INSTITUTO TECNOLÓGICO DE COSTA RICA

ESCUELA DE QUÍMICA

CARRERA DE INGENIERÍA AMBIENTAL

Proyecto Final de Graduación para optar por el grado de Licenciatura en Ingeniería Ambiental

"Modelación de la calidad del agua superficial con algoritmos de aprendizaje automático y el índice ICA-NSF"

José Andrés Gómez Mejía

CARTAGO, noviembre, 2025

“Modelación de la calidad del agua superficial con algoritmos de aprendizaje automático y el índice ICA-NSF”

Informe presentado a la Escuela de Química del Instituto Tecnológico de Costa Rica como requisito parcial para optar por el título de Ingeniero Ambiental con el grado de licenciatura

Miembros del tribunal

TEC | Tecnológico de Costa Rica
Firmado digitalmente por LAURA HERNANDEZ ALPIZAR (FIRMA)
Fecha: 2025.12.06 16:02:16 -06'00'

M.Sc. Laura Hernández Alpízar
Director(a)

ESMERALDA ISABEL VARGAS MADRIGAL (FIRMA)
Firmado digitalmente por ESMERALDA ISABEL VARGAS MADRIGAL (FIRMA)
Fecha: 2025.12.08 10:33:01 -06'00'

Lic. Esmeralda Vargas Madrigal
Lector 1

TEC | Tecnológico de Costa Rica
Firmado digitalmente por CARLOS ENRIQUE CALLEJA AMADOR (FIRMA)
Fecha: 2025.12.10 07:53:05 -06'00'

M.Sc. Carlos Enrique Calleja Amador
Lector 2

TEC | Tecnológico de Costa Rica
Firmado digitalmente por DAVID ISASI HERNANDEZ PARRA (FIRMA)
Fecha: 2025.12.09 16:54:30 -06'00'

M.Sc. David Hernández Parra
Coordinador(a) COTRAFIG

GUILLERMO DE JESUS CALVO BRENES (FIRMA)
Digitally signed by GUILLERMO DE JESUS CALVO BRENES (FIRMA)
Date: 2025.12.11 14:03:31 -06'00'

Dr. Guillermo de Jesús Calvo Brenes
Director Escuela de Química

TEC | Tecnológico de Costa Rica
Firmado digitalmente por DIANA ALEXANDRA ZAMBRANO PIAMBA (FIRMA)
Fecha: 2025.12.11 09:01:49 -06'00'

M.Sc. Diana Zambrano Piamba
Coordinadora Carrera de Ingeniería Ambiental

DEDICATORIA

Le dedico este trabajo final de graduación a mi familia, que me ha amado y apoyado toda mi vida. Sin ustedes: mamá, Tita, Tata, Ronald y Pablo, no habría llegado tan lejos. Les debo cada uno de mis logros, especialmente a mi mamá. Los amo mucho.

AGRADECIMIENTOS

Le agradezco profundamente a cada uno de los profesores y profesoras que me guiaron durante la carrera. Especialmente gracias a todo el personal de la Escuela de Química y la carrera de Ingeniería Ambiental. Le agradezco también a mis a amistades y compañeros de clase. De ustedes aprendí muchísimo y me siento agradecido por coincidir y compartir tanto con ustedes.

Gracias al Instituto Tecnológico de Costa Rica, a las universidades públicas y al Estado Social de Derecho que me permitieron estudiar una ingeniería de la más alta calidad. Gracias a todas las personas que trabajan por Costa Rica, en las aulas, en las oficinas, en los hospitales, en el campo, en las construcciones, en las fábricas… gracias por construir y mantener este país.

Gracias a la ingeniera Esmeralda Vargas Madrigal y al ingeniero Leonardo Cascante Chavarría de la Dirección de Agua del MINAE por su apoyo en la práctica profesional y en la tesis. Gracias por su disposición y por compartir tanto material conmigo.

Mi más sincero y eterno agradecimiento a mi profesora tutora, Laura Hernández Alpízar, quien me dio la oportunidad de trabajar con ella desde mi segundo año de carrera. Gracias por tantas oportunidades, lecciones de vida y por guiarme en este proceso de la tesis y en todos los demás trabajos que hemos hecho (y en los que nos faltan).

Gracias especialmente a los profesores que han contribuido a mi desarrollo: la profesora Diana Zambrano que desde la coordinación ha sido un apoyo inmenso, a la profesora Laura Quesada de la cual he aprendido un montón y a quien le agradezco toda su confianza en mí, al profesor David Hernández por su acompañamiento en todo este proceso del TFG, a la profesora Mary Luz Barrios por su apoyo desde el CIPA, al profesor Eric Romero quien me ayudó con muchas dudas estadísticas y al profesor Carlos Calleja quien nutrió este trabajo.

Gracias a mis amistades de toda la vida, especialmente a Leo, José Pablo, Mariano y Saúl. Gracias por estar ahí a pesar del tiempo y la distancia, gracias por escucharme hablar de este y otros trabajos.

Y el más grande agradecimiento a mi familia: mi mamá Saray, mi abuela María Esther, mi abuelo Manuel, mis tíos Ronald y Pablo. Especialmente gracias también a mi novia Valentina, quien me ha acompañado tanto tiempo. Literalmente a tu lado escribí gran parte de esta tesis y de otros artículos. Gracias por tu amor, por tu apoyo, por tu paciencia y por enseñarme tantas cosas. Te amo.

TABLA DE CONTENIDO

LISTA DE CUADROS

LISTA DE FIGURAS

LISTA DE SIGLAS Y ACRÓNIMOS

CAT: categorías simplificadas de calidad

DA: Dirección de Agua

DBO: Demanda biológica de oxígeno (5 días, 25 °C)

DT: *Decision Tree*, árbol de decisión

EDA: *Exploratory data analysis*, análisis exploratorio de datos

FOS: concentración de ion fosfato

ICA: Índice de calidad de agua

ICA-NSF: Índice de calidad de agua de la *National Sanitation Foundation*

ICA-NSF-CR: Índice de calidad de agua de la *National Sanitation Foundation* adaptado para Costa Rica

MINAE: Ministerio de Ambiente y Energía de Costa Rica

ML: *Machine learning*, aprendizaje automático

NIT: concentración de ion nitrato

NSF: *National Sanitation Foundation*

PNMCCAS: Plan Nacional de Monitoreo de la Calidad de los Cuerpos de Agua Superficiales

PSO: Porcentaje de saturación de oxígeno disuelto

RF: *Random Forest*, bosque aleatorio

SHAP: *SHapley Additive exPlanations*

ST: concentración de sólidos totales

XGB: *eXtreme Gradient Boosting*

1 RESUMEN

Resumen

La evaluación integral de la calidad de los cuerpos de agua superficiales puede realizarse utilizando índices de calidad de agua (ICA) diseñados a partir de un conjunto de parámetros fisicoquímicos con diferente peso estadístico en el índice. La distribución de los pesos dentro de un ICA es un elemento crítico, porque determina la influencia de cada parámetro en una clasificación final coherente con los elementos que componen el índice. La Dirección de Agua del Ministerio de Ambiente y Energía de Costa Rica ha empezado a aplicar el ICA-NSF adaptado al país con seis parámetros: saturación de oxígeno disuelto, demanda biológica de oxígeno, pH, ion nitrato, ion fosfato y sólidos totales. La distribución de los pesos fue definida en los Estados Unidos, por lo que se desconoce si es aplicable al territorio costarricense, ya que podría no contemplar la incidencia real de los parámetros. Se puede modelar la calidad del agua con algoritmos de aprendizaje automático (machine learning) y datos reales del Plan Nacional de Monitoreo de la Calidad de los Cuerpos de Agua Superficiales 2021-2024, tomando como base las categorías de calidad de un ICA, y extraer las importancias de los parámetros a lo interno del modelo. Se modeló con tres algoritmos de clasificación: Decision Tree, Random Forest y eXtreme Gradient Boosting. Los modelos fueron evaluados mediante validación cruzada estratificada para la exactitud, precisión, sensibilidad y F1. Luego de dos iteraciones para eliminar la influencia de los pesos originales, se escogió el modelo de Random Forest a partir de su rendimiento e interpretabilidad intrínseca. Este modelo asignó los pesos: 0,36 para DBO, 0,27 para ion fosfato, 0,15 para sólidos totales, 0,11 para porcentaje de saturación de oxígeno disuelto, 0,06 para ion nitrato y pH. Estos pesos objetivos establecen una línea base a partir de los datos del monitoreo nacional, que permitiría una aplicación más confiable del índice en Costa Rica.

Palabras clave

Análisis de datos, ciencia de datos, datos ambientales, monitoreo de ríos, pesos objetivos, pesos de un ICA, Random Forest

Abstract

Comprehensive assessment of surface water bodies can be performed using Water Quality Indices (WQI). The distribution of weights within a WQI is critical, as it determines the influence of each parameter on the final classification. The Water Directorate of the Ministry of Environment and Energy of Costa Rica has begun applying the NSF-WQI, adapted to the country with six parameters: dissolved oxygen saturation, biological oxygen demand, pH, nitrates, phosphates, and total solids. The distribution of weights was defined in the United States; therefore, it is unknown whether it is applicable to Costa Rica, as it may not reflect the actual incidence of the parameters. Water quality can be modeled using machine learning algorithms and real data from the National Plan for Monitoring the Quality of Surface Water Bodies 2021-2024, based on the quality categories of a WQI, and the importance of the parameters can be extracted within the model. Three classification algorithms were applied for modeling: Decision Tree, Random Forest and eXtreme Gradient Boosting. The models were evaluated using stratified cross-validation with the Accuracy, Precision, Recall, and F1 metrics. After two iterations to eliminate the influence of the original weights, the Random Forest model was selected based on its performance and intrinsic interpretability. This model assigned the following weights: 0.36 for BOD, 0.27 for phosphates, 0.15 for total solids, 0.11 for dissolved oxygen saturation, and 0.06 for nitrates and pH. These objective weights establish a baseline grounded on the national monitoring data, which can support a more trustworthy application of the index in Costa Rica.

Keywords

Data analysis, data science, environmental data, river monitoring, objective weights, WQI weights, Random Forest

2 INTRODUCCIÓN

Los cuerpos de agua superficial sirven de fuente para diversas actividades humanas como el abastecimiento de agua potable, el riego agrícola, la ganadería, la industria, la generación hidroeléctrica y proyectos de infraestructura [1]. Además, estos ecosistemas sustentan una amplia diversidad de organismos vivos [2]. Por ello, se debe asegurar que el agua superficial tenga la calidad óptima para los diversos usos que se le pretende dar, es decir, que se encuentre en las condiciones químicas, físicas y biológicas adecuadas [3]. Estas condiciones pueden ser afectadas por diversas fuentes de contaminación, que dependen del desarrollo humano y las actividades productivas de la región [4, 5].

La calidad del agua superficial puede evaluarse mediante parámetros fisicoquímicos que sirvan como indicadores de contaminación; sin embargo, su interpretación aislada no brinda una visión integral del estado del cuerpo de agua [6]. Por ello desde la década de 1960, se han implementado índices de calidad de agua (ICA) que sinteticen múltiples parámetros fisicoquímicos en un único valor de calidad [3]. Un ICA se constituye de cinco elementos [6]: 1) los indicadores, 2) las funciones de subcalidad, que transforman los valores medidos de cada parámetro en una escala adimensional de 0 a100 que representa el puntaje de calidad específico del indicador, 3) una distribución de pesos que asigna importancia a cada indicador, 4) una función de agregación que combina los valores de subcalidad y sus pesos respectivos en un puntaje numérico dentro del mismo indicador, 5) un sistema de clasificación con base en el puntaje obtenido.

La distribución de pesos es un elemento crítico ya que determina en gran medida la clasificación final de la calidad del agua [7]. En algunas metodologías, los pesos se asignan a partir de criterios de personas expertas dada las importancias conceptuales de los indicadores [8], lo cual es susceptible a errores ya que los pesos podrían no ser representativos de la variabilidad estadística que tienen esos parámetros en la región [9, 10]. Como alternativa, se han desarrollado métodos objetivos, como algoritmos de aprendizaje automático, que asignan pesos con base en la importancia estadística de cada indicador en la clasificación a partir de datos reales [7].

La Dirección de Agua del Ministerio de Ambiente y Energía de Costa Rica utilizó el índice de calidad de agua de la *National Sanitation Foundation* (NSF) de los Estados Unidos (ICA-NSF), propuesto en 1973 con pesos establecidos subjetivamente [6]. Este índice está siendo aplicado en Costa Rica con seis parámetros y una distribución de pesos definida a partir del índice original [11,

12]: porcentaje de saturación de oxígeno disuelto (0,23), pH (0,17), demanda biológica de oxígeno (0,17), ion fosfato (0,15), ion nitrato (0,15) y sólidos totales (0,13); sin embargo, esta distribución podría optimizarse para la clasificación de las aguas nacionales. Para lo anterior, se utilizaron algoritmos de aprendizaje automático para determinar una distribución de los pesos de los parámetros ajustada a la base de datos reales.

Este trabajo final de graduación contribuye al cumplimiento de Objetivo de Desarrollo Sostenible **6. Agua limpia y saneamiento**. Específicamente en las metas: **6.3** (indicador 6.3.2) y **6.5** (indicador 6.5.1).

2.1 Objetivos

Objetivo general

Evaluar el uso de algoritmos de aprendizaje automático para obtener un mejor ajuste del índice ICA-NSF como recurso para clasificar la calidad del agua superficial en Costa Rica.

Objetivos específicos

1. Seleccionar, entre los algoritmos de aprendizaje automático más utilizados en el modelaje ambiental, el de mejor rendimiento con base en la métrica de clasificación de sensibilidad, y su interpretabilidad intrínseca.
2. Proponer los pesos relativos de mejor ajuste para los parámetros del índice ICA-NSF obtenidos con el algoritmo de aprendizaje automático de mejor rendimiento e interpretabilidad.

3 MARCO TEÓRICO

3.1 Calidad del agua superficial

La calidad del agua superficial se entiende como el grado de contaminación en relación con su capacidad para sustentar la vida y su idoneidad para usos humanos, como: el consumo, la industria, el comercio, agricultura, ganadería, acuicultura, turismo, fuerza hidráulica, entre otros [3, 13]. Este es un concepto relativo, ya que un mismo cuerpo de agua puede tener la calidad suficiente -es decir, cumplir con estándares fisicoquímicos y biológicos- para ciertos usos, pero incumplirlos para otros. La calidad, entonces, se fundamenta en la evaluación de múltiples parámetros que sirvan como indicadores de contaminación [14], [15].

Por otro lado, medir múltiples parámetros para obtener una visión completa del estado de calidad de un cuerpo de agua puede representar un reto logístico y financiero [6], especialmente en países en desarrollo [16]. Por ello se vuelve vital optimizar la selección de parámetros, tomando en cuenta las fuentes de contaminación específicas, la variabilidad de los parámetros y la información que cada indicador provee [7, 17]. Lo ideal es que los parámetros que se midan sean representativos de la contaminación existente, para evitar medir parámetros de poca pertinencia y disminuir la redundancia [9].

Los países han incluido el concepto de calidad dentro de sus normativas y regulaciones ambientales, como parte de los compromisos estatales para asegurar un ambiente sano y equilibrado. El reto ha sido tener un marco actualizado que permita evaluar la calidad del agua sistemáticamente en una región (ya sea a nivel nacional, en divisiones internas o en cuerpos de agua individuales), además de una escala que permita comparaciones tanto temporales como espaciales [8, 18]. Esto es útil para regular los usos potenciales de cada cuerpo de agua con el fin de evitar amenazas ambientales y a la salud, así como para plantear estrategias que permitan restaurar los cuerpos de agua [14, 19]. En respuesta a esa necesidad, los índices de calidad de agua (ICA) han surgido como herramientas que permiten integrar múltiples parámetros en un único valor y clasificar los cuerpos de agua según una escala de calidad, lo que facilita la evaluación comparativa y el monitoreo [3].

3.1.1 Índices de calidad de agua (ICA)

Los índices de calidad de agua (ICA) son herramientas para evaluar la calidad del agua de forma sistemática en una cierta región [6]. Los ICA convierten valores de mediciones de parámetros fisicoquímicos en un único valor numérico que se asocia a una escala de categorías de calidad [6]. Su utilidad radica en estandarizar la evaluación, lo cual es esencial en monitoreos a gran escala y para el cumplimiento de normativas ambientales [11].

Por otro lado, los ICA tienen dificultades para abarcar la complejidad de los cuerpos de agua superficiales, ya que reducen fenómenos complejos en un único puntaje [8]. Además, distintos ICA pueden asignar calidades contradictorias a un mismo cuerpo de agua [9, 11]. Por ello, se ha propuesto el desarrollo de índices locales, donde se considere la variabilidad estadística de los parámetros y las principales fuentes de contaminación [6, 9].

La mayoría de los ICA presentan cinco componentes básicos [15]:

1. **Conjunto de parámetros:** los indicadores son seleccionados con base en la opinión de personas expertas, la disponibilidad para medirlos (costo y disponibilidad de equipo), el objetivo del índice y las características ambientales de la región [3, 6].
2. **Funciones de subcalidad:** cada parámetro es transformado en una escala sin unidades por medio de una función matemática específica [15]. Esto permite obtener una cuantificación de la subcalidad, o la calidad específica, de cada parámetro [14]. Estas funciones pueden ser determinadas a partir de criterios de personas expertas, quienes definen funciones de subcalidad en función de la medición del parámetro, además de límites nacionales o internacionales [3].
3. **Distribución de pesos:** a cada parámetro se asigna un peso según su importancia en la calidad [3]. La distribución de pesos, aunque no es homogénea, suma en total 1. Los pesos pueden ser asignados por métodos subjetivos, donde personas expertas son consultadas, o por métodos objetivos, donde se utilizan herramientas estadísticas [6]. Se espera que los pesos sean representativos de la contaminación presente en la región y aborden los objetivos del índice [15]. Adaptar una distribución de pesos propia de un país a otro podría ocasionar errores, debido a las diferencias en las fuentes de contaminación [20, 21].
4. **Fórmula de agregación:** una función final considera las combinaciones de las funciones de subcalidad y los pesos de todos los parámetros para obtener el puntaje del índice [15].

Este valor cuantifica la calidad del punto de muestreo en una escala definida, generalmente de 0 a 100 [6].

5. **Escala de clasificación:** se asigna una categoría de calidad a partir del puntaje obtenido de la fórmula de agregación [15], lo que permite interpretar cualitativamente el estado del cuerpo de agua [6].

Cada uno de los componentes del índice aporta su peso en la incertidumbre predictiva del modelo [22]. La asignación de pesos es una fuente significativa de error [6, 15, 23]. En particular, los métodos subjetivos pueden incorporar sesgos [8] por no considerar la distribución estadística de los parámetros en la región [20, 21].

3.2 Construcción de índices con métodos subjetivos

La determinación de pesos por métodos subjetivos implica la consulta a personas expertas y demás partes interesadas para que den su opinión sobre la importancia de los parámetros escogidos para el índice, con base en la importancia teórica y los objetivos del índice [6]. Existen dos formas principalmente utilizadas para obtener una distribución de pesos a partir de los múltiples criterios de las personas entrevistadas para un ICA: método Delphi y el proceso analítico jerárquico (AHP, por sus siglas en inglés) [20, 21].

El método Delphi se basa en la recopilación anónima y estructurada de opiniones de personas expertas [20]. Su aplicación para la construcción de ICAs implica la consulta a las partes interesadas (científicos, representantes de organizaciones, políticos) por medio de cuestionarios, encuestas y discusiones grupales sobre la importancia de cada parámetro, cuyas respuestas se organizan estadísticamente [21]. Una vez establecidas las opiniones, los pesos numéricos se asignan mediante escalas y se normalizan en un rango [20, 24]. Esta técnica permite que los pesos reflejen el conocimiento experto y las opiniones de las partes interesadas en el índice [24]; no obstante, puede transmitir los sesgos de las personas entrevistadas y requiere de tiempo y recursos [25].

Por su parte, el proceso analítico jerárquico es un método también aplicado en la construcción de ICAs. Este se utiliza la toma de decisiones a partir de múltiples variables que se ha aplicado para asignar pesos a parámetros [6]. Se basa en comparaciones por pares (*pairwise comparison*), donde

personas expertas y partes interesadas reportan las variables en orden de importancia [26]. Estas comparaciones se organizan en matrices y se aplican cálculos para obtener los pesos finales [27].

3.2.1 ICA-NSF

En 1965 se desarrolló uno de los primeros índices de calidad de agua gracias al aporte de la *National Sanitation Foundation* (NSF) de los Estados Unidos, por lo que el índice lleva su nombre: ICA-NSF [28]. Se utilizó el método Delphi para seleccionar los parámetros y sus respectivos pesos (CUADRO 3.1) [6, 14].

CUADRO 3.1
PARÁMETROS Y PESOS DEL ICA-NSF

Parámetro	Peso
Saturación de oxígeno disuelto (%)	0,17
Coliformes fecales (NMP / 100 mL)	0,16
pH	0,11
Demanda biológica de oxígeno (5 días, mg / L)	0,11
Ion nitrato (mg / L)	0,10
Ion fosfato (mg / L)	0,10
Temperatura (°C)	0,10
Turbiedad (NTU)	0,08
Sólidos totales (mg / L)	0,07

Asimismo, las personas consultadas definieron funciones que representasen la variación del nivel de calidad del agua con respecto a cada parámetro seleccionado [28]. Posteriormente se combinaron todas las funciones para obtener una función total de subcalidad por parámetro [28]. Con las funciones y los pesos definidos, se estableció una fórmula de agregación aditiva, donde se suman los productos de las subcalidades y los pesos respectivos para cada parámetro en la sumatoria total (ecuación 1) [6].

$$ICA = \sum_{i=1}^{n} s_i w_i \qquad 1$$

Donde *n* es el número de parámetros (en este caso, nueve), *s* es la subcalidad y *w* es el peso.

En 1973, los mismos autores decidieron cambiar la función de agregación de una aritmética a una multiplicativa (ecuación 2) [6]. Esta nueva función es más sensible a valores extremos, por lo que

se evita que alguna variable con una puntuación muy baja pueda reducir significativamente el resultado final [3]; no obstante, el resultado final sigue dependiendo de cada uno de los pesos [6].

$$ICA = \prod_{i=1}^{n} {s_i}^{w_i} \qquad 2$$

El ICA-NSF ha sido ampliamente utilizado como índice general de calidad en múltiples países [3]. En algunos de ellos, los parámetros y los pesos son modificados para adaptar la metodología a las condiciones locales [29]. Esto permite evitar los sesgos procedentes de su construcción subjetiva, ya que los parámetros y sus pesos fueron concebidos originalmente para las condiciones de contaminación del agua de los Estados Unidos [3, 30].

3.2.2 ICA-NSF-CR

El ICA-NSF fue adaptado para Costa Rica por el Instituto Costarricense de Electricidad (ICE) en 2015 [12]. Posteriormente, la Dirección de Agua del Ministerio de Ambiente y Energía (DA-MINAE) lo implementó en el Plan Nacional de Monitoreo de la Calidad de los Cuerpos de Agua Superficiales (PNMCCAS) para el período 2021-2024 [12]. Esta nueva versión, ICA-NSF-CR, mantiene seis de los nueve parámetros originales (**CUADRO 3.2**). Los pesos de los parámetros eliminados fueron redistribuidos equitativamente entre los seis restantes, por lo que se preserva el orden de importancia original [31].

CUADRO 3.2
PARÁMETROS Y PESOS DEL ICA-NSF-CR

Parámetro	Peso
Oxígeno disuelto (% de saturación)	0,23
pH	0,17
Demanda biológica de oxígeno (5 días, mg / L)	0,17
Ion nitrato (mg / L)	0,15
Ion fosfato (mg / L)	0,15
Sólidos totales (mg / L)	0,13

Cada uno de los parámetros aporta información sobre el estado del cuerpo de agua e indica algún tipo de contaminación presente (

CUADRO 3.3) [16, 32].

CUADRO 3.3

RELEVANCIA AMBIENTAL DE LOS INDICADORES DEL ICA-NSF-CR

Parámetro	Indicador
Oxígeno disuelto	Contaminación orgánica (incluida materia fecal), aireación del agua
pH	Vertidos industriales o mineros, tipo de suelo también afecta
Demanda bioquímica de oxígeno	Contaminación orgánica, incluida materia fecal
Ion nitrato	Contaminación por fertilizantes
Ion fosfato	Contaminación por fertilizantes, detergentes y jabones
Sólidos totales	Cambio de uso de suelo, erosión

El índice mantuvo las funciones de subcalidad (**Fig. 3.1**) y de agregación originales [31]. Asimismo, el ICA-NSF-CR presenta cinco categorías de clasificación basadas en el puntaje (

CUADRO 3.4). Las categorías representan la calidad general del cuerpo de agua y van desde "Muy mala" hasta "Excelente".

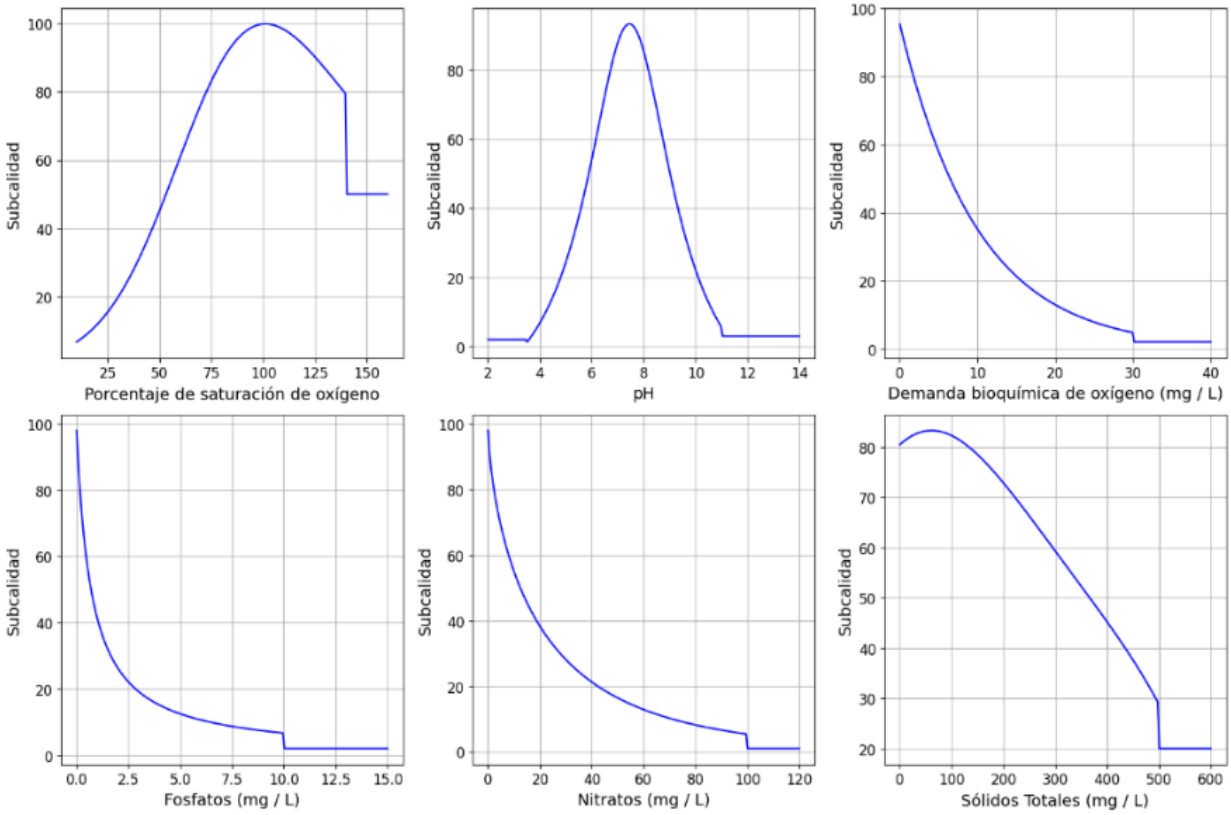

Fig. 3.1. Funciones de subcalidad del ICA-NSF-CR.

CUADRO 3.4

CLASIFICACIÓN DEL AGUA SUPERFICIAL SEGÚN EL ICA-NSF-CR

Puntaje ICA-NSF-CR	Calidad	Color
91 - 100	Excelente	Azul
71 < 90	Buena	Verde
51 < 70	Regular	Amarillo
26 < 50	Mala	Anaranjado
0 < 26	Muy mala	Rojo

Nota: el valor del puntaje se redondea al número entero más cercano para definir la clasificación.

3.3 Construcción de índices con métodos objetivos

Los pesos de los parámetros de un ICA también pueden obtenerse a partir de métodos objetivos, los cuales consideran la importancia empírica de cada parámetro a partir de datos reales de la región estudiada [25]. Estos métodos, al no considerar datos subjetivos, evitan los sesgos de personas que no conocen a profundidad las fuentes de contaminación de la región [3].

Por otro lado, los pesos obtenidos reflejan la realidad observada y no necesariamente los objetivos del índice [27]. Se debe considerar que los pesos objetivos no son, por sí mismos, los más adecuados. Significan que están libres de sesgos, tanto negativos como positivos; por lo que pueden aplicarse directamente o usarse como una base y ajustarlos según los objetivos del índice.

Para determinar pesos objetivamente, se han utilizado técnicas estadísticas que evalúan la relevancia de cada parámetro en la clasificación, como el análisis de componentes principales (PCA, por sus siglas en inglés) y el método de entropía (EWM). El PCA identifica las combinaciones de variables que concentran mayor varianza dentro del conjunto de datos, estos son los componentes principales [17, 30]. A partir de estos componentes, se determinan los parámetros que más contribuyen a la variabilidad y se les asigna pesos en función de su aporte [17, 30]. Cuanta más varianza aporte un parámetro, más representativo se considera y, por lo tanto, más importancia debería tener en el ICA [33].

Un problema del PCA es que, al únicamente basarse en los componentes principales, puede ignorar información necesaria, ya que no se consideran todos los datos [34]. Además, este método asume que la relaciones entre las variables son lineales, por lo que, si existen relaciones no lineales, los componentes principales podrían no representar adecuadamente la variabilidad de los datos [35].

El EWM, por su lado, se basa en el concepto de la entropía de información, que en este contexto se interpreta como una medida de la información asociada a la distribución de los datos de un indicador. Una distribución con alta variabilidad resulta matemáticamente en una menor entropía de información e implica una mayor cantidad de información útil para la toma de decisiones [36]. Así, para un conjunto de muestras, si un parámetro fisicoquímico varía mucho entre muestras diferentes, el EWM le asigna un mayor peso porque proporciona una mayor cantidad de información útil y poder de discriminación [36]. Entonces, para el caso de un ICA, los parámetros con menor entropía recibirían los pesos más altos, al ser las variables que más discriminan en el conjunto de datos [37]. Este método tiene la limitación de asignarle menor entropía a los parámetros con valores más altos y por lo tanto mayor peso, lo que puede subestimar la importancia de aquellos con distribuciones sesgadas hacia valores bajos, además de ser susceptible a valore atípicos [37-39].

Fundamentalmente, los dos métodos se basan en la variabilidad de los parámetros dentro del conjunto de datos. Aunque esto evita los sesgos de los métodos subjetivos, implica que se basan únicamente en la dispersión estadística de los parámetros sin considerar la importancia objetiva en la clasificación. Los algoritmos de aprendizaje automático (ML, por sus siglas en inglés) también se han utilizado como métodos objetivos para determinar pesos [7, 27, 40]. Sin embargo, estos asignan la importancia según el aporte real de cada parámetro en la clasificación, a partir de la estructura del algoritmo [27, 41]. Este método se acerca más a un proceso cognitivo humano para clasificar un cuerpo de agua [41].

3.4 Aprendizaje automático (ML)

El aprendizaje automático o *machine learning* (ML) se refiere a un conjunto de herramientas computacionales que en un proceso de aprendizaje iterativo permiten modelar fenómenos complejos. Los algoritmos se ajustan a sí mismos a través de procesos de aprendizaje continuo, donde los algoritmos se ajustan a sí mismos a través de la retroalimentación continua con datos de entrenamiento y prueba [42].

En la clasificación supervisada, los datos de entrenamiento están etiquetados, es decir, cada observación incluye un conjunto de variables de entrada (por ejemplo, parámetros fisicoquímicos) y su correspondiente categoría de salida, que es la base de la modelación (por ejemplo, una clase de calidad del agua) [10]. El algoritmo, para cada observación, toma las variables de entrada y les

asigna una clase, la cual compara con la verdadera. De esta forma, aprende patrones y relaciones a medida que procesa los datos y se minimiza el error en la clasificación [41].

El aprendizaje automático ha tenido múltiples aplicaciones en las áreas de la ingeniería ambiental [43, 44], especialmente en la gestión del recurso hídrico y de aguas superficiales [45]. Asimismo, se han aplicado algoritmos de ML para determinar objetivamente los pesos de parámetros en ICAs [27, 40]. Entre los algoritmos más utilizados en la modelación de la calidad del agua superficial se encuentran los árboles de decisión o *Decision Tree* (DT) [46-48], bosque aleatorio o *Random Forest* (RF) [10, 40, 49],y *eXtreme Gradient Boosting* (XGBoost o XGB) [7, 40, 47, 49]. Estos han presentado altos rendimientos e interpretabilidad [50], lo que los vuelve útiles para obtener los pesos de los parámetros del ICA-NSF-CR.

3.4.1 Árbol de decisión

Un árbol de decisión es un algoritmo de aprendizaje automático que crea divisiones rectangulares en el espacio vectorial mediante puntos de corte definidos a partir de las variables independientes del modelo (**Fig. 3.2**) [41]. Estas regiones se conocen como nodos, los cuales son sucesivamente ramificados hasta alcanzar un nivel en el que ya no se realizan más subdivisiones. En este punto, se forman los nodos terminales u hojas, que representan las clasificaciones finales del algoritmo.

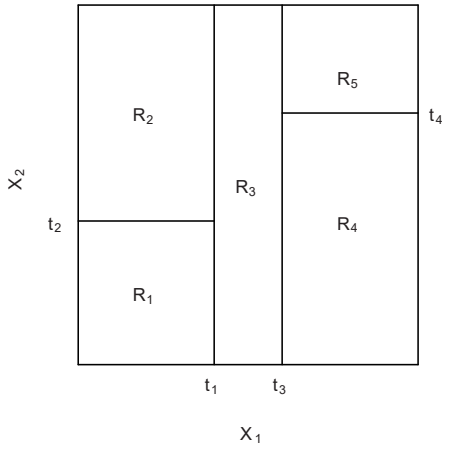

Fig. 3.2. Representación matemática del algoritmo de árbol de decisión. Las regiones R_1, …, R_4 se construyen a partir de los valores t_1, …, t_4 de las variables independientes x_1 y x_2 [41].

Como la segmentación del espacio se basa en un conjunto de reglas en los nodos, el algoritmo se puede representar mediante un diagrama similar a un árbol (**Fig. 3.3)**. En ese diagrama, se empieza con la totalidad de los datos en la copa del árbol y se van dividiendo según las decisiones tomadas en los nodos intermedios, hasta clasificar todos los datos en alguna hoja. Cada hoja del árbol

representa una clase, definida por la clase mayoritaria en esa región. Una misma clase puede estar representada por múltiples hojas.

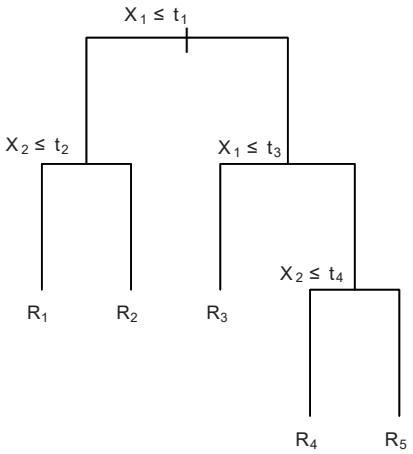

Fig. 3.3. Representación gráfica de un árbol de decisión. Tomado de [41].

El algoritmo puede utilizar ciertos criterios para definir las reglas de partición en los nodos. Uno de esos criterios es el índice Gini (ecuación 3), que se destaca por su buen rendimiento y bajo costo computacional [51]. Este valor mide la pureza de cada nodo y se dice que un nodo es puro cuando únicamente tiene datos de una misma clase [52]. Si se tiene *K* clases, el índice de Gini se define según la ecuación 3 [41] para una clase *k* en un nodo *m*:

$$G_m = \sum_{k=1}^{K} \hat{p}_{m,k}\left(1 - \hat{p}_{m,k}\right) \qquad 3$$

En la ecuación 3 $\hat{p}_{m,k}$ representa la proporción de datos de la clase *k* en el nodo *m*. En cada nodo se prueban todos los *X* parámetros y para cada parámetro *x* se prueban distintos valores de corte *s*. Con base en ello, se escogen los parámetros y valores que más disminuyen el índice de Gini. Cuando el valor de G en un nodo es 0, significa que ese nodo es puro, ya que $\hat{p}_{m,k} = 1$.

En cada nodo se dividen los datos en dos subconjuntos: el nodo izquierdo donde están los datos que cumplen $x \leq s$, y el nodo derecho donde cumplen $x > s$. De tal forma que el índice de Gini de una partición se puede calcular según la ecuación 4 [52].

$$G_{partición,x} = \frac{N_{izq.}}{N} G_{izq.} + \frac{N_{der.}}{N} G_{der.} \qquad 4$$

Donde N es la cantidad total de datos en el nodo original; $N_{izq.}$ y $N_{der.}$, la cantidad de datos en los nodos izquierdo y derecho, respectivamente; y $G_{izq.}$ y $G_{der.}$, los índices de Gini obtenidos en los nodos izquierdo y derecho, respectivamente, a partir de la ecuación 3. El algoritmo hace varias iteraciones sobre el conjunto de datos N y selecciona el parámetro x y su respectivo valor de s que minimizan el $G_{partición}$, es decir, que obtienen nodos más puros. Ese proceso se aplica en todas las divisiones sucesivas hasta que se llega a un punto final definido o cuando todos los nodos son completamente puros.

Los parámetros que permiten una mayor reducción del índice de Gini se consideran los más importantes dentro del modelo, ya que son los que más influyen en el árbol para clasificar correctamente [41]. La importancia I de un parámetro x se puede calcular según la ecuación 5 [53].

$$I_x = \sum_{m=1}^{M_x} \frac{N_m}{N} \Delta G_m \qquad 5$$
$$Donde: \Delta G_m = G_m - G_{partición,m}$$

Donde m es un nodo que pertenece al conjunto de nodos M_x en el que la variable x fue usada como criterio de partición; N_m, el número de datos en el nodo; y ΔG_m, la disminución del índice de Gini en el nodo m. El algoritmo reporta la importancia normalizada en una escala de 0 a 1. Esta forma de obtener la importancia de un parámetro se basa en su capacidad de reducir la impureza del árbol y optimizar la clasificación [41, 52, 53], a diferencia de los otros métodos objetivos que consideran únicamente la varianza del parámetro.

Si bien los árboles de decisión son algoritmos simples, de fácil interpretación y pueden manejar relaciones complejas, tienen alta varianza, lo que significa que ligeros cambios en el conjunto de datos podrían generar desviaciones significativas en los resultados [41, 54]. Esto limita la capacidad de generalización del algoritmo y les resta confianza a las importancias de los parámetros [50].

3.4.2 Random Forest

Se pueden obtener predicciones más robustas si se consideran múltiples árboles de decisión y se promedia el resultado [41]. Esta técnica se conoce como *bagging*, que es aplicada por el algoritmo Random Forest [55]. Este método crea un conjunto de árboles de decisión independientes entre sí y con diferentes subconjuntos de datos [56].

A cada árbol se le asigna un subconjunto de los parámetros para evitar la correlación entre los árboles y prevenir la dependencia excesiva de características específicas [56]. Por lo que RF resulta una buena herramienta para determinar importancias de parámetros, ya que otorga una imagen global y menos sesgada de la influencia de cada parámetro en la clasificación [50, 57].

El algoritmo genera *D* árboles y a cada uno le asigna un subconjunto de parámetros X', tal que X' sea un subconjunto de *X*, que representa la totalidad de parámetros del modelo. Para el resultado final, RF combina todos los árboles en un único clasificador (ecuación 6) [56].

$$P(k|X') = \frac{1}{D}\sum_{d=1}^{D} p_d(k|X'_d) : X' \subset X \qquad 6$$

Donde $P(k|X')$ es la probabilidad final del bosque de clasificar en una categoría k, a partir del subconjunto X' de parámetros. Este resultado considera las probabilidades p_d de cada árbol *d* para reportar la clasificación final, o la clase más probable, después de que cada árbol fue individualmente optimizado con base en el índice de Gini.

Para determinar la importancia de los parámetros, RF suma la reducción del índice de Gini (ecuación 6) para cada parámetro en cada árbol [41]. De esta forma, los parámetros más importantes serán aquellos que logren una mayor reducción del índice de Gini en todo el bosque.

3.4.3 eXtreme Gradient Boosting

Mientras que RF genera múltiples árboles independientes y pondera el resultado final (*bagging*), XGBoost produce inicialmente árboles pequeños y de bajo rendimiento que sucesivamente mejora y amplía (*boosting*), por lo que va analizando los resultados intermedios para tener árboles finales óptimos [41]. El método considera el error de las predicciones, a través de la optimización de la función de pérdida *L* (ecuación 7) [58, 59].

$$\hat{k}(X) = \sum_{d=1}^{D} \lambda f_d(X'_d) \qquad 7$$

$$Donde: f_d(X'_d) = \sum_{n=1}^{N} \left[L\left(k_n, \hat{k}_n\right)\right] + \Omega$$

Donde $\hat{k}$ es la clasificación final otorgada por algoritmo a partir de los X parámetros, λ es el factor de aprendizaje, que indica la tasa de optimización del algoritmo, f_d es la función del árbol d que representa las separaciones en el espacio para clasificar, L es la función de pérdida que mide el error entre la categoría predicha k con la categoría real $\hat{k}$ para el dato n, y Ω es la función de regularización que impide el sobreajuste del modelo.

Por medio de la minimización de la función L, el algoritmo identifica los parámetros y puntos de corte óptimos con los cuales ir construyendo árboles cada vez mejores. Para obtener el resultado final, XGBoost pondera cada árbol, que indica la tasa de optimización del algoritmo [41]. Mientras más alto sea, mayor será la contribución de cada nuevo árbol.

En cuanto a la determinación de importancias, XGBoost utiliza el mismo método que RF [60]; no obstante, su tipo de optimización implica que puede ignorar algunas variables por completo o depender mayoritariamente en unas pocas [61]. El algoritmo tiende a priorizar los parámetros más importantes en los primeros árboles, a diferencia de RF que permite a todos influir en diferentes árboles [41]. Esto lo vuelve poco adecuado para determinar pesos objetivos, a pesar de haber sido utilizado para ello debido a su alto rendimiento [40].

3.4.4 Interpretabilidad y extracción de importancias

Debido a sus características, los algoritmos de ML pueden utilizarse para modelar un fenómeno y extraer información importante sobre él [44]. Para ello, los algoritmos deben ser interpretables, es decir, que permitan al usuario comprender cómo llegan a sus resultados y qué variables tuvieron más influencia [47, 50]. La interpretabilidad de los modelos permite comprender el cómo y por qué de la asignación de pesos a cada parámetro, lo que justifica su incorporación a las normativas [47, 55].

Los algoritmos implementados en este estudio presentan niveles diferenciados de interpretabilidad con base en cómo asignan importancia a las variables. En la literatura se encuentra que el árbol de decisión se considera altamente interpretable, mientras que bosque aleatorio tiene una interpretabilidad moderada y XGBoost, una moderada-baja [50, 62]. Se puede utilizar el método SHAP (SHapley Additive exPlanations) para complementar las importancias internas de cada modelo. Esta técnica evalúa la contribución de cada variable en predicciones individuales, a diferencia de las importancias internas que presentan una influencia general en el modelo [63].

Utilizar SHAP permite mostrar cuáles variables son más importantes y determinar cómo sus magnitudes afectan la probabilidad de que una muestra sea clasificada en cierta categoría [55, 64]. De esta forma, se interpreta qué tanto influye una variable en el modelo y también cómo lo hace.

3.4.5 Validación cruzada y métricas de rendimiento

Los algoritmos de clasificación se evalúan con los datos de prueba luego de ser entrenados. No obstante, los datos de prueba pueden contener sesgos internos, por lo que evaluar con una sola parte de los datos podría dar resultados engañosos [65]. Se puede aplicar el método de validación cruzada para una evaluación más rigurosa, donde se divide aleatoriamente todo el conjunto de datos en múltiples particiones (*folds*). Si se definen k particiones, se utilizan k-1 para entrenar el modelo y la partición restante, para evaluarlo según las métricas; al finalizar se reportan los resultados promedio de las k pruebas [65]. Además, se puede utilizar la validación cruzada estratificada para mantener la proporción original de clases en cada partición.

El rendimiento puede evaluarse con las métricas: exactitud (*accuracy*), precisión (*precision*), sensibilidad (*recall)* y puntaje F1 (*F1 score*) [66]. Para calcular las métricas, se define una de las clases como la clase positiva, que será la de interés para clasificar correctamente, y la otra será la negativa. En modelaciones de calidad de agua, la clase positiva debería ser la que representa el peor estado, ya que es preferible reducir los falsos negativos para impedir la sobreestimación de la calidad.

La exactitud indica el porcentaje total de predicciones correctas, pero puede no ser representativa cuando hay desbalance en la cantidad de datos por categorías [10]. La precisión y sensibilidad miden la cantidad de verdaderos positivos, pero la primera se utiliza para reducir los falsos positivos; mientras que la segunda, para los falsos negativos [54]. El puntaje F1 es la media armónica de precisión y sensibilidad, por lo que ofrece un balance entre ambos tipos de error.

4 METODOLOGÍA

Se utilizó la base de datos del Plan Nacional de Monitoreo de la Calidad de los Cuerpos de Agua Superficiales que contiene datos para diferentes ríos del país que abarcan la mayoría de las cuencas del territorio (**Fig. 4.1**), correspondiente al periodo 2021-2024 [12]. Esta contiene mediciones de parámetros fisicoquímicos con la correspondiente clasificación del ICA-NSF-CR.

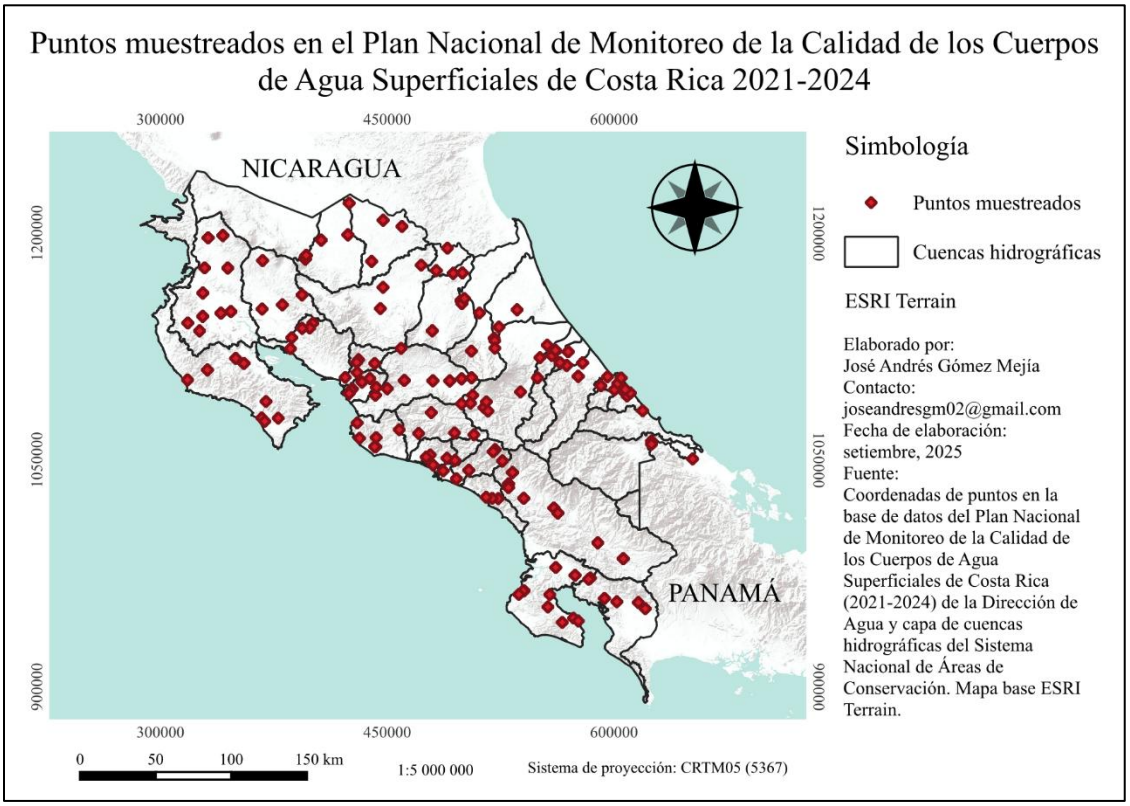

Fig. 4.1. Sitios de muestreo del PNMCCAS 2021-2024.

Los datos se integraron en una hoja de cálculo dinámica en Microsoft Excel que permitió ajustar los pesos de los parámetros iterativamente conforme a los resultados de las modelaciones y observar cambios en las clasificaciones. La base de datos con los pesos originales se exportó en formato CSV para su tratamiento en Python (versión 3.11.11),

Se realizó un preprocesamiento para eliminar valores erróneos y filas vacías, para un total de 12 eliminaciones. La distribución de clases original contiene tres categorías -verde, amarillo y naranja-. Sin embargo, para la clasificación supervisada se construyó una clasificación simplificada binaria para la clasificación supervisada. Se asignó la clase "1" a los registros de mayor calidad -azul o verde- y la clase "0", a los de menor calidad -amarillo, naranja o rojo-. Esta metodología es similar a un enfoque previo aplicado en Irlanda, donde se utilizaron modelos de

clasificación supervisada binaria para obtener los pesos internos de los parámetros [40]. En ese caso, se utilizaron otros criterios para etiquetar los datos, ya que se estaba construyendo un ICA desde cero. Por lo tanto, esa parte específica del etiquetado no aplica en el presente trabajo, donde se parte de un índice existente que se desea optimizar a partir de una escala definida.

Posteriormente, se realizó un análisis exploratorio de datos (EDA), que incluyó: estadísticas descriptivas generales, evaluación de distribuciones, pruebas de normalidad de Shapiro-Wilk, pruebas de Mann-Whitney para comparar medianas entre las clases, análisis de correlaciones de Spearman y análisis de relaciones por pares. Las pruebas estadísticas se realizaron con el paquete "stats" de la librería "scipy" (versión 1.16.0). El EDA permitió caracterizar las variables y sus relaciones, así como obtener insumos para interpretar las modelaciones [67, 68].

Se aplicaron los algoritmos *Decision Tree Classifier* (DT) y *Random Forest Classifier* (RF) del paquete "scikit-learn" (versión 1.6.1), y *eXtreme Gradient Boosting Classifier* (XGB) del paquete "xgboost" (versión 2.1.4). Se utilizaron los hiperparámetros por defecto de los paquetes.

Los algoritmos fueron evaluados mediante validación cruzada estratificada con 10x10 particiones y se aplicaron las métricas de rendimiento: exactitud, precisión, sensibilidad y F1, según las ecuaciones 8 a 11. Es decir, se realizaron 10 iteraciones de la validación cruzada, cambiando la semilla aleatoria, de tal forma que se obtuvieron 100 resultados por métrica para cada modelo. Se consideró a la categoría "0" como la clase positiva, ya que indica la presencia de contaminación significativa.

Exactitud

$$Exac. = \frac{VP + VN}{VP + VN + FP + FN} \tag{8}$$

Precisión

$$Prec. = \frac{VP}{VP + FP} \tag{9}$$

Sensibilidad

$$Sens. = \frac{VP}{VP + FN} \tag{10}$$

F1

$$F1 = 2 \times \frac{Precisión \times Sensibilidad}{Precisión + Sensibilidad} \quad (11)$$

Donde:

- Verdaderos Positivos (VP): número de muestras clasificadas correctamente como clase positiva.
- Verdaderos Negativos (VN): número de muestras clasificadas correctamente como clase negativa.
- Falsos Positivos (FP): número de muestras clasificadas incorrectamente como clase positiva.
- Falsos Negativos (FN): número de muestras clasificadas incorrectamente como clase negativa.

Posteriormente, se realizó una prueba no paramétrica de Friedman para muestras no independientes con el fin de evaluar si existía diferencia en la mediana de los resultados de sensibilidad entre los tres modelos. Se aplicó esa prueba porque los resultados de sensibilidad tenían distribución no normal. Se decidió evaluar esta métrica en específico porque representa la proporción de datos de la clase positiva clasificados correctamente y para que los pesos del índice no tiendan a sobreestimar la calidad. Una baja sensibilidad indicaría una capacidad deficiente para clasificar datos de baja calidad [69].

Cuando existió diferencia significativa entre las tres medianas, se aplicó una prueba post hoc de rangos con signo de Wilcoxon para las comparaciones por pares [70, 71]. La prueba se utilizó para determinar entre cuáles modelos existe la diferencia significativa en el resultado de sensibilidad. Cuando no se encontró un modelo significativamente mejor, se consideró la interpretabilidad y confiabilidad de los pesos según el siguiente orden de prioridad: 1. RF, 2. XGBoost y 3. Árbol de decisión. Por último, los modelos se entrenaron con la totalidad de los datos para aprovechar toda la información y los pesos internos de los parámetros se extrajeron con la función "feature_importance" de cada algoritmo.

Los pesos de la primera modelación están sesgados por los pesos asignados en el ICA-NSF-CR, ya que las categorías de calidad se construyeron con base en el puntaje numérico, por lo que se planteó un proceso iterativo para corregir esa influencia. Los pesos internos del modelo

seleccionado se integraron en la hoja de cálculo dinámica para obtener una distribución de clases actualizada. El nuevo conjunto de datos se utilizó para entrenar al algoritmo seleccionado y obtener otra nueva distribución de pesos. El proceso iterativo se mantuvo hasta converger en una distribución de clases de calidad que no sea significativamente distinta que la anterior, según la prueba estadística de proporciones.

El flujo de procesamiento aquí utilizado se presenta el flujo de procesamiento (**Fig. 4.2**).

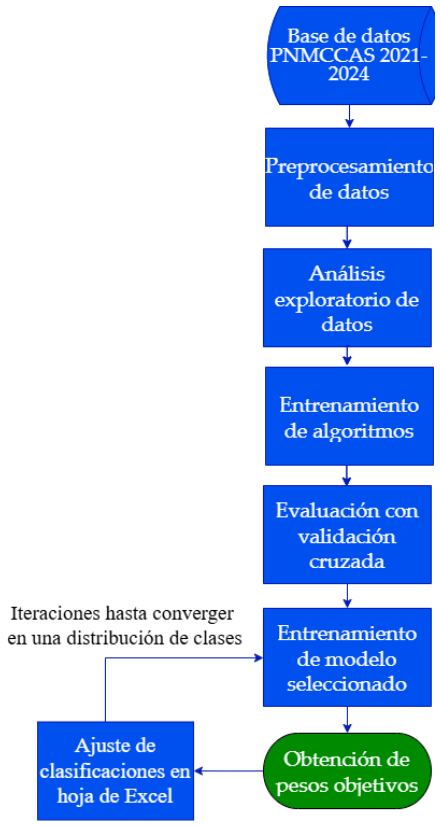

Fig. 4.2. Diagrama de flujo de la metodología.

Finalmente, se analizaron los valores SHAP del modelo seleccionado como exitoso con las funciones "TreeExplainer" y "shap_values" del paquete "shap" (versión 0.48.0). Para ello, se definió un conjunto de prueba estratificado correspondiente al 25% de los datos, con la función "train_test_split". Los modelos se probaron con ese conjunto y se analizaron cada una de las predicciones hechas, extrayendo de ellas los valores SHAP para la clase 0. Estos valores se utilizaron para comprender a mayor profundidad la influencia de cada variable predictiva.

5 RESULTADOS Y DISCUSIÓN

5.1 Análisis exploratorio de datos (EDA)

La base de datos del PNMCCAS contiene originalmente 337 datos en la clase “verde”, 179 en la “amarilla” y 23 en la “naranja” (**Fig. 5.1**). No se tienen datos en las categorías extremas de calidad: “azul” (excelente) y “rojo” (muy mala).

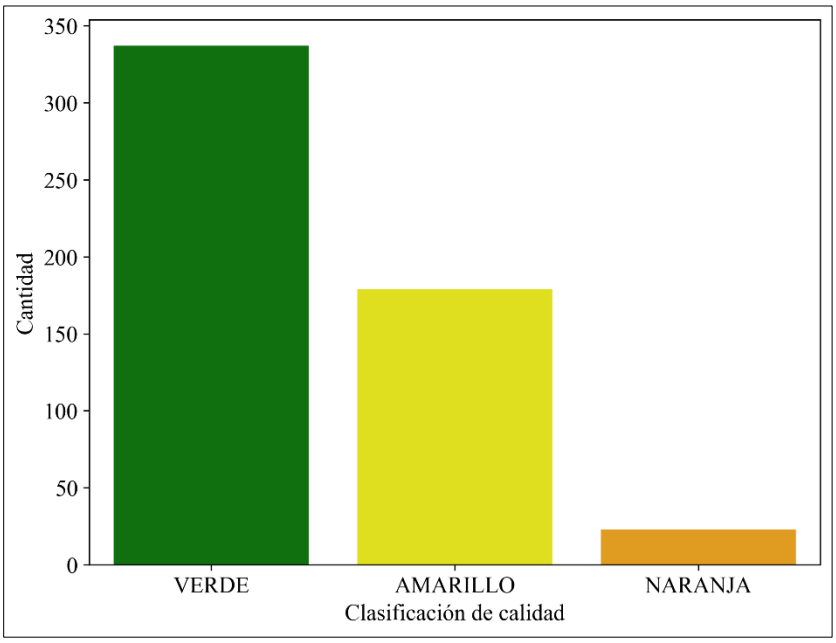

Fig. 5.1. Distribución de las clases de calidad en la base de datos PNMCCAS 2021-2024.

En la construcción de la categorización simplificada, se obtiene una distribución más homogénea con 337 puntos en la categoría 1 (los puntos de la clase “verde”) y 202 en la categoría 0 (los puntos de las clases “amarilla” y “naranja”) (**Fig. 5.2**).

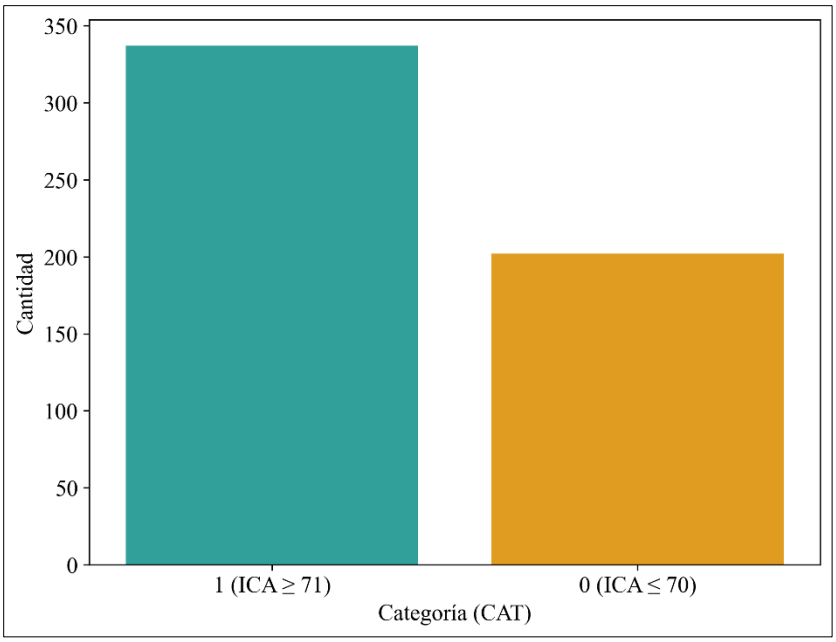

Fig. 5.2. Distribución simplificada de categorías en la base de datos PNMCCAS 2021-2024.

Se presenta la descripción estadística de las variables independientes del ICA-NSF (**; es decir**, una mayor presencia de valores extremos altos y bajos.

CUADRO 5.1). El valor del coeficiente de variación es de especial interés, ya que indica que los parámetros ion fosfato (FOS), demanda biológica de oxígeno (DBO) e ion nitrato (NIT) tienen los datos con mayor dispersión; es decir, una mayor presencia de valores extremos altos y bajos.

CUADRO 5.1

ESTADÍSTICAS DESCRIPTIVAS DE LA BASE DE DATOS DEL PNMCCAS 2021-2024

Variable	Porcentaje de saturación de oxígeno	pH	Demanda bioquímica de oxígeno	Concentración de ion fosfato	Concentración de ion nitrato	Concentración de sólidos totales	Puntaje del índice
Abreviatura	**PSO**	**PH**	**DBO**	**FOS**	**NIT**	**ST**	**ICA**
Unidades	%	-	mg / L	mg / L	mg / L	mg / L	mg / L
Cantidad	539	539	539	539	539	539	539
Media	90,76	7,76	5,48	0,88	1,71	218,6	71,6
Desviación estándar	17,47	0,67	5,76	1,23	2,18	205,1	10,4
Coeficiente de variación	0,19	0,09	1,05	1,39	1,27	0,94	0,14
Mínimo	10	3,97	0,01	0,06	0,05	3	26,6
25%	84,15	7,4	1,9945	0,3	0,53	115,2	65,7
50%	95	7,86	3,5	0,6	1,19	185,7	74,6
75%	101	8,19	7	0,72	1,5	275,4	79,8

Máximo	145,4	9,4	57,6	18,1	15,41	3014	86,6

A continuación, se visualizan las gráficas de distribución para las variables analizadas con los resultados de las pruebas Shapiro-Wilk (**Fig. 5.3**). Se aplica la prueba a cada distribución para determinar si existe normalidad ($\alpha = 0{,}05$), con base en las siguientes hipótesis: H_0: La distribución es normal, o H_1: La distribución no es normal. Como todos los valores p son menores que 0,05, se rechaza la hipótesis nula para cada prueba. Se concluye con un 95% de confianza que ninguna de las distribuciones presenta comportamiento normal. Esto indica el uso pruebas y métodos no paramétricos, así como algoritmos no lineales para modelar el fenómeno.

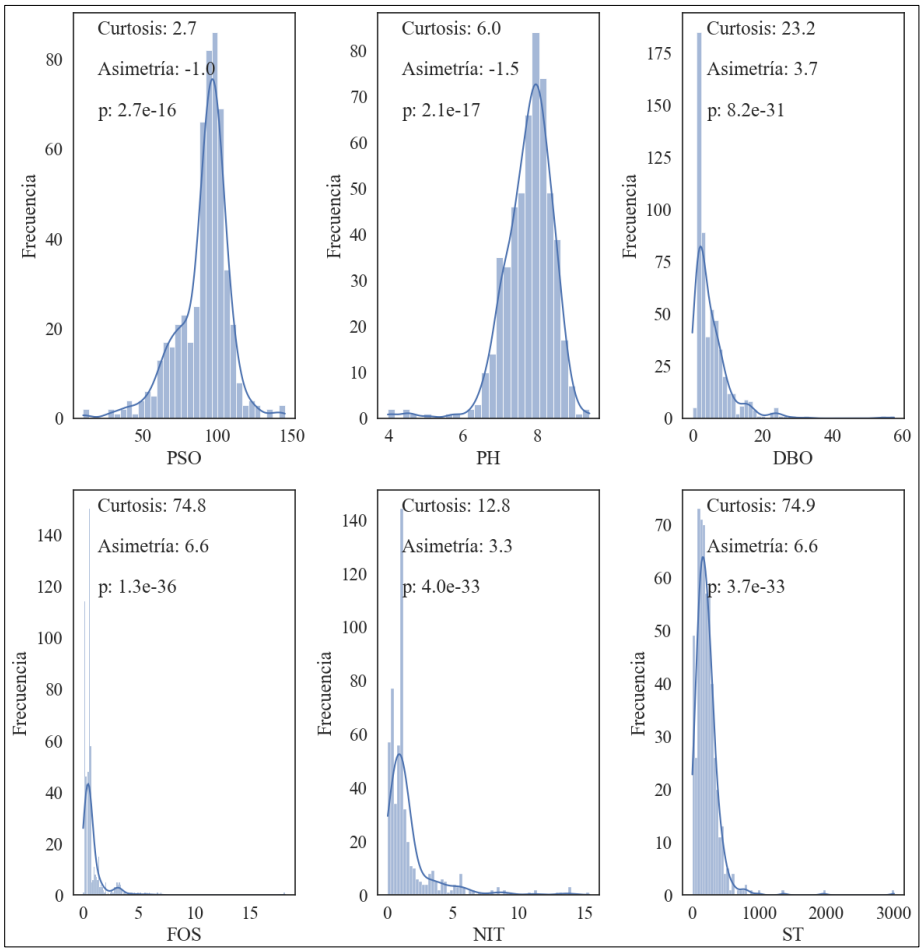

Fig. 5.3. Distribución estadística y resultados de la prueba de Shapiro-Wilk para los parámetros analizados.

Para determinar si las dos clases de calidad son diferentes entre sí, se aplica la prueba Mann-Whitney a una cola para los parámetros en los que valores más altos están asociados con una menor calidad (DBO, FOS, NIT y ST), bajo las siguientes hipótesis (**CUADRO 5.2**).: H_0: la mediana del parámetro en la clase 0 no es significativamente mayor que la mediana en la clase 1 ($\tilde{x}_0 = \tilde{x}_1$); H_1:

la mediana del parámetro en la clase 0 es significativamente mayor que la mediana en la clase 1 ($\tilde{x}_0 > \tilde{x}_1$). Los parámetros PSO y el pH, cuyos valores extremos tanto altos como bajos pueden estar asociados a mala calidad (**Fig. 3.1**), se evaluaron con la prueba a dos colas, bajo las hipótesis: H_0: $\tilde{x}_0 = \tilde{x}_1$; H_1: $\tilde{x}_0 \neq \tilde{x}_1$.

CUADRO 5.2

RESULTADOS DE LA PRUEBA DE MANN-WHITNEY PARA LA COMPARACIÓN DE MEDIANAS ENTRE LAS CATEGORÍAS SIMPLIFICADAS DE CALIDAD PARA TODOS LOS PARÁMETROS DEL ICA-NSF-CR

Parámetro	Mediana		*p* (α = 0,05)	
	Categoría 0	Categoría 1	Una cola (derecha)	Dos colas
PSO	90,10	96,30	-	**5,11E-10**
pH	7,80	7,86	-	0,22
DBO	7,10	2,52	**4,28E-37**	-
FOS	0,97	0,60	**8,28E-24**	-
NIT	1,19	0,92	**2,52E-12**	-
ST	255,70	154,36	**1,77E-17**	-

Nota: en negrita los valores de p menores al α = 0,05.

Para los parámetros DBO, FOS, NIT y ST, se obtiene valores de p significativamente bajos en la prueba a una cola, lo que indica que las medianas de la clase 0 son estadísticamente mayores que las de la clase 1, para un 95% de confianza. Por otro lado, el parámetro PSO también mostró una diferencia significativa entre clases en la prueba a dos colas ($p = 5{,}11\times10^{-10}$), lo que indica una diferencia significativa en las medianas al 95% de confianza. Por otro lado, el pH no presenta una diferencia estadísticamente significativa (p = 0,22) al 95% de confianza, lo que sugiere que este parámetro no permitiría distinguir de forma clara entre las dos categorías. En general, los resultados indican que la clasificación simplificada propuesta es válida para modelar, ya que existen diferencias estadísticas claras entre ambas categorías que los algoritmos podrían detectar.

Para analizar las correlaciones internas, se visualiza una matriz de correlación de Spearman (**Fig. 5.4**). Para las modelaciones es más relevante analizar la correlación entre los parámetros y las categorías de calidad. Por ello, se realizan pruebas de correlación de Spearman para determinar cuáles variables tienen una correlación significativa con las categorías de calidad (**CUADRO 5.3**),

bajo las siguientes hipótesis: H_0: no existe correlación monotónica significativa ($\rho = 0$); H_1: existe correlación monotónica significativa ($\rho \neq 0$).

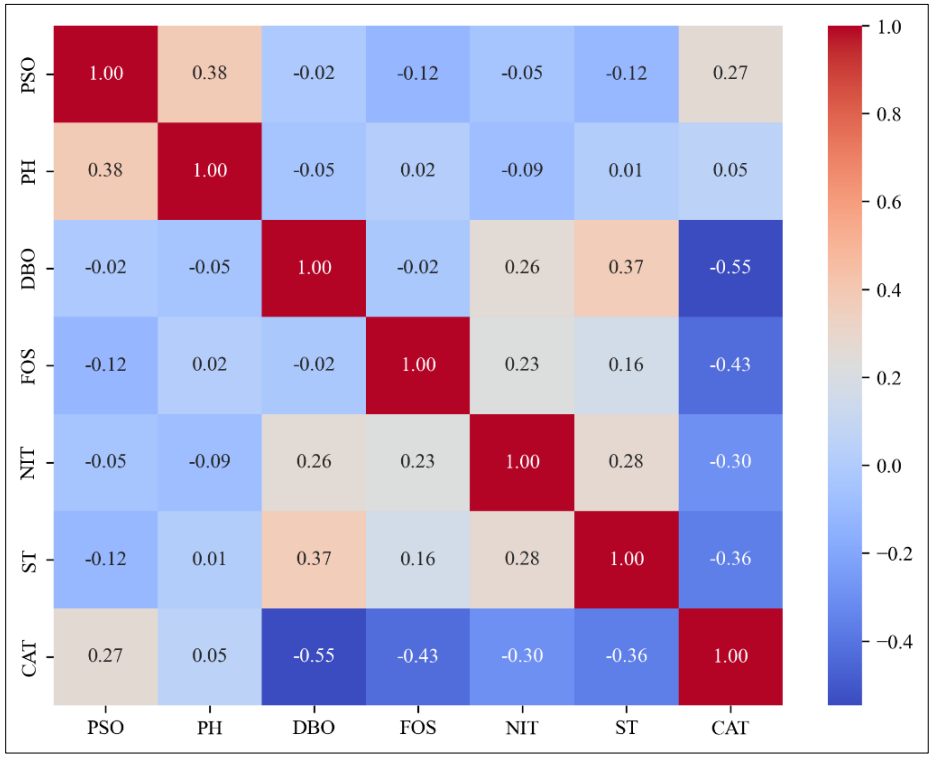

Fig. 5.4. Matriz de correlación de Spearman entre parámetros fisicoquímicos del ICA-NSF-CR y categorías simplificadas de calidad.

CUADRO 5.3
RESULTADOS DE PRUEBAS DE CORRELACIÓN DE SPEARMAN ENTRE PARÁMETROS Y CATEGORÍAS SIMPLIFICADAS DE CALIDAD

Parámetro	Correlación (ρ)	*p* (α = 0,05)
PSO	0,27	**2,56E-10**
pH	0,05	0,225
DBO	-0,55	**2,93E-43**
FOS	-0,43	**9,17E-26**
NIT	-0,30	**1,72E-12**
ST	-0,36	**2,90-18**

Todos los parámetros, excepto el pH, muestran una correlación de Spearman significativa al 95% de confianza. Esto es previsible, ya que las categorías de calidad se construyen a partir de los

variables del índice, por lo que están, en mayor o menor medida, correlacionadas. El pH no obtiene correlación significativa ya que no presenta valores extremos de calidad, lo cual también se refleja en su bajo coeficiente de variación. Se observa que las variables más correlacionadas con las categorías 1 y 0 son la DBO, el ion fosfato y los ST. Esto a pesar de que no son las variables que tienen más peso en el índice (**CUADRO 3.2**). Es de notar que estas variables no son las que tienen más peso en el índice y aun así podrían tener mayor importancia en la clasificación. Por otro lado, el pH tiene un peso considerable en el índice, pero tiene la correlación más baja y estadísticamente no significativa. Sin embargo, una baja correlación no implica necesariamente que esos parámetros vayan a tener poca relevancia en las modelaciones, ya que los algoritmos de ML van más allá de relaciones lineales o monotónicas [57].

Continuando con el EDA, se evaluó la capacidad que tienen los diferentes pares de variables para separar según la clasificación simplificada mediante un gráfico de pares (**Fig. 5.5**). En la diagonal se observan las distribuciones univariadas de cada parámetro separadas según las dos clases; en el resto de las celdas se presentan las dispersiones bivariadas separadas por clases.

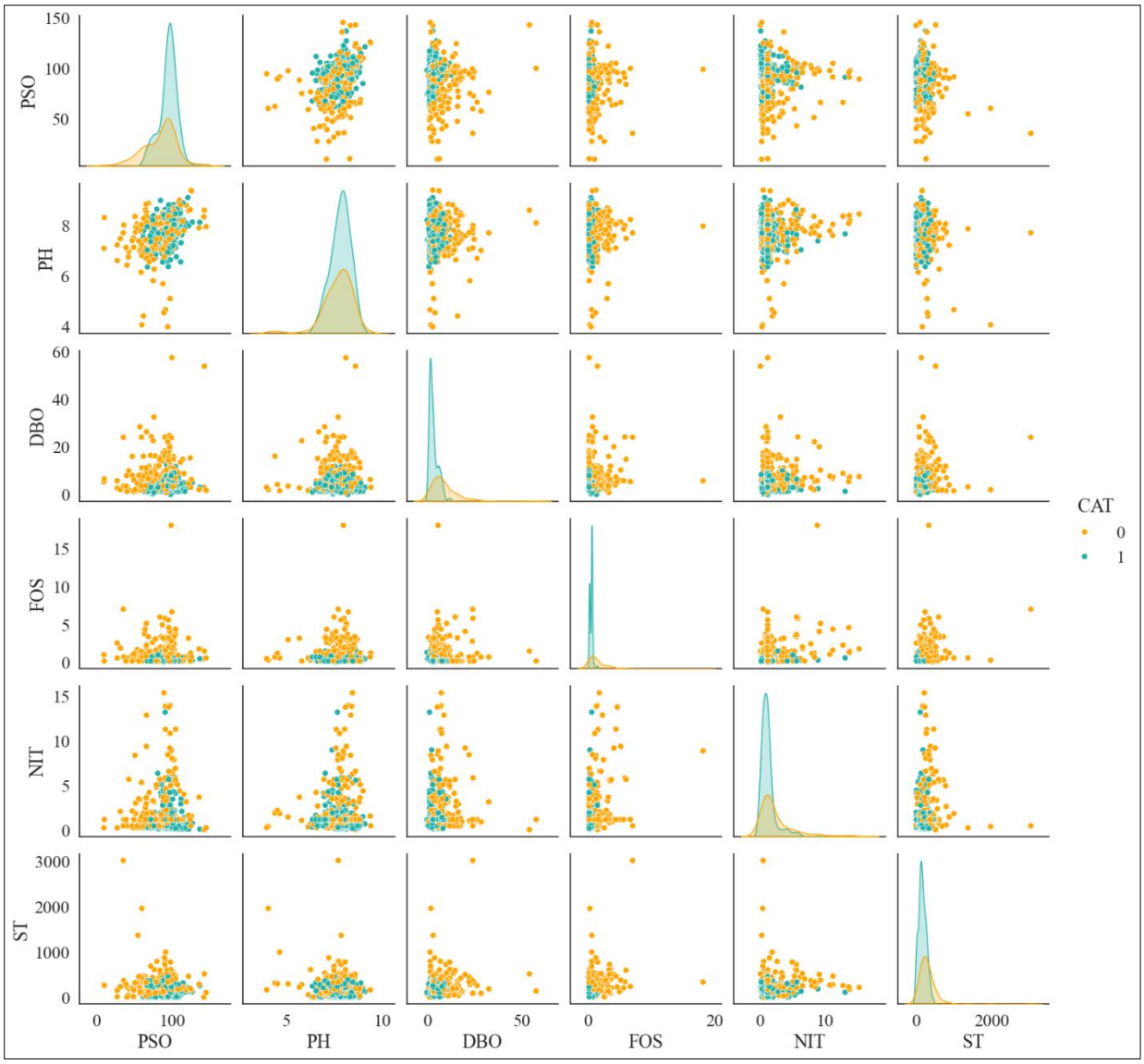

Fig. 5.5. Dispersión de las categorías simplificadas en función de los pares de variables fisicoquímicas del PNMCCAS. Donde el color naranja corresponda a la categoría 0 y el turquesa, a la categoría 1.

En cuanto a las distribuciones univariadas, PSO y pH presentan distribuciones solapadas entre las dos categorías, lo que sugiere que no existe una separación clara al considerar estos parámetros por sí solos. Por otro lado, en los casos de DBO, FOS, NIT y ST las densidades de la categoría 0 están más desplazadas hacia la derecha, lo que indica que valores más altos de estos parámetros representan mala calidad.

Respecto a las distribuciones bivariadas, las combinaciones de DBO versus FOS, NIT y ST y las de ST versus DBO y FOS muestran dispersiones de los puntos de categoría 0 hacia regiones de valores altos de ambos ejes, mientras que los de categoría 1 se agrupan cerca del origen. Esto

sugiere que combinaciones de esos parámetros podrían separar satisfactoriamente ambas categorías. En cambio, las combinaciones con PSO y PH no consiguen separar adecuadamente las categorías.

En general, también se observa en todas las distribuciones que los datos de la categoría 1 se encuentran más juntos, mientras que los de la categoría 0 están más dispersos. Eso podría indicar una mayor facilidad en la modelación para clasificar datos de categoría 1, que representan una mejor calidad de agua.

Al visualizar con mayor detalle la combinación entre DBO, ion fosfato y sólidos totales, se identifica una separación entre las dos categorías con algunos puntos solapados (**Fig. 5.6**). Esto concuerda con lo obtenido en la prueba de correlación, en la que se observa que estas tres variables están más directamente correlacionadas con las categorías. También se observa que algunos datos de la categoría 1 se separan únicamente a lo largo de un eje, ya sea el de DBO, el de ion fosfato o el de ST; mientras que los datos de categoría 0 se concentran en los valores más bajos de las tres variables. El análisis exploratorio indica que la combinación conjunta de estos parámetros permitiría discriminar entre las categorías.

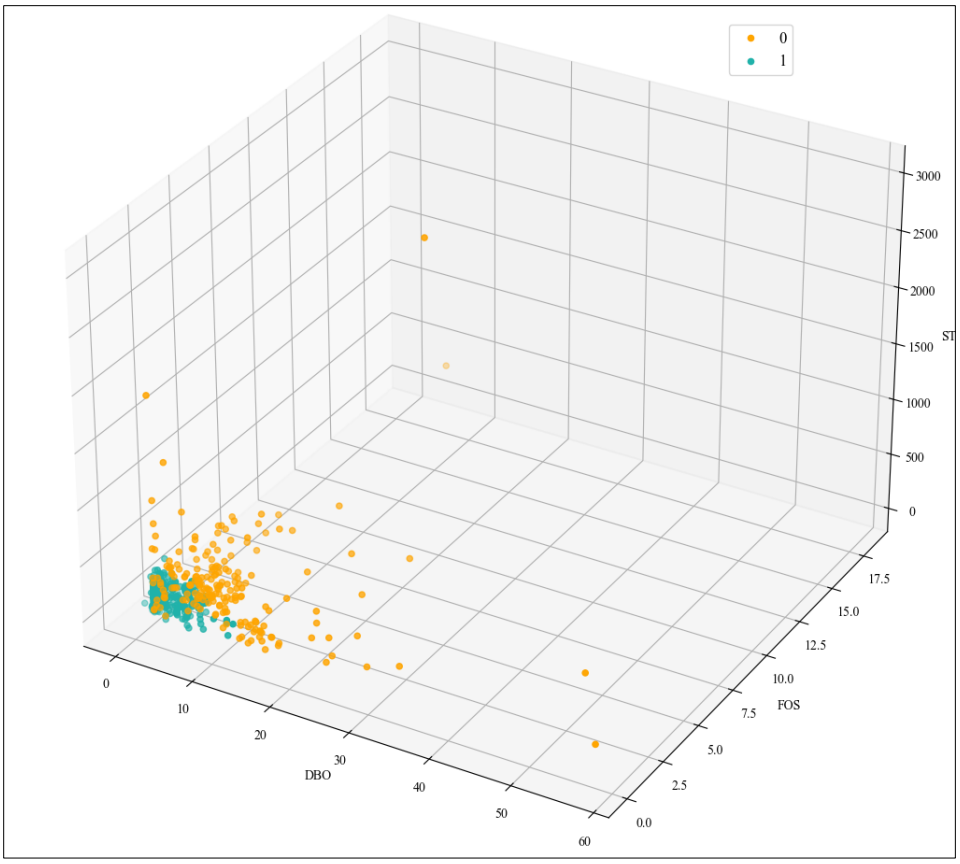

Fig. 5.6. Dispersión tridimensional de las categorías simplificadas en función de las variables DBO, FOS y ST.

El EDA muestra aspectos clave que sirven como insumo para interpretar los resultados de las modelaciones. La DBO, el ion fosfato y el ion nitrato son los parámetros con mayor variabilidad. Además, la DBO, el ion fosfato y los sólidos totales muestran una alta correlación con las categorías de calidad, así como una capacidad para separar los datos de ambas clases. Estos hallazgos indican que la DBO, ion fosfato y los sólidos totales podrían tener una mayor relevancia en las modelaciones.

5.2 Modelación inicial

Con los tres modelos: DT, RF y XGB, se obtiene un rendimiento similar al evaluarlos con la base de datos original mediante validación cruzada estratificada **(Fig. 5.7**). El promedio de las cuatro métricas, para los tres algoritmos, se encuentra por encima de 0,80 e incluso algunos por encima de 0,90 (**CUADRO 5.4**), por lo que se consideran modelos de buen rendimiento [49, 55]. Para la categoría 0 se observa que la sensibilidad es consistentemente inferior que la exactitud y la

precisión, lo cual se puede deber a la menor representación y mayor dispersión de los datos así categorizados, por lo que son más difíciles de clasificar.

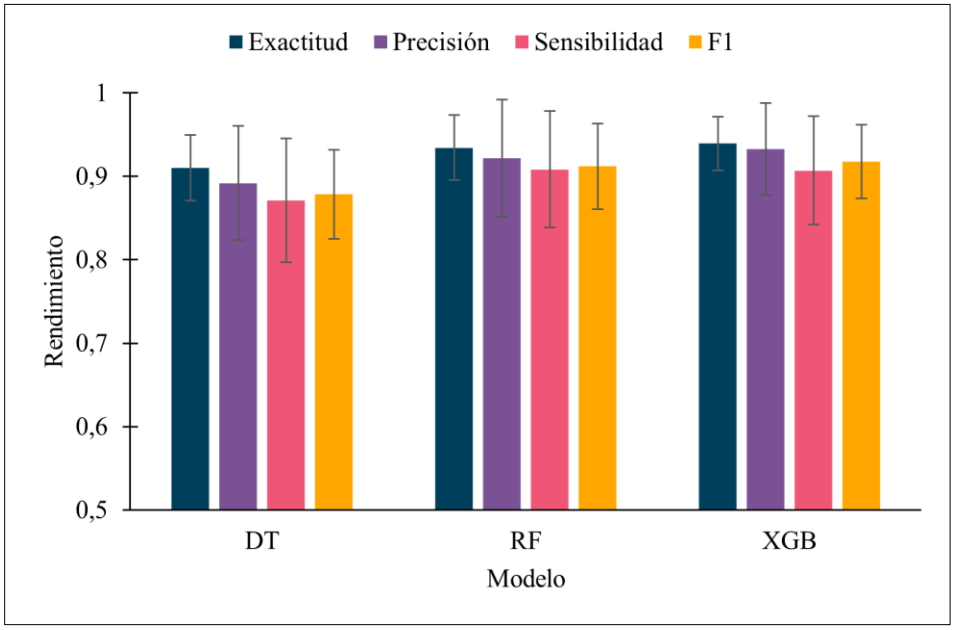

Fig. 5.7. Resultados promedio de la validación cruzada estratificada para la modelación inicial.

CUADRO 5.4
RESULTADOS PROMEDIO DE LA VALIDACIÓN CRUZADA PARA LA MODELACIÓN INICIAL

Métrica	Modelo		
	DT	RF	XGB
Exactitud	0,910 ± 0,040	0,934 ± 0,039	0,939 ± 0,032
Precisión	0,892 ± 0,068	0,922 ± 0,070	0,933 ± 0,055
Sensibilidad	0,871 ± 0,074	0,908 ± 0,070	0,907 ± 0,065
F1	0,879 ± 0,053	0,912 ± 0,051	0,918 ± 0,044

En la **Fig. 5.8** se muestran gráficos de cajas y bigotes para los resultados de sensibilidad, ya que esta es la métrica de mayor interés.

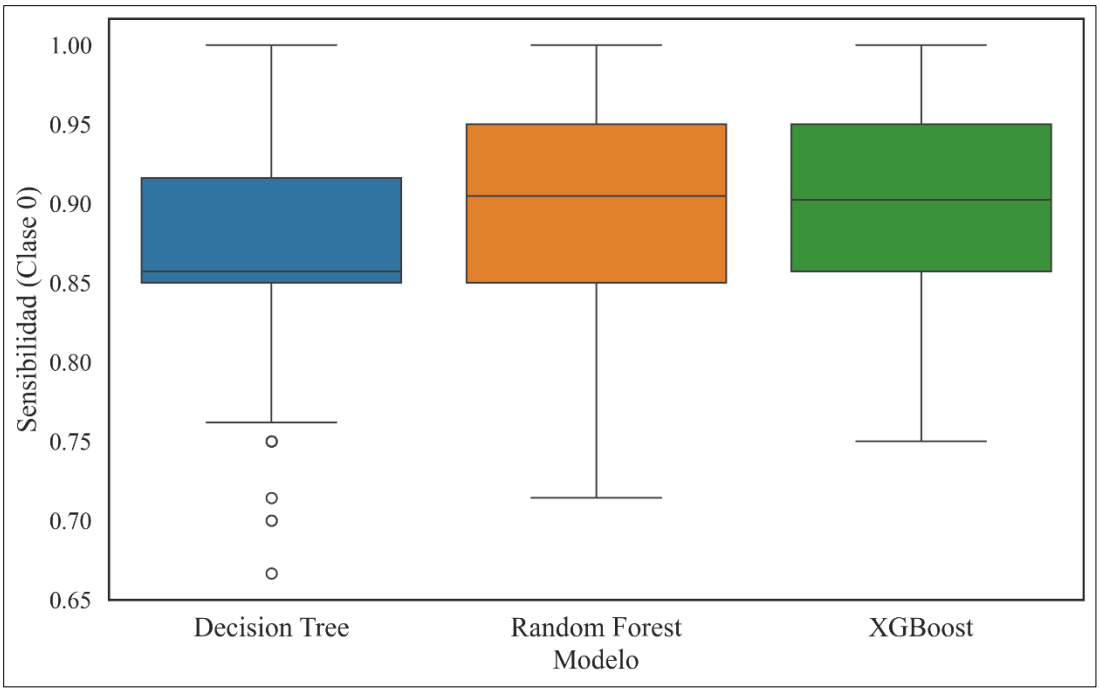

Fig. 5.8. Resultados promedio de sensibilidad para la modelación inicial.

Para determinar cuál modelo tiene significativamente mejor rendimiento, se debe utilizar una prueba estadística que compare los tres modelos considerando que los resultados no son independientes entre sí, ya que se evaluaron con las mismas particiones de los datos. Primero se realiza una prueba de Shapiro-Wilk ($\alpha = 0{,}05$) para determinar si los resultados de Sensibilidad de cada modelo tienen distribución normal. Como en los tres casos se obtuvo un $p < 0{,}05$, se rechaza la hipótesis nula para los tres modelos. Se concluye con un 95% de confianza que ninguna distribución de resultados de Sensibilidad presenta comportamiento normal.

CUADRO 5.5
RESULTADOS DE LA PRUEBA DE NORMALIDAD SHAPIRO-WILK APLICADA A LOS VALORES DE SENSIBILIDAD PARA LA MODELACIÓN INICIAL

Modelo	*p* ($\alpha = 0{,}05$)
DT	**1,6E-03**
RF	**8,3E-06**
XGB	**2,6E-06**

Se utiliza la prueba no paramétrica de Friedman para comparar los resultados promedio de Sensibilidad ($\alpha = 0{,}05$), ya que los resultados presentan distribución no normal y son no independientes. Donde H_0: las medianas de la Sensibilidad de los tres modelos son iguales; H_1: al

menos una de las medianas de la Sensibilidad es diferente. Se obtiene un $p = 7{,}36\times10^{-8}$, por lo que se rechaza la hipótesis nula. Se concluye a un 95% de confianza que existe diferencia significativa en la mediana de los resultados de sensibilidad en al menos un modelo.

Consecuentemente, se aplica una prueba post hoc de rangos con signo de Wilcoxon para comparaciones por pares ($\alpha = 0{,}05$), bajo las hipótesis: H_0: no hay diferencia significativa en la distribución de los resultados de sensibilidad entre los dos modelos que se comparan, H_1: sí existe diferencia significativa en la distribución de los resultados de sensibilidad entre los dos modelos. Los resultados indican que se rechaza la hipótesis nula para las comparaciones entre DT y RF, y DT y XGB; mientras que se acepta la hipótesis nula para la comparación entre RF y XGB **(CUADRO 5.6**). Es decir, no existe diferencia significativa entre los resultados de Sensibilidad de RF y XGB, para un 95% de confianza. Por lo tanto, la sensibilidad de RF y XGB es significativamente distinta a la de DT, pero no existe diferencia significativa entre RF y XGB.

CUADRO 5.6
RESULTADOS DE LA PRUEBA POST HOC DE RANGOS CON SIGNO DE WILCOXON PARA LA COMPARACIÓN DE RESULTADOS SENSIBILIDAD DE LA MODELACIÓN INICIAL

Comparación	*p* ($\alpha = 0{,}05$)
DT vs RF	**6,36E-07**
DT vs XGB	**3,08E-06**
RF vs XGB	0,812

Seguidamente, se comparan los pesos internos de los tres modelos. Los tres coincidieron en el mismo orden de importancia interna de las variables (**Fig. 5.9**). El ion fosfato tiene la mayor relevancia, seguido de la DBO, el PSO, los ST, el ion nitrato y finalmente el pH. Esto indica que los modelos se basaron principalmente en los valores de ion fosfato y DBO para aprender a clasificar los datos, mientras que el pH y el ion nitrato tuvieron menor influencia. En comparación con los pesos originales del índice, la importancia relativa de la DBO, el ion fosfato y los sólidos totales aumentó; mientras que disminuyó para el PSO, el pH y el ion nitrato.

Si bien los tres algoritmos coincidieron en el orden, XGB asignó un peso relativamente mayor al ion fosfato. Esto concuerda con el funcionamiento teórico del algoritmo, que optimiza los árboles de decisión en el proceso de *boosting* con las variables que tuvieron mayor importancia en los primeros árboles, lo que genera pesos menos homogéneos y potencialmente sesgados. Por el otro

lado, RF da una distribución de pesos más homogénea, ya que en el proceso de *bagging* genera árboles independientes con subconjuntos aleatorios de datos y variables, por lo que prueba todos los parámetros sin priorizar excesivamente alguno.

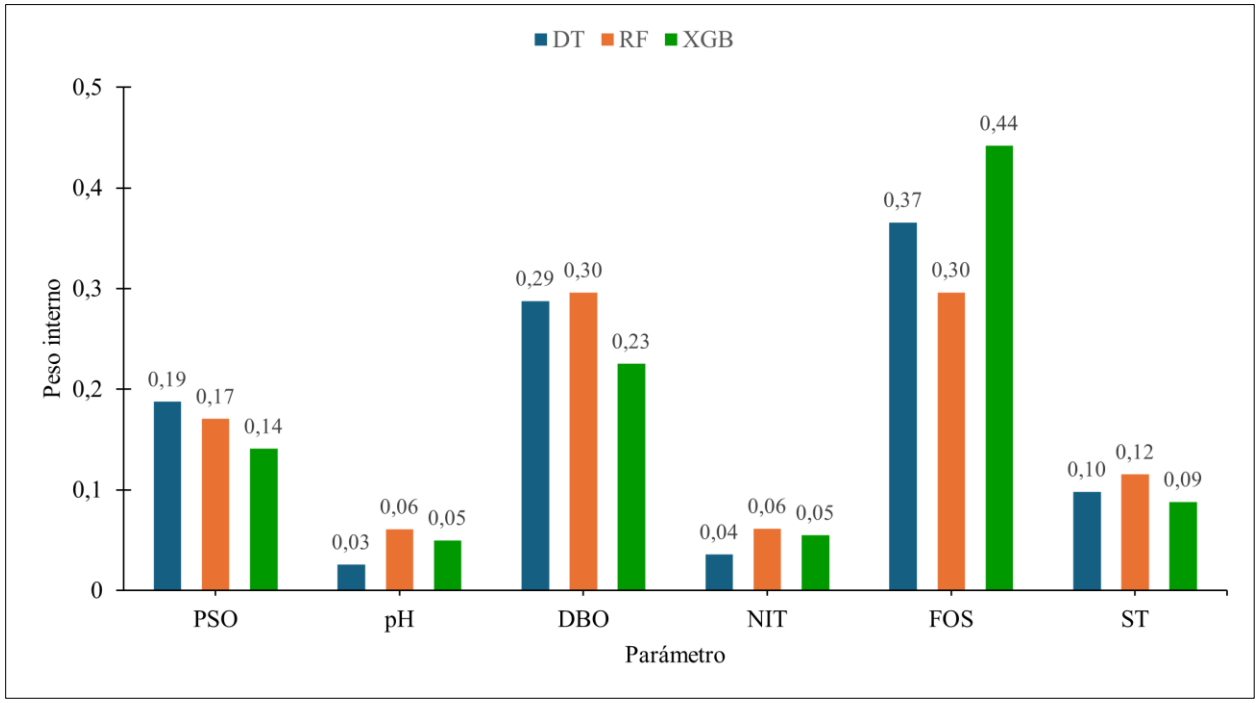

Fig. 5.9. Pesos internos obtenidos en las primeras modelaciones.

Como los pesos en RF son más homogéneos, pueden cubrir un amplio escenario de esquemas de contaminación. Un ICA con una distribución de pesos desigual, como la otorgada por XGB, podría ignorar escenarios con múltiples escenarios de contaminación ya que opacaría los parámetros con menor peso. Considerando lo anterior, el modelo de RF ofrece una mayor interpretabilidad y estimaciones más confiables de la importancia de las variables en comparación a XGB, por el funcionamiento interno del algoritmo.

Aunque se nota un desacoplamiento de las importancias relativas en comparación con el índice, los pesos obtenidos pueden estar sesgados por los originales, ya que la categorización de calidad utilizada para el entrenamiento está definida por el índice. Por lo tanto, se debe aplicar el proceso iterativo propuesto. Los pesos internos del modelo de RF se seleccionan para incorporarlos en la hoja de cálculo dinámica y así obtener un nuevo conjunto de datos con clasificaciones actualizadas.

5.3 Análisis del primer conjunto de datos ajustado

La sustitución por los pesos internos del modelo RF en el ICA-NSF-CR da lugar a un nuevo conjunto de datos con una distribución de clases actualizadas. Notablemente, aparecen datos en la categoría “rojo” y se incrementa la cantidad de las clases “naranja” y “amarillo”. La distribución final es de 0 observaciones en la clase “Azul”, 237 en “Verde”, 223 en “amarillo”, 73 en “naranja” y 6 en “rojo”. Aplicando la clasificación simplificada, se invierte la relación entre las categorías 0 y 1, donde ahora se tienen 302 y 237 datos, respectivamente. En general, el ajuste de pesos resulta en una clasificación más conservadora de la calidad del agua, lo que sugiere que los pesos originales podrían estar atenuando la relevancia de ciertos parámetros en la clasificación.

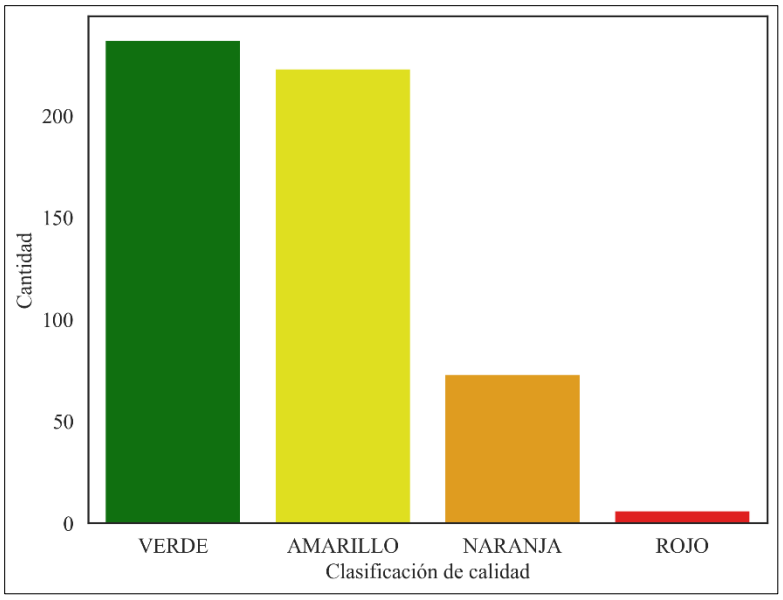

Fig. 5.10. Distribuciones de clases del primer conjunto de datos ajustado.

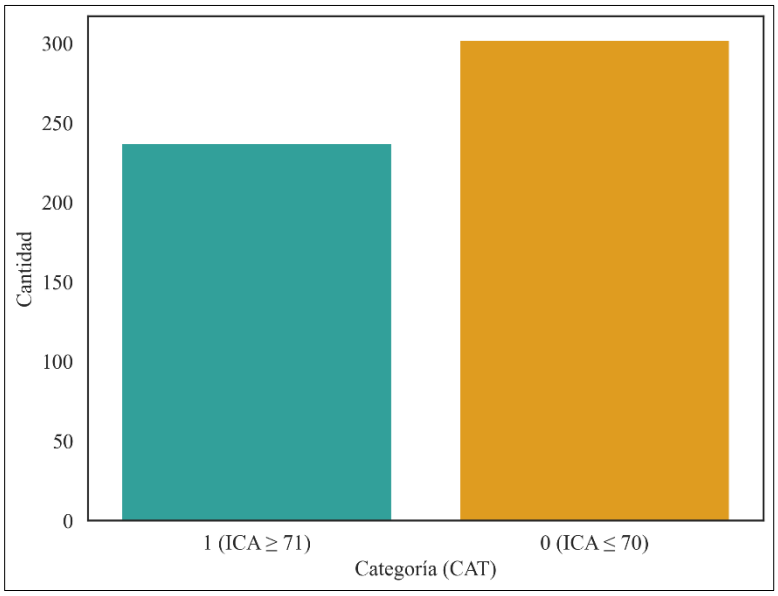

Fig. 5.11. Distribución de categorías 1 y 0 en el primer conjunto de datos ajustado.

Para evaluar si esta nueva distribución de categorías 1 y 0 es significativamente diferente que la original se aplica una prueba de proporciones. Para ello se compara la proporción de datos de categoría 0 (π) en los dos conjuntos de datos, bajo las siguientes hipótesis ($\alpha = 0{,}05$): H_0: $\pi_{original} = \pi_{nuevo}$, H_1: $\pi_{original} \neq \pi_{nuevo}$. Los resultados indican un $p < 0{,}05$, por lo que se rechaza la hipótesis nula (**CUADRO 5.7**). Por lo tanto, se asegura con un 95% de confianza que existe una diferencia significativa en la proporción de datos de categoría 0 entre el conjunto original y el conjunto

ajustado. Ese resultado indica que la sustitución de los pesos originales del índice por los obtenidos mediante RF genera un cambio estructural en la distribución de clases. El proceso se realiza iterativamente hasta alcanzar una distribución de categorías que no sea significativamente distinta que la anterior, buscando la convergencia del ajuste en la distribución de pesos.

CUADRO 5.7
RESULTADOS DE LA PRUEBA DE PROPORCIONES DE CATEGORÍA 0 ENTRE EL CONJUNTO DE DATOS ORIGINAL Y EL AJUSTADO CON LOS PRIMEROS PESOS

Conjunto de datos	Proporción datos categoría 0	p ($\alpha = 0{,}05$)
Original	0,375	**1,51E-09**
Ajustado 1	0,560	

5.4 Optimización de la modelación

El nuevo conjunto de datos se prueba nuevamente con los tres algoritmos con el objetivo de observar si se obtiene una diferencia significativa en el desempeño entre algoritmo. Los resultados de la validación cruzada muestran un rendimiento alto (> 90%) y similar en las cuatro métricas para los tres modelos (**Fig. 5.12**). Como resultado del ajuste en la distribución de pesos y de categorías, se obtiene una mayor homogeneidad en los resultados (**CUADRO 5.7**).

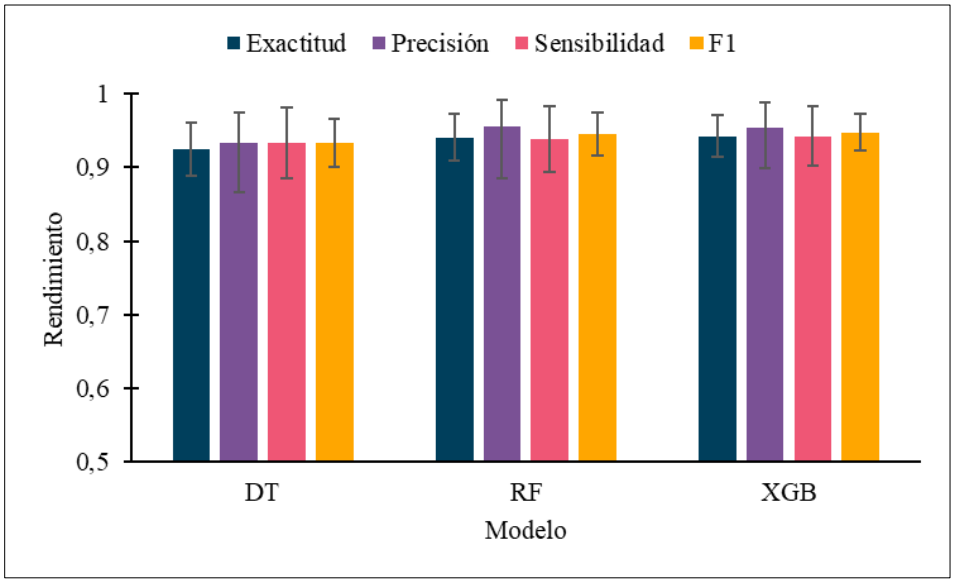

Fig. 5.12. Resultados promedio de la validación cruzada estratificada para la clasificación ajustada.

CUADRO 5.8
RESULTADOS PROMEDIO DE LA VALIDACIÓN CRUZADA ESTRAFICADA PARA LA CLASIFICACIÓN AJUSTADA

Métrica	Modelo		
	DT	RF	XGB
Exactitud	0,925 ± 0,036	0,941 ± 0,032	0,943 ± 0,028
Precisión	0,934 ± 0,040	0,956 ± 0,036	0,955 ± 0,033
Sensibilidad	0,933 ± 0,048	0,939 ± 0,045	0,943 ± 0,041
F1	0,933 ± 0,033	0,946 ± 0,029	0,948 ± 0,025

Los resultados específicos de la sensibilidad (**Fig. 5.13**) muestran una mayor similitud entre sí en comparación con las primeras modelaciones. Se procede a comparar si existe diferencia significativa entre los modelos.

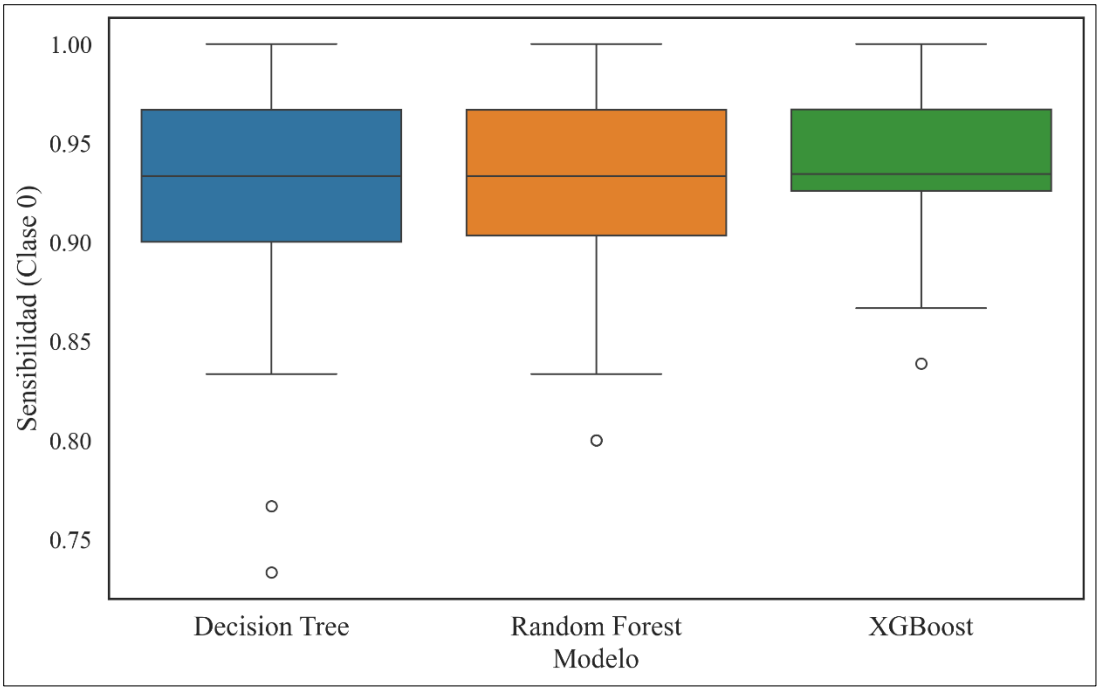

Fig. 5.13. Resultados promedio de sensibilidad para la clasificación ajustada.

Al aplicar la prueba Shapiro-Wilk ($\alpha = 0,05$) a los resultados de sensibilidad de cada modelo se obtiene un $p < 0,05$ en los tres casos (**CUADRO 5.9**), por lo que se rechaza la hipótesis nula para los tres modelos. Se concluye con un 95% de confianza que ninguna distribución de resultados de Sensibilidad en las segundas modelaciones presenta comportamiento normal. Por lo tanto, también se debe aplicar la prueba no paramétrica de Friedman ($\alpha = 0,05$) para datos no independientes.

CUADRO 5.9

RESULTADOS DE LA PRUEBA DE NORMALIDAD SHAPIRO-WILK APLICADA A LOS VALORES DE SENSIBILIDAD PARA LA MODELACIÓN CON LA BASE DE DATOS AJUSTADA

Modelo	p ($\alpha = 0,05$)
DT	5,1E-07
RF	3,8E-06
XGB	1,0E-05

La prueba de Friedman arroja un p = 0,052 por lo que se acepta la hipótesis nula. Se concluye a un 95% de confianza que no existe diferencia significativa en la mediana de los resultados de Sensibilidad entre los modelos. Esto indica que el ajuste en la base de datos genera una mayor facilidad predictiva de todos los modelos, incluyendo DT que es el más básico y la base de los otros dos.

El orden de las importancias internas de los parámetros cambia con respecto a las primeras modelaciones (**Fig. 5.14**). Para el caso de RF, el parámetro de mayor relevancia es DBO, seguido de ion fosfato, sólidos totales, PSO, ion nitrato y pH. El nuevo orden de importancias internas continúa siendo diferente al de los pesos originales. Considerando que no hubo diferencia significativa en el rendimiento entre modelos: al igual que en la modelación anterior, se selecciona el modelo RF como el óptimo en este caso, al considerar su interpretabilidad y su manejo de las importancias internas.

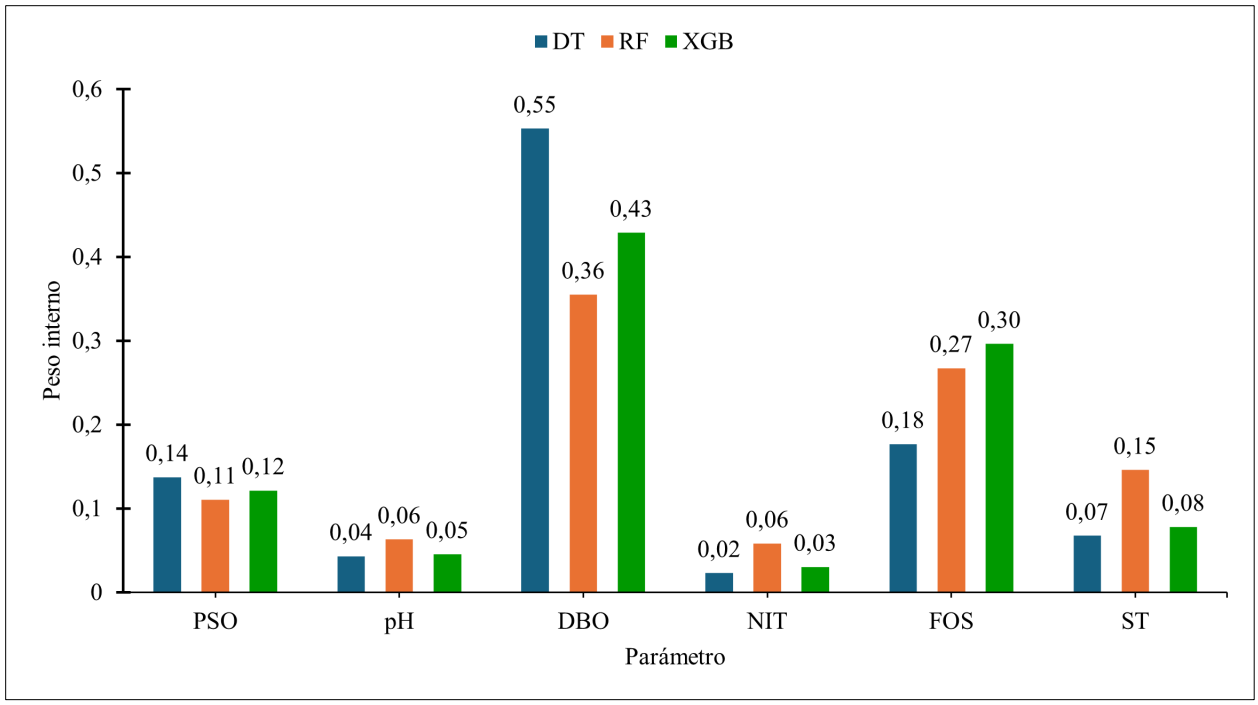

Fig. 5.14. Pesos internos obtenidos en la modelación con clasificación ajustada.

5.5 Análisis del segundo conjunto de datos ajustado

La nueva distribución de clases contiene 0 datos en la clase "azul", 216 en "verde", 223 en "amarillo", 91 en "naranja y 9 en "rojo" (**Fig. 5.15**). Para la clasificación simplificada se tienen 216 observaciones en la categoría 1 y 323 en la 0 (**Fig. 5.16**). Con respecto al anterior conjunto ajustado, disminuye la cantidad de datos en "verde", aumentan en "naranja" y "rojo", y se mantienen igual en "amarillo". Por lo tanto, se observa un aumento en números absolutos de la categoría 0; sin embargo, debe evaluarse si ese aumento es significativo.

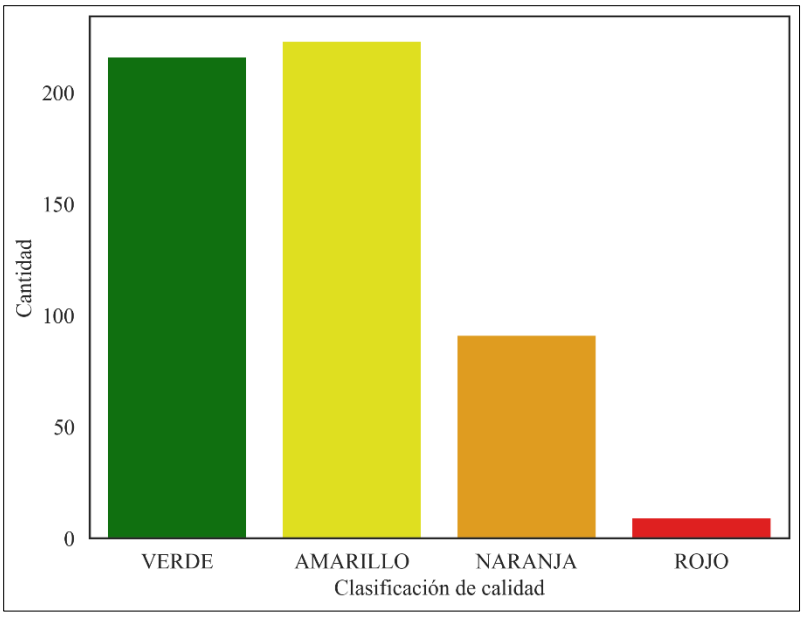

Fig. 5.15. Distribución de clases en el segundo conjunto de datos ajustado.

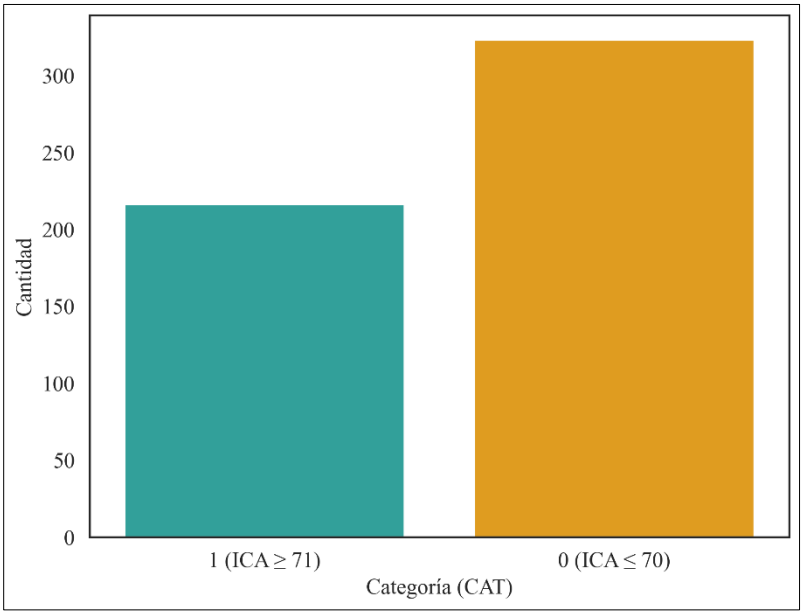

Fig. 5.16. Distribución de categorías 1 y 0 en el segundo conjunto de datos ajustado.

La prueba de proporciones ($\alpha = 0{,}05$) da como resultado un $p = 0{,}217$, por lo que se rechaza la hipótesis nula (**.**

CUADRO 5.10). Se afirma con un 95% de confianza que no existe diferencia significativa en la proporción de datos categoría 0 entre los dos conjuntos ajustados. Por lo tanto, no existe diferencia entre ambos conjuntos ajustados de datos y las iteraciones finalizan en este punto.

CUADRO 5.10

RESULTADOS DE LA PRUEBA DE PROPORCIONES DE CATEGORÍA 0 ENTRE LOS DOS CONJUNTOS DE DATOS AJUSTADOS

Conjunto de datos	Proporción datos categoría 0	p (α = 0,05)
Ajustados 1	0,560	0,217
Ajustados 2	0,599	

5.6 Análisis de los pesos objetivos

Los pesos utilizados para obtener el conjunto final de datos clasificados se proponen ajuste del ICA-NSF-CR (**CUADRO 5.11**). Estos corresponden a las importancias internas de los parámetros en el modelo de RF en la modelación final.

CUADRO 5.11

PESOS OBJETIVOS PARA EL ICA-NSF-CR OBTENIDOS CON RANDOM FOREST

Parámetro	Peso
Demanda bioquímica de oxígeno (5 días, mg / L)	0,36
Ion fosfato (mg / L)	0,27
Sólidos totales (mg / L)	0,15
Oxígeno disuelto (% saturación)	0,11
Ion nitrato (mg / L)	0,06
pH	0,06

Para complementar el análisis de los pesos internos del modelo, se presentan los resultados del procesamiento SHAP (**Fig. 5.17**). En la gráfica, cada punto corresponde a una predicción realizada por el modelo. El eje horizontal indica la contribución del parámetro al resultado, expresada como valor SHAP, donde los valores positivos aumentan la probabilidad de clasificar en la categoría 0, mientras que valores negativos la reducen. En general, cuanto mayor sea el valor absoluto del SHAP, mayor es la influencia del parámetro en esa predicción [57]. El color de cada punto representa el valor numérico del parámetro en esa observación, donde los valores bajos se muestran de color azul y los altos, de rojo.

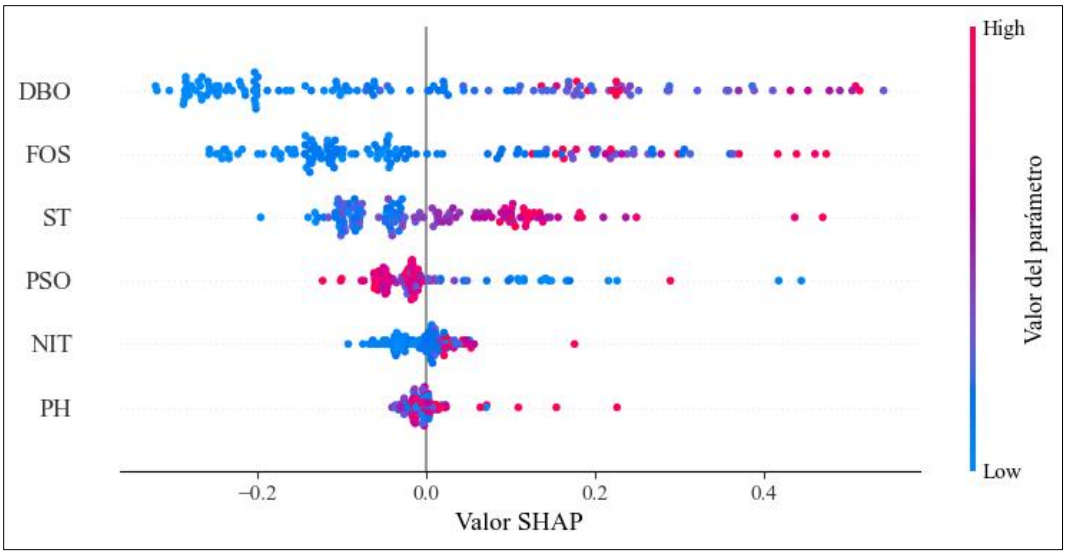

Fig. 5.17. Valores SHAP del modelo para clasificar en categoría 0.

Los parámetros se encuentran en orden de importancia, de mayor a menor, según sus valores SHAP. Se observa, en primer lugar, que el orden coincide con el de las importancias internas (**CUADRO 5.11**). Ese orden indica que la DBO, FOS y los ST presentan un amplio rango de valores SHAP, lo representa su alta importancia predictiva. Para esos tres parámetros puede observarse que valores altos de estos parámetros tienden a asociarse con SHAP positivos, por lo que favorecen la clasificación en la categoría 0; mientras que los valores bajos favorecen la predicción de la categoría 1, al tener SHAP negativos. Esto también señala que la presencia de valores extremos en las distribuciones de los parámetros fortalece la capacidad de los modelos para clasificar. Estos valores, al estar asociados a condiciones de calidad más definidas, proporcionan señales que permiten al algoritmo encontrar umbrales de decisión más adecuados [57].

Que un parámetro tenga una importancia interna o valor SHAP bajo, no significa que sea innecesario de medir. Todos los parámetros tienen su importancia ambiental intrínseca, ya que indican algún tipo de condición específica del agua. El monitoreo de todos los parámetros es necesario para asegurar una evaluación integral del cuerpo de agua, donde se evalúen todos los posibles riesgos. Los pesos de un ICA se ajustan para evitar que uno o varios parámetros enmascaren a otros, sin que esto desmerite a los de menor peso [3, 34].

En el contexto específico de este estudio, el pH y el ion nitrato tienen los pesos objetivos más bajos debido a que su variabilidad estadística no presenta la suficiente información para que los algoritmos puedan clasificar adecuadamente con base en ellos [27, 55, 57]. Por el otro lado, la

DBO, ion fosfato y sólidos totales sí tenían la suficiente información para entrenar clasificadores de buen rendimiento. Como se observó en el EDA, estos tres parámetros justamente tienen los mayores coeficientes de variación, las mayores correlaciones con las categorías simplificadas de calidad y permiten separar espacialmente los datos de ambas clases. La información que cada parámetro logra transmitir está relacionada con su incidencia en el territorio [10, 40], por lo que se puede analizar la relevancia ambiental de cada parámetro según las condiciones de Costa Rica, para comprender los pesos internos del modelo.

La demanda biológica de oxígeno, que fue considerada la variable más importante en los modelos (0,36), está relacionada con la contaminación por materia orgánica y fecal. En Costa Rica, para el 2024 [72], el 75% de la población utilizaba tanque séptico y solamente un 25% gestionaba las excretas de forma segura, lo que genera un impacto recurrente a los cuerpos de agua superficiales cercanos al no haber un tratamiento adecuado de las aguas residuales domésticas [73]. En consecuencia, múltiples estudios en el país han señalado que la DBO es uno de los parámetros de mayor alteración en los monitoreos [2, 74-76]. Por lo tanto, el peso otorgado por el modelo tiene un sustento ambiental claro, dada la relevancia de la DBO y su relación con la presencia de materia fecal.

La concentración de ion fosfato ocupa el segundo lugar en importancia (0,27). Este nutriente se encuentra naturalmente en algunos cuerpos de agua superficiales, pero múltiples actividades humanas pueden generar concentraciones excesivas y provocar eutrofización [77]. Se ha señalado que la fertilización intensiva y las aguas jabonosas mal tratadas pueden aumentar la concentración del ion fosfato [78, 79]. En Costa Rica se ha reportado una alta tasa de contaminación por aguas jabonosas debido a que muchos hogares no conectan esas tuberías con el sistema de tratamiento de aguas residuales [74, 80]. También la aplicación excesiva de fertilizantes en muchos lugares del país libera grandes cantidades de elementos fosforados y otros nutrientes al ambiente [81], donde debe considerarse que el sector agrícola es el mayor generador de aguas residuales en el país [82].

La concentración de ion nitrato también indica presencia de nutrientes y de contaminación por fertilizantes, pero tiene menor importancia en los modelos (0,06). Esto se puede deber a que la influencia de la concentración de ion fosfato opaca a la concentración de ion nitrato en los modelos, ya que transmite más información [73, 83]. El ion fosfato es menos soluble en agua así que tiene un impacto más local y mayor persistencia en el ambiente [84]. Por ello, el ion fosfato es

considerado un nutriente más limitante en los ecosistemas acuáticos superficiales que el ion nitrato [85-87]. Además, la contaminación por aguas jabonosas es una fuente adicional de contaminación por ion fosfato especialmente en zonas urbanas [74, 88-90]. Asimismo, el ion nitrato presentan un menor coeficiente de variación, menor correlación con las categorías de calidad y no logra separar adecuadamente las clases, según lo analizado en el EDA.

Los sólidos totales muestran una relevancia media-alta en el modelo, con un peso de 0,15. Los ST no necesariamente indican contaminación, ya que pueden provenir de fuentes naturales como la erosión y el arrastre de sedimentos [2, 73, 91]. En Costa Rica se ha detectado un aumento de fenómenos que pueden ocasionar episodios de altas concentraciones de ST, como: vertidos de aguas residuales, cambios de uso del suelo, especialmente movimientos de tierra y construcciones, además de desviación, uso y dragado ilegal de cauces y nacientes [2, 92]. Por lo tanto, los valores extremos de ST pueden estar relacionados con actividades humanas y la degradación de la cobertura natural [73, 93], aunque una alta concentración no necesariamente indique baja calidad.

El porcentaje de saturación de oxígeno disuelto tiene una importancia media en el modelo (0,11). Este parámetro depende del flujo del río y su aireación, por lo que en ciertas ocasiones no está relacionado directamente con la calidad [83, 94]. En ríos con alto caudal y movimiento, puede haber valores adecuados de oxígeno disuelto aún en presencia de contaminación [95]. Por otro lado, una correcta aireación permite que la materia orgánica y otros contaminantes sean degradados, así que un bajo porcentaje de oxígeno disuelto puede indicar contaminación y una poca capacidad del río para depurarse [2]. Es un parámetro con alta relevancia ambiental para el monitoreo continuo de cuerpos de agua, pero a nivel general del territorio no presenta alta significancia relativa en comparación con los otros parámetros [2, 73, 74, 96].

El pH tiene la menor importancia en el modelo, con un 0,06. Este parámetro es vital en el monitoreo de cuerpos de agua superficiales, ya que aumentos o disminuciones abruptas de pH pueden afectar gravemente la vida acuática. No obstante, los valores extremos están asociados con contaminación por efluentes industriales y actividades mineras [97, 98]. Estas son fuentes minoritarias de contaminación en Costa Rica, por lo que la incidencia general es muy baja y los valores suelen ser estables entre 7 y 8 [2, 73, 74, 96]. Debido a eso, los modelos no pueden discernir entre las categorías de calidad con base en el pH [10, 99]. Lo mismo sucede con la concentración

de ion nitrato que, si bien es indicador de un tipo de contaminación grave, su relativa baja incidencia en la base de datos no permite utilizarlo como discriminador de la calidad.

Cabe aclarar que cada punto de datos, al ser un punto específico de un río, representa un contexto particular, influenciado por distintas fuentes de contaminación e interacciones ecosistémicas y fisicoquímicas. Los pesos obtenidos con el modelo son generales, ya que nacen del procesamiento de la base de datos completa, por lo que reflejan una visión conjunta del estado de la calidad del agua superficial en Costa Rica. Es decir, la resolución de los pesos objetivos es baja, ya que están ajustados considerando el estado general del país.

Esto significa que, si bien los pesos objetivos propuestos son útiles como referencia nacional, no necesariamente abordan las características locales de cada cuerpo de agua. Es pertinente considerar que los pesos en un ICA deben responder al objetivo del índice. Los pesos objetivos carecen de esa dimensión, ya que se obtienen a partir de operaciones matemáticas aplicadas a los datos, sin que haya un proceso humano de discernimiento. Los pesos objetivos propuestos, entonces, deben servir como una base para futuros cambios en el ICA-NSF-CR con el fin de optimizarlo más al territorio costarricense.

6 CONCLUSIONES Y RECOMENDACIONES

6.1 Conclusiones

Las modelaciones con algoritmos de aprendizaje automático permiten optimizar el ICA-NSF-CR, ya que generan pesos objetivos para los parámetros a partir de su importancia en la clasificación en categorías de calidad. Estos pesos objetivos pueden servir para ajustar el índice e incidir en la normativa nacional. Estos pesos objetivos también complementan el criterio profesional con base en los datos.

Los pesos obtenidos mediante el modelo seleccionado reflejan los tipos de contaminación más concernientes en la actualidad para los ríos del país. Esto es así porque los modelos de aprendizaje automático aprenden de los patrones en los datos, así que pueden aplicarse para conocer las relaciones entre parámetros fisicoquímicos y la calidad ambiental.

La metodología de ajuste iterativo de la distribución de pesos permitió mejorar la eficiencia de los algoritmos con base en los datos disponibles. Este enfoque generó conjuntos de datos con una distribución más equilibrada entre categorías, lo que permitió un aprendizaje más estable y pesos más representativos.

Además, la metodología propuesta contribuye a un mejor ajuste de los modelos a las condiciones nacionales, donde los valores extremos en los parámetros fisicoquímicos son menos frecuentes que en otros contextos internacionales. Por lo tanto, al refinar la clasificación iterativamente, se obtiene una distribución de pesos más sensible para un seguimiento de la calidad.

La selección del modelo óptimo para esta aplicación específica no depende únicamente del rendimiento, sino también de su funcionamiento interno e interpretabilidad. Esto adquiere relevancia en el modelaje ambiental, donde se necesita información racional para la toma de decisiones.

El modelo de RF demostró ser el más adecuado para modelar la calidad del agua superficial, debido a que otorgó una distribución de pesos más homogénea entre los parámetros. Una distribución de pesos balanceada permite abarcar una mayor variedad de condiciones de calidad del agua sin depender en exceso de un único parámetro.

6.2 Recomendaciones

Se recomienda, con base en el producto principal de esta tesis, redistribuir los pesos en el ICA-NSF-CR para darle más relevancia a los parámetros DBO e ion fosfato, por lo menos. El ajuste debe hacerse sin descuidar los parámetros que tuvieron menores pesos objetivos pero que son de gran importancia para caracterizar la calidad ambiental y el monitore de la actividad antropogénica, como el ion nitrato.

A nivel metodológico, se podría explorar otros algoritmos, como redes neuronales y métodos de ensamblaje más complejos, con el objetivo de incrementar el rendimiento, aun cuando su interpretabilidad sea menor. En estos casos, se pueden utilizar técnicas como SHAP y LIME para obtener pesos objetivos a partir de modelos con poca interpretabilidad.

Asimismo, se recomienda valorar la optimización de hiperparámetros con una base de datos más amplia. Para bases de datos relativamente pequeñas, como la utilizada en este estudio, el beneficio de un ligero aumento de rendimiento no subsana el costo computacional; no obstante, podría ser factible con una mayor cantidad de datos.

Los resultados obtenidos están ajustados a la realidad nacional, esto implica que otros países podrían obtener resultados distintos. Por lo tanto, se recomienda implementar la metodología utilizando bases de datos propias y representativas de cada territorio.

La ampliación de la red nacional de monitoreo, incorporando un mayor volumen de datos vinculados con indicadores socioeconómicos, ambientales y climáticos, así como análisis de series temporales, podría fortalecer significativamente la capacidad predictiva de los algoritmos. Se recomienda desarrollar estudios en esta línea con el fin de planificar políticas públicas y de ordenamiento territorial. Con los recursos financieros y operativos limitados, esta expansión del monitore puede hacerse de manera gradual, priorizando cuencas críticas, muestreos estacionales y parámetros de mayor peso.

7 REFERENCIAS

[1] W. Musie y G. Gonfa, "Fresh water resource, scarcity, water salinity challenges and possible remedies: A review," *Heliyon*, vol. 9, no. 8, 2023, doi: 10.1016/j.heliyon.2023.e18685.

[2] G. Pérez Gómez, V. Alvarado García, J.A. Rodríguez Rodríguez, F. Herrera y R. Sánchez Gutiérrez, "Calidad fisicoquímica y microbiológica del agua superficial del río Grande de Tárcoles, Costa Rica: un enfoque ecológico," *URJ*, vol. 13, no. 1, 2021, doi: 10.22458/urj.v13i1.3148.

[3] S. Chidiac, P. El Najjar, N. Ouaini, Y. El Rayess y D. El Azzi, "A comprehensive review of water quality indices (WQIs): history, models, attempts and perspectives," *Rev. Environ. Sci. Bio/Technol.*, vol. 22, no. 2, pp. 349–395, 2023, doi: 10.1007/s11157-023-09650-7.

[4] P.K. Singh, et al., "Critical review on toxic contaminants in surface water ecosystem: sources, monitoring, and its impact on human health," *Environ. Sci. Pollut. Res.*, vol. 31, no. 45, pp. 56428–56462, 2024, doi: 10.1007/s11356-024-34932-0.

[5] Bijay-Singh y E. Craswell, "Fertilizers and nitrate pollution of surface and ground water: an increasingly pervasive global problem," *SN Applied Sciences*, vol. 3, no. 4, pp. 518, 2021, doi: 10.1007/s42452-021-04521-8.

[6] M.G. Uddin, S. Nash y A.I. Olbert, "A review of water quality index models and their use for assessing surface water quality," *Ecol.Ind.*, vol. 122, pp. 107218, 2021, doi: 10.1016/j.ecolind.2020.107218.

[7] M.G. Uddin, S. Nash, A. Rahman y A.I. Olbert, "A comprehensive method for improvement of water quality index (WQI) models for coastal water quality assessment," *Water Res.*, vol. 219, pp. 118532, 2022, doi: 10.1016/j.watres.2022.118532.

[8] S. Gupta y S.K. Gupta, "A critical review on water quality index tool: Genesis, evolution and future directions," *Ecol. Inf.*, vol. 63, pp. 101299, 2021, doi: 10.1016/j.ecoinf.2021.101299.

[9] M. Kachroud, F. Trolard, M. Kefi, S. Jebari y G. Bourrié, "Water Quality Indices: Challenges and Application Limits in the Literature," *Water*, vol. 11, no. 2, 2019, doi: 10.3390/w11020361.

[10] L. Hernández-Alpízar y J.A. Gómez-Mejía, "Modeling Surface Water Quality Using K-Nearest Neighbors and Random Forest," presentado en *2024 IEEE 6th Int. Conf. on BioInsp. Process. (BIP)*, pp. 1–5, 2024, doi: 10.1109/BIP63158.2024.10885395.

[11] M.E. Pérez-Villanueva, et al., "An integrative water quality evaluation in two surface water bodies from a tropical agricultural region in Cartago, Costa Rica," *Environ. Sci. Pollut. Res.*, vol. 29, no. 15, pp. 21968–21980, 2022, doi: 10.1007/s11356-021-17283-y.

[12] E. Vargas Madrigal, comunicación personal, ene., 2025.

[13] Presidencia de la República y Ministerio de Ambiente y Energía. Decreto Ejecutivo Nº 32868, "Canon por Concepto de Aprovechamiento de Aguas," San José, Costa Rica, Ago. 24, 2005. [En línea]. Disponible: https://pgrweb.go.cr/scij/Busqueda/Normativa/Normas/nrm_texto_completo.aspx?nValor1=1&nValor2=56341.

[14] G. Calvo-Brenes, *Índices E Indicadores Sobre La Calidad Del Agua,* 1 ed. Cartago: Ed. Tecnológica de Costa Rica, 2018. .

[15] M.M.M. Syeed, M.S. Hossain, M.R. Karim, M.F. Uddin, M. Hasan y R.H. Khan, "Surface water quality profiling using the water quality index, pollution index and statistical methods: A critical review," *Environ. Sustainability Indic.*, vol. 18, pp. 100247, 2023, doi: 10.1016/j.indic.2023.100247.

[16] E.F. Franco, R. Ramos, A. Ovando-Javier, E. Montero-Espaillat, S. Bonilla y A. Veda, "Sensores de calidad de agua para el control de la contaminación fisicoquímica en los acuíferos de Latinoamérica: una revisión," *CAC*, vol. 6, no. 1, pp. 45–70, 2023, doi: 10.22206/cac.2023.v6i1.pp45-70.

[17] B. Nath Roy, et al., "Principal component analysis incorporated water quality index modeling for Dhaka-based rivers," *City Environ. Interact.*, vol. 23, pp. 100150, 2024, doi: 10.1016/j.cacint.2024.100150.

[18] M.G. Uddin, et al., "Assessing the impact of COVID-19 lockdown on surface water quality in Ireland using advanced Irish water quality index (IEWQI) model," *Env. Pollut.*, vol. 336, pp. 122456, 2023, doi: 10.1016/j.envpol.2023.122456.

[19] Z. Yu, X. Sun, L. Yan, S. Yu, Y. Li y H. Jin, "Analysis of the Water Quality Status and Its Historical Evolution Trend in the Mainstream and Major Tributaries of the Yellow River Basin," *Water*, vol. 16, no. 17, pp. 2413, 2024, doi: 10.3390/w16172413.

[20] D.K. Lukhabi, P.K. Mensah, N.K. Asare, T. Pulumuka-Kamanga y K.O. Ouma, "Adapted Water Quality Indices: Limitations and Potential for Water Quality Monitoring in Africa," *Water*, vol. 15, no. 9, 2023, doi: 10.3390/w15091736.

[21] D. Kumar, R. Kumar, M. Sharma, A. Awasthi y M. Kumar, "Global water quality indices: Development, implications, and limitations," *Total Env. Advances*, vol. 9, pp. 200095, 2024, doi: 10.1016/j.teadva.2023.200095.

[22] M.G. Uddin, S. Nash, A. Rahman y A.I. Olbert, "A novel approach for estimating and predicting uncertainty in water quality index model using machine learning approaches," *Water Res.*, vol. 229, pp. 119422, 2023, doi: 10.1016/j.watres.2022.119422.

[23] L. Chemeri, J. Cabassi, M. Taussi y S. Venturi, "Development and testing of a new flexible, easily and widely applicable chemical water quality index (CWQI)," *J.Environ.Manage.*, vol. 348, pp. 119383, 2023, doi: 10.1016/j.jenvman.2023.119383.

[24] Z. Liu, et al., "Groundwater Quality Evaluation of the Dawu Water Source Area Based on Water Quality Index (WQI): Comparison between Delphi Method and Multivariate Statistical Analysis Method," *Water*, vol. 13, no. 8, 2021, doi: 10.3390/w13081127.

[25] H. Rajkumar, P.K. Naik y M.S. Rishi, "A comprehensive water quality index based on analytical hierarchy process," *Ecol.Ind.*, vol. 145, pp. 109582, 2022, doi: 10.1016/j.ecolind.2022.109582.

[26] A.P. Mishra, S. Singh, M. Jani, K.A. Singh, C.B. Pande y A.M. Varade, "Assessment of water quality index using Analytic Hierarchy Process (AHP) and GIS: a case study of a struggling Asan River," *Int.J.Environ.Anal.Chem.*, vol. 104, no. 5, pp. 1159–1171, 2024, doi: 10.1080/03067319.2022.2032015.

[27] F. Ding, et al., "Optimization of water quality index models using machine learning approaches," *Water Res.*, vol. 243, pp. 120337, 2023, doi: 10.1016/j.watres.2023.120337.

[28] R.M. Brown, N.I. McClelland, R.A. Deininger y R.G. Tozer, "A water quality index-do we dare," *Water and sewage works*, vol. 117, no. 10, pp. 339–343, 1970.

[29] R. Noori, R. Berndtsson, M. Hosseinzadeh, J.F. Adamowski y M.R. Abyaneh, "A critical review on the application of the National Sanitation Foundation Water Quality Index," *Env. Pollut.*, vol. 244, pp. 575–587, 2019, doi: 10.1016/j.envpol.2018.10.076.

[30] M.S. Gradilla-Hernández, et al., "Assessment of the water quality of a subtropical lake using the NSF-WQI and a newly proposed ecosystem specific water quality index," *Environ Monit Assess*, vol. 192, no. 5, pp. 296, 2020, doi: 10.1007/s10661-020-08265-7.

[31] J.A. López Zúñiga, comunicación personal, feb., 2025.

[32] N.A. Misman, M.F. Sharif, A.J.K. Chowdhury y N.H. Azizan, "Water pollution and the assessment of water quality parameters: a review," *Desalin. Water Treat.*, vol. 294, pp. 79–88, 2023, doi: 10.5004/dwt.2023.29433.

[33] M. Tripathi y S.K. Singal, "Allocation of weights using factor analysis for development of a novel water quality index," *Ecotoxicol. Environ. Saf.*, vol. 183, pp. 109510, 2019, doi: 10.1016/j.ecoenv.2019.109510.

[34] G. Uddin, S. NashyA.I. Olbert, "Optimization of parameters in a water quality index model using principal component analysis," presentado en *39th IAHR World Congr.*, pp. 5739–5744, 2022, doi: 10.3850/IAHR-39WC2521711920221326.

[35] F.L. Gewers, et al., "Principal Component Analysis: A Natural Approach to Data Exploration," *ACM Comput.Surv.*, vol. 54, no. 4, pp. 1–34, 2021, doi: 10.1145/3447755.

[36] F. Ding, L. Chen, C. Sun, W. Zhang, H. Yue y S. Na, "An upgraded groundwater quality evaluation based on Hasse diagram technique & game theory," *Ecol.Ind.*, vol. 140, pp. 109024, 2022, doi: 10.1016/j.ecolind.2022.109024.

[37] W. Zhe, X. Xigang y Y. Feng, "An abnormal phenomenon in entropy weight method in the dynamic evaluation of water quality index," *Ecol.Ind.*, vol. 131, pp. 108137, 2021, doi: 10.1016/j.ecolind.2021.108137.

[38] Y. Zhu, D. Tian y F. Yan, "Effectiveness of Entropy Weight Method in Decision-Making," *Mathematical Problems in Eng*, vol. 2020, no. 1, pp. 3564835, 2020, doi: 10.1155/2020/3564835.

[39] Q. Bao, Z. Yuxin, W. Yuxiao y Y. Feng, "Can Entropy Weight Method Correctly Reflect the Distinction of Water Quality Indices?" *Wat. Resour. Manag.*, vol. 34, no. 11, pp. 3667–3674, 2020, doi: 10.1007/s11269-020-02641-1.

[40] M.G. Uddin, "Development of a novel water quality index model using data science approaches," Tesis de doctorado, Univ. de Galway, 2023. [En línea]. Disponible: http://hdl.handle.net/10379/17786.

[41] G. James, D. Witten, T. Hastie, R. Tibshirani y J. Taylor, "Tree-Based Methods," en *An Introduction to Statistical Learning*, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani y Jonathan Taylor, Eds., Cham, Suiza: Springer Cham, 2023, cap. 8, pp. 331–366. [En línea]. Disponible: https://hastie.su.domains/ISLP/ISLP_website.pdf.download.html.

[42] S. Gupta, D. Aga, A. Pruden, L. Zhang y P. Vikesland, "Data Analytics for Environmental Science and Engineering Research," *Environ.Sci.Technol.*, vol. 55, no. 16, pp. 10895–10907, 2021, doi: 10.1021/acs.est.1c01026.

[43] L. Hernández-Alpízar, J.A. Gómez-Mejía y M.B. Argüello-Vega, "Inteligencia artificial, machine learning y SIG en ingeniería ambiental: tendencias actuales," *TM*, vol. 37, no. 7, pp. 87–96, 2024, doi: 10.18845/tm.v37i7.7304.

[44] S. Zhong, et al., "Machine Learning: New Ideas and Tools in Environmental Science and Engineering," *Environ.Sci.Technol.*, vol. 55, no. 19, pp. 12741–12754, 2021, doi: 10.1021/acs.est.1c01339.

[45] M. He, Q. Qian, X. Liu, J. Zhang y J. Curry, "Recent Progress on Surface Water Quality Models Utilizing Machine Learning Techniques," *Water*, vol. 16, no. 24, 2024, doi: 10.3390/w16243616.

[46] B.Q. Lap, et al., "Predicting Water Quality Index (WQI) by feature selection and machine learning: A case study of An Kim Hai irrigation system," *Ecol. Inf.*, vol. 74, pp. 101991, 2023, doi: 10.1016/j.ecoinf.2023.101991.

[47] S. Ramya, S. Srinath y P. Tuppad, "Comprehensive analysis of multiple classifiers for enhanced river water quality monitoring with explainable AI," *Case Stud. Chem. Environ. Eng.*, vol. 10, pp. 100822, 2024, doi: 10.1016/j.cscee.2024.100822.

[48] S.B.H.S. Asadollah, A. Sharafati, D. Motta y Z.M. Yaseen, "River water quality index prediction and uncertainty analysis: A comparative study of machine learning models," *J. Environ. Chem. Eng.*, vol. 9, no. 1, pp. 104599, 2021, doi: 10.1016/j.jece.2020.104599.

[49] M.G. Uddin, S. Nash, A. Rahman y A.I. Olbert, "Performance analysis of the water quality index model for predicting water state using machine learning techniques," *Process Saf.Environ.Prot.*, vol. 169, pp. 808–828, 2023, doi: 10.1016/j.psep.2022.11.073.

[50] H. Visser, et al., "What drives the ecological quality of surface waters? A review of 11 predictive modeling tools," *Water Res.*, vol. 208, pp. 117851, 2022, doi: 10.1016/j.watres.2021.117851.

[51] O. Rahmati, M. Avand, P. Yariyan, J.P. Tiefenbacher, A. Azareh y D.T. and Bui, "Assessment of Gini-, entropy- and ratio-based classification trees for groundwater potential modelling and prediction," *Geocarto Int.*, vol. 37, no. 12, pp. 3397–3415, 2022, doi: 10.1080/10106049.2020.1861664.

[52] Y. Lu, T. YeyJ. Zheng, "Decision Tree Algorithm in Machine Learning," presentado en *2022 IEEE Int. Conf. Advances Elect. Eng. Comput. Appl.*, pp. 1014–1017, 2022, doi: 10.1109/AEECA55500.2022.9918857.

[53] Z. Zhou y G. Hooker, "Unbiased Measurement of Feature Importance in Tree-Based Methods," *ACM Trans.Knowl.Discov.Data*, vol. 15, no. 2, 2021, doi: 10.1145/3429445.

[54] A. Elmotawakkil, N. Enneya, S.K. Bhagat, M.M. Ouda y V. Kumar, "Advanced machine learning models for robust prediction of water quality index and classification," *J. Hydroinf.*, vol. 27, no. 2, pp. 299–319, 2025, doi: 10.2166/hydro.2025.290.

[55] N.G. Rezk, S. Alshathri, A. Sayed y E. El-Din Hemdan, "EWAIS: An Ensemble Learning and Explainable AI Approach for Water Quality Classification Toward IoT-Enabled Systems," *Processes*, vol. 12, no. 12, 2024, doi: 10.3390/pr12122771.

[56] S. Lee, C. Lee, K. G. Mun y D. Kim, "Decision Tree Algorithm Considering Distances Between Classes," *IEEE Access*, vol. 10, pp. 69750–69756, 2022, doi: 10.1109/ACCESS.2022.3187172.

[57] R.K. Makumbura, et al., "Advancing water quality assessment and prediction using machine learning models, coupled with explainable artificial intelligence (XAI) techniques like shapley additive explanations (SHAP) for interpreting the black-box nature," *Results Eng.*, vol. 23, pp. 102831, 2024, doi: 10.1016/j.rineng.2024.102831.

[58] M. Niazkar, et al., "Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023)," *Env. Model. Software*, vol. 174, pp. 105971, 2024, doi: 10.1016/j.envsoft.2024.105971.

[59] T. Chen y C. Guestrin, "XGBoost: A Scalable Tree Boosting System," presentado en *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[60] J. Tian, Y. Jiang, J. Zhang, Z. Wang, J.J. Rodríguez-Andina y H. Luo, "High-Performance Fault Classification Based on Feature Importance Ranking-XgBoost Approach with Feature Selection of Redundant Sensor Data," *Curr. Chin. Sci.*, vol. 2, no. 3, pp. 243–251, 2022, doi: 10.2174/2210298102666220318100051.

[61] P. Devan y N. Khare, "An efficient XGBoost–DNN-based classification model for network intrusion detection system," *Neural Comput. Appl.*, vol. 32, no. 16, pp. 12499–12514, 2020, doi: 10.1007/s00521-020-04708-x.

[62] G. James, D. Witten, T. Hastie, R. Tibshirani y J. Taylor, "Statistical Learning," en *An Introduction to Statistical Learning*, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani y Jonathan Taylor, Eds., Cham, Suiza: Springer Cham, 2023, cap. 2, pp. 15–68. [En línea]. Disponible: https://hastie.su.domains/ISLP/ISLP_website.pdf.download.html.

[63] S.M. Lundberg y S. Lee, "A unified approach to interpreting model predictions," presentado en *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–4777, 2017.

[64] C. Liu, et al., "DDWQI: A novel water quality index based on data-driven approaches," *Ecol. Ind.*, vol. 178, pp. 113850, 2025, doi: 10.1016/j.ecolind.2025.113850.

[65] G. James, D. Witten, T. Hastie, R. Tibshirani y J. Taylor, "Resampling Methods," en *An Introduction to Statistical Learning*, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani y Jonathan Taylor, Eds., Cham, Suiza: Springer Cham, 2023, cap. 5, pp. 201–228. [En línea]. Disponible: https://hastie.su.domains/ISLP/ISLP_website.pdf.download.html.

[66] M. Grandini, E. Bagli y G. Visani, "Metrics for multi-class classification: an overview," *arXiv*, 2020, doi: 10.48550/arXiv.2008.05756.

[67] P.N. Bhowmik, K. Saini, N.T. Sai Priya, P. Anand y B. Ateş, "A Scalable Machine Learning Framework for Hydrological Water Quality Monitoring Using Physicochemical and Microbial Parameters," *Water*, vol. 17, no. 14, pp. 2158, 2025, doi: 10.3390/w17142158.

[68] J.P. Nair y M.S. Vijaya, "Exploratory Data Analysis of Bhavani River Water Quality Index Data," en *Proceedings of International Conference on Communication and Computational Technologies*, Sandeep Kumar, Saroj Hiranwal, S. D. Purohit y Mukesh Prasad, Eds., Singapore: Springer Nature, 2023, cap. 74, pp. 971–987. [En línea]. Disponible: https://link.springer.com/chapter/10.1007/978-981-19-3951-8_74.

[69] L.H. Alpízar y J.A. Gómez Mejía, "Machine Learning for Effluent Quality Classification in Wastewater Treatment," presentado en *2025 IEEE Colombian Conf. Appl. Comput. Intell.*, pp. 1–5, August 25-27, 2025, doi: 10.1109/ColCACI67437.2025.11230782.

[70] H. Kim, "Statistical notes for clinical researchers: Nonparametric statistical methods: 2. Nonparametric methods for comparing three or more groups and repeated measures," *Restor Dent Endod*, vol. 39, no. 4, 2014, doi: 10.5395/rde.2014.39.4.329.

[71] A. Benavoli, G. Corani y F. Mangili, "Should we really use post-hoc tests based on mean-ranks?" *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 152–161, 2016, doi: 10.48550/arXiv.1505.02288.

[72] D. Mora-Alvarado y C.F. Portuguez-Barquero, "Agua para uso y consumo humano, saneamiento e higiene de Costa Rica en el contexto de américa latina al 2022," *TM*, vol. 37, no. 8, pp. 36–49, 2024, doi: 10.18845/tm.v37i8.6854.

[73] K. Arce-Villalobos, R. Sánchez-Gutiérrez, J. Centeno-Morales, R. Marín-León y J.A. Rodríguez-Rodríguez, "Calidad del agua superficial y presiones socioambientales en la microcuenca alta del río Poás," *Uniciencia*, vol. 36, no. 1, pp. 1–23, 2022, doi: 10.15359/ru.36-1.24.

[74] R.d.l.Á Camacho-Jiménez, "Plan de gestión comunitaria para la microcuenca del Río Toyogres: estrategias para la reducción de la contaminación en ríos de Oreamuno y Cartago, Costa Rica," Tesis de licenciatura, Ins. Tec. de Costa Rica, 2024. [En línea]. Disponible: https://repositoriotec.tec.ac.cr/handle/2238/15677.

[75] K.D. Salazar Corrales, "Evaluación fisicoquímica y microbiológica como insumo para el mejoramiento de la gestión socioambiental del recurso hídrico en la parte media y baja de la Cuenca del Río Tempisque, Guanacaste, Costa Rica," Tesis de licenciatura, Univ. Nac. de Costa Rica, 2022. [En línea]. Disponible: http://hdl.handle.net/11056/23210.

[76] K. Navarro Salas y Y. Monge Fernández, "Evaluación de la influencia de las actividades socioeconómicas en el caudal, calidad del agua y comunidades de macroinvertebrados bentónicos en el río Maravilla, Cartago, Costa Rica," Tesis de licenciatura, Univ. Nac. de Costa Rica, 2021. [En línea]. Disponible: http://hdl.handle.net/11056/18896.

[77] M. Devlin y J. Brodie, "Nutrients and Eutrophication," en *Marine Pollution – Monitoring, Management and Mitigation*, Amanda Reichelt-Brushett, Eds., Cham: Springer Nature Switzerland, 2023, cap. 4, pp. 75–100. [En línea]. Disponible: https://doi.org/10.1007/978-3-031-10127-4_4.

[78] J. Wang, T. Huang, Q. Wu, C. Bu y X. Yin, "Sources and Cycling of Phosphorus in the Sediment of Rivers along a Eutrophic Lake in China Indicated by Phosphate Oxygen Isotopes," *ACS Earth Space Chem.*, vol. 5, no. 1, pp. 88–94, 2021, doi: 10.1021/acsearthspacechem.0c00298.

[79] G. He, Q. Lao, G. Jin, Q. Zhu y F. Chen, "Increasing eutrophication driven by the increase of phosphate discharge in a subtropical bay in the past 30 years," *Front. Mar. Sci.*, vol. 10, 2023, doi: 10.3389/fmars.2023.1184421.

[80] H. Madrigal-Solís, et al., "What do we Think About Water? Public Perception of the Current Situation of Water Resources in Costa Rica: an Indicator of Water Understanding and Management," *Uniciencia*, vol. 34, no. 1, pp. 159–188, 2020, doi: 10.15359/ru.34-1.10.

[81] K. Chacón Araya y S. González Rosales, *Agricultura: Impactos Y Desafíos Para La Seguridad Alimentaria Y La Sostenibilidad Ambiental En Costa Rica,* San José, Costa Rica: CONARE-PEN, 2023. [En línea]. Disponible: https://hdl.handle.net/20.500.12337/10155.

[82] M.F. Vargas González, *Tendencias Y Desafíos De La Gestión De Los Recursos Hídricos Para El Ambiente Y El Desarrollo Humano Sostenible En Costa Rica,* San José, Costa Rica: CONARE-PEN, 2025. [En línea]. Disponible: https://hdl.handle.net/20.500.12337/10971.

[83] N. Gil-Rodas, et al., "A comparative study of several types of indices for river quality assessment," *Water Qual. Res. J.*, vol. 58, no. 3, pp. 169–183, 2023, doi: 10.2166/wqrj.2023.029.

[84] I. Meghea, "Statistical Methods and Models for Pollutant Control in Municipal Surface Waters," *Water*, vol. 15, no. 23, pp. 4178, 2023, doi: 10.3390/w15234178.

[85] Y. Chen, et al., "Phosphorus – The main limiting factor in riverine ecosystems in China," *Sci. Total Environ.*, vol. 870, pp. 161613, 2023, doi: 10.1016/j.scitotenv.2023.161613.

[86] P. Gogoi, et al., "An integrative water quality index and multivariate modeling approach to assess surface water quality, trophic status and nutrient source apportionment in a large tropical reservoir, Hirakud–the longest earthen dam in Asia," *Appl. Water Sci.*, vol. 15, no. 8, pp. 219, 2025, doi: 10.1007/s13201-025-02517-y.

[87] C. Noriega, H. Varona, C. Medeiros, A. Hounsou-Gbo, J. Araujo y M. Araujo, "Carbon, Nitrogen, and Phosphorus Fluxes in Sixty Tropical Brazilian Rivers: Current Status, Stoichiometry and Trends," *Water Air Soil Pollut.*, vol. 235, no. 7, pp. 465, 2024, doi: 10.1007/s11270-024-07271-6.

[88] J.D. Bolaños-Alfaro, G. Cordero-Castro and G. Segura-Araya, "'Determinación de nitritos, nitratos, sulfatos y fosfatos en agua potable como indicadores de contaminación ocasionada por el hombre, en dos cantones de Alajuela (Costa Rica)," *TM*, vol. 30, no. 4, -10-01, pp. 15–2710.18845/tm.v30i4.3408.

[89] L.L. Freire, A.C. Costa y I.E.L. Neto, "Effects of rainfall and land use on nutrient responses in rivers in the Brazilian semiarid region," *Environ. Monit. Assess.*, vol. 195, no. 6, pp. 652, 2023, doi: 10.1007/s10661-023-11281-y.

[90] H. Mahadevan, K.A. Krishnan, R.R. Pillai y S. Sudhakaran, "Assessment of urban river water quality and developing strategies for phosphate removal from water and wastewaters:

Integrated monitoring and mitigation studies," *SN Appl. Sci.*, vol. 2, no. 4, pp. 772, 2020, doi: 10.1007/s42452-020-2571-0.

[91] H. Chang, Y. Makido y E. Foster, "Effects of land use change, wetland fragmentation, and best management practices on total suspended solids concentrations in an urbanizing Oregon watershed, USA," *J. Environ. Manage.*, vol. 282, pp. 111962, 2021, doi: 10.1016/j.jenvman.2021.111962.

[92] Consejo Nacional de Rectores y Programa Estado de la Nación, "Armonía con la naturaleza," en *Estado de la Nación 2024*, Karen Chacón Araya y Leonardo Merino Trejos, Eds., San José, Costa Rica: CONARE-PEN, 2024, cap. 4, pp. 165–228.

[93] G. von Rückert, C.C. Figueredo, J.E.d.F. Barros y R.R.d. Silva, "Water quality and land use in Ipanema Stream Watershed (Doce River Basin/Brazil): effects of urbanization," *RBRH*, vol. 29, pp. e25, 2024, doi: https://doi.org/10.1590/2318-0331.292420230142.

[94] L.M. Soto-Castro, "Análisis del nivel de contaminación en la cuenca media alta del río Guápiles, Pococí," Tesis de licenciatura, Inst. Tec. de Costa Rica, 2019. [En línea]. Disponible: https://repositoriotec.tec.ac.cr/handle/2238/10702.

[95] C.N. Tran, C. Yossapol, N. Tantemsapya, P. Kosa y P. Kongkhiaw, "Water Quality Simulation and Dissolved Oxygen Change Scenarios in Lam Takhong River in Thailand," *J. Sustainable Dev. Energy Water Environ. Syst.*, vol. 10, no. 1, pp. 1–13, 2022, doi: 10.13044/j.sdewes.d9.0389.

[96] G. Calvo-Brenes y K. Salazar-Céspedes, "Estrategia de monitoreo hídrico comunitario para la microcuenca río Jorco basado en el análisis de indicadores fisicoquímicos, microbiológicos y biológicos de la calidad de agua," *TM*, pp. 181–193, 2023, doi: 10.18845/tm.v36i4.6456.

[97] S. Kumi, D. Adu-Poku y F. Attiogbe, "Dynamics of land cover changes and condition of soil and surface water quality in a Mining–Altered landscape, Ghana," *Heliyon*, vol. 9, no. 7, 2023, doi: 10.1016/j.heliyon.2023.e17859.

[98] F. Lemessa, B. Simane, A. Seyoum y G. Gebresenbet, "Assessment of the Impact of Industrial Wastewater on the Water Quality of Rivers around the Bole Lemi Industrial Park (BLIP), Ethiopia," *Sustainability*, vol. 15, no. 5, pp. 4290, 2023, doi: 10.3390/su15054290.

[99] B.M. Saalidong, S.A. Aram, S. Otu y P.O. Lartey, "Examining the dynamics of the relationship between water pH and other water quality parameters in ground and surface water systems," *PLos ONE*, vol. 17, no. 1, pp. e0262117, 2022, doi: 10.1371/journal.pone.0262117.

APÉNDICE 1. CÓDIGO DE PYTHON PARA MODELAR CON LOS ALGORITMOS

```
# Importamos las librerías necesarias
import numpy as np
import pandas as pd
from sklearn.model_selection import StratifiedKFold, cross_validate
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
from sklearn.metrics import make_scorer, precision_score, recall_score, f1_score

# Cargar datos
# df es el dataframe con los datos
X = df.drop(columns=['CLASE', 'CAT', 'ICA']) # Variables excluidas
y = df['CAT'] # Categorías binarias de calidad

# Definir los modelos
# Se definió la semilla como 42 para asegurar la repetibilidad de los resultados
models = {
    'Decision Tree': DecisionTreeClassifier(random_state=42),
    'Random Forest': RandomForestClassifier(random_state=42),
    'XGBoost': XGBClassifier(random_state=42)
}

# Definir las métricas con make_scorer
# Se utiliza 0 como categoría positiva
scoring = {
    'accuracy': 'accuracy',
```

```
    'precision': make_scorer(precision_score, pos_label=0, average='binary'),
    'recall': make_scorer(recall_score, pos_label=0, average='binary'),
    'f1': make_scorer(f1_score, pos_label=0, average='binary')
}

# Número de folds e iteraciones para tener 100 resultados por métrica por modelo
n_splits = 10
n_iterations = 10

# Diccionarios para almacenar los resultados de las métricas por modelo
results = {name: {metric: [] for metric in scoring.keys()} for name in models.keys()}
recall_scores = {name: [] for name in models.keys()} # Para almacenar solo los resultados de recall

# Para verificar que el código va corriendo bien
print(f"Realizando validación cruzada estratificada ({n_splits}-fold) con {n_iterations} iteraciones...")

for iteration in range(n_iterations):
    print(f"\nIteración {iteration + 1}/{n_iterations}")
    # Crear un nuevo StratifiedKFold con un random_state diferente para cada iteración
    skf = StratifiedKFold(n_splits=n_splits, shuffle=True, random_state=42 + iteration)

    for name, model in models.items():
        cv_results = cross_validate(model, X, y, cv=skf, scoring=scoring)

        for metric in scoring.keys():
            results[name][metric].extend(cv_results[f'test_{metric}'])
```

```
        # Almacenar los puntajes de recall para las pruebas estadisticos
        recall_scores[name].extend(cv_results['test_recall'])

# Mostrar resultados promedio y desviación estándar
print("\n-- Resultados promedio con desviación estándar --")
for name, metric_results in results.items():
    print(f"\n-- {name} --")
    for metric, scores in metric_results.items():
        mean_score = np.mean(scores)
        std_score = np.std(scores)
        print(f"{metric.capitalize()}: Mean = {mean_score:.4f}, Std Dev = {std_score:.4f}")

# Crear un dataframe con los resultados para el análisis estadístico
df_results = pd.DataFrame(recall_scores)
df_results['Subject'] = range(len(df_results))

# Las pruebas estadísticas pueden realizarse aquí en Python o exportar el df y hacerlas en R
```