]. Additionally, [2] stated that some statistical models suffer due to the need for linearity of the regression coefficient when used with nonlinear variables, thus affecting the predictive power of the models.

Modeling water quality becomes challenging due to the number of physical, chemical, and biological parameters involved in the study of this substance. Due to its value, there is a need for conservation and monitoring. The relevance of water quality monitoring, prediction, and evaluation are fundamental for the management of rivers and their resources [1]. Consequently, the impact of its degradation is observed in the economic, social, and health fields. The use of regression techniques for this purpose is hindered by the aforementioned factors. For this reason, several tools have been developed over the years, with the most representative and easily understandable being the Water Quality Index.

The development and improvement of water quality indices began in the mid-1960s. Notable researchers in this field include [3], the first to present water quality as a numerical index; [4], who presented a map representing the water quality of the river in the Bavaria region in Germany with colors; [5, 6], who applied and developed Horton and Ledman's ideas in the first index used by government agencies in the United States of America.

DOI: 10.5281/zenodo.10862055

In the subsequent decades, various quality indices were implemented or modified according to application needs. In 1971, [7] developed an index indicating organic pollution in surface waters; [8] suggested an index with 5 quality classes for water; [9] introduced an index classifying waters for recreational use between values 0 and 1; [10] developed an index using a rank with non-parametric variables in its methodology; [11] presented the water quality index for the state of Oregon; [12] enhanced the environmental quality index. Many other indices and their applications can be found in review works by [13, 14].

Machine learning (ML) and artificial intelligence (AI) models offer flexibility that allows better adaptation to non-linear data, producing more adjusted predictions when compared to classical techniques [15]. In recent years, AI has rapidly developed and been applied in various areas of knowledge. Works such as: [14] using AI techniques in the field of education; [16] In the health area; [17] in marketing; [18] in Industry 4.0; [19] in the food industry and with [20] in hydrology.

Within the field of hydrology, the use of AI techniques, especially machine learning algorithms, still has plenty of room for development and application. Models with a large number of parameters become challenging to solve due to their complexity and non-linearity, requiring significant computational power and the application of stochastic methods. The application of ML techniques has been solving many of these problems [21]. Many models for predicting WQI are sensitive to the dataset used for modeling [22]. Water quality prediction becomes more accurate when using deep learning and ML techniques, thanks to their treatment of nonlinear and unstable variables [23]. Thus, ML models can achieve great success in predicting hydrological processes quickly, accurately, and with easy interpretation [24].

In the case of Brazil, particularly in the state of São Paulo, efforts for monitoring water quality have been ongoing since 1975. In that year, the Environmental Company of the State of São Paulo (CETESB) implemented the surface water quality index. One of the state's most important rivers is the Tietê River. According to [25], it holds significance in the food industry, transportation, navigation, and electricity generation, significantly impacting the state's economic development. It is also highlighted that this river crosses the state with the highest population density and GDP in the country, presenting degradation characteristics in line with the economic and urban influence of the region it traverses, providing a general view of degradation present in other state rivers.

The objective of this work was to predict the WQI of the Tietê River using 11 machine learning algorithms to analyze and statistically compare their performances, identifying which ones perform better and could be used for water quality index prediction. Additionally, it also aimed to reduce the number of parameters to obtain an optimized prediction.

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 regression and the fourth the conclusion of the work.

II. MATERIALS AND METHODS

Figure 1, illustrates the stages of the work, starting from data collection conducted in reports available on the internet, through training and testing with two assembled datasets, and finally, the phase of evaluating the performance of the models using regression and classification metrics.

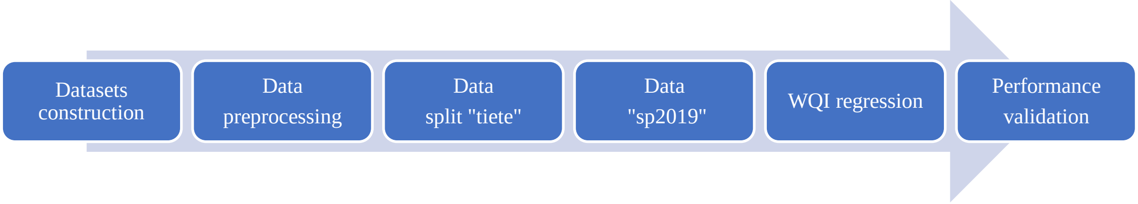

Figure 1. Overall process of working.

2.1. Characterization of the Study Area

The Tietê River is an intrastate river located in the state of São Paulo, crossing it from East to West with a length of 1,100 km, as shown in Figure 2. The river is divided into 6 sub-basins: Tietê/Batalha, Lower Tietê, Upper Tietê, Piracicaba, Sorocaba/Middle Tietê, and Tietê/Jacaré.

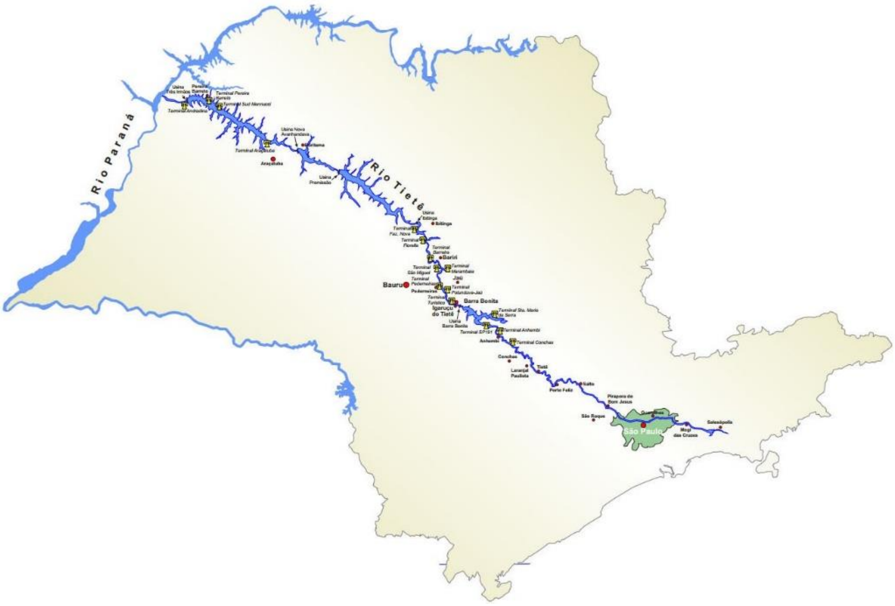

Figure 2. Tietê River

2.2. Construction of Datasets

For the dataset called “tiete”, time series data from 1994 to 2019 for 9 physical, chemical, and biological parameters were collected: 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) at the 78 measurement points that are located on the Tietê River or the closest points in the flow locations of the tributaries that make up the river basins. For the dataset named "sp2019," data from 2019 for all measurement points in all hydrographic basins of the state of São Paulo were used, excluding points that are part of the "tiete" dataset.

The aforementioned data can be accessed through the CETESB link (https://cetesb.sp.gov.br/aguas-interiores/publicacoes-e-relatorios/), where annual reports on surface water quality (WQI) and other indices used by the company are available.

The WQI used by the company is represented by a value with a range of 0 to 100. It has 5 water quality classes: “terrible” ranges from 0 and ≤ 19, represented by the color purple; “bad” ranges from > 19 and ≤ 36, marked by the color red; “regular” is classified for values > 36 and ≤ 51, characterized by the color yellow; “good” falls between values > 51 and ≤ 79, with the color green; and “excellent” for WQI greater than > 79, with the color blue being used.

Using the data cleaning techniques (handling missing and inconsistent data), the “tiete” dataset totaled 7101 samples and the “sp2019” dataset obtained 2332 samples. The “tiete” dataset can be obtained at the link (https://zenodo.org/records/10357787) [26].

2.3. Programs Used, Outlier Detection and Data Transformation

2.3.1. Programs used

The simulations were developed using Anaconda Navigator 2.3.2 (https://anaconda.org/anaconda/anaconda-navigator), Jupyter Notebook 6.4.8 (https://jupyter.org/), Python 3.9.12 (https://www.python.org/), Scikit-learn 1.2.2 (https://scikit-learn.org/stable/index.html#), Yellowbrick 1.5 (https://www.scikit-yb.org/en/latest/index.html), and Pandas 2.0.1 (https://pandas.pydata.org/).

2.3.2. Outlier detection

The Isolation Forest algorithm was used for outlier detection, with all hyperparameters kept as default. The "tiete" dataset had 7101 samples, reduced to 6545 after applying the technique. The dataset was then split into 80/20 for training and testing. The trainset had 5236 samples (80%), and the testset had 1309 samples (20%).

2.3.3. Data transformation

Normally, when working with machine learning, there is a need to standardize different variables into scales. In this case, we choose to use data transformation techniques, with the most common being: min-max and z-score. In this work, the Yeo-Johnson transformation technique was used to perform the data transformation, as this family of power transformation presents a good response to negative data or zeros [27].

2.4. Artificial Intelligence Techniques

2.4.1. Classical machine learning techniques for regression

The techniques used were: KNN – K Nearest Neighbors, ElastiNetCV, LSVM – Linear Support Vector Machine, SVM – Support Vector Machine, MLP – Multilayer Perceptron, DT – Decision Tree. The parameters of each technique were kept as default, as the optimization of the hyperparameters did not yield relevant results to the study, except for SVM, which in relation to parameter C, which indicates a regularization penalty, was modified from value 1 to value 20. For DT, the maximum depth parameter (max_depth) was set to 3 to avoid excessive growth leading to overfitting, and ccp_alpha, an internal minimal cost-complexity parameter, was changed from 0 to 0.150. Further information about the parameters and functionality of the techniques can be obtained by accessing the scikit-learn library documentation on the website (https://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model).

2.4.2. Ensemble machine learning techniques for regression

The machine learning ensemble techniques were ADA – AdaBoost, BAG – Bagging, ET – Extra trees, GDB – Gradient Boosting, and RF – Random Forest. The hyperparameters of the algorithms were kept as default. For more information, refer to the scikit-learn library link (https://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model).

2.5. Evaluation Models

The metrics used to compare the regressors were: Mean Squared Error (MSE), equation 1, which calculates the average between all points and the regression model.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_r - Y_p\right)^2 \tag{1}$$

The Root Mean Squared Error (RMSE), equation 2, which calculates the square root of the average of the errors between the predicted values and the actual values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_r - Y_p\right)^2} \tag{2}$$

The Mean Absolute Error (MAE), equation 3, which calculates the mean absolute error between the actual values and the predicted values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_r - Y_p\right| \tag{3}$$

And the classification metrics were Precision, equation 4, which 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).

$$P = \frac{T_p}{T_p + F_p} \tag{4}$$

Recall, equation 5, is the ratio of true positives (T_p) to the sum of false negatives (F_n) and true positives (T_p).

$$R = \frac{T_p}{T_p + F_n} \tag{5}$$

F1, equation 6, is the harmonic mean between precision and recall.

$$F1 = 2\ x\ \frac{P\ x\ R}{P + R} \tag{6}$$

Balanced accuracy, equation 7, 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 [28].

$$BA = \frac{1}{2}\left(\frac{T_p}{T_p + F_n} + \frac{T_n}{T_n + F_p}\right) \tag{7}$$

III. RESULTS AND DISCUSSION

3.1. Descriptive Statistics of the Datasets

Tables 1 and 2 present the statistical summary of the variables for each dataset. The "tiete" dataset was treated with outlier removal, while the "sp2019" dataset had no treatment, only the removal of missing or inconsistent values.

Table 1. 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

Table 2. 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	2	332
tc (NMP/100mL)	371620	2444414	2	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

3.2. Evaluation of Models Using 9 Variables

3.2.1. Trainset (overall and cross-validation)

Table 3 summarizes the values of the MSE, RMSE, and MAE metrics of the ML algorithms for the trainset, as well as their performance when using the 10-fold cross-validation technique.

Table 3. Performance of ML models for the “tiete” dataset trainset

	Overall			**Cross-validation (10 folds)**		
Model	**Mse**	**Rmse**	**Mae**	**Mse**	**Rmse**	**Mae**
KNN	9.084	3.014	2.011	13.939 ± 2.835	3.716 ±0.363	2.503 ± 0.145
ElasticNet	20.614	4.540	3.483	20.729 ± 2.625	4.544 ± 0.279	3.489 ± 0.071
LSVM	21.223	4.607	3.416	21.298 ± 2.811	4.606 ± 0.295	3.427 ± 0.082
SVM	4.819	2.195	0.963	5.847 ± 2.459	2.373 ± 0.464	1.171 ± 0.086
MLP	5.159	2.271	1.234	6.336 ± 2.467	2.477 ± 0.448	1.368 ± 0.102
DT	51.054	7.147	5.141	57.616 ± 4.020	7.586 ± 0.264	5.404 ± 0.157
ADA	39.518	6.286	5.274	39.249 ± 1.621	6.264 ± 0.130	5.204 ± 0.145
BAG	3.468	1.862	1.238	15.528 ± 2.643	3.927 ± 0.328	2.750 ± 0.103
ET	1.018	1.009	0.540	6.724 ± 2.582	2.553 ± 0.451	1.432 ± 0.101
GDB	5.223	2.285	1.444	7.808 ± 2.324	2.767 ± 0.392	1.674 ± 0.077
RF	38.229	6.183	4.637	39.456 ± 2.347	6.279 ± 0.188	4.708 ± 0.102

The Bagging and Extra Tree algorithms exhibited the two best performances when considering the overall metric values. However, the same is not reflected when analyzing these algorithms with cross-validation techniques. Small values are observed when using the overall MSE metric (3.468 and 1.018), and values five times higher (15.528 ± 2.643 and 6.724 ± 2.582) after applying cross-validation for metric calculation. This characteristic reflects an overfitting of the ML techniques to the trainset data.

The SVM algorithm obtained the best absolute values for the metrics both overall and in cross-validation. However, there is a difference of more than two times between the RMSE and MAE values, both overall (2.195 and 0.963) and in cross-validation (2.373 ± 0.464 and 1.171 ± 0.086), which could indicate the presence of samples with very high variability in the trainset, thus complicating the algorithm's regression process.

The MLP and GB algorithms demonstrated interesting performances. With small RMSE and MAE values and a ratio between them less than twice both for the metrics in general: MLP (2.271 and 1.234) and GB (2.285 and 1.444) and in the cross-validation: MLP (2.477 ± 0.478 and 1.368 ± 0.102) and GB (2.767 ± 0.392 and 1.674 ± 0.077), thus indicating small variances. This result is expected when working with real-world data.

3.2.2. Testset and sp2019

The final evaluation of ML models using 9 variables occurred in two stages: first, using the testset of the "tiete" dataset, and second, using the "sp2019" dataset. Tables 4 and 5 summarize the performance of ML algorithms.

Table 4. Performance of ML models for the “tiete” dataset testset.

Model	**Mse**	**Rmse**	**Mae**	**Accuracy**	**Precision**	**Recall**	**F1**
KNN	12.662	3.558	2.403	0.882	0.892	0.892	0.892
ElasticNet	18.769	4.332	3.380	0.809	0.847	0.833	0.834
LSVM	19.516	4.418	3.347	0.827	0.849	0.837	0.837
SVM	5.145	2.268	1.142	0.936	0.941	0.940	0.940
MLP	5.365	2.316	1.312	0.923	0.931	0.930	0.931
DT	50.744	7.123	5.239	0.750	0.765	0.788	0.765
ADA	39.752	6.305	5.267	0.497	0.465	0.633	0.536
BAG	14.398	3.795	2.732	0.844	0.872	0.869	0.868
ET	5.750	2.398	1.398	0.921	0.932	0.931	0.932
GDB	6.875	2.622	1.656	0.906	0.922	0.922	0.921
RF	35.289	5.940	4.540	0.817	0.826	0.826	0.822

Table 5. Performance of ML models for the "sp2019" dataset.

Model	Mse	Rmse	Mae	Accuracy	Precision	Recall	F1
KNN	23.326	4.830	3.436	0.805	0.862	0.865	0.862
ElasticNet	19.089	4.369	3.203	0.800	0.868	0.869	0.863
LSVM	18.990	4.358	3.046	0.819	0.872	0.874	0.869
SVM	4.809	2.193	1.167	0.902	0.948	0.948	0.948
MLP	5.529	2.351	1.418	0.902	0.940	0.940	0.940
DT	68.481	8.275	6.171	0.711	0.763	0.777	0.728
ADA	48.947	6.996	5.752	0.482	0.636	0.755	0.690
BAG	18.740	4.329	3.261	0.819	0.888	0.887	0.885
ET	7.968	2.823	1.796	0.873	0.927	0.926	0.926
GDB	7.772	2.788	1.832	0.864	0.917	0.917	0.917
RF	46.337	6.807	5.219	0.797	0.833	0.837	0.834

The GB, MLP, and SVM algorithms demonstrated the best performance on the testset of the "tiete" dataset and the "sp2019" dataset. They achieved RMSE values below 3.0 and MAE values below 2.0 in both datasets. It should also be noted that all algorithms used have RMSE/MAE ratio values below 2.0. The BAG and ET algorithms presented MSE values compatible with the MSE values in cross-validation, but still 5 times higher than the overall MSE values.

Table 6 summarizes the maximum accuracy value of the ML algorithms per water quality class for both datasets.

Table 6. Performance of ML models for the CM technique.

	Trainset					Sp2019				
Model	0	1	2	3	4	0	1	2	3	4
KNN	197	308	119	413	131	77	163	251	1320	205
ElasticNet	185	245	125	434	102	87	139	241	1358	201
LSVM	189	237	125	428	117	87	138	241	1340	231
SVM	200	327	145	424	134	72	180	340	1361	257
MLP	196	323	139	427	133	78	181	329	1358	246
DT	198	285	29	387	132	74	181	14	1305	238
ADA	0	279	118	432	0	0	144	277	1338	0
BAG	163	310	122	40	122	75	165	278	1346	203
ET	198	321	140	431	129	73	169	331	1349	237
GB	197	326	125	428	131	74	174	305	1350	235
RF	200	281	86	382	132	76	163	201	1267	244
Total	211	348	165	446	139	91	204	374	1392	270

0 – "terrible", 1 – "bad", 2 – "regular", 3 – "good" and 4 – "excellent".

Among the top-performing algorithms (SVM, MLP, GB), SVM obtained the highest number of correct predictions for 4 water classes ("terrible", "bad", "regular" and "excellent"), while GB performed better only in the "good" class using the testset. In the validation with the "sp2019" dataset, the SVM obtained more correct predictions for the "regular", "good" and "excellent" classes and the MLP got more hits for the "terrible" and "bad" classes.

3.3. Evaluation of models after variable reduction

The Spearman correlation technique was employed to assess the relationship between the independent variables and the dependent variable. Figure 3 presents the Spearman correlation coefficient between the response variable and the independent variables. It was observed that the variables: "tc", "bod", "tn" and "tp" show a strong negative correlation, while "do" has a strong but positive correlation with the response variable "WQI". And the variables "ph" and "temp" have a very weak positive correlation with the "WQI".

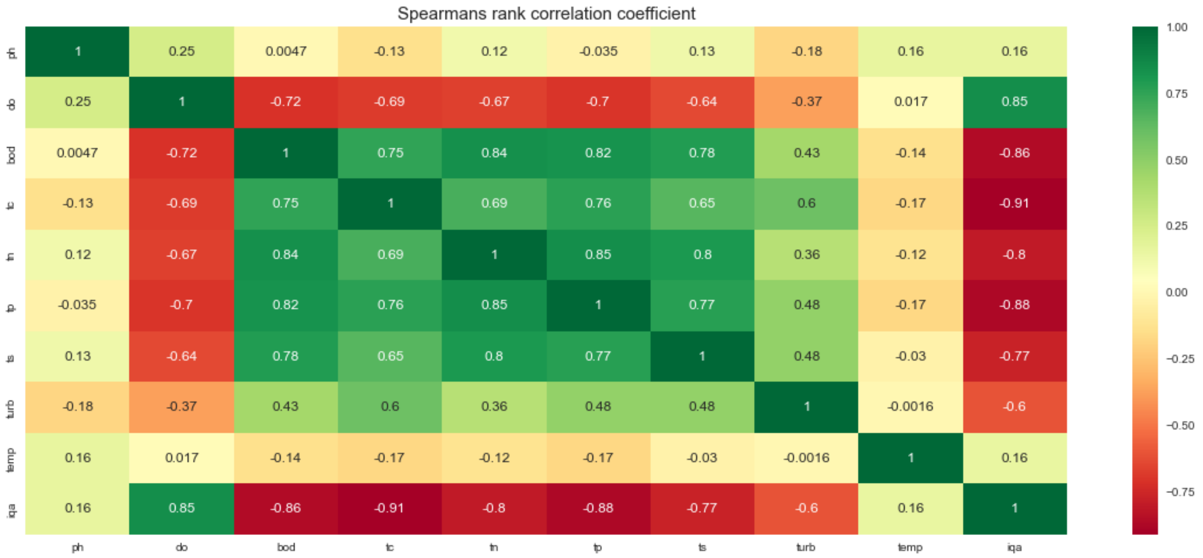

Figure 3. Spearman's rank correlation.

3.3.1. Regressors with: 7 (tc, bod, do, tp, turb, tn, ts), 5 (tc, bod, do, tp, tn), and 4 (tc, bod, do and tp) features.

The performances for the SVM, MLP, and GD models of the MSE, RMSE and MAE metrics are summarized in table 7. The metrics values were obtained for regressors with 7 features (excluding "ph" and "temp"), with 5 features (excluding "ph", "temp", "turb", "ts"), and with 4 features ("tc", "bod", "do" and "tp" that have the strongest correlations).

Table 7. Performance of 7-, 5-, and 4-feature ML models for the trainset.

Model	Mse	Rmse	Mae	Mse	Rmse	Mae
SVM^7	5.566	2.359	1.181	6.262 ± 1.938	2.475 ± 0.368	1.309 ± 0.068
SVM^5	9.154	3.026	1.676	9.624 ± 2.148	3.086 ± 0.319	1.747 ± 0.083
SVM^4	9.505	3.083	1.735	9.764 ± 2.053	3.110 ± 0.302	1.782 ± 0.085
MLP^7	5.843	2.417	1.411	6.757 ± 2.296	2.568 ± 0.406	1.472 ± 0.064
MLP^5	9.219	3.036	1.860	9.876 ± 2.003	3.129 ± 0.297	1.896 ± 0.066
MLP^4	9.748	3.122	1.946	10.144 ± 1.931	3.172 ± 0.284	1.929 ± 0.087
GDB^7	5.661	2.379	1.512	8.059 ± 2.233	2.815 ± 0.370	1.727 ± 0.068
GDB^5	8.196	2.863	1.862	11.019 ± 1.905	3.308 ± 0.275	2.099 ± 0.066
GDB^4	8.417	2.901	1.877	11.052 ± 1.893	3.314 ± 0.269	2.088 ± 0.073

*7, 5, and 4 – number of features.

It is noted that there was an increase in the metric values, but this was expected since, with the reduction of variables, the regressors exhibit a decrease in the ability to describe the phenomenon. Despite the increase in value, it was very small and not significant, still obtaining an acceptable regression with 4 variables.

The metrics calculated for the testset and sp2019 are summarized in tables 8 and 9.

Table 8. Performance of 7-, 5-, and 4-feature ML models for the testset.

Model	Mse	Rmse	Mae	Accuracy	Precision	Recall	F1
SVM^7	5.415	2.327	1.255	0.926	0.932	0.932	0.932
SVM^5	8.386	2.896	1.749	0.911	0.918	0.916	0.916
SVM^4	8.647	2.941	1.783	0.911	0.919	0.917	0.917
MLP^7	5.919	2.433	1.460	0.917	0.929	0.929	0.929
MLP^5	8.668	2.944	1.909	0.891	0.907	0.906	0.906
MLP^4	9.070	3.012	2.011	0.888	0.907	0.905	0.906
GDB^7	7.140	2.672	1.711	0.907	0.922	0.923	0.922
GDB^5	9.665	3.109	2.049	0.890	0.904	0.904	0.904
GDB^4	9.730	3.119	2.071	0.891	0.905	0.905	0.905

*7, 5, and 4 – number of features

Table 9. Performance of 7-, 5-, and 4-feature ML models for the sp2019.

Model	Mse	Rmse	Mae	Accuracy	Precision	Recall	F1
SVM7	6.132	2.476	1.429	0.904	0.943	0.943	0.942
SVM5	10.490	3.239	2.153	0.886	0.916	0.916	0.916
SVM4	10.833	3.291	2.215	0.877	0.912	0.913	0.912
MLP7	7.413	2.723	1.875	0.878	0.924	0.924	0.924
MLP5	11.910	3.451	2.50	0.869	0.910	0.909	0.909
MLP4	12.071	3.474	2.602	0.856	0.902	0.901	0.901
GDB7	8.838	2.973	2.018	0.863	0.914	0.914	0.914
GDB5	13.239	3.638	2.554	0.844	0.897	0.897	0.896
GDB4	13.799	3.715	2.601	0.846	0.896	0.897	0.895

*7, 5, and 4 – number of features.

Table 10 shows the accuracy per water quality class for the models developed.

Table 10. CM performance of 7-, 5-, and 4-feature models.

	Testset					Sp2019				
Model	0	1	2	3	4	0	1	2	3	4
SVM7	200	321	139	426	134	75	182	338	1351	251
SVM5	195	317	135	418	134	79	179	296	1329	253
SVM4	196	318	135	418	133	78	179	291	1333	247
MLP7	191	321	141	434	129	77	180	324	1350	223
MLP5	190	321	129	420	126	75	186	292	1331	236
MLP4	192	321	127	421	124	75	179	293	1326	228
GDB7	197	324	127	430	130	74	176	305	1344	232
GDB5	195	317	122	419	130	73	173	277	1332	236
GDB4	196	314	124	421	129	75	173	275	1334	233
Total	211	348	165	446	139	91	204	374	1392	270

0 – "terrible", 1 – "bad", 2 – "regular", 3 – "good" and 4 – "excellent", *7, 5, and 4 – number of features, green (win) and red (draw).

For the SVM, MLP, and GB models, the accuracy varied similarly. Regarding the classes, there was a predominance of the SVM algorithm for the "excellent" class in both datasets. In the testset, the SVM algorithm obtained the highest number of correct predictions for each class in 6 instances, the MLP in 5, the GB in 1, and there were 3 draws. For sp2019, the SVM algorithm achieved the highest number of correct predictions 9 times, MLP got 3, GB got 2, and there was 1 draw.

3.4. Comparing with literature

Considering the results obtained in the literature in other studies, it is observed that [29] used ML algorithms for classification and prediction of water quality index. The two bests were RF and GB, with F1 metrics of 0.50 and 0.53 for RF and GB. MAE and RMSE were RF (2.30 and 3.09) and GB (1.96 and 2.68), respectively. [21] presented a study on multiple regression, with total dissolved solids and electrical conductivity as dependent variables and 8 independent variables. The MAE and RMSE metrics of the algorithms were, respectively: ANN (9.56 and 11.76), MLR (8.22 and 8.25), and MNLR (6.67 and 12.76).

Finally, [30], worked with 16 ML algorithms to implement Iran's water quality index. The two bests were the hybrid algorithms: bagging-RF with MAE (1.51) and RMSE (2.78), and bagging-RT with MAE (1.87), and RMSE (2.71). [22] developed an ecosystem optimized with deep learning for classifying and predicting water quality. The RMSE values obtained by the algorithms used in the prediction were: ANN (0.7158); LR (2.7129); RF (2.9155); GB (2.6134); and SVM (2.6230). The algorithms used to classify the F1 values were: MLP (84.83); KNN (89.33); DT (85.99); and LR (87.04). [31] studied the impact of soil cover on groundwater quality. Two algorithms were trained, RF and ANN. Accuracy for the RF ranged between 0.81 and 0.787; MAE, 0.343 and 0.334; and RMSE, 0.397 and 0.387. As for the ANN algorithm, the variation ranged from 0.60 to 0.775 for accuracy, from 0.43 to 0.372 for MAE, and from 0.459 to 0.427 for RMSE.

The GB, MLP, and SVM algorithms also had the best performance when observing the classification metrics, F1, and balanced accuracy, with values above 0.90 for both metrics in the testset. For the sp2019 dataset of the 3 algorithms, only GB obtained a value below 0.90 in the accuracy balanced metric.

It is observed that the results obtained in the datasets ("testset" and "sp2019") did not show a significant variation in values when there was a reduction in variables. Regarding classification metrics, they mostly remained close to 0.90, except for balanced accuracy, which approached 0.85 when the number of variables was reduced. Thus, the reduction in the number of variables did not significantly affect the new models obtained.

Comparing the values obtained in this study (tables 4, 5, 8, and 9) with the values observed in the bibliography, it is noted that even with the reduction of variables from 9 to 4, it is still possible to predict the water quality index within the standard of metrics obtained in other studies. In this way, the SVM, MLP, and GB algorithms can be used as an alternative to predict the WQI of surface waters.

IV. CONCLUSION

It is concluded that the best algorithms presented in this study for predicting the water quality index were: SVM, MLP, and GB. The three achieved equivalent statistical performance. They showed robustness in predicting the WQI, even when subjected to variable reduction, as the difference between the metrics was not significant. Therefore, one of these ML algorithms, trained with 4 variables, can be used, provided they are the ones with the strongest correlations with the response variable. Finally, it was found that AI algorithms are useful tools in predicting WQI values, as they are robust in interpreting phenomena that deviate from linearity.

ACKNOWLEDGEMENTS

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

REFERENCES

[1]. Najah Ahmed, A., Binti Othman, F., Abdulmohsin Afan, H., Khaleel Ibrahim, R., Ming Fai, C., Shabbir Hossain, M., Ehteram, M., & Elshafie, A. (2019). Machine learning methods for better water quality prediction. *Journal of Hydrology*, *578*. https://doi.org/10.1016/j.jhydrol.2019.124084

[2]. 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

[3]. 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.

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

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

[6]. 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

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

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

[9]. Walski TM, Parker FL (1974) Consumer's water quality index. *Journal* Environ Eng Div ASCE 100(3):593–611

[10]. Harkins RD (1974). An objective water quality index. *Journal Water Pollut Control Fed* 46(3):588–591

[11]. Dunnette DA (1979) A geographically variable water quality index used in Oregon. *Journal Water Pollut Control Fed* 51(1):53–61

[12]. Steinhart CE, Schierow LJ, Sonzogni WC (1982). An environmental quality index for the Great Lakes. *Water Resour Bull* 18(6):1025–1031

[13]. 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

[14]. 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

[15]. Ho, J. Y., Afan, H. A., El-Shafie, A. H., Koting, S. B., Mohd, N. S., Jaafar, W. Z. B., Lai Sai, H., Malek, M. A., Ahmed, A. N., Mohtar, W. H. M. W., Elshorbagy, A., & El-Shafie, A. (2019). Towards a time and cost effective approach to water quality index class prediction. *Journal of Hydrology*, *575*, 148–165. https://doi.org/10.1016/j.jhydrol.2019.05.016

[16]. 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

[17]. 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

[18]. 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

[19]. 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

[20]. 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

[21]. Shah, M. I., Javed, M. F., & Abunama, T. (2020). Proposed formulation of surface water quality and modelling using gene expression, machine learning, and regression techniques. https://doi.org/10.1007/s11356-020-11490-9/Published

[22]. 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

[23]. 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

[24]. Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., & Al-Shamma’a, A. (2022). Water quality classification using machine learning algorithms. *Journal of Water Process Engineering*, *48*. https://doi.org/10.1016/j.jwpe.2022.102920

[25]. 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

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

[27]. 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

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

[29]. Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. *Water (Switzerland)*, *11*(11). https://doi.org/10.3390/w11112210

[30]. Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of the Total Environment*, *721*. https://doi.org/10.1016/j.scitotenv.2020.137612

[31]. Anjum, R., Ali, S. A., & Siddiqui, M. A. (2023). Assessing the Impact of Land Cover on Groundwater Quality in a Smart City Using GIS and Machine Learning Algorithms. *Water, Air, and Soil Pollution*, *234*(3). https://doi.org/10.1007/s11270-023-06198-8

AUTHORS

Mario Elias Carvalho do Nascimento PhD candidate in the Postgraduate Program 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