Universidad de Lima

Facultad de Ingeniería y Arquitectura

Carrera de Ingeniería de Sistemas

MACHINE LEARNING: COMPARISON OF ALGORITHMS FOR DETERMINING WATER QUALITY IN THE RÍMAC RIVER

Tesis para optar el Título Profesional de Ingeniero de Sistemas

Juan Miguel Marroquin Peralta

Código 20141988

Asesor

Yvan Jesus Garcia Lopez

Lima – Perú

Octubre de 2021

Research Article

Machine Learning: Comparison of Algorithms for Determining Water Quality in the Rímac River

Marroquin-Peralta J. M.[1], Garcia-Lopez, Y.J.[2], Taquia-Gutierrez, J.A.[3]
[1]20141988@aloe.ulima.edu.pe
Faculty of Engineering and Architecture
University of Lima, Perú
Av. Javier Prado Este 4600, Santiago de Surco, 15023, Perú
[2]ygarcia@ulima.edu.pe
https://orcid.org/0000-0001-9577-4188
Faculty of Engineering and Architecture
University of Lima, Perú
Av. Javier Prado Este 4600, Santiago de Surco, 15023, Perú,
[3]jtaquia@ulima.edu.pe
http://orcid.org/0000-0002-1711-6603
Faculty of Engineering and Architecture
University of Lima, Perú
Av. Javier Prado Este 4600, Santiago de Surco, 15023, Perú

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 23 May 2021

Abstract: The evaluation of the quality of the water in rivers is necessary to manage the efficiency of its use, being necessary to carry out physicochemical and biological analyzes to determine its healthiness, but it implies in its determination of a series of parameters that use various analytical methods that often they are tedious and time consuming to calculate. The present study makes a comparison of machine learning models such as Multiple Linear Regression (MLR), Neural Network Backpropagation (BPNN) and Support Vector Regression (SVR) to estimate Dissolved Oxygen (DO) and Biochemical Oxygen Demand (BOD) to determine the quality of the water of the Rímac river. Water samples were collected from 26 stations and non-point sources of contamination along the Rímac River with 624 records made during the years 2010 to 2012. The physical and chemical parameters introduced in the models include pH, turbidity, total dissolved solids, temperature, electrical conductivity, dissolved oxygen, biochemical oxygen demand, chemical oxygen demand, hardness, chloride, sulfate, calcium, magnesium, and nitrate. The dependent variables of the output models include biochemical oxygen demand (BOD) and dissolved oxygen (DO). The independent variables that were selected for the BOD, these were: pH, EC, turbidity, Nitrites, TOC, COD, iron, and chlorides. For DO, they were temperature, Nitrites, COD, Nitrates, STD, Chlorides and Total Solids. Both dependent parameters have 8 independent variables and the highest correlation coefficient values. The models were trained for learning and validation of 70% and 30% of the data set, respectively. The BPNN presented for the estimation of BOD, with 16 hidden nodes, values of $R^2 = 0.857$ for training and 0.481 for the test phase; For the estimation of DO, with 8 hidden nodes, this was $R^2 = 0.768$ in training and test phase of 0.605. These values were higher than the MLR and SVR, which showed that the BPNN was the best selection. Finally, the classification of water quality as Good, Fair and Poor obtained a precision of 0.88 with a sensitivity of 0.86 and an f1-score of 85%, which evidenced its effectiveness when carrying out this process.

Keywords: Water quality, artificial neural network, multiple linear regression, support vector regression, Rímac river

1. INTRODUCTION

In Peru, in several provinces of the Coast region you can find different valleys with cities, rural areas, agricultural areas, among other sectors. Over the years, these sectors have been growing; therefore, there is an increase in the demand for water and its quality. But one of the main problems in Peru is the quality of the water; since it is considered that there is an annual discharge of 960.5 million cubic meters of drainage on surface, underground and marine water, of which 64.0% belongs to domestic drains, 5.6% industrial drains, 4, 4% from fishery drains, 25.4% from mining effluents and 0.2% from oil effluents (National Water Authority [ANA], 2015).

According to the Chillón Rímac Lurín Water Observatory (2019), for the year 2016, the upper basin of the Rímac River presents regular quality with high levels of metals, mainly: arsenic, manganese, iron and lead. The Huaycoloro river ravine has poor quality, due to high levels of biochemical demand for oxygen, phosphorus, arsenic, iron, and thermotolerant coliforms. This is due to the different anthropic pollutants that cause the deterioration of its quality; for example, wastewater and irrigation dumping, solid waste

dumps, etc. On the other hand, Mayca Zegarra (2019) determined that the water quality of the Rímac river in the Chicla district presents parameters with concentrations below the standards, which are pH, Dissolved Oxygen and Conductivity, in addition, the Manganese metal showed values higher than standard. In addition, the sources of contamination that affect most of these concentrations were identified, that is, domestic residual discharges, industrial residuals from mining, contamination due to accumulation of wastelands, due to the presence of solid waste dumps.

Currently, the ANA (2018) uses a methodology to determine the ICA-PE Water Quality Index at the different monitoring points, based on various parameters, which are carried out along the Rímac River. This methodology gives a qualitative result that is expressed on five scales from poor to excellent; however, it is not a prediction model, since it is carried out with the parameters that are evaluated at that moment. The determination of the water parameters takes between 1 to 10 days after taking the samples.

Artificial neural networks can be used as an effective tool for estimating river water quality and could also be used in other areas to improve understanding of river pollution trends (Adeniran, Adelodun, & Ogunshina, 2016). Therefore, in this study, the Backpropagation neural network (BPNN), Multiple Linear Regression (MLR) and Support Vector Regression (SVR) were constructed to estimate the values of biochemical oxygen demand and dissolved oxygen, individually, using the base of data on the physical and chemical parameters of the Rímac River between 2010 and 2012. This study was carried out independently of the methodology used by the ANA; that is, it was not related to it, proposing a different methodology for estimating water quality. The results obtained were compared between each model. The one with the highest precision was used to determine the level of water quality through a proposed comparative table, which will result in a qualitative value.

2. STATE OF ART

According to Sierra (2011), there are two ways to describe water quality: i) measuring physical, chemical or biological variables, and ii) using water quality indices. The data that are obtained through measurements can be carried out in the field or in laboratories, since, subsequently, these data must be analyzed. The results may vary due to contamination through the direct or indirect introduction of substances or energy to the aquatic environment, causing problems in living organisms, effect on human health, etc.

The application of Machine Learning (ML) models for the prediction of water quality has been carried out in several rivers on different continents; for example, Adeniran et al. (2016) applied backpropagation artificial neural networks (BPNN) for the modeling of biochemical oxygen demand (BOD) and dissolved oxygen (DO) in the Asa river, Nigeria, with correlation coefficient (r) 0.953 and 0.956 respectively; used 6 physical and 9 chemical parameters chosen by statistical analysis. However, Raheli et al. (2017) used the BPNN to estimate the BOD and DO, but applied the firefly optimizer algorithm (FF), showing the effectiveness of this tool when obtaining results of 2.514 mg / L, 35.296%, 0.778 and 0.818 for RMSE, RMSE (%), R and WI (Willmott index). On the other hand, Areerachakul et al. (2011) analyzed the performance of a neural network using the Levenberg-Marquardt algorithm in the channels of the Dusit district, Bangkok, which estimated Dissolved Oxygen by entering 10 parameters; obtaining r equal to 0.84, mean square error (MSE) at 0.78 and mean absolute error (MAE) at 0.7.

Wen et al. (2012) also used the BPNN but applied the Bayesian regularization training algorithm to estimate the same parameter in the Heihe River. The results obtained were r of 0.9654, 0.9841 and 0.9680, and RMSE of 0.4272, 0.3667 and 0.4570 for training, validation, and testing, respectively. Contrary to Omar (2017) who used only the BP to estimate the DO with R values of 0.885, 0.869 and 0.885 for the training, validation, and testing phase, followed by the MSE with values of 1.031, 0.143 and 0.133 for the phases indicated respectively.

The presence of several ML models can give us different results among them, prevailing those that show greater precision; To achieve the model that provides the best results, a comparison is made between them, where Olyaie et al. (2017) made the comparison between neural networks (Backpropagation and Radial Base Function), Linear Genetic Programming (LGP) and Support Vector Machine (SVM) for the estimation of Dissolved Oxygen in the Delaware River, resulting in the SVM with the highest value of r = 0.991 to that of LGP, 0.934. Another case where SVM stood out was in the Wen-Rui Tang River by Ji et

al. (2017), when compared with the multiple linear regression (MLR), the BPNN and the generalized regression neural network (RNRG), where the SVM obtained the values of 0.9416 mg / L, 0.8646 and 0.8763 for MSE, R2, and NS; furthermore, that ammonia nitrogen was the most significant variable.

For the case of reservoirs, Chen and Liu (2013) compared the Backpropagation and ANFIS (Adaptive Neural-based fuzzy inference system) neural networks against the MLR, where ANFIS showed better r results with values of 0.89 for the training phases and then the other two models for the prediction of DO in the Feitsuit reservoir, Taiwan. However, Nemati et al. (2015) used these last three methods to also predict DO in the Tai Po River, Hong Kong; with the difference that the evaluation of the performance of the models was added the Nash-Sutcliffe efficiency criterion; giving as results of r for the training and testing phase, 0.796 and 0.798 in the case of ANN as having better results than MLR and ANFIS.

Csábrágia et al. (2019) applied the Multiple Linear Regression (MLR) model, Radial Base Function neural network (RBF) and Generalized Regression neural network (GRNN) to discover which is the best method to predict OD. They included the runoff parameter as an input value and proposed three types of configurations in the training and test data of the Tisza River. As a result, the GRNN had a better result of r2 = 0.31 than the other models. Antanasijević et al. (2013) also used the GRNN but compared it with the Backpropagation neural network (BPNN) and the Recurrent neural network (RNN) for the prediction of OD in the Danube River. However, the GRNN model had a high value of r2 (0.99), much better than RNN (0.87) and BPNN (0.83). Another case study where they used GRNN, was applied by Ji et al. (2017), comparing it with the BPNN and the SVM to estimate the DO, with the difference that this last model had greater precision when obtaining results of MSE = 0.9416 mg / L, R2 = 0.8646 and NS = 0.8763 in the test phase.

On the other hand, Sarkar and Pandey (2015) had configurations in the training data, proposing three neural networks depending on the three sets of input variables that would later enter the Backpropagation network, choosing the case study the Yamuna River, city of Mathura, India. The highest performance was obtained by the second configuration where the values of RMSE, R, DC were 1.71, 0.907, 0.822 respectively during training and 1.52, 0.928, 0.856 during the test phase. Xiao et al. (2017) also used BPNN, with the difference of the combination of the activation functions purelin, logsig and tansig for the estimation of OD, which was later compared with the autoregression, gray model and SVM, where the RN obtained estimated values with lower at 5% of the limit error.

Ahmed (2017) evaluated the possibility of predicting DO with only two input parameters, Biochemical Oxygen Demand (BOD) and Chemical Oxygen Demand (COD), in the Surma River, Bangladesh. He used the RBF neural network and the Pre-powered network; where, Ahmed confirmed that DO can be predicted with a small number of variables, yielding acceptable precision with values of correlation coefficient (0.944), MSE (1.009) and E (0.966) for the RBFN model. Additionally, Ahmed and Shah (2017), estimated the BOD using the ANFIS model considering 10 water parameters as independent variables, where they obtained for the test phase the values of 1.01, 1.67, 83.12 and 0.885 for MAE, MSE, EFF and R.

Iglesias et al. (2014) did not predict DO, they focused on predicting Turbidity, since it is a key variable for water quality. Using the Backpropagation neural network, they concluded that temperature is the parameter most influenced by turbidity, and network performance increases when a synergistic variable with r = 0.8 is additionally entered. Wang et al. (2017) refer to the application of the LSTM neural network for water quality prediction, which was then compared with the Backpropagation neural network (BPNN) and the Online Sequential Extreme Machine Learning (OS-EML) model. The chosen model was used to predict Dissolved Oxygen (DO) and Total Phosphorus (FT), where the RN LSTM obtained as a result of a correlation coefficient of 0.89, and its RMSE value was lower than that of the other models, in conclusion, this neural network is more generalized. On the other hand, Zhu and Heddam (2019) compared the multilayer perceptron neural network (MLPNN) and extreme learning machine (ELM) for the estimation of DO in three different rivers, showing that these models can be applied in rivers with less impact, as long as it does not have a higher concentration of contamination since it is difficult for the models to consider the impact of anthropogenic influences.

Bayram and Kankal (2015) applied neural networks to estimate DO, comparing it with regression analysis (RA). They used the multilayer neural network with results of RMSE = 0.623 mg / L and MAE

= 0.710 mg / L against the RA values that were 0.967 mg / L and 0.787 mg / L respectively. Therefore, the multilayer NR presented values closer to 0, interpreting it as the most feasible. Finally, Csábrági et al. (2017) adapted four models for the estimation of OD, these were multiple linear regression (MLR), multilayer perceptron neural network (MPNN), radial function neural network (RFNN) and generalized regression neural network (GRNN), where they agreed on that non-linear models showed better results than linear models. The GRNN showed better performance and RFNN better results with the RMSE, MAE, IA and R2 indicators.

The studies analyzed in this section helped to understand how the DO and BOD estimation could be carried out, since the authors consider that they are essential values to determine the water quality in rivers.

3. BACKGROUND

The quality of the water is one of the main characteristics of a river, the quality of the water must be simulated and predicted through mathematical models. If the predicted quality is not satisfactory, some changes or precautions should be implemented. To avoid this unwanted trend, controlling water pollution has become essential to maintain the sustainability of water resources. Water quality can be assessed using a series of critical parameters carefully selected to represent the level of contamination of the water body of interest and reflect its overall water quality status. It is for this reason that mathematical models are used, which allow a more adequate evaluation of the behavior of water quality in a natural stream. In addition, the models allow the creation of fundamental future scenarios for the planning and proper management of natural resources. The mathematical models applied in the modeling of water to determine its quality are non-linear and complex, naturally due to the processes to be replicated. Today, with the use of machine learning algorithms it has simplified their use for the determination of water quality.

3.1 Streeter-Phelps Water Quality Model

There is a great variety of water quality models, with which it is possible to establish a discharge behavior in a receiving water body. The main differences lie in the transport processes (advection, dispersion) and reaction (transformation of the water quality determinants) that they include and therefore in the assumptions they make. Harold Streeter and Earle Phelps carried out studies in the Ohio River between 1914 and 1916 to be able to carry out the mathematical modeling of Dissolved Oxygen. This mathematical model is related to two main mechanisms that define dissolved oxygen in a surface water channel that receives wastewater discharge: (i) Oxidation of biodegradable organic matter, and (ii) Reaeration of oxygen (Joaquín Suárez, 2008).

3.1.1. Biodegradable organic matter oxidation

The Biochemical Oxygen Demand is a chemical parameter that allows to determine the content of organic matter in a water sample. This parameter is measured by how much oxygen is required by microorganisms to degrade, oxidize, stabilize, etc. organic matter. The results are given in milligrams per liter (mg / L) of oxygen consumed (Sierra Ramírez, 2011).

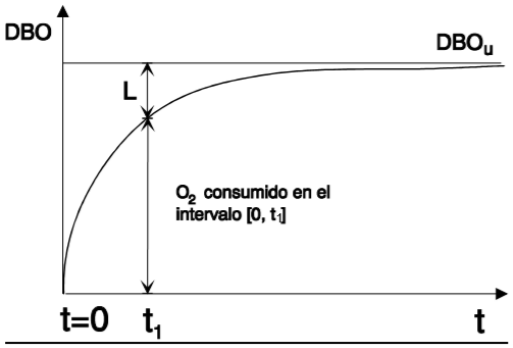

Figure 1. Biochemical oxidation
Source: Sierra Ramírez, 2011

Figure 1 shows the biochemical oxidation process, where the *BOD* exerted (oxygen consumed) per unit of time due to the oxidation of organic matter is represented, which is directly proportional to the amount of organic matter that remains to be oxidized in a certain moment.

$$\frac{dOD}{dt} = -K_1.L$$

where K1 is the degradation parameter that varies with the type of water and with the degree of wastewater treatment. (Joaquín Suárez, 2008).

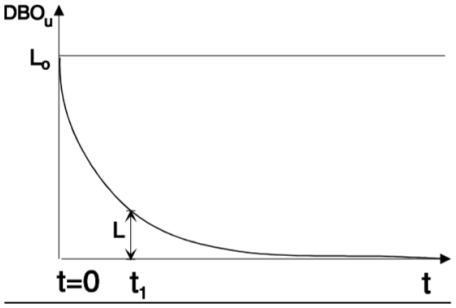

Figure 2. *BOD* remnant in time
Source: Sierra Ramírez, 2011

If what you want is to demonstrate the remaining *BOD* (oxygen not consumed) in time, figure 2 shows this, then the equation would be as:

$$\frac{dL_t}{dt} = -K_1.L_t$$

where *L* is the *BOD* not consumed at instant *t* (Joaquín Suárez, 2008).

If we integrate the equation; therefore, the formula to determine the amount of *BOD* exerted at instant t is as follows:

$$DBO_t = DBO_u(1 - e^{-tK_1})$$

3.1.2. Surface Reaeration

The low concentration of Dissolved Oxygen (DO) is reflected in an unbalanced ecosystem, fish kills, odors and other aesthetic annoyances. It is considered the most important parameter in the diagnosis of the state of contamination of an aquatic ecosystem. The problem can be summarized as the dumping of organic and inorganic waste into the water causing the decrease in DO which interferes with beneficial uses of the water. The most important discharges are the discharges of industrial waste and domestic sewage, oxidizable forms of nitrogen and nutrients that stimulate the growth of phytoplankton (Sierra Ramírez, 2011).

DO concentration reflects the balance between oxygen production and oxygen consumption, processes in the aquatic ecosystem. This depends on many factors such as temperature, salinity, oxygen depletion, oxygen source, and others. The DO level is the health criterion, which is frequently used for the control of water quality in different aquatic ecosystems such as reservoirs and wetlands (Olyaie, Zare Abyaneh, & Danandeh Mehr, 2017).

To determine the degree of contamination by biodegradable organic matter in a river, the amount of dissolved oxygen must be measured.

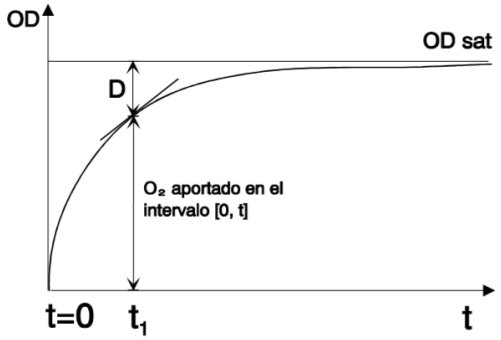

Figure 3. Surface Reaeration
Source: Sierra Ramírez, 2011

3.2. Multiple Linear Regression

The MLR learns the relationship between dependent variables and a dependent one. It offers simple and easily interpretable models; however, it can lead to inaccurate models that predict poorly in the presence of a non-linear or non-additive relationship. In the linear case, the functional relationship between the dependent variable and its predictors is estimated by minimizing the residual sum of squares (Nemati 2015).

The best MLR equation is based on the high multiple correlation coefficient, the lowest standard deviation, and the magnitude of the F-radius. The general model of the MLR is expressed as:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_q x_{qi}$$

where yi is the dependent variable, n is the sample size and $i = 1,\ldots, n$; $x_1, x_2,\ldots, x_q$ are the dependent variables; $x_{1i}, x_{2i},\ldots, x_{qi}$ are the measured values; β_0 is a constant; and $\beta_1, \beta_2,\ldots, \beta_3$ are the multiple regression coefficients (Ouma et al., 2020)..

3.3. Neural Network Backpropagation

It is a forward feeding neural network with one or more hidden layers, which are between the input and the output. It is the most popular type of network that generally uses the backward propagation error technique to train the network configuration. The meaning of feedforward is that the data flows in one direction from the input layer to the output layer. This type of network is trained with the backpropagation learning algorithm. (Olyaie et al., 2015). The single layer BPNN training process is relatively straightforward because the loss function is applied as a direct function of the weights, allowing a simple gradient to be applied. For multilayers, the loss is a complicated function made up of the weights of the nearby layers. The backpropagation algorithm uses the differential calculation chain rule, which calculates the error gradient in terms of sums of squares of local gradient over the various paths from a node to the output (Aggarwal, 2018).

The output of a neuron (Ahmed, 2017) can be expressed as: output = $f(n)$

Where:

$$n = \sum_{j=1}^{R} \omega_j x_j + b$$

$x_1, x_2,\ldots, x_R$ are the input values; $\omega_1, \omega_2,\ldots, \omega_R$ are the weights of the neuron; b is bias value and $f(n)$ is activation function. The activation function ReLu (Rectified Linear Unit) is used in the architecture of the neural network for the hidden layers (Aggarwal, 2018), expressed as: $f(n) = max\{n, 0\}$

The output y of the output node can be calculated as:

$$y = \sum_{i=0}^{z} \left(\omega_{j,i(2)} max\left(\sum_{j=1}^{R} x_{j\omega_{i,j(1)}} + b_{i(1)}, 0 \right) \right) + b_{1(2)}$$

where R is the number of input nodes, z is the number of hidden nodes, $\omega_{i,\ j\ (1)}$ is the weight of the first layer between the input j and the hidden i, $\omega_{i,\ j(2)}$ the weight of the second layer between hidden neuron i and exit neuron, $b_{i(2)}$ is the biased weight for hidden neuron i and $b_{1(2)}$ is the biased weight for exit neuron.

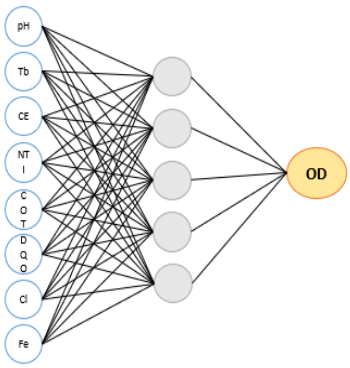

Figure 4. Neural network Model
Source: Own elaboration

3.4. Support Vector Regression

In principle, SVM was developed to solve classification problems, later its use has been extended to applications of regression by estimation functions, generating the Support Vector Regression (SVR). It estimates an output variable based on a set of input variables. The development includes the principle of Structural Risk Minimization (SRM), which is superior to the traditional Empirical Risk Minimization (ERM) principle, used for conventional neural networks. The SRM minimizes the upper limit of the expected risk, opposite to the other principle that minimizes the error of the training data (Olyaie et al., 2015).

The regression of the SVM can be expressed as:

$$f(n) = w \cdot \phi(x) + b$$

where w is the weight vector, $\phi\ (x)$ the nonlinear transfer function and b the bias. The coefficients of w and b can be determined by minimizing the regularized hazard function:

$$\text{Minimize: } \frac{1}{2}\|w\|^2 + c\sum_i^n(\varepsilon_i; \varepsilon_i *)$$

$$\text{Subject to } \begin{cases} y_i - w \cdot \phi(x) - b \leq \varepsilon + \varepsilon_i \\ w \cdot \phi(x) + b - y_i \leq \varepsilon_i + \varepsilon_i * \\ \varepsilon_i; \varepsilon_i *, \ \ i = 1, 2, 3, \ldots, n \end{cases}$$

where c is the regularization parameter, εi and εi * are stationary variables. To solve the optimization problem, the Lagrangian multipliers are used:

$$\text{Maximize: } -\frac{1}{2}\sum_{i=1}^n\sum_{i=1}^n(a_i - a_i *)(a_j - a_j *)K(x_i, x_j) - \sum_{i=1}^n(a_i - a_i *) + \sum_{i=1}^n(a_i - a_i *)y_i$$

$$\text{Subjet to } \sum_{i=1}^n(a_i - a_i *),\ 0 \leq a_i \leq C \ \ 0 \leq a_i * \leq C$$

where $K(x_i,x)$ is a function of .

Imposing the Karush-Kuhn-Tucker *(KKT)* optimizing condition, w * is obtained as:

$$w^* = \sum_{i=1}^{n}(a_i - a_i *) \cdot K(x_i, x)$$

Therefore, the SVR expression is:

$$f(x) = \sum_{i=1}^{n} (a_i - a_i *) \cdot K(x_i, x) + b$$

For the kernel function four options can be applied, these are: polynomial, sigmoid, linear and RBF. The latter maps the input vectors into a non-linear multi-dimensional space; in addition, it could model complicated non-linear relationships. The RBF is easier to use than the polynomial and sigmoid kernels, due to the fewer tunable parameters available. Its formula is:

$$K(x_i, x) = exp(-g\|x_i - x\|^2)$$

where *g* is the adjustable kernel parameter (Ji et al., 2017; Olyaie et al., 2016).

3.6. Performance Evaluation

To evaluate the performance of the explained models, some statistical measurements should be considered depending on the models that are chosen. Those measurements considered were correlation coefficient (R), root mean square error (RMSE), mean absolute error (MAE) and Nashe-Sutcliffe coefficient of efficiency.

The R measures the degree of correlation between the experimental variable and the measure (Ahmed, 2017). This is in the range of -1 to 1. If R = 0, there is a non-linear relationship, on the other hand, if R = 1 or -1, there is a perfect positive or negative linear relationship (Olyaie et al., 2017). The equation is:

$$R = \frac{\sum_{i=1}^{n}(DO_m(i) - \overline{DO}_m)(DO_e(i) - \overline{DO}_e)}{\sqrt{\sum_{i=1}^{n}(DO_m(i) - \overline{DO}_m)^2 \sum_{i=1}^{n}(DO_e(i) - \overline{DO}_e)^2}}$$

The RMSE measures the number of squares of errors between the estimated and measured value (Ahmed, 2017). It is calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(DO_e(i) - DO_m(i))^2}{n}}$$

The MAE verifies the robustness of the implemented model, offers an indication about the overestimation and underestimation of a model (Omar, 2017).

$$MAE = \frac{\sum_{i=1}^{n}|DO_e(i) - DO_m(i)|}{n}$$

The NS has been used to evaluate the performance of hydrological models (Ahmed, 2017). It is an alternative to R as a measure of relative error where it is sensitive to the differences between the measured and estimated mean and variance (Olyaie et al. 2017):

$$NS = 1 - \frac{\sum_{i=1}^{n}(DO_m(i) - DO_e(i))^2}{\sum_{i=1}^{n}(DO_m(i) - \overline{DO}_m)^2}$$

where $DO_m\ (i)$ and $DO_m\ (i)$ are the measured and estimated *DO*, respectively. The *DO* is the mean, and n is the number of data considered.

4. METHODOLOGY

The methodology seeks to determine the water quality of a point in a given place in a riverbed. The first point that was made was the identification of the river, which would form the case study, since the information with the water parameters that the entities have in their databases over several years will be extracted from that place. After obtaining the information, the parameter selection analysis was carried out, since not all the parameters are essential to the model. After that, they entered the three models chosen to meet the main objective, these are Multiple Linear Regression (MLR), Neural Network Backpropagation (BPNN) and Support Vector Regression (SVR), implemented in Python language with Jupyter. These models had a particular objective, the estimation of Dissolved Oxygen and Biochemical Oxygen Demand, independently. However, the results of the three models were not taken, but the one

with the highest precision in its results and the best reliability was chosen.The base of the two objective parameters of the chosen model was obtained. These two parameters went through a classification table prepared based on the Water Quality Standards (ECA), this classification is 4 categories of water quality: awfully bad, bad, good, incredibly good. This classification corresponds to each sample carried out. Figure 6 shows the methodology in a simplified and structured way.

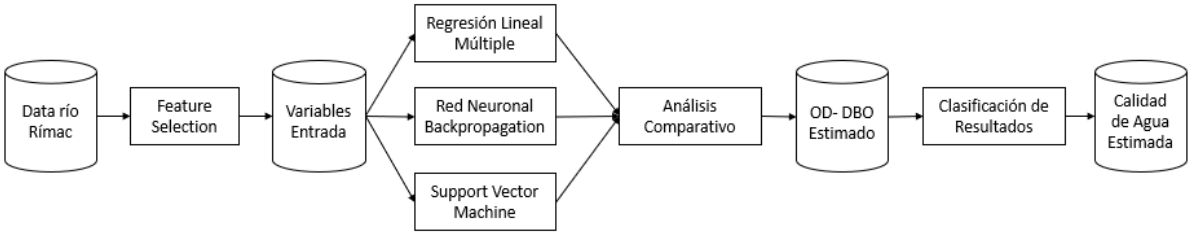

Figure 5. Proposed methodology for estimating water quality
Source: Own elaboration

4.1. Study area

The Rímac River is born on the western slope of the Andes Mountains in the Pacay mountain range at a maximum altitude of 5,508 m.a.s.l. approximately 132 km northeast of the city of Lima, flowing into the Pacific Ocean, through Callao (Chillón Rímac Lurín Water Observatory, 2019). It is the most important river in the department, being the capital of the republic the main consumer of surface water and aquifer. In figure 6 we observe the hydrographic map of the Rímac river basin, where mining settlements are verified along its channel, causing contamination in its waters.

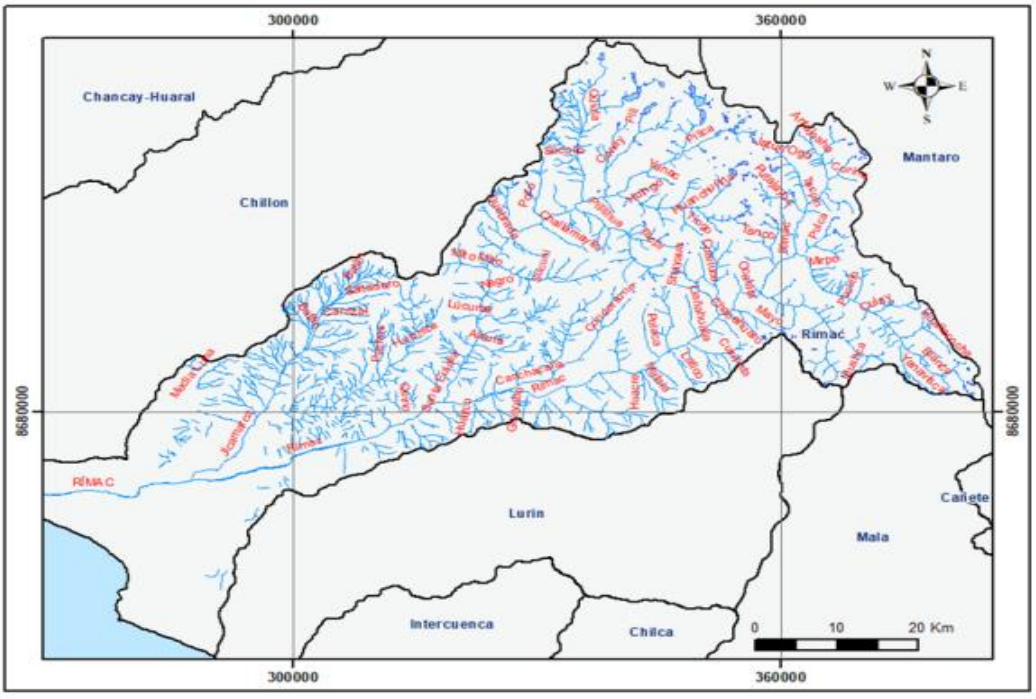

Figure 6. Hydrographic map of the Rímac river basin
Source: Ministry of Energy and Mines (MEM)

In addition, there is a presence of mining activity in the upper basin, evacuating some discharges directly to the river (General Directorate of Environmental Health, [DIGESA], 2011). DIGESA monthly monitors the Rímac river basin with 26 stations, where samples are evaluated with up to 32 physical, chemical and biological parameters such as dissolved oxygen, biochemical oxygen demand, arsenic, cadmium, copper, total chromium, iron, manganese. , lead, mercury, zinc, hydrogen potential (pH), thermotolerant coliforms, among others. The set of parameters used for this study consists of 624 records between the years 2010 and 2012. The values of this set are quantitative, there are no qualitative values.

4.2. Parameter selection

Dutta and Chaki (2012) provide us with which are the parameters that affect water quality and are essential in the application of the different Machine Learning models. These are: pH, DO, BOD, Total dissolved solids, Electrical Conductivity, Specific Conductance, Water level, Temperature, Turbidity, Tidal effect, Calcium, Magnesium, Nitrate, Phosphate, Sulfate, Ammonia, Chloride, rainfall, mining, and industry.

In the case of the Rímac River, the selection of parameters was made not by the identification carried out in other studies but based on the relationship that these parameters maintain with the dependent variables. First, those parameters that did not have a sufficient amount of data to enter the model were removed, leaving only the following: pH, temperature (° C), specific conductivity (EC), turbidity, total nitrogen (NT), Nitrites (NO2-), Nitrates (NO3-), Chloride (Cl-), Total Organic Carbon (TOC), Chemical Oxygen Demand (COD), Total Solids, Total Suspended Solids (TSS), Total Dissolved Solids (TDS), Aluminum, and Iron, among others.

Table 1. Basic statistics of the water quality parameters in the Rímac River.2010-2012

Variables	**Statistical Measure**				**Correlation Pearson**	
	Mean	**Std deviation**	**Minimum**	**Maximum**	**BOD**	**DO**
Temperature	14.257	5.898	4.39	29.6	**0.573**	-0.130
Nitrite	0.048	0.183	0	1.986	**0.411**	**-0.246**
TOC	2.984	4.856	0.3	41.27	**0.292**	**-0.327**
COD	28.062	52.12	4	472	**0.401**	**-0.181**
Nitrate	0.692	0.548	0.123	5.046	**0.413**	-0.143
C.E.	583.692	318.269	11	1890	0.250	**-0.300**
STD	463.526	270.792	120	1495	**0.350**	-0.151
Chlorides	22.493	61.559	0.31	673.5	**0.312**	**-0.175**
Turbidity	124.976	207.724	0	1205	0.281	**0.190**
Total Solids	549.032	325.481	148	1630	**0.373**	-0.066
NT	1.403	1.702	0.12	14.02	0.278	-0.141
Iron	4.359	10.781	0.03	61.35	0.182	**0.215**
Aluminum	1.292	1.859	0.045	18.08	0.180	-0.131
STS	85.506	148.957	5	854	0.178	0.131
pH	8.81	0.51	7.13	9.86	-0.047	**-0.224**
Zinc	1.118	2.181	0.009	23.51	-0.160	0.070
Arsenic	0.027	0.021	0.002	0.145	-0.123	0.023
Manganese	1.501	5.001	0.013	36.6	-0.127	0.000

Own elaboration

Table 1 shows the statistical data of these parameters. However, as there is many independent variables, those that maintain the highest correlation coefficient with each dependent variable were chosen, having a total of seven independent variables for each model. The independent variables with a correlation coefficient greater than 0.290 were selected for the BOD, these were: pH, EC, turbidity, Nitrites, TOC, COD, iron and chlorides. For DO, the variables had a correlation coefficient greater than ± 0.170. which were: temperature, Nitrites, COD, Nitrates, TDS, Chlorides and Total Solids. Both dependent parameters have 8 independent variables with the highest correlation coefficient values. Figure 7 shows the correlation coefficients of the water parameters classified between selected and not selected.

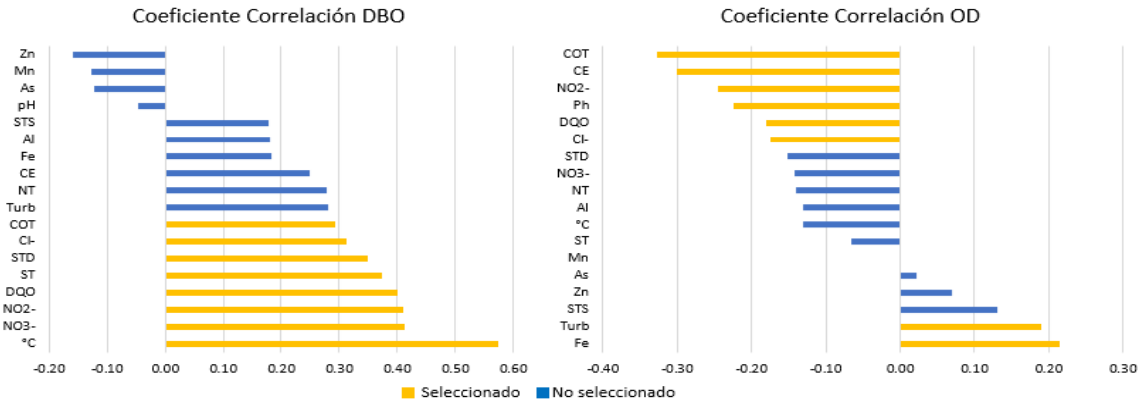

Figure 7. Correlation coefficients between parameters and independent variables
Source: Own elaboration

4.4. Construction of DO and BOD estimation models.

The dataset went through a normalization process for greater benefit and rapid convergence. Feature Scaling was used, where the normalized value is calculated from the division of the subtraction of the unnormalized value and its minimum value, and the subtraction of the maximum and minimum value of the attribute set. The result of the normalization was in a range between 0 and 1. After that, the data were divided into two groups, data train and data test, where the percentage of data each had is 80% and 20% respectively. this division was made randomly.

Python language libraries were used for the implementation of MLR, BPNN and SVR. With respect to BPNNs, the input layer has eight nodes for the DO and BOD architecture, this due to the selection made in the previous section. Ji et al. (2017) indicates that the appropriate number of nodes in the hidden layer should be in the range between (α + ⟦2n⟧ ^ (1/2)) and (2n + 1), where α is the number of output nodes and n, the number of input nodes. Therefore, the determination of the number of hidden nodes was carried out in the range of 5 and 17, through trial and error to obtain the quantity that most favored the prediction of the dependent variable; This layer had the ReLu activation function; finally, the output layer was the estimated value of DO and BOD, independently, the linear activation function was used.

The determination of cost *(C)* and gamma *(g)* is important for increasing SVR performance. Grid Search was applied to obtain these parameters, of which the chosen ones are presented as the optimum of the model. The range of *C* was between 1 and 100, for *g* between 0.0001 and 0.1 with a jump of 0.001, and for epsilon it was between 0.1 and 0.5. The RBF function was used for the kernel.

4.5. Determination of water quality

Supreme Decree No. 004-2017-MINAM indicates that for rivers within Category 1-2A, the DO must be greater than or equal to 5 mg / L, and for the BOD it must be less than 5 mg / L, complying with the Water Quality Standards (ECA). On the other hand, Regalado et al. (2008) indicates that the DO being less than 4 mg / L is considered bad, greater than this and less than 8 mg / L is acceptable and greater than the latter is good. For the BOD, if it is less than 3 mg / L it is exceptionally good, greater than this and less than 5 mg / L is acceptable, and greater than the latter is bad.

Table 2 shows the proposal to classify the quality of the water qualitatively, taking into consideration the cases of the previous paragraph. This is interpreted as follows: if the DO is greater than or equal to 5, it is assigned True, if it is less than 5, False; If the BOD is less than or equal to 5, True is assigned, otherwise, False. If both obtain True, it is qualified as Good quality, if both are False, it is classified as Bad; and if one is different from the other, the classification is Regular. This classification was applied after obtaining the estimated values of DO and BOD of the model that obtained more precise results compared to the other two models.

Table 2. Determination of the proposed water quality.

Water quality	**DO (mg/L)**	**BOD (mg/L)**
	≥ 5	**≤ 5**
Good	True	True
Regular	False True	True False
Bad	False	False

Own elaboration

5. RESULTS

In this study, Machine Learning models were performed to estimate each dependent variable, dissolved oxygen (DO) and biochemical oxygen demand (BOD), through the parameters of Temperature, Nitrites, Total Organic Carbon, Chemical Oxygen Demand, Nitrates, Electrical Conductivity, Total Dissolved Solids and Chlorides, in the Rímac River. These models were developed using Multiple Linear Regression (MLR), Neural Network Backprogation (BPNN) and Support Vector Regression (SVR), implemented in Python language with Jupyter.

Equation (1) was used to estimate the DO value for the training and testing phases of the MLR. The performance criteria of the model for both phases are shown in Table 3. Figure 8 presents the measured and estimated DO values for the mentioned phases. The MAE, RMSE, R and NS values for the training phase were 0.117 mg / L, 0.147 mg / L, 0.633 and 0.400, while the values for the test phase were 0.151 mg / L, 0.195 mg / L, 0.245 and 0.038 respectively.

$$DO = 0.736 - 0.881 \times COT - 0257 \times CE - 0.216 \times NO2 - 0.206 \times pH + 0.039 \times Fe + 0.341 \times Turb - 0.407 \times DQO + 1.122 \times Cl$$

Table 3. Performance criterion of the MLR model

Output	**Training**				**Test**			
	MAE	**RMSE**	**R**	**NS**	**MAE**	**RMSE**	**R**	**NS**
DO	0.117	0.147	0.633	0.400	0.151	0.195	0.245	0.038
BOD	0.106	0.145	0.708	0.502	0.129	0.184	0.473	0.164

Own elaboration

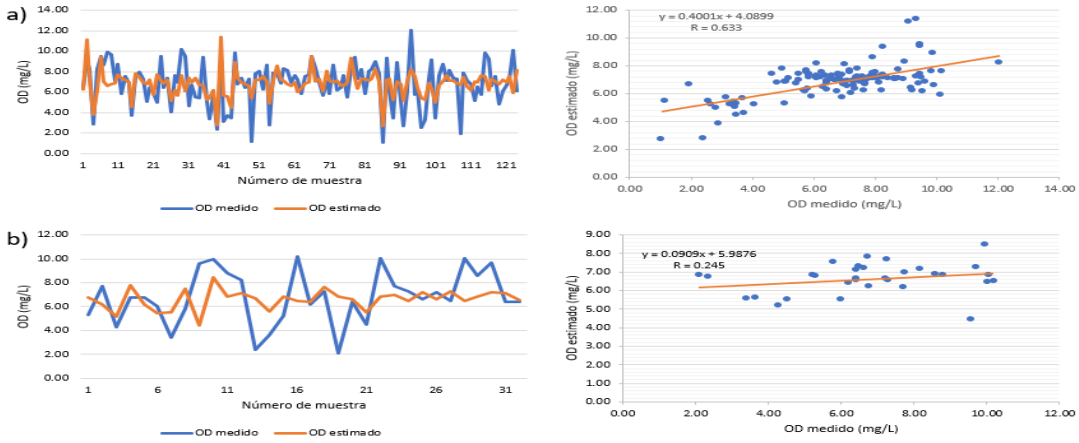

Figure 8. Comparison of graphs of DO measured and estimated using MLR for a training data and b test data.

Source: Own elaboration

For the estimation of the BOD, equation (2) was applied for the training and testing phase. The criteria are also shown in Table 3. The measured and estimated values of the BOD were presented in Figure 9. In the training phase, the MAE, RMSE, R and NS values were obtained, these were 0.106 mg / L, 0.145 mg / L, 0.708 and 0.502, while the values for the test phase were 0.129 mg / L, 0.184 mg / L, 0.473 and 0.164, respectively.

$$BOD = 0.015 + 0.487 \times °C + 0.205 \times NO3 + 0.247 \times NO2 + 0.475 \times DQO - 0.015 \times ST + 0.111 \times STD - 0.840 \times STD + 0.266 \times Cl$$

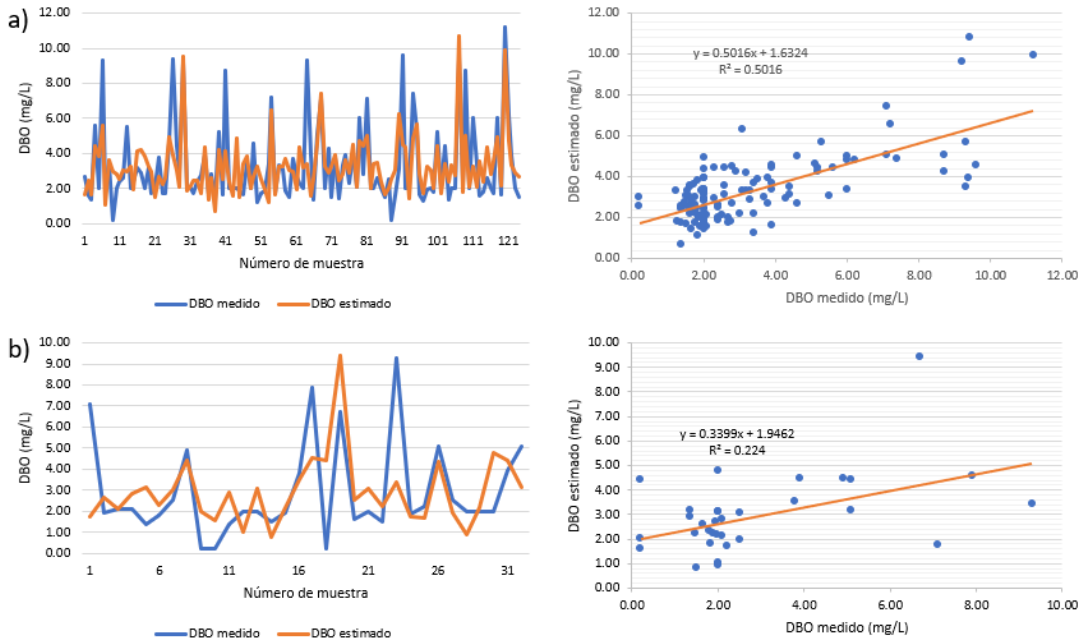

Figure 9. Comparison of graphs of measured and estimated BOD using MLR for a training data and b test data.

Source: Own elaboration

Two BPNN models were built to estimate DO and BOD, independently. For the BPNN DO the activation function ReLu was used in the hidden layer and the linear function between the hidden layer and the output. The network was trained with 1000 epochs and a batch size of 10. Table 4 shows the performance criteria of the different architectures chosen. These architectures were selected from a range of hidden nodes of 5 and 17, where three architectures were highlighted with the best results..

Table 4. BPNN performance criteria for estimating DO.

Layers	**Training**				**Test**			
	MAE	**RMSE**	**R**	**NS**	**MAE**	**RMSE**	**R**	**NS**
8 – 8 – 1	0.082	0.122	0.768	0.583	0.117	0.162	0.605	0.337
8 – 11 – 1	0.078	0.119	0.776	0.602	0.126	0.168	0.540	0.283
8 – 12 – 1	0.083	0.123	0.765	0.579	0.126	0.168	0.548	0.280

Own elaboration

The selection of the BPNN DO (8 - 8 - 1) is due to the most optimal values in the test and training phase. The MAE, RMSE, R and NS values for the training phase were 0.082 mg / L, 0.122 mg / L, 0.768 and 0.583, while in the test phase these were 0.117 mg / L, 0.162 mg / L, 0.605 and 0.337, respectively. Figure 10 shows the comparison between the measured results and the estimates obtained from the model for the DO training and testing phase. The measured and estimated values are compared in Figure 10.

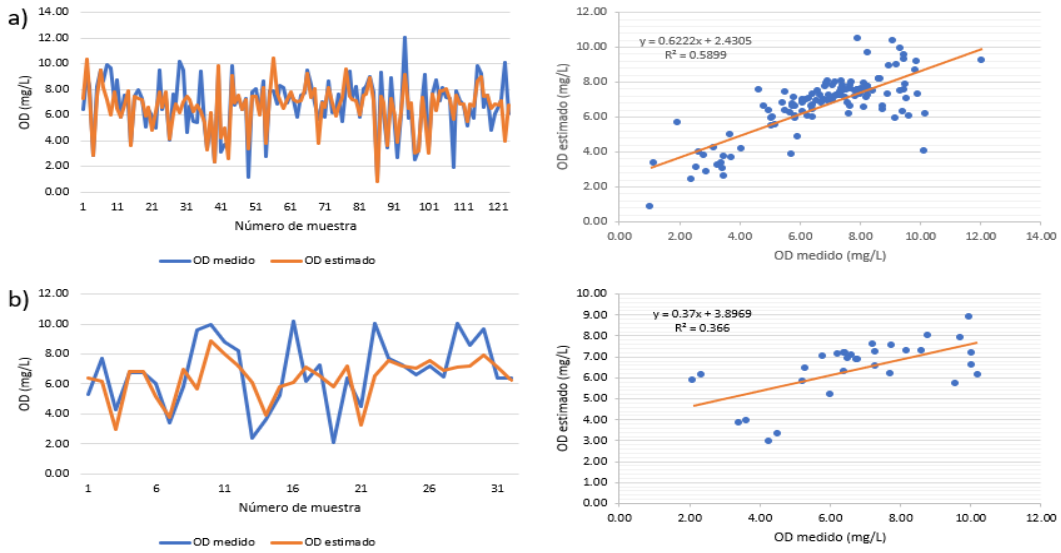

Figure 10. Comparison of graphs of DO measured and estimated using BPNN for a training data and b test data

Source: Own elaboration

The same hyperparameters as the BPNN DO model were taken for the BOD case, where the number of hidden nodes with more precise values were 15, 16 and 17, as shown in Table 5. However, the one that stood out the best between them was 16 hidden nodes. The measured and estimated values of the BOD of this number of hidden nodes were presented in Figure 11. In the training phase, the MAE, RMSE, R and NS values were obtained, these were 0.062 mg / L, 0.106 mg / L, 0.857 and 0.733, while the values for the test phase were 0.111 mg / L, 0.182 mg / L, 0.481 and 0.183 respectively.

Table 5. BPNN performance criteria for estimating BOD.

Layers	Training				Test			
	MAE	RMSE	R	NS	MAE	RMSE	R	NS
8 - 15 - 1	0.071	0.119	0.820	0.661	0.110	0.182	0.498	0.187
8 - 16 - 1	0.062	0.106	0.857	0.733	0.111	0.182	0.481	0.183
8 - 17 - 1	0.066	0.113	0.844	0.695	0.118	0.194	0.406	0.174

Own elaboration

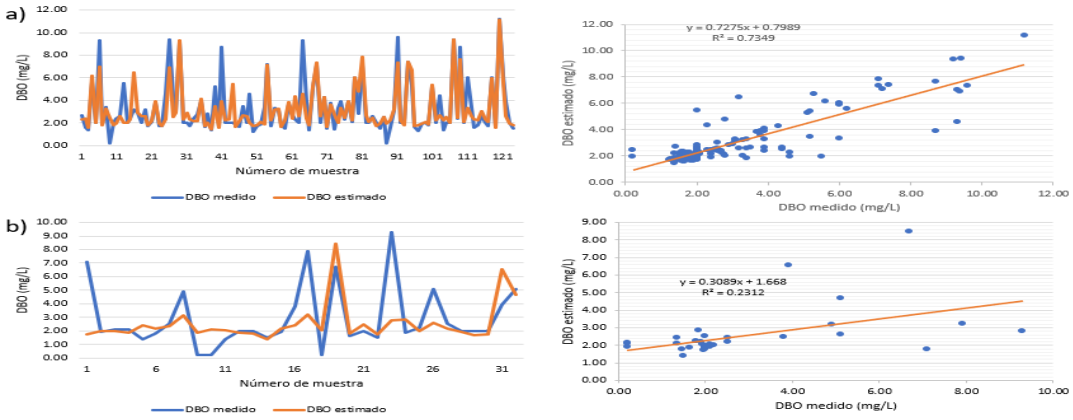

Figure 11. Comparison of graphs of measured and estimated BOD using BPNN for a training data and b test data

Source: Own elaboration

Grid Search was used to determine the SVR hyperparameters, where C = 6, γ = 0.0991 and e = 0.3 were the optimized values for the DO estimation. The performance he had is shown in Table 6. The MAE, RMSE, R and NS values were 0.127 mg / L, 0.161 mg / L, 0.541 and 0.278 for the training phase; On the other hand, for the test phase, these values were 0.144 mg / L, 0.181 mg / L, 0.465 and 0.165. The comparison between the measured and estimated values is observed in Figure 12.

Table 6. SVR performance criteria.

Dependent variable	C, γ, e	Training				Test			
		MAE	RMSE	R	NS	MAE	RMSE	R	NS
DO	6, 0.0991, 0.3	0.127	0.161	0.541	0.278	0.144	0.181	0.465	0.165
BOD	1, 0.0991, 0.1	0.110	0.153	0.694	0.442	0.122	0.181	0.444	0.197

Own elaboration

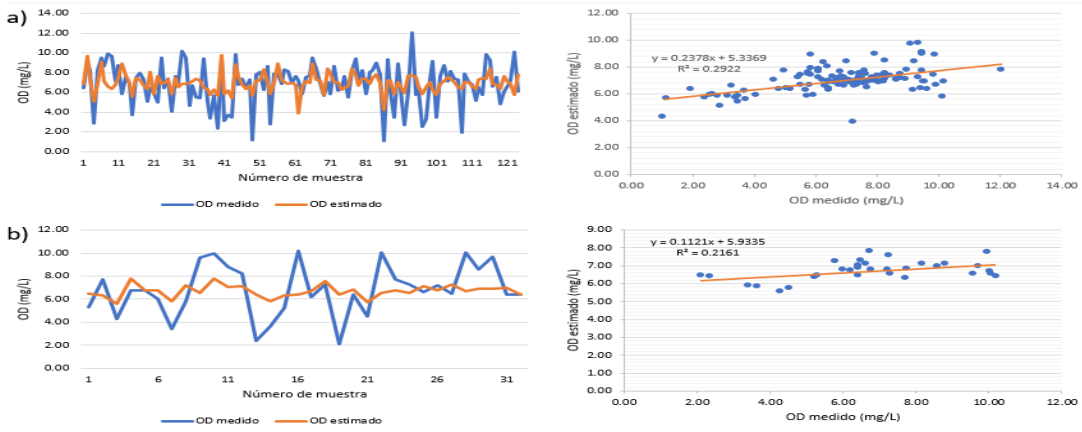

Figure 12. Comparison of graphs of DO measured and estimated using SVR for a training data and b test data

Source: Own elaboration

The BOD estimation also used Grid Search to determine the optimal hyperparameters, their values were C = 1, γ = 0.0991 and e = 0.1. The performance can be observed in Table 6. The MAE, RMSE, R and NS values for the training phase were 0.110 mg / L, 0.153 mg / L, 0.694 and 0.442, for the test phase these were 0.122 mg / L, 0.181 mg / L, 0.444 and 0.197, respectively. Figure 13 shows the graphs of the comparison between the measured and estimated values of the BOD.

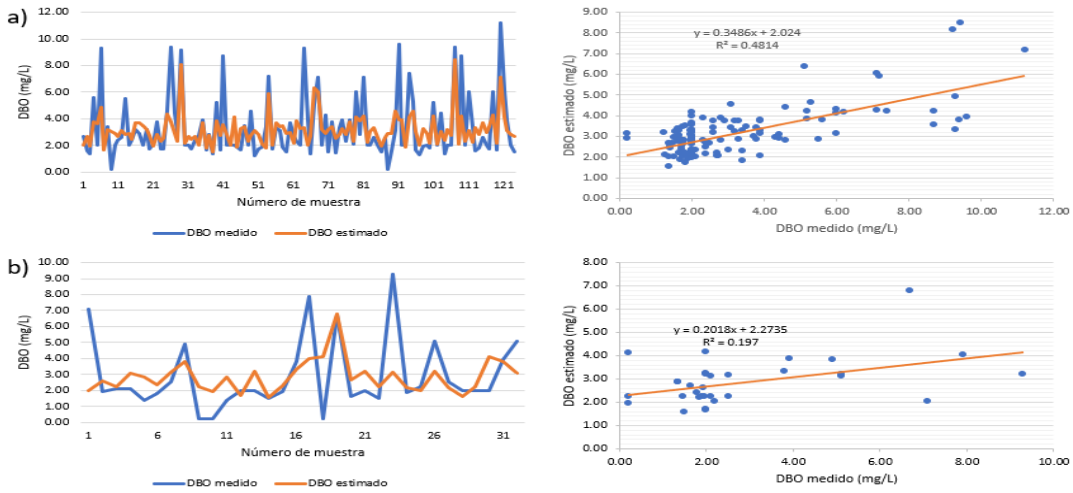

Figure 13. Comparison of graphs of measured and estimated BOD using SVR for a training data and b test data

Source: Own elaboration

6. DISCUSSION

The input variables for the model were divided into two groups because each group is intended for a dependent variable. These variables were pH, specific conductivity (EC), turbidity, nitrites, total organic carbon (TOC), chemical oxygen demand (COD), iron and chlorides, for BOD, with a correlation coefficient greater than +0.29. For DO, the variables with a coefficient greater than ± 0.17 were temperature, nitrites, COD, nitrates, total dissolved solids (TDS), chlorides and total solids (TS). The reason for the selection of these two correlation values is due to the number of parameters that have greater than indicated, which in total are 10 for the BOD and DO, independently. In addition to this, it should be considered that Antanasijević et al. (2013) and Csábrági et al. (2019) indicated that pH and EC are relevant due to their simple, easy and continuous measurement in monitoring stations; furthermore, its change in a negative direction reflects the entry of wastewater and the increase in the decomposition of organic material. Most of the studies analyzed include pH, EC and temperature within the model because they present correlation coefficient values greater than 0.3; This coefficient is used mainly to determine the essential parameters for the model.

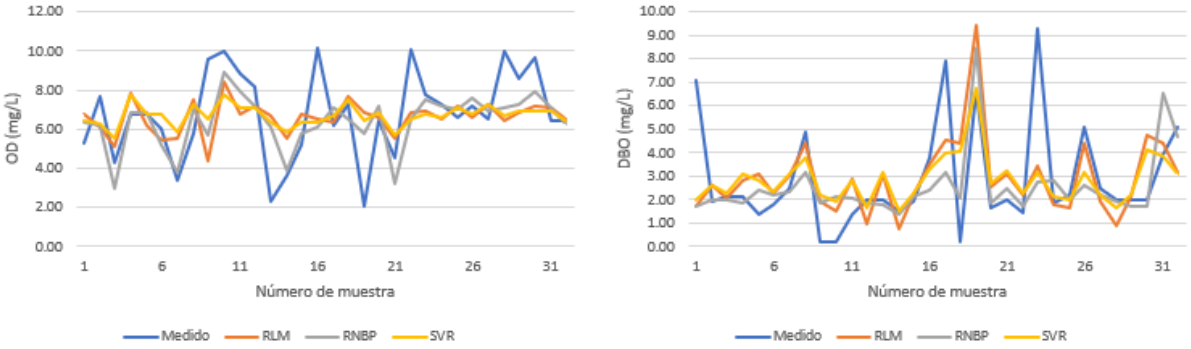

Figure 14. Comparison of models chosen in the estimation of DO and BOD in the test phase.
Source: Own elaboration

Obtaining the results in the previous section leads to the comparison of the models to meet the stated objective. Table 7 shows the values of the statistical indicators of each of the MLR, BPNN and SVR models for estimating DO. Regarding R in the test phase, it can be ordered as MLR <SVR <BPNN, like that obtained by Ji et al. (2017) with the difference that the SVM obtained the highest value. Additionally, other authors concluded that the performance of the MLR was below the BPNN and SVM (Csábrágia et al., 2019; Chen and Liu, 2013; Csábrági et al., 2017; Nemati et al., 2015).

It is observed that the BPNN has the R values of 0.768 and 0.605 for the training and test phase respectively, being higher than the SVR. This is not the case for Ji et al (2017) because they obtained an R of 0.9298 for the SVM showing that it estimates the DO in rivers with low oxygen; contrary to the value obtained from the SVR in 0.465 mg / L in the test phase. However, Chen and Liu (2013) obtained an R of 0.5 for the BPNN close to the ANFIS model that they compared. Furthermore, the MAE and RMSE values of the developed network are closer to 0 for both phases than the other models. Therefore, the BPNN was chosen for the estimation of DO, which afterwards entered its values to determine the water quality. Figure 14 shows the comparison of the values between the mentioned models for the DO.

Table 7.

Output	**Training**				**Test**			
	MAE	**RMSE**	**R**	**NS**	**MAE**	**RMSE**	**R**	**NS**
MLR	0.117	0.147	0.633	0.400	0.151	0.195	0.245	0.038
BPNN	0.082	0.122	0.768	0.583	0.117	0.162	0.605	0.337

Performance comparison for DO estimation.

SVR	0.127	0.161	0.541	0.278	0.144	0.181	0.465	0.165

Own elaboration

On the other hand, for the estimation of the BOD, Table 8 shows the comparison of the statistical indicators between the MLR, BPNN and SVR. Like the DO, the BPNN showed values of R = 0.857 in the training phase, being higher than 0.708 and 0.694 of the MLR and SVR, respectively. For the MAE and RMSE, these two are close to 0, with a difference of 0.5 less than the other two models. The comparison of the values between the models to estimate the BOD is shown in Figure 13. The BPNN demonstrated its effectiveness in predicting the BOD as in the case studies of the Langat River (Raheli et al., 2017) and in the Asa river (Adeniran et al., 2016), with the difference that they implemented different learning algorithms, which were firefly and Levenberg-Marquardt, respectively.

Table 8.

Output	**Training**				**Test**			
	MAE	**RMSE**	**R**	**NS**	**MAE**	**RMSE**	**R**	**NS**
MLR	0.106	0.145	0.708	0.502	0.129	0.184	0.473	0.164
BPNN	0.062	0.106	0.857	0.733	0.111	0.182	0.481	0.183
SVR	0.110	0.153	0.694	0.442	0.122	0.181	0.444	0.197

Performance comparison for BOD estimation

Own elaboration

After the comparison of the models and the choice for DO and BOD, the results obtained from the test phase entered the classification of water quality in Table 3. The determination of the performance of this second classification process was performed using the Confusion Matrix. Said matrix can be seen in figure 13 where the good classification is the one with the greatest amount present and with effective precision.

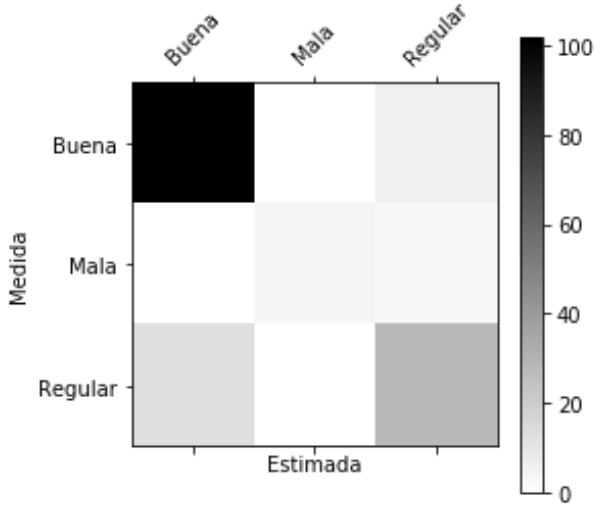

Figure 13. Confusion matrix of water quality in the Rímac river
Source: Own elaboration

The good category had a precision of 0.89 and a sensitivity of 0.94, showing that the model handles the class almost perfectly. On the other hand, the Regular has a precision of 0.76 and 0.68, these values being not close to one and below the previous category, it does not manage to obtain a correct class in the

classification. Finally, the Bad category, having a precision of 1 and sensitivity of 0.57, does not allow the exact detection of the class, but it is exceptionally reliable when it does.

In summary, the average obtained in precision and sensitivity were 0.88 and 0.86, which shows that the model correctly classifies most of the entered set, in addition, it presents an f1-score of 85% considered acceptable.

Table 10. Water quality classification report

	Precision	**Sensibility**	**F1-score**	**Support**
Good	0.89	0.94	0.91	108
Regular	0.76	0.68	0.72	7
Bad	1.00	0.57	0.73	41
Mean	0.88	0.86	0.85	156

Own elaboration

Figure 14 shows us the comparison between the measured and estimated water quality, where the greater presence of good quality is observed in the estimated part, which must be lower due to the values obtained from the information sources, there is also a difference in regular and poor quality; However, the measured quality values are slightly higher than the estimated.

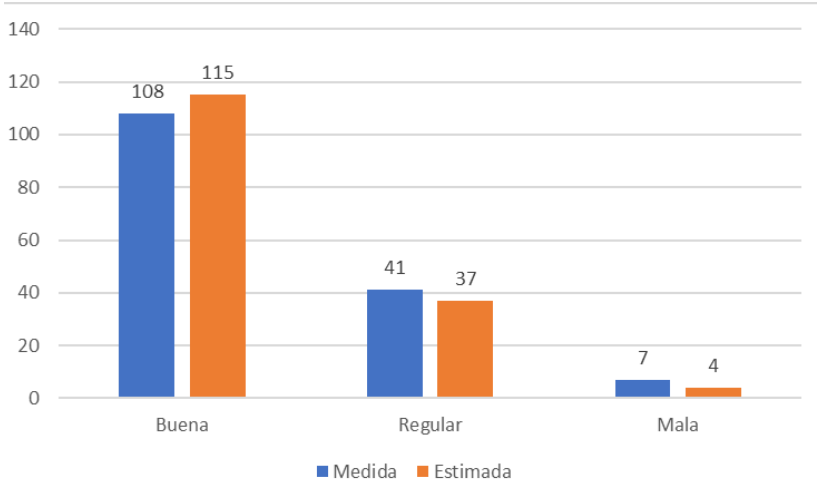

Figure 14. Comparison of measured and estimated water quality
Fuente: Own elaboration

7. CONCLUSIONS

In the present study, three Machine Learning (ML) models were compared to estimate the biochemical oxygen demand (BOD) and dissolved oxygen (DO), which was subsequently used as a model to determine the water quality in the Rímac river (Peru). These ML models were Multiple Linear Regression (MLR), Neural Network Backpropagation (BPNN) and Support Vector Regression (SVR). Water quality parameters such as pH, specific conductivity (EC), turbidity, nitrites, total organic carbon, chemical oxygen demand (COD), iron and chlorides entered the models to estimate the BOD, while parameters such as temperature, Nitrites, COD, nitrates, total dissolved solids, chlorides, and total solids estimated the DO. The results obtained in the comparison of the three models demonstrated the effectiveness of each one in estimating BOD and DO; However, the performance of the MLR and SVR did not exceed that of the BPNN, since it presented for the estimation of the BOD, with 16 hidden nodes, MAE, RMSE and R values such as 0.062 mg / L, 0.106 mg / L, 0.857 for training and 0.111 mg / L, 0.182 mg / L, 0.481 for the test phase. For the estimation of DO, with 8 hidden nodes, these were .082 mg / L, 0.122 mg / L, 0.768 in training, while in the test phase these were 0.117 mg / L, 0.162 mg / L, 0.605. Finally, the classification of water quality as Good, Fair and Bad obtained a precision of 0.88 with a

sensitivity of 0.86 and an f1-score of 85%, demonstrating its effectiveness when carrying out this process, comparing it with the quality determined with the measured values. In summary, the analysis presented in this study shows that BPNN was superior to SVR and MLR in estimating DO and BOD independently, and subsequently the precision of its results in determining water quality was favorable.

8. FUTURE WORK

The scope for determining the quality of water in rivers can not only be determined with the biochemical oxygen and dissolved oxygen demand values, but other chemical and physical elements can be used just as important as the two mentioned; Furthermore, the use of microbiological parameters provides another way to determine quality, such as larvae, algae, among others. On the other hand, parameter calibration can be applied to obtain other parameters that have a greater relationship with the estimated parameters of the study. Like the use of Machine Learning models contemplated in the research literature that were not used in the application of the general objective of this study.

REFERENCES

1. Acosta, R., Ríos, B., Rieradevall, M & Prat,N. (2009). Propuesta de un protocolo de evaluación de la calidad ecológica de ríos andinos (CERA) y su aplicación a dos cuencas en Ecuador y Perú. Limnetica 28: 35-64
2. Adeniran, K. A., Adelodun, B., & Ogunshina, M. (2016). Artificial neural network modelling of biochemical oxygen demand and dissolved oxygen of rivers: Case study of Asa River. *Environmental Research, Engineering and Management*, 72(3), 59-74.
3. Aggarwal, C. C. (2018). Neural networks and deep learning. *Springer*, 10, 978-3.
4. Agrawal, A., Pandey, S. R. & Sharma, B. (2010). Water pollution with special reference to pesticide contamination in India. Journal of Water Resources and Protection, 2, 432-448. https://doi.org/10.4236/jwarp.2010.25050.
5. Ahmed, A. M. (2017). Prediction of dissolved oxygen in Surma River by biochemical oxygen demand and chemical oxygen demand using the artificial neural networks (ANNs). *Journal of King Saud University-Engineering Sciences*, 29(2), 151-158.
6. Ahmed, A. A. M., & Shah, S. M. A. (2017). Application of adaptive neuro-fuzzy inference system (ANFIS) to estimate the biochemical oxygen demand (BOD) of Surma River. *Journal of King Saud University - Engineering Sciences*, 29(3), 237–243
7. Antanasijević, D., Pocajt, V., Povrenović, D., Perić-Grujić, A., & Ristić, M. (2013). Modelling of dissolved oxygen content using artificial neural networks: Danube River, North Serbia, case study. *Environmental Science and Pollution Research International*, 20(12), 9006-9013.
8. Areerachakul, S., Junsawang, P., & Pomsathit, A. (2011). Prediction of dissolved oxygen using artificial neural network. *International Conference on Computer Communication and Management*. 5, 524-528.
9. Autoridad Nacional del Agua. (2018). Estado situacional de los recursos hídricos en las cuencas Chillón, Rímac y Lurín 2016/2017.
10. Autoridad Nacional del Agua. (2018). Metodología para la determinación del índice de calidad de agua ICA-PE, aplicado a los cuerpos de agua continentales superficiales. Lima.
11. Bayram, A., & Kankal, M. (2015). Artificial Neural Network Modeling of Dissolved Oxygen Concentration in a Turkish Watershed. *Polish Journal of Environmental Studies*, 24(4).
12. Chen, W. B., & Liu, W. C. (2014). Artificial neural network modeling of dissolved oxygen in reservoir. *Environmental Monitoring & Assessment*, 186(2), 1203-1217.

13. Csábrági, A., Molnár, S., Tanos, P., & Kovács, J. (2017). Application of artificial neural networks to the forecasting of dissolved oxygen content in the Hungarian section of the river Danube. *Ecological Engineering*, 100, 63–72.

14. Csábrági, A., Molnár, S., Tanos, P., Kovács, J., Molnár, M., Szabó, I., & Hatvani, I. G. (2019). Estimation of dissolved oxygen in riverine ecosystems: Comparison of differently optimized neural networks. *Ecological Engineering*, 138, 298-309.

15. Decreto Supremo N° 004-2017-MINAM, Aprueban Estándares de Calidad Ambiental (ECA) para Agua y establecen Disposiciones Complementarias. (7 de junio de 2017). http://www.minam.gob.pe/wp-content/uploads/2017/06/DS-004-2017-MINAM.pdf

16. Dutta, P. & Chaki, R. (2012). A survey of data mining applications in water quality management. *Proceedings of the CUBE International Information Technology Conference*, 470-475.

17. Iglesias, C., Torres, J. M., Nieto, P. G., Fernández, J. A., Muñiz, C. D., Piñeiro, J. I., & Taboada, J. (2014). Turbidity prediction in a river basin by using artificial neural networks: a case study in northern Spain. *Water resources management*, 28(2), 319-331.

18. Ji, X., Shang, X., Dahlgren, R. A., & Zhang, M. (2017). Prediction of dissolved oxygen concentration in hypoxic river systems using support vector machine: a case study of Wen-Rui Tang River, China. *Environmental Science and Pollution Research*, 24(19), 16062-16076.

19. Mayca Zegarra, G. C. G. (2019). Calidad de agua del río Rimac sector Chicla, provincia de Huarochiri, departamento de Lima. Universidad Nacional Federico Villareal.

20. Nemati, S., Fazelifard, M. H., Terzi, Ö., & Ghorbani, M. A. (2015). Estimation of dissolved oxygen using data-driven techniques in the Tai Po River, Hong Kong. *Environmental earth sciences*, 74(5), 4065-4073.

21. Observatorio del Agua Chillón Rímac Lurín. (2019). Diagnóstico Inicial para el Plan de Gestión de Recursos Hídricos de las cuencas Chillón, Rímac, Lurín y Chilca. Lima.

22. Olyaie, E., Abyaneh, H. Z., & Mehr, A. D. (2017). A comparative analysis among computational intelligence techniques for dissolved oxygen prediction in Delaware River. *Geoscience Frontiers*, 8(3), 517-527.

23. Omar, K. A. (2017). Prediction of dissolved oxygen in Tigris River by water temperature and biological oxygen demand using Artificial Neural Networks (ANNs). *Journal of Duhok University*, 691-700.

24. Ouma, Y. O., Okuku, C. O., & Njau, E. N. (2020). Use of Artificial Neural Networks and Multiple Linear Regression Model for the Prediction of Dissolved Oxygen in Rivers: Case Study of Hydrographic Basin of River Nyando, Kenya. *Complexity*, 2020.

25. Raheli, B., Aalami, M. T., El-Shafie, A., Ghorbani, M. A., & Deo, R. C. (2017). Uncertainty assessment of the multilayer perceptron (MLP) neural network model with implementation of the novel hybrid MLP-FFA method for prediction of biochemical oxygen demand and dissolved oxygen: a case study of Langat River. *Environmental Earth Sciences,* 76(14).

26. Regalado, M. A., Peralta, R. E., & González, R. C. A. (2008). Cómo hacer un modelo matemático. *Temas de Ciencia y Tecnología*, 12(35), 9-18.

27. Sarkar, A., & Pandey, P. (2015). River water quality modelling using artificial neural network technique. *Aquatic Procedia*, 4, 1070-1077.

28. Sierra Ramírez, C. A. (2011). *Calidad del agua: Evaluación y diagnóstico*. Sello Editorial de la Universidad de Medellín. Colombia.

29. Villamarín, C., Rieradevall, M., Paul, J., Barbour, M. T. & Prat, N. (2013). A tool to assess the ecological condition of tropical high Andean streams in Ecuador and Peru: the IMEERA index. Ecological indicators 29: 79-92.

30. Wang, Y., Zhou, J., Chen, K., Wang, Y., & Liu, L. (2017). Water quality prediction method based on LSTM neural network. *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, 1-5.

31. Wen, X., Fang, J., Diao, M., & Zhang, C. (2012). Artificial neural network modeling of dissolved oxygen in the Heihe River, Northwestern China. *Environmental monitoring and assessment*, 185(5), 4361-4371.

32. Xiao, Z., Peng, L., Chen, Y., Liu, H., Wang, J., & Nie, Y. (2017). The Dissolved Oxygen Prediction Method Based on Neural Network. *Complexity*.

33. Zhu, S., & Heddam, S. (2019). Prediction of dissolved oxygen in urban rivers at the Three Gorges Reservoir, China: extreme learning machines (ELM) versus artificial neural network (ANN). *Water Quality Research Journal*.