water

MDPI

Article

Simple Prediction of an Ecosystem-Specific Water Quality Index and the Water Quality Classification of a Highly Polluted River through Supervised Machine Learning

Alberto Fernández del Castillo [1,†], Carlos Yebra-Montes [2,†], Marycarmen Verduzco Garibay [1], José de Anda [3], Alejandro Garcia-Gonzalez [4,*] and Misael Sebastián Gradilla-Hernández [1,*]

1 Tecnologico de Monterrey, Escuela de Ingenieria y Ciencias, Av. General Ramon Corona 2514, Nuevo México, Zapopan CP 45138, Jalisco, Mexico; a01273977@itesm.mx (A.F.d.C.); a01228191@itesm.mx (M.V.G.)
2 ENES-León, Universidad Nacional Autónoma de México, Blvd. UNAM 2011, Predio el Saucillo y El Potrero, León CP 37684, Guanajuato, Mexico; carlosyebra@comunidad.unam.mx
3 Unidad de Tecnología Ambiental, Centro de Investigación y Asistencia en Tecnología y Diseño del Estado de Jalisco, A. C. Av. Normalistas 800, Colinas de la Normal, Guadalajara CP 44270, Jalisco, Mexico; janda@ciatej.mx
4 Tecnologico de Monterrey, Escuela de Medicina y Ciencias de la Salud, Av. General Ramon Corona 2514, Nuevo Mexico, Zapopan CP 45138, Jalisco, Mexico
* Correspondence: alexgargo@tec.mx (A.G.-G.); msgradilla@tec.mx (M.S.G.-H.)
† These authors contributed equally to this work.

Abstract: Water quality indices (WQIs) are used for the simple assessment and classification of the water quality of surface water sources. However, considerable time, financial resources, and effort are required to measure the parameters used for their calculation. Prediction of WQIs through supervised machine learning is a useful and simple approach to reduce the cost of the analysis through the development of predictive models with a reduced number of water quality parameters. In this study, regression and classification machine-learning models were developed to estimate the ecosystem-specific WQI previously developed for the Santiago-Guadalajara River (SGR-WQI), which involves the measurement of 17 water quality parameters. The best subset selection method was employed to reduce the number of significant parameters required for the SGR-WQI prediction. The multiple linear regression model using 12 parameters displayed a residual square error (RSE) of 3.262, similar to that of the multiple linear regression model using 17 parameters (RSE = 3.255), which translates into significant savings for WQI estimation. Additionally, the generalized additive model not only displayed an adjusted R^2 of 0.9992, which is the best fit of all the models evaluated, but also fitted the rating curves of each parameter developed for the original algorithm for the SGR-WQI calculation with great accuracy. Regarding the classification models, an overall proportion of 93% and 86% of data were correctly classified using the logistic regression model with 17 and 12 parameters, respectively, while the linear discriminant functions using 12 parameters correctly classified an overall proportion of 84%. The models evaluated were found to be efficient in predicting the SGR-WQI with a reduced number of parameters as complementary tools to extend the current water quality monitoring program of the Santiago-Guadalajara River.

Keywords: water quality index prediction; regression and classification algorithms; Santiago-Guadalajara River

Citation: Fernández del Castillo, A.; Yebra-Montes, C.; Verduzco Garibay, M.; de Anda, J.; Garcia-Gonzalez, A.; Gradilla-Hernández, M.S. Simple Prediction of an Ecosystem-Specific Water Quality Index and the Water Quality Classification of a Highly Polluted River through Supervised Machine Learning. *Water* **2022**, *14*, 1235. https://doi.org/10.3390/w14081235

Academic Editors: Zheng Duan and Babak Mohammadi

Received: 10 March 2022
Accepted: 7 April 2022
Published: 12 April 2022

1. Introduction

Rivers all around the world are a vital source of freshwater for the environment, the economy, and society [1–3]. However, these freshwater ecosystems are currently affected by several problems such as overexploitation, climate change, and anthropogenic pollution [4]. The water quality of rivers is affected by natural (rainfall and erosion) and anthropogenic processes (such as urbanization, agriculture, and manufacturing) [5–7].

Water quality is a nonlinear and nonstationary phenomenon, which encompasses complex relationships between natural and anthropogenic processes and, thus, continuous water quality monitoring is fundamental to develop strategies to remediate and preserve rivers and maintain their sustainable management [8]. However, monitoring programs require numerous measurements of different water quality parameters at different sampling points and at different sampling times, and, consequently, large databases are generated [9]. Thus, complex analyses and interpretation are necessary to assess the environmental situation of a waterbody [10,11]. To overcome this problem, water quality indices (WQIs) are recognized as useful tools to ease water quality visualization, interpretation, and communication [12]. WQIs have been widely applied to assess and classify the water quality of surface and groundwater sources and to help the managers of water resources to make more effective decisions [13–16]. WQIs are calculated by the integration of multiple physical, chemical, or biological parameters specifically selected for their significance to the water quality of water sources in general [17,18]. Moreover, recent efforts have focused on developing ecosystem-specific water quality indices developed for the specific conditions of a given water source and following local limits/standards/guidelines for the protection of aquatic life [19,20]. These ecosystem-specific WQIs are developed to adequately reflect spatial and temporal variations in agreement with the local territorial context [21,22].

Because considerable time, effort, and financial resources are required to measure the water quality indicators (physical, chemical, and biological) that are included in WQIs' algorithms, there is a need for practical computational approaches to estimate WQIs accurately and efficiently [23]. The costs associated with water quality monitoring can be reduced if WQIs can be estimated based on smaller sets of water quality parameters and, consequently, water quality monitoring could be extended to wider catchment areas of the water sources in question or to monitor other water bodies.

Machine learning (ML) has been extensively applied for the monitoring and control of several engineering processes using several algorithms such as linear regression, logistic regression, decision trees, vector support machines, and artificial neural networks, among others. These methods are used to perform data regression and classification tasks, which aim to determine the association between a set of variables called input variables (regressors) or characteristics (features) with an output variable (prediction). In the case of regression algorithms, a numerical value of the output variable is predicted, and, in the case of classification algorithms, this output is a categorical variable. ML algorithms must be adjusted to execute such tasks. Initially, a step known as training is performed by changing the values of the predictor variables, considering the goal of reducing the error between the actual and the predicted value of the output variable. A training data set is used for this purpose and the greater the amount of data available for the task, the better the performance of the prediction algorithms. Once that error-minimizing model is obtained, it is tested using the testing dataset, which should be different from the training dataset [24].

The objective of this study was to evaluate the potential use of simple structure supervised ML methods to predict the ecosystem-specific WQI previously developed for the Santiago-Guadalajara River (SGR-WQI) in Mexico [20]. This WQI is currently used by the local government of Jalisco, Mexico, to communicate the water quality status and trends to the general society [25]. The main contribution of this research is the use of supervised ML algorithms (multiple linear regression and generalized additive models for regression tasks, and logistic regression and linear discriminant analysis for classification tasks) for the prediction of the SGR-WQI, using less water quality parameters to ease the time and costs associated with water quality monitoring and, consequently, to extend the number of sampling points that are regularly monitored in this large basin.

2. Literature Review

Both supervised, with a priori knowledge of the actual value or the class of the output variable, and unsupervised training methods, which focus on pattern recognition

without the involvement of a target output attribute, have been developed for water quality prediction and, in both cases, optimization techniques are used to minimize the prediction error [26–28]. Adaptive neuro-fuzzy inference system (ANFIS) and artificial neural networks (ANN) have been the most implemented ML methods during the last decade for the prediction of water quality [29]. Likewise, other complex ML algorithms, such as decision trees and support vector machines have been implemented [23,30,31].

For instance, ML models (probabilistic neural network, k-nearest neighbor, and support vector machine) were used to predict the WQI of Karoon River, Iran. The results showed that when none of the nine input parameters of the WQI were removed, all models displayed the same results, however, the PNN displayed the best performance (accuracy of 94.57%) when parameters were removed [32]. Another study tested a back propagation neural network (BPNN), an adaptive neuro-fuzzy inference system (ANFIS), a support vector regression (SVR), and a multilinear regression (MLR) for the prediction of the WQI at three stations across the Yamuna River, India. Although the unsupervised models displayed a better performance, the MLR displayed a correlation coefficient above 0.9 in most cases [33]. Likewise, a different study tested eight algorithms including multilinear regression (MLR), random forest (RF), M5P tree, random subspace (RSS), additive regression (AR), ANN, SVR, and locally weighted linear regression (LWLR) to predict WQI of the groundwater in Illizi region, southeast Algeria. As a result, the MLR model displayed a higher accuracy compared to the rest of the models [34].

While these complex models can generate a wider variety of mathematical structures to estimate the response, the use of simple-structured supervised models is preferred when models are applied for inference purposes (in addition to the prediction goal), due to its simpler interpretation. For example, in a multiple linear regression model, the relationship between Y and X_1, X_2, ... , X_p, will be easy to understand as they are linearly correlated, while in flexible methods, such as ANN, the association between the response and any individual predictor is quite complex [35].

3. Materials and Methods

3.1. Site Description and Data Collection

The Santiago-Guadalajara River (SGR) originates in Lake Chapala (the largest lake in Mexico) and flows around the Guadalajara Metropolitan Area (GMA), which is the second largest metropolitan area in Mexico, on its way to the Pacific Ocean. The Lerma River, which originates in the State of Mexico, is part of the Lenna-Chapala-Santiago hydrographic system, with a length greater than 700 km. The Lerma-Chapala-Santiago is one of the largest hydrological systems in Mexico, where intensive agricultural and industrial activities are concentrated. The Ahogado stream receives industrial discharges, as well as municipal discharges from the GMA, until it merges with the Santiago Guadalajara River.

The Santiago-Guadalajara River has a length of 433 km and an average flow of 320 m^3/s. The Santiago-Guadalajara River section (with a catchment area of ~10,016 km^2) expands from the tributary basin of the Zula River and is characterized by high levels of industrial, agricultural, and livestock activities, and large urban areas [36]. The water quality in this river has been affected by point and non-point sources of pollution, originating from different industries: crop fields, urban settlements, and municipal landfills, among others [36]. The sanitation systems within the basin have been reported to be insufficient to treat the wastewater generated by the population [36–38], and the public health hazards generated by the water quality condition of the river have caused concern with the different public and private agencies. As a result, the Water Commission of the Government of the State of Jalisco (CEA Jalisco) monitors 44 water quality parameters on a monthly basis at 20 sampling points located in the Río Santiago-Guadalajara basin (Figure 1) [39].

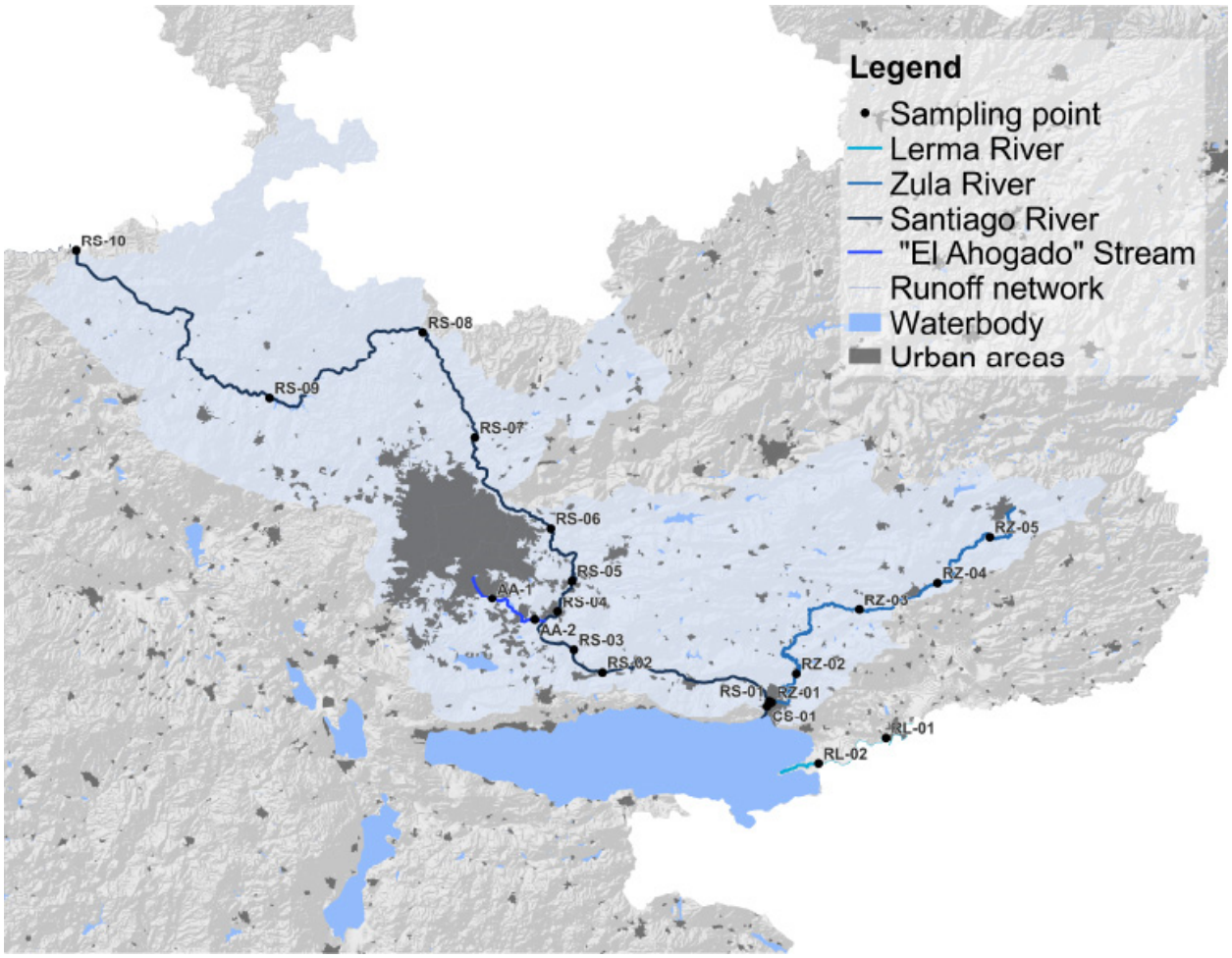

Figure 1. Sampling points along the Santiago-Guadalajara River (SGR).

Sampling points RS01–RS10, which are located within the Santiago River, and sampling points AA01–AA03, located within the Ahogado stream (an urban stream that is an important tributary to the Santiago River), have been monitored since 2009 to date. Sampling points RZ01–RZ05, located within the Zula River, and sampling points RL01–RL02, located at the Lerma-River, have been monitored since March 2020.

3.2. SGR-WQI Algorithm

The SGR-WQI algorithm was specifically developed for the Santiago-Guadalajara River [20]. The methodology for WQI calculation started with a dataset of 51 parameters, and principal component analysis (PCA) was applied to reduce the number of parameters. Seventeen water quality parameters were included in the WQI algorithm as these were found to be the best indicators of the water quality of the river (Table 1). Then, rating curves were developed by [20] for each parameter included in the SGR-WQI algorithm. The rating curves are functions that rate each water quality measurement with a value between the range of 0 (very poor quality) and 100 (excellent quality). These rated values are referred to as sub-indices, Qi. The rating curves, developed by [20], incorporated the local legal limits applicable to the Santiago-Guadalajara River and provide the SGR-WQI with high sensitivities to parameter values that are outside the thresholds established for the protection of aquatic life.

Table 1. Water quality parameters and their assigned weights for the WQI calculation.

Number	Parameter	Abbreviation	Weight
1	Cadmium	Cd	0.057091
2	Chromium	Cr	0.068067
3	Biological oxygen demand	BOD_5	0.066625
4	Dissolved oxygen	DO	0.064003
5	Fecal coliforms	FC	0.047843
6	Fluoride	FL	0.076765
7	Fats, oils, and grease	FOG	0.045367
8	Mercury	Hg	0.032107
9	Ammonia	NH_3	0.071960
10	Nitrates	NO_3	0.088895
11	Lead	Pb	0.042887
12	Hydrogen potential	pH	0.044139
13	Total suspended solids	TSS	0.060231
14	Sulfides	SULF	0.058181
15	Total dissolved solids	TDS	0.079982
16	Temperature	TEMP	0.045952
17	Zinc	Zn	0.049905

The WQI was then calculated as a weighted average of the parameter sub-indices, as shown by the following equation:

$$\mathrm{WQI} = \sum_{k=1}^{17} w_k Qi_k \quad (1)$$

where $0 \leq Qi_k \leq 100$ is the sub-index value of the kth parameter obtained through the rating curve, and w_k is the weight of the kth parameter (Table 1). For the development of the WQI algorithm, multivariate methods (PCA, discriminant analysis, and analysis of variance) were used to select the set of parameter weights, w_k, that best reflected temporal and spatial variability, as well as the annual cycle and trends in the water quality of the river [20]. Each WQI observation was then assigned to a water quality class (WQC) according to the ranges shown in Table 2. A detailed description of the WQI development and algorithm can be found in [20].

Table 2. Classification ranges [20].

WQI Range	Water Quality Class (WQC)
0–25	Very bad
25–50	Bad
50–70	Medium
70–90	Good
90–100	Excellent

3.3. Data Splitting, Training, and Testing Datasets

The complete data matrix was divided into two data subsets. The training subset included water quality data from 2009 to 2020 (with a defined length of 26,367 observations) and was used to calculate the WQI values (1551 WQI observations in total) and to develop the regression and classification models implemented. This subset was used for training because the same data subset was used for the SGR-WQI development [20]. The models developed were then evaluated with a second dataset (testing dataset), which included water quality measurements from January 2021 to April 2021 (1156 observations), which corresponded to 68 WQI observations, to assess the predictive capacity of the models. This data subset was used for testing because it had not been previously used for the SGR-WQI development. Additionally, a 10-fold cross-validation method was implemented to avoid

over-fitting [26]. The accuracies of the derived models were evaluated through the mean squared error (MSE), residual square error (RSE), and the proportion of correctly classified observations in the case of the classification models. The overall methodological approach is summarized in Figure 2.

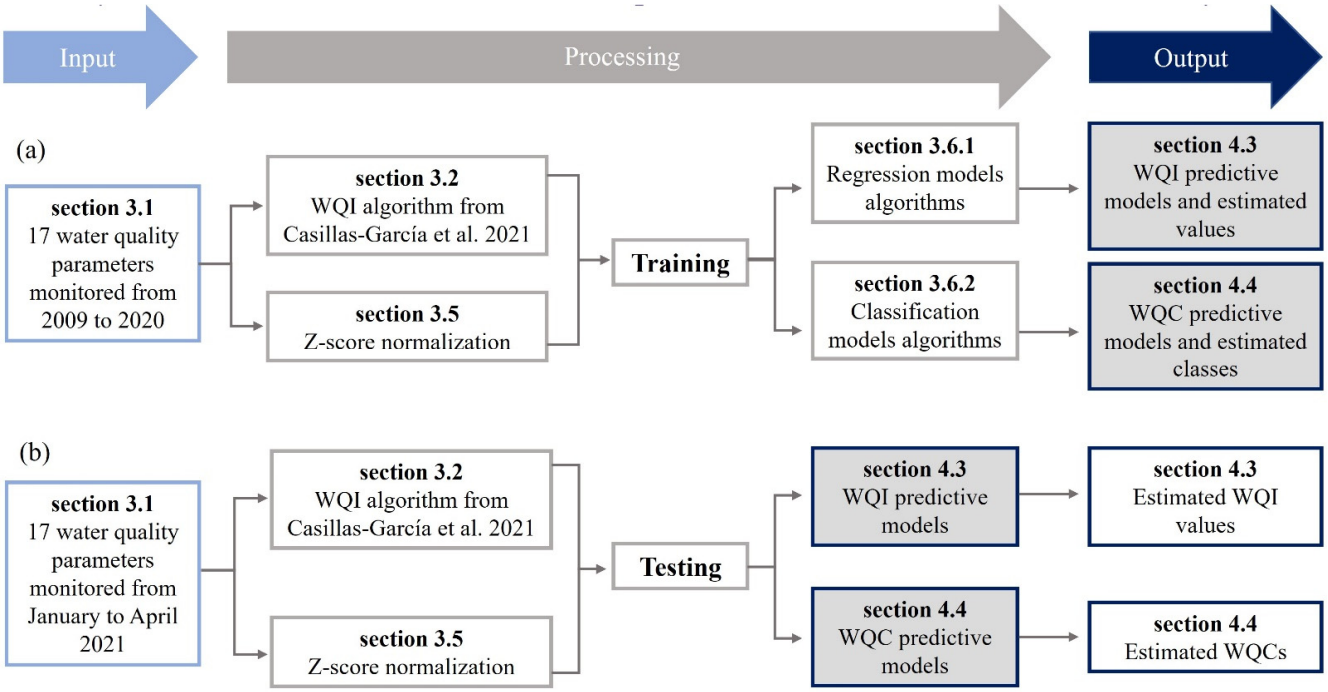

Figure 2. Overall methodological approach for training and testing the predictive models. (**a**) training approach and (**b**) testing approach.

3.4. Outlier Detection

Outliers were detected graphically, using a plot of studentized residuals, and then eliminated to proceed with the phases of model training and testing. Distribution graphs were then developed to observe the number of observations belonging to each WQC.

3.5. Z-Score Normalization

Because of the highly different scales used to measure the several water quality parameters, a normalization process was applied to convert both datasets (training and testing) to a default scale, prior to the phases of model training and testing. A z-score normalization was used for this purpose (Equation (2)). The result of this transformation is a dataset with values varying, ideally between −3 and +3 [40].

$$z_{score} = \frac{x - \mu}{\sigma} \quad (2)$$

where x is the value of a particular observation, μ represents the water quality parameter mean, and σ is the corresponding standard deviation.

3.6. Models Development

3.6.1. Regression Models

Regression models are mathematical expressions that relate to two or more quantitative variables. The variable to be predicted is called the 'response' variable, while the variable(s) used for the estimation are called 'predictor(s)'. These models provide meaningful information for the understanding of causes and responses occurring in natural systems [41]. Simple linear regression models (see "Simple Linear Regression" part) were applied to analyze the relationship between water quality parameters and WQI values,

which provide information on how the individual behavior of a water quality parameter affects the overall quality of the water body (as expressed by the WQI). Multiple linear regression models (see "Multiple Linear Regression" part) were then developed to approximate linear relations between all the water quality parameters and the WQI values, which, contrary to simple linear regression, contribute to the understanding of how the complete set of water quality parameters affects water quality. Ridge and Lasso regressions (see "Lasso and Ridge Regression" part) were then applied to determine if linear models with a better fit could be developed with a reduced number of parameters. Additionally, the selection of significant parameters was achieved using the best subset-selection method, including both the forward and backward methods (see "Parameter Reduction" part). Finally, generalized additive models (see "Generalized Additive Models" part) were used to describe a non-linear relation between all the water quality parameters and the WQI, because the SGR is a natural ecosystem affected by seasonality, geography, and anthropogenic activity, and a complex behavior is expected.

Simple Linear Regression

A simple linear regression model is a simple approach for predicting a quantitative response Y with a single predictor variable X [35]. A linear relationship is assumed between X and Y, according to Equation (3). In this study, simple linear regression models were developed taking Y as the WQI (as calculated by the algorithm described in Section 3.2) and X. as the 17 water quality parameters used to calculate the WQI.

$$\widehat{WQI} = \beta_0 + \beta_1 X \quad (3)$$

Multiple Linear Regression

Multiple linear regression is a method used to predict a response variable Y as a function of more than one predictor X_p, assigning a coefficient β for each predictor variable, as expressed by Equation (4) [35,42]. Coefficient β is a reference of how significant the parameter influence over the WQI is. To develop these models, WQI was considered as Y and the 17 water quality parameters, included for the WQI calculation, were used as predictor variables X_p.

$$\widehat{WQI} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = \beta_0 + \sum_{k=1}^{p} \beta_k X_k \quad (4)$$

where β_0 is the intercept, β_p represents the estimated regression coefficients for each of the water quality parameters, and X_p represents the predictor variables [35]. The regression coefficients were estimated using the least squares fitting procedure minimizing the residual sum of squares.

$$RSS = \sum_{i=1}^{n} \left(WQI_i - \widehat{WQI_i} \right)^2 = \sum_{i=1}^{n} \left(WQI_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad (5)$$

Lasso and Ridge Regression

Lasso and Ridge regressions are models that reduce the regression coefficients towards zero to improve the fit of the model and reduce the variance [35]. These models fit a parametric linear regression model, as represented by Equation (4), but use a different fitting procedure to estimate the coefficients β_0, β_1, …, β_p. The values of the Ridge and Lasso regression coefficients are those that best fit Equations (6) and (7), respectively. Lasso regression is a variant of the Ridge regression, the only difference being the regression

penalty, which is expressed by the β_j^2 coefficients on the Ridge regression model and by the $\left|\beta_j\right|$ model coefficients on the Lasso regression [35].

$$\sum_{i=1}^{n}\left(WQI_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2, \quad (6)$$

$$\sum_{i=1}^{n}\left(WQI_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right| = RSS + \lambda\sum_{j=1}^{p}\left|\beta_j\right|, \quad (7)$$

In both cases, $\lambda \geq 0$ is a tuning parameter to be determined separately. The second terms, $\lambda\sum_j \beta_j^2$ and $\lambda\sum_j\left|\beta_j\right|$, are referred to as the shrinkage penalty. The penalty term has no effect when $\lambda = 0$, and both regressions (Ridge and Lasso) will produce the least square estimates. As $\lambda \to \infty$, the impact of the shrinkage penalty grows and the ridge regression and lasso coefficients will approach zero.

Parameter Reduction

The best subset selection method was applied to the dataset to evaluate the statistical significance of the 17 water quality parameters over the WQI. This algorithm evaluates a separate least squares regression for each possible combination of p predictors to select the best model from 2^p possibilities. The algorithm fits all models for $k = 1, 2, \cdots, p$ that contain k predictors and select the best models of each subset based on the highest R^2. Then, the best model is selected evaluating the adjusted R^2 and the Bayesian information criterion (BIC) [35].

The selection of significant parameters was completed using stepwise methods. Initially, the forward stepwise method was used. This procedure begins with a model containing no predictors and subsequently adds predictors, one at a time, until all the predictors are in the model, for $k = 1, 2, \cdots, (p-1)$, and chooses the best models based on the highest R^2. Then, a final single model was selected using the BIC and the adjusted R^2 [35].

Additionally, the backward stepwise selection method was implemented. This model begins with all p predictors in the least-squares model and removes the least useful predictors one at a time, for $k = p, (p-1), \cdots, 1$, then chooses k models with the highest R^2. In the end, the selection of a single best model is based on the BIC and the adjusted R^2. Backward selection requires several samples n larger than the number of variables p [35].

Generalized Additive Models

Generalized additive models provide a general frame for extending a standard linear model by allowing non-linear functions for each of the variables while maintaining additivity. The following structure was used for the WQI prediction as a function of the water quality parameters $X_1, \ldots X_p$.

$$\widehat{WQI} = \sum_{j=1}^{k_1}\beta_{1j}\phi_{1j}(X_1) + \sum_{j=1}^{k_2}\beta_{2j}\phi_{2j}(X_2) + \cdots + \sum_{j=1}^{k_p}\beta_{pj}\phi_{pj}(X_p), \quad (8)$$

where the base functions ϕ_{ij} are allowed to be, for example, polynomials, natural cubic splines, smooth splines, or even linear models. Thus, the model is the addition of linear combinations of base functions, hence, the contribution and significance of each parameter can be evaluated as in the case of the MLR.

3.6.2. Classification Models

The WQI expresses the water quality by a single number between 0 and 100, which is used to assign a WQC (Table 2) depending on the range of the calculated WQI. The WQC is a qualitative classification that enables communication of the water quality in a more understandable way [43].

Classification models consider a qualitative response variable Y, contrary to linear regression models for which the response is quantitative. In this study, the qualitative response variable is the WQC. The models applied for classification in this work predict the probability of an observation, or a set of observations, to belong to a specific class. The basis for making this classification is the prediction of the probability of each of the categories of a qualitative variable (such as good, medium, or bad water quality). Consequently, these models behave like regression models, as the categorical variable is converted to a numerical value represented by the probability [35].

Logistic Regression

A logistic regression model was developed to calculate the probability p of a Y set of observations of water quality parameters, belonging to a particular WQC. In this sense, there are two possible options, as shown by Equations (5) and (6) [35].

$$p(X) = Pr(Y = 0|\, X) \tag{9}$$

$$p(X) = Pr(Y = 1|\, X) \tag{10}$$

where Y is the WQC and X represents the 17 water quality parameters used for the WQI calculation. For the training phase, the WQC was determined based on the WQI values obtained by the original WQI algorithm developed by Casillas-García et al. (2021) [19] using the training data set (measurements made before 2020). Because this model only considers two classifications for Y, 0 was assigned to a medium quality (Equation (9)) and 1 was assigned to bad quality (Equation (10)). Other categories were not considered because, historically, there were null (for the excellent and good WQCs) or very few (for the very bad WQC) observations corresponding to these classes, as shown in Table 3.

Table 3. WQC observed within the training and testing data sets.

	WQC					
	Excellent	**Good**	**Medium**	**Bad**	**Very Bad**	**Total**
Training data set	0	0	369	1178	4	1551
Test data set	0	0	12	55	1	68

Linear Discriminant Analysis

Linear discriminant analysis (LDA) was used to classify the water quality observations into k classes ($k \geq 2$). In this study, only three classes (medium, bad, and very bad) were considered since the other categories have never been observed; the SGR has been a highly polluted river for more than a decade [20]. LDA is used to determine the probability of an observation, or a set of observations, belonging to each class, as shown by Equation (11).

$$f_k(X) \equiv Pr(X|Y = k) \tag{11}$$

where Y is the qualitative response variable (in this case the WQC), X represents the set of water quality parameters, and k represents the classes (medium, bad, and very bad). Linear discriminant analysis may be used to determine the probability of an observation belonging to the kth class given the predictors' observations. Then, the Bayes' theorem, given in Equation (12), states:

$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{k} \pi_l f_l(x)} \tag{12}$$

where $f_k(x)$ is the density function and π_k is the prior probability of an observation belonging to the kth class [35]. A diagram explaining the methodology for model development and testing is shown in Figure 3. The regression and classification models implemented in this study are summarized in Figure 3.

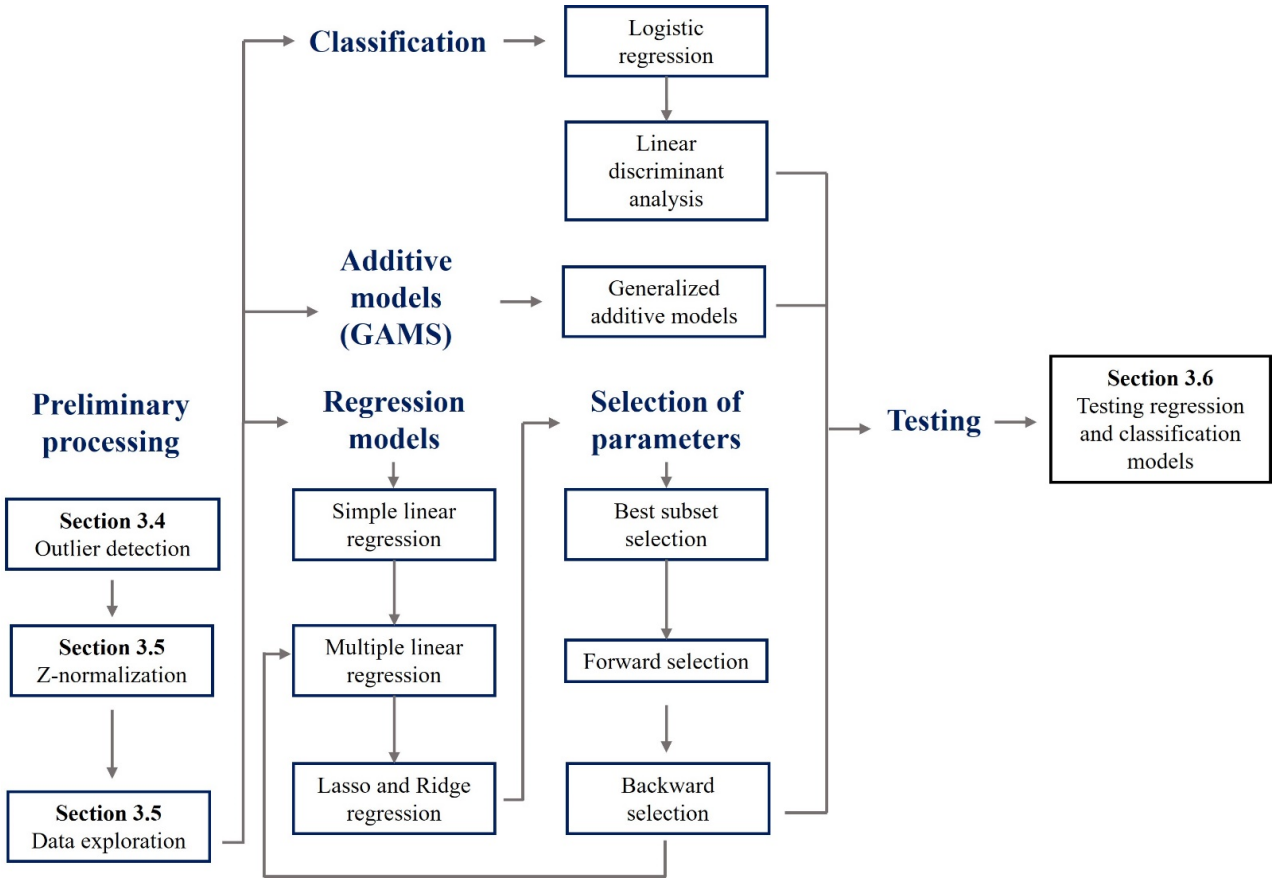

Figure 3. Training and testing of prediction models.

3.7. Software

All calculation and computational analyses were performed using the RStudio software, version 1.4.1717. The packages used were: stats version 4.1.0, MASS version 7.3-54, leaps version 3.1, glmnet version 4.1-2, mgcv version 1.8-35, ggplot2 version 3.3.4, car version 3.0-10, dplyr version 1.0.7, and base version 4.1.0.

4. Results

4.1. Outliers Detection

Four outliers and one observation with high leverage were identified by plotting studentized residuals. Outliers are shown in red in Figure 4. This graph shows the atypical behavior of the outliers, compared with the rest of the observations. These observations belonged to the good WQC and were eliminated from the original dataset.

4.2. Data Exploration and Classification

As previously stated, the Santiago-Guadalajara River is a highly polluted urban river that receives discharges from industrial, agricultural, livestock, and urban sources [44]. Consequently, most of the WQI observations are classified into the bad and medium WQCs, while very few observations fall into the very bad WQC. It is important to note that only these three categories (medium, bad, and very bad) were present in both datasets. The distribution curves of the WQI observations within the training and testing data sets show that no observations were classified into the good or excellent WQCs (Figure 5).

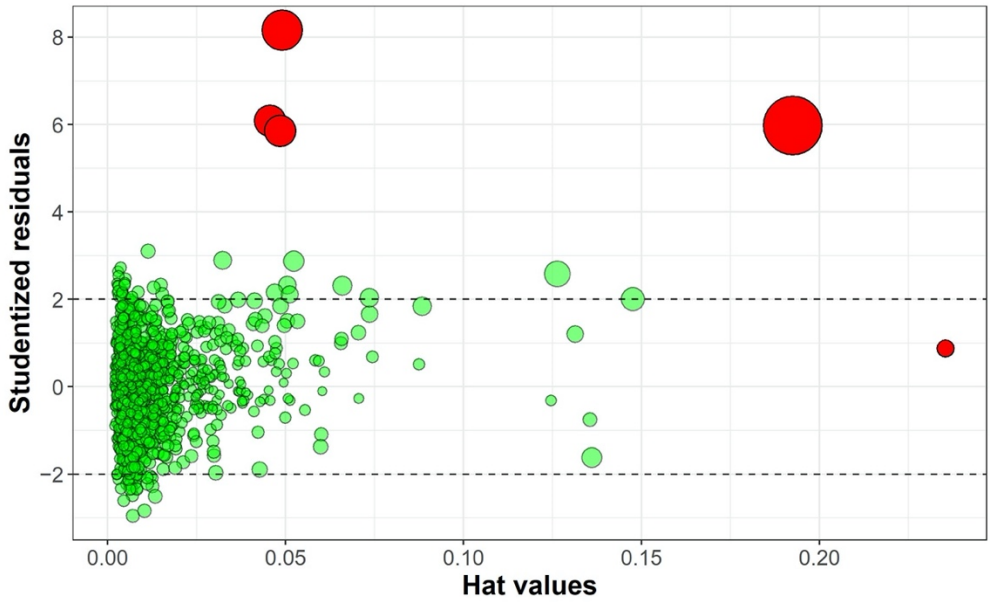

Figure 4. Hat values vs. studentized residuals (green dots are normal range observations and red dots are outliers).

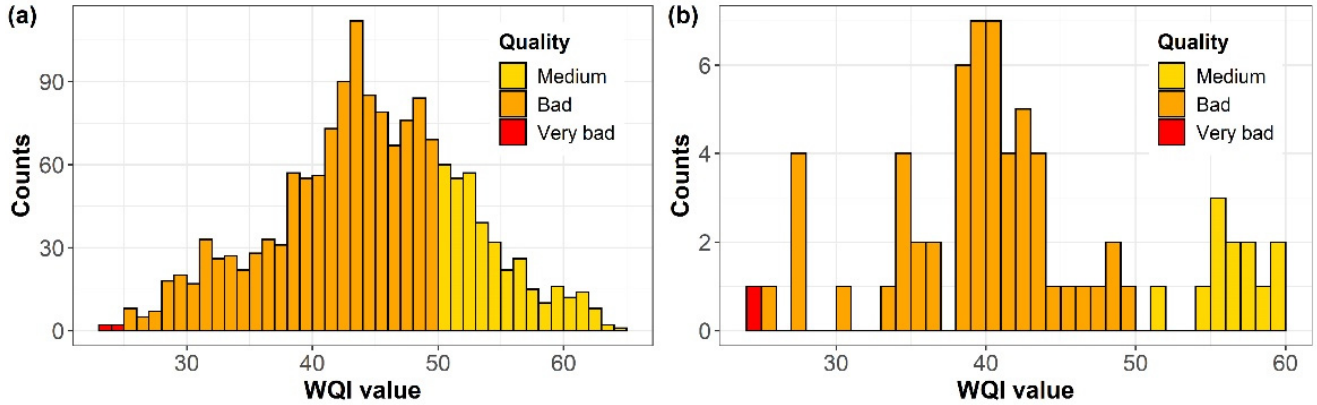

Figure 5. (**a**) Distribution of the WQI observations in the training data set; (**b**) distribution of the WQI observations in the testing data set.

4.3. Regression Models

4.3.1. Simple Linear Regression

Figure 6 presents the distribution curve and the simple linear regression of two individual parameters, pH (Figure 6a,b) and TDS (Figure 6c,d). The TDS linear regression displayed a negative slope with an adjusted R^2 of 0.4425 and R^2 of 0.4429, while the pH linear regression displayed an adjusted R^2 value of -0.00063 and R^2 of 1.033×10^{-5}. These two parameters correspond to the highest (TDS) and lowest (pH) R^2 values obtained within all 17 of the simple linear regression models developed (one for each water quality parameter). Even the highest R^2 is very low (0.4425), which shows that the adjustment is very poor for all the models. These results prove that none of the parameters affect the WQI significantly by themselves, thus, the WQI variations are caused by changes in multiple parameters, as proposed by Casillas-García et al. (2021) for the original WQI algorithms. This fact was expected as it is difficult to estimate a response variable (calculated with 17 parameters) by using only one of them. However, an initial analysis of the fit of these models was useful to proceed with the development of more complex ones.

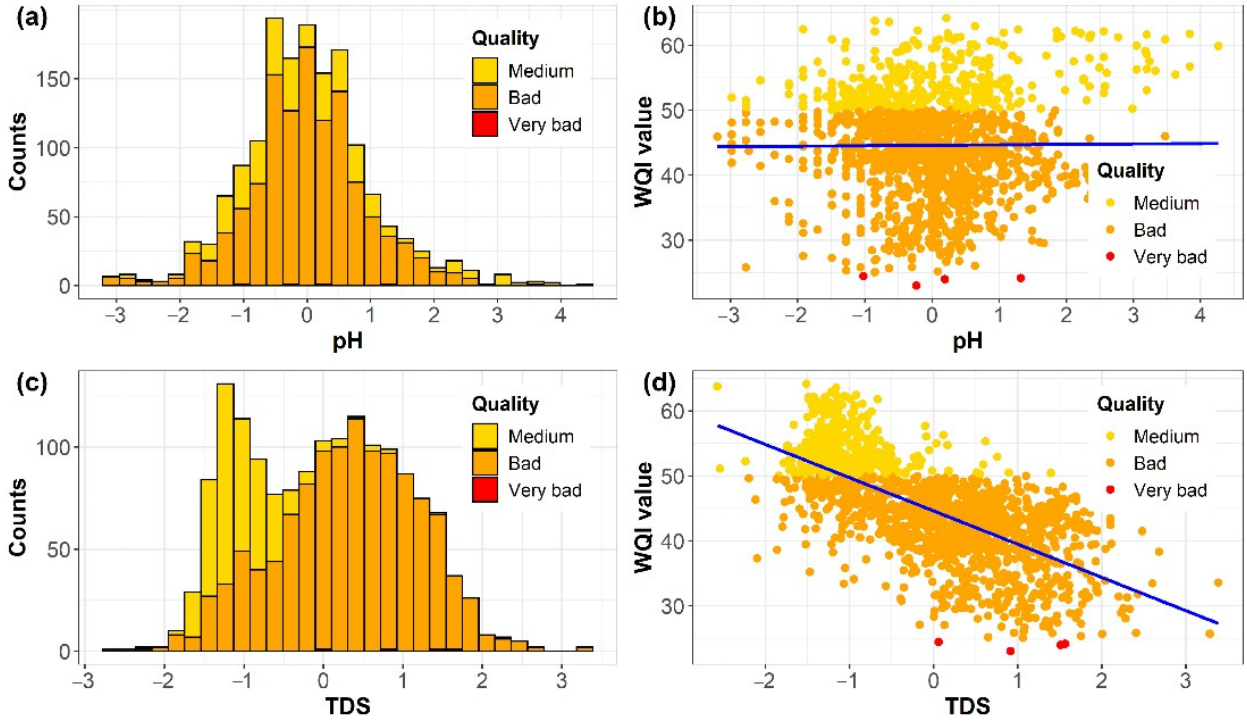

Figure 6. Distributions (**a**,**c**) and simple linear regressions (**b**,**d**) for pH and TDS.

4.3.2. Multiple Linear Regression

The multiple linear regression model used to predict the WQI values (Equation (4)), using all 17 water quality parameters displayed a residual standard error (RSE) of 3.255, an R^2 of 0.8217, and an adjusted R^2 of 0.8197. Additionally, 14 of the 17 parameters were found to be significant (Table 4). The dispersion graphs of the residuals and the fitted values are shown in Figure 7. A better fit was obtained with the multiple linear regression compared to all simple linear regression models, however, the fit was still insufficient to accurately predict the WQI values.

Table 4. Multiple linear regression models' coefficients and *p*-values.

	(a) with 17 Parameters		(b) with 12 Parameters	
Parameter	**Coefficient**	***p*-Value**	**Coefficient**	***p*-Value**
Intercept	44.593	<0.001	44.593	<0.001
Cd	−1.073	<0.001	−1.158	<0.001
Cr	−0.078	0.363		
BOD_5	−0.868	<0.001	−0.827	<0.001
DO	2.293	<0.001	2.321	<0.001
FC	−0.463	<0.001	−0.444	<0.001
FL	−1.894	<0.001	−1.909	<0.001
FOG	−0.690	<0.001	−0.739	<0.001
Hg	−0.020	0.811		
NH_3	−0.580	<0.001	−0.563	<0.001
NO_3	0.113	0.239		
Pb	−0.569	<0.001	−0.594	<0.001
pH	−0.285	0.007	−0.315	0.003
SST	−1.536	<0.001	−1.600	<0.001
SULF	0.187	0.071		
TDS	−2.598	<0.001	−2.603	<0.001
TEMP	0.949	<0.001	0.918	<0.001
Zn	−0.205	0.039		

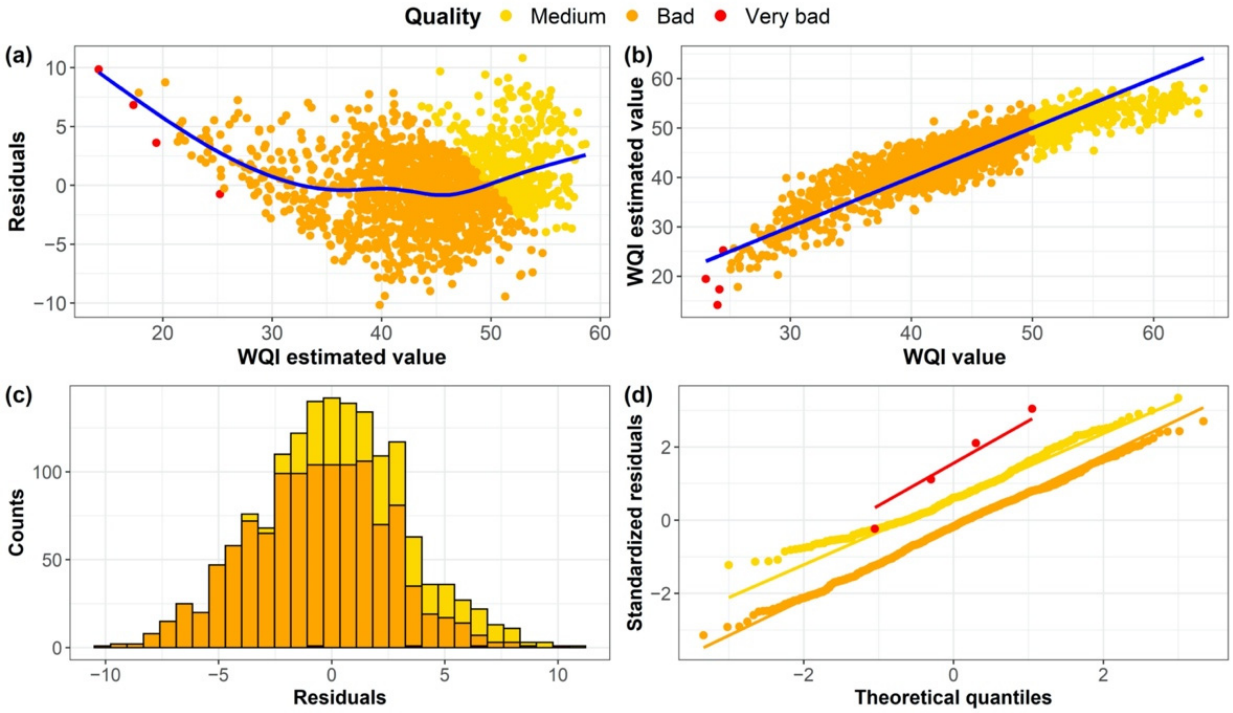

Figure 7. (**a**) Model residuals (**b**) fitted values for the multiple linear regression using the 17 parameters within the training data set, (**c**) histogram of the residuals, and (**d**) QQPlot of standardized residuals.

As only 14 parameters were statistically significant in this model (Table 4), the adjustment could be improved through parameter reduction. The residuals plot in Figure 7a exhibits a quadratic pattern and provides a strong indication of nonlinearity. Nevertheless, normality assumption was checked as shown in Figure 7c,d and was assessed using a Shapiro–Wilk normality test (p value = 0.1892). Heteroscedasticity was detected using a Breusch–Pagan test (p value = 2.32×10^{-16}), thus a constant variance cannot be assumed. Additionally, the NO_3 and SULF coefficients were positive, which is inverse to the logical behavior, as a decrease in the river's water quality is expected when the concentration of this parameters increases. The coefficient value is indicative of the influence of the parameter on the WQI behavior; parameters with positive coefficients positively affect the WQI (such as DO and TEMP), while parameters with negative coefficients diminish the WQI. However, positive coefficients were determined for NO_3 and SULF, which is not expected since a rise in the concentration of these parameters should decrease the WQI. For these reasons, non-significant and inconsistent parameters were excluded for the development of a second multiple linear regression (Table 4).

The second multiple linear regression model (using only 12 parameters) displayed an RSE of 3.262, an R^2 of 0.8205, and an adjusted R^2 of 0.8191. The fit of this new model was found to be almost equal to that of the multiple linear regression model including all 17 parameters. As the model adjustment did not change significantly, the prediction of the SGR-WQI could be achieved by a reduced number of parameters without a significant loss of accuracy. Likewise, the residuals plot (Figure 8a) exhibits the same quadratic pattern indicating non-linearity; normality assumption was also observed and evaluated by a Shapiro–Wilk normality test (p value = 0.1743). Once again, heteroscedasticity was detected using a Breusch–Pagan test (p value = 3.54×10^{-12}). However, as the fit did not improve after eliminating non-significant and inconsistent parameters, linear models could not consistently describe the relationship between the WQI and the water quality parameters. The multiple linear regression model (using 12 parameters) was evaluated using the testing data subset, and a RSE of 11.3664 was obtained. Although this value is indicative of WQI

values predicted with good accuracy, a better WQI prediction could be obtained with non-linear models. Also, the coefficients of the parameters indicate which parameters have a higher influence on the model, which are TDS > DO > FL > TSS > Cd; high content of TDS and TSS are standard indications of water contamination and are related to high amounts of other pollutants, DO is very important for aquatic ecosystems since very low levels can cause fish death, FL has gained relevance in last decades since it is related to health problems such as dental and skeletal fluorosis. Figure 8 shows the residuals and fitted values of the multiple linear regression model using 12 parameters.

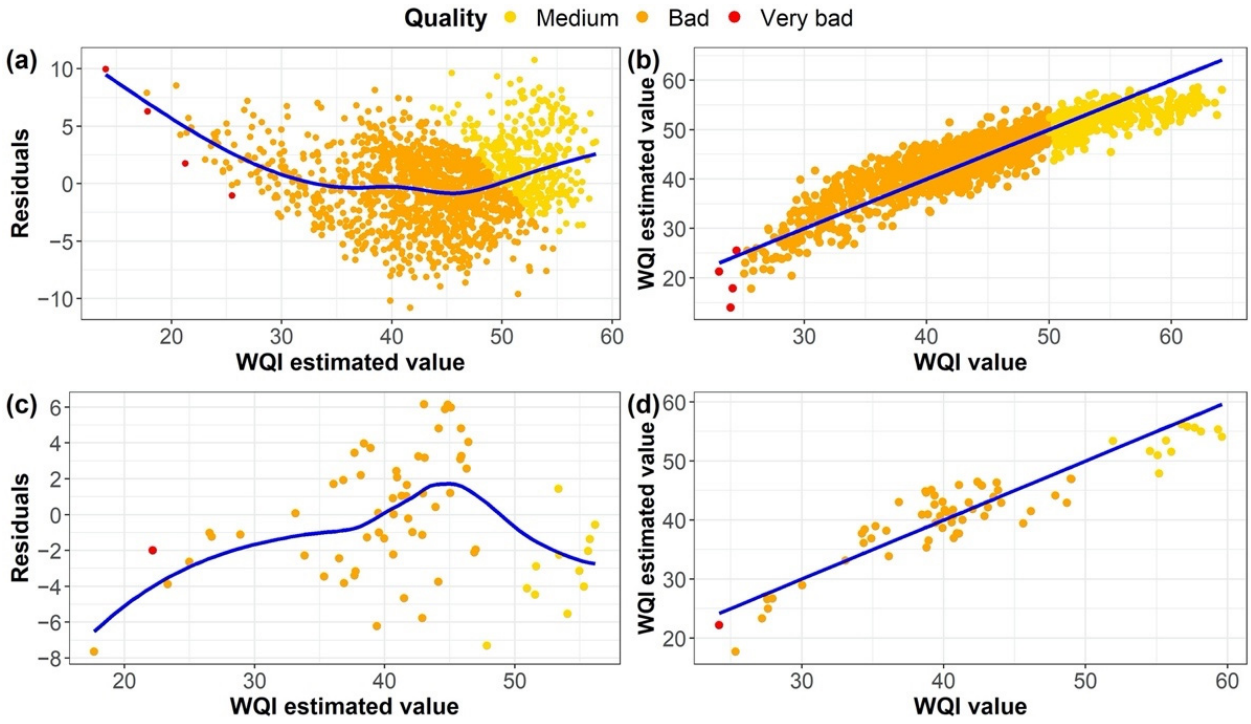

Figure 8. Residuals and fitted values of the multiple linear regression model using 12 parameters. (**a**) Plot of residuals versus the predicted values (using the training data set). (**b**) Real WQI values versus predicted WQI values using the training data set. (**c**) Plot of residuals versus predicted WQI values (using the testing data set). (**d**) Real WQI values versus predicted WQI values (using the testing data set).

4.3.3. Ridge and Lasso Regression

A series of Ridge and Lasso regression models were performed, varying the values of the tuning parameter λ to find the minimum test MSE. These errors were calculated for each value within the range $1 \times 10^{-5} \leq \lambda \leq 1 \times 10^{5}$. Figure 9 shows that the errors minimize for values of $\lambda < 0.001$. Given that these tuning values are close to zero, the penalty term has no effect and, thus, the Ridge and Lasso regression coefficients are like those obtained by the least squares method (Equation (5)). Ridge and Lasso regressions seek to improve the fit of the multiple linear regression model testing iterations in which the coefficient of some parameters is set to zero. The results show that the model is only benefited by eliminating five parameters (Cr, Hg, NO_3, SULF, and Zn), which were those regarded as inconsistent or not significant in the initial multiple linear regression using 17 parameters.

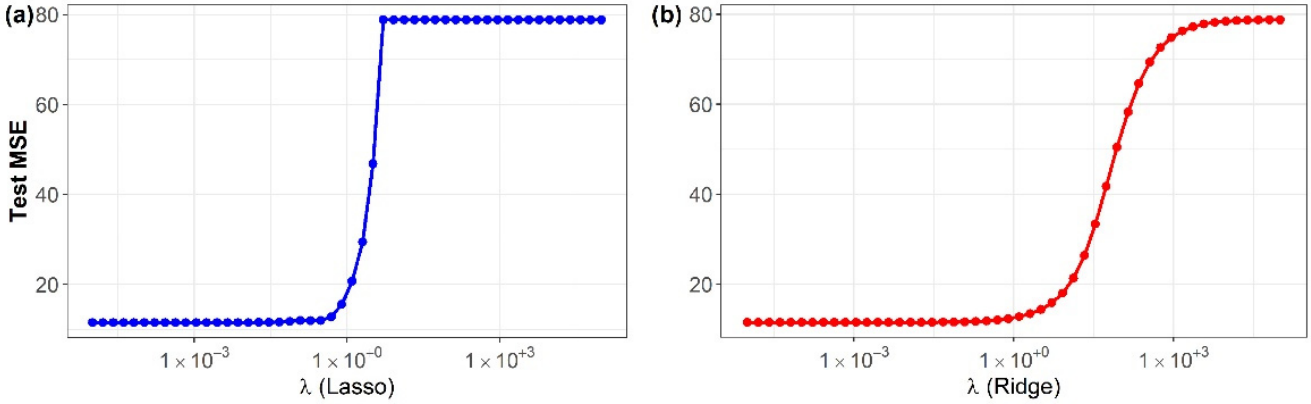

Figure 9. Behavior of the test MSE for the (**a**) Lasso and (**b**) Ridge regressions.

4.3.4. Best Subset Selection, Forward and Backward Stepwise Selection

Figure 10 shows the adjusted R^2 and the BIC for each best model with k predictors. For example, the best simple linear regression model $\mathcal{M}_1$ is given considering TDS as the predictor variable (Figure 1). The maximum adjusted R^2 and the best BIC are reached at 12 predictors (red squares in Figure 10), and, at this point, there is little improvement if additional variables are included. The best subset selection included 12 selected predictors (Cd, BOD_5, DO, CF, FL, FOG, NH_3, Pb, pH, TSS, TDS, and TEMP), which were also used for the multiple linear regression model, including 12 parameters, as they displayed significant coefficients (Table 4). The forward and backward stepwise selections were performed and the same subsets of parameters were obtained at each step, thus, these models are equivalent. These results confirm that the 12 parameters previously selected for the MLR provide a better fit.

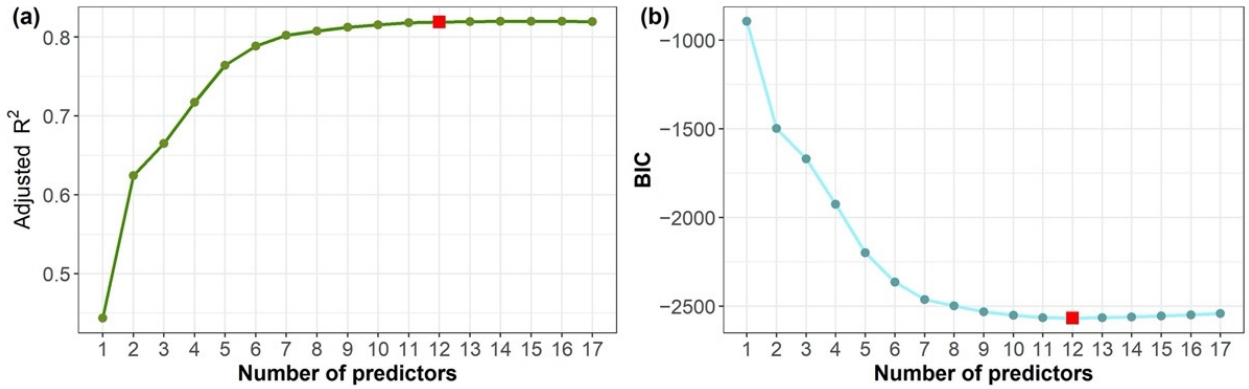

Figure 10. (**a**) Adjusted R^2 and (**b**) BIC vs. the number of predictors obtained by the best subset selection method.

4.3.5. Generalized Additive Model

To improve the accuracy of the WQI prediction, a generalized additive model (Equation (8)) was fitted using natural cubic splines and linear combinations as base functions. Figure 11a shows the plot of residuals for the training data set where there is no discernible pattern. These results have a better accuracy compared to the multiple linear regressions, as an adjusted R^2 of 0.9992 and a MSE of 1.026 were obtained using the test data set. The points in the scatterplots (Figure 11c,d) are practically on the identity function, thus, WQI $\approx \widehat{\text{WQI}}$.

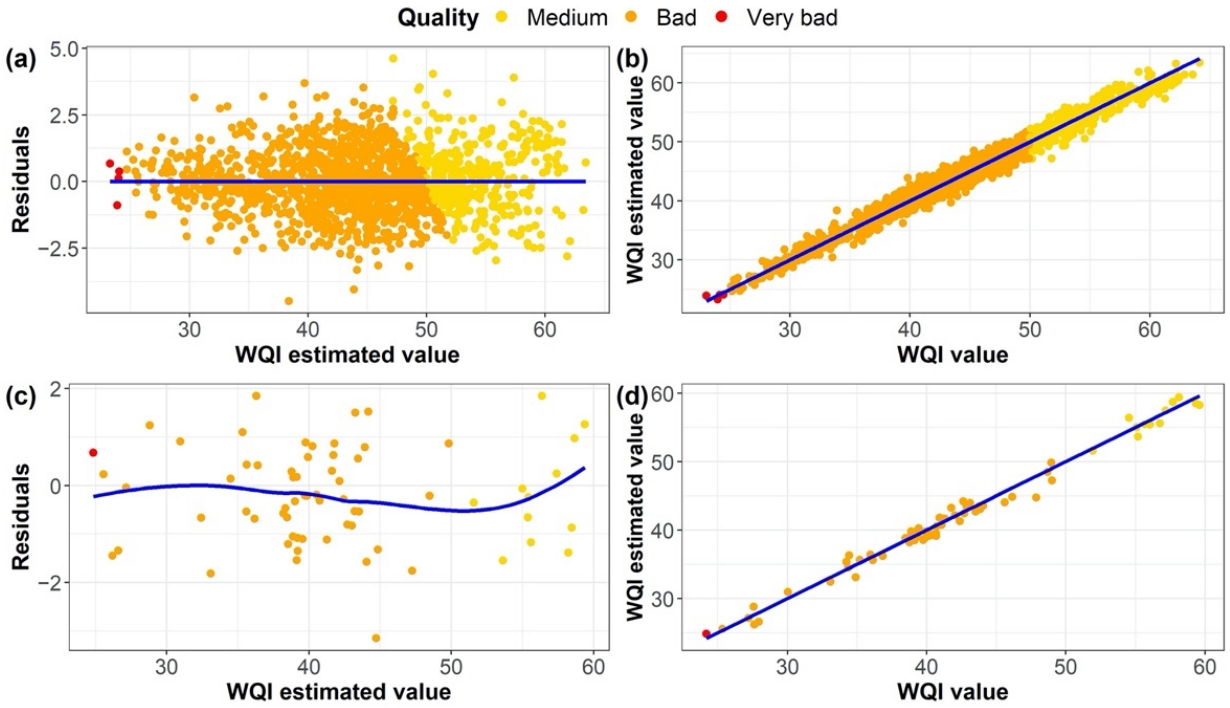

Figure 11. Generalized additive models: (**a**) Plot of residuals versus the predicted values (using the training data subset). (**b**) Real WQI values versus predicted WQI values using the training data subset. (**c**) Plot of residuals versus predicted WQI values (using the testing data subset). (**d**) Real WQI values versus predicted WQI values (using the testing data subset).

An additional characteristic of this model is its capability to model how each parameter affects the WQI. The model in Equation (8) can be rewritten as:

$$\widehat{WQI} = \sum_{k=1}^{p} f_k(X_k), \tag{13}$$

where each function f_k can be fitted by natural splines or step linear models. The original function for the WQI can be considered as a model of this type:

$$\widehat{WQI} = \sum_{k=1}^{p} w_k Q_{ik}(X_k) = \sum_{k=1}^{p} \widehat{Q_{ik}}(X_k), \tag{14}$$

The generalized additive model describes the behavior of the rating curves, such that, as a general setting, we have

$$w_k Q_{ik}(X_k) \approx f_k(X_k) + c_k \tag{15}$$

where c_k is a constant for each parameter, such that $\sum_k c_k \approx 0$. Figure 12 shows these trends for DBO_5, DO, pH, and TDS. The generalized additive models not only improve the accuracy of the WQI prediction but also fit the rating curves (plus a constant) of each parameter developed for the original algorithm by Casillas-García et al. (2021) [20] with great accuracy. As shown in Figure 12, the base functions and the rating curves corresponding to DBO_5, DO, pH, and TDS display the same trend, however, the base functions exhibit deviations that can be attributed to the constant c_k effect. This is a remarkable result, considering that Casillas-García et al. 2021 [20] developed such curves by considering the maximum permissible values of each parameter, as well as the distribution of historical water quality observations.

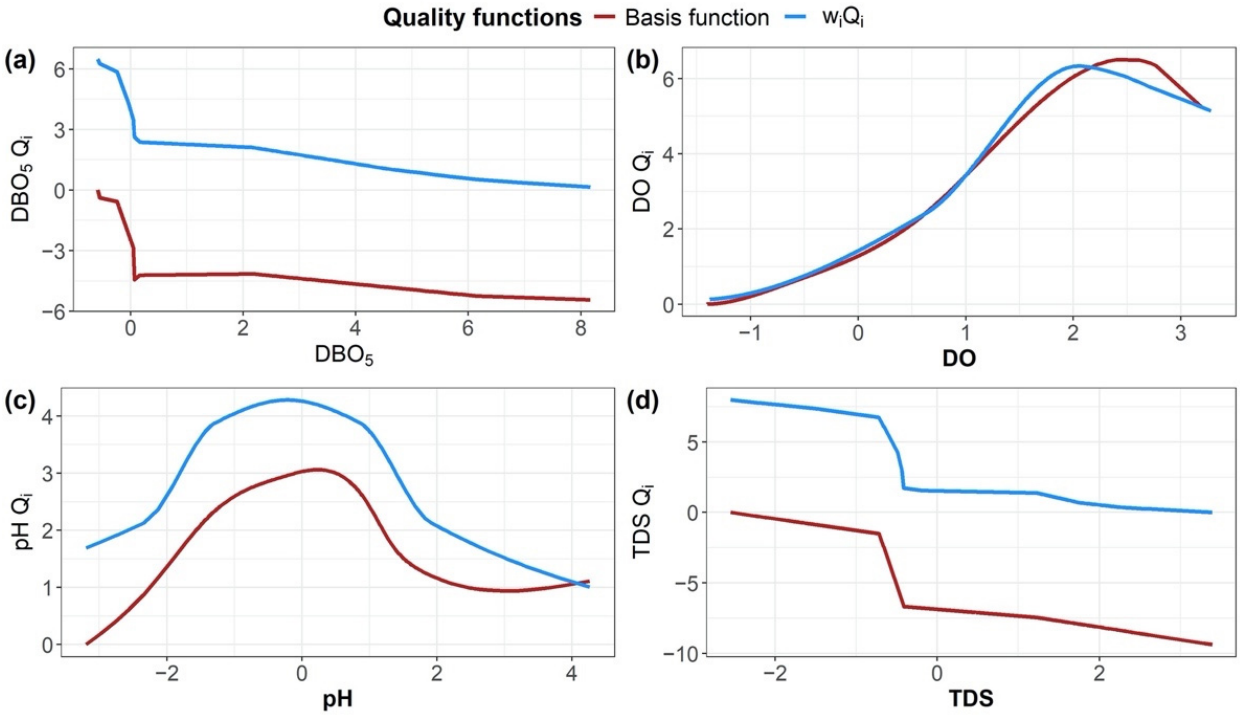

Figure 12. Comparison of the weighted rating curve (WiQi) developed for WQI calculation by Casillas-García et al., (2021) [20], marked in blue, and the base functions obtained by the generalized additive model for (**a**) BOD_5, (**b**) DO, (**c**) pH, and (**d**) TDS, marked in red.

4.4. Classification Models

The results of the classification models were compared with the true classification (Table 3) defined by the ranges for the WQCs estimated by the SGR-WQI algorithm (Table 2).

4.4.1. Logistic Regression Model

The logistic regression model correctly classified a proportion of 84% (311 observations) into the medium water quality class, when comparing the model's classification results with the true classification obtained through the original SGR-WQI algorithm. In addition, a proportion of 96% (1130 observations) was correctly classified into the bad water quality class. The overall performance of this model was 93%. These results are summarized in Table 5.

Table 5. Summary of the results achieved by the logistic regression model.

			True Classification		Proportion of Correctly Classified Observations
			Medium	Bad	
Model classification	Training data set (17 parameters)	Bad	58	1130	0.959
		Medium	311	48	0.843
	Training data set (12 parameters)	Bad	98	1099	0.933
		Medium	271	79	0.734
	Testing dataset	Bad	1	55	1
		Medium	11	0	0.916

The Cd, BOD_5, DO, FL, FOG, Pb, pH, SST, SULF, TDS, TEMP, and Zn parameter results were significant for this model. It is important to note that the significant parameters in this model coincide with previous methods (the multiple linear regression and best subset selection). Only two categories were included to be classified by this model, the medium WQC was set if the calculated conditional probability was >0.5 (Equation (9)),

while the bad WQC corresponded to a probability <0.5 (Equation (10)). As shown in Figure 13, highly reliable WQC predictions were achieved by the logistic regression model. However, the observations classified into the medium WQC (yellow dots) with a conditional probability >0.5 are observations that were misclassified, as these actually belong to the bad WQC (true classification). Conversely, the observations classified into the medium WQC (orange dots) with a conditional probability <0.5 are observations belonging to the medium WQC that were misclassified.

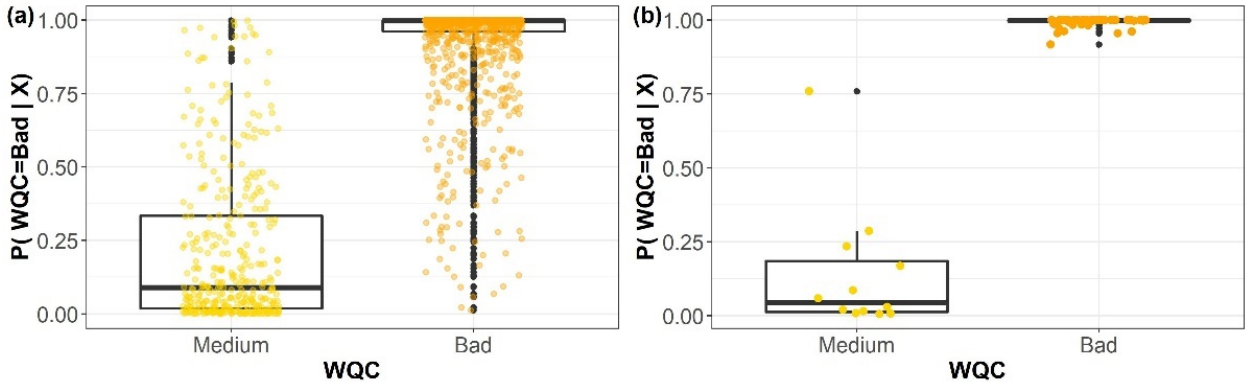

Figure 13. Conditional probability of the classified observations for (**a**) the training data set and (**b**) testing data set.

To reduce the number of parameters required to make the water quality classification, a second logistic regression model was developed. The parameters selected for this second logistic regression model were Cd, Cr, BOD_5, DO, FC, FL, FOG, Hg, NH_3, NO_3, Pb, and Zn (12 parameters). These parameters were selected by Roy's first root statistic criteria [45] and are consistent with those selected by the best subset selection (Section 3.5). As a result, proportions of 73% (271 observations) and 93% (1099 observations) were classified correctly into the medium and bad water quality class, respectively (Table 5), with a global performance of 86% of correctly classified observations. This logistic regression displayed a lower proportion of correct classifications in comparison with the logistic regression model including 17 parameters (84% for medium and 96% for bad classifications). However, this difference may not be significant if the cost-benefit of predicting the WQC using only 12 parameters is considered. Finally, the logistic regression models were applied to test the data subset and a proportion of 100% (55 observations) was correctly classified into the bad WQC and a proportion of 92% (11 observations) was correctly classified into the medium WQC. Both logistic regression models (with 17 and 12 parameters) exhibited the same results for testing.

4.4.2. Linear Discriminant Analysis

The linear discriminant functions, derived from the discriminant analysis, correctly classified a proportion of 81% (298 observations) into the medium WQC and a proportion of 92% (1130 observations) was correctly classified into the bad WQC (Table 6), resulting in an overall performance of 87%. All the observations belonging to the very bad WQC were correctly classified by the model, however, 13 additional observations were misclassified in this category. This error could be attributed to the low number of observations in this category, which makes the accurate adjustment of the linear discriminant functions difficult. Figure 14 shows the differences between the results of the classification displayed by the original SGR-WQI calculation algorithm and the linear discriminant functions (Figure 14a,b, respectively). The first linear discriminant function (LD1) in the x-axis distinguished between the medium and bad quality observations while the second linear discriminant function (LD2) distinguished between medium/bad and very bad quality observations.

Table 6. Summary of discriminant analysis model.

			True Classification			Proportion of Correctly Classified Data
			Medium	Bad	Very Bad	
Model classification	Training data set (17 parameters)	Medium	298	77	0	0.807
		Bad	71	1088	0	0.923
		Very bad	0	13	4	1
	Training dataset (12 parameters)	Medium	210	74	0	0.569
		Bad	159	1092	0	0.937
		Very bad	0	12	4	1
	Test data set (17 parameters)	Medium	11	0	0	0.916
		Bad	1	55	1	1
		Very bad	0	0	0	0
	Test data set (12 parameters)	Medium	10	0	0	0.833
		Bad	2	55	1	1
		Very bad	0	0	0	0

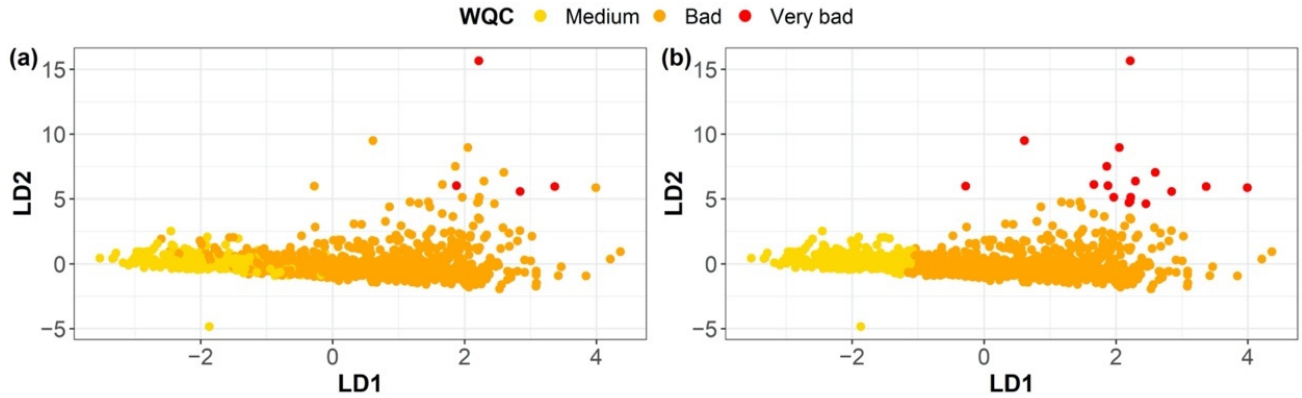

Figure 14. (**a**) Real WQC original obtained by the original WQI calculation algorithm; (**b**) predicted WQC by the linear discriminant functions using 17 parameters. Both using the training data subset.

A second discriminant analysis was performed with the 12 parameters selected by Roy's first root statistic criteria [45] and the best subset selection (Section 3.5). This new model correctly classified a proportion of 57% (210 observations) and 94% (1092 observations) into the medium and bad WQCs, respectively (Table 6), with a corresponding global performance of 84% of correctly classified observations. Similarly, for the model with 17 parameters, all observations belonging to the very bad WQC were correctly assigned, however, 12 additional observations were misclassified into this category. Figure 15 shows the differences between the results of the classifications displayed by the original WQI calculation algorithm and the linear discriminant functions (using 12 parameters). The classification achieved by the model with 12 parameters was similar to that achieved by the model with 17 parameters. Finally, the LDA models (17 and 12 parameters) were applied to the testing data set and they both correctly classified a proportion of 100% (55 observations) into the bad WQC and proportions of 92% (11 observations) and 83% (10 observations) into the medium WQC, respectively.

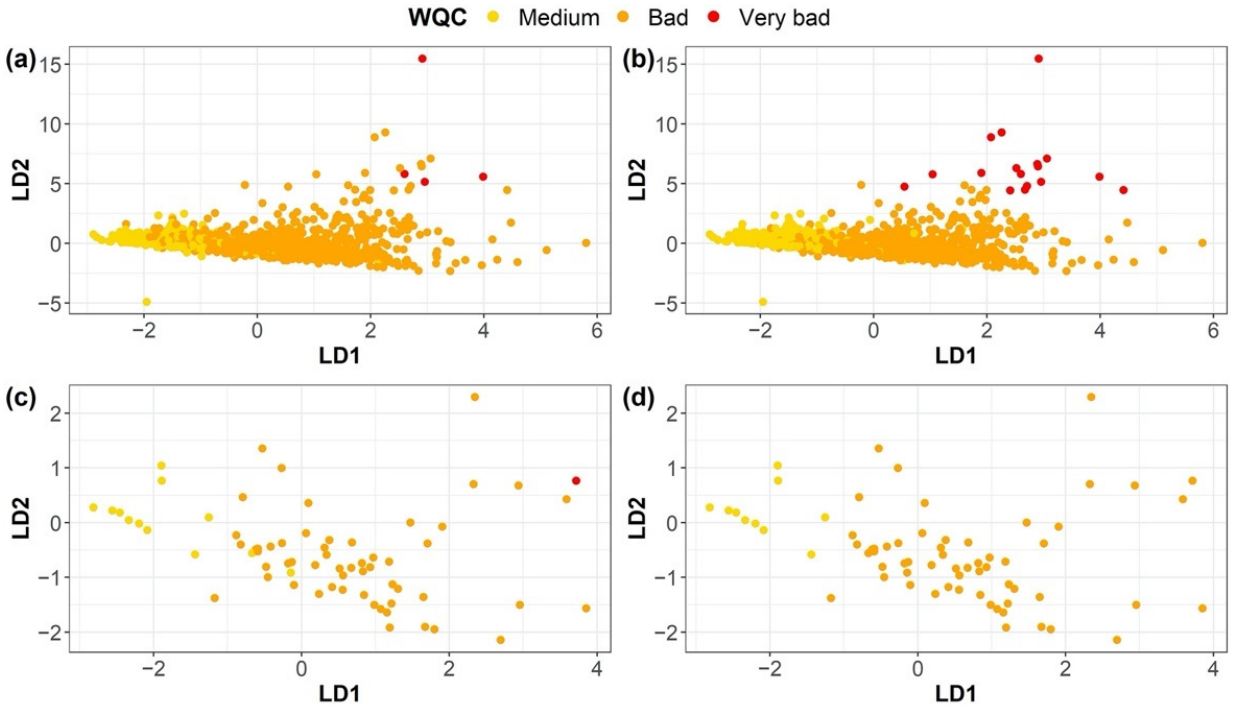

Figure 15. (**a**) Real WQCs obtained by the original WQI calculation algorithm (using the training data subset). (**b**) Predicted WQCs achieved by the model with 12 parameters (using the training data subset). (**c**) Real WQCs obtained by the original WQI calculation algorithm (using the testing data subset). (**d**) Predicted WQCs achieved by the model with 12 parameters (using the testing data subset).

5. Discussion

Simple linear regression models were developed to analyze how the individual behavior of a water quality parameter affects the WQI, however, none of these linear models displayed a good fit, indicating that none of the parameters can be used to predict the WQI by itself. The multiple linear regression model using 17 parameters obtained a residual standard error (RSE) of 3.255, an R^2 of 0.8217, and an adjusted R^2 of 0.8197. The adjustment was better than that reported by Ahmed et al. (2019) for a multiple linear regression model using only four water quality parameters, implemented to predict a WQI which displayed a RSE of 11.7550 and an R^2 of 0.6573.

Because the multiple linear regression model including 12 parameters displayed little variation in accuracy, compared with the previous multiple linear regression including 17 parameters, the specific SGR-WQI developed by Casillas-García et al. (2021), which comprises a complex calculation algorithm, can be estimated using less water quality parameters through a simpler algorithm. This fact opens up the possibility to expand the monitoring to additional sampling points within the Santiago-Guadalajara River basin with a lower budget. However, it is important to maintain the monitoring of the 20 current sampling points with the 17 original parameters as the water quality trends could change over time, especially if corrective actions are applied to reduce contamination. The SGR-WQI original algorithm was developed with a self-adaptive approach, based on the historical distribution of water quality data [20]. If the trends of these data changes over time, the SGR-WQI calculation will adapt accordingly to maintain its reliability and specificity. In the same way, upgrading the machine-learning models developed will be required if new trends in water quality are presented. Corrective actions have recently been implemented and will be further implemented in the coming years to improve the water quality in the Santiago-Guadalajara River. If the water quality in the Santiago-Guadalajara River improves in the future and the ranges not reported in the current dataset are reached, new supervised machine-learning models will be necessary to adjust to the new behavior, as the

current calculations are only reliable for the current and past water quality conditions of the river. The generalized additive model displayed an adjusted R^2 of 0.9992 and an MSE of 1.026. These results are comparable to those reported by Asadollah et al. (2021) using an extra tree regression to predict a WQI with 10 parameters, which displayed a R^2 of 0.99 and an RSE of 1.37. Similarly, Hameed et al. (2017) used a neural network with a radial base function using six parameters to estimate a WQI and obtained an R^2 of 0.9872 and RSE of 0.0157. The results obtained by the generalized additive model are relevant as they prove that WQIs can be estimated using simpler algorithms than those reported previously (decisions trees and neural networks). The use of simpler models is advantageous to reduce the time and effort required for WQI prediction.

Regarding classification models, the logistic regression model developed here correctly classified 93% and 86% observations for the 17 and 12 parameter models, respectively. The linear discriminant functions correctly classified 87% and 84% observations for the 17 and 12 parameter models respectively. Ho et al. (2019) used a decision tree algorithm to predict WQC and achieved 84% accuracy by their best configuration using five parameters; the accuracy was reduced to 81.8% and 77.3% when four and three parameters were used. In addition, Ahmed et al. (2019) achieved an R^2 of 0.56 for a classification model using a multi-layer perceptron algorithm with four parameters. The logistic linear regression model developed in this study displayed a better performance for WQC prediction compared with the linear discriminant functions and other algorithms reported previously, however, this model can only be applied for the prediction of two classes, as most of the observations on SGR belong to medium and bad categories.

In this study, we focused on testing simple-structured supervised ML models with linear or quasi-linear structures as there was previous knowledge of the highly linear structure of the SGR-WQI algorithm [32–34]. Furthermore, the results herein presented indicate that in fact the WQI can be predicted with these models. The generalized additive model displayed the best performance of the models here tested. However, the MLR model using 12 parameters is a better option if the main goal is parameter reduction. Additionally, the models applied in this work are easy to reproduce for water quality evaluation due to their relatively simple structure and practical programming.

6. Conclusions

An accurate, efficient, practical, and cost-effective water quality framework is vital to preserve and establish remedial actions for water bodies. The prediction of the SGR-WQI through machine-learning algorithms is a solid approach to analyze and estimate water quality behavior. From the models evaluated in this research, the generalized additive model and the logistic regression methods displayed better performance in estimating WQI and WQC, respectively. However, for input reduction, the MLR model using 12 parameters is a better option.

A good model must be able to differentiate the intrinsic data variability (caused by the sampling and measurement techniques) from actual changes in the water quality trend. The wider the dataset used, the more precise the model will be. In our case, the monthly monitoring data from the last 10 years was used, however, this monitoring data only produced WQI observations within two water quality classes (more than 99% of the observations belonged to the medium or bad WQCs). Thus, if the new monitoring data produces WQI observations belonging to different WQCs, the models developed here could become less precise. In this sense, fuzzy logic could be a better approach to process a wider range of WQI inputs. Additionally, more complex supervised models, such as ANNs, could provide more precise responses and should be tested in future research. However, these models have reduced interpretability in comparison to the supervised models herein presented and have the potential of overfitting a quasi-linear system.

It is important to highlight that the machine-learning models do not substitute the SGR-WQI algorithm; these methodologies were developed as complementary tools to extend the current water quality monitoring program of the Santiago-Guadalajara River.

The monitoring of all 17 water quality parameters is essential for the estimation of the SGR-WQI in the main channel of the river. However, through machine-learning models a good estimation can be achieved using 12 parameters. The SGR-WQI estimation, in turn, can be used to expand the monitoring to additional sampling points in the main channel of the Santiago-Guadalajara River or its tributaries. Additionally, some future perspectives for expanding the scope of the application of the machine-learning models for the SGR-WQI prediction is the use of a state observer with a differential network for further parameter reduction. If SGR-WQI can be estimated with accuracy, using only parameters measured in situ by sensors, online real-time estimation of the SGR-WQI would be feasible. This extended monitoring would require less time and a lower budget, as this methodology requires less water quality parameters, no laboratory analytics, and a simpler algorithm to do the calculations. The expanded monitoring would provide more detailed information on the watershed situation, which is useful for developing more efficient and precise corrective and preventive action plans to improve the water quality of the river.

Author Contributions: Conceptualization, A.F.d.C., C.Y.-M. and M.S.G.-H.; Data curation, C.Y.-M. and M.V.G.; Formal analysis, A.F.d.C., C.Y.-M. and M.V.G.; Funding acquisition, A.G.-G. and M.S.G.-H.; Investigation, A.F.d.C., C.Y.-M., M.V.G. and M.S.G.-H.; Methodology, A.F.d.C., C.Y.-M. and M.V.G.; Project administration, A.G.-G. and M.S.G.-H.; Resources, A.G.-G. and M.S.G.-H.; Software, A.F.d.C. and C.Y.-M.; Supervision, J.d.A.; Visualization, A.F.d.C.; Writing—original draft, A.F.d.C.; Writing—review & editing, A.F.d.C. and M.S.G.-H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the "Jalisco Scientific Development Fund (FODECIJAL) to Attend State Problems 2020", project name: "Predictive support system for the recovery and comprehensive management of water in the upper basin of the Santiago River, based on a dynamic model and Artificial Intelligence and Machine Learning tools", grant code: 8962-2020.

Data Availability Statement: Water quality data used in this study are published by the Jalisco state government, available in "Sistema de Calidad del Agua" at http://info.ceajalisco.gob.mx/sca/ (accessed on 1 July 2021).

Acknowledgments: The authors acknowledge the transparency of the Jalisco state government, especially the State Water Commission (CEA) of the Secretariat of Integral Water Management (SGIA) and the Secretariat of Environment and Territorial Development (SEMADET), who gave access to all the data presented.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Grabowski, R.C.; Gurnell, A.M. Hydrogeomorphology—Ecology Interactions in River Systems. *River Res. Appl.* **2016**, *32*, 139–141. [CrossRef]
2. Das Gupta, A. Implication of Environmental Flows in River Basin Management. *Phys. Chem. Earth Parts A/B/C* **2008**, *33*, 298–303. [CrossRef]
3. Pandhiani, S.M.; Sihag, P.; Shabri, A.B.; Singh, B.; Pham, Q.B. Time-Series Prediction of Streamflows of Malaysian Rivers Using Data-Driven Techniques. *J. Irrig. Drain. Eng.* **2020**, *146*, 04020013. [CrossRef]
4. Brack, W.; Dulio, V.; Ågerstrand, M.; Allan, I.; Altenburger, R.; Brinkmann, M.; Bunke, D.; Burgess, R.M.; Cousins, I.; Escher, B.I.; et al. Towards the Review of the European Union Water Framework Directive: Recommendations for More Efficient Assessment and Management of Chemical Contamination in European Surface Water Resources. *Sci. Total Environ.* **2017**, *576*, 720–737. [CrossRef]
5. Bhatti, N.B.; Siyal, A.A.; Qureshi, A.L.; Bhatti, I.A. Socio-Economic Impact Assessment of Small Dams Based on T-Paired Sample Test Using SPSS Software. *Civ. Eng. J.* **2019**, *5*, 153–164. [CrossRef]
6. Cordier, C.; Guyomard, K.; Stavrakakis, C.; Sauvade, P.; Coelho, F.; Moulin, P. Culture of Microalgae with Ultrafiltered Seawater: A Feasibility Study. *SciMedicine J.* **2020**, *2*, 56–62. [CrossRef]
7. Singh, B.; Sihag, P.; Deswal, S. Modelling of the Impact of Water Quality on the Infiltration Rate of the Soil. *Appl. Water Sci.* **2019**, *9*, 15. [CrossRef]
8. Kachroud, M.; Trolard, F.; Kefi, M.; Jebari, S.; Bourrié, G. Water Quality Indices: Challenges and Application Limits in the Literature. *Water* **2019**, *11*, 361. [CrossRef]

9. Tiyasha; Tung, T.M.; Yaseen, Z.M. A Survey on River Water Quality Modelling Using Artificial Intelligence Models: 2000–2020. *J. Hydrol.* **2020**, *585*, 124670. [CrossRef]
10. Behmel, S.; Damour, M.; Ludwig, R.; Rodriguez, M.J. Water Quality Monitoring Strategies—A Review and Future Perspectives. *Sci. Total Environ.* **2016**, *571*, 1312–1329. [CrossRef]
11. Ouyang, Y. Evaluation of River Water Quality Monitoring Stations by Principal Component Analysis. *Water Res.* **2005**, *39*, 2621–2635. [CrossRef] [PubMed]
12. Abbasi, T.; Abbasi, S.A. Chapter 1—Why Water-Quality Indices. In *Water Quality Indices*; Abbasi, T., Abbasi, S.A., Eds.; Elsevier: Amsterdam, The Netherlands, 2012; pp. 3–7. ISBN 978-0-444-54304-2.
13. Ewaid, S.H.; Abed, S.A.; Kadhum, S.A. Predicting the Tigris River Water Quality within Baghdad, Iraq by Using Water Quality Index and Regression Analysis. *Environ. Technol. Innov.* **2018**, *11*, 390–398. [CrossRef]
14. Lumb, A.; Sharma, T.C.; Bibeault, J.-F. A Review of Genesis and Evolution of Water Quality Index (WQI) and Some Future Directions. *Water Qual. Expo. Health* **2011**, *3*, 11–24. [CrossRef]
15. Debels, P.; Figueroa, R.; Urrutia, R.; Barra, R.; Niell, X. Evaluation of Water Quality in the Chillán River (Central Chile) Using Physicochemical Parameters and a Modified Water Quality Index. *Environ. Monit. Assess.* **2005**, *110*, 301–322. [CrossRef] [PubMed]
16. Mohebbi, M.R.; Saeedi, R.; Montazeri, A.; Azam Vaghefi, K.; Labbafi, S.; Oktaie, S.; Abtahi, M.; Mohagheghian, A. Assessment of Water Quality in Groundwater Resources of Iran Using a Modified Drinking Water Quality Index (DWQI). *Ecol. Indic.* **2013**, *30*, 28–34. [CrossRef]
17. Bordalo, A.A.; Teixeira, R.; Wiebe, W.J. A Water Quality Index Applied to an International Shared River Basin: The Case of the Douro River. *Environ. Manag.* **2006**, *38*, 910–920. [CrossRef] [PubMed]
18. 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.* **2007**, *7*, 315–328. [CrossRef]
19. Rangeti, I.; Dzwairo, B.; Barratt, G.J.; Otieno, F.A.O. Ecosystem-Specific Water Quality Indices. *Afr. J. Aquat. Sci.* **2015**, *40*, 227–234. [CrossRef]
20. Casillas-García, L.F.; de Anda, J.; Yebra-Montes, C.; Shear, H.; Díaz-Vázquez, D.; Gradilla-Hernández, M.S. Development of a Specific Water Quality Index for the Protection of Aquatic Life of a Highly Polluted Urban River. *Ecol. Indic.* **2021**, *129*, 107899. [CrossRef]
21. Tyagi, S.; Sharma, B.; Singh, P.; Dobhal, R. Water Quality Assessment in Terms of Water Quality Index. *Am. J. Water Resour.* **2013**, *1*, 34–38. [CrossRef]
22. Gradilla-Hernández, M.S.; de Anda, J.; Garcia-Gonzalez, A.; Montes, C.Y.; Barrios-Piña, H.; Ruiz-Palomino, P.; Díaz-Vázquez, D. Assessment of the Water Quality of a Subtropical Lake Using the NSF-WQI and a Newly Proposed Ecosystem Specific Water Quality Index. *Environ. Monit. Assess.* **2020**, *192*, 296. [CrossRef] [PubMed]
23. Asadollah, S.B.H.S.; Sharafati, A.; Motta, D.; Yaseen, Z.M. River Water Quality Index Prediction and Uncertainty Analysis: A Comparative Study of Machine Learning Models. *J. Environ. Chem. Eng.* **2021**, *9*, 104599. [CrossRef]
24. Braiek, H.B.; Khomh, F. On Testing Machine Learning Programs. *J. Syst. Softw.* **2020**, *164*, 110542. [CrossRef]
25. Estrategia. Available online: http://riosantiago.jalisco.gob.mx/estrategia (accessed on 28 August 2021).
26. Peters, T. Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control. *Contemp. Phys.* **2019**, *60*, 320. [CrossRef]
27. Di, Z.; Chang, M.; Guo, P.; Li, Y.; Chang, Y. Using Real-Time Data and Unsupervised Machine Learning Techniques to Study Large-Scale Spatio–Temporal Characteristics of Wastewater Discharges and Their Influence on Surface Water Quality in the Yangtze River Basin. *Water* **2019**, *11*, 1268. [CrossRef]
28. Alloghani, M.; Al-Jumeily, D.; Mustafina, J.; Hussain, A.; Aljaaf, A.J. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In *Supervised and Unsupervised Learning for Data Science*; Berry, M.W., Mohamed, A., Yap, B.W., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 3–21. ISBN 978-3-030-22475-2.
29. Ighalo, J.O.; Adeniyi, A.G.; Marques, G. Artificial Intelligence for Surface Water Quality Monitoring and Assessment: A Systematic Literature Analysis. *Model. Earth Syst. Environ.* **2021**, *7*, 669–681. [CrossRef]
30. Hameed, M.; Sharqi, S.S.; Yaseen, Z.M.; Afan, H.A.; Hussain, A.; Elshafie, A. Application of Artificial Intelligence (AI) Techniques in Water Quality Index Prediction: A Case Study in Tropical Region, Malaysia. *Neural Comput. Appl.* **2017**, *28*, 893–905. [CrossRef]
31. 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.; et al. Towards a Time and Cost Effective Approach to Water Quality Index Class Prediction. *J. Hydrol.* **2019**, *575*, 148–165. [CrossRef]
32. Dezfooli, D.; Hosseini-Moghari, S.-M.; Ebrahimi, K.; Araghinejad, S. Classification of Water Quality Status Based on Minimum Quality Parameters: Application of Machine Learning Techniques. *Model. Earth Syst. Environ.* **2018**, *4*, 311–324. [CrossRef]
33. Abba, S.I.; Pham, Q.B.; Saini, G.; Linh, N.T.T.; Ahmed, A.N.; Mohajane, M.; Khaledian, M.; Abdulkadir, R.A.; Bach, Q.-V. Implementation of Data Intelligence Models Coupled with Ensemble Machine Learning for Prediction of Water Quality Index. *Environ. Sci. Pollut. Res.* **2020**, *27*, 41524–41539. [CrossRef]
34. Kouadri, S.; Elbeltagi, A.; Islam, A.R.M.T.; Kateb, S. Performance of Machine Learning Methods in Predicting Water Quality Index Based on Irregular Data Set: Application on Illizi Region (Algerian Southeast). *Appl. Water Sci.* **2021**, *11*, 190. [CrossRef]

35. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer: New York, NY, USA, 2013; Volume 112.
36. Flores Díaz, A.C.; Bollo Manent, M.; Hernández Santana, J.R.; Montaño Salazar, R.; Morales Manilla, L.M.; Ortiz Rivera, A.; Hillon Vega, Y.T.; Lemoine Rodríguez, R.; Bautista Andalón, M.; Amador García, A.; et al. Situación Ambiental de La Cuenca Del Río Santiago-Guadalajara 2017. Available online: https://www.researchgate.net/publication/325654707_Situacion_ambiental_de_la_cuenca_del_Rio_Santiago_Guadalajara (accessed on 12 October 2021).
37. Belmont, L.S. Ciudad e Industria En La Zona Metropolitana de Guadalajara: Un Caos Que Consume La Cuenca Del Río Santiago. *Ciudad Paz-ando* **2016**, *9*, 55–70. [CrossRef]
38. Rizo-Decelis, L.D.; Andreo, B. Water Quality Assessment of the Santiago River and Attenuation Capacity of Pollutants Downstream Guadalajara City, Mexico. *River Res. Appl.* **2016**, *32*, 1505–1516. [CrossRef]
39. Sistema de Calidad Del Agua-CEA Jalisco. Available online: http://info.ceajalisco.gob.mx/sca/ (accessed on 13 August 2021).
40. Jayalakshmi, T.; Santhakumaran, A. Statistical Normalization and Back Propagation for Classification. *Int. J. Comput. Theory Eng.* **2011**, *3*, 1793–8201.
41. Valentini, M.; dos Santos, G.B.; Muller Vieira, B. Multiple Linear Regression Analysis (MLR) Applied for Modeling a New WQI Equation for Monitoring the Water Quality of Mirim Lagoon, in the State of Rio Grande Do Sul—Brazil. *SN Appl. Sci.* **2021**, *3*, 70. [CrossRef]
42. Ahmed, U.; Mumtaz, R.; Anwar, H.; Shah, A.A.; Irfan, R.; García-Nieto, J. Efficient Water Quality Prediction Using Supervised Machine Learning. *Water* **2019**, *11*, 2210. [CrossRef]
43. Azhar, S.C.; Aris, A.Z.; Yusoff, M.K.; Ramli, M.F.; Juahir, H. Classification of River Water Quality Using Multivariate Analysis. *Procedia Environ. Sci.* **2015**, *30*, 79–84. [CrossRef]
44. McCulligh, C.; Tetreault, D.; Martínez, P. Conflicto y Contaminación: El Movimiento Socio-Ecológico En Torno al Río Santiago. In *Gobernanza y Gestión del Agua en el Occidente de México: La Metrópoli de Guadalajara*; ITESO: Tlaquepaque, Mexico, 2012; pp. 129–172.
45. Duarte Silva, A.P. Discarding Variables in a Principal Component Analysis: Algorithms for All-Subsets Comparisons. *Comput. Stat.* **2002**, *17*, 251–271. [CrossRef]