# Aprendizaje automático para la predicción de calidad de agua potable

## Machine learning for predicting drinking water quality

***Andrea C. Aguilar Aguilar*[1]**
***Francisco F. Obando-Díaz*[2]**

**DOI:** https://doi.org/10.18041/1909-2458/ingeniare.28.6215

**RESUMEN**

La conservación y el cuidado del agua es uno de los problemas medioambientales más importantes en la actualidad. La calidad de agua hace referencia a los valores apropiados de los parámetros fisicoquímicos y/o biológicos del agua para un uso específico. Su monitoreo proporciona información útil a fin de procesarla por herramientas de aprendizaje automático con fines predictivos. Este documento tiene como objetivo presentar una revisión de las técnicas de aprendizaje automático utilizadas en la estimación de la calidad de agua. Los trabajos investigativos muestran que las redes neuronales (RN), los sistemas de inferencia neurodifusa (Anfis) y las máquinas de vectores de soporte (MVS) son las técnicas predictivas más utilizadas. Los resultados obtenidos en las medidas de exactitud evidencian la viabilidad de estimar la calidad de agua en ríos, cuencas y lagos, entre otros.

***Palabras clave:*** Análisis de datos; Aprendizaje automático; Calidad de agua; Predicción; Inteligencia artificial.

**ABSTRACT**

Water conservation and care is one of the most important environmental problems today. Water quality refers to the appropriate values of the physicochemical and / or biological parameters of the water for a specific use and its monitoring provides useful information to be processed by machine learning tools for predictive purposes. This document aims to present a review of machine learning techniques used in estimating water quality. Research works show that neural networks (RN), neuro diffuse inference systems (Anfis), and support vector machines (MVS) are the most widely used predictive techniques, the results obtained in the accuracy measures show the viability of estimate the quality of water in rivers, basins, lakes, among others.

***Keywords:*** Data analysis; Machine learning; Water quality; Prediction; Artificial intelligence.



**Como citar este artículo:** A. Aguilar Aguilar y F. Obando-Dìaz, Aprendizaje automático para la predicción de calidad de agua potable, ingeniare, vol. 2, n.º 28, jun. 2020.

*1 Ingeniera en Automatización Industrial, estudiante de la Maestría en Electrónica y Telecomunicaciones, Universidad del Cauca, Popayán, Colombia, ORCID: https://orcid.org/0000-0001-5695-2973. Correo: acaguilar@unicauca.edu.co*
*2 Magíster en Electrónica y Telecomunicaciones, Docente Universidad del Cauca, Popayán, Colombia, ORCID: https://orcid.org/0000-0001-5666-6969. Correo: fobando@unicauca.edu.co*

*Fecha de recepción: 22 de diciembre de 2019 • Fecha de aceptación: 14 de marzo de 2020*

## 1. INTRODUCCIÓN

El agua potable como recurso natural limitado es una fuente vital para la supervivencia del ser humano y de otras especies. Nuestro planeta está compuesto de, aproximadamente, 70 % de agua, pero de esta solo un 3 % es agua dulce y se encuentra contenida, en su mayoría, en aguas subterráneas y casquetes polares [1]. La problemática ambiental que se ha incrementado en los últimos años hace que sus consecuencias se evidencien en la reducción del acceso y la disponibilidad al agua potable [2].

El término *calidad de agua* se asocia a un conjunto de parámetros físicos, químicos y biológicos cuyas mediciones proporcionan la información sobre el estado en el que se encuentra un cuerpo de agua [3] Garantizar las condiciones apropiadas para el consumo y tener una gestión eficiente de este recurso en cuanto a distribución, aprovechamiento y tratamiento del agua son algunos de los temas que más preocupan a las organizaciones mundiales y a la comunidad científica que trabaja por su conservación y cuidado [4].

Dentro de los procesos de análisis y control del agua, las acciones de monitoreo que se realizan normalmente ya no son estrategias suficientes para garantizar su calidad [5]. La medición de parámetros puede resultar una tarea compleja en la medida en que se requieren diferentes procesos, equipos y personal capacitado para realizar la toma de datos, por lo que contar con información de un sistema representa una enorme ventaja si esta se analiza de forma eficiente.

En este sentido, la predicción de la calidad de agua tiene un gran aporte en el campo medioambiental, así como en los sectores sociales y económicos que dependen de este preciado líquido [6]. La inserción de la tecnología y la inteligencia artificial han permitido desarrollar tanto algoritmos como técnicas de predicción que hacen posible estimar las condiciones de calidad de un cuerpo de agua a partir de datos que han sido recolectados previamente [7]. El presente trabajo propone un modelo híbrido predictivo capaz de utilizar datos y conocimiento para brindar los resultados, enriqueciéndolo, en el caso que así lo requiera, con recomendaciones que faciliten la toma de decisiones. Se utilizaron técnicas de Inteligencia Artificial para representar en un esquema ontológico el conocimiento obtenido al aplicar reglas de asociación. Tradicionalmente, los problemas de predicción se resuelven mediante modelos estadísticos de regresión tales como regresión simple o múltiple, dependiendo del número de variables [8]. Otro modelado que se encuentra con frecuencia son los árboles de regresión para series temporales [9]. Entre las técnicas utilizadas en la inteligencia artificial, las redes neuronales (RN) encabezan la lista de las más difundidas [10].

Este artículo tiene como objetivo presentar una revisión de técnicas utilizadas en la estimación de parámetros y calidad de agua, así como los planteamientos futuros de las tecnologías para este campo. El documento inicia con una breve descripción de la evaluación de la calidad de agua, la predicción y

el aprendizaje automático, luego se expone la revisión de trabajos relacionados con la estimación para calidad de agua, la discusión de los resultados encontrados, las conclusiones y las referencias.

## 2. METODOLOGÍA

La metodología para la selección de trabajos que se consideran en este documento estableció como criterios de inclusión artículos originales de los últimos cuatro años en idioma inglés, de las categorías calidad de agua potable, estimación de parámetros fisicoquímicos y/o biológicos, índices de calidad de agua y técnicas de aprendizaje automático. No se consideraron temáticas de predicción de caudal, consumo y distribución de agua, estimación de índices de contaminación y usos del agua diferentes al de consumo humano. La cadena de búsqueda general fue: “prediction” AND “water quality” OR “water quality index, WQI” AND “machine learning” en las bases de datos bibliográficas ScienceDirect y SpringerLink. El protocolo se desarrolló con base en las recomendaciones para revisiones sistemáticas [11]. Los trabajos encontrados se importaron al gestor de referencias Mendeley, en el cual se eliminaron duplicados y se clasificaron los artículos por tópicos de interés: parámetro estimado, toma de datos y variables de entrada (modelado), correlación de parámetros, técnicas de predicción y estrategias de validación. Este texto de investigación cuenta con más de cincuenta artículos seleccionados producto de la metodología expuesta.

## 3. EVALUACIÓN DE LA CALIDAD DE AGUA

La calidad de agua puede clasificarse de acuerdo con el uso final al que se destine. Es importante destacar que una vez haya cumplido su función, ella retorna nuevamente al sistema hidrológico, por lo que los tratamientos de potabilización son vitales a fin de minimizar riesgos potenciales por contaminación. De manera más práctica, los análisis de calidad de agua se basan en las mediciones de parámetros sobre fuentes hídricas (ríos, lagos, aguas subterráneas, etc.) realizadas por organismos medioambientales de control y se nutren con la información proporcionada por diferentes sectores en una recolección de datos sistémica, la cual puede utilizar los indicadores para su representación [12]. El agua que se destina para el consumo humano debe cumplir con los criterios admisibles reglamentados para cada parámetro. En el caso de los índices, estos se construyen a partir de dos o más parámetros; los índices de calidad (ICA) y contaminación del agua (ICO) son los más comunes [13]. La figura 1 muestra los escenarios de evaluación de la calidad de agua.

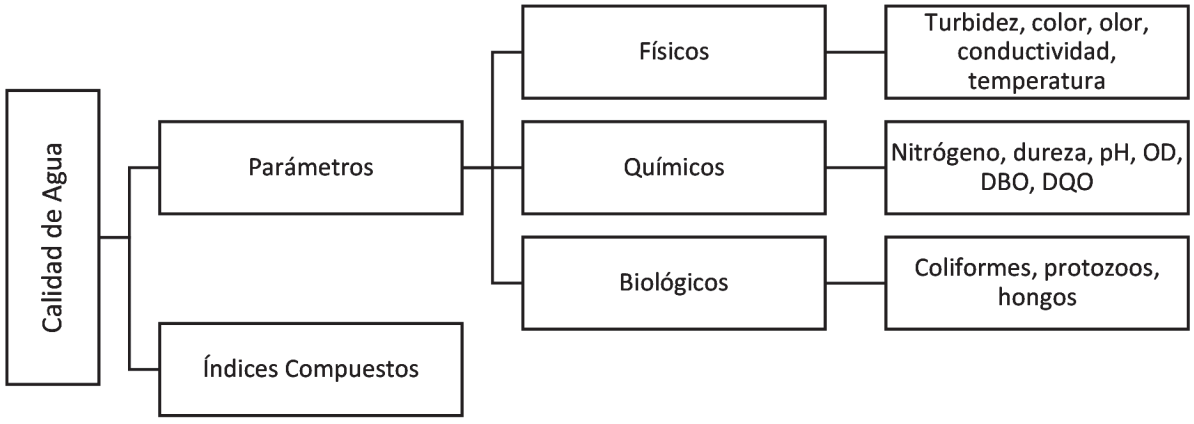


**Figura 1. Evaluación para calidad de agua**

Fuente: elaboración propia.

### 3.1. Estimación de variables y aprendizaje automático

El concepto de *predicción* se enfoca en la extracción de información de datos reales previos de un proceso a fin de predecir patrones de comportamiento o tendencias de posibles eventos futuros. Su aplicación se da en diferentes campos de la ciencia y en fenómenos naturales, no obstante, las tareas de predicción pueden llegar a ser complejas debido al número de variables, el grado de iteración y la dinámica desconocida del fenómeno que se estudia [14].

El procesamiento y el análisis de datos que se efectúa en un aprendizaje automático se lleva a cabo con una alta velocidad y con una mínima intervención humana en la toma de decisiones. Dependiendo de los requerimientos del problema es posible escoger entre distintos métodos y técnicas disponibles, capaces de seguir operando con alto rendimiento, incluso cuando se adicionan más valores durante su ejecución [15]. El aprendizaje supervisado es uno de los más comunes en este campo y se utiliza, generalmente, cuando se conocen los parámetros de la salida deseada; entre las tareas más frecuentes se encuentran la regresión y la clasificación. Los algoritmos no supervisados ajustan su modelo utilizando solo la información de entrada y no están predispuestos operativamente por los valores de salida esperados, lo que permite identificar o agrupar estructuras de un conjunto de datos [16]. En el proceso de aprendizaje, para el caso de los algoritmos supervisados, es posible identificar dos fases en las que es necesario dividir el total de datos en dos conjuntos: pruebas y entrenamiento, o mejor conocidos como *testing and training* [17].

Durante la fase de entrenamiento se construye el modelo utilizando uno de los dos conjuntos de datos a fin de supervisar la variable a estimar. De esta manera, el modelo aprende sobre las posibles causas que influyen en su comportamiento. En la fase de pruebas se verifica la validez del modelo sobre el otro

conjunto, se calcula el error entre las predicciones del modelo y los valores reales. La fase de pruebas también permite evitar el sobreajuste que representa un ajuste muy bueno a los datos para los que se conoce el resultado esperado pero bajo rendimiento en nuevas estimaciones.

Otra estrategia utilizada para evitar el sobreajuste es la validación cruzada, en la cual se divide el conjunto de entrenamiento en *k* subconjuntos; una vez seleccionado un subconjunto *k* como conjunto de prueba, los datos restantes se utilizan como datos de entrenamiento, repitiendo el proceso para *k* iteraciones [18]. La medición de la precisión entre los valores reales y las estimaciones se realiza utilizando las medidas de exactitud, algunas de las cuales son MAPE, MAE, RMSE y [19].

### 3.2. Aprendizaje automático para la estimación de calidad de agua

Hoy en día es posible extraer una gran cantidad de información valiosa sobre los fenómenos que ocurren. En el caso de los ecosistemas hídricos, las investigaciones relacionadas con la estimación de variables utilizando técnicas de aprendizaje automático se han incrementado en los últimos años, lo que ha permitido obtener avances importantes. Estas estrategias también benefician la captura de datos que, en su mayoría, se realizan de forma digital y por métodos manuales, facilitando el estudio de cuerpos de agua en lugares remotos. La tabla 1 resume las características y las técnicas de los trabajos seleccionados de bases de datos bibliográficas como ScienceDirect y SpringerLink.

**Tabla 1. Investigaciones recientes sobre estimación de calidad de agua**

| ID | Proyecto | Técnica | Parámetro estimado | Parámetros de entrada | Validación |
|---|---|---|---|---|---|
| 1. | Predicción del Índice de calidad de agua en la cuenca del río Peak, Malasia [20]. | Red neuronal | ICA | OD, ST, pH, NH3-NL, T°, CE, Turbidez, D S, TS, NO3, Cl, PO4, As, Zn, Ca, Fe, K, Mg, Na, OG, E-Coli, Coliformes, Cd, Cr, Pb | MSE |
| 2. | Estimación del Índice de calidad de agua potable-agua subterránea, Bardaskan [21]. | Red neuronal de inferencia difusa (Anfis) | ICA | DT, CaH, Turbidez, pH, T°, TDS, CE, ALK, Mg, Ca, K, Na, Sulfato, Bicarbonato, Fluoruro, NO3-, NO2-, Cl-. | MAE<br>$R^2$ |
| 3. | Predicción de parámetros de calidad para una represa, Cheongpyeong [22]. | Red neuronal | T°, OD, pH, CE, TN, TP turbidez y clorofila | T°, OD, pH, CE, TN, TP | RMSE<br>$R^2$ |
| 4. | Estimación de parámetros sobre un río en Irán [23]. | Anfis-híbrido | CE, TDS, sodio, dureza carbonatos, dureza total | CE, TDS, SAR, CH, TH, pH, Nap, Cl, Carbonato, Sulfato, Mg y Ca | RMSE<br>MAPE<br>$R^2$ |
| 5. | Predicción de fósforo y nitrógeno en los drenajes de un lago en Manzala [24]. | Anfis | Fosforo<br>Nitrógeno | Caudal, pH SST, CE, TDS, T°, OD, Turbidez | RMSE<br>$R^2$ |
| 6. | Estimación de la calidad de agua en un embalse [25]which is a frequently used metric of water quality in reservoirs. Data collected over ten years (1995-2016. | Red neuronal<br>Máquina de vectores de soporte, árbol de regresión, regresión lineal. | Índice de estado trófico de Carlson | T°, DBO, SS, DQO, NH3, y variables categóricas, temporada y lugar | RMSE<br>MAE<br>MAPE<br>$R^2$ |

| ID | Proyecto | Técnica | Parámetro estimado | Parámetros de entrada | Validación |
|---|---|---|---|---|---|
| 7. | Predicción del índice de calidad del agua en el río Tigris, Bagdad [26]and its values were used as the dependent variable in stepwise multiple linear regression (MLR. | Regresión lineal | ICA | Turbidez, CE DQO, dureza y pH | $R^2$ |
| **8.** | Estimación de parámetros de calidad de agua del rio Tireh [27]. | Red neuronal<br>Máquinas de vectores de soporte<br>Red neuronal-modificado | Ca, Cl, CE, HCO3, Mg, Na, So4, TDS, pH | Ca, Cl, EC, HCO3, Mg, Na, So4, TDS, pH | RMSE<br>$R^2$ |
| **9.** | Estimación del oxígeno disuelto, ríos urbanos, China [28]. | Máquinas de aprendizaje extremo (ELM)<br>Red neuronal perceptrón multicapa | Oxígeno disuelto | Combinaciones de T°, pH, DO, índice de permanganato, NH3–N, CE, DQO, TN y TP | RMSE<br>MAE<br>$R^2$ |
| **10.** | Estimación de Índice de calidad de agua sobre un lago, China [29]. | Máquinas de vectores de soporte- híbrido | ICA | pH, HCO3, TP, TN, DBO, NH3, -N, Fe, Cu, Zn, fenol, DO, TDS, Cl, SO4, Na, Ca, Mg, COD, PO4, Cr | RMSE<br>$R^2$ |
| **11.** | Predicción de la demanda bioquímica de oxigeno, Argelia [30]. | Anfis | DBO | TIN,COD,O2,TDS,PO4<br>Combinación | RMSE<br>MAE<br>$R^2$ |
| **12.** | Estimación de CO2 para un reservorio [31]. | Red neuronal<br>modificado | CO2 | Clorofila, T° carbono orgánico disuelto, TP, CO2. | RMSE<br>MAE<br>$R^2$ |
| **13.** | Estimación de parámetros para evaluar aguas residuales, EE. UU. [32]. | MVS<br>Arboles de regresión | TSS,TDS,DQO, DBO | TSS,TDS,DQO, DBO | RMSE<br>$R^2$ |
| **14.** | Predicción de bioindicadores sobre un río, China [33]. | MVS | Bioindicadores | EC,DO,BOD5,COD NH3-N, TP, Hidromorfologia | MSE<br>$R^2$ |
| **15.** | Estimación microbiana del agua lago Noruega [34]enhancing their applicability in full-scale plants require investigation of their capabilities and limitations in key aspects of the water supply chain. This study comprehensively evaluates the performances of three artificial neural network (ANN. | Redes neuronales<br>MVS | Coliformes | pH, T°, CE, Turbidez, color, alcalinidad, coliformes | MSE |

Fuente: elaboración propia.

En la tabla 1 se presentan los trabajos de estimación de parámetros e índices de calidad de agua sobre diferentes cuerpos de agua, así como las técnicas utilizadas y los parámetros de entrada sobre los cuales se ha realizado la toma de datos, generalmente en sitio. En [20] se expone una estrategia para estimar en tiempo real el índice de calidad de agua sobre el río Peak en Malasia. Como se ha mencionado, una de las ventajas de la estimación es facilitar el acceso y el procesamiento de la información; para esto no se tienen en cuenta los parámetros de DBO (demanda biológica de oxígeno) y DQO (demanda química de oxígeno), ya que para estas no es posible obtener un valor por medición directa. En cuanto

a las técnicas, se utiliza una red neuronal y múltiples redes neuronales, con lo cual se consigue mejorar los resultados de desempeño.

Otro ejemplo se da en [22], en el que se estima la temperatura, el oxígeno disuelto, el pH, la conductividad, la TN, la TP, la turbidez y la clorofila en una represa utilizando una red neuronal. Se obtienen buenos resultados de RMSE y para siete de los ocho parámetros estimados. En [23] se realiza un análisis de correlación con el fin de determinar el mejor conjunto de parámetros de entrada para el modelado. Se compararon los resultados entre las técnicas Anfis y Anfis híbrida, es decir, una combinación con la optimización de enjambre de partículas y colonia de hormigas, evidenciando un mejor desempeño en esta última; ejemplos similares se muestran en [28] y [30].

En cuanto a los índices, formados por dos o más parámetros, en [21] y [26]and its values were used as the dependent variable in stepwise multiple linear regression (MLR se estima el índice de calidad de agua sobre una fuente subterránea y un río a partir de diferentes parámetros de entrada aplicando las técnicas Anfis y regresión lineal, respectivamente. En el control de riesgos por contaminación es posible evaluar los drenajes y los vertimientos a una fuente de agua. La estimación puede aplicarse tanto en índices de calidad como de contaminación o en parámetros específicos, los cuales pueden ser importantes para un estudio o como referencia de control; la estimación del fósforo y el nitrógeno en un lago es un ejemplo de este tipo de análisis [24].

La calidad del agua también se ve afectada por factores externos que en algunos estudios se toman en consideración a fin de poder mejorar los resultados. Estas variables se identifican como categóricas y pueden estar relacionadas con la distribución geográfica, las estaciones del año y hasta información socioeconómica del sector. En [25] se evalúan diferentes técnicas de aprendizaje computacional tales como RN, MVS, árboles de decisión y regresión lineal, a fin de estimar un índice de calidad muy característico en embalses. En el estudio, además, se comparan los desempeños de diferentes *software* de modelado.

De forma similar, en [29] los datos hiperespectrales de teledetección contribuyen a controlar la calidad de los efluentes en la estimación del índice de calidad, y en [31] la estimación del $CO^2$ se da a partir de parámetros y datos categóricos utilizando redes neuronales modificadas.

Estudios comparativos de técnicas de aprendizaje automático también se consideran en el campo medioambiental. En algunas de ellas es posible encontrar cambios de la estructura original, como es el caso de las redes neuronales que se combinan con otras estrategias para potenciar sus resultados. En [27] se realiza un estudio comparativo de RN, MVS y RN híbridas. Otra comparación se da en [32], en el que se estiman parámetros para evaluar la calidad de agua residual de vertimientos en cuencas, al comparar el desempeño de las MVS y los árboles de regresión.

La calidad de agua no es exclusiva para el consumo del ser humano. Los ecosistemas acuáticos también requieren que el agua cumpla ciertas condiciones que garanticen su conservación. Además de los parámetros fisicoquímicos, los bioindicadores pueden proporcionar información valiosa para controlar la calidad de agua dulce. En [33] los indicadores se estiman a partir de parámetros fisicoquímicos e información biológica del cuerpo de agua aplicando la técnica de MVS. De igual manera, se estima la calidad microbiana de un lago comparando dos técnicas de aprendizaje automático [34].

### 3.3. Evaluación del desempeño

Para evaluar la precisión en las estimaciones de un modelo predictivo es posible utilizar las medidas de exactitud, en donde es el valor real, el estimado y el número de muestras de datos, algunas de estas se enlistan y describen a continuación.

- *Media de la desviación porcentual absoluta (MAPE)*. Mide en términos porcentuales el error absoluto, muy efectivo en el momento de identificar diferencias entre modelos; no se afecta por valores estimados o reales (0 % representa un ajuste perfecto). Se calcula a partir de la ecuación 1.

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y-y'}{y}\right| \qquad \text{(Ecuación 1)}$$

- *Error absoluto medio (MAE*). Mide el promedio de las medias absolutas entre los valores reales y los estimados. Es un valor lineal y no es muy sensible frente a valores atípicos; está dado por la ecuación 2.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y'-y|^2}{n} \qquad \text{(Ecuación 2)}$$

- *Error cuadrático medio (MSE*). Mide el error cuadrado promedio entre el valor estimado y el valor real para cada punto, y su resultado no es negativo (ecuación 3); es de más utilidad cuando se trata de grades errores puesto que un valor de MSE alto también puede representar un buen ajuste.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y'-y)^2 \qquad \text{(Ecuación 3)}$$

- *Raíz del error cuadrático medio (RMSE)*. Para dos conjuntos de datos, el RMSE mide el tamaño del error. Es la raíz cuadrada de la suma de errores entre un valor estimado y uno observado o real. Es eficiente al revelar diferencias muy notables y se da en términos de la variable analizada. Esta dada por ecuación 4.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y'-y)^2} \qquad \text{(Ecuación 4)}$$

- *Coeficiente de determinación ()*. Evalúa la calidad del modelo al proporcionar información sobre qué tan bien el modelo se aproxima a los valores observados. Se obtiene de la ecuación 5. El numerador

representa la suma de cuadrados de los residuos y el denominador corresponde a la suma total de cuadrados, y se da entre 0 y 1, donde 1 denota que las estimaciones de regresión se ajustan perfectamente a los datos.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y' - y)^2} \qquad \text{(Ecuación 5)}$$

En la tabla 2 se presenta un resumen de las técnicas, de los parámetros estimados y las medidas de exactitud utilizadas en los trabajos seleccionados.

**Tabla 2. Comparación de características técnicas de predicción**

| ID | Algoritmo de predicción | Parámetros de salida | Exactitud | | | | |
|---|---|---|---|---|---|---|---|
| | | | RMSE | MAPE (%) | MSE | MAE | R |
| 1. | Redes neuronales | ICA-indica de calidad de agua | — | — | 0,9090 | — | 0,9340 |
| | Múltiples redes neuronales | | — | — | 0,1740 | — | 0,1156 |
| 2. | Anfis | ICA | 2,89 | — | — | 0,923 | 0,2808 |
| 3. | Redes neuronales | T° | 0,360 | — | | | 0,998 |
| 4. | Anfis | CE | 4,30 | 7,73 | — | — | 0,91 |
| | Anfis-híbrido | CE | 3,50 | 4,69 | — | — | 0,97 |
| 5. | Anfis | Fósforo | 0,023 | — | — | — | 0,94 |
| | | Nitrógeno | 1,109 | — | — | — | 0,92 |
| 6. | Redes neuronales | Índice de estado trófico de Carlson | 4,644 | 7,721 | — | 3,622 | 0,865 |
| | Máquinas de vectores de soporte | | 5,035 | 8,090 | — | 3,814 | 0,840 |
| | Arboles de regresión | | 5,080 | 8,534 | — | 3,991 | 0,835 |
| | Regresión lineal | | 5,115 | 8,351 | — | 3,936 | 0,835 |
| 7. | Regresión lineal | ICA | — | — | — | — | 0,974 |
| 8. | Redes neuronales | Ca -calcio | 0,295 | — | — | — | 0,84 |
| | Máquinas de vectores de soporte | | 0,193 | — | — | — | 0,94 |
| | Redes neuronales-modificado | | 0,313 | — | — | — | 0,85 |
| 9. | Extreme Machine Learning | OD | 0,518 | — | — | 0,359 | 0,870 |
| | MLPNN multilayer perceptron neural network | | 0,365 | — | — | 0,262 | 0,937 |
| 10. | MVS-híbrido | ICA | 165,91 | — | —— | — | 0,92 |
| 11. | Anfis | DBO | 3,2991 | — | —— | 2,3715 | 0,8906 |
| 12. | Red neuronal modificado | CO2 | 418,48 | — | — | 295,34 | 0,61 |
| 13. | MVS | TSS | 1049 | — | — | — | 0,97 |
| | Arboles de regresión | | 3486 | — | — | — | 0,906 |
| 14. | MVS | Bioindicador | — | — | 87,72 | — | 0,98 |
| 15. | Redes neuronales | Coliformes | — | — | 84,57 | — | — |
| | MVS | | — | — | 140,09 | — | — |

Fuente: elaboración propia.

De acuerdo con la Tabla 2, en la evaluación del desempeño, las medidas de y RMSE son las más utilizadas, seguidas por MAE, MAPE y MSE. Los resultados de estimación de las técnicas de aprendizaje automático se encuentran con valores por encima de 0,61 y alcanzan valores de 0,998 para , lo que muestra que es posible estimar parámetros o índices de calidad de agua con muy buena fiabilidad. Se observa también que los resultados de exactitud mejoran en los estudios comparativos en los que se contrasta la estructura original con una híbrida [19], [27] y [24].

Las técnicas más utilizadas de acuerdo con la información encontrada son las RN, MVS, Anfis, regresión lineal y árboles de regresión. Se evidencia también que, en la mayoría de los casos, los resultados de las técnicas híbridas son superiores a los que se obtienen con la técnica tradicional; algunas alternativas dependerán, entonces, del grado de complejidad que se pueda tener. La regresión lineal permite crear un modelo que describe la relación entre una variable de respuesta basada en una o más variables predictoras. En los árboles de regresión, la salida del modelo se estima con base en el aprendizaje de las reglas de decisión inferidas de las características de los datos. Las máquinas de vectores de soporte construyen un hiperplano a partir de un conjunto de muestras categorizadas, y el algoritmo puede predecir a qué categoría pertenece una nueva muestra. La técnica Anfis integra las redes neuronales y la lógica difusa, su sistema de inferencia responde a reglas difusas y es ideal para sistemas no lineales [35].

Otro elemento destacable es el número de datos utilizados para la estimación. Aunque no se especifica un valor mínimo o máximo para el modelado, es indispensable contar con una buena cantidad de datos, ya que normalmente estos se dividen tanto para la etapa de pruebas como de entrenamiento y validación. Pese a que no todos los artículos muestran una información detallada referente a la cantidad de datos, en la figura 2 se presenta una distribución en años del tamaño de la información utilizada en cada caso.

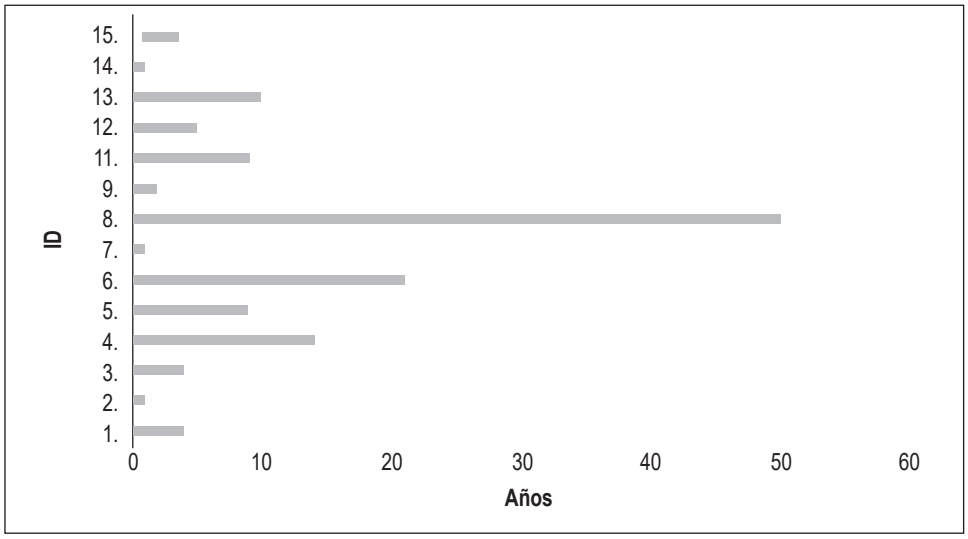


**Figura 2. Datos utilizados en la estimación de calidad de agua. Trabajos seleccionados**

Fuente: elaboración propia.

En la figura 2 se observa que los proyectos trabajan con bases de datos iguales o superiores a un año, once artículos entre uno y diez años y tres estudios con catorce, veintiuno y más de cincuenta años de información recolectada. El estudio ID [25], que no se muestra en el gráfico, presenta una distribución espacial y no temporal, es decir, en un día se tomó una muestra en 48 puntos diferentes a lo largo del río, lo que evidencia dos tipos de distribuciones para los análisis de calidad de agua en ríos [36].

## 4. DISCUSIÓN

El aprendizaje automático se ha convertido en una buena herramienta para los procesos de estimación de calidad de agua, con mejores resultados en comparación con las técnicas estadísticas tradicionales. El uso de la tecnología facilita el tratamiento de los datos y la precisión de los modelos que se construyen. Pese a que se puede encontrar una gran variedad de estrategias, las redes neuronales han abarcado este campo con buenos resultados [37]. Los desafíos se centran ahora en combinar sus propiedades para modelar sistemas con características no lineales y no estacionarias. En el caso de los algoritmos genéticos (GA) y la optimización por enjambre de partículas (PSO), estos se emplean en la selección de subconjuntos y la optimización de parámetros de entrada [38]. De esta manera, los algoritmos de aprendizaje pueden realizar posteriormente las tareas de predicción.

Las MVS muestran mayor eficiencia en su entrenamiento, menor probabilidad de sobreajuste y un mejor comportamiento cuando no hay suficiente información de entrada. Las RN se aplican cuando existe una posible relación entre las entradas y las salidas del sistema, son flexibles y tienen buena respuesta ante patrones no lineales imprevistos [39]. Las técnicas Anfis, una combinación entre redes neuronales y lógica difusa, permiten incorporar conocimiento *a priori* mediante reglas difusas [40]. Estudios comparativos de técnicas pueden aportar información relevante a la hora de escoger la estrategia de modelado [41] to continually provide water to consumers with appropriate quality, quantity and pressure, water utilities require accurate and appropriate short-term water demand (STWD, algunos de estos evalúan el desempeño de las técnicas, como es el caso de las redes neuronales frente a las máquinas de vectores de soporte [42] y otros métodos como, por ejemplo, los bayesianos [43], [44].

La estimación para calidad de agua se da, entonces, tanto para parámetros fisicoquímicos como biológicos e índices compuestos (p. ej., el ICA). La selección de la variable de salida dependerá de los objetivos de estudio, con qué datos se cuenta y el tipo de fuente hídrica que se analiza. Los valores de correlación son útiles para identificar las relaciones entre parámetros y así elegir la mejor combinación de entrada para el modelado. Algunas de ellas, como el TDS, se encuentran fuertemente correlacionadas con la conductividad, así como los sólidos totales con los sólidos suspendidos y disueltos. Otras relaciones importantes se dan entre turbidez, color conductividad y solidos totales, y entre DBO, oxígeno, temperatura y pH [45].

De acuerdo con los trabajos seleccionados, el pH, la conductividad, el oxígeno disuelto y la demanda bioquímica de oxígeno son los parámetros que más se estiman en calidad de agua, lo cual puede indicar que la incidencia sobre otros parámetros es significativa. La alcalinidad, por ejemplo, es un indicador de la capacidad de amortiguación del medio (resistencia a las variaciones en el pH) y es causada por la presencia de iones de bicarbonato, carbonato e hidroxilo, por lo que el pH aumenta más rápido en aguas altamente alcalinas [46]. Las temperaturas bajas favorecen los niveles de oxígeno disuelto en el agua que está relacionado con minerales tales como los carbonatos de calcio y magnesio. Además, la conductividad es sensible a la temperatura y puede apreciarse en los valores de correlación de los parámetros [47].

Factores socioeconómicos y geográficos pueden influir en la calidad de agua, lo que implica considerar estas variables categóricas como información importante para mejorar la exactitud de las predicciones[48]. Si bien los algoritmos de aprendizaje automático son capaces de procesar diferentes variables de entrada y un número considerable de datos en beneficio de la exactitud y la precisión, no siempre es fácil realizar la medición de todos los parámetros, porque se requieren equipos especializados, análisis posteriores en un laboratorio y las fuentes se encuentran en lugares remotos; por tanto, es importante la selección de los parámetros que mejor representen la dinámica del sistema. La implementación de los modelos predictivos a fin de obtener la información en tiempo real también puede representar una ventaja para el estudio de cuerpos de agua en zonas de difícil acceso [49].

Las técnicas de aprendizaje automático se aplican, en su mayoría, sobre ríos, lagos y fuentes de agua en movimiento. El agua almacenada, por su parte, se utiliza en sistemas de distribución, tratamiento y reserva que requieren, además, un control riguroso para su consumo humano. Las herramientas predictivas pueden ser útiles en este caso no solo para la predicción de parámetros, sino también en la estimación de los tiempos en los que se puedan conservar las condiciones mínimas de calidad [50]. Si bien en los sistemas hídricos se generan múltiples reacciones, estas no suelen evidenciarse de forma inmediata al hacer que los tiempos de respuesta sean largos y varíen de un medio a otro, lo que puede guardar relación, entre otros factores, con el volumen del agua [51].

De acuerdo con el análisis de la literatura encontrada, es posible aprovechar las ventajas predictivas en el estudio de la evolución temporal de parámetros fisicoquímicos, a fin de cuantificar el tiempo estimado en el que se pueden conservar las propiedades óptimas de una masa de agua para un uso definido.

## 5. CONCLUSIONES

Las técnicas de aprendizaje automático de mayor aplicabilidad en el recurso hídrico, de acuerdo con la revisión bibliográfica realizada son las RN, MVS y Anfis con porcentajes del 36 %, 24 % y 16 %, respectivamente; el 24% restante corresponde a la implementación de otro tipo de estrategias. Es evidente que el modelado híbrido es una herramienta mejorada que arroja buenos resultados en comparación

con técnicas predictivas tradicionales. Algunas propuestas investigativas podrían girar en torno a la comparación entre métodos híbridos, en la construcción de híbridos con otras metodologías e incluir características del entorno, así como la implementación de los modelos para aplicaciones en tiempo real, lo que contribuye a facilitar las actividades de muestreo. Las técnicas aquí mencionadas tienen diferentes ventajas, así como limitaciones; la selección dependerá, entonces, de las características del problema que se desea abordar y de las estrategias a implementar para mejorar la exactitud de las predicciones.

La estimación de calidad sobre agua almacenada es poco frecuente, sin embargo, estar en capacidad de conocer y cuantificar los tiempos en los que un cuerpo de agua puede permanecer en condiciones deseadas representa una ventaja para los procesos que manejan este tipo de depósitos.

En la investigación se encuentran aspectos comunes que podrían constituir una metodología en el proceso de predicción y que se convierten, además, en tópicos de interés. La medición y el tratamiento de datos, las fases de entrenamiento, pruebas, validación y ajuste son algunas de las etapas que se identifican. Por otra parte, los valores de correlación muestran el grado de relación entre parámetros y ayuda a determinar el conjunto de variables de entrada que brinden mejores resultados. En la validación, es posible utilizar diferentes medidas de exactitud, no obstante, el coeficiente de determinación es un factor válido para determinar el buen ajuste entre los valores reales y estimados.

## REFERENCIAS

[1] D. García, “La crisis del agua, pasado y presente: Memorias de Foro”, *Let. con Cienc. tecnológica*, pp. 64-73, 2018.

[2] A. Nazemi y K. Madani, “Urban water security: emerging discussion and remaining challenges”, *Sustainable Cities and Society*, vol. 41. pp. 925-928, 2018, doi: 10.1016/j.scs.2017.09.011

[3] A. Gómez-Gutiérrez, M. J. Miralles, I. Corbella, S. García, S. Navarro y X. Llebaria, “La calidad sanitaria del agua de consumo”, *Gaceta Sanitaria*, vol. 30. pp. 63-68, 2016, doi: 10.1016/j.gaceta.2016.04.012

[4] WHO, *Safer water, better health*. World Health Organization, 2019.

[5] E. Terneus-Jácome y P. Yánez, “Principios Fundamentales en torno a la calidad del agua, el uso de bioindicadores acuáticos y la restauración ecológica fluvial en Ecuador”, *La Granja Rev. Ciencias la Vida*, vol. 27, n.º 1, pp. 36-50, 2018, doi: 10.17163/lgr.n27.2018.03

[6] B. M. Brentan, E. Luvizotto, M. Herrera, J. Izquierdo y R. Pérez-García, “Hybrid regression model for near real-time urban water demand forecasting”, *J. Comput. Appl. Math.*, vol. 309, pp. 532-541, 2017, doi: 10.1016/j.cam.2016.02.009

[7] R. Cruz Guerrero, M. Alonso Lavernia y A. Franco Árcega, “Hybrid predictive model and recommendations with techniques of data mining and artificial intelligence”, *Program. Matemática y Softw.*, vol. 9, n.º 3, 2017, pp. 18-24.

[8] I. N. Gómez Miranda y G. A. Peñuela Mesa, "Revisión de los métodos estadísticos multivariados usados en el análisis de calidad de aguas," *Mutis*, vol. 6, n.º 1, 2016, pp. 54-63.

[9] E. Núñez, E. W. Steyerberg y J. Núñez, "Regression Modeling Strategies", *Rev. Española Cardiol. (English Ed.)*, vol. 64, n.º 6, pp. 501-507, 2011. doi: 10.1016/j.rec.2011.01.017

[10] I. D. López, A. Figueroa y J. C. Corrales, "Un mapeo sistemático sobre predicción de calidad del agua mediante técnicas de inteligencia computacional", *Rev. Ing. Univ. Medellín*, vol. 15, n.º 28, 2016. pp. 35-52.

[11] G. Urrútia y X. Bonfill, "Declaración Prisma: una propuesta para mejorar la publicación de revisiones sistemáticas y metaanálisis", *Med. Clin. (Barc).*, vol. 135, n.º 11, 2010. pp. 507-511.

[12] S. P. Gorde y M. V Jadhav, "Assessment of water quality parameters : a review", *Int. J. Eng. Res. Appl.*, vol. 3, n.º 6, 2013, pp. 2029-2035.

[13] M. Castro, J. Almeida, J. Ferrer y D. Diaz, "Indicadores de la calidad del agua: evolución y tendencias a nivel global," *Ing. Solidar.*, vol. 9, n.º 17, 2015, doi: 10.16925/in.v9i17.811

[14] Y.-F. Kao, R. Venkatachalam, Y.-F. Kao y R. Venkatachalam, "Human and machine learning," *Comput. Econ.*, 2018. pp. 1-21, doi: 10.1007/s10614-018-9803-z

[15] G. Luo, "A review of automatic selection methods for machine learning algorithms and hyper-parameter values", *Netw. Model. Anal. Heal. Informatics Bioinforma.*, vol. 5, n.º 1, 2016, p. 18. doi: 10.1007/s13721-016-0125-6

[16] H. S. Dhiman, D. Deb y V. E. Balas, "Supervised machine learning models based on support vector regression", en *Supervised Machine Learning in Wind Forecasting and Ramp Event Prediction*, Elsevier, 2020, pp. 41-60.

[17] S. Günnemann, "Machine Learning Meets Databases", *Datenbank Spektrum*, vol. 17, 2017, pp. 77-83. doi: 10.1007/s13222-017-0247-8

[18] D. N. Sotiropoulos y G. A. Tsihrintzis, *Machine learning paradigms. Artificial immune systems and their applications in software personalization*. 2016.

[19] Y. Liu, Y. Liang, S. Liu, D. S. Rosenblum, Y. y Zheng, S., "Predicting urban water quality with ubiquitous data," *arXiv Prepr. arXiv1610.09462*, 2016.

[20] N. O. Al-Musawi y F. M. Al-Rubaie, "Prediction and assessment of water quality index using neural network model and gis case study: Tigris River in Baghdad City," *Appl. Res. J.*, vol. 3, n.º 11, 2017, pp. 343-353.

[21] M. RadFard *et al.*, "Protocol for the estimation of drinking water quality index (DWQI) in water resources: Artificial neural network (ANFIS) and Arc-Gis", *MethodsX*, vol. 6, 2019, pp. 1021-1029. doi: 10.1016/j.mex.2019.04.027

[22] I. won Seo, S. H. Yun y S. Y. Choi, "Forecasting Water Quality Parameters by ANN Model Using Pre-processing Technique at the Downstream of Cheongpyeong Dam", *Procedia Eng.*, vol. 154, 2016, pp. 1110-1115. doi: 10.1016/j.proeng.2016.07.519

[23] A. Azad, H. Karami, S. Farzin, S.-F. Mousavi y O. Kisi, "Modeling river water quality parameters using modified adaptive neuro fuzzy inference system", *Water Sci. Eng.*, vol. 12, n.º 1, 2019, pp. 45-54. doi: 10.1016/j.wse.2018.11.001

[24] G. K. Kang, J. Z. Gao y G. Xie, "Data-driven Water Quality Analysis and Prediction: A Survey," 2017, doi: 10.1109/BigDataService.2017.40

[25] J.-S. Chou, C.-C. Ho y H.-S. Hoang, "Determining quality of water in reservoir using machine learning", *Ecol. Inform.*, vol. 44, 2018, pp. 57-75. doi: 10.1016/j.ecoinf.2018.01.005

[26] S. Hussein Ewaid, S. Ali Abed y S. A. Kadhum, "Predicting the Tigris River water quality within Baghdad, Iraq by using water quality index and regression analysis", *Environ. Technol. Innov.*, vol. 11, 2018, pp. 390-398. doi: 10.1016/j.eti.2018.06.013

[27] A. H. Haghiabi, A. H. Nasrolahi y A. Parsaie, "Water quality prediction using machine learning methods", *Water Qual. Res. J. Canada*, vol. 53, n.º 1, 2018, pp. 3-13. doi: 10.2166/wqrj.2018.025

[28] S. Zhu y S. Heddam, "Prediction of dissolved oxygen in urban rivers at the Three Gorges Reservoir, China: extreme learning machines (ELM) versus artificial neural network (ANN)", *Water Qual. Res. J.*, 2019, doi: 10.2166/wqrj.2019.053

[29] X. Wang, F. Zhang y J. Ding, "Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China", *Sci. Rep.*, vol. 7, n.º 1, 2017. doi: 10.1038/s41598-017-12853-y

[30] B. Khaled, A. Abdellah, D. Noureddine, S. Heddam y A. Sabeha, "Modelling of biochemical oxygen demand from limited water quality variable by anfis using two partition methods," *Water Qual. Res. J. Canada*, vol. 53, n.º 1, 2018, pp. 24–40. doi: 10.2166/wqrj.2017.015.

[31] Z. Chen, X. Ye, and P. Huang, "Estimating Carbon Dioxide (CO2) Emissions from Reservoirs Using Artificial Neural Networks," *Water*, vol. 10, n.º 1, 2018, p. 26. doi: 10.3390/w10010026.

[32] F. Granata, S. Papirio, G. Esposito, R. Gargano y G. de Marinis, "Machine learning algorithms for the forecasting of wastewater quality indicators", *Water (Switzerland)*, vol. 9, n.º 2, 2017, pp. 1-12, doi: 10.3390/w9020105.

[33] J. Fan *et al.*, "Predicting bio-indicators of aquatic ecosystems using the support vector machine model in the Taizi River, China", *Sustain.*, vol. 9, n.º 6, 2017, pp. 1-11, doi: 10.3390/su9060892

[34] H. Mohammed, A. Longva y R. Seidu, "Predictive analysis of microbial water quality using machine-learning algorithms", *Environ. Res. Eng. Manag.*, vol. 74, n.º 1, 2018, pp. 7–20.

[35] P. Harrington, *Machine Learning in Action*. Manning Publications Co., 2012.

[36] I. Duerr *et al.*, "Forecasting urban household water demand with statistical and machine learning methods using large space-time data: a comparative study", *Environ. Model. Softw.*, vol. 102, 2018, pp. 29-38, doi: 10.1016/j.envsoft.2018.01.002

[37] J. M. Hunter *et al.*, "Framework for developing hybrid process-driven, artificial neural network and regression models for salinity prediction in river systems", *Hydrol. Earth Syst. Sci.*, vol. 22, n.º 5, 2018, pp. 2987-3006, doi: 10.5194/hess-22-2987-2018

[38] C. M. Chew, M. K. Aroua y M. A. Hussain, "A practical hybrid modelling approach for the prediction of potential fouling parameters in ultrafiltration membrane water treatment plant", *J. Ind. Eng. Chem.*, vol. 45, pp. 145-155, 2017, doi: 10.1016/j.jiec.2016.09.017

[39] S. Liu, H. Tai, Q. Ding, D. Li, L. Xu y Y. Wei, "A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction", *Math. Comput. Model.*, vol. 58, 2013, pp. 458-465,. doi: 10.1016/j.mcm.2011.11.021

[40] A. Solgi, A. Pourhaghi, R. Bahmani y H. Zarei, "Improving SVR and ANFIS performance using wavelet transform and PCA algorithm for modeling and predicting biochemical oxygen demand (BOD)", *Ecohydrol. Hydrobiol.*, vol. 17, n.º 2, 2017, pp. 164-175, doi: 10.1016/j.ecohyd.2017.02.002

[41] A. Anele *et al.*, "Overview, comparative assessment and recommendations of forecasting models for short-term water demand prediction", *Water*, vol. 9, n.º 11, 2017, p. 887, doi: 10.3390/w9110887

[42] N. Sánchez Anzola, "Máquinas de soporte vectorial y redes neuronales artificiales en la predicción del movimiento USD/COP spot intradiario", *Odeon*, n.º 9, 2016, p. 113, doi: 10.18601/17941113.n9.04

[43] M. Sakizadeh, "Artificial intelligence for the prediction of water quality index in groundwater systems", *Model. Earth Syst. Environ.*, vol. 2, n.º 1, 2016, p. 8. doi: 10.1007/s40808-015-0063-9

[44] G. Carvajal, D. J. Roser, S. A. Sisson, A. Keegan y S. J. Khan, "Bayesian belief network modelling of chlorine disinfection for human pathogenic viruses in municipal wastewater", *Water Res.*, vol. 109, 2017, pp. 144–154, doi: 10.1016/j.watres.2016.11.008.

[45] Z. M. Yaseen *et al.*, "Hybrid adaptive neuro-fuzzy models for water quality index estimation", *Water Resour. Manag.*, vol. 32, n.º 7, 2018, pp. 2227–2245,. doi: 10.1007/s11269-018-1915-7.

[46] A. H. Divya and P. A. Solomon, "Effects of Some Water Quality Parameters Especially Total Coliform and Fecal Coliform in Surface Water of Chalakudy River," *Procedia Technol.*, vol. 24, 2016, pp. 631–638,. doi: 10.1016/j.protcy.2016.05.151.

[47] G. A. H. Sallam and E. A. Elsayed, "Estimating relations between temperature, relative humidity as independed variables and selected water quality parameters in Lake Manzala, Egypt," *Ain Shams Eng. J.*, vol. 9, n.º 1, 2018, pp. 1–14,. doi: 10.1016/j.asej.2015.10.002.

[48] M. Hino, E. Benami, and N. Brooks, "Machine learning for environmental monitoring," *Nat. Sustain.*, vol. 1, n.º 10, 2018, pp. 583-588, doi: 10.1038/s41893-018-0142-9.

[49] D. S. Manu y A. K. Thalla, "Artificial intelligence models for predicting the performance of biological wastewater treatment plant in the removal of Kjeldahl Nitrogen from wastewater", *Appl. Water Sci.*, vol. 7, n.º 7, 2017, pp. 3783-3791, doi: 10.1007/s13201-017-0526-4

[50] X. Xin, K. Li, B. Finlayson y W. Yin, "Evaluation, prediction, and protection of water quality in Danjiangkou Reservoir, China", *Water Sci. Eng.*, vol. 8, n.º 1, 2015, pp. 30-39, doi: 10.1016/j.wse.2014.11.001

[51] D. P. C. Peters, K. M. Havstad, J. Cushing, C. Tweedie, O. Fuentes y N. Villanueva-Rosales, "Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology", *Ecosphere*, vol. 5, n.º 6, 2014, doi: 10.1890/ES13-00359.1