PREDICCIÓN DEL PH Y EL DBO EN EL RÍO CAUCA UTILIZANDO TÉCNICAS DE MACHINE LEARNING: ANTICIPANDO CAMBIOS CRÍTICOS DE CALIDAD

DANIEL SOTELO DE LA CRUZ. CÓDIGO: 9019380
WEIMAR CORTES MONTIEL. CÓDIGO: 9014286

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director(a)
CRISTIAN ALEJANDRO TORRES VALENCIA

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JUNIO DE 2026

FICHA RESUMEN

TÍTULO DEL PROYECTO: PREDICCIÓN DEL PH Y EL DBO EN EL RÍO CAUCA UTILIZANDO TÉCNICAS DE MACHINE LEARNING: ANTICIPANDO CAMBIOS CRÍTICOS DE CALIDAD

1.	ÉNFASIS	Sistemas y Computación
2.	TIPO DE PROYECTO	Proyecto aplicado
3.	ÁREA DE TRABAJO:	Ciencia de datos
4.	ESTUDIANTE(S):	Daniel Sotelo de la Cruz Weimar Cortes Montiel
5.	CORREO ELECTRÓNICO:	dansotelo@javerianacali.edu.co weimarcortes@javerianacali.edu.co
6.	DIRECCIÓN Y TELÉFONO:	Mz B Casa 141 Barrio Santa Mónica (Pasto, Nariño), 3142771557 Calle 8ª #9c-47 Barrio Chipre (Manizales, Caldas), 3154003627
7.	DIRECTOR:	Cristian Alejandro Torres Valencia
8.	VINCULACIÓN DEL DIRECTOR:	Docente de planta de la PUJC, Departamento de electrónica y ciencias de la computación
9.	CORREO ELECTRÓNICO DEL DIRECTOR:	cristian.torres@javerianacali.edu.co
10.	PALABRAS CLAVE (al menos 5):	pH, DBO (Demanda Bioquímica de Oxígeno), Río Cauca, Calidad del Agua, Machine Learning, Modelado Predictivo, Ciencia de Datos
11.	ODS QUE APLICA EL PROYECTO (Agenda 2030):	6. Agua limpia y saneamiento
12.	FECHA DE INICIO:	1 de septiembre 2025
13.	RESUMEN:	Dada la creciente preocupación por la calidad del agua en Colombia, este proyecto de maestría desarrolló modelos predictivos para estimar los niveles de pH y Demanda Bioquímica de Oxígeno (DBO) en el río Cauca, con énfasis en la cuenca alta, integrando técnicas de ciencia de datos y aprendizaje automático. El estudio partió del análisis y preprocesamiento de datos históricos oficiales, que incluyeron variables fisicoquímicas como pH, DBO, oxígeno disuelto, temperatura, conductividad y nutrientes. La metodología contempló la partición de los datos en conjuntos de entrenamiento y prueba (80-20 %), con validación cruzada repetida para garantizar la robustez de la evaluación. Se entrenaron y ajustaron modelos de regresión como Regresión Lineal Múltiple, Árboles de Decisión, Random Forest, Gradient Boosting (XGBoost), SVR y Perceptrón Multicapa (MLP), optimizados mediante Grid Search y Optuna. Su desempeño se evaluó con métricas como RMSE, MAE y R^2, así como mediante curvas de aprendizaje para detectar sobreajuste. Se estimaron con precisión los valores de pH y DBO del río Cauca, y mediante análisis SHAP se identificaron las variables

fisicoquímicas de mayor incidencia: para la DBO, los sulfatos (mg SO_4/l), bicarbonatos (mg $CaCO_3$/l) y conductividad eléctrica (µS/cm) resultaron las más determinantes; para el pH, los nitratos (mg N-NO_3/l) y bicarbonatos (mg $CaCO_3$/l) concentraron la mayor contribución explicativa. El trabajo se centró en la ejecución técnica de los modelos, sin desarrollos externos ni visualizaciones interactivas. Se priorizó el análisis cuantitativo de los resultados, documentando el proceso completo y aportando evidencia sobre la aplicabilidad del aprendizaje automático en el monitoreo ambiental. Esta investigación fortaleció herramientas analíticas para la gestión del recurso hídrico local y ofrece una base metodológica replicable, útil como insumo académico o institucional, sin pretensión de implementación automatizada o en tiempo real.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE IMÁGENES

LISTA DE ANEXOS[1]

[1] Dada la extensión de los reportes técnicos y el código fuente, los anexos se presentan como documentos complementarios independientes al cuerpo del trabajo, identificados según la numeración de esta lista.

1. INTRODUCCIÓN

La gestión sostenible del recurso hídrico constituye uno de los desafíos ambientales más urgentes del siglo XXI. En Colombia, la presión creciente de las actividades agrícolas, industriales y urbanas ha generado un deterioro progresivo en la calidad del agua de los principales sistemas fluviales, entre ellos el río Cauca, cuya relevancia ecológica, económica y social lo convierte en un eje vital para la región andina. El aumento de las descargas contaminantes, junto con la variabilidad hidrológica y climática, ha modificado parámetros fisicoquímicos fundamentales como el potencial de hidrógeno (pH) y la Demanda Bioquímica de Oxígeno (DBO), indicadores esenciales del equilibrio biogeoquímico del ecosistema acuático.

Pese a los esfuerzos institucionales, las metodologías tradicionales de monitoreo ambiental presentan limitaciones en cobertura temporal y espacial, lo que dificulta la detección temprana de anomalías y la predicción de eventos críticos de contaminación. En este contexto, la incorporación de herramientas de Ciencia de Datos e Inteligencia Artificial (IA) se ha consolidado como una alternativa prometedora para fortalecer la gestión ambiental. Estas técnicas permiten procesar grandes volúmenes de información, descubrir patrones ocultos en los datos y desarrollar modelos predictivos capaces de anticipar comportamientos complejos del sistema hídrico.

A partir de esta necesidad, el presente estudio se orientó a responder la pregunta de cómo predecir eficazmente los niveles de pH y DBO en el río Cauca mediante el análisis de registros y variables fisicoquímicas asociadas. Para ello, se implementó una metodología integral de Ciencia de Datos, que comprendió la preparación y depuración de un conjunto de datos multianual, la identificación de variables significativas mediante indicadores de colinealidad (VIF) y Eliminación Recursiva de Características (RFE), y la evaluación comparativa de un portafolio de algoritmos de regresión.

Entre los modelos desarrollados se incluyeron la Regresión Lineal Múltiple (RLM), empleada como referencia base para contrastar el comportamiento de los métodos no lineales, junto con árboles de decisión (Decision Tree Regressor), bosques aleatorios (Random Forest), métodos de ensamble como Gradient Boosting y XGBoost, redes neuronales tipo Perceptrón Multicapa (MLP) y Máquinas de Vectores de Soporte (SVR). Asimismo, se evaluaron variantes optimizadas mediante selección de características (RFE, VIF y Unión RFE), lo que permitió examinar el efecto de la reducción dimensional sobre el rendimiento predictivo. Todos los modelos fueron calibrados y comparados con base en métricas de desempeño (R^2, RMSE y MAE), utilizando técnicas avanzadas de optimización de hiperparámetros como Grid Search y Optuna para garantizar su robustez, estabilidad y precisión.

Los resultados obtenidos demostraron la viabilidad y pertinencia del uso del Machine Learning en el monitoreo ambiental. Para la variable DBO, el modelo Random Forest optimizado alcanzó un coeficiente de determinación R^2 del 91%, con un error cuadrático medio (RMSE) de 3.21 mg O_2/L, evidenciando una notable capacidad para reproducir los patrones de contaminación orgánica. Las

variables con mayor influencia fueron los sulfatos, bicarbonatos y la conductividad eléctrica, lo que refuerza la relación entre la mineralización del agua y los procesos de degradación orgánica. En contraste, la predicción del pH presentó una menor varianza explicada (R^2 = 0.40), coherente con su estabilidad fisicoquímica; sin embargo, el modelo XGBoost optimizado logró un RMSE de 0.30 unidades, identificando a los bicarbonatos y nitratos como factores determinantes del equilibrio ácido-base del sistema.

En conjunto, el estudio evidencia el potencial de la Ciencia de Datos como herramienta de apoyo a la toma de decisiones ambientales, permitiendo anticipar comportamientos críticos del río Cauca y fortalecer los programas de monitoreo y gestión del recurso hídrico. Más allá de los resultados numéricos, el proyecto ofrece un marco analítico para comprender las interacciones entre los procesos naturales y las presiones antrópicas, y demuestra que el aprendizaje automático puede integrarse de manera efectiva en la planificación ambiental a escala regional.

El documento se estructura de la siguiente manera: el Capítulo 2 presenta la contextualización del proyecto, el planteamiento del problema, los objetivos y los fundamentos teóricos del Machine Learning aplicados al modelado de la calidad del agua. El Capítulo 3 describe el proceso de preparación y depuración del conjunto de datos histórico; el Capítulo 4 aborda el análisis exploratorio y la selección de variables; el Capítulo 5 expone el entrenamiento y optimización de los modelos; y el Capítulo 6 detalla la evaluación, comparación e interpretabilidad de los resultados obtenidos. Finalmente, el Capítulo 7 recoge las conclusiones principales y propone líneas futuras de investigación orientadas a mejorar la predicción y gestión sostenible del recurso hídrico en sistemas fluviales complejos.

2. CONTEXTUALIZACIÓN DEL PROYECTO

2.1. DEFINICIÓN DEL PROBLEMA

La limitada capacidad de los sistemas de monitoreo tradicionales para anticipar las variaciones en la calidad del agua del Río Cauca plantea la necesidad de desarrollar métodos predictivos más eficientes. Ante esta situación, el problema central de esta investigación se centró en responder a la pregunta: ¿Cómo predecir eficazmente los niveles de pH y Demanda Bioquímica de Oxígeno (DBO) a partir del análisis de datos y variables fisicoquímicas, utilizando para ello modelos de Machine Learning?

Resolver este desafío implicó un proceso metodológico riguroso que abarcó desde la preparación de los conjuntos de datos y la identificación de las variables fisicoquímicas más influyentes, hasta el entrenamiento y la evaluación comparativa de diversos modelos de regresión para asegurar la fiabilidad de las predicciones.

2.1.1. PLANTEAMIENTO DEL PROBLEMA

La calidad del agua del Río Cauca presenta un deterioro progresivo debido al aumento de la presión ambiental asociada a la actividad agrícola, industrial y urbana en los departamentos del Cauca, Valle del Cauca y Risaralda, pertenecientes a la cuenca alta del río. Esta degradación se manifiesta en la alteración de parámetros críticos como el pH y la Demanda Bioquímica de Oxígeno (DBO), que son indicadores fundamentales de la calidad del agua y reflejan directamente el impacto de las actividades humanas en el ecosistema fluvial [1].

Sin embargo, el monitoreo tradicional de estos indicadores suele ser limitado en cobertura temporal y espacial, lo que dificulta una comprensión anticipada de los cambios en la calidad del agua. Por un lado, el pH es especialmente sensible a las descargas industriales, las escorrentías agrícolas cargadas de fertilizantes y pesticidas, y los vertimientos de aguas residuales sin tratamiento adecuado, que pueden modificar drásticamente las condiciones químicas del agua. Por otra parte, la DBO es un indicador de la cantidad de materia orgánica presente en el agua, cuya descomposición requiere grandes cantidades de oxígeno, afectando negativamente la vida acuática y los procesos de autodepuración del río [2].

En este contexto, el uso de modelos de Machine Learning representa una alternativa prometedora para aprovechar datos históricos y realizar predicciones más eficientes y automatizadas con base en parámetros que muestran variaciones significativas a lo largo del tiempo, reflejando el impacto continuo de fuentes de contaminación puntuales y difusas en diferentes tramos del río.

No obstante, existe la necesidad de predecir de manera confiable los valores de pH y DBO en distintos tramos del río, utilizando modelos que permitan anticipar los efectos de la contaminación sobre la calidad del agua. Anticipar sus alteraciones mediante modelos predictivos basados en el aprendizaje automático permite detectar patrones de contaminación y apoyar la toma de decisiones; en este sentido resulta clave construir un modelo que, a partir del análisis de

variables fisicoquímicas, logren estimar con precisión los valores PH y DBO. [3].

2.1.2. FORMULACIÓN DEL PROBLEMA

¿Cómo predecir los niveles de pH y Demanda Bioquímica de Oxígeno (DBO) en el Río Cauca, a partir del análisis de datos y variables fisicoquímicas utilizando modelos de Machine Learning?

2.1.3. SISTEMATIZACIÓN

- ¿Cómo preparar los conjuntos de datos de calidad del agua del Río Cauca para el análisis predictivo?

- ¿Cuáles son las variables fisicoquímicas que influyen en los niveles de pH y DBO en el río Cauca?

- ¿Cómo entrenar diferentes modelos de regresión para la predicción de los niveles de pH y DBO en el río Cauca?

- ¿Cómo evaluar los modelos de regresión para la predicción de pH y DBO en el Río Cauca?

2.2. OBJETIVOS

2.2.1. OBJETIVO GENERAL

Predecir los niveles de pH y Demanda Bioquímica de Oxígeno (DBO) en el Río Cauca, a partir del análisis de datos y variables fisicoquímicas utilizando modelos de Machine Learning.

2.2.2. OBJETIVOS ESPECÍFICOS

- Preparar los conjuntos de datos de calidad del agua del Río Cauca para el análisis predictivo.

- Identificar las variables fisicoquímicas significativas que influyen en los niveles de pH y DBO en el Río Cauca.

- Entrenar diferentes modelos de regresión para la predicción de los niveles de pH y DBO en el Río Cauca, a partir de datos de variables fisicoquímicas.

- Evaluar los modelos de regresión para la predicción de pH y DBO en el Río Cauca, mediante métricas de regresión.

2.3. MARCO DE REFERENCIA

En esta sección se presentan los fundamentos teóricos y contextuales que sustentan la investigación. Se inicia con el Marco Teórico, donde se detalla la base en Ciencia de Datos y Aprendizaje Automático que guía el modelado predictivo. Posteriormente, en la sección de Antecedentes, se contextualiza la problemática de la calidad del agua, se definen sus indicadores clave y se exponen las limitaciones de los enfoques tradicionales.

2.3.1. MARCO TEÓRICO

RECURSOS HÍDRICOS

Los recursos hídricos constituyen un pilar fundamental para el desarrollo sostenible, el bienestar social y la preservación de la vida en el planeta. Dentro de estos, los ecosistemas fluviales, como los ríos, no solo son fuentes primordiales de agua dulce para consumo humano, agricultura e industria, sino que también albergan una considerable biodiversidad y cumplen funciones ecológicas vitales [4]. La integridad de estos ecosistemas es, por lo tanto, crucial.

Sin embargo, estos valiosos recursos se encuentran bajo una presión creciente debido a diversas actividades antrópicas, siendo la contaminación hídrica una de las amenazas más significativas a nivel global [5]. La gestión ambiental de las cuencas hidrográficas se erige como un proceso indispensable, que busca equilibrar el desarrollo humano con la protección y conservación de los recursos naturales [6]. En este marco, el concepto de calidad del agua adquiere una relevancia central, definiéndose como la composición del agua en la medida en que esta se ve afectada por procesos naturales y por las actividades humanas, determinando su idoneidad para un uso específico [7].

En consecuencia, la conservación y gestión de los recursos acuáticos requiere herramientas eficaces para evaluar su estado y dinámica. El monitoreo ambiental y la interpretación de indicadores fisicoquímicos se convierten en elementos esenciales que permiten vincular la teoría ecológica con la acción práctica y la toma de decisiones informadas.

Para evaluar dicha calidad, el monitoreo, entendido como la recolección y análisis sistemático de datos, se convierte en una herramienta esencial que permite diagnosticar el estado de los cuerpos de agua [1]. La interpretación de estos datos se basa en indicadores fisicoquímicos que proporcionan información cuantitativa sobre las características del agua. Entre la multitud de variables que pueden ser monitoreadas, el potencial de hidrógeno (pH) y la Demanda Bioquímica de Oxígeno (DBO) se consideran ejes centrales de análisis. El pH es una medida de la acidez o alcalinidad del agua, crucial para la mayoría de los procesos químicos y biológicos [8], mientras que la DBO es uno de los indicadores más importantes de contaminación por materia orgánica biodegradable, reflejando la presión sobre el oxígeno disuelto del ecosistema [1]. Otros parámetros como la temperatura, la conductividad y los nutrientes son también fundamentales para contextualizar sus variaciones [2].

La inherente complejidad de los sistemas hídricos, con múltiples variables fisicoquímicas interactuando de manera no lineal, a menudo excede la capacidad de los enfoques analíticos tradicionales para generar predicciones robustas. Esta limitación subraya la necesidad de recurrir a metodologías avanzadas capaces de modelar estas interacciones y extraer patrones significativos a partir de grandes volúmenes de datos. Es en este contexto donde la Ciencia de Datos emerge como un campo que proporciona las herramientas para abordar tales desafíos [3].

La complejidad inherente de los procesos fisicoquímicos y la naturaleza no lineal de las

interacciones entre variables ambientales requieren el uso de modelos predictivos basados en técnicas de aprendizaje automático. Estos modelos permiten capturar patrones subyacentes en los datos, estimar valores faltantes y predecir comportamientos futuros con mayor precisión. En este contexto, resulta fundamental establecer un flujo de preprocesamiento y depuración que garantice la calidad estadística de la información antes del entrenamiento de los algoritmos.

FUNDAMENTOS TEÓRICOS DEL PREPROCESAMIENTO DE DATOS

La construcción de modelos predictivos robustos requiere un flujo riguroso de preparación y depuración de datos. En este proceso, la estandarización del lenguaje mediante un diccionario de variables constituye una práctica fundamental, ya que formaliza la definición semántica, las unidades de medida y los rangos de calidad de cada parámetro, garantizando la consistencia y reproducibilidad del análisis [3]. Asimismo, la aplicación de umbrales mínimos de completitud —como requerir al menos el 80% de registros válidos por variable— es reconocida en la literatura como un criterio de rigor estadístico para asegurar la representatividad de los predictores.

No obstante, en datos ambientales del mundo real, es frecuente que variables cruciales no alcancen este umbral, dada la presencia inevitable de valores ausentes. Así, en lugar de descartar información valiosa, se recurre a técnicas de imputación que preserven las relaciones subyacentes.

Entre estas técnicas, la imputación por K-Vecinos más Cercanos (KNN) constituye un enfoque multivariado ampliamente utilizado [9], [10], [11], . A diferencia de métodos univariados simples, este algoritmo reconstruye cada valor faltante a partir del contexto provisto por las observaciones más similares, preservando la estructura de correlación del conjunto de datos. Formalmente, para un registro i con el atributo x_{ij} faltante, se localiza el conjunto de vecinos $N_{\mathrm{k}}(i)$ sobre las variables observadas utilizando una métrica en el espacio estandarizado (z-scores), con distancia euclidiana ponderada:

$$d(x_i, x_\ell) = \sqrt{\sum_{m \in O_{i\ell}} w_m (x_{im} - x_{\ell m})^2}$$

donde $O_{i\ell}$ indica las variables simultáneamente observadas en i y ℓ, y w_m permite penalizar/ponderar predictores. El valor faltante se estima como un promedio ponderado por distancia de los vecinos:

$$\hat{x}_{ij} = \frac{\sum_{\ell \in N_k(i)} w_{i\ell} w_{\ell j}}{\sum_{\ell \in N_k(i)} w_{i\ell}}, w_{i\ell} = \frac{1}{(d(x_i, x_\ell) + \varepsilon)^p}$$

con $\varepsilon > 0$, para evitar divisiones por cero y p controlando el decrecimiento del peso (típicamente $p \in [1, 2]$) [10]. Esta formulación mantiene las dependencias multivariadas y reduce el sesgo frente a métodos univariados simples [12].

Más allá de los valores ausentes, los registros ambientales presentan otro desafío crítico: la

presencia de valores extremos u atípicos que pueden distorsionar significativamente los modelos predictivos. Para este segundo problema, la literatura reconoce dos transformaciones robustas de uso extendido: la transformación logarítmica $x' = log(x + c)$, aplicable cuando existe asimetría positiva, y la winsorización, técnica que acota los valores extremos sin eliminar observaciones [10]. Formalmente, si q_α y $q_{1-\alpha}$ son los cuantiles α y $1 - \alpha$, la versión winsorizada de una variable se define como:

$$x_i^{(win)} = min\{max(x_i, q_\alpha), q_{1-\alpha}\}$$

Con α usualmente entre 0.5% y 5% según la severidad de colas. A diferencia del trimming (que descarta datos), la winsorización preserva el tamaño muestral y el orden relativo en la zona central, reduce la varianza de estimadores al acotar la función de influencia de observaciones extremas y atenúa la dominancia de outliers en el aprendizaje, a costa de introducir un sesgo controlado en las colas [9], [10]. La combinación de estandarización, imputación KNN y tratamiento robusto de valores atípicos es ampliamente reconocida como un esquema que estabiliza las distribuciones y mejora la robustez estadística de los predictores previo al modelado [9].

Más allá de garantizar que los datos sean estadísticamente robustos mediante estandarización e imputación, existe una dimensión conceptual crítica: reconocer que no todas las variables disponibles contribuyen equitativamente a la predicción. La selección rigurosa de variables constituye una etapa crítica en el diseño de modelos predictivos, orientada a identificar el subconjunto de predictores más informativo para cada variable objetivo, reduciendo la dimensionalidad y mejorando la interpretabilidad del modelo. El análisis de correlación permite cuantificar la fuerza y dirección de la asociación entre predictores y variables objetivo mediante los coeficientes de Pearson —para relaciones lineales— y de Spearman —para relaciones monotónicas no necesariamente lineales—; este último resulta más robusto ante valores atípicos y, por tanto, frecuentemente preferido en datos ambientales complejos [12].

Sin embargo, cuando se evalúan múltiples predictores simultáneamente, surge una complicación: la multicolinealidad entre predictores puede desestabilizar los coeficientes del modelo y dificultar la interpretación de la importancia de cada variable. El Factor de Inflación de la Varianza (VIF) es la métrica estándar para su diagnóstico, cuantificando cuánto aumenta la varianza de un coeficiente estimado debido a su colinealidad con otros predictores. Su aplicación iterativa —eliminando sucesivamente las variables con mayor VIF hasta que todas se encuentren por debajo de un umbral predefinido— produce un conjunto de características más parsimonioso e interpretable [10], [12].

Complementariamente, la Eliminación Recursiva de Características (RFE, por sus siglas en inglés) es un método wrapper que evalúa subconjuntos de variables en función de su contribución al desempeño del modelo base. Propuesto originalmente por Guyon et al. en el contexto de SVM-RFE, el algoritmo entrena iterativamente sobre subconjuntos decrecientes de predictores, descartando en cada paso las variables de menor importancia según los pesos del modelo [13].

Su fortaleza radica en capturar interacciones multivariadas que los métodos de filtro no detectan, siendo especialmente valorado en conjuntos de datos de alta dimensionalidad [13].

APRENDIZAJE SUPERVISADO Y MODELOS DE REGRESIÓN

El aprendizaje automático ofrece un paradigma para construir modelos que aprenden patrones directamente de los datos, superando las limitaciones de los enfoques estadísticos tradicionales en la captura de interacciones no lineales. Dentro de este paradigma, el aprendizaje supervisado abarca los problemas en los que el modelo se entrena a partir de pares entrada-salida conocidos, con el objetivo de generalizar hacia nuevas observaciones. Cuando la variable de salida es continua —como los niveles de pH o DBO—, el problema se enmarca en la categoría de regresión supervisada [11], [14].

Un elemento clave en la implementación de modelos supervisados es el uso de *pipelines*, estructuras que encadenan secuencialmente las etapas de transformación de datos y entrenamiento del modelo. Su principal ventaja es la prevención de la fuga de datos (*data leakage*), garantizando que las transformaciones aplicadas al conjunto de entrenamiento —como el escalado— se apliquen de manera idéntica y sin contaminación sobre los datos de prueba y nuevos datos [11], [14]. Esto asegura la reproducibilidad y consistencia de los experimentos.

La literatura en modelado predictivo de calidad del agua reconoce un espectro de algoritmos de regresión, que va desde modelos lineales interpretables hasta ensambles de alta capacidad y redes neuronales. La evaluación comparativa de estos modelos es una práctica estándar, ya que permite determinar empíricamente cuál se adapta mejor a la complejidad inherente de los datos ambientales.

- **Regresión Lineal Múltiple:** es el modelo de referencia (*baseline*) por excelencia en tareas de regresión supervisada. Su formulación asume una relación lineal entre las variables predictoras y la variable objetivo, lo que la convierte en un estimador altamente interpretable. Aunque esta suposición puede resultar simplista frente a las interacciones no lineales propias de los sistemas ambientales, su desempeño establece el umbral mínimo que cualquier modelo más complejo debe superar para justificar su mayor costo computacional e interpretativo [15]. La Tabla 1 resume los parámetros clave que controlan su configuración en la biblioteca Scikit-Learn, junto con sus rangos típicos de uso.

Tabla 1. Parámetros clave de regresión lineal múltiple en Scikit-Learn

PARÁMETRO (SCIKIT-LEARN)	EXPLICACIÓN	RANGO TÍPICO (MIN–MAX)
fit_intercept	Incluye o no la constante β0	True / False
normalize (obsoleto)	Normaliza predictores	True / False

El conocimiento y correcto ajuste de los hiperparámetros es fundamental para obtener modelos efectivos, ya que una configuración inadecuada puede derivar en sobreajuste (*overfitting*) —cuando el modelo memoriza el ruido del conjunto de entrenamiento— o

subajuste (*underfitting*) —cuando no captura la estructura subyacente de los datos—. La optimización de hiperparámetros es, por tanto, una práctica transversal en el aprendizaje supervisado: independientemente del algoritmo empleado, el ajuste fino de estos parámetros permite maximizar la capacidad predictiva del modelo frente al problema específico que se aborda [14].

- **Árbol de Decisión:** es un modelo no paramétrico que representa el conocimiento en forma de reglas jerárquicas, dividiendo recursivamente el espacio de variables predictoras en regiones cada vez más homogéneas respecto a la variable objetivo. Su estructura arbórea lo convierte en uno de los modelos más interpretables del aprendizaje supervisado, ya que las reglas de decisión pueden visualizarse y seguirse de manera intuitiva [16]. No obstante, su principal limitación teórica es la alta tendencia al sobreajuste (*overfitting*): sin restricciones, el árbol puede crecer hasta capturar el ruido del conjunto de entrenamiento en lugar de la señal subyacente, comprometiendo su capacidad de generalización sobre datos no vistos [16], [17].

  Para controlar este comportamiento, la literatura identifica un conjunto de hiperparámetros estructurales que regulan la profundidad y granularidad del árbol. La profundidad máxima (*max_depth*), el número mínimo de muestras requeridas para dividir un nodo (*min_samples_split*) y el número mínimo de muestras en cada hoja (*min_samples_leaf*) actúan como mecanismos de regularización implícita, limitando la complejidad del modelo. El parámetro *max_features* —que restringe el número de variables evaluadas en cada división— introduce aleatoriedad controlada que favorece la generalización [17]. La Tabla 2 resume estos parámetros y sus rangos típicos de configuración.

Tabla 2. Parámetros clave del árbol de decisión

PARÁMETRO (SCIKIT-LEARN/XGBOOST)	EXPLICACIÓN	RANGO TÍPICO (MIN–MAX)
max_depth	Profundidad máxima del árbol	1 – 30 (o None)
min_samples_split	Mínimo de muestras para dividir nodo	2 – 20
min_samples_leaf	Mínimo de muestras en cada hoja	1 – 10
max_features	Nº de variables por división	1 – total, "sqrt", "log2"

  El balance entre profundidad y restricciones de hoja determina el equilibrio entre sesgo y varianza del modelo: árboles más profundos reducen el sesgo, pero incrementan la varianza, mientras que árboles más restringidos generalizan mejor a costa de mayor simplificación. Este principio, conocido en la literatura como el bias-variance tradeoff, es central en la teoría del aprendizaje estadístico y aplica de manera transversal a todos los modelos supervisados [16].

- **Random Forest:** es un método de ensamble diseñado para mitigar la tendencia al sobreajuste propia de los árboles de decisión individuales. Su principio teórico se fundamenta en la construcción de una multitud de árboles durante el entrenamiento, cada uno operando sobre

un subconjunto de datos muestreado con reemplazo (bootstrap) y utilizando subconjuntos aleatorios de características en cada división [18]. La predicción final se obtiene como el promedio de las predicciones de todos los árboles individuales, mecanismo que reduce la varianza del modelo sin incrementar proporcionalmente el sesgo [19]. Este proceso de promediado produce estimadores más robustos y generalizables, razón por la cual el *Random Forest* es reconocido como uno de los algoritmos de mayor desempeño en datos tabulares [19].

Desde el punto de vista teórico, el comportamiento del ensamble está determinado por dos grupos de hiperparámetros: los que controlan la estructura de cada árbol individual — profundidad máxima, mínimo de muestras por nodo y por hoja, y número de variables evaluadas por división— y los que gobiernan la diversidad del ensamble, principalmente el número de árboles (*n_estimators*) y la activación del muestreo *bootstrap*. La diversidad entre árboles es condición necesaria para que el promediado reduzca efectivamente la varianza: árboles muy similares entre sí aportan poca ganancia al ensamble [18]. La Tabla 3 resume estos hiperparámetros y sus rangos típicos en la literatura [18], [19].

Tabla 3. Parámetros clave de random forest

PARÁMETRO (SCIKIT-LEARN/XGBOOST)	EXPLICACIÓN	RANGO TÍPICO (MIN–MAX)
n_estimators	Nº de árboles	100 – 1000
max_depth	Profundidad máxima	5 – 50 (o None)
min_samples_split	Igual a árbol	2 – 20
min_samples_leaf	Igual a árbol	1 – 10
max_features	Nº de variables por división	"sqrt", "log2", 0.1–1.0
bootstrap	Muestreo con reemplazo	True / False

El equilibrio entre el número de árboles y las restricciones estructurales de cada uno refleja el bias-variance tradeoff a escala de ensamble: mayor número de árboles estabiliza las predicciones, mientras que restricciones de profundidad y muestreo controlan la complejidad individual. La activación del bootstrap incrementa la diversidad del ensamble al exponer cada árbol a una muestra distinta del espacio de datos, lo que se traduce en mayor capacidad de generalización frente a conjuntos de datos de tamaño limitado [18], [19].

- **Regresión de Vectores de Soporte (SVR):** se fundamenta en un principio distinto al de los modelos de regresión clásicos: en lugar de minimizar el error en todos los puntos de entrenamiento, busca encontrar una función que se ajuste a los datos dentro de un margen de tolerancia definido por el parámetro épsilon (ε). ignorando los errores cuya magnitud no supere dicho umbral. Este enfoque, conocido como pérdida ε-insensible, concentra la atención del modelo en los puntos que se encuentran en el borde o fuera del margen —

denominados vectores de soporte—, lo que confiere al modelo robustez ante valores atípicos y ruido en los datos [20].

Una característica teórica central del SVR es el uso de funciones *kernel*, que permiten mapear implícitamente los datos a un espacio de características de mayor dimensión sin necesidad de calcular explícitamente dicha transformación. La Función de Base Radial (RBF) es la más utilizada en la literatura, dada su capacidad para modelar relaciones no lineales complejas. El SVR es especialmente valorado en espacios de alta dimensionalidad y en conjuntos de datos de tamaño moderado, donde su formulación basada en márgenes le otorga ventajas de generalización frente a otros métodos [20].

Los hiperparámetros que gobiernan el comportamiento del SVR reflejan directamente sus fundamentos teóricos: el parámetro *C* controla el equilibrio entre la amplitud del margen y la penalización a los errores fuera de él; *epsilon* define la anchura de la zona de tolerancia; *kernel* determina el tipo de transformación del espacio de características; y *gamma* regula la influencia de cada punto de entrenamiento en la función de decisión para kernels no lineales [20]. La Tabla 4 resume estos parámetros y sus rangos típicos reportados en la literatura.

Tabla 4. Parámetros clave de regresión de vectores de soporte

PARÁMETRO (SCIKIT-LEARN/XGBOOST)	EXPLICACIÓN	RANGO TÍPICO (MIN–MAX)
C	Regularización, controla margen/error	0.1 – 1000
epsilon	Tolerancia en el margen	0.001 – 1.0
kernel	Tipo de función	'linear', 'rbf', 'poly', 'sigmoid'
gamma	Influencia de cada punto (kernels no lineales)	1e-4 – 10

La relación entre C y epsilon determina el bias-variance tradeoff en el SVR: valores bajos de C producen modelos más simples con mayor margen, pero mayor tolerancia al error, mientras que valores altos de gamma incrementan la sensibilidad del modelo a cada punto individual, aumentando el riesgo de sobreajuste. La selección del kernel y sus parámetros asociados define en última instancia la complejidad de las relaciones que el modelo es capaz de capturar [20].

- **Perceptrón Multicapa (MLP):** es una arquitectura de red neuronal artificial de tipo feedforward, compuesta por capas de nodos interconectados donde cada neurona aplica una transformación lineal a sus entradas seguida de una función de activación no lineal. Al apilar múltiples capas ocultas, el MLP adquiere la capacidad de aproximar cualquier función continua —propiedad conocida como el teorema de aproximación universal—, lo que le confiere un alto poder expresivo para capturar relaciones complejas en los datos [21]. Esta capacidad, sin embargo, conlleva un mayor costo computacional durante el entrenamiento y una mayor sensibilidad a la escala de las características y a la configuración de hiperparámetros, en comparación con modelos más simple [22].

Desde el punto de vista teórico, la arquitectura del MLP está determinada por tres dimensiones fundamentales: la estructura de capas (*hidden_layer_sizes*), que define la profundidad y anchura de la red; las funciones de activación, que introducen la no linealidad necesaria para modelar patrones complejos —siendo ReLU y tanh las más empleadas en la literatura—; y el algoritmo de optimización (*solver*), que gobierna la actualización de pesos mediante retro propagación del error [21]. La regularización L2 (*alpha*) actúa como mecanismo de control del sobreajuste al penalizar los pesos de gran magnitud, mientras que la tasa de aprendizaje (*learning_rate_init*) y el número máximo de iteraciones (*max_iter*) determinan la convergencia del proceso de entrenamiento [22]. La Tabla 5 resume estos hiperparámetros y sus rangos típicos reportados en la literatura.

Tabla 5. Parámetros clave de perceptrón multicapa (MLP)

PARÁMETRO (SCIKIT-LEARN)	EXPLICACIÓN	RANGO TÍPICO (MIN–MAX)
hidden_layer_sizes	Nº y tamaño de capas ocultas	1–3 capas, 10–200 neuronas
activation	Función de activación	'relu', 'tanh', 'logistic', 'identity'
solver	Algoritmo de optimización	'adam', 'sgd', 'lbfgs'
alpha	Regularización L2	1e-5 – 1.0
learning_rate_init	Tasa de aprendizaje	1e-4 – 0.1
max_iter	Iteraciones de entrenamiento	200 – 2000

La profundidad de la red —número de capas ocultas— y la anchura de cada capa determinan la capacidad representacional del modelo: redes más profundas pueden capturar jerarquías de características, pero incrementan el riesgo de sobreajuste y la dificultad de optimización. El equilibrio entre capacidad expresiva y generalización constituye el desafío central en el diseño de arquitecturas MLP, y su resolución depende tanto de la naturaleza del problema como del volumen de datos disponibles [21], [22].

- **XGBoost (Extreme Gradient Boosting):** es un algoritmo de ensamble basado en la técnica de gradient boosting, cuya arquitectura se distingue fundamentalmente del Random Forest por su naturaleza secuencial: en lugar de construir árboles de forma independiente, cada nuevo modelo se entrena para corregir los errores residuales del modelo anterior. Este enfoque aditivo —conocido como boosting de gradiente— permite minimizar iterativamente una función de pérdida diferenciable, lo que se traduce en niveles de precisión superiores a los obtenidos por métodos de ensamble paralelo en muchos problemas de predicción con datos tabulares [23], [24].

  Una característica teórica distintiva de XGBoost es la incorporación explícita de términos de regularización en su función objetivo, que penalizan tanto la complejidad de cada árbol individual como la magnitud de los pesos de las hojas. Esta regularización integrada —L1 (*reg_alpha*) y L2 (*reg_lambda*)— diferencia a XGBoost de las implementaciones clásicas de gradient boosting y contribuye a su robustez frente al sobreajuste [23]. Adicionalmente,

mecanismos como el submuestreo de observaciones (*subsample*) y de variables (colsample_bytree) en cada iteración introducen diversidad en los árboles construidos, reduciendo la varianza del ensamble de manera análoga al Random Forest [24].

Los hiperparámetros de XGBoost operan en dos niveles teóricos: los que controlan la estructura de cada árbol —profundidad máxima, umbral mínimo de reducción de pérdida (*gamma*)— y los que gobiernan la dinámica del proceso de boosting —tasa de aprendizaje (*learning_rate*) y número de iteraciones (*n_estimators*)—. La tasa de aprendizaje actúa como factor de contracción que escala la contribución de cada árbol, forzando al algoritmo a construir soluciones más conservadoras y generalizables [23]. La Tabla 6 resume estos hiperparámetros y sus rangos típicos reportados en la literatura.

Tabla 6. parámetros clave de XGBoost

PARÁMETRO (XGBOOST)	EXPLICACIÓN	RANGO TÍPICO (MIN–MAX)
n_estimators	Nº de árboles secuenciales	100 – 2000
max_depth	Profundidad máxima de cada árbol	3 – 15
learning_rate	Tasa de aprendizaje	0.01 – 0.3
subsample	Porcentaje de datos por árbol	0.5 – 1.0
colsample_bytree	Proporción de variables por árbol	0.5 – 1.0
gamma	Reducción mínima de pérdida para dividir	0 – 10
reg_alpha	Regularización L1	0 – 10
reg_lambda	Regularización L2	0 – 10

El equilibrio entre la tasa de aprendizaje y el número de árboles secuenciales ilustra el bias-variance tradeoff en el contexto del boosting: tasas de aprendizaje bajas con mayor número de iteraciones producen modelos más estables, mientras que la profundidad de cada árbol determina la capacidad del modelo para capturar interacciones de alto orden entre variables predictoras En este sentido, resulta fundamental no solo entender las particularidades de cada algoritmo sino también comparar sus características frente a los demás métodos disponibles. De esta forma, se puede seleccionar la opción más adecuada para cada desafío específico en la modelización ambiental y en la predicción de datos complejos [23], [24].

La literatura en modelado predictivo reconoce que ningún algoritmo de regresión es universalmente superior: el desempeño de cada método depende de la naturaleza de los datos, la dimensionalidad del problema y las relaciones entre variables [23]. Por ello, la evaluación comparativa de múltiples algoritmos —desde modelos lineales interpretables hasta ensambles de alta capacidad y redes neuronales— constituye una práctica estándar en ciencia de datos aplicada, que permite identificar empíricamente el enfoque más adecuado para cada contexto. La Tabla 7 sintetiza los fundamentos teóricos, fortalezas y limitaciones de los algoritmos considerados en este estudio en el marco de la modelización ambiental.

Tabla 7. Conjunto de algoritmos usados

MODELO	FUNDAMENTO TEÓRICO	FORTALEZAS	DEBILIDADES
Regresión Lineal Múltiple	Asume una relación lineal entre predictores y objetivo.	Modelo sencillo e interpretable.	Suposición lineal que puede no ajustarse a relaciones más complejas.
Árbol de Decisión	Decide dividiendo espacios según reglas jerárquicas.	Fácil de entender y explicar.	Alta tendencia al sobreajuste.
Random Forest	Combina múltiples árboles, promediando predicciones.	Alta precisión y robustez.	Menor interpretabilidad comparado con árboles únicos.
SVR (Support Vector Regression)	Busca hiperplanes con margen en espacios de alta dimensión.	Bueno en datos con relaciones no lineales y en espacios cortos.	Requiere cuidadoso ajuste de parámetros (C, kernels, gamma).
Perceptrón Multicapa (MLP)	Redes neuronales con capas interconectadas.	Capacidad para modelar relaciones complejas.	Necesita muchos datos y entrenamiento costoso.
XGBoost	Boosting secuencial, ajustando errores iterativamente.	Rendimiento de vanguardia y regularización incorporada.	Requiere ajuste fino de hiperparámetros.

OPTIMIZACIÓN DE HIPERPARÁMETROS EN MODELOS DE REGRESIÓN

En los modelos de aprendizaje automático, los hiperparámetros son configuraciones externas al proceso de entrenamiento —como el número de árboles en un *Random Forest* o la tasa de aprendizaje en XGBoost— que no se aprenden a partir de los datos, sino que deben definirse previamente. Su elección tiene un impacto directo sobre la capacidad predictiva y la generalización del modelo, razón por la cual la optimización de hiperparámetros es reconocida como una etapa crítica en el diseño de sistemas de aprendizaje supervisado [25].

El método clásico de optimización es la búsqueda exhaustiva en rejilla (*Grid Search*), que evalúa sistemáticamente todas las combinaciones posibles dentro de un espacio de búsqueda predefinido. Aunque garantiza encontrar el óptimo dentro del espacio considerado, su costo computacional crece exponencialmente con el número de hiperparámetros y valores evaluados [25], [26]. Frente a esta limitación, la optimización bayesiana —implementada en frameworks como Optuna— constituye una alternativa teóricamente más eficiente: construye un modelo probabilístico del espacio de búsqueda a partir de los resultados de ensayos anteriores y utiliza este modelo para decidir qué configuraciones explorar a continuación, equilibrando la exploración de regiones desconocidas y la explotación de regiones prometedoras [26], [27]. Este paradigma reduce sustancialmente el número de evaluaciones necesarias para aproximarse al óptimo global frente a la búsqueda exhaustiva [27].

El *Grid Search* constituye una estrategia de búsqueda exhaustiva para la optimización de

hiperparámetros, fundamentada en la exploración sistemática de todas las combinaciones posibles dentro de un espacio discreto de valores previamente definido. El desempeño de cada configuración se estima mediante validación cruzada, técnica que divide el conjunto de entrenamiento en *k* particiones para obtener una estimación robusta y menos sesgada del error de generalización [25]. Aunque su carácter exhaustivo garantiza una búsqueda completa y reproducible —lo que lo convierte en un referente de comparación entre configuraciones—, su costo computacional crece exponencialmente con el número de hiperparámetros y valores evaluados, constituyendo su principal limitación teórica frente a métodos más eficientes [26]. La Tabla 8 resume los parámetros que gobiernan este método y sus rangos típicos reportados en la literatura.

Tabla 8. Parámetros clave del método Grid Search (Scikit-Learn)

PARÁMETRO (SCIKIT-LEARN)	EXPLICACIÓN	RANGO TÍPICO (MIN–MAX)
param_grid	Diccionario con las combinaciones de hiperparámetros que serán exploradas sistemáticamente.	Depende del modelo (por ejemplo, profundidad, tasa de aprendizaje, número de estimadores).
cv	Número de particiones para validación cruzada.	3 – 10
scoring	Métrica de evaluación empleada para seleccionar el mejor modelo.	'r2', 'neg_mean_squared_error', 'neg_mean_absolute_error'
n_jobs	Núcleos de CPU utilizados para ejecución paralela.	-1 (todos los disponibles)
verbose	Nivel de detalle mostrado durante la búsqueda.	0 – 3
refit	Reentrena el mejor modelo con todos los datos tras la búsqueda.	True / False

Optuna representa un enfoque moderno de optimización de hiperparámetros fundamentado en la optimización bayesiana y la búsqueda adaptativa. Su principio teórico central es el uso de algoritmos de muestreo probabilístico que construyen un modelo sustituto (surrogate model) del espacio de búsqueda a partir de los ensayos previos, priorizando las regiones con mayor probabilidad de contener configuraciones óptimas [26]. Este mecanismo de aprendizaje secuencial distingue a Optuna de los métodos exhaustivos: en lugar de evaluar el espacio de manera uniforme, equilibra dinámicamente la exploración —evaluación de regiones poco conocidas— y la explotación —refinamiento de regiones prometedoras—, reduciendo sustancialmente el número de evaluaciones necesarias para aproximarse al óptimo [26]. Un elemento teórico adicional de Optuna es el uso de pruners, mecanismos de poda que interrumpen anticipadamente los ensayos cuyo desempeño parcial sugiere baja probabilidad de mejorar los resultados ya obtenidos. Esta capacidad de poda temprana incrementa la eficiencia del proceso sin comprometer la calidad de la búsqueda [26], [27]. La Tabla 9 resume los parámetros que gobiernan el comportamiento de este método y sus rangos típicos en la literatura.

Tabla 9. Parámetros clave del método Optuna

PARÁMETRO (OPTUNA)	EXPLICACIÓN	RANGO TÍPICO (MIN–MAX)
n_trials	Número máximo de combinaciones o ensayos evaluados.	50 – 500
direction	Dirección de optimización: maximizar o minimizar una métrica objetivo.	'maximize' / 'minimize'
sampler	Algoritmo de muestreo utilizado para seleccionar combinaciones de hiperparámetros.	'TPESampler', 'RandomSampler'
pruner	Método de poda para detener ensayos poco prometedores y ahorrar tiempo.	'MedianPruner', 'SuccessiveHalvingPruner'
timeout	Tiempo máximo total de búsqueda (en segundos).	Opcional (por ejemplo, 3600)
seed	Semilla aleatoria para garantizar reproducibilidad.	Entero positivo

La evaluación rigurosa de los modelos generados tras la optimización de hiperparámetros requiere herramientas que permitan diagnosticar tanto la capacidad predictiva como la estabilidad y generalización de cada algoritmo. La optimización por sí sola no garantiza un desempeño adecuado si el modelo incurre en sobreajuste o pérdida de capacidad de generalización sobre datos no vistos. Por ello, la literatura en aprendizaje automático reconoce un conjunto de técnicas complementarias orientadas a evaluar el rendimiento real de los modelos, interpretar su comportamiento interno y garantizar la reproducibilidad de los resultados.

La elección de la estrategia de optimización se definió de acuerdo con la complejidad del espacio de hiperparámetros de cada modelo. Grid Search se aplicó en modelos con combinaciones discretas y acotadas, mientras que la optimización bayesiana con Optuna se reservó para modelos con mayor cantidad de parámetros y rangos de búsqueda más amplios. Esta decisión permitió mantener un uso eficiente de los recursos computacionales sin perder trazabilidad en el proceso de ajuste. La comparabilidad entre modelos se garantizó mediante el uso de la misma partición de datos, la misma métrica objetivo, el mismo esquema de validación y las mismas métricas finales de evaluación.

EVALUACIÓN Y VALIDACIÓN DE MODELOS PREDICTIVOS

La evaluación cuantitativa del rendimiento de modelos predictivos se fundamenta en métricas que comparan los valores estimados por el modelo $(\hat{y}_i)$ con los valores reales observados (y_i) para cada observación *i*. Las tres métricas más empleadas en la literatura de regresión supervisada son el Error Absoluto Medio (MAE), la Raíz del Error Cuadrático Medio (RMSE) y el Coeficiente de Determinación (R^2) [28], [29].

El Error Absoluto Medio, se define como:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|\,,$$

Esta métrica calcula la media de las magnitudes de los errores en valor absoluto, donde *n* representa el número total de observaciones. Su interpretación es directa, ya que el resultado se expresa en las mismas unidades de la variable de salida, proporcionando una medida del error promedio sin penalización diferencial por errores de gran magnitud [28], [29].

La Raíz del Error Cuadrático Medio (RMSE) se obtiene como la raíz cuadrada del promedio de los errores al cuadrado:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}\,,$$

A diferencia del MAE, el RMSE penaliza de forma cuadrática los errores de mayor magnitud, lo que lo hace más sensible a valores atípicos en las predicciones. Valores bajos de RMSE indican mayor precisión y constituyen una métrica estándar para la comparación entre modelos y configuraciones de Hiperparámetro [28], [29].

El Coeficiente de Determinación (R^2) expresa la proporción de la variabilidad total de los datos que es explicada por el modelo:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\,,$$

donde $\bar{y}$ representa el valor medio de las observaciones reales. Este coeficiente toma valores entre 0 y 1, siendo 1 el ajuste perfecto. No obstante, la literatura advierte que un valor elevado de R^2 no garantiza la ausencia de sobreajuste, por lo que su interpretación debe realizarse en conjunto con MAE y RMSE [28], [29].

Más allá de estas métricas de desempeño global, existe un aspecto crítico que determina si el modelo realmente generalizará a datos nuevos: la capacidad de distinguir entre dos riesgos opuestos: el sobreajuste (overfitting) —cuando el modelo captura ruido en lugar de patrones— y el subajuste (underfitting) —cuando carece de capacidad para representar la complejidad de los datos— son los dos riesgos centrales que la evaluación busca detectar y cuantificar [30].

Las curvas de aprendizaje (*learning curves*) son una herramienta diagnóstica consolidada en la literatura para identificar estos fenómenos. Presentan la evolución del error del modelo en los conjuntos de entrenamiento y validación en función del tamaño muestral: una brecha amplia y persistente entre ambas curvas es indicativa de sobreajuste, mientras que la convergencia en niveles altos de error sugiere subajuste. Su análisis permite tomar decisiones sobre la complejidad

del modelo y la necesidad de datos adicionales [30].

Garantizar que un modelo generalice correctamente no es suficiente. En contextos aplicados, especialmente en la gestión ambiental, es fundamental entender por qué el modelo realiza sus predicciones. Un modelo puede tener métricas excelentes y, aun así, basarse en relaciones que los expertos del dominio considerarían contradictorias o no interpretables. Aquí es donde emerge la interpretabilidad como dimensión crítica.

En la literatura de aprendizaje automático aplicado, especialmente en contextos donde las predicciones deben ser comprensibles y coherentes con el conocimiento del dominio, el campo de la Inteligencia Artificial Explicable (XAI, por sus siglas en inglés) agrupa un conjunto de técnicas diseñadas para revelar el comportamiento interno de modelos que, de otro modo, operarían como sistemas opacos [31].

La Importancia por Permutación (*Permutation Importance*) es una de estas técnicas: estima la importancia global de cada variable midiendo la degradación del desempeño del modelo cuando sus valores son aleatorizados, rompiendo así la asociación entre el predictor y la variable objetivo [31]. Para un análisis de mayor granularidad, los valores SHAP (*SHapley Additive exPlanations*) —fundamentados en la teoría de juegos cooperativos— descomponen cada predicción individual en contribuciones atribuibles a cada variable, permitiendo explicar no solo qué predictores son globalmente relevantes, sino la dirección e intensidad de su influencia en cada caso específico [32]. Ambas técnicas se complementan con los Gráficos de Dependencia Parcial (PDP), que visualizan el efecto marginal de una o dos variables sobre la predicción del modelo, revelando la naturaleza de la relación aprendida —lineal, monotónica o no lineal— independientemente del resto de predictores [33].

La integración de métricas de desempeño, diagnóstico de generalización e interpretabilidad configura un marco evaluativo que trasciende la simple medición de la precisión predictiva, convirtiendo al modelo en un instrumento de generación de conocimiento sobre los procesos subyacentes al fenómeno estudiado [32], [33].

2.3.2. ANTECEDENTES

El antecedente más próximo geográfica y temáticamente al presente estudio es el trabajo de Bueno-Zabala, Torres-Lozada y Delgado-Cabrera [34], quienes evaluaron el comportamiento del pH en el agua tratada del Río Cauca destinada al abastecimiento de Santiago de Cali, ciudad que depende de este afluente para el 80% de su suministro hídrico, equivalente a aproximadamente 2.3 millones de habitantes. El estudio, desarrollado con datos del período 2010-2013, aplicó siete índices de estabilización fisicoquímica —entre ellos el Índice de Saturación de Langelier (ISL), el Índice de Estabilidad de Ryznar (IER) y la Relación de Larson (RL)— para caracterizar el equilibrio carbonato-calcio del agua tratada. Los resultados evidenciaron valores promedio de pH entre 6.50 y 7.17 unidades, notablemente por debajo del rango recomendado por la OMS (8.0-8.5), con una alcalinidad media de apenas 21 mg$CaCO_3$/L, confirmando el carácter moderadamente agresivo y

corrosivo del agua para las tuberías del sistema de distribución. Si bien este estudio constituye la única referencia científica publicada sobre el comportamiento del pH directamente en el Río Cauca, su alcance se limita a métodos analíticos fisicoquímicos clásicos, sin incorporar técnicas de aprendizaje automático ni abordar la predicción de la Demanda Bioquímica de Oxígeno. Esta limitación representa la brecha más directa que el presente proyecto busca cerrar, al proponer modelos predictivos basados en ML sobre el mismo sistema fluvial.

En el contexto académico colombiano, un grupo de investigación de la Universidad del Cauca —en colaboración con la Universidad de Medellín— desarrolló dos contribuciones complementarias y estrechamente relacionadas. En primer lugar, López, Figueroa y Corrales [35] llevaron a cabo un mapeo sistemático de la literatura sobre predicción de calidad del agua mediante técnicas de inteligencia computacional, publicado en la Revista Ingenierías de la Universidad de Medellín. El estudio seleccionó 39 trabajos de bases de datos como Elsevier, IEEE, ACM y Springer, e identificó las Redes Neuronales Artificiales (RNA), las Máquinas de Vectores de Soporte (MVS) y los sistemas inmunes artificiales como las técnicas más empleadas en el campo. Una conclusión fundamental del mapeo fue la identificación de una brecha crítica en la literatura: ninguno de los trabajos revisados consideraba la aplicabilidad de sus modelos a diferentes usos del recurso hídrico de forma simultánea, limitándose a predecir la calidad para un único uso específico. A partir de este hallazgo, el mismo autor principal desarrolló, como trabajo complementario de pasantía , un mecanismo adaptativo de predicción de calidad del agua evaluado sobre datos históricos del Instituto CINARA, orientado a predecir simultáneamente la calidad del agua para usos como consumo humano, riego agrícola, uso pecuario e industrial, a partir de variables fisicoquímicas que incluyen pH, DBO, oxígeno disuelto, conductividad eléctrica, turbidez, nitratos y sulfatos, aplicando RNA, modelos ARIMA y MVS. Esta línea de investigación del grupo de la Universidad del Cauca establece un precedente metodológico directo para el presente estudio, al demostrar tanto la necesidad como la viabilidad de aplicar técnicas computacionales avanzadas para la predicción de parámetros de calidad del agua en cuencas de la región andina colombiana.

Complementando el panorama desde una perspectiva bibliográfica más amplia, Aguilar y Obando-Díaz [28] presentaron en la revista INGENIARE de la Universidad Libre de Barranquilla una revisión sistemática de técnicas de aprendizaje automático aplicadas a la predicción de calidad de agua potable, analizando más de 50 estudios publicados entre 2016 y 2020 en las bases de datos ScienceDirect y SpringerLink. La revisión identificó que las RNA representan el 36% de los estudios analizados, seguidas de MVS y sistemas Anfis con el 24% cada uno, y los árboles de regresión con el 16%. Entre los hallazgos más relevantes para el presente proyecto, se destaca que el Random Forest alcanzó valores de R^2=0.93 en la predicción conjunta de pH, DBO y oxígeno disuelto utilizando temperatura y OD como variables de entrada clave, posicionándose como el algoritmo de mayor desempeño en el contexto de agua potable. Adicionalmente, la revisión evidenció que los métodos híbridos superan consistentemente a las técnicas individuales, y que la calidad y extensión de los datos históricos constituye el factor más determinante del desempeño predictivo, con valores de R^2 que oscilan entre 0.61 y 0.998 según el contexto y el volumen de datos

disponibles. Estas evidencias respaldan la decisión metodológica del presente proyecto de evaluar comparativamente múltiples algoritmos de regresión —en lugar de apostar por uno único— y de combinar métricas como R^2, RMSE y MAE para garantizar una evaluación robusta e integral del desempeño.

Los estudios en español revisados permiten establecer que, en el contexto colombiano y latinoamericano, existe una línea de investigación consolidada orientada a la predicción de parámetros de calidad del agua mediante técnicas de inteligencia computacional, con énfasis en RNA, MVS y métodos híbridos. No obstante, estos trabajos no incorporan técnicas avanzadas de interpretabilidad como SHAP, no aplican estrategias de control de multicolinealidad como el VIF o la Eliminación Recursiva de Características, y ninguno aborda específicamente la predicción simultánea de pH y DBO en el Río Cauca con un enfoque comparativo y reproducible.

La literatura internacional en lengua inglesa ha avanzado significativamente en estas dimensiones —particularmente en el uso de métodos de ensamble, optimización bayesiana de hiperparámetros e interpretabilidad explicable—, constituyendo el referente metodológico complementario que se presenta a continuación.

Desde una perspectiva de revisión internacional, Zhu et al. [36] publicaron en la revista Eco-Environment & Health (Elsevier, 2022) una revisión exhaustiva de la aplicación del aprendizaje automático en la evaluación de la calidad del agua, cubriendo agua superficial, subterránea, potable, residual y marina en múltiples contextos geográficos. El estudio documenta que los métodos de ensamble —Random Forest, Gradient Boosting y XGBoost— alcanzan sistemáticamente los mayores valores de R^2 con los menores errores reportados, mientras que arquitecturas de aprendizaje profundo como LSTM muestran ventajas particulares en la captura de dinámicas temporales no estacionarias. Entre los parámetros más frecuentemente modelados en agua superficial, el oxígeno disuelto, la DBO y el pH concentran la mayor parte de los estudios revisados, con desempeños que superan R^2=0.95 en los mejores casos. Crucialmente, los autores señalan la falta de interpretabilidad como la principal limitación de los modelos más complejos, identificando esta dimensión como una de las brechas prioritarias a resolver para que el aprendizaje automático pueda ser adoptado con confianza en la gestión ambiental institucional. Esta conclusión sustenta directamente la incorporación del análisis SHAP en el presente proyecto.

En cuanto a aplicaciones específicas sobre parámetros individuales de calidad del agua, Al-Sultani et al. [37] aplicaron Redes Neuronales Artificiales para predecir la idoneidad del agua subterránea para riego agrícola en la ciudad de Babilonia, Iraq, utilizando pH, conductividad eléctrica, sulfatos, cloruros, magnesio, calcio, potasio, sodio y bicarbonatos como variables de entrada sobre 111 pozos muestreados en 2019. Los modelos alcanzaron valores de R^2 entre 0.918 y 0.984 para la predicción de sólidos disueltos totales, porcentaje de sodio y relaciones de adsorción iónica, con errores relativos inferiores al 5.6% en todos los casos. Este estudio confirma que las ANN logran alta precisión predictiva a partir de variables fisicoquímicas básicas similares a las disponibles en la base de datos histórica del Río Cauca —especialmente la presencia de iones como calcio,

magnesio y sulfatos—, validando el enfoque de ML en contextos hídricos distintos al consumo humano y respaldando la generalización metodológica adoptada en el presente trabajo.

Avanzando hacia modelos de mayor complejidad, Islam et al. [38] desarrollaron un framework de aprendizaje automático para la predicción del Índice de Calidad del Agua (WQI) publicado en Applied Agriculture Sciences (2024), evaluando Gradient Boosting, Random Forest, XGBoost, CatBoost, Extra Trees y AdaBoost sobre un conjunto de datos que incluye pH, sodio, calcio, potasio, sulfatos y sólidos disueltos totales como predictores. El Gradient Boosting obtuvo el mejor desempeño individual con R^2=0.94 y RMSE=0.06, mientras que el ensamble apilado alcanzó R^2=0.9952. El análisis SHAP incorporado en este estudio reveló que los sulfatos y el calcio se encuentran entre las variables de mayor peso explicativo en la predicción del WQI —hallazgo especialmente relevante para el presente proyecto, donde los sulfatos emergieron también como la variable más influyente en la predicción de DBO en el Río Cauca, con un valor SHAP medio de 1.86.

En la misma línea metodológica, Hoque et al. [39] evaluaron ocho técnicas de regresión supervisada —incluyendo Regresión Lineal, Ridge, SVR, Random Forest, Extra Trees y ANN— sobre 1,991 registros históricos de calidad del agua de 233 estaciones de monitoreo en India, utilizando TDS, turbidez, pH y conductividad eléctrica como variables predictoras del WQI. El SVR alcanzó R^2=0.92 y RMSE=0.11, posicionándose como uno de los algoritmos más competitivos para la predicción de índices de calidad del agua en entornos con datos históricos de densidad moderada. El estudio demostró además que la ingeniería de características —específicamente la incorporación de índices derivados de calidad— mejora significativamente el desempeño de todos los modelos evaluados, principio que respalda la estrategia de selección de variables mediante RFE adoptada en el presente proyecto.

Finalmente, en el estado del arte más reciente, Choudhary et al. [29] publicaron en Scientific Reports de Nature (2025) un framework integrado de ensamble apilado con interpretabilidad basada en SHAP para la predicción continua del WQI en tiempo real, entrenado sobre 1,987 registros de ríos de India del período 2005-2014, con oxígeno disuelto, DBO, pH, conductividad, nitratos y coliformes como predictores. El ensamble apilado —compuesto por seis modelos base (XGBoost, CatBoost, Random Forest, Gradient Boosting, Extra Trees y AdaBoost) y Regresión Lineal como meta-learner validado con 5-fold cross-validation— alcanzó R^2=0.9952, MAE=0.7637 y RMSE=1.0704, superando en todas las métricas a cada modelo individual. El análisis SHAP confirmó que el oxígeno disuelto y la DBO son los parámetros de mayor contribución explicativa al WQI, seguidos por conductividad y pH, demostrando que la interpretabilidad basada en SHAP no sacrifica precisión predictiva, sino que la complementa con transparencia para la toma de decisiones ambientales. Este trabajo representa el estándar metodológico más avanzado disponible en la literatura y motivó directamente la incorporación de SHAP como herramienta de interpretabilidad en el presente proyecto.

En consecuencia, la Tabla 10 presenta una síntesis de los estudios más relevantes analizados,

destacando las variables objetivo, los modelos utilizados, sus métricas de rendimiento y los aportes específicos que cada investigación ofrece al diseño del presente estudio. Esta tabla consolida la base empírica y comparativa sobre la cual se sustenta la elección de los algoritmos de predicción empleados, permitiendo visualizar de forma estructurada los enfoques, resultados y líneas de avance que conforman el estado del arte actual en la modelación de calidad del agua.

Tabla 10. Resumen Comparativo de Estudios Clave en la Predicción de Calidad del Agua con Machine Learning.

DOCUMENTO	VARIABLE OBJETIVO	MODELOS USADOS	MEJOR RENDIMIENTO	VARIABLES DE ENTRADA CLAVE	APORTE PRINCIPAL AL PROYECTO ACTUAL
Aprendizaje automático para la predicción de calidad de agua potable [28]	pH, DBO, OD	Regresión lineal, RF, SVR	RF (R^2=0.93, RMSE bajo)	pH, DBO, T, OD	Demuestra la eficacia de Random Forest y la importancia de combinar métricas (R^2, RMSE) para evaluar modelos en calidad de agua potable
Monitoreo y medición del ajuste del pH del agua tratada del Rio Cauca mediante Índices de estabilización [34]	pH	Índices de Estabilización	N/A (Método Analítico-Descriptivo)	pH, T, Ca, Alcalinidad, TDS, Cl^-, SO_4^{2-}.	Identifica la baja alcalinidad como factor crítico de inestabilidad y justifica la necesidad de modelos predictivos para anticipar cambios en el equilibrio químico.
Un mapeo sistemático sobre predicción de calidad del agua mediante modelos de	DBO, OD, pH, TDS	Regresión, KNN, RF, SVM	RF (R^2=0.91, RMSE=0.12)	Datos fisicoquímicos históricos	Realiza un mapeo sistemático sobre ML supervisado en calidad de agua,

DOCUMENTO	VARIABLE OBJETIVO	MODELOS USADOS	MEJOR RENDIMIENTO	VARIABLES DE ENTRADA CLAVE	APORTE PRINCIPAL AL PROYECTO ACTUAL
aprendizaje supervisado [35]					comparando múltiples modelos y métricas clásicas
Anexo B Predicción Adaptativa de la Calidad del Agua Mediante Técnicas de Inteligencia Computacional [40]	pH, EC, T	SVR	SVR (R^2=0.92, RMSE=0.13)	pH, EC, T	Presenta el uso de SVR para la predicción de variables clave con ajuste de hiperparámetros y comparación de métricas
Improving Water Quality Index Prediction Using Regression Learning Models [39]	WQI, pH, OD, TDS	Regresión, SVR, RF, MLP	SVR (R^2=0.92, RMSE=0.11)	TDS, turbidez, pH, EC	Mejorando la predicción del Índice de calidad del agua usando regresión
Machine Learning-Driven Water Quality Index Prediction [38]	pH, TDS, Na, Ca	Gradient Boosting, Random Forest, XGBoost	Gradient Boosting (R^2=0.94, RMSE=0.06)	pH, TDS, Na, Ca, K, SO_4^{2-}	Predicción del Índice de calidad del agua impulsada por aprendizaje automático
A review of the application of machine learning in water quality prediction [36]	pH, EC, WQI	Regresión, Ensambles, Redes neuronales	Ensambles (alta R^2, bajo RMSE)	Parámetros fisicoquímicos variados	Revisión de la aplicación del aprendizaje automático en la predicción de la calidad del agua
Predicting water quality index	WQI	XGBoost, CatBoost, RF,	Ensamble Apilado	OD, DBO, EC, pH	Introduce el concepto de

DOCUMENTO	VARIABLE OBJETIVO	MODELOS USADOS	MEJOR RENDIMIENTO	VARIABLES DE ENTRADA CLAVE	APORTE PRINCIPAL AL PROYECTO ACTUAL
using stacked ensemble regression and SHAP based explainable artificial intelligence [29]		GBM, Extra Trees, AdaBoost	(Stacking) (R^2=99%)		meta-aprendizaje (Stacking) y utiliza SHAP para confirmar la importancia de DBO y pH.
Artificial Neural Network Assessment of Groundwater Quality for Agricultural Use in Babylon City: An Evaluation of Salinity and Ionic Composition [37]	TDS, SAR, MAR, Na% (para uso agrícola)	ANN	ANN (R^2 hasta 98%)	pH, EC, iones (Mg^{2+}, Na^+, Ca^{2+}, etc.)	Demuestra la alta precisión de las ANN para predecir parámetros de calidad del agua a partir de variables fisicoquímicas, validando el enfoque de ML en un contexto de uso agrícola del agua.

Los términos técnicos y abreviaciones señalados en la presente tabla se definen en el ANEXO A. Diccionario de variables para proyectos similares . En síntesis, los antecedentes revisados confirmaron que el aprendizaje automático constituye una herramienta consolidada para la predicción de parámetros de calidad del agua, con resultados consistentemente superiores a los métodos estadísticos clásicos. No obstante, se evidenció una brecha en la literatura respecto a la aplicación simultánea de estos modelos sobre el pH y la DBO en el río Cauca, con énfasis en la interpretabilidad de los resultados. El presente estudio respondió a esta necesidad, aportando un enfoque reproducible y explicable adaptado a las condiciones particulares de esta cuenca.

3. PREPARACIÓN Y CARGA DE DATOS

Este capítulo describe el proceso de preparación de los datos de calidad del agua del río Cauca, etapa fundamental para garantizar que la información utilizada en el modelado predictivo fuera coherente, completa y representativa. El tratamiento adecuado de los datos constituye un requisito previo al entrenamiento de cualquier modelo de aprendizaje automático, dado que la presencia de valores faltantes, variables irrelevantes o inconsistencias en los registros puede comprometer significativamente la capacidad predictiva de los algoritmos empleados.

Como parte inicial del proceso, se visualizó el estado bruto y la estructura original de la base de datos, lo que permitió identificar variables nulas, redundantes o que carecían de relevancia para los objetivos del análisis. En esta etapa exploratoria se examinaron las primeras filas y columnas de la base de datos, como muestra la Tabla 11, donde se muestran las primeras 5 filas y 5 columnas de las 56 disponibles.

Tabla 11. Vista inicial de la base de datos de calidad del agua del Río Cauca.

FECHA DE MUESTREO	ESTACIONES	pH	TEMPERATURA (°C)	DEMANDA BIOQUÍMICA DE OXIGENO (mg O2/l)	...	CAUDAL (m3/s)
12/19/1998 12:00:00 AM	YOTOCO	7.1	4.1	4.2	...	NaN
12/19/1998 12:00:00 AM	MEDIACANOA	7	2	3	...	NaN
12/19/1998 12:00:00 AM	PASO DE LA TORRE	7	22.9	5	...	NaN
05/09/1990 0:00	ANTES SUAREZ	6.6	NaN	0.5	...	NaN
1/10/1990 0:00	ANTES RIO OVEJAS	6.7	NaN	2.1	...	NaN
5 rows × 56 columns						

Una vez realizada la revisión exploratoria y visualizada la estructura básica de la base de datos, se procede a una depuración inicial, eliminando variables que, por definición del proyecto, no aportarán directamente al análisis predictivo, como la fecha de muestreo y la estación[2]. Este paso asegura que el conjunto de trabajo se enfoque exclusivamente en las variables fisicoquímicas y ecológicas de interés para el modelado.

Con el propósito de complementar esta revisión exploratoria y fortalecer la interpretación ambiental del conjunto de variables, a continuación, se presenta el *Diccionario de variables*

[2] Se especifica en el proyecto que la variable tiempo y la variable estación no se usarían

fisicoquímicas del Río Cauca.

Este compendio reúne la totalidad de los parámetros analizados en el estudio, junto con sus unidades de medida, definiciones concisas y su relación con los procesos de contaminación desde un enfoque ecológico. Su incorporación dentro del documento permite disponer de una referencia técnica integral que facilita la comprensión del papel de cada variable en el análisis predictivo desarrollado.

En este sentido es necesario aclarar que el presente estudio no contempla en sus alcances un análisis de series de tiempo, por tanto y como ya se mencionó anteriormente la fecha de muestreo y la estación fueron usadas para verificar la estructura de la base de datos y no se incorporan al conjunto final de variables predictoras. Esta decisión metodológica focaliza el estudio al desarrollo de modelos predictivos con datos tabulares, a partir de variables fisicoquímicas y ecológicas de interés. Bajo este enfoque, la base de datos cuenta con una granularidad a nivel de evento de muestreo puntual, lo que significa que no se aplicaron técnicas de agregación temporal. De este modo, la unidad de análisis se define como la muestra independiente, permitiendo que los algoritmos evalúen las relaciones fisicoquímicas directas en cada instante de registro.

Tabla 12. Diccionario de variables fisicoquímicas del Río Cauca[3]

CLASIFICACIÓN	PARÁMETRO (UNIDAD)	DEFINICIÓN CONCISA	RELACIÓN CON LA CONTAMINACIÓN (ENFOQUE ECOLÓGICO)
Parámetros Físicos Generales	TEMPERATURA (°C)	Medida de la energía calorífica del agua	Contaminación Térmica: Altera el metabolismo, la reproducción y la supervivencia de la vida acuática. Reduce drásticamente el oxígeno disuelto.
	COLOR (UPC)	Apariencia visual del agua por materia disuelta o suspendida	Contaminación Química/Orgánica: Indica presencia de efluentes industriales (tintes, químicos) o alta carga de materia orgánica en descomposición. Bloquea la luz solar, afectando la fotosíntesis.
	TURBIEDAD (UNT)	Medida de la falta de claridad del agua por partículas en suspensión	Contaminación por Sólidos: Refleja la cantidad de sólidos suspendidos por erosión (deforestación, obras civiles) o descargas de aguas residuales. Afecta la respiración de los peces

[3] Fuente: Elaboración propia a partir de Baird, Rice y Posavec [41]

CLASIFICACIÓN	PARÁMETRO (UNIDAD)	DEFINICIÓN CONCISA	RELACIÓN CON LA CONTAMINACIÓN (ENFOQUE ECOLÓGICO)
			(branquias) y destruye hábitats.
	SÓLIDOS TOTALES (MG/L)	Suma de toda la materia disuelta y suspendida que queda al evaporar el agua	Medida general de la carga de materiales por vertidos o erosión.
	SÓLIDOS SUSPENDIDOS TOTALES (MG/L)	Partículas sólidas no disueltas que son retenidas por un filtro	Indicador clave de aguas residuales y erosión. Transporta contaminantes y daña hábitats acuáticos.
	SÓLIDOS DISUELTOS TOTALES (MG/L)	Sales y minerales que permanecen disueltos tras filtrar el agua	Afecta el sabor. Puede indicar contaminación salina o vertidos industriales (sales, metales).
	CONDUCTIVIDAD ELÉCTRICA (µS/CM)	Capacidad del agua para conducir electricidad debido a los iones disueltos (SDT)	Indicador rápido de la cantidad de SDT y contaminación por sales o vertidos iónicos.
Parámetros Químicos Generales y de Contaminación Orgánica	PH	Medida del grado de acidez o alcalinidad del agua en una escala de 0 a 14	Desviaciones del rango natural (6.5-8.5) indican contaminación ácida (drenaje minero, lluvia ácida) o alcalina (vertidos industriales), afectando la vida acuática y la toxicidad de otros químicos.
	DEMANDA BIOQUÍMICA DE OXIGENO (MG O2/L)	Cantidad de oxígeno que los microorganismos consumen para descomponer la materia orgánica biodegradable	Es el indicador principal de contaminación por materia orgánica (aguas residuales). Una DBO alta agota el oxígeno del agua, provocando la muerte de peces y vida acuática.
	DEMANDA QUÍMICA DE OXIGENO (MG O2/L)	Cantidad de oxígeno requerido para oxidar químicamente toda la materia orgánica (biodegradable y no biodegradable)	Mide la carga orgánica total. Es útil para detectar contaminación industrial con químicos no biodegradables. Una alta relación DQO/DBO sugiere contaminación química.

CLASIFICACIÓN	PARÁMETRO (UNIDAD)	DEFINICIÓN CONCISA	RELACIÓN CON LA CONTAMINACIÓN (ENFOQUE ECOLÓGICO)
	OXIGENO DISUELTO (MG O2/L)	Cantidad de oxígeno gaseoso (O_2) que se encuentra disuelto en el agua, esencial para la vida acuática	Es el indicador más directo de la salud de un río. La contaminación orgánica (alta DBO) reduce el oxígeno, causando hipoxia (bajo O_2) y la muerte masiva de peces.
Parámetros de Dureza y Alcalinidad	DUREZA TOTAL (MG CACO3/L)	Medida de la concentración total de iones minerales multivalentes, principalmente Calcio (Ca^{2+}) y Magnesio (Mg^{2+}), disueltos en el agua	La dureza es principalmente de origen natural (geológico). No es un indicador directo de contaminación, pero niveles anómalos pueden señalar vertidos industriales. Modula la toxicidad de metales pesados para la vida acuática.
	DUREZA CÁLCICA (MG CACO3/L)	Fracción de la dureza total que es aportada específicamente por las sales de calcio (Ca^{2+})	Indica la contribución del calcio a la dureza. Es de origen geológico. Cambios drásticos pueden estar asociados a vertidos de industrias que usan sales de calcio (construcción, alimentos).
	DUREZA MAGNÉSICA (MG CACO3/L)	Fracción de la dureza total que es aportada específicamente por las sales de magnesio (Mg^{2+})	Indica la contribución del magnesio a la dureza. De origen geológico. Alteraciones pueden sugerir vertidos de industrias metalúrgicas o de fertilizantes.
	CALCIO (MG CA/L)	Concentración del ion calcio (Ca^{2+}), uno de los principales cationes en aguas naturales y componente esencial de la dureza	Es un nutriente esencial, no un contaminante per se. Su origen es geológico. Concentraciones extremadamente altas pueden estar ligadas a vertidos industriales.
	MAGNESIO (MG MG/L)	Concentración del ion magnesio (Mg^{2+}), el segundo catión más importante que contribuye a la dureza del agua	Nutriente esencial. Su origen es geológico. No se considera contaminante, aunque valores muy elevados pueden indicar vertidos de ciertas industrias o escorrentía agrícola.
	ALCALINIDAD TOTAL (MG CACO3/L)	Medida de la capacidad del agua para neutralizar ácidos (capacidad buffer), principalmente debida a	Es un indicador de la estabilidad del pH. Aguas con muy baja alcalinidad son vulnerables a la lluvia ácida. Cambios bruscos

CLASIFICACIÓN	PARÁMETRO (UNIDAD)	DEFINICIÓN CONCISA	RELACIÓN CON LA CONTAMINACIÓN (ENFOQUE ECOLÓGICO)
		la presencia de bicarbonatos, carbonatos e hidróxidos	pueden ser causados por vertidos industriales (ácidos o alcalinos) que agotan esta capacidad.
	BICARBONATOS (MG CACO3/L)	Es la principal forma de alcalinidad en aguas naturales. Se originan por la disolución de rocas carbonatadas por el CO_2 del agua	Componente clave del ciclo del carbono acuático y principal agente buffer. Su concentración es natural. Una disminución drástica indica la adición de un ácido fuerte, señal de contaminación.
Iones y Compuestos Inorgánicos (No Metálicos)	CIANUROS (µG CN-/L)	Anión (CN^-) extremadamente tóxico. Su presencia en aguas naturales es casi nula	Indicador inequívoco de contaminación industrial, principalmente de minería (oro), galvanoplastia y siderurgia. Es altamente letal para la vida acuática al inhibir la respiración celular.
	FLUORUROS (MG F-/L)	Ión fluoruro (F^-) que se encuentra de forma natural en el agua por la disolución de minerales	Principalmente de origen geológico. Niveles anómalamente altos pueden indicar contaminación por industrias de aluminio, fertilizantes o semiconductores. Es tóxico para la vida acuática en altas concentraciones.
	SULFUROS (MG S2-/L)	Compuestos de azufre en su estado de oxidación más bajo, a menudo presentes como el gas tóxico sulfuro de hidrógeno (H_2S) con olor a "huevo podrido"	Su presencia indica condiciones anóxicas (sin oxígeno) por descomposición de materia orgánica. Es un claro signo de contaminación severa por aguas residuales o industriales (curtiembres).
	SÍLICE (MG SI/L)	Dióxido de silicio (SiO_2) disuelto en el agua, proveniente de la meteorización natural de las rocas	Es un componente natural y no un indicador de contaminación. Actúa como un nutriente esencial para organismos acuáticos como las diatomeas (algas con caparazón de sílice).

CLASIFICACIÓN	PARÁMETRO (UNIDAD)	DEFINICIÓN CONCISA	RELACIÓN CON LA CONTAMINACIÓN (ENFOQUE ECOLÓGICO)
	CLORUROS (MG CL-/L)	Ión cloruro (Cl^-), una de las sales más comunes y solubles en el agua	Niveles elevados en aguas dulces son un claro indicador de contaminación por aguas residuales, efluentes industriales o lixiviados de rellenos sanitarios. Afecta el equilibrio osmótico de los organismos.
	SULFATOS (MG SO4/L)	Ión sulfato (SO_4^{2-}), presente de forma natural en el agua por disolución de minerales como el yeso	Altas concentraciones son un indicador clave de drenaje ácido de mina, una forma grave de contaminación. También puede provenir de vertidos industriales (textiles, papeleras).
Nutrientes	NITRÓGENO TOTAL (MG N/L)	Suma de todas las formas de nitrógeno presentes en el agua (orgánico, amoniacal, nitritos y nitratos)	Es el indicador general de la carga de nitrógeno. El exceso de nitrógeno, junto con el fósforo, causa eutrofización: un crecimiento explosivo de algas que agota el oxígeno y mata la vida acuática.
	NITRÓGENO AMONIACAL (MG N-NH3/L)	Forma de nitrógeno (amonio/amoníaco) que proviene de la descomposición de materia orgánica	Es un indicador directo de contaminación reciente por aguas residuales domésticas o escorrentía ganadera. Es altamente tóxico para los peces, especialmente a pH elevado.
	NITRITOS (MG N-NO_2/L)	Forma intermedia y muy inestable del nitrógeno durante el proceso de nitrificación (de amoníaco a nitrato)	Su presencia indica un proceso de contaminación activa y en curso. Es tóxico para la vida acuática porque interfiere con el transporte de oxígeno en la sangre (metahemoglobinemia).
	NITRATOS (MG N-NO3/L)	Forma final y más estable de la oxidación del nitrógeno. Es un nutriente principal para las plantas y algas	Indicador de contaminación por fertilizantes agrícolas, aguas residuales tratadas y lixiviados de sistemas sépticos. Es el principal contribuyente a la eutrofización en muchos ríos.

CLASIFICACIÓN	PARÁMETRO (UNIDAD)	DEFINICIÓN CONCISA	RELACIÓN CON LA CONTAMINACIÓN (ENFOQUE ECOLÓGICO)
	FÓSFORO TOTAL (MG P/L)	Suma de todas las formas de fósforo en el agua (ortofosfatos, polifosfatos y fósforo orgánico)	Es el indicador general de la carga de fósforo. Generalmente, es el nutriente limitante en aguas dulces; pequeñas adiciones pueden causar una eutrofización severa y rápida.
	FOSFATOS (MG PO4/L)	Forma de fósforo directamente asimilable por las plantas y algas (fósforo reactivo)	Indicador directo de contaminación por detergentes, aguas residuales y fertilizantes. Es el principal motor del crecimiento explosivo de algas que conduce a la eutrofización.
Metales (Totales y Disueltos)	HIERRO TOTAL (MG FE/L)	Suma de todas las formas de hierro (disueltas y particuladas) presentes en el agua	Principalmente de origen geológico. Niveles altos indican drenaje ácido de mina o vertidos industriales (siderurgia). El hierro particulado puede cubrir los lechos de los ríos, afectando hábitats.
	HIERRO DISUELTO (MG FE/L)	Fracción del hierro que se encuentra soluble en el agua. Es la forma biodisponible y la que reacciona químicamente	En aguas oxigenadas, la concentración de hierro disuelto es naturalmente baja. Niveles elevados son un signo de condiciones de bajo oxígeno (anoxia) en el fondo o de contaminación reciente.
	MANGANESO TOTAL (MG MN/L)	Suma de todas las formas de manganeso (disueltas y particuladas) en el agua	De origen natural, pero altas concentraciones están asociadas a drenaje de minas y vertidos industriales. Es más soluble en condiciones de bajo oxígeno. Es un micronutriente esencial pero tóxico en exceso.
	MANGANESO DISUELTO (MG MN/L)	Fracción del manganeso que se encuentra soluble en el agua	Niveles altos son un indicador claro de condiciones anóxicas en el sedimento o en el fondo del cuerpo de agua, lo que promueve su liberación. Es la forma directamente tóxica para la vida acuática.

CLASIFICACIÓN	PARÁMETRO (UNIDAD)	DEFINICIÓN CONCISA	RELACIÓN CON LA CONTAMINACIÓN (ENFOQUE ECOLÓGICO)
	SODIO TOTAL (MG NA/L)	Suma de todas las formas de sodio en el agua. Dado que es muy soluble, es prácticamente igual al sodio disuelto	Principalmente de origen natural. Niveles elevados en aguas dulces indican contaminación por aguas residuales, vertidos industriales o intrusión salina. Afecta el equilibrio osmótico de los organismos.
	SODIO DISUELTO (MG NA/L)	Concentración del ion sodio (Na^+), un catión mayor en la mayoría de las aguas	Al ser la forma predominante, su interpretación es idéntica a la del Sodio Total. Es un indicador clave de la salinización de aguas dulces, un problema ecológico creciente.
	POTASIO TOTAL (MG K/L)	Suma de todas las formas de potasio en el agua. Al ser muy soluble, es prácticamente igual al potasio disuelto	Principalmente de origen natural (geológico). Niveles elevados en aguas dulces son un indicador de contaminación por fertilizantes (ceniza y potasa) y aguas residuales.
	POTASIO DISUELTO (MG K/L)	Concentración del ion potasio (K^+), un catión mayor y nutriente esencial para las plantas	Al ser la forma predominante, su interpretación es idéntica a la del Potasio Total. Es un indicador clave de la escorrentía agrícola y de la influencia de aguas residuales.
	COBRE TOTAL (MG CU/L)	Suma de todas las formas de cobre (disueltas y particuladas) presentes en el agua	El cobre es un micronutriente esencial pero altamente tóxico para la vida acuática en concentraciones bajas. Su presencia indica contaminación por minería, fungicidas o vertidos industriales (galvanoplastia).
	COBRE DISUELTO (MG CU/L)	Fracción del cobre que se encuentra soluble en el agua. Es la forma más biodisponible y tóxica para los organismos acuáticos	Indicador directo de la toxicidad inmediata del cobre en el ecosistema. Su concentración es crítica para determinar el riesgo para peces e invertebrados.

CLASIFICACIÓN	PARÁMETRO (UNIDAD)	DEFINICIÓN CONCISA	RELACIÓN CON LA CONTAMINACIÓN (ENFOQUE ECOLÓGICO)
	ZINC TOTAL (MG ZN/L)	Suma de todas las formas de zinc (disueltas y particuladas) en el agua	El zinc es un micronutriente esencial que se vuelve tóxico en altas concentraciones. Su presencia indica contaminación por minería, efluentes industriales (galvanizado) y lixiviados de llantas.
	ZINC DISUELTO (MG ZN/L)	Fracción del zinc que se encuentra soluble en el agua, siendo la forma más biodisponible y tóxica	Indicador directo de la toxicidad inmediata del zinc en el ecosistema. Su concentración depende fuertemente del pH y la dureza del agua.
	CADMIO TOTAL (MG CD/L)	Suma de todas las formas de cadmio (disueltas y particuladas). Es un metal pesado sin función biológica conocida	El cadmio es extremadamente tóxico y se bioacumula. Su presencia indica contaminación por minería, fertilizantes fosfatados o efluentes industriales (baterías, pigmentos).
	CADMIO DISUELTO (MG CD/L)	Fracción del cadmio que se encuentra soluble en el agua. Es la forma más biodisponible y tóxica	Indicador directo de la toxicidad inmediata del cadmio. La toxicidad varía con la dureza del agua (es más tóxico en aguas blandas).
	CROMO TOTAL (MG CR/L)	Suma de todas las formas de cromo, principalmente Cromo (III), un micronutriente, y Cromo (VI), altamente tóxico	Su presencia indica contaminación por vertidos industriales (curtiembres, galvanoplastia, textiles). El Cromo (VI) es la forma de mayor preocupación ecológica por su alta toxicidad y carcinogenicidad.
	CROMO DISUELTO (MG CR/L)	Fracción soluble del cromo, que incluye la forma más tóxica, el Cromo (VI)	Mide la fracción de cromo que está inmediatamente disponible para ser absorbida por los organismos acuáticos, representando el riesgo tóxico agudo en el ecosistema.

CLASIFICACIÓN	PARÁMETRO (UNIDAD)	DEFINICIÓN CONCISA	RELACIÓN CON LA CONTAMINACIÓN (ENFOQUE ECOLÓGICO)
	NÍQUEL TOTAL (MG NI/L)	Suma de todas las formas de níquel (disueltas y particuladas) en el agua	El níquel es un micronutriente que se vuelve tóxico en altas concentraciones. Su presencia indica contaminación por minería y vertidos industriales (aleaciones, galvanoplastia, baterías).
	NÍQUEL DISUELTO (MG NI/L)	Fracción del níquel que se encuentra soluble en el agua, siendo la forma más biodisponible y tóxica	Indicador directo del riesgo tóxico del níquel para la vida acuática. Su toxicidad es dependiente de la dureza del agua, siendo mayor en aguas blandas.
	PLOMO TOTAL (MG PB/L)	Suma de todas las formas de plomo (disueltas y particuladas). Es un metal pesado sin función biológica, altamente tóxico	El plomo es extremadamente tóxico y se bioacumula. Su presencia indica contaminación por minería, efluentes industriales, lixiviados o tuberías antiguas.
	PLOMO DISUELTO (MG PB/L)	Fracción del plomo que se encuentra soluble en el agua. Es la forma más biodisponible y tóxica	Indicador directo de la toxicidad inmediata del plomo en el ecosistema. La toxicidad varía con la dureza del agua (es más tóxico en aguas blandas).
	MERCURIO (µG HG/L)	Suma de todas las formas de mercurio, un metal pesado extremadamente tóxico que se biomagnifica en la cadena alimentaria	Su presencia indica contaminación por minería de oro (artesanal e industrial), quema de combustibles fósiles y efluentes industriales. Se convierte en metilmercurio, una potente neurotoxina.
Parámetros Biológicos	COLIFORMES TOTALES (NMP/100 ML)	Grupo de bacterias que se encuentran en el ambiente (suelo, agua, vegetación) y en las heces de animales de sangre caliente	Son un indicador general de la integridad del sistema. Su presencia en grandes cantidades sugiere una posible contaminación por una fuente externa (escorrentía, aguas residuales) y la posible presencia de patógenos.

CLASIFICACIÓN	PARÁMETRO (UNIDAD)	DEFINICIÓN CONCISA	RELACIÓN CON LA CONTAMINACIÓN (ENFOQUE ECOLÓGICO)
	COLIFORMES FECALES (NMP/100 ML)	Subgrupo de coliformes totales que sobrevive a temperaturas elevadas, más específico de las heces de animales de sangre caliente	Indicador directo de contaminación fecal por aguas residuales domésticas, sistemas sépticos defectuosos o escorrentía ganadera. Su presencia implica un alto riesgo sanitario por patógenos.

La incorporación de este diccionario dentro del cuerpo del documento no solo facilita la consulta directa de los parámetros evaluados, sino que también permite establecer una base coherente para los procedimientos de depuración y normalización de los datos.

Este diccionario también hace parte del código realizado para este proyecto, por medio de una herramienta interactiva que está enfocada en la facilidad de la consulta técnica[4].

Con una depuración inicial y un diccionario instaurado, se observa en la Tabla 13, que la base de datos queda compuesta únicamente por variables relevantes y contextualizadas para el estudio, optimizando así el análisis posterior.

Tabla 13. Vista del conjunto de datos luego de la eliminación de columnas no utilizadas (Fecha de muestreo y Estaciones).

pH	TEMPERATURA (°C)	DEMANDA BIOQUÍMICA DE OXIGENO (mg O2/l)	DEMANDA QUÍMICA DE OXIGENO (mg O2/l)	OXIGENO DISUELTO (mg O2/l)	...	CAUDAL (m3/s)
7.1	4.1	4.2	NaN	1.5	...	NaN
7	2	3	NaN	1.81	...	NaN
7	22.9	5	NaN	2	...	NaN
6.6	NaN	0.5	5.2	5.6	...	NaN
6.7	NaN	2.1	24	6	...	NaN
5 rows × 54 columns						

Finalizado el proceso de depuración y validación de las variables, se procedió a la transformación de los parámetros a formato numérico, estandarizando la codificación y el uso de decimales. Durante este ajuste, se eliminaron las columnas cuyos registros eran completamente nulos. De esta forma, la base de datos quedó conformada por 39 variables numéricas aptas para modelado y análisis estadístico, como se muestra en la Tabla 14.

[4] Esta ayuda visual solo se verá en el código y en el GitHub del proyecto, debido a que su formato y tamaño no permite una visualización técnica dentro de este documento.

Tabla 14. Variables numéricas depuradas con número de registros no nulos y tipo de dato en el conjunto de datos de calidad del agua del Río Cauca.

#	COLUMNA	NON-NULL COUNT	DTYPE
0	pH	2208 non-null	float64
1	TEMPERATURA (°C)	1956 non-null	float64
2	COLOR (UPC)	1882 non-null	float64
3	TURBIEDAD (UNT)	2090 non-null	float64
4	COLIFORMES FECALES (NMP/100 ml)	1870 non-null	float64
5	CAUDAL (m3/s)	222 non-null	float64
…	…	…	…
37	COLIFORMES FECALES (NMP/100 ml)	1870 non-null	float64
38	CAUDAL (m3/s)	222 non-null	float64
dtypes: float64(39)			

Posterior a la conversión de las variables a valores numéricos, se procedió con el cálculo de estadísticas descriptivas para cada parámetro de calidad del agua. Este análisis incluyó medidas de tendencia central (media, mediana), dispersión (desviación estándar, rango intercuartílico), y forma de distribución (asimetría, curtosis), así como el porcentaje de datos válidos. Esto permite comprender la variabilidad y distribución de los parámetros, además de identificar patrones anómalos o posibles outliers fundamentales para la calidad de los análisis posteriores.

Este enfoque no solo provee un panorama integral del comportamiento de los datos, sino que también habilita la representación rigurosa de los resultados en tablas y figuras que facilitan su interpretación y comparación.

Tabla 15. Extracto de la tabla de estadísticas descriptivas

VARIABLE	N.VALID	MEAN	MEDIAN	STD.DEV	MIN	MAX	SKEWNESS
pH	2208	7.04	7.07	0.41	4.1	9.7	-0.1
TEMPERATURA (°C)	1956	22.02	22.4	3.6	0	32.7	-2.34
COLOR (UPC)	1882	135.45	70	244.58	0	2956	6.16
TURBIEDAD (UNT)	2090	130.73	65	171.03	1	1900	2.86
SOLIDOS TOTALES (mg SST/l)	2194	275.19	206	227.62	0	2361	3.14
SOLIDOS SUSPENDIDOS TOTALES (mg SS/l)	2188	165.2	92.65	202.91	1.2	2112	3.06
DEMANDA BIOQUÍMICA DE OXIGENO (mg O2/l)	2092	5.22	3.1	16.25	0.1	427	16.25
DEMANDA QUÍMICA DE OXIGENO (mg O2/l)	2066	29.89	20.73	37.29	1.46	706	7.24

OXIGENO DISUELTO (mg O2/l)	2179	4.1	3.84	2.96	0	51.5	5.94
…	…	…	…	…	…	…	…
DUREZA CÁLCICA (mg CaCO3/l)	2166	27.37	26	11.5	0	126.5	1.54

En la Tabla 15 se exponen estas estadísticas descriptivas para un subconjunto representativo de variables fisicoquímicas, incluyendo pH, temperatura, color, turbiedad y sólidos totales, así como parámetros bioquímicos y de dureza. La tabla íntegra, que abarca las 39 variables evaluadas, se encuentra en el ANEXO B. Estadísticas de las variables fisicoquímicas del Rio Cauca.

En relación con los resultados más relevantes, se observa que el pH presenta una media de 7.04 y baja dispersión (desviación estándar de 0.41), lo que indica una relativa estabilidad y valores cercanos al rango natural (6.5–8.5). Por su parte, la Demanda Bioquímica de Oxígeno (DBO) muestra una media de 5.22 mg/L y una mayor dispersión, reflejando variaciones significativas en la carga de materia orgánica biodegradable. Entre las variables complementarias, la turbiedad y el color registran altos niveles de variabilidad, lo que sugiere la presencia de sólidos suspendidos y materia orgánica en descargas puntuales. Finalmente, el oxígeno disuelto presenta un valor medio de 4.10 mg/L, considerado bajo, lo cual guarda relación directa con los incrementos en la DBO y evidencia un posible riesgo para la vida acuática.

Con el fin de garantizar la calidad del análisis, se estableció un umbral mínimo de completitud del 80% para las variables consideradas. Aquellas que no alcanzaron este criterio fueron identificadas y separadas del conjunto de trabajo. En total se detectaron 12 variables con menos del 80% de registros válidos, dentro de las cuales se encuentran parámetros como Manganeso total, Nitrógeno total, Nitrógeno amoniacal, Cromo total, Níquel total, Plomo total y Caudal, entre otros.

Tabla 16. Filtro de variables con baja completitud de datos (menos del 80%)

VARIABLE	PCT.VALID	N.VALID	TIPO DE DATO
MANGANESO TOTAL (mg Mn/l)	70.72	1594	float64
COBRE TOTAL (mg Cu/l)	23.82	537	float64
ZINC TOTAL (mg Zn/l)	46.10	1039	float64
NITRÓGENO TOTAL (mg N/l)	58.65	1322	float64
NITRÓGENO AMONIACAL (mg N-NH3/l)	57.14	1288	float64
FOSFORO TOTAL (mg P/l)	79.72	1797	float64
FOSFATOS (mg PO4/l)	49.38	1113	float64
CADMIO TOTAL (mg Cd/l)	4.44	100	float64
CROMO TOTAL (mg Cr/l)	8.65	195	float64
NÍQUEL TOTAL (mg Ni/l)	9.14	206	float64
PLOMO TOTAL (mg Pb/l)	7.63	172	float64
CAUDAL (m3/s)	9.85	222	float64

En la Tabla 16 se muestran los resultados del filtrado, incluyendo el porcentaje de completitud, el número de registros válidos y el tipo de dato asociado a cada variable.

Con el fin de garantizar la calidad y representatividad de los análisis, se adoptó un umbral mínimo de 80 % de completitud para las variables consideradas. Este valor se estableció siguiendo las guías de aseguramiento de calidad de datos ambientales de la U.S. Environmental Protection Agency [30], las cuales definen la completitud (completeness) como la proporción de datos válidos obtenidos respecto a los planificados y recomiendan establecer metas explícitas por proyecto.

Asimismo, este criterio es consistente con programas de control y monitoreo de calidad de agua que aplican umbrales operativos de ≥ 80 % para garantizar la fiabilidad estadística y la reproducibilidad de los resultados, como se evidencia en el OARS Water Quality Monitoring Program QAPP [42].

La adopción de este umbral permite reducir sesgos derivados de registros incompletos, asegurar una base muestral suficiente para los cálculos estadísticos y mantener la trazabilidad del proceso analítico. Luego de aplicar el umbral, el conjunto de datos quedó conformado por 27 variables fisicoquímicas, que constituyen la base para el desarrollo de las siguientes fases metodológicas. Estas variables conservan un número suficiente de registros válidos para permitir su análisis y modelado, manteniendo representatividad en la calidad del agua del Río Cauca.

Tabla 17. Resultado después de aplicar el filtro de completitud

pH	TEMPERATURA (°C)	DEMANDA BIOQUÍMICA DE OXIGENO (mg O2/l)	DEMANDA QUÍMICA DE OXIGENO (mg O2/l)	OXIGENO DISUELTO (mg O2/l)	...	COLIFORMES FECALES (NMP/100 ml)
7.1	4.1	4.2	NaN	1.5	...	NaN
7	2	3	NaN	1.81	...	NaN
7	22.9	5	NaN	2	...	NaN
6.6	NaN	0.5	5.2	5.6	...	23
6.7	NaN	2.1	24	6	...	NaN
5 rows × 27 columns						

En la Tabla 17 se muestra una vista de las primeras filas del conjunto resultante, donde se observan variables objetivo como pH y DBO, junto con parámetros complementarios tales como turbiedad, oxígeno disuelto, conductividad eléctrica, sulfatos y coliformes fecales, entre otros.

Con las actividades desarrolladas en esta primera fase se logró consolidar un conjunto de datos histórico del Río Cauca depurado, estandarizado y filtrado bajo criterios de calidad. A partir de un total inicial de 54 variables, se eliminaron aquellas sin información y se convirtieron todas a formato numérico, obteniendo 39 variables válidas. Posteriormente, al aplicar el umbral de completitud del 80%, se conformó un dataset final de 27 variables fisicoquímicas, dentro de las cuales permanecen las dos variables objetivo de este estudio: pH y DBO.

El proceso incluyó la integración del diccionario de variables, el cálculo de estadísticas descriptivas y la identificación de parámetros con alta dispersión, como turbiedad y color, que aportan información relevante sobre la variabilidad del sistema. De este modo, la Fase 1 deja como resultado un conjunto de datos confiable y listo para ser explorado en mayor profundidad en la Fase 2, orientada al análisis exploratorio y la selección de predictores significativos.

Con este conjunto de variables seleccionadas, el estudio está en condiciones óptimas para avanzar hacia el análisis exploratorio y la identificación de predictores significativos para los modelos de calidad de agua, asegurando robustez y representatividad estadística.

4. ANÁLISIS EXPLORATORIO Y SELECCIÓN DE VARIABLES

En la primera etapa del análisis exploratorio se evaluó la presencia de valores faltantes en las variables seleccionadas. En la Figura 1 se presenta el mapa de valores nulos correspondiente al conjunto de datos antes de la imputación; el color amarillo indica ausencia de datos y el púrpura representa registros válidos.

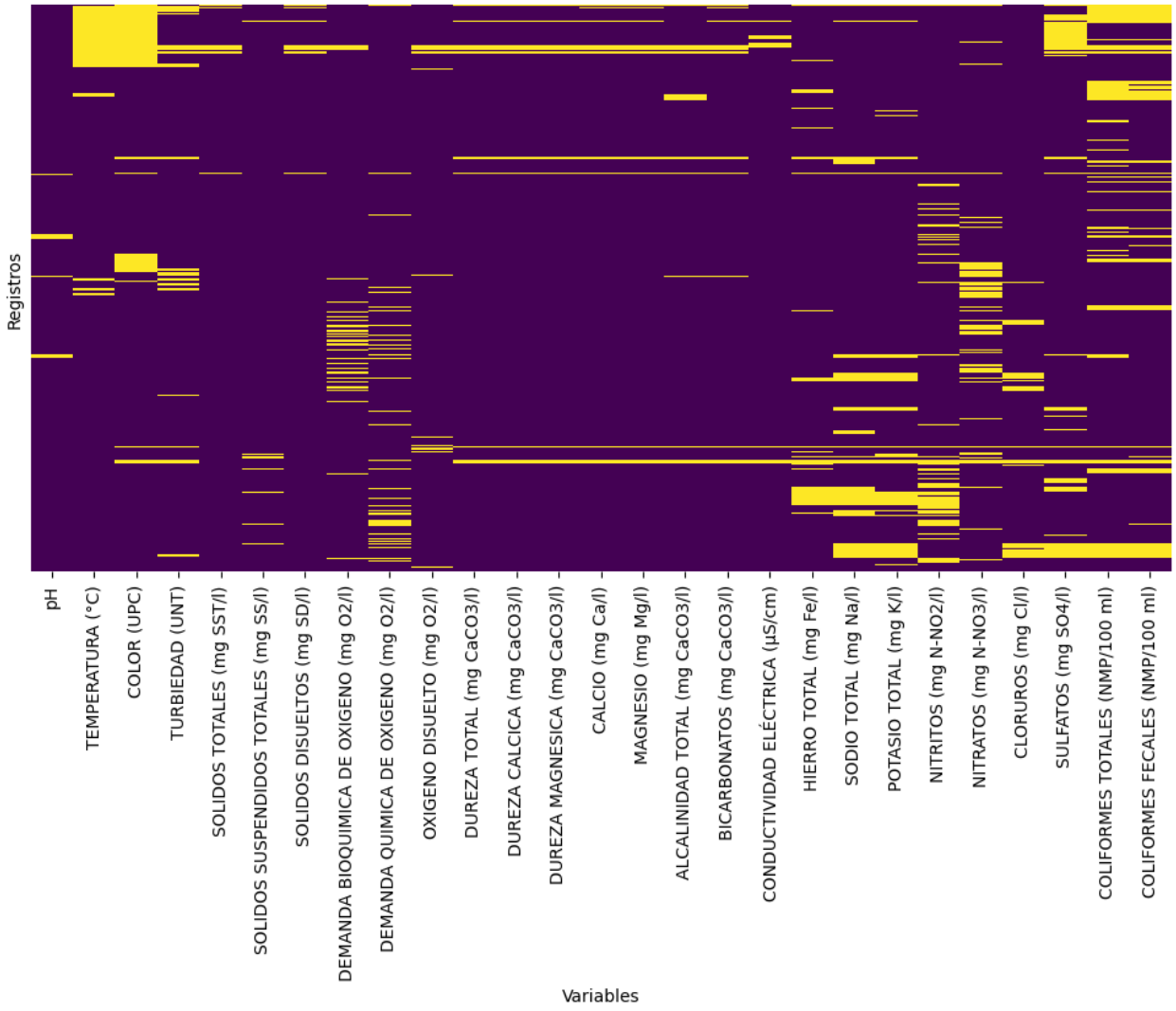

Figura 1. Mapa de calor de valores nulos

Este diagnóstico permitió identificar que variables como Color, Sólidos Disueltos, Nitratos, Coliformes y Demanda Bioquímica de Oxígeno (DBO) presentan vacíos considerables, mientras que parámetros como pH, Conductividad Eléctrica y Dureza Total mantienen registros más completos. Esta visualización evidencia la necesidad de aplicar técnicas de imputación durante el

procesamiento previo al modelado. Para evitar fuga de información, dicha imputación se realizó posterior a la partición entre entrenamiento y prueba, ajustando el imputador con el conjunto de entrenamiento y aplicándolo posteriormente al conjunto de prueba, de esta manera el procedimiento conserva la independencia del conjunto de evaluación y reduce el sesgo derivado de los datos faltantes.

Por ello, se realiza una imputación general en la base de datos mediante el algoritmo K-Nearest Neighbors (KNN) con un parámetro de $k = 5$. Esta configuración se escogió para capturar la similitud entre observaciones evitando la sobre-suavización que podría perder diferenciación. De esta forma, el modelo tiende a generar imputaciones representativas de la estructura original de los datos [30].

La elección de k=5 responde a recomendaciones metodológicas que buscan un equilibrio óptimo: evitar imputaciones poco representativas con valores bajos de k, y prevenir la pérdida de diferenciación con valores excesivos. De este modo, el conjunto de datos queda listo para avanzar en el análisis exploratorio, selección de variables y construcción de modelos predictivos con confianza en la integridad de la información.

El mapa posterior muestra la completa ausencia de valores faltantes en las variables seleccionadas, lo que confirma la efectividad del método de imputación. Este procedimiento preserva la estructura y coherencia estadística del dataset, estimando los valores ausentes a partir

de vecinos reales, y facilitando así análisis más robustos y confiables en etapas posteriores.

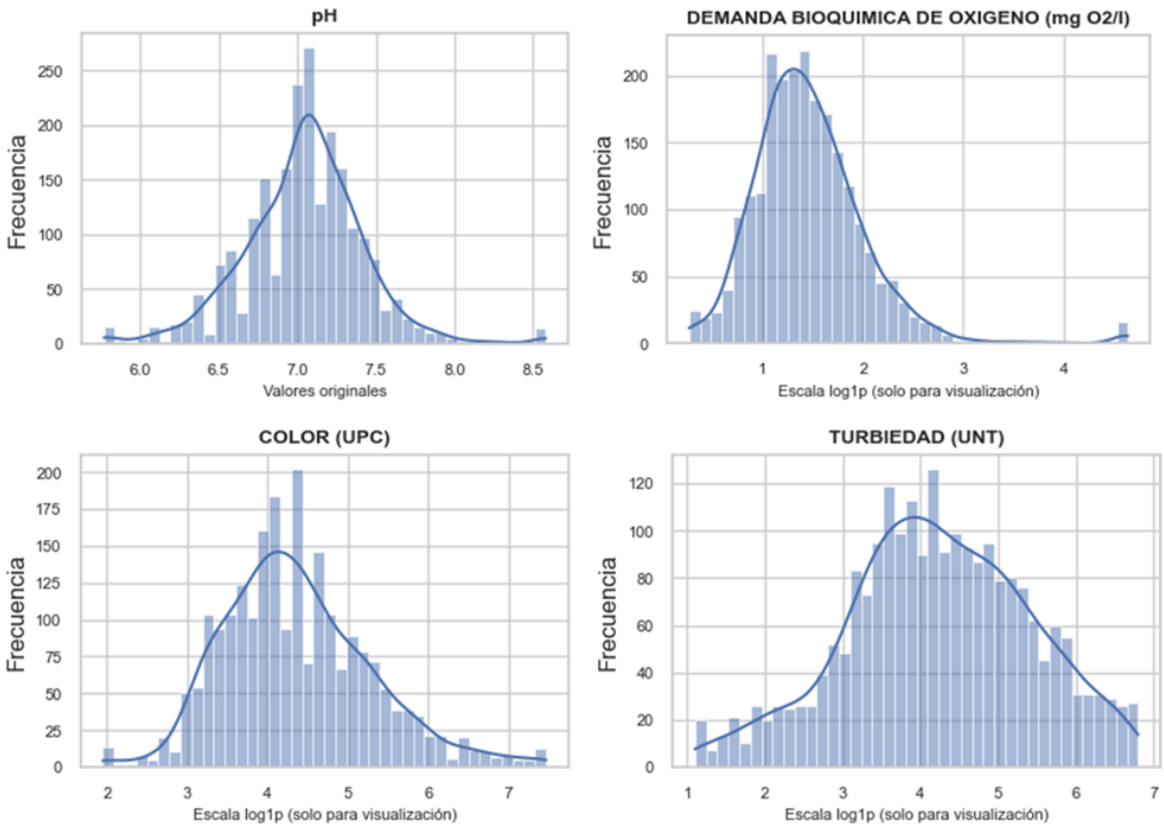

Figura 2. Distribuciones de variables fisicoquímicas con ajuste logarítmico y visualización para el Río Cauca

En la Figura 2 se presentan las distribuciones de las variables fisicoquímicas pH, Demanda Bioquímica de Oxígeno (DBO), Color y Turbidez, visualizadas tras aplicar ajustes logarítmicos y procedimientos de winsorización con el fin de reducir el efecto de valores extremos. Las distribuciones de las demás variables analizadas se incluyen en el ANEXO C. Graficas de histogramas con asimetría. Este análisis permite identificar patrones característicos de cada variable y posibles sesgos en su comportamiento.

Se observa que el pH mantiene una distribución relativamente simétrica, concentrada alrededor de valores cercanos a 7, lo que refleja condiciones estables dentro de los rangos naturales del agua superficial. En contraste, la DBO presenta una distribución sesgada hacia valores altos, indicativa de posibles episodios de contaminación orgánica en determinados registros. Por su parte, Color y Turbidez exhiben colas largas, lo que sugiere la presencia de sólidos y materia orgánica en suspensión con variaciones entre estaciones.

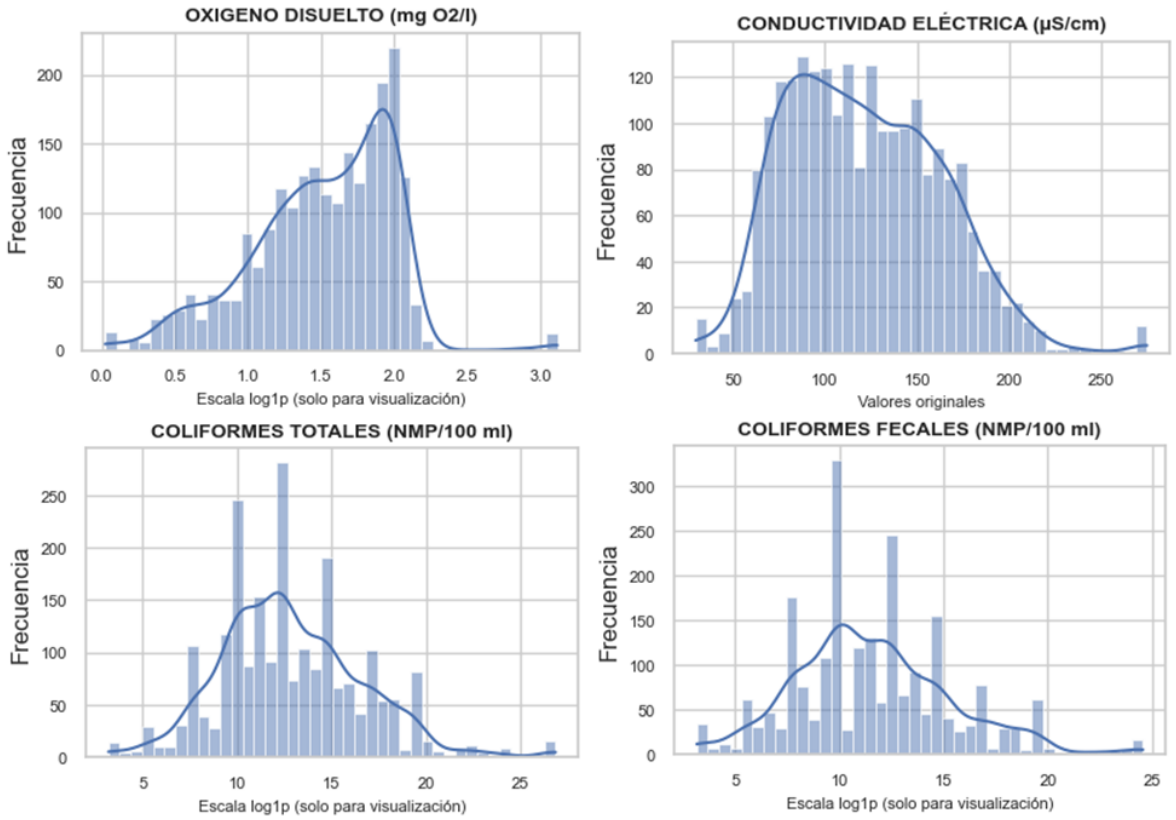

Figura 3. Distribuciones de variables fisicoquímicas y microbiológicas con ajuste logarítmico y visualización de Oxígeno Disuelto, Conductividad Eléctrica, Coliformes Totales y Coliformes Fecales en el Río Cauca

En la Figura 3 se presentan las distribuciones de los parámetros Oxígeno Disuelto y Conductividad Eléctrica, los cuales exhiben comportamientos sesgados, posiblemente asociados a descargas puntuales o a condiciones naturales de ciertos tramos del río. Por su parte, las variables microbiológicas (Coliformes Totales y Coliformes Fecales) muestran una alta dispersión y picos pronunciados, lo que refleja la presencia intermitente de fuentes de contaminación fecal.

El análisis de estas distribuciones constituye un primer acercamiento a la variabilidad y los posibles sesgos presentes en los datos. Sin embargo, para lograr una caracterización más precisa de los valores extremos y comprender su comportamiento entre estaciones, se recurre a representaciones complementarias, como los diagramas de caja (boxplots), incluidos en el ANEXO D. Diagramas de cajas por variable.

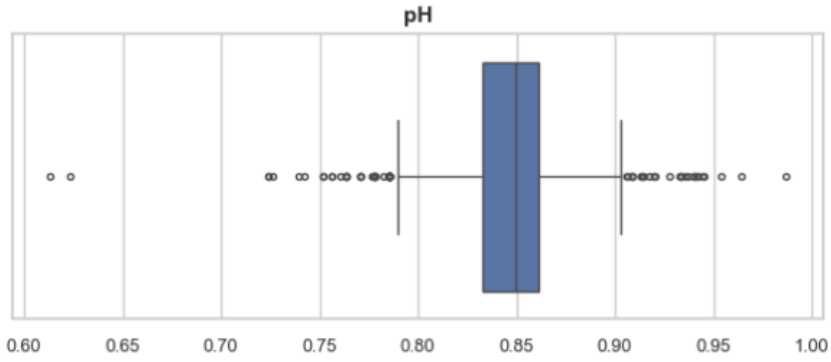

Figura 4. Distribución y valores atípicos de la variable pH en el Río Cauca

En la Figura 4 se observa la distribución del pH, la cual se concentra alrededor de valores normalizados cercanos a 0.85, con un rango intercuartílico estrecho y presencia limitada de valores atípicos en ambos extremos. Este comportamiento evidencia una alta estabilidad del pH en el conjunto de datos, reflejando condiciones químicas estables y dentro de los rangos naturales del agua superficial del Río Cauca. Los pocos valores extremos identificados podrían asociarse a variaciones locales o a procesos de mezcla en determinados puntos del cauce. El pH actúa como parámetro de referencia fisicoquímica, permitiendo contrastar la variabilidad observada en las demás variables analizadas.

En la Figura 5 se muestra la distribución de la Demanda Bioquímica de Oxígeno (DBO), caracterizada por un sesgo positivo y una dispersión mayor en comparación con el pH. La mediana se ubica alrededor de 0.5 en la escala normalizada, con una concentración principal de valores bajos y una cola extendida hacia la derecha. Este comportamiento indica la presencia de eventos puntuales de incremento en la carga orgánica, posiblemente relacionados con descargas domésticas, vertimientos industriales o acumulación de materia orgánica biodegradable. Los valores atípicos altos refuerzan la existencia de episodios irregulares de contaminación, mientras que la mayoría de las observaciones se mantienen dentro de rangos moderados. La DBO constituye un indicador sensible de la presión antrópica sobre la calidad del agua, complementando el equilibrio fisicoquímico representado por el pH.

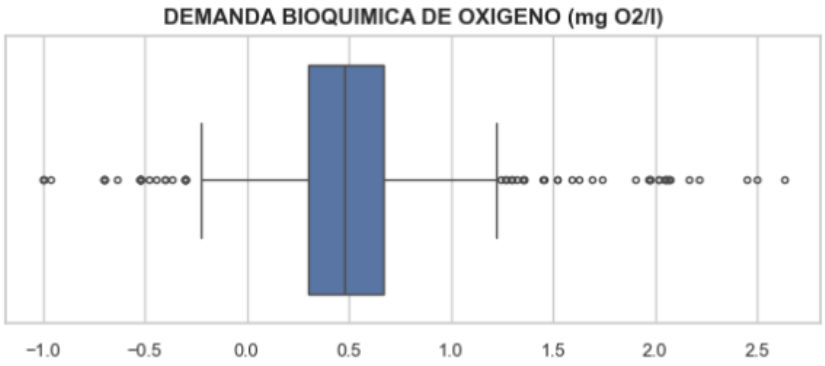

Figura 5. Distribución y valores atípicos de la variable Demanda Bioquímica de Oxígeno (DBO) en el Río Cauca

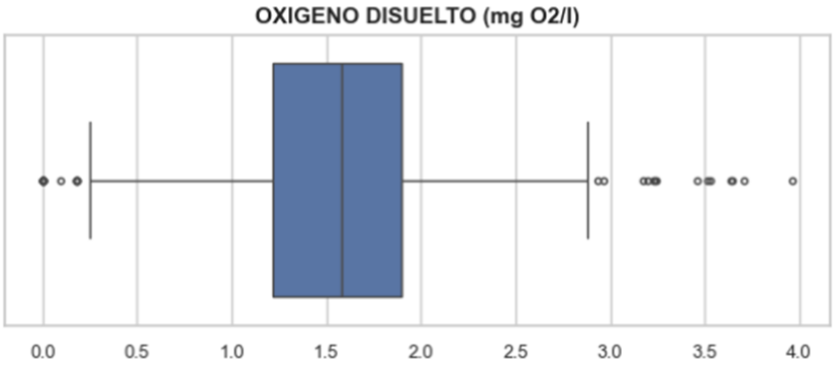

Figura 6. Distribución y valores atípicos de la variable Oxígeno Disuelto (OD) en el Río Cauca

En la Figura 6 se presenta la distribución del Oxígeno Disuelto (OD), la cual muestra una mediana cercana a 1.6 en la escala normalizada y una dispersión moderada con ligera asimetría hacia valores bajos. Este comportamiento sugiere la existencia de periodos o zonas con menor disponibilidad de oxígeno, posiblemente asociados a procesos de descomposición de materia orgánica o a reducida turbulencia en ciertos tramos del río. Se observan algunos valores atípicos altos y bajos, lo que refleja fluctuaciones locales en la dinámica de oxigenación del sistema. En términos ecológicos, esta variabilidad indica un equilibrio dinámico entre los procesos de consumo y producción de oxígeno, estrechamente relacionado con los patrones de Demanda Bioquímica de Oxígeno (DBO) analizados previamente.

En la Figura 7 se observa la distribución de la Turbidez (UNT), la cual presenta una mediana cercana a 1.8 en la escala normalizada y una dispersión amplia, con una caja más extendida en comparación con las variables analizadas previamente. Este patrón indica una alta variabilidad en la presencia de sólidos suspendidos, probablemente asociada a fenómenos de escorrentía, erosión de márgenes y aportes de sedimentos en periodos de lluvia. Los valores atípicos identificados hacia ambos extremos reflejan episodios puntuales de incremento o reducción de partículas en suspensión, lo que sugiere cambios estacionales o localizados en la dinámica del cauce. La Turbidez se configura como un indicador clave de la estabilidad hidrosedimentaria,

complementando la interpretación del pH, la DBO y el Oxígeno Disuelto en el diagnóstico general de calidad del agua.

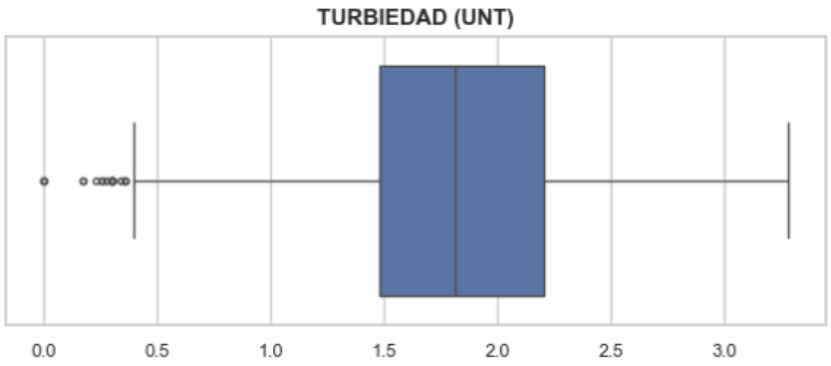

Figura 7. Distribución y valores atípicos de la variable Turbidez (UNT) en el Río Cauca

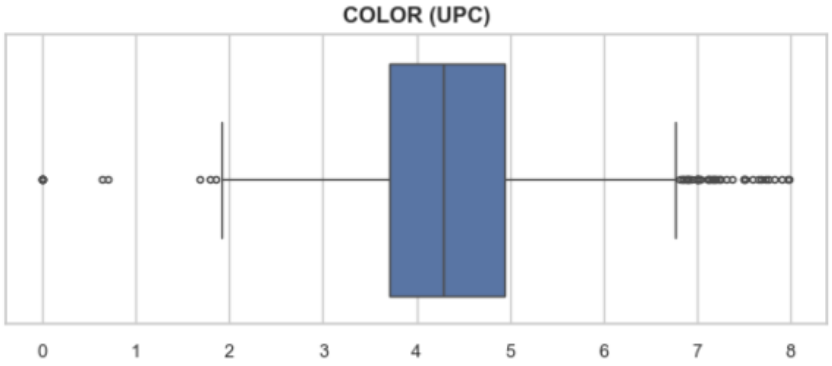

Figura 8. Distribución y valores atípicos de la variable Color (UPC) en el Río Cauca

En la Figura 8 se presenta la distribución de la variable Color (UPC), caracterizada por una mediana próxima a 4.3 en la escala normalizada y una dispersión amplia con presencia de múltiples valores atípicos, especialmente hacia el extremo superior. Este comportamiento indica una variabilidad significativa en la concentración de materia orgánica disuelta y compuestos metálicos, que inciden directamente en el color aparente del agua. Los valores elevados podrían asociarse a procesos de lixiviación de suelos, descargas de aguas residuales o presencia de taninos derivados de la vegetación ribereña. En contraste, los valores bajos reflejan condiciones de mayor transparencia y menor aporte orgánico. El Color complementa la información aportada por la Turbidez, ya que ambos parámetros describen dimensiones distintas (disuelta y particulada) de la carga orgánica y sedimentaria del sistema fluvial.

La identificación y caracterización de los valores atípicos (outliers) resulta fundamental para el análisis predictivo, ya que aportan información crítica sobre episodios extremos en la calidad del agua. En lugar de eliminarse, estos valores se conservaron visibles en esta etapa, dado que forman parte del comportamiento natural del sistema y permiten comprender mejor la dinámica de los

eventos de contaminación.

Tabla 18. Análisis complementario para Nitritos y Conductividad Eléctrica.

VARIABLE	Q1	MEDIANA	Q3	MAX	ASIMETRÍA	CURTOSIS	OUTLIERS (%)
Nitritos (mg N-NO2/l)	0.01	0.03	0.09	150000	41.43	1810.84	19.79
Conductividad Eléctrica (µS/cm)	87.5	117	151	4259	28.37	1052.21	0.62

En la Tabla 18 se presentan los resultados del análisis complementario para las variables Nitritos (mg N-NO_2/l) y Conductividad Eléctrica (µS/cm), seleccionadas por su elevado grado de dispersión y asimetría. Ambas variables mostraron comportamientos altamente sesgados, con presencia de colas largas y picos pronunciados que evidencian la existencia de registros extremos.

Los Nitritos exhiben una asimetría muy marcada (skewness = 41.43) y una curtosis extrema (kurtosis = 1810.84), lo que confirma su tendencia a concentrar valores cercanos a cero y, simultáneamente, a generar episodios con concentraciones inusualmente elevadas. Se identificaron 446 valores atípicos, equivalentes al 19.79% del total de registros, lo que demuestra una fuerte inestabilidad asociada a procesos de contaminación nitrogenada.

Por su parte, la Conductividad Eléctrica, aunque también presenta una asimetría considerable (skewness = 28.37) y una curtosis alta (kurtosis = 1052.21), mostró un número significativamente menor de atípicos (14 registros, equivalentes al 0.62% del total). Esto sugiere que la mayoría de los valores permanecen dentro de un rango más estable, con desviaciones extremas vinculadas a eventos puntuales de aporte salino o mineralización.

Estos resultados resaltan la importancia de tratar de forma diferenciada las variables problemáticas, ya que, mientras los Nitritos reflejan patrones recurrentes de inestabilidad, la Conductividad Eléctrica se asocia principalmente a anomalías aisladas que no alteran sustancialmente la estructura general de la distribución.

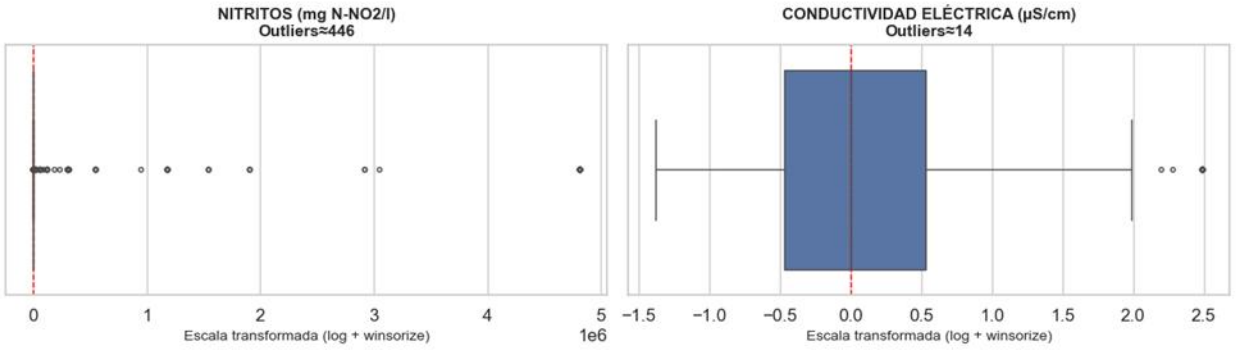

Figura 9. Distribución refinada de variables problemáticas (Nitritos y Conductividad Eléctrica) con ajuste logarítmico y winsorización en el Río Cauca

En la Figura 9 se muestran las distribuciones refinadas de dos variables que presentaron un comportamiento especialmente problemático: Nitritos (mg N-NO2/l) y Conductividad Eléctrica (µS/cm)[5]. Para ambas se aplicaron transformaciones logarítmicas combinadas con winsorización, lo que permitió atenuar parcialmente la dispersión y resaltar la magnitud de los valores atípicos.

En el caso de los Nitritos, se identificaron aproximadamente 446 outliers, lo que equivale al 19.79% de los registros. La concentración de valores cercanos a cero y la aparición de observaciones extremadamente elevadas reflejan un patrón de alta inestabilidad y episodios críticos de contaminación nitrogenada, que deben ser tratados con especial cuidado en el modelado posterior.

Por su parte, la Conductividad Eléctrica presentó únicamente 14 outliers (0.62%), lo que indica que la mayor parte de los registros se concentra en un rango estable, con algunos picos aislados posiblemente asociados a descargas puntuales de sales y minerales disueltos. En este caso, la variable conserva mayor robustez estadística, y los valores extremos no alteran de manera sustancial su distribución central, por lo que se mantiene en el conjunto de predictores.

La comparación entre ambas variables evidencia la necesidad de adoptar enfoques diferenciados: mientras que en los nitritos los outliers forman parte de la dinámica recurrente del sistema y reflejan un proceso ambiental relevante, en la conductividad corresponden más a eventos esporádicos que pueden ser considerados como anomalías de menor impacto.

Una vez caracterizados los valores atípicos y diferenciada su naturaleza para cada parámetro, el siguiente paso consistió en examinar las relaciones internas entre las variables del conjunto de estudio. Para ello, se calcularon matrices de correlación, con el propósito de identificar asociaciones lineales y no lineales relevantes para el modelado posterior.

Tabla 19. Pares de variables altamente correlacionadas según coeficientes de Pearson y Spearman en el Río Cauca.

Columnas usadas para correlación entre predictores: 25 Objetivos excluidos del mapa entre predictores: ['DEMANDA BIOQUÍMICA DE OXIGENO (mg O2/l)', 'pH']			
Pares altamente correlacionados (Pearson):			
	VARIABLE_1	VARIABLE_2	ρ[6]
285	NITRITOS (mg N-NO2/l)	NITRATOS (mg N-NO3/l)	0.9992
69	SOLIDOS TOTALES (mg SST/l)	SOLIDOS SUSPENDIDOS TOTALES	0.9459

[5] Las variables Nitritos (mg N-NO2/l) y Conductividad Eléctrica (µS/cm) fueron seleccionadas para un análisis detallado debido a que, en el diagnóstico preliminar de distribuciones y boxplots, presentaron los mayores niveles de dispersión y sesgo. Esta característica las convierte en casos representativos para ilustrar el tratamiento de valores atípicos y su impacto en la consistencia del dataset.

[6] Correlación de Spearman

		(mg SS/l)	
209	CALCIO (mg Ca/l)	MAGNESIO (mg Mg/l)	0.8898
196	DUREZA MAGNÉSICA (mg CaCO3/l)	MAGNESIO (mg Mg/l)	0.8623
Pares altamente correlacionados (Spearman):			
	VARIABLE_1	VARIABLE_2	ρ
196	DUREZA MAGNÉSICA (mg CaCO3/l)	MAGNESIO (mg Mg/l)	0.9892
181	DUREZA CÁLCICA (mg CaCO3/l)	CALCIO (mg Ca/l)	0.9597
234	ALCALINIDAD TOTAL (mg CaCO3/l)	BICARBONATOS (mg CaCO3/l)	0.9332
299	COLIFORMES TOTALES (NMP/100 ml)	COLIFORMES FECALES (NMP/100 ml)	0.8907
69	SOLIDOS TOTALES (mg SST/l)	SOLIDOS SUSPENDIDOS TOTALES (mg SS/l)	0.8637
164	DUREZA TOTAL (mg CaCO3/l)	DUREZA CÁLCICA (mg CaCO3/l)	0.8178
165	DUREZA TOTAL (mg CaCO3/l)	DUREZA MAGNÉSICA (mg CaCO3/l)	0.8040

Los resultados de este análisis se sintetizan en la Tabla 19, que presenta los pares de variables con correlación significativamente alta según los coeficientes de Pearson y Spearman, tras excluir las variables objetivo pH y DBO del análisis entre predictores.

El hallazgo más contundente lo entregó Pearson: Nitritos y Nitratos mostraron una correlación prácticamente perfecta (r = 0.999), lo que en términos estadísticos equivale a decir que ambas variables miden exactamente lo mismo desde el punto de vista del modelo. Incluir las dos sería como pesar una maleta dos veces y esperar resultados distintos. En esa misma línea, los Sólidos Totales y los Sólidos Suspendidos Totales compartieron una correlación de r = 0.946, mientras que Calcio y Magnesio —ambos componentes estructurales de la dureza del agua— registraron valores entre 0.86 y 0.89, reflejo directo de su origen geológico común.

Spearman, más robusto ante valores extremos, reforzó este panorama y amplió el mapa de redundancias. Las correlaciones más fuertes involucraron a las variables de dureza: Dureza Magnésica con Magnesio (ρ = 0.989), Dureza Cálcica con Calcio (ρ = 0.960) y Alcalinidad Total con Bicarbonatos (ρ = 0.933). La razón es química antes que estadística: estas variables no son independientes en la naturaleza —unas son fracciones o derivaciones directas de las otras—, por lo que su comportamiento conjunto en el río Cauca era esperable. La relación entre Coliformes Totales y Coliformes Fecales (ρ = 0.891) siguió la misma lógica: los fecales son un subconjunto de los totales, y el río no distingue entre ellos a la hora de contaminar.

Estos resultados evidenciaron la necesidad de aplicar criterios formales de selección de variables para depurar la redundancia antes del modelado. En particular, la altísima correlación entre Nitritos y Nitratos —sumada a la inestabilidad estadística de los Nitritos documentada en el análisis previo de distribuciones— respaldó con doble argumento la decisión de eliminar los Nitritos como predictor: no solo eran redundantes, sino también problemáticos.

Este análisis preliminar condujo a un refinamiento adicional mediante la visualización de las

matrices de correlación de Pearson y Spearman, que permitieron observar de manera más clara los patrones de redundancia entre predictores y sustentar con evidencia gráfica las decisiones de exclusión de variables.

La Figura 10 traduce los coeficientes numéricos de la Tabla 19 en un lenguaje visual: cuanto más intenso el rojo, mayor la correlación positiva entre dos variables; cuanto más azul, más fuerte la relación inversa. El blanco es silencio estadístico — dos variables que no se escuchan entre sí.

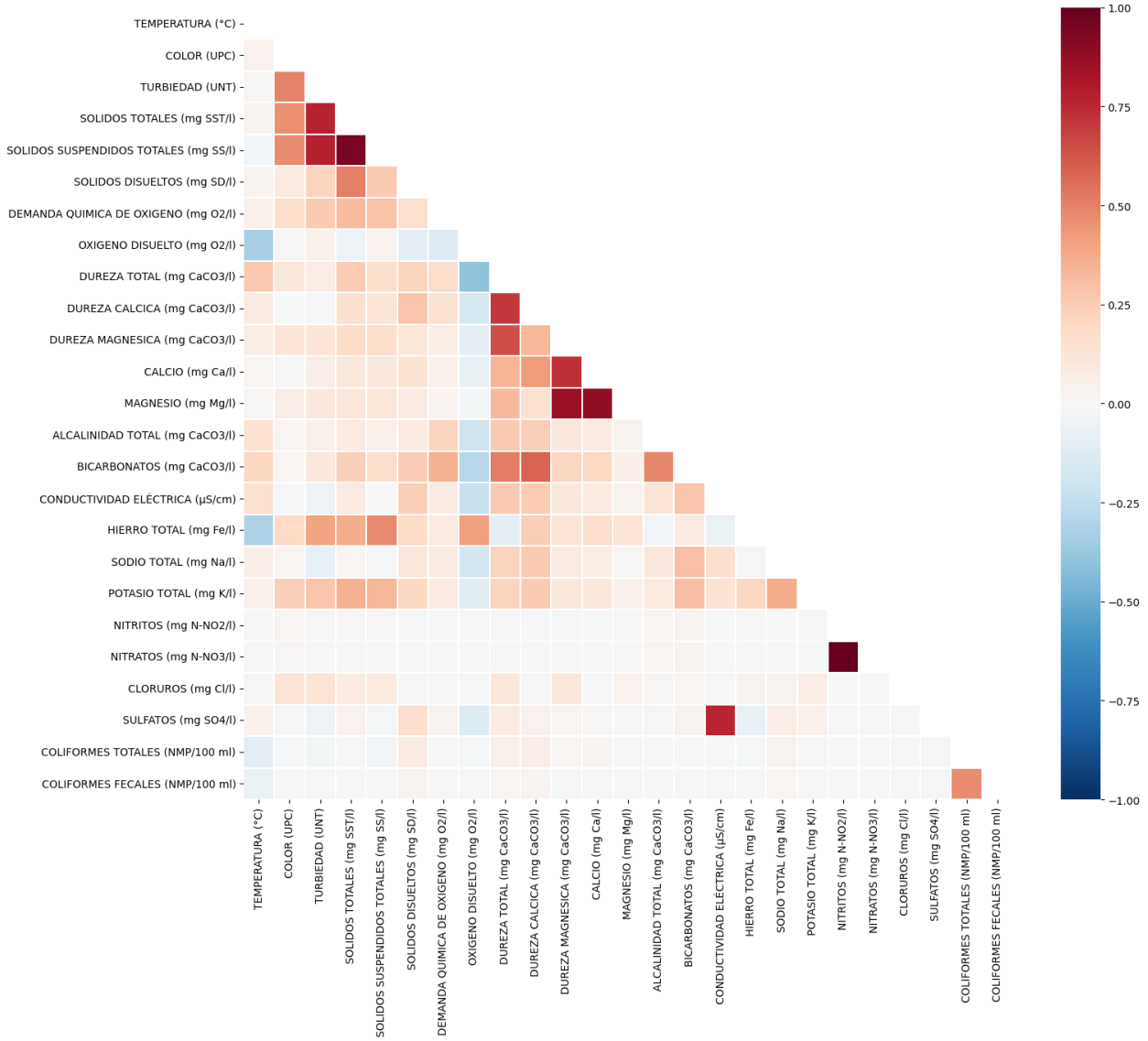

Figura 10. Matriz de correlación de Pearson entre predictores fisicoquímicos del Río Cauca

El primer patrón que salta a la vista es el cuadro rojo oscuro en la esquina inferior derecha: la correlación entre Nitritos y Nitratos (r ≈ 0.999). Es el valor más alto de toda la matriz y el más fácil

de interpretar químicamente — ambas variables son formas distintas de la misma cadena de transformación del nitrógeno en el agua, por lo que medir una es, en términos prácticos, medir la otra.

En la zona superior izquierda se distingue otro bloque cálido: Sólidos Totales y Sólidos Suspendidos Totales ($r \approx 0.946$). Su redundancia también es lógica, ya que los sólidos suspendidos son una fracción de los totales, y en el Río Cauca representan la parte dominante de esa suma.

Hacia el centro de la matriz, el grupo conformado por Dureza Cálcica, Dureza Magnésica, Calcio y Magnesio forma un bloque de tonos rojizos (r entre 0.86 y 0.96), evidenciando que la dureza del agua está directamente escrita en la concentración de estos dos iones. De manera similar, Alcalinidad Total y Bicarbonatos ($r \approx 0.933$) aparecen como un par inseparable, lo cual refleja que los bicarbonatos son el componente principal de la capacidad buffer del agua.

En contraste, el Oxígeno Disuelto muestra celdas ligeramente azuladas frente a varios parámetros orgánicos y de sólidos, anticipando su comportamiento inverso frente a la carga contaminante — una señal que se confirmará con mayor claridad en la correlación con la DBO.

La Figura 11 presenta la misma arquitectura visual que la anterior, pero con una diferencia fundamental: aquí el color ya no mide linealidad sino monotonía. El morado intenso indica que cuando una variable sube, la otra también lo hace de forma consistente — aunque no necesariamente a la misma velocidad ni en proporción constante. El naranja señala la dirección contraria. Y es precisamente en esos matices donde Spearman aporta información que Pearson no captura.

Los grandes bloques morados del grupo de dureza — Dureza Cálcica, Dureza Magnésica, Calcio y Magnesio — se mantienen igual de sólidos que en Pearson, confirmando que estas redundancias son robustas independientemente del método utilizado. Lo mismo ocurre con Alcalinidad Total y Bicarbonatos, cuya asociación se sostiene bajo ambos coeficientes.

La diferencia más llamativa aparece en la esquina inferior derecha: el cuadro oscuro de Coliformes Totales y Coliformes Fecales ($\rho \approx 0.891$) es visualmente más prominente en Spearman que en Pearson. Esto indica que la relación entre ambos indicadores microbiológicos es monótonamente fuerte, aunque no perfectamente lineal — algo esperable en variables biológicas con alta variabilidad y distribuciones asimétricas.

El otro hallazgo que vale la pena destacar es el patrón naranja del Oxígeno Disuelto frente a varios parámetros de carga orgánica y sólidos. Spearman lo hace más visible: a medida que la contaminación orgánica aumenta, el oxígeno disponible decrece de forma sistemática, incluso cuando esa relación no es estrictamente lineal. Este comportamiento es ecológicamente coherente y refuerza la relevancia del Oxígeno Disuelto como predictor clave en el modelado posterior.

En conjunto, Spearman no contradice a Pearson — lo complementa. Donde Pearson detecta linealidad, Spearman confirma monotonía; y donde Pearson subestima, Spearman expone. Por

eso se adoptó como criterio principal para el filtro de correlaciones en las etapas siguientes.

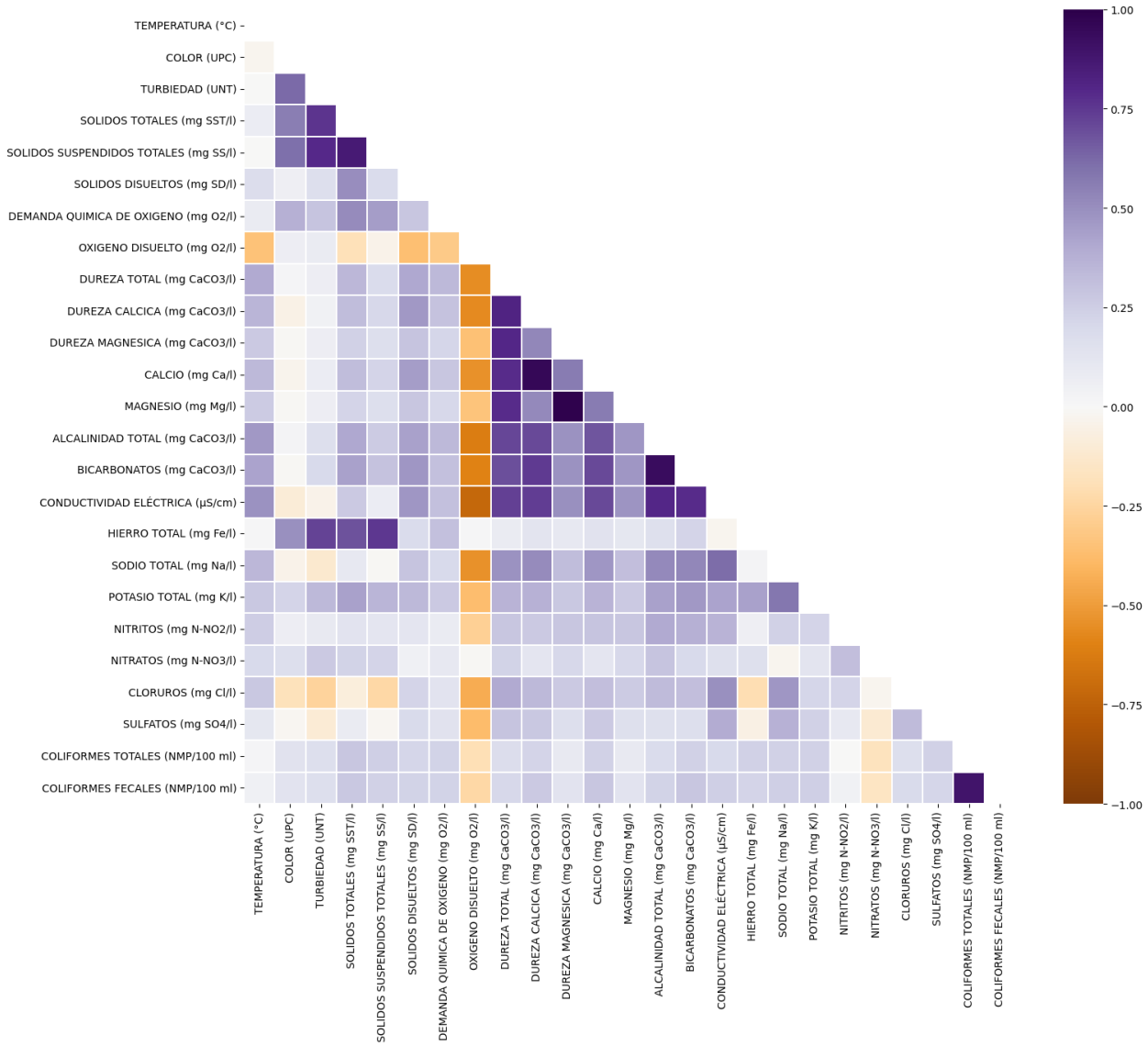

Figura 11. Matriz de correlación de Spearman entre predictores fisicoquímicos del Río Cauca

La elección de Spearman como coeficiente principal no fue arbitraria — respondió a las características propias de los datos del Río Cauca. Las variables fisicoquímicas analizadas presentan distribuciones con marcada asimetría, colas largas y una proporción considerable de valores atípicos, condiciones que vulneran los supuestos de normalidad y linealidad que Pearson requiere para ser confiable. Cuando esos supuestos no se cumplen, el coeficiente de Pearson puede inflar o subestimar correlaciones reales, porque trabaja con los valores absolutos de los

datos y es sensible a los extremos.

Spearman, en cambio, opera sobre los rangos. No le importa cuánto vale un dato, sino su posición relativa dentro del conjunto. Esa propiedad lo hace robusto frente a outliers y válido sin necesidad de asumir normalidad, lo que lo convierte en una elección metodológicamente más honesta para sistemas ambientales como este [31].

Una vez analizada la estructura interna de correlaciones entre predictores, el análisis dio un giro estratégico: ¿cuáles variables explican mejor el comportamiento de la Demanda Bioquímica de Oxígeno? La Tabla 20 presenta los diez predictores con mayor correlación de Spearman frente a la DBO, y el resultado es revelador tanto por lo que encabeza la lista como por lo que aparece en sentido contrario.

Tabla 20. Correlación de Spearman entre los predictores y la DBO (objetivo) top 10

VARIABLE	Spearman DBO
BICARBONATOS (mg CaCO3/l)	0.5269
CONDUCTIVIDAD ELÉCTRICA (µS/cm)	0.5007
OXIGENO DISUELTO (mg O2/l)	-0.4975
ALCALINIDAD TOTAL (mg CaCO3/l)	0.4884
CALCIO (mg Ca/l)	0.4608
DUREZA CÁLCICA (mg CaCO3/l)	0.4584
SODIO TOTAL (mg Na/l)	0.3813
DUREZA TOTAL (mg CaCO3/l)	0.3575
COLIFORMES FECALES (NMP/100 ml)	0.3548
POTASIO TOTAL (mg K/l)	0.3360

Los Bicarbonatos (ρ = 0.527) y la Conductividad Eléctrica (ρ = 0.501) lideran las correlaciones positivas, lo que sugiere que la carga de sales disueltas y la capacidad buffer del agua tienden a crecer junto con la materia orgánica biodegradable. Una interpretación posible es que los vertimientos que elevan la DBO —aguas residuales, escorrentía orgánica— también aportan minerales y sales que incrementan la conductividad y alteran el equilibrio carbonatado del río. La Alcalinidad Total (ρ = 0.488), el Calcio (ρ = 0.461) y la Dureza Cálcica (ρ = 0.458) refuerzan esta lectura desde la composición iónica del agua.

El tercer valor de la tabla, sin embargo, es el más informativo de todos: el Oxígeno Disuelto con ρ = −0.498. Su signo negativo cuenta una historia que cualquier ecólogo reconocería de inmediato — cuando la materia orgánica aumenta y la DBO sube, el oxígeno disponible para la vida acuática cae. No es una coincidencia estadística; es el mecanismo central de la contaminación orgánica en ríos. Este resultado confirma al Oxígeno Disuelto como el predictor con mayor carga ecológica del conjunto.

Finalmente, la presencia de Coliformes Fecales (ρ = 0.355) entre los diez primeros establece el

vínculo entre la contaminación de origen fecal y el incremento de la carga orgánica, cerrando un ciclo coherente: más aguas residuales, más bacterias, menos oxígeno, más DBO.

Con el mapa de predictores para la DBO definido, el análisis se desplazó hacia la segunda variable objetivo: el pH. El contraste entre ambas tablas es, en sí mismo, un hallazgo — porque el pH no responde de la misma manera.

La Tabla 21 presenta los diez predictores con mayor correlación de Spearman frente al pH, y lo primero que llama la atención es la magnitud de los valores: el coeficiente más alto apenas alcanza ρ = 0.306, correspondiente a los Nitratos. En comparación con la DBO, donde los Bicarbonatos llegaban a ρ = 0.527, el pH aparece como una variable notablemente más esquiva — ningún predictor individual la domina de forma clara.

Tabla 21. Correlación de Spearman entre los predictores y la pH (objetivo) top 10

VARIABLE	Spearman pH
NITRATOS (mg N-NO3/l)	0.3060
ALCALINIDAD TOTAL (mg CaCO3/l)	0.2852
BICARBONATOS (mg CaCO3/l)	0.2433
TEMPERATURA (°C)	0.2354
DUREZA CÁLCICA (mg CaCO3/l)	0.1725
NITRITOS (mg N-NO2/l)	0.1703
CALCIO (mg Ca/l)	0.1640
CONDUCTIVIDAD ELÉCTRICA (µS/cm)	0.1583
DUREZA TOTAL (mg CaCO3/l)	0.1534
SULFATOS (mg SO4/l)	-0.1481

Esto no es una limitación del análisis; es una propiedad química del sistema. El pH del agua es el resultado de múltiples equilibrios simultáneos — carbonatados, biológicos, geológicos, hidrológicos — que se compensan entre sí de manera dinámica. Ninguna variable sola puede capturar esa complejidad, lo que explica por qué las correlaciones son moderadas en el mejor de los casos.

Aun así, la estructura de la tabla es coherente. El sistema carbonatado vuelve a aparecer en los primeros lugares — Alcalinidad Total (ρ = 0.285), Bicarbonatos (ρ = 0.243) — porque estos compuestos son precisamente los reguladores naturales del pH en aguas superficiales. Su presencia indica que el río tiene cierta capacidad de amortiguar cambios de acidez, y que cuando esa capacidad se expande, el pH tiende a subir moderadamente. La Temperatura (ρ = 0.235) también aparece, lo que refleja la dependencia térmica de los equilibrios de disolución de gases y reacciones ácido-base en el agua.

El único valor negativo de la tabla es el de los Sulfatos (ρ = −0.148), cuya asociación inversa con el pH es ambientalmente interpretable: altas concentraciones de sulfatos pueden estar ligadas a

procesos de acidificación, como el drenaje ácido de actividades mineras, que reducen el pH del sistema.

En conjunto, la comparación entre las Tabla 20 y Tabla 21 deja una conclusión metodológica importante: predecir la DBO y predecir el pH son problemas de distinta naturaleza. La DBO tiene predictores relativamente dominantes; el pH exige modelos capaces de integrar señales débiles de múltiples fuentes simultáneamente.

Con las correlaciones entre predictores y variables objetivo ya establecidas, el siguiente paso fue resolver la redundancia interna del conjunto. Mantener variables que esencialmente miden lo mismo no aporta información adicional al modelo — solo introduce ruido y puede distorsionar la interpretación de los coeficientes. Para identificar y eliminar estas redundancias, se aplicó un filtro de colinealidad basado en Spearman, con umbral de $|\rho| \geq 0.85$.

Tabla 22. Pares de variables altamente correlacionadas ($|\rho| \geq 0.85$) según Spearman y variables eliminadas por colinealidad en el Río Cauca.

VARIABLE 1	VARIABLE 2	CORRELACIÓN
DUREZA MAGNÉSICA (mg CaCO3/l)	MAGNESIO (mg Mg/l)	0.99
DUREZA CÁLCICA (mg CaCO3/l)	CALCIO (mg Ca/l)	0.96
ALCALINIDAD TOTAL (mg CaCO3/l)	BICARBONATOS (mg CaCO3/l)	0.93
COLIFORMES TOTALES (NMP/100 ml)	COLIFORMES FECALES (NMP/100 ml)	0.89
SÓLIDOS TOTALES (mg SST/l)	SÓLIDOS SUSPENDIDOS TOTALES (mg SS/l)	0.86
DUREZA TOTAL (mg CaCO3/l)	DUREZA CÁLCICA (mg CaCO3/l)	0.82
DUREZA TOTAL (mg CaCO3/l)	DUREZA MAGNÉSICA (mg CaCO3/l)	0.80

La Tabla 22 presenta los pares problemáticos detectados. Los casos más extremos son también los más interpretables: la Dureza Magnésica y el Magnesio ($\rho = 0.99$) son prácticamente la misma variable expresada en unidades distintas, al igual que la Dureza Cálcica y el Calcio ($\rho = 0.96$). En ambos casos, la dureza es simplemente una forma derivada de expresar la concentración del ion correspondiente. La Alcalinidad Total y los Bicarbonatos ($\rho = 0.93$) siguen la misma lógica: en aguas naturales como el Río Cauca, los bicarbonatos constituyen la fracción dominante de la alcalinidad, por lo que ambas variables cuentan la misma historia química.

Para cada par se aplicó una decisión de conservación basada en la relevancia explicativa frente a las variables objetivo: se eliminaron las variables con menor correlación con DBO y pH, priorizando además aquellas con interpretación ecológica más directa. El resultado fue la eliminación de cinco variables — Dureza Magnésica, Dureza Cálcica, Alcalinidad Total, Coliformes Totales y Sólidos Suspendidos Totales — conservando sus contrapartes más informativas. Tras esta depuración, el conjunto quedó conformado por 20 predictores estadísticamente independientes y ambientalmente coherentes, listos para la siguiente etapa de selección mediante VIF.

El VIF cuantifica cuánto se infla la varianza de los coeficientes de regresión debido a la colinealidad

con el resto de predictores. Un valor cercano a 1 indica independencia casi total; valores superiores a 10 señalan una dependencia tan fuerte que el predictor se vuelve estadísticamente problemático. El umbral adoptado en este estudio fue VIF < 10.

Al ejecutar el análisis por primera vez sobre el conjunto resultante del filtro Spearman, los Nitritos registraron un VIF de 616.9 — un valor extraordinario que reflejaba su dependencia casi total con los Nitratos, ya documentada en el análisis de Pearson (r ≈ 0.999). La decisión fue inmediata: los Nitritos fueron eliminados.

Tabla 23. Resultados del análisis de multicolinealidad mediante VIF (Variance Inflation Factor) y variables conservadas para el modelado del Río Cauca.

Variable	VIF	Variable	VIF
CALCIO (mg Ca/l)	5.69	POTASIO TOTAL (mg K/l)	1.45
MAGNESIO (mg Mg/l)	5.62	COLOR (UPC)	1.44
SOLIDOS TOTALES (mg SST/l)	3.82	SODIO TOTAL (mg Na/l)	1.3
TURBIEDAD (UNT)	3.09	DEMANDA QUÍMICA DE OXIGENO (mg O2/l)	1.27
CONDUCTIVIDAD ELÉCTRICA (µS/cm)	2.93	OXIGENO DISUELTO (mg O2/l)	1.24
SULFATOS (mg SO4/l)	2.57	CLORUROS (mg Cl/l)	1.06
BICARBONATOS (mg CaCO3/l)	1.97	COLIFORMES FECALES (NMP/100 ml)	1.01
DUREZA TOTAL (mg CaCO3/l)	1.85	NITRATOS (mg N-NO3/l)	1
HIERRO TOTAL (mg Fe/l)	1.77	TEMPERATURA (°C)	0.38
SOLIDOS DISUELTOS (mg SD/l)	1.54		

Tras esa depuración, los resultados de la iteración siguiente se presentan en la Tabla 23 El valor máximo de VIF corresponde al Calcio (5.69), seguido del Magnesio (5.62), ambos por debajo del umbral crítico. En el extremo opuesto, variables como Nitratos (VIF = 1.00) y Temperatura (VIF = 0.38) muestran independencia prácticamente total respecto al resto del conjunto.

El mensaje de la tabla es claro: ninguna variable restante genera distorsión estadística en el conjunto. Sin embargo, tener 19 predictores independientes no implica que todos sean igualmente útiles para predecir el pH o la DBO — algunos aportan información relevante, otros simplemente acompañan. Para identificar cuáles realmente importan, se aplicó como siguiente paso la Eliminación Recursiva de Características (RFE), diseñada para evaluar la importancia de cada variable directamente desde la perspectiva del modelo predictivo. A diferencia de un filtro estadístico como el VIF, el RFE no mide dependencia entre predictores sino contribución explicativa real frente a cada variable objetivo.

Para que esta selección no dependiera del criterio de un único algoritmo, se diseñó un esquema de votación con tres estimadores de naturaleza distinta — Regresión Lineal, Ridge y SVR lineal. Una variable solo era retenida si obtenía el respaldo de al menos dos de los tres modelos, asegurando que su relevancia fuera consistente y no un artefacto de un método específico.

A diferencia de un filtro estadístico como el VIF, el RFE evalúa la importancia de cada variable

desde la perspectiva del modelo predictivo: elimina iterativamente los predictores menos influyentes y conserva solo aquellos que aportan capacidad explicativa real. Para que esta selección no dependiera del criterio de un único algoritmo, se diseñó un esquema de votación con tres estimadores de naturaleza distinta — Regresión Lineal, Ridge y SVR lineal. Una variable solo era retenida si obtenía el respaldo de al menos dos de los tres modelos, asegurando que su relevancia fuera consistente y no un artefacto de un método específico.

La Figura 12 sintetiza visualmente el resultado del proceso de selección de variables a través de los cuatro escenarios experimentales definidos para la fase de modelado.

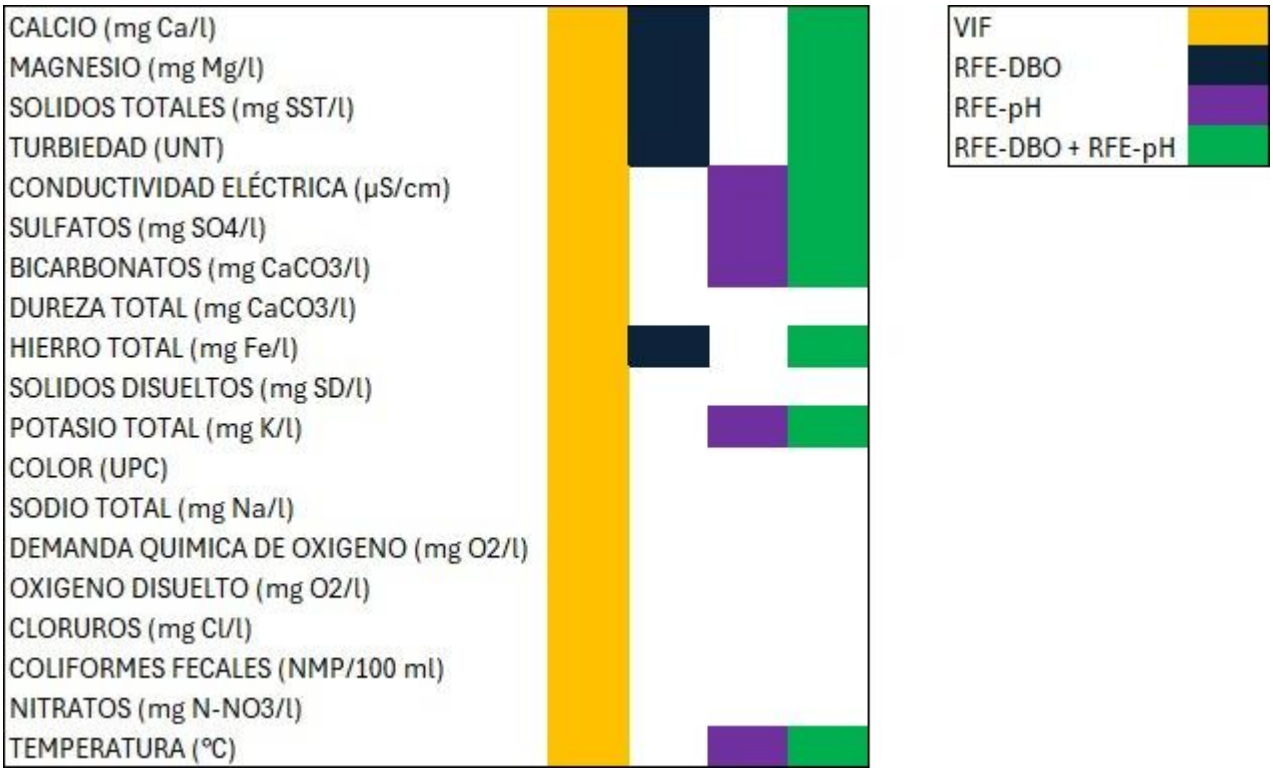

Figura 12. Escenarios para modelado

El escenario VIF — representado en amarillo — corresponde al conjunto base: las 19 variables que superaron el filtro de colinealidad y conforman el punto de partida. Los tres escenarios restantes son el resultado del RFE aplicado sobre ese conjunto.

El escenario RFE-DBO retuvo cinco predictores: Calcio, Hierro Total, Magnesio, Sólidos Totales y Turbiedad. Este conjunto refleja variables asociadas a la carga mineral y la presencia de sólidos en suspensión, parámetros que caracterizan la composición física del agua y están vinculados a los procesos de transporte y sedimentación de materia orgánica biodegradable.

El escenario RFE-pH seleccionó cinco predictores: Bicarbonatos, Conductividad Eléctrica, Potasio Total, Sulfatos y Temperatura. Este conjunto captura el sistema carbonatado y la dinámica iónica

que regula el equilibrio ácido-base del agua, coherente con los mecanismos conocidos de control del pH en sistemas hídricos naturales.

Un hallazgo relevante emerge de la comparación entre ambos conjuntos: no comparten ninguna variable. RFE-DBO y RFE-pH seleccionaron predictores completamente distintos, lo que confirma que predecir estas dos variables no es el mismo problema y que los factores que gobiernan la carga orgánica del río son diferentes de los que regulan su equilibrio ácido-base.

El escenario Unión integra los diez predictores seleccionados en los dos escenarios RFE, construyendo un conjunto amplio que permite evaluar modelos generales fundamentados en variables robustas para ambas variables objetivo. Por su parte, las nueve variables que solo aparecen en el escenario VIF no obtuvieron respaldo mayoritario en ninguno de los tres estimadores del RFE, lo que sugiere que su aporte explicativo, aunque presente, no resulta determinante frente a los predictores seleccionados.

Estos cuatro escenarios constituyeron el marco experimental para la fase de entrenamiento de modelos descrita en el siguiente capítulo.

5. ENTRENAMIENTO DE MODELOS DE REGRESIÓN

Con los predictores depurados, independientes y organizados en cuatro escenarios experimentales, el análisis llegó al momento central del estudio: entrenar los modelos y dejar que los datos hablen. El objetivo de esta fase no fue simplemente ajustar algoritmos, sino comparar enfoques de distinta naturaleza — lineales, basados en árboles, de ensamble y redes neuronales — para identificar cuál captura mejor las relaciones entre los parámetros fisicoquímicos del Río Cauca y las variables objetivo pH y DBO.

Para garantizar una evaluación objetiva, los datos se dividieron en un 80% para entrenamiento y un 20% para prueba. Se implementaron seis algoritmos: Regresión Lineal, Árbol de Decisión (CART), Bosque Aleatorio, Máquina de Vectores de Soporte (SVR), Perceptrón Multicapa (MLP) y XGBoost. Cada uno fue entrenado sobre los cuatro escenarios de variables, y su desempeño se evaluó mediante tres métricas complementarias: R^2, MAE y RMSE. Las configuraciones técnicas empleadas se detallan en la Tabla 24.

Tabla 24. Modelos y configuraciones empleadas en el proceso de modelado

MODELO	BIBLIOTECA / MÉTODO	CONFIGURACIÓN EMPLEADA	OBSERVACIONES
Regresión Lineal (LR)	Pipeline con StandardScaler y sklearn.linear_model. LinearRegression	StandardScaler(); LinearRegression()	Modelo lineal base. Se estandarizan las variables para facilitar la comparabilidad de coeficientes.
Árbol de Decisión (CART)	sklearn.tree.DecisionTreeRegressor	random_state = 42	Modelo no lineal basado en particiones. Se ejecuta con parámetros por defecto y semilla fija.
Bosque Aleatorio (RF)	sklearn.ensemble.RandomForestRegressor	random_state = 42; n_jobs = -1; n_estimators = 100 por defecto	Modelo ensamblado usado como referencia robusta. Usa procesamiento paralelo.
Máquina de Vectores de Soporte (SVR RBF)	Pipeline con StandardScaler y sklearn.svm.SVR	StandardScaler(); kernel = "rbf"	Modelo no lineal con kernel radial. Requiere estandarización previa.
Perceptrón Multicapa (MLP)	Pipeline con StandardScaler y sklearn.neural_network.MLPRegressor	StandardScaler(); random_state = 42; max_iter = 600; hidden_layer_sizes = (100,) por defecto; activation = "relu" por defecto; solver = "adam" por defecto; alpha = 0.0001 por defecto	Red neuronal base. Se amplía el número máximo de iteraciones para favorecer convergencia.

MODELO	BIBLIOTECA / MÉTODO	CONFIGURACIÓN EMPLEADA	OBSERVACIONES
XGBoost	xgboost.XGBRegressor	random_state = 42; tree_method = "hist"; n_jobs = -1; max_depth = 6; learning_rate = 0.1; n_estimators = 400; subsample = 0.8; colsample_bytree = 0.8; reg_lambda = 1.0; verbosity = 0	Modelo ensamblado de boosting. Se usa una configuración base controlada antes de procesos de optimización.

La Tabla 24 presenta los seis algoritmos implementados junto con sus configuraciones técnicas. Más allá de los parámetros específicos, lo que define el diseño experimental es la diversidad deliberada de enfoques: cada modelo representa una forma distinta de entender los datos.

La Regresión Lineal actúa como referencia base — el modelo más simple posible, que asume relaciones proporcionales entre predictores y variable objetivo. Su inclusión no es ingenua: si los modelos complejos no la superan por un margen considerable, la complejidad adicional no se justifica. El Árbol de Decisión (CART) introduce no linealidad sin abandonar la interpretabilidad, particionando el espacio de predictores en regiones homogéneas. El Bosque Aleatorio extiende esa lógica mediante un ensamble de cientos de árboles, reduciendo la varianza y ganando robustez frente al ruido.

La Máquina de Vectores de Soporte con kernel RBF captura relaciones no lineales desde una perspectiva geométrica, proyectando los datos a espacios de mayor dimensión donde las relaciones se vuelven separables. El Perceptrón Multicapa representa el enfoque de redes neuronales — con una capa oculta de 100 neuronas, activación ReLU y optimización Adam, configurado con 600 iteraciones máximas para garantizar convergencia. Finalmente, XGBoost es el modelo con la configuración más elaborada: profundidad controlada, submuestreo de datos y variables, y regularización L2, estableciendo un punto de partida sólido antes de su posterior optimización con Optuna.

Un elemento transversal a todos los modelos es random_state = 42, que garantiza la reproducibilidad completa de los resultados. Los tres modelos sensibles a la escala — Regresión Lineal, SVR y MLP — se implementaron dentro de un pipeline con StandardScaler, asegurando que ninguna variable dominara por unidades de medida y no por información real

Para cuantificar y comparar el rendimiento de los modelos, se emplearon tres métricas estándar: el Coeficiente de Determinación (R^2), que mide qué proporción de la variabilidad de los datos explica el modelo; el Error Absoluto Medio (MAE), que expresa la magnitud promedio del error en las mismas unidades de la variable objetivo; y el Error Cuadrático Medio Raíz (RMSE), seleccionado como criterio principal de selección dado que penaliza con mayor fuerza los errores de gran magnitud — los más costosos en un sistema ambiental.

La Figura 13 y Figura 14 presentan los resultados comparativos para la predicción de la Demanda Bioquímica de Oxígeno (DBO), mostrando respectivamente el R^2 y los errores RMSE y MAE de los cinco modelos con mejor desempeño dentro del conjunto experimental.

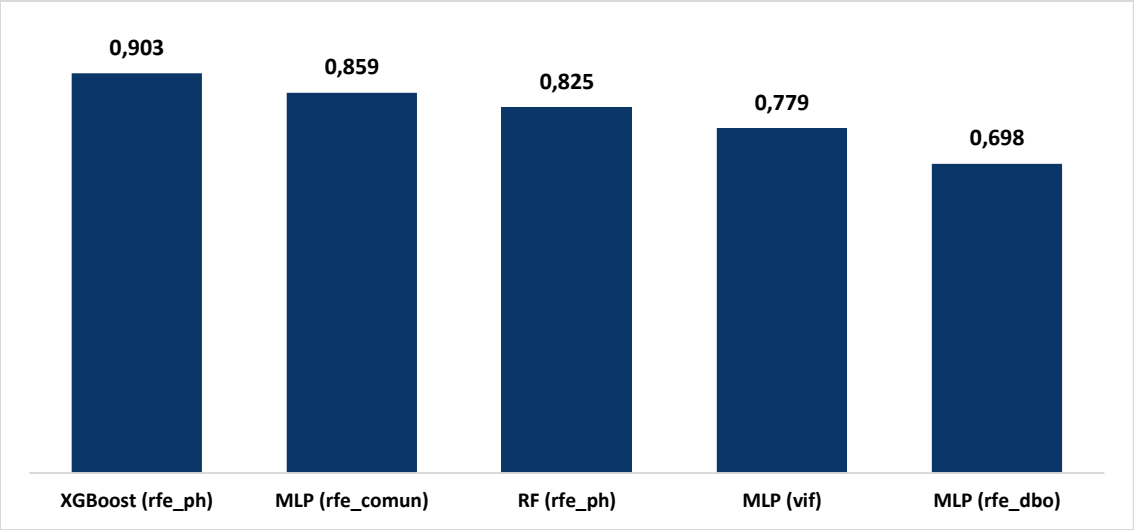

Figura 13. Desempeño comparativo de los cinco modelos con mayor R^2 para la predicción de la Demanda Bioquímica de Oxígeno (DBO)

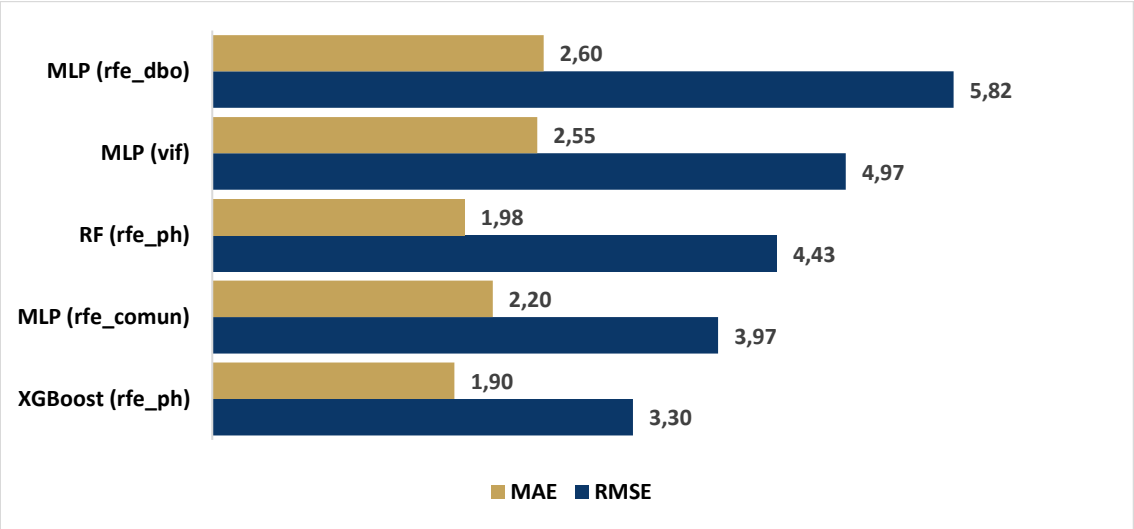

Figura 14. Comparación de errores RMSE y MAE de los mismos modelos destacados en la predicción de DBO

La Figura 13 presenta el coeficiente de determinación (R^2) de los cinco modelos con mejor ajuste en la predicción de la DBO. El resultado más destacado es el XGBoost entrenado con el escenario RFE-pH, que alcanza un R^2 de 0.903 — explicando más del 90% de la variabilidad de la DBO a partir de los predictores fisicoquímicos. Le siguen el MLP con el escenario de unión (RFE-Común, R^2=0.859) y el Bosque Aleatorio también con RFE-pH (R^2=0.825).

La Figura 14 complementa este análisis con las métricas de error, y confirma la jerarquía: XGBoost (RFE-pH) registra el RMSE más bajo (3.30) y el MAE más reducido (1.90), consolidándose como el modelo con mejor precisión global. En el extremo opuesto, MLP (RFE-DBO) presenta el mayor RMSE (5.82) y MAE (2.60), a pesar de haber sido entrenado con las variables específicamente seleccionadas para la DBO.

Este último punto merece atención. El escenario RFE-DBO — diseñado con predictores directamente correlacionados con la DBO como Oxígeno Disuelto, Coliformes Fecales y Sólidos Totales — no logró superar al escenario RFE-pH, cuyas variables fueron seleccionadas originalmente para predecir el pH. Los predictores del sistema carbonatado y la carga mineral disuelta —Bicarbonatos, Conductividad Eléctrica, Nitratos y Temperatura— demostraron mayor capacidad explicativa frente a la DBO que los indicadores de contaminación orgánica directa. Este hallazgo sugiere que la dinámica de la materia orgánica biodegradable en el Río Cauca está más estrechamente ligada al equilibrio químico del sistema que a los indicadores biológicos convencionales.

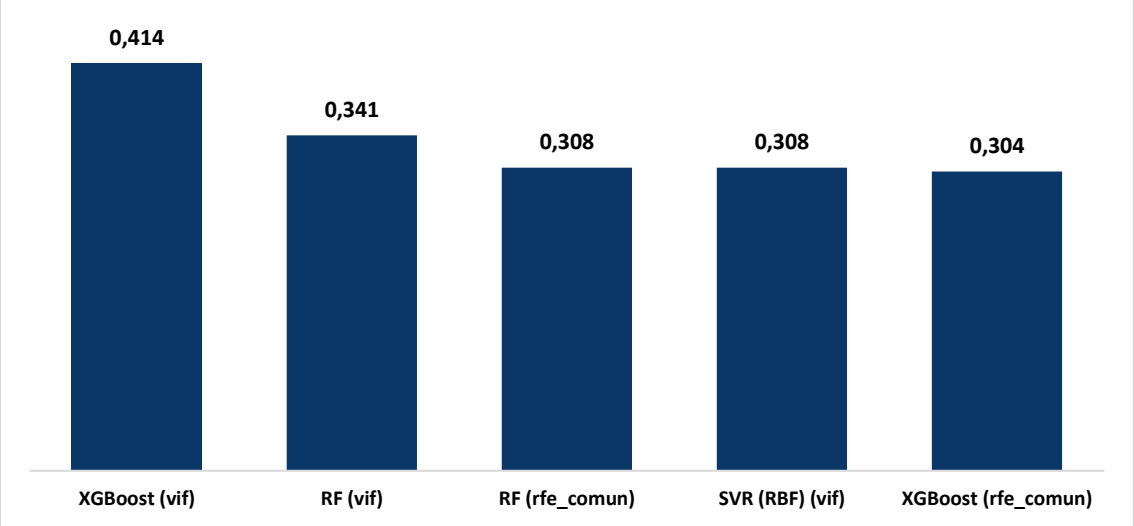

Figura 15. Top 5 modelos con mayor R^2 para la predicción del pH

La Figura 15 presenta el R^2 de los cinco modelos con mejor ajuste para la predicción del pH, y el contraste con los resultados de la DBO es inmediato: el valor más alto apenas alcanza 0.414, correspondiente al XGBoost entrenado con el escenario VIF. Lo que para la DBO era un resultado por encima del 90%, para el pH representa el techo del conjunto experimental. Este dato no es un fracaso metodológico — es una confirmación de lo anticipado en el análisis de correlaciones: el pH del Río Cauca responde a múltiples equilibrios simultáneos que ningún conjunto reducido de predictores logra capturar de forma dominante.

La Figura 16 refuerza esta lectura desde los errores. XGBoost (VIF) registra el RMSE más bajo (0.30) y el MAE más reducido (0.21), consolidándose como el modelo más preciso también en esta

dimensión. Los valores de error son pequeños en términos absolutos — lo que es esperable dada la escala acotada del pH — pero el R^2 bajo indica que los modelos describen correctamente el nivel general de la variable sin capturar bien su variabilidad interna.

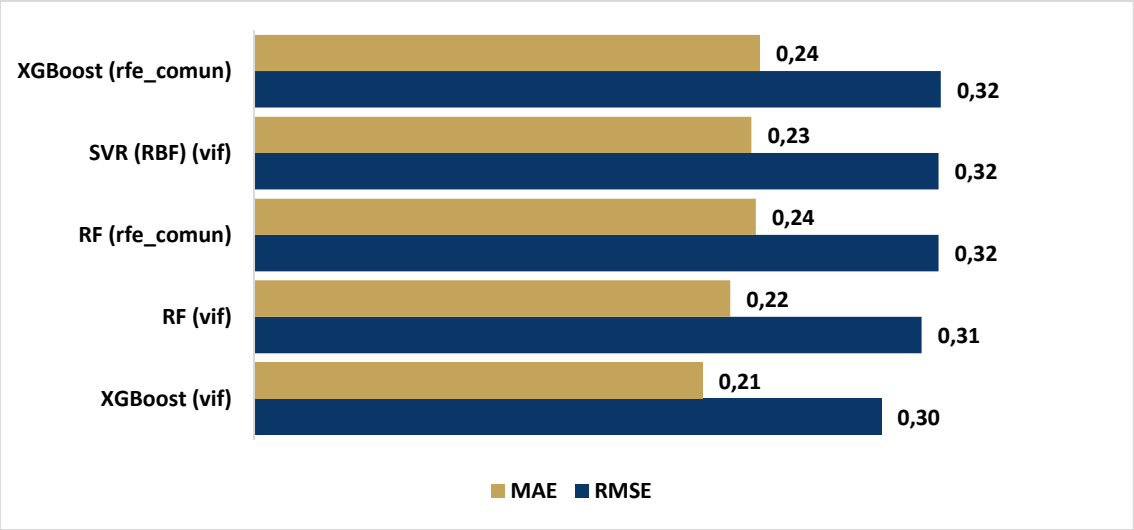

Figura 16. Comparación de errores RMSE y MAE de los modelos destacados en la predicción de pH

Dos observaciones adicionales emergen de estas figuras. Primero, todos los modelos del top 5 utilizan el escenario VIF o RFE-Común — el escenario RFE-pH, diseñado específicamente para el pH, no aparece entre los mejores. Esto sugiere que para una variable tan multifactorial como el pH, disponer de un conjunto más amplio de predictores resulta más ventajoso que uno reducido y especializado. Segundo, a diferencia de los resultados para DBO donde el MLP dominaba el ranking, aquí los métodos de ensamble — XGBoost y Random Forest — ocupan las primeras posiciones junto con el SVR de kernel radial, lo que indica que el pH responde mejor a modelos que trabajan con estructuras de decisión combinadas.

En conjunto, los resultados de pH y DBO establecen una conclusión metodológica clara: ambas variables requieren modelos no lineales y de ensamble para ser estimadas con precisión, pero representan problemas de distinta complejidad predictiva.

Los resultados detallados de todos los modelos y escenarios — incluyendo R^2, RMSE y MAE para cada combinación — se consolidan en el ANEXO E. Comparativa de precisión entre algoritmos de aprendizaje automático para la estimación de la calidad del agua (DBO) y ANEXO F. Comparativa de precisión entre algoritmos de aprendizaje automático para la estimación de la calidad del agua (pH). donde puede consultarse el comportamiento completo de los algoritmos bajo todas las configuraciones evaluadas.

Los modelos entrenados en la fase anterior usaban configuraciones base — funcionales, pero no optimizadas. Para extraer el máximo rendimiento de cada algoritmo, se implementó una segunda fase de ajuste de hiperparámetros con dos estrategias diferenciadas según la naturaleza de cada

modelo.

Para los cinco modelos tradicionales — Regresión Lineal, CART, Random Forest, SVR y MLP — se aplicó búsqueda exhaustiva en rejilla (Grid Search). El principio es simple: se define una cuadrícula de valores posibles para cada hiperparámetro y se evalúan sistemáticamente todas las combinaciones mediante validación cruzada. El ganador es la configuración que minimiza el error de validación. La Tabla 25 detalla los parámetros y rangos evaluados para cada modelo.

La Tabla 25 presenta los parámetros y rangos evaluados para cada modelo. La amplitud de la cuadrícula refleja la complejidad de cada algoritmo: desde la Regresión Lineal — con un único parámetro binario que define si se ajusta o no el intercepto — hasta el Random Forest, que requiere sintonizar simultáneamente cinco dimensiones: número de árboles, profundidad máxima, criterios de división y fracción de variables por nodo.

Tabla 25. Parámetros y rangos evaluados mediante el método Grid Search.

MÉTODO / MODELO	PARÁMETRO	DESCRIPCIÓN	RANGO / VALORES EVALUADOS
LinearRegression (con StandardScaler)	m__fit_intercept	Define si se ajusta el intercepto	[True, False]
DecisionTreeRegressor (CART)	max_depth	Profundidad máxima del árbol	[3, 5, 8, 10, None]
	min_samples_split	Muestras mínimas para dividir nodo	[2, 5, 10]
	min_samples_leaf	Muestras mínimas por hoja	[1, 2, 4]
RandomForestRegressor (RF)	n_estimators	Número de árboles	[100, 300, 500]
	max_depth	Profundidad máxima	[5, 10, 20, None]
	min_samples_split	Muestras mínimas para dividir nodo	[2, 5]
	min_samples_leaf	Muestras mínimas por hoja	[1, 2]
	max_features	Número de variables consideradas por split	['sqrt', 'log2', 0.5]
SVR (RBF)	m__C	Parámetro de regularización	[0.1, 1, 10, 100]
	m__epsilon	Margen de tolerancia	[0.01, 0.1, 0.5]
	m__gamma	Parámetro del kernel	['scale', 'auto']

MÉTODO / MODELO	PARÁMETRO	DESCRIPCIÓN	RANGO / VALORES EVALUADOS
MLPRegressor	m__hidden_layer_sizes	Arquitectura de la red	[(50,), (100,), (100,50), (200,100)]
	m__activation	Función de activación	['relu', 'tanh']
	m__alpha	Regularización L2	[0.0001, 0.001, 0.01]
	m__learning_rate_init	Tasa de aprendizaje inicial	[0.001, 0.01]

Para los modelos basados en árboles, los rangos de profundidad incluyen la opción None, que permite crecimiento irrestricto. Esta decisión es deliberada: en algunos escenarios, un árbol sin poda captura patrones relevantes que una profundidad fija truncaría prematuramente. El riesgo de sobreajuste asociado a esta configuración queda controlado por la validación cruzada, que penaliza modelos que memorizan el entrenamiento, pero fallan en datos nuevos.

En el SVR, el parámetro C recorre cuatro órdenes de magnitud — de 0.1 a 100 — explorando desde modelos con alta tolerancia al error hasta configuraciones de margen muy estricto. El parámetro gamma, que controla el alcance de influencia de cada punto de entrenamiento, se evalúa en sus modos automáticos ('scale' y 'auto'), delegando en el algoritmo la estimación óptima según la escala de los datos.

El MLP es el modelo con mayor riqueza explorada en términos de arquitectura: desde una red compacta de 50 neuronas hasta una configuración de dos capas con 200 y 100 neuronas respectivamente. Esta variación permite evaluar si la complejidad adicional de redes más profundas se traduce en mejora real de predicción o simplemente en mayor costo computacional sin ganancia significativa.

Donde el Grid Search recorre una cuadrícula fija de combinaciones, Optuna opera de forma distinta: aprende. Cada ensayo informa al siguiente, concentrando la búsqueda progresivamente en las regiones del espacio de hiperparámetros que muestran mayor potencial. Este enfoque bayesiano, implementado mediante el muestreador TPE (*Tree-structured Parzen Estimator*) con semilla fija (seed=42), ejecutó 60 iteraciones sobre el XGBoost — suficientes para explorar el espacio de forma adaptativa sin incurrir en costos computacionales excesivos.

Tabla 26. Parámetros y rangos de búsqueda empleados en la optimización bayesiana con Optuna

MÉTODO / MODELO	PARÁMETRO	DESCRIPCIÓN	RANGO / VALORES EVALUADOS
XGBoost (Optuna)	n_estimators	Número de árboles secuenciales	100 – 2000
	max_depth	Profundidad máxima de cada árbol	3 – 12
	learning_rate	Tasa de aprendizaje	0.01 – 0.3, escala logarítmica
	subsample	Proporción de datos usada por árbol	0.5 – 1.0

MÉTODO / MODELO	PARÁMETRO	DESCRIPCIÓN	RANGO / VALORES EVALUADOS
	colsample_bytree	Proporción de variables usadas por árbol	0.5 – 1.0
	reg_alpha	Regularización L1	1e-3 – 10.0, escala logarítmica
	reg_lambda	Regularización L2	1e-3 – 10.0, escala logarítmica
	min_child_weight	Peso mínimo por nodo hijo	1 – 8
	gamma	Reducción mínima de pérdida para particionar	0.0 – 2.0

La Tabla 26 presenta los nueve parámetros optimizados y sus rangos de búsqueda. El contraste con el Grid Search es inmediato: en lugar de valores discretos predefinidos, Optuna trabaja con rangos continuos — algunos de ellos en escala logarítmica, como la tasa de aprendizaje (0.01–0.3) y los términos de regularización L1 y L2 (1e-3 a 10.0). Esta escala logarítmica es intencional: en parámetros donde pequeñas variaciones en magnitudes bajas tienen mayor impacto que grandes variaciones en magnitudes altas, explorar en escala logarítmica garantiza una cobertura más representativa del espacio real de búsqueda.

El rango de n_estimators se amplió hasta 2000 árboles — cinco veces el valor base utilizado en la configuración inicial — mientras que la profundidad máxima se exploró entre 3 y 12 niveles. Los parámetros de submuestreo (subsample y colsample_bytree) controlan qué fracción de datos y variables se usa en cada árbol, funcionando como mecanismos de regularización implícita que reducen el sobreajuste. Finalmente, gamma y min_child_weight añaden criterios adicionales de poda, evitando que el modelo genere particiones innecesarias sobre ruido estadístico.

Con las estrategias de optimización aplicadas, el siguiente paso fue comparar el rendimiento real de los modelos ajustados. Las figuras a continuación presentan los resultados obtenidos tras el proceso de calibración, evaluando el desempeño de los algoritmos en términos de R^2, RMSE y MAE sobre los conjuntos de prueba — la medida definitiva de qué tan bien generaliza cada modelo frente a datos que nunca vio durante el entrenamiento.

En la Figura 17 presenta el R^2 de los cinco mejores modelos tras la fase de optimización[7], y el panorama cambió respecto a los resultados base. El Bosque Aleatorio entrenado con el escenario RFE-pH mediante Grid Search emerge como el modelo con mayor capacidad de ajuste, alcanzando un R^2 de 0.898. Le sigue XGBoost optimizado con Optuna (R^2=0.88), y en tercer lugar aparece un nuevo protagonista: el Árbol de Decisión (CART) con RFE-pH (R^2=0.865), que antes de la

[7] Todos los resultados obtenidos se encuentran en ANEXO G. Comparación de modelos y escenarios de predicción para la Demanda Bioquímica de Oxígeno (DBO) después de Grid Search y Optuna

optimización no figuraba entre los primeros puestos. Este ascenso evidencia cuánto puede ganar un modelo aparentemente simple cuando sus hiperparámetros se calibran correctamente.

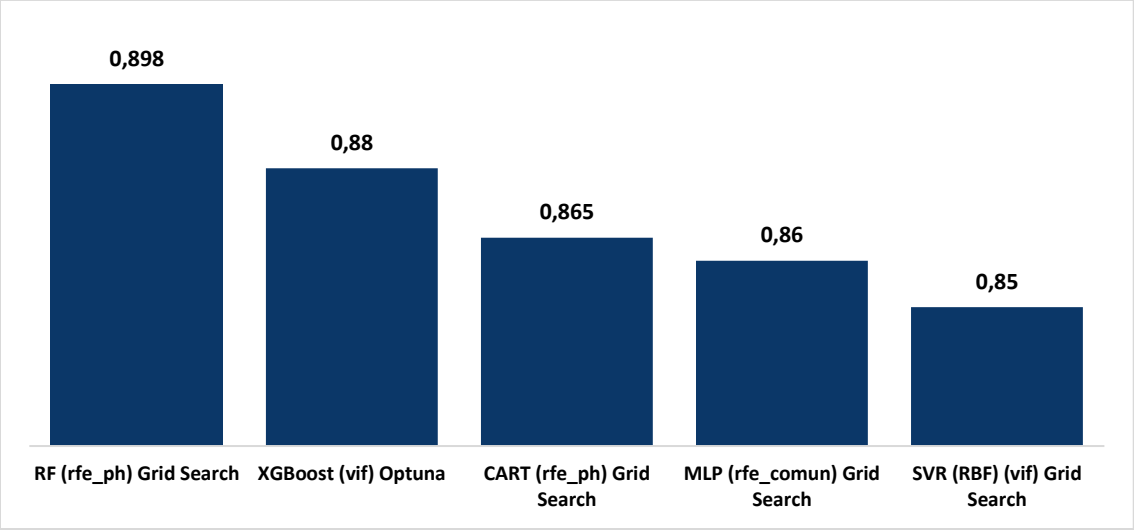

Figura 17. Comparación de métricas de rendimiento (R2) de los mejores modelos optimizados para la predicción de DBO

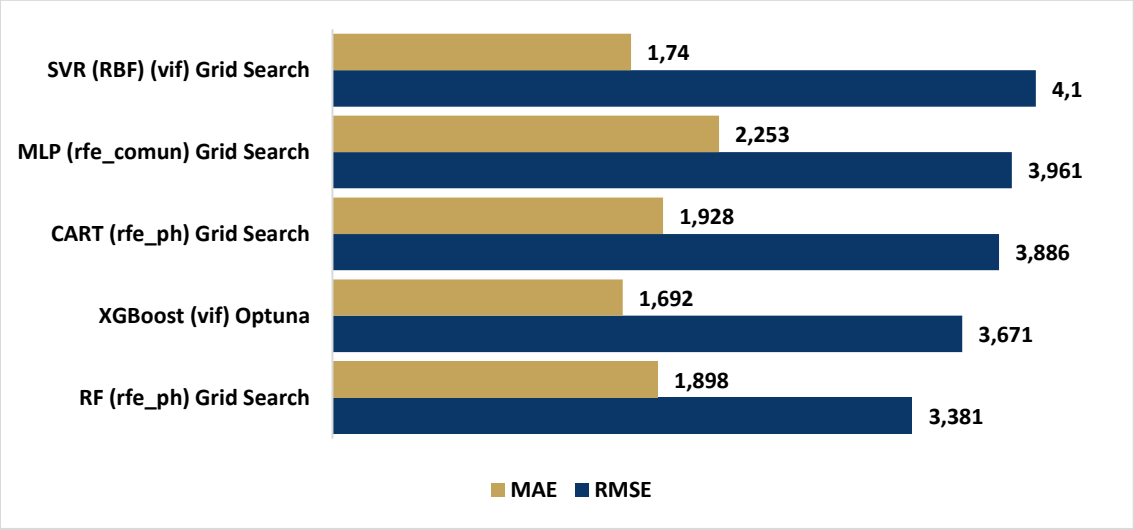

Figura 18. Comparación de métricas de rendimiento (RMSE y MAE) de los mejores modelos optimizados para la predicción de Demanda Bioquímica de Oxígeno (DBO)

La Figura 18 completa el análisis con los errores RMSE y MAE. El RF (RFE-pH) confirma su liderazgo con el RMSE más bajo del grupo (3.381) y un MAE de 1.898, lo que lo consolida como el mejor modelo en ambas dimensiones simultáneamente. XGBoost con Optuna sigue de cerca (RMSE=3.671), mientras que el SVR presenta el caso más paradójico del conjunto: a pesar de figurar en el top 5 de R^2, registra el RMSE más alto (4.10), lo que indica una mayor dispersión en

los errores individuales — penalizada precisamente por ser la métrica más sensible a valores extremos.

Dos conclusiones emergen de este bloque. Primero, la optimización redistribuyó el ranking: modelos que antes no destacaban — como CART — ganaron terreno, mientras que el MLP, que dominaba los resultados base, cedió posiciones. Segundo, el escenario RFE-pH se consolida como el conjunto de variables más potente para predecir la DBO, apareciendo en los dos modelos líderes. Las variables del equilibrio químico del río siguen explicando mejor la carga orgánica que los indicadores biológicos directos.

El análisis de la Demanda Bioquímica de Oxígeno demostró que es posible alcanzar niveles de ajuste superiores al 89% con los modelos adecuados. La pregunta que siguió fue si ese mismo desempeño se replicaría para el pH — la segunda variable objetivo del estudio. La respuesta, como anticipaba el análisis de correlaciones, fue que no: el pH juega con reglas distintas.

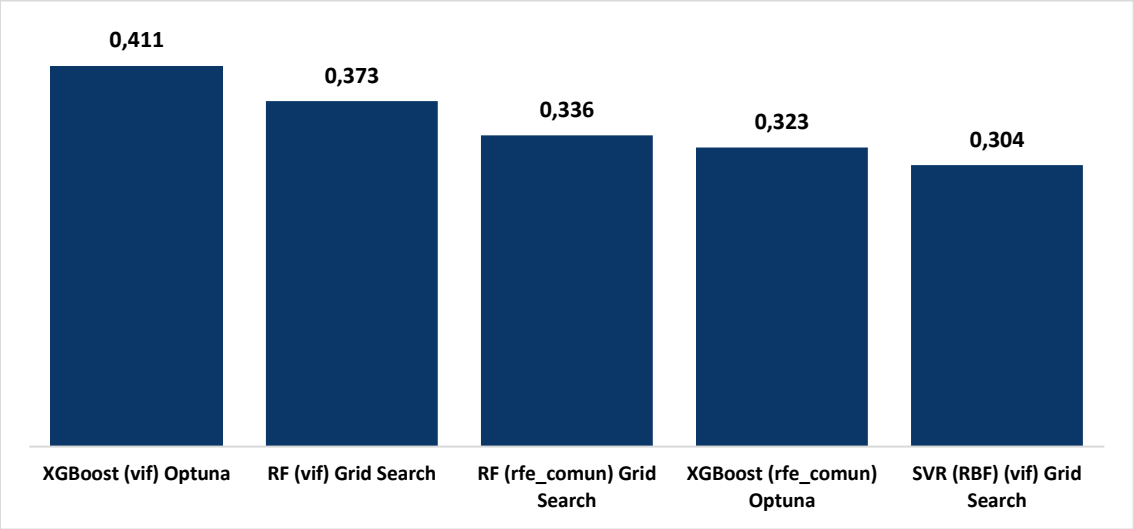

Figura 19. Comparación de métricas de rendimiento (R2) de los mejores modelos optimizados para la predicción de pH

La Figura 19 presenta el R^2 de los cinco mejores modelos para pH tras la optimización[8], y el resultado más revelador no está en quién gana — sino en cuánto gana. XGBoost con Optuna sobre el escenario VIF lidera con un R^2 de 0.411, seguido del Random Forest con Grid Search (0.373). Ambos resultados son estadísticamente sólidos dentro del conjunto experimental, pero el techo de predicción del pH permanece en un nivel significativamente inferior al de la DBO.

La Figura 20 confirma el liderazgo de XGBoost (VIF) Optuna también en los errores: registra el

[8] Todos los resultados obtenidos se encuentran en ANEXO H. Comparación de modelos y escenarios de predicción para el pH después de Grid Search y Optuna

RMSE más bajo (0.297) y el MAE más reducido (0.21). En términos absolutos, estos errores son pequeños — coherentes con la escala acotada del pH — pero el R^2 bajo indica que los modelos describen correctamente el nivel general de la variable sin capturar su variabilidad interna con precisión.

El hallazgo más importante de este bloque no es técnico sino interpretativo: la optimización no cambió sustancialmente el panorama del pH. El mejor R^2 antes de la calibración era 0.414; después es 0.411. La diferencia es marginal. Esto confirma que el límite predictivo del pH en este estudio no es un problema de configuración de modelos — es una propiedad del sistema. El pH del Río Cauca responde a equilibrios simultáneos de naturaleza geológica, biológica e hidrológica que el conjunto de predictores disponible no alcanza a representar completamente, independientemente del algoritmo o la estrategia de optimización empleada.

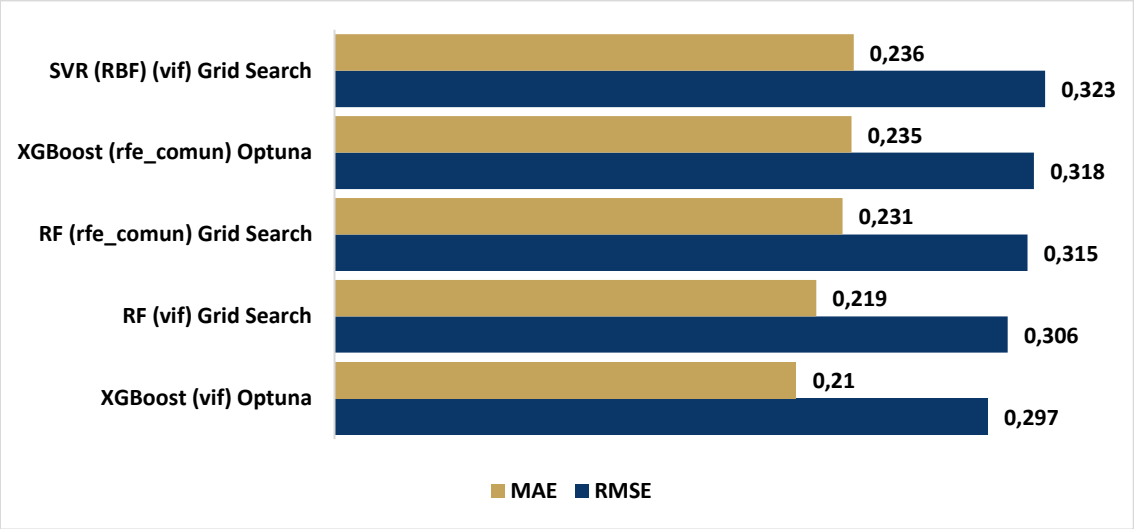

Figura 20. Comparación de métricas de rendimiento (RMSE y MAE) de los mejores modelos optimizados para la predicción de pH

Este resultado, lejos de ser una limitación menor, constituye un hallazgo metodológico relevante: establece con evidencia empírica el horizonte de predicción alcanzable para el pH bajo las condiciones de monitoreo actuales del río.

En conjunto, los resultados del entrenamiento y optimización de modelos establecen una distinción clara entre las dos variables objetivo. La DBO demostró ser predecible con alta precisión — el Bosque Aleatorio con el escenario RFE-pH alcanzó un R^2 de 0.898 y un RMSE de 3.381 tras la calibración con Grid Search, confirmando que la carga orgánica del río tiene predictores fisicoquímicos bien definidos. El pH, en cambio, mostró un comportamiento más complejo: el mejor modelo — XGBoost con Optuna sobre el escenario VIF — alcanzó un R^2 de 0.411, un valor que no mejora sustancialmente con la optimización y que refleja la naturaleza multifactorial de esta variable.

Estos resultados no solo identifican los mejores modelos para cada objetivo, sino que plantean

una pregunta legítima: ¿qué variables están impulsando esas predicciones? Para responderla, la siguiente etapa del análisis se adentra en la interpretabilidad de los modelos mediante valores DE Permutation Importance, SHAP[9] y Curvas de Aprendizaje, con el fin de comprender qué factores fisicoquímicos del Río Cauca pesan más en cada predicción y qué tan coherentes son esos pesos con el conocimiento ambiental del sistema.

[9] *SHapley Additive exPlanations*

6. EVALUACIÓN, SELECCIÓN E INTERPRETABILIDAD

Identificar el mejor modelo es solo la mitad del trabajo. La otra mitad es entender por qué funciona — qué variables está usando, cuánto peso le asigna a cada una y si esa lógica interna tiene sentido desde el punto de vista ambiental. Esta etapa buscó precisamente eso: no solo comparar rendimiento, sino abrir los modelos y leer lo que hay adentro.

Tabla 27. Modelos seleccionados

OBJETIVO	ESCENARIO	MODELO	R^2	RMSE	MAE	TÉCNICA
DBO	RFE-pH	RF	0.90	3.38	1.90	Grid
pH	VIF	XGBoost	0.41	0.30	0.21	Optuna

La Tabla 27 consolida los dos modelos seleccionados como referencia para el análisis de interpretabilidad. Para la DBO, el Bosque Aleatorio con Grid Search sobre el escenario RFE-pH alcanzó un R^2=0.90, RMSE=3.38 mg O_2/L y MAE=1.90 mg O_2/L, empleando únicamente cinco predictores de alta relevancia estadística. Para el pH, el XGBoost optimizado con Optuna sobre el escenario VIF obtuvo R^2=0.41, RMSE=0.30 y MAE=0.21, utilizando el conjunto completo de 19 variables. Ambos modelos combinan buen desempeño, parsimonia metodológica y capacidad de generalización — condiciones necesarias para que su interpretación sea confiable.

DEMANDA BIOQUÍMICA DE OXIGENO (mg O2/l)

El primer análisis de interpretabilidad aplicado al modelo RF (RFE-pH) fue la Importancia por Permutación (Permutation Importance), una técnica que mide cuánto se deteriora el desempeño del modelo cuando los valores de una variable se mezclan aleatoriamente. Si al desordenar una variable el error aumenta mucho, esa variable era crítica; si el error apenas cambia, su contribución es marginal.

La Tabla 28 y Figura 21 presentan los resultados para las cinco variables del modelo, y el hallazgo es contundente: los Sulfatos dominan con una importancia media de 0.9534 — un valor que deja al resto del conjunto prácticamente en el margen. La Temperatura ocupa el segundo lugar con 0.1184, menos de una octava parte del valor de los Sulfatos, seguida de Bicarbonatos (0.0486), Conductividad Eléctrica (0.0304) y Potasio Total (0.0149).

Este resultado tiene una lectura ambiental directa. Los Sulfatos en el Río Cauca son un marcador de actividad antropogénica — drenaje ácido de minas, efluentes industriales de sectores como textiles y papeleras. Su dominio en la predicción de la DBO sugiere que los episodios de mayor carga orgánica biodegradable en el río están estrechamente asociados a fuentes de contaminación industrial, donde los sulfatos y la materia orgánica se vierten de forma simultánea. La Temperatura complementa esta lectura desde la biología: a mayor temperatura, los procesos de descomposición microbiana se aceleran, incrementando el consumo de oxígeno y, por ende, la DBO.

Tabla 28. Permuted Importance del mejor modelo para DBO

VARIABLE	IMPORTANCIA MEDIA	DESVIACIÓN ESTÁNDAR
SULFATOS (mg SO4/l)	0.9534	0.0449
TEMPERATURA (°C)	0.1184	0.0225
BICARBONATOS (mg CaCO3/l)	0.0486	0.0166
CONDUCTIVIDAD ELÉCTRICA (µS/cm)	0.0304	0.0137
POTASIO TOTAL (mg K/l)	0.0149	0.0045

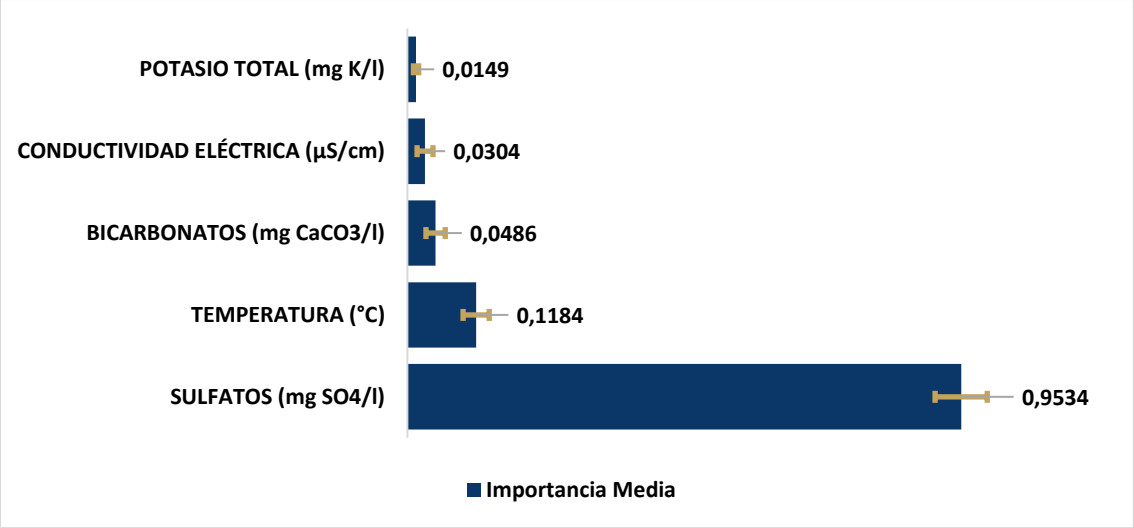

Figura 21. Permuted Importance del mejor modelo para DBO

La baja desviación estándar de los Sulfatos (0.0449) frente a su alta importancia confirma que este resultado es estable — no es un artefacto de una corrida específica del modelo sino una señal consistente en todas las evaluaciones.

Mientras la Importancia por Permutación evalúa el impacto global de cada variable sobre el error del modelo, los valores SHAP van un paso más allá: cuantifican la contribución marginal de cada predictor en cada predicción individual, promediando luego esos aportes para obtener una medida de importancia global. La diferencia no es solo técnica — es interpretativa. SHAP distribuye el crédito de forma más equitativa entre variables correlacionadas, ofreciendo una imagen más fiel de cómo el modelo toma sus decisiones.

La Figura 22 presenta los valores SHAP medios absolutos para el modelo RF (RFE-pH) en la predicción de la DBO. El contraste con la Figura 21 es inmediato: donde la Importancia por Permutación concentraba casi todo el peso en los Sulfatos, SHAP revela un sistema más distribuido. Los Sulfatos siguen liderando (1.1106), pero la Temperatura (0.9056), los Bicarbonatos (0.8912) y la Conductividad Eléctrica (0.6617) aparecen como contribuyentes sustanciales — no como variables secundarias sino como pilares activos del modelo.

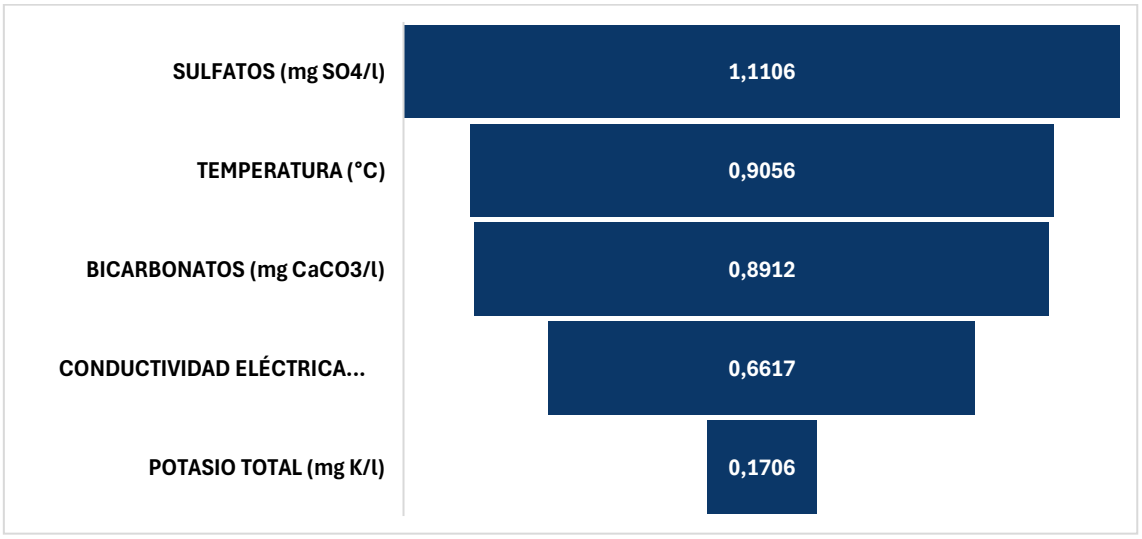

Figura 22. Importancia global de características (SHAP) con valor medio absoluto para el modelo RandomForest [pH-RFE] en la predicción de Demanda Bioquímica de Oxígeno (DBO)

Esta distribución tiene coherencia ambiental. Los Sulfatos marcan la huella industrial y minera; la Temperatura regula la velocidad de los procesos de descomposición microbiana; los Bicarbonatos y la Conductividad Eléctrica reflejan la carga iónica del sistema, vinculada a los vertimientos que simultáneamente elevan la DBO. En conjunto, estas cuatro variables construyen un retrato fisicoquímico de la presión antrópica sobre el río. El Potasio Total (0.1706), aunque presente, actúa como señal complementaria de menor peso explicativo.

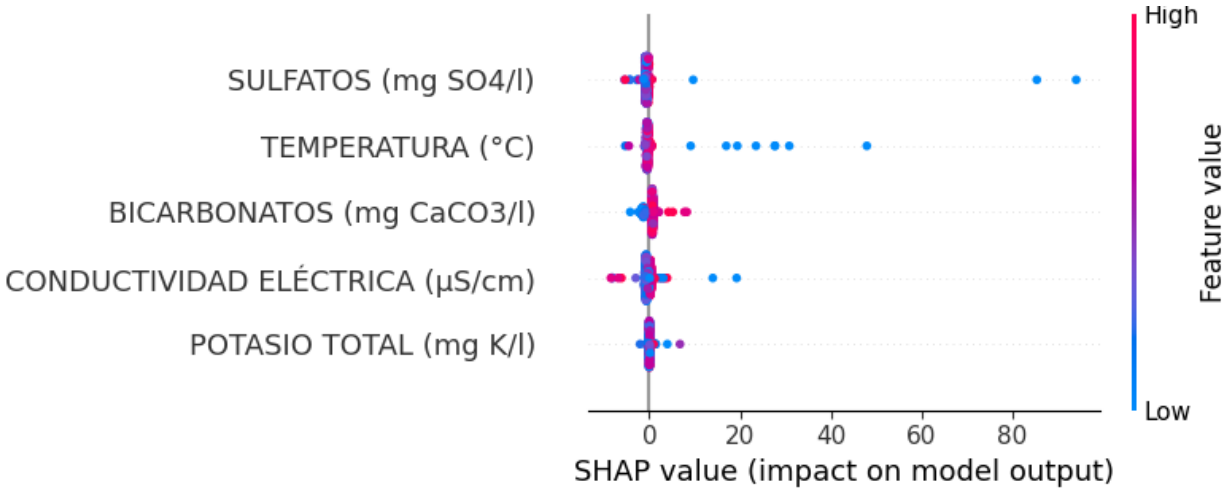

Figura 23. Distribución del impacto de las características (SHAP) para el modelo RandomForest [RFE-pH] en la predicción de Demanda Bioquímica de Oxígeno (DBO)

La Figura 22 mostraba los promedios — la Figura 23 muestra la historia completa. El gráfico de dispersión SHAP (beeswarm plot) representa el impacto individual de cada variable en cada

predicción del conjunto de prueba. Cada punto es una observación; su posición horizontal indica cuánto y en qué dirección influyó esa variable en esa predicción específica; su color refleja si el valor de la variable era alto (rosa) o bajo (azul).

El patrón más revelador corresponde a los Sulfatos. La gran mayoría de las observaciones se concentra cerca de cero — la mayor parte del tiempo, los sulfatos tienen un impacto moderado sobre la predicción de DBO. Pero hay puntos extremos que se proyectan hacia la derecha con valores SHAP de entre 60 y 90, muy por encima del resto. Estos outliers representan episodios puntuales de concentraciones elevadas de sulfatos donde el modelo reconoce una señal de contaminación intensa y ajusta la predicción de DBO al alza de forma drástica. Son eventos, no la norma — y por eso la distribución es tan asimétrica.

La Temperatura muestra un patrón más continuo: los puntos rosas (valores altos de temperatura) tienden a ubicarse a la derecha del eje, confirmando que períodos más cálidos se asocian consistentemente con mayor DBO. Bicarbonatos y Conductividad Eléctrica presentan distribuciones más compactas, con impactos moderados pero presentes en la mayoría de las observaciones. El Potasio Total permanece agrupado cerca del origen — su influencia es real pero pequeña en términos individuales.

En conjunto, esta figura revela que la DBO en el Río Cauca no sigue un patrón uniforme: la mayor parte del tiempo el sistema opera dentro de rangos predecibles, pero existen eventos episódicos — capturados principalmente por los Sulfatos — donde la carga contaminante se dispara y el modelo los detecta con precisión.

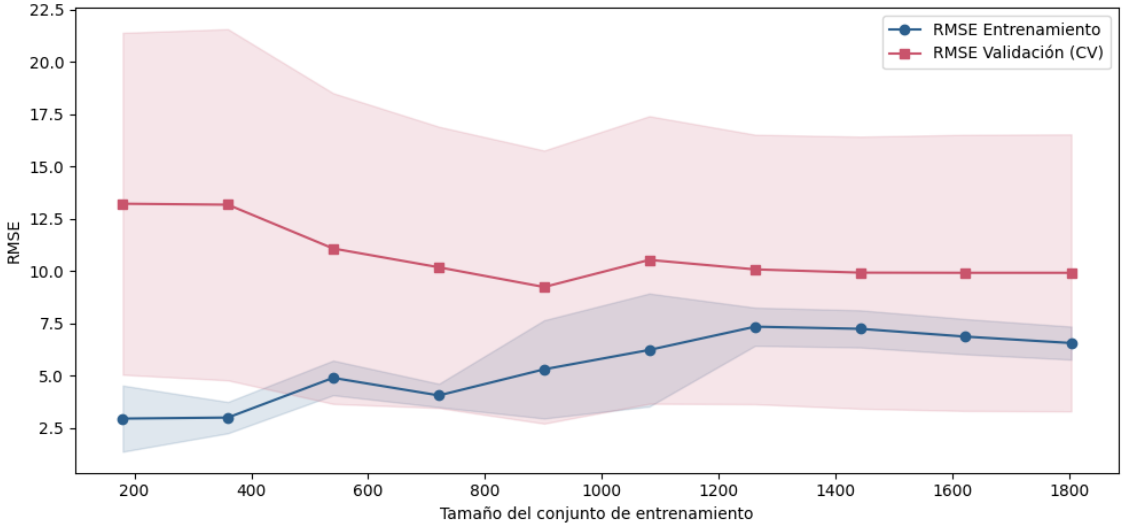

Figura 24. Curva de aprendizaje (RMSE) del modelo RandomForest [GS] para la predicción de Demanda Bioquímica de Oxígeno (DBO)

La Figura 24 presenta la curva de aprendizaje del modelo RF con Grid Search, una herramienta que permite evaluar cómo evoluciona el desempeño a medida que aumenta el volumen de datos

de entrenamiento. La línea azul representa el RMSE sobre los datos de entrenamiento; la línea roja, el RMSE de validación cruzada. Las bandas sombreadas indican la variabilidad de cada estimación.

El comportamiento es el esperado para un modelo de ensamble bien calibrado. Con pocos datos (alrededor de 200 registros), el RMSE de entrenamiento es bajo — el modelo memoriza fácilmente un conjunto pequeño — mientras que el RMSE de validación es alto e inestable, reflejado en la amplia banda rosa. A medida que el tamaño del conjunto crece, ambas curvas convergen: el error de entrenamiento sube moderadamente (el modelo ya no puede memorizar) y el error de validación desciende de forma sostenida, estabilizándose cerca de 10 mg O_2/L hacia los 1800 registros.

La brecha que persiste entre ambas curvas al final de la escala indica cierto grado de sobreajuste residual — el modelo sigue ajustando mejor sobre los datos que ha visto que sobre los nuevos. Sin embargo, la tendencia convergente sugiere que este comportamiento mejoraría con mayor volumen de datos históricos, lo que abre una recomendación directa para programas futuros de monitoreo del Río Cauca.

El análisis de la DBO, desde la selección de variables hasta la interpretabilidad del modelo final, configura un resultado coherente de principio a fin. El Bosque Aleatorio con el escenario RFE-pH no solo alcanzó el mejor desempeño predictivo del estudio (R^2=0.898, RMSE=3.38), sino que reveló que los Sulfatos, la Temperatura, los Bicarbonatos y la Conductividad Eléctrica son los factores fisicoquímicos con mayor peso explicativo sobre la carga orgánica del río — una lectura que conecta directamente con las fuentes de presión antrópica documentadas en la cuenca. Con este panorama establecido, el análisis continúa con la segunda variable objetivo: el pH.

pH (potencial de hidrógeno)

El pH no fue un problema de un solo factor — y sus resultados de interpretabilidad lo confirman. A diferencia de la DBO, donde los Sulfatos concentraban la mayor parte del peso explicativo, el análisis del pH revela un sistema distribuido: múltiples variables contribuyen de forma simultánea, ninguna domina de forma absoluta, y la predicción depende del equilibrio entre todas ellas.

La Tabla 29 y la Figura 25 presentan los resultados de Importancia por Permutación para el modelo XGBoost (VIF), mostrando los diez predictores con mayor contribución a la predicción del pH. El contraste con los resultados de la DBO es inmediato y significativo.

Tabla 29. Permuted Importance del mejor modelo para pH

VARIABLE	IMPORTANCIA MEDIA	DESVIACIÓN ESTÁNDAR
BICARBONATOS (mg CaCO3/l)	0.2478	0.037
NITRATOS (mg N-NO3/l)	0.1152	0.0275
TEMPERATURA (°C)	0.0721	0.0098
OXIGENO DISUELTO (mg O2/l)	0.0605	0.0158

VARIABLE	IMPORTANCIA MEDIA	DESVIACIÓN ESTÁNDAR
POTASIO TOTAL (mg K/l)	0.0576	0.0154
CLORUROS (mg Cl/l)	0.051	0.0091
SOLIDOS TOTALES (mg SST/l)	0.0467	0.0094
TURBIEDAD (UNT)	0.0463	0.0081
CALCIO (mg Ca/l)	0.0447	0.011
COLIFORMES FECALES (NMP/100 ml)	0.0344	0.0053

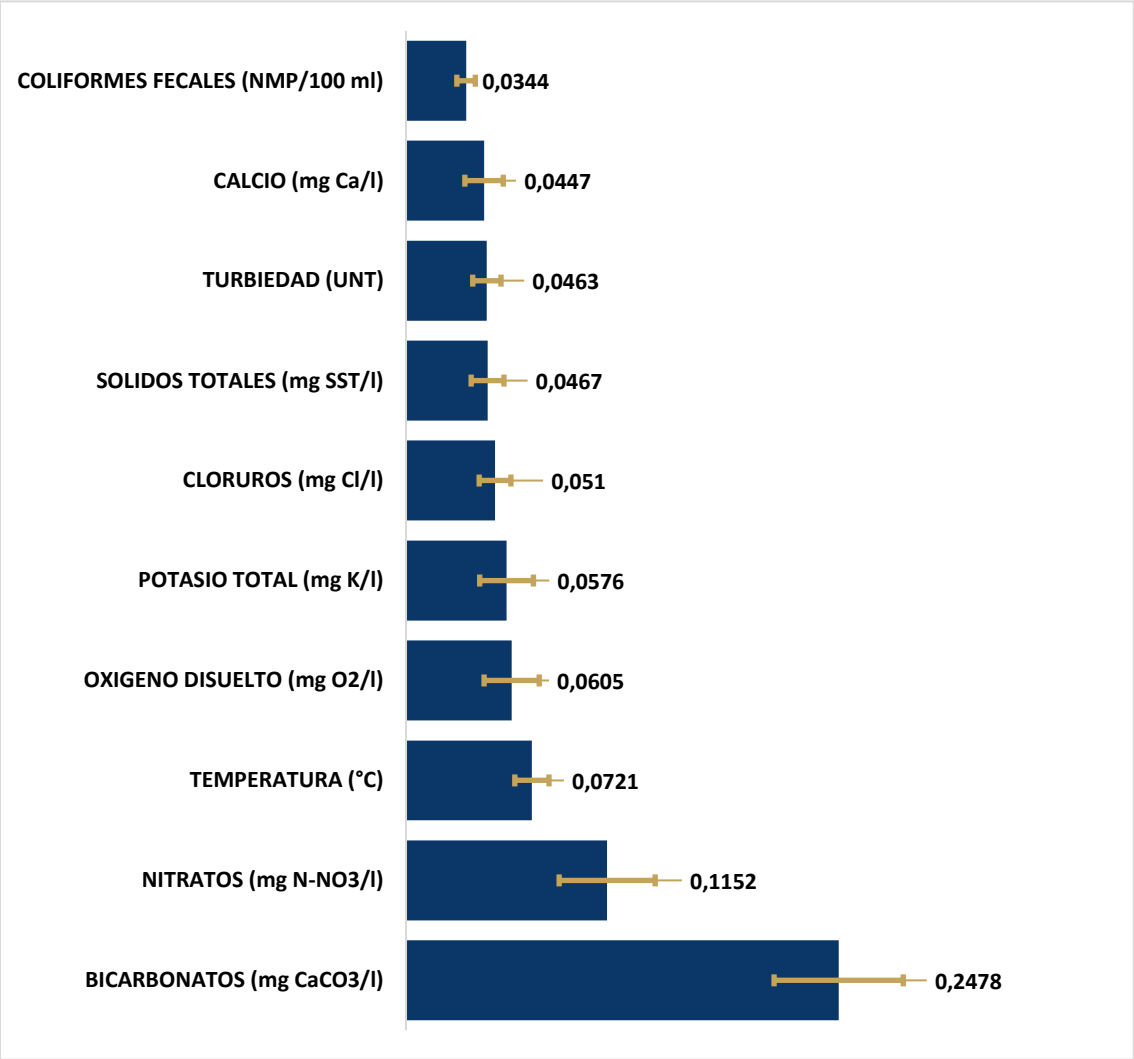

Figura 25. Permuted Importance del mejor modelo para pH

Los Bicarbonatos lideran con una importancia media de 0.2478 — un valor que refleja su rol central como reguladores del equilibrio ácido-base en aguas naturales. Sin embargo, a diferencia de los Sulfatos en la DBO, los Bicarbonatos no aplastan al resto: los Nitratos (0.1152), la

Temperatura (0.0721), el Oxígeno Disuelto (0.0605) y el Potasio Total (0.0576) mantienen contribuciones relevantes y relativamente próximas entre sí. La distancia entre el primer y el último valor de la tabla es de aproximadamente siete veces — en la DBO esa misma distancia era de más de sesenta.

Esta distribución más uniforme tiene una lectura directa: el pH del Río Cauca no obedece a un único driver sino a un conjunto de equilibrios fisicoquímicos que operan en paralelo. Los Bicarbonatos regulan la capacidad buffer; los Nitratos reflejan la actividad agrícola y el ciclo del nitrógeno; la Temperatura condiciona las reacciones de disolución de gases; el Oxígeno Disuelto conecta con los procesos biológicos aeróbicos; y variables como Cloruros, Sólidos Totales y Coliformes Fecales aportan señales adicionales de presión antrópica sobre el sistema.

La baja desviación estándar en todos los predictores confirma que estos resultados son estables y reproducibles — no artefactos de una corrida específica del modelo.

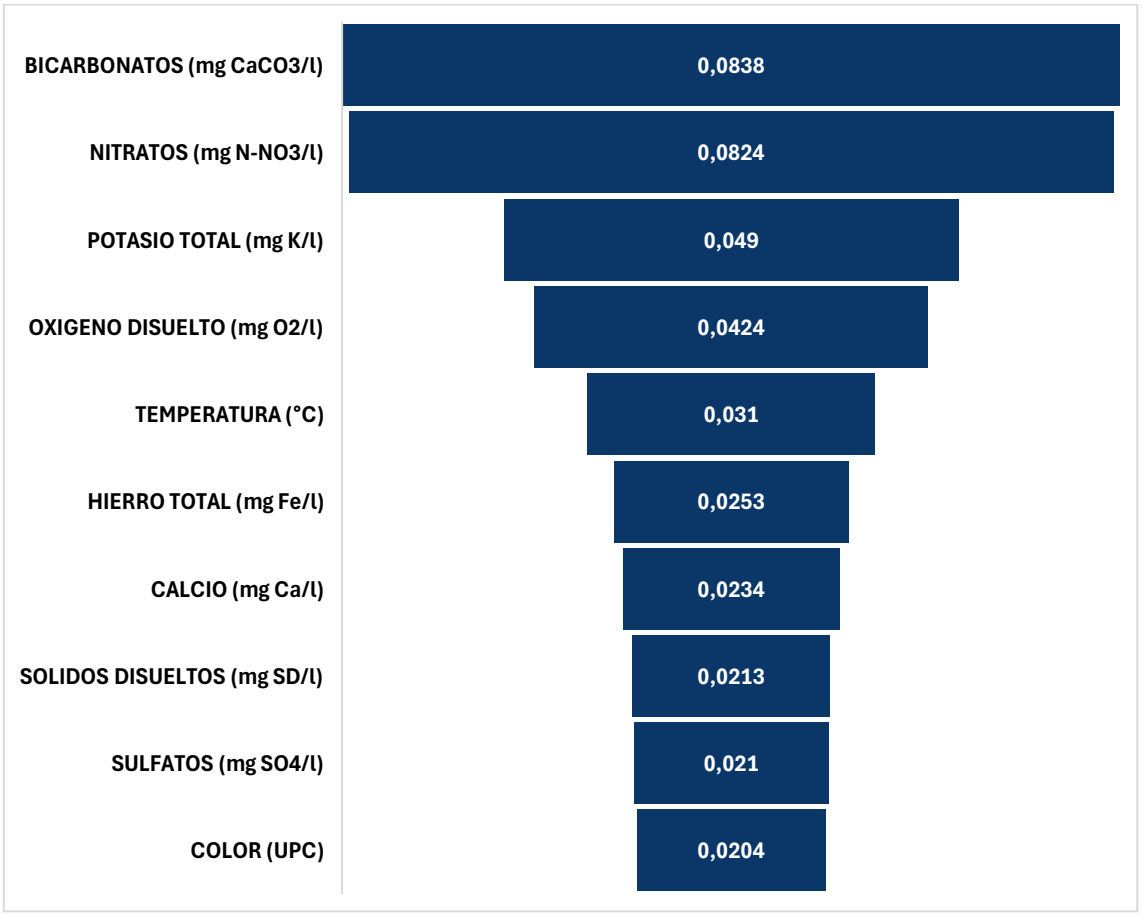

Figura 26. Importancia global de características (SHAP) con valor medio absoluto para el modelo XGBoost [VIF] en la predicción de pH

La Figura 27 presenta los valores SHAP medios absolutos para el modelo XGBoost (VIF) en la predicción del pH, y refuerza con mayor claridad lo que la Permutation Importance ya anticipaba: el pH no tiene un conductor principal — tiene un conjunto de variables que gobiernan en condiciones de equilibrio.

Bicarbonatos (0.0838) y Nitratos (0.0824) aparecen prácticamente empatados en el primer lugar, una diferencia de apenas 0.0014 que los convierte en codominantes del modelo. Esta proximidad es significativa: mientras la Permutation Importance mostraba a los Bicarbonatos con el doble de importancia que los Nitratos, SHAP revela que ambos contribuyen de forma casi idéntica cuando se evalúan las predicciones individuales. Los Bicarbonatos aportan desde el equilibrio buffer; los Nitratos, desde la actividad agrícola y el ciclo del nitrógeno en la cuenca.

Le siguen Potasio Total (0.049), Oxígeno Disuelto (0.0424) y Temperatura (0.031) — variables que mantienen sus posiciones respecto al análisis anterior. Más abajo aparecen Hierro Total (0.0253), Sólidos Disueltos (0.0213) y Color (0.0204), predictores que no figuraban en los primeros lugares de la Permutation Importance pero que SHAP identifica como contribuyentes activos a través de interacciones con otras variables. El Hierro Total, en particular, tiene relevancia química en sistemas con actividad minera: su oxidación genera iones H^+ que acidifican el medio, un proceso que el modelo captura de forma implícita.

Dos contrastes merecen destacarse. Primero, los valores absolutos de SHAP para el pH son notablemente más pequeños que los de la DBO — el máximo es 0.0838 frente a 1.1106 — lo que refleja que el modelo realiza ajustes más moderados por variable, coherente con la menor capacidad predictiva global. Segundo, los Sulfatos, que dominaban completamente la predicción de la DBO, aparecen aquí en novena posición con un valor de apenas 0.021 — evidencia de que los factores que impulsan la carga orgánica del río y los que regulan su equilibrio ácido-base son fundamentalmente distintos.

La Figura 27 presenta la distribución individual de los valores SHAP para el modelo XGBoost (VIF) en la predicción del pH, y la diferencia con el beeswarm de la DBO es evidente desde el primer vistazo: no hay outliers extremos, no hay barras que se proyecten decenas de unidades hacia la derecha. Todo el espectro de impactos individuales se concentra entre −0.75 y +0.5, reflejando un sistema donde ningún evento puntual genera una predicción radicalmente distinta del resto.

El patrón más legible corresponde a los Bicarbonatos. Los puntos azules — valores bajos de bicarbonatos — se desplazan hacia la izquierda del eje, indicando que cuando la concentración de bicarbonatos cae, el modelo predice un pH más bajo. Los puntos rosas — valores altos — se ubican a la derecha, elevando la predicción. Esta direccionalidad clara es exactamente lo que la química del agua anticipa: los bicarbonatos son el principal agente buffer del sistema, y su ausencia o reducción debilita la capacidad del río para resistir la acidificación.

El resto de las variables muestra distribuciones más compactas y simétricas, con puntos dispersos en ambas direcciones del eje. Esta bidireccionalidad no es ruido — indica relaciones condicionadas al contexto: una misma concentración de Nitratos o de Oxígeno Disuelto puede

impulsar el pH hacia arriba o hacia abajo dependiendo del estado general del sistema en ese momento. Es precisamente este tipo de interacción compleja la que justifica el uso de XGBoost sobre modelos lineales para predecir el pH.

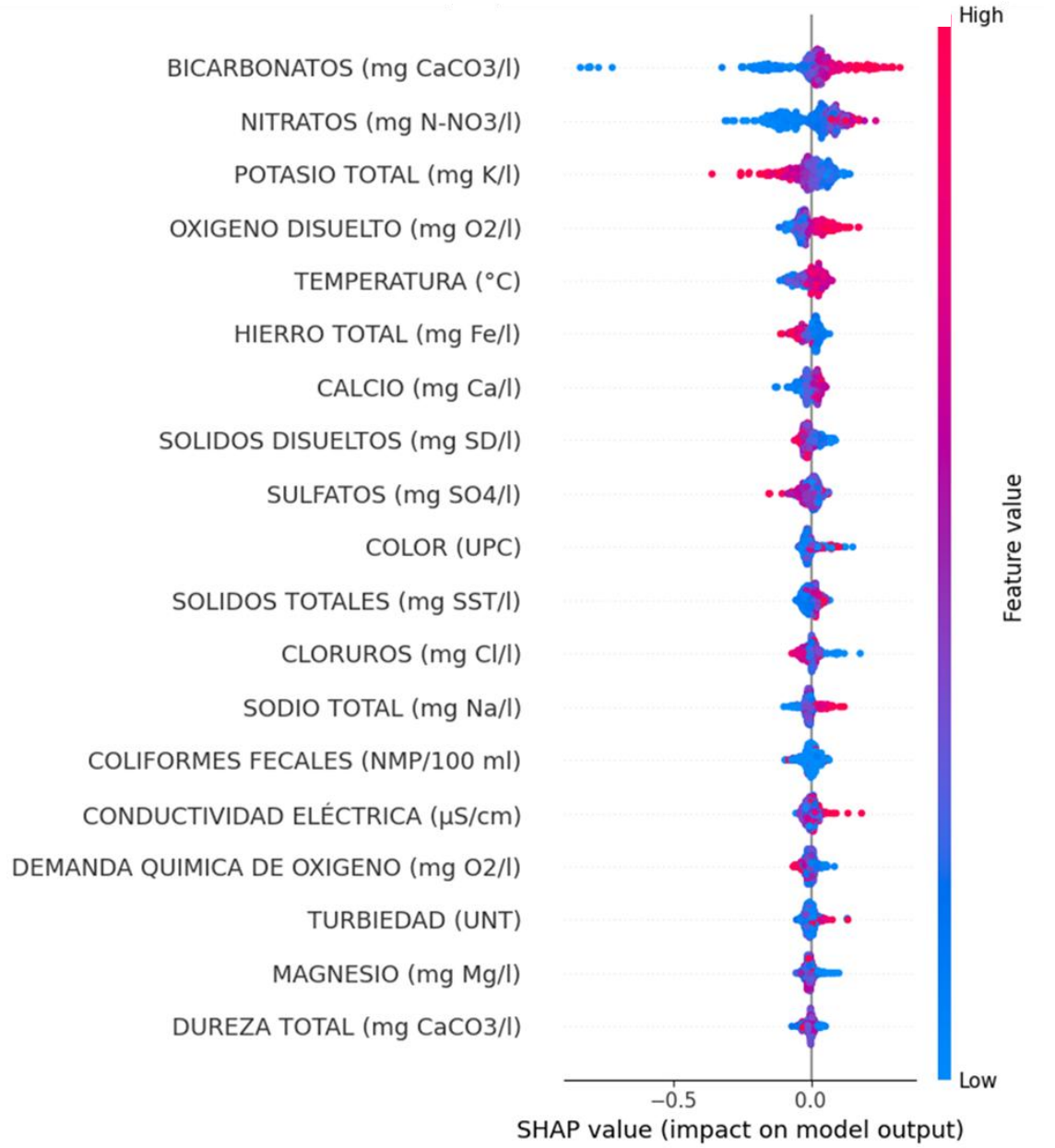

Figura 27. Distribución del impacto de las características (SHAP) para el modelo XGBoost [VIF] en la predicción de pH

En conjunto, el beeswarm confirma lo que los análisis anteriores anticipaban: el pH del Río Cauca es un fenómeno de equilibrio múltiple. No hay un evento dominante ni una variable que cuente toda la historia — hay un conjunto de factores que negocian permanentemente el estado químico

del agua.

Finalmente, la Figura 28 presenta la curva de aprendizaje del XGBoost para pH, y cuenta una historia distinta a la de la DBO — y más incómoda de leer.

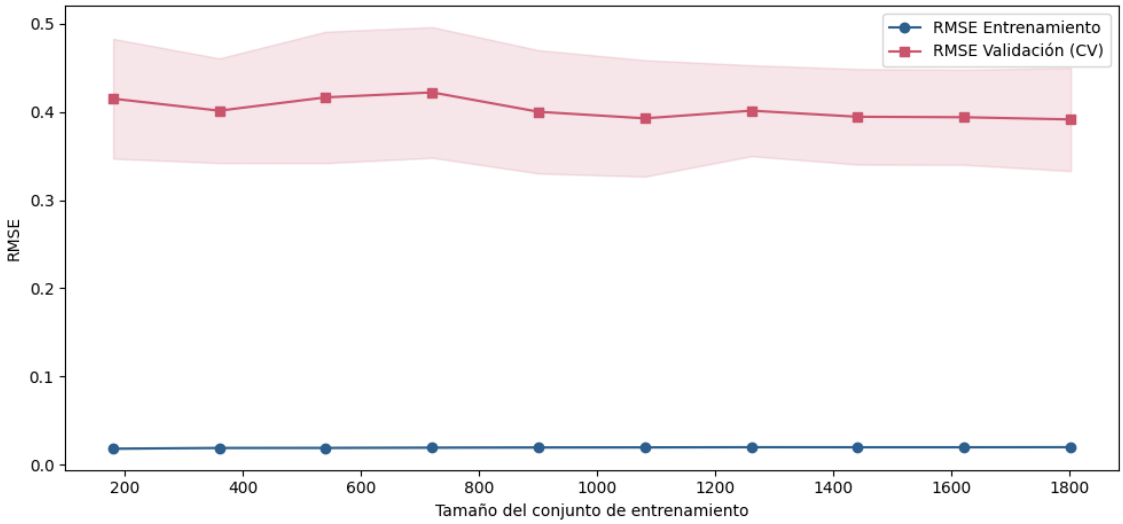

Figura 28. Curva de aprendizaje (RMSE) del modelo XGBoost [GS] para la predicción de pH

La línea azul de entrenamiento permanece prácticamente plana en cero a lo largo de toda la escala: el modelo ajusta los datos de entrenamiento con un error casi nulo independientemente del tamaño del conjunto. La línea roja de validación, en cambio, oscila entre 0.39 y 0.42 sin mostrar una tendencia descendente clara. A diferencia de la DBO, donde ambas curvas convergían progresivamente, aquí la brecha entre entrenamiento y validación se mantiene constante — no se cierra con más datos.

Este patrón describe un sobreajuste pronunciado: el modelo aprendió el conjunto de entrenamiento con precisión casi perfecta, pero esa precisión no se transfiere a datos nuevos. En el caso de XGBoost, este comportamiento es esperable cuando la variable objetivo tiene una dinámica compleja que los predictores disponibles no logran representar completamente. Agregar más observaciones no resuelve el problema si la información que falta no está contenida en las variables medidas.

La lectura práctica es directa: el techo de predicción del pH no es un problema de volumen de datos ni de configuración del modelo — es una consecuencia de las variables disponibles en el sistema de monitoreo actual. Para mejorar sustancialmente la predicción del pH en el Río Cauca, sería necesario incorporar variables adicionales que capturen los equilibrios geológicos, hidrológicos y biológicos que hoy no están representados en el conjunto de datos. El resumen del código se encuentra en el ANEXO I. Funcionamiento básico del código.

7. CONCLUSIONES Y TRABAJOS FUTUROS

CONCLUSIONES

Este estudio demostró que predecir la calidad del agua en el Río Cauca con técnicas de aprendizaje automático es posible — pero no trivial. Los resultados obtenidos no son solo métricas de desempeño: son el reflejo de un proceso metodológico riguroso que comenzó mucho antes del primer modelo, en la forma en que se trataron, depuraron y seleccionaron los datos. Las conclusiones que siguen recorren ese proceso por etapas, desde la construcción del conjunto de datos hasta la interpretación ambiental de los modelos finales.

La base de datos original del Río Cauca contenía 56 variables y más de dos décadas de registros históricos — un punto de partida rico pero desigual. Tras eliminar variables administrativas no informativas (fecha de muestreo y estación) y convertir todos los parámetros a formato numérico, el conjunto quedó conformado por 39 variables válidas. La aplicación de un umbral de completitud del 80% redujo ese número a 27, eliminando parámetros con registros insuficientes para garantizar representatividad estadística, entre ellos el Caudal, con apenas un 9.85% de datos válidos.

La imputación de valores faltantes mediante KNN con k=5 preservó la estructura estadística del conjunto sin introducir sesgos artificiales, estimando cada valor ausente a partir de los cinco registros más similares. Este paso fue determinante para garantizar que los modelos recibieran un conjunto de datos coherente y completo.

La selección de variables se ejecutó en tres etapas encadenadas. El filtro de correlación de Spearman con umbral de $|\rho| \geq 0.85$ eliminó cinco variables redundantes — entre ellas Dureza Magnésica, Dureza Cálcica y Alcalinidad Total — conservando en cada par la variable con mayor capacidad explicativa frente a las variables objetivo. El análisis de VIF detectó y eliminó los Nitritos, cuyo valor de 616.9 evidenciaba una dependencia casi total con los Nitratos. Finalmente, el esquema de votación del RFE con tres estimadores distintos — Regresión Lineal, Ridge y SVR lineal — generó cuatro escenarios experimentales que organizaron los 19 predictores resultantes según su relevancia específica para pH y DBO.

Este proceso de depuración progresiva no fue un trámite técnico — fue una decisión metodológica que determinó la calidad de todo lo que vino después. Un conjunto de variables redundantes o incompletas habría comprometido la estabilidad de los modelos y la interpretabilidad de sus resultados.

El conjunto de datos depurado y los escenarios de variables definidos establecieron las condiciones para la fase de modelado, donde seis algoritmos de distinta naturaleza competirían por capturar la dinámica fisicoquímica del sistema fluvial.

El entrenamiento comparativo de seis algoritmos — Regresión Lineal, CART, Random Forest, SVR, MLP y XGBoost — sobre los cuatro escenarios de variables permitió establecer con evidencia

empírica qué enfoques son más adecuados para predecir la calidad del agua en el Río Cauca. La diversidad metodológica fue deliberada: incluir modelos lineales junto a modelos de ensamble y redes neuronales permitió cuantificar la ganancia real que aporta la complejidad adicional.

Los resultados confirmaron que las relaciones entre los parámetros fisicoquímicos del río y las variables objetivo son marcadamente no lineales. Para la DBO, los modelos de ensamble y redes neuronales superaron consistentemente a la Regresión Lineal, con el XGBoost sobre el escenario RFE-pH alcanzando un R^2 de 0.903 antes de la optimización. Un hallazgo inesperado fue que el escenario RFE-DBO — diseñado específicamente para predecir la DBO — produjo los peores resultados entre los cinco modelos destacados, mientras que el escenario RFE-pH, construido con variables seleccionadas para el pH, generó las mejores predicciones. Este resultado sugiere que los factores del equilibrio químico del sistema — Bicarbonatos, Conductividad Eléctrica, Nitratos y Temperatura — tienen mayor capacidad explicativa sobre la carga orgánica del río que los indicadores biológicos directos seleccionados por el RFE para la DBO.

Tras la optimización de hiperparámetros mediante Grid Search y Optuna, el Bosque Aleatorio con el escenario RFE-pH emergió como el mejor modelo para DBO, alcanzando R^2=0.898 y RMSE=3.381 mg O_2/L — resultado que combina alta precisión con un conjunto reducido de apenas cinco variables predictoras. Para el pH, el XGBoost optimizado con Optuna sobre el escenario VIF obtuvo el mejor desempeño con R^2=0.411 y RMSE=0.297.

La diferencia entre ambas variables objetivo constituye uno de los hallazgos más relevantes del estudio. La DBO demostró ser predecible con alta precisión; el pH estableció un techo de predicción que la optimización de hiperparámetros no logró superar de forma significativa — el mejor R^2 antes de la calibración era 0.414 y después fue 0.411. La curva de aprendizaje del modelo de pH confirmó este diagnóstico: la brecha entre el error de entrenamiento y el de validación no se cierra con más datos, lo que indica que el límite predictivo no es un problema de volumen ni de configuración algorítmica, sino de la información contenida en las variables disponibles. El pH responde a equilibrios geológicos, biológicos e hidrológicos que el monitoreo actual del río no captura completamente.

Con los modelos seleccionados y su desempeño cuantificado, el análisis avanzó hacia la pregunta más importante desde el punto de vista ambiental: ¿qué variables están impulsando esas predicciones y qué nos dicen sobre la salud del río?

Conocer el desempeño de un modelo es necesario, pero no suficiente. En un sistema ambiental como el Río Cauca, donde los resultados deben orientar decisiones de monitoreo y gestión, la pregunta crítica no es solo cuánto acierta el modelo sino por qué acierta — qué variables está usando y si esa lógica interna tiene coherencia ecológica. Para responderla, se aplicaron dos técnicas complementarias de interpretabilidad: Importancia por Permutación y valores SHAP.

Para la DBO, los resultados fueron contundentes. La Importancia por Permutación identificó a los Sulfatos como el predictor dominante con un valor de 0.9534 — más de ocho veces superior al segundo predictor, la Temperatura (0.1184). Los valores SHAP confirmaron el liderazgo de los

Sulfatos (1.1106) pero revelaron un sistema más distribuido, donde Temperatura (0.9056), Bicarbonatos (0.8912) y Conductividad Eléctrica (0.6617) también contribuyen de forma sustancial. El análisis de distribución individual de los valores SHAP mostró que la mayor parte de las observaciones genera impactos moderados y predecibles, pero que existen episodios puntuales de concentraciones extremas de sulfatos donde el modelo detecta señales de contaminación intensa y ajusta la predicción de DBO de forma drástica. Esta dinámica episódica es consistente con las fuentes de presión antrópica documentadas en la cuenca: los Sulfatos son un marcador de drenaje ácido de minas y efluentes industriales, mientras que la Temperatura regula la velocidad de los procesos de descomposición microbiana que consumen el oxígeno disuelto.

Para el pH, el panorama fue radicalmente distinto. Tanto la Importancia por Permutación como los valores SHAP revelaron un sistema donde ninguna variable domina de forma absoluta. Los Bicarbonatos lideraron ambos análisis como reguladores del equilibrio buffer del agua, pero los Nitratos — reflejo de la actividad agrícola y el ciclo del nitrógeno — aparecieron esencialmente empatados en el análisis SHAP (0.0838 vs 0.0824). Variables adicionales como el Hierro Total — cuya oxidación genera iones H^+ que acidifican el medio en contextos mineros — aportaron señales secundarias que el modelo capturó implícitamente. La distribución de los valores SHAP individuales confirmó la ausencia de eventos extremos: el pH se regula de forma continua y multifactorial, sin los picos episódicos que caracterizan a la DBO.

La comparación entre ambas variables a través de las técnicas de interpretabilidad estableció una conclusión ambiental de fondo: la carga orgánica del Río Cauca está dominada por eventos puntuales de contaminación industrial y minera, mientras que el equilibrio ácido-base del sistema responde a procesos difusos y permanentes de naturaleza geológica, agrícola y biológica. Ambos fenómenos requieren estrategias de monitoreo y gestión diferenciadas.

Con la interpretabilidad de los modelos establecida y sus implicaciones ambientales identificadas, el análisis cierra con una visión integradora que conecta los hallazgos técnicos con las necesidades del sistema de monitoreo del Río Cauca.

Los resultados de este estudio no son solo métricas de desempeño — son una lectura cuantitativa de la dinámica ambiental del Río Cauca. Tomados en conjunto, los modelos de predicción de DBO y pH configuran un retrato coherente de un sistema hídrico bajo presión antrópica múltiple, donde la carga orgánica y el equilibrio ácido-base responden a fuentes de contaminación distintas, con dinámicas distintas y horizontes de predicción distintos.

La superioridad del Bosque Aleatorio para la DBO y del XGBoost para el pH frente a modelos lineales confirma lo que la revisión de Zhu et al. [36] establece para el estado del arte global: los métodos de ensamble son consistentemente los más eficaces para capturar las relaciones no lineales propias de los sistemas de calidad del agua, y su combinación con técnicas robustas de selección de variables — como el esquema de votación RFE implementado en este estudio — potencia aún más su capacidad predictiva. El R^2=0.898 alcanzado para la DBO se sitúa en el rango

superior de los resultados reportados en la literatura para parámetros de demanda de oxígeno en ríos.

El caso del pH merece una lectura diferente. El techo de predicción de R^2=0.411 no es una anomalía metodológica — es una propiedad documentada del sistema. Bueno-Zabala et al. [34], en su análisis del agua tratada del Río Cauca mediante índices de estabilización, identificaron que el agua de esta cuenca tiende a ser naturalmente insaturada y corrosiva debido a sus bajos niveles de alcalinidad total. Las variables que emplearon para caracterizar ese equilibrio inestable — bicarbonatos, calcio, conductividad eléctrica, temperatura — son exactamente las mismas que los modelos SHAP identificaron como los predictores más influyentes del pH en este estudio. Esta convergencia entre un enfoque fisicoquímico clásico y un modelo de aprendizaje automático sobre el mismo sistema hídrico no es casual: es evidencia de que los modelos están capturando señales reales del río, no artefactos estadísticos. Y si aun así el R^2 no supera el 0.41, es porque la complejidad natural del pH del Río Cauca excede lo que el conjunto de variables monitoreadas actualmente puede representar.

Este estudio demostró que el aprendizaje automático es una herramienta válida, interpretable y ambientalmente coherente para el monitoreo predictivo de la calidad del agua del Río Cauca. La combinación de un pipeline reproducible — desde la depuración de datos hasta la interpretabilidad con SHAP — con una estrategia de selección de variables metodológicamente rigurosa, produce modelos que no solo predicen, sino que explican. Y en un sistema tan complejo y presionado como este río, explicar es tan importante como predecir.

TRABAJOS FUTUROS

Los resultados de este estudio abren líneas de trabajo concretas, identificadas precisamente por los límites que el análisis encontró.

La predicción del pH es el frente más urgente. Con un R^2=0.411 como techo actual, la mejora más prometedora no pasa por cambiar el algoritmo — pasa por ampliar el conjunto de variables. Incorporar parámetros que capturen directamente el equilibrio carbonatado del sistema, como el CO_2 disuelto o índices de estabilización derivados de los trabajos de Bueno-Zabala et al. [34] — Índice de Saturación de Langelier, Índice de Estabilidad de Ryznar — podría aportar información que las variables fisicoquímicas convencionales no logran representar. De igual forma, el Caudal, eliminado en este estudio por su baja completitud (9.85%), es un predictor hidrológico de alto valor potencial: un programa de monitoreo que garantice registros continuos de caudal permitiría incorporarlo en futuras versiones del modelo.

En cuanto al volumen de datos, la curva de aprendizaje de la DBO mostró que el modelo aún se beneficia de registros adicionales. Ampliar la frecuencia de muestreo, incorporar nuevas estaciones de monitoreo o integrar datos de afluentes clave de la cuenca podría reducir el RMSE de validación y acercar ambas curvas hacia la convergencia. Esto es especialmente relevante para el pH, cuya curva de aprendizaje no mostró mejora con más datos bajo el conjunto de variables

actual — pero que podría responder de forma diferente con predictores más informativos.

Una dimensión no explorada en este estudio es la espacial. Los datos se trataron como un conjunto homogéneo sin distinguir entre estaciones de monitoreo, lo que implica que el modelo aprende un comportamiento promedio del río. Desarrollar modelos específicos por tramo o por estación permitiría capturar dinámicas locales — como la influencia de afluentes contaminados, zonas de mezcla o tramos con mayor presión minera — que hoy quedan invisibles en el promedio global.

Desde el punto de vista metodológico, explorar arquitecturas de aprendizaje profundo con memoria temporal — como redes LSTM (*Long Short-Term Memory*) — representaría un paso natural hacia modelos que no solo usen el estado actual del agua sino su trayectoria reciente. La calidad del agua en un río es un proceso dinámico: lo que ocurrió ayer condiciona lo que ocurre hoy, y los modelos que ignoran esa dimensión temporal dejan sobre la mesa información valiosa.

Finalmente, el pipeline desarrollado en este estudio — depuración, imputación, selección de variables, entrenamiento, optimización e interpretabilidad — constituye una base reproducible que podría extenderse a otras variables objetivo del Río Cauca, como el Oxígeno Disuelto o la Turbiedad, y adaptarse a otras cuencas hidrográficas colombianas con estructuras de monitoreo similares. La transferibilidad metodológica es, en sí misma, una contribución que los estudios futuros podrían validar y expandir.

8. REFERENCIAS BIBLIOGRÁFICAS

[1] K. Wu *et al.*, «Drivers of 2013–2020 ozone trends in the Sichuan Basin, China: Impacts of meteorology and precursor emission changes», *Environ. Pollut.*, vol. 300, may 2022, doi: 10.1016/j.envpol.2022.118914.

[2] S. Chidiac, P. E. Najjar, N. Ouaini, Y. E. Rayess, y D. E. Azzi, «A comprehensive review of water quality indices (WQIs): history, models, attempts and perspectives», *Reviews in Environmental Science and Biotechnology*, vol. 22, n.º 2. Springer Science and Business Media B.V., pp. 349-395, junio de 2023. doi: 10.1007/s11157-023-09650-7.

[3] Q. Xu, Y. Shi, J. Bamber, Y. Tuo, R. Ludwig, y X. X. Zhu, «Physics-aware Machine Learning Revolutionizes Scientific Paradigm for Machine Learning and Process-based Hydrology», jul. 2024, [En línea]. Disponible en: http://arxiv.org/abs/2310.05227

[4] D. Dudgeon *et al.*, «Freshwater biodiversity: Importance, threats, status and conservation challenges», *Biological Reviews of the Cambridge Philosophical Society*, vol. 81, n.º 2. pp. 163-182, mayo de 2006. doi: 10.1017/S1464793105006950.

[5] Y. Shen, H. Cao, M. Tang, y H. Deng, «The human threat to river ecosystems at the watershed scale: An ecological security assessment of the songhua river basin, Northeast China», *Water Switz.*, vol. 9, n.º 3, 2017, doi: 10.3390/w9030219.

[6] G. Wang *et al.*, «Integrated watershed management: evolution, development and emerging trends», *J. For. Res.*, vol. 27, n.º 5, pp. 967-994, oct. 2016, doi: 10.1007/s11676-016-0293-3.

[7] M. Meybeck, «Looking for water quality», *Hydrol. Process.*, vol. 19, n.º 1, pp. 331-338, ene. 2005, doi: 10.1002/hyp.5778.

[8] S. K. Dewangan, S. K. Shrivastava, V. Tigga, y M. Lakra, «Review Paper on the Role of pH in Water Quality: Implications for Aquatic Life, Human Health, and Environmental Sustainability», *Int. Adv. Res. J. Sci. Eng. Technol. ISO*, vol. 3297, 2007, doi: 10.17148/IARJSET.2023.10633.

[9] P. Dixneuf, F. Errico, y M. Glaus, «A computational study on imputation methods for missing environmental data», 2021.

[10] S. K. Kwak y J. H. Kim, «Statistical data preparation: Management of missing values and outliers», *Korean Journal of Anesthesiology*, vol. 70, n.º 4. Korean Society of Anesthesiologists, pp. 407-411, agosto de 2017. doi: 10.4097/kjae.2017.70.4.407.

[11] L. Sasse *et al.*, «On Leakage in Machine Learning Pipelines», 2002. [En línea]. Disponible en: https://hunch.net/?p=22

[12] S. Senthilnathan, «Usefulness of correlation analysis», International Training Institute Papua New Guine, 2019. [En línea]. Disponible en: https://ssrn.com/abstract=3416918https://ssrn.com/abstract=34169182https://ssrn.com/a

bstract=3416918

[13] N. V. Sharma y N. S. Yadav, «An optimal intrusion detection system using recursive feature elimination and ensemble of classifiers», *Microprocess. Microsyst.*, vol. 85, p. 104293, sep. 2021, doi: 10.1016/j.micpro.2021.104293.

[14] S. Kapoor y A. Narayanan, «Leakage and the reproducibility crisis in machine-learning-based science», *Patterns*, vol. 4, n.º 9, sep. 2023, doi: 10.1016/j.patter.2023.100804.

[15] C. Lavanya, M. Nikitha, N. Swetha, K. Nikhitha, y L. Hussein, «Assessment and estimation of water quality using multi-linear regression», en *E3S Web of Conferences*, EDP Sciences, may 2024. doi: 10.1051/e3sconf/202452903007.

[16] H. Blockeel, L. Devos, B. Frénay, G. Nanfack, y S. Nijssen, «Decision trees: from efficient prediction to responsible AI», *Frontiers in Artificial Intelligence*, vol. 6. Frontiers Media SA, 2023. doi: 10.3389/frai.2023.1124553.

[17] E. Halabaku y E. Bytyçi, «Overfitting in Machine Learning: A Comparative Analysis of Decision Trees and Random Forests», *Intell. Autom. Soft Comput.*, vol. 39, n.º 6, pp. 987-1006, 2024, doi: 10.32604/iasc.2024.059429.

[18] S. M. Alomani, N. I. Alhawiti, y A. Alhakamy, «Prediction of Quality of Water According to a Random Forest Classifier», *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, n.º 6, pp. 892-899, 2022, doi: 10.14569/IJACSA.2022.01306105.

[19] J. Xu *et al.*, «An alternative to laboratory testing: Random forest-based water quality prediction framework for inland and nearshore water bodies», *Water Switz.*, vol. 13, n.º 22, nov. 2021, doi: 10.3390/w13223262.

[20] X. Su, X. He, G. Zhang, Y. Chen, y K. Li, «Research on SVR Water Quality Prediction Model Based on Improved Sparrow Search Algorithm», *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/7327072.

[21] W. Zhi, A. P. Appling, H. E. Golden, J. Podgorski, y L. Li, «Deep learning for water quality», *Nature Water*, vol. 2, n.º 3. Springer Nature, pp. 228-241, marzo de 2024. doi: 10.1038/s44221-024-00202-z.

[22] D. Costa *et al.*, «Water quality estimates using machine learning techniques in an experimental watershed», *J. Hydroinformatics*, vol. 26, n.º 11, pp. 2798-2814, nov. 2024, doi: 10.2166/hydro.2024.132.

[23] T. Chen y C. Guestrin, «XGBoost: A scalable tree boosting system», en *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, ago. 2016, pp. 785-794. doi: 10.1145/2939672.2939785.

[24] M. Imani, A. Beikmohammadi, y H. R. Arabnia, «Comprehensive Analysis of Random Forest

and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels», *Technologies*, vol. 13, n.º 3, mar. 2025, doi: 10.3390/technologies13030088.

[25] T. Akiba, S. Sano, T. Yanase, T. Ohta, y M. Koyama, «Optuna: A Next-generation Hyperparameter Optimization Framework», en *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, jul. 2019, pp. 2623-2631. doi: 10.1145/3292500.3330701.

[26] R. Turner *et al.*, «Bayesian Optimization is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020», ago. 2021, [En línea]. Disponible en: http://arxiv.org/abs/2104.10201

[27] S. Shekhar, A. Bansode, y A. Salim, «A Comparative study of Hyper-Parameter Optimization Tools», ene. 2022, [En línea]. Disponible en: http://arxiv.org/abs/2201.06433

[28] A. C. A. Aguilar y F. F. O.- Díaz, «APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DE CALIDAD DE AGUA POTABLE», *Ingeniare*, n.º 28, pp. 47-62, may 2020, doi: 10.18041/1909-2458/ingeniare.28.6215.

[29] R. Choudhary, A. Kumar, P. C, M. M. Naik, M. Choudhury, y N. A. Khan, «Predicting water quality index using stacked ensemble regression and SHAP based explainable artificial intelligence», *Sci. Rep.*, vol. 15, n.º 1, dic. 2025, doi: 10.1038/s41598-025-09463-4.

[30] T. Viering y M. Loog, «The Shape of Learning Curves: a Review», nov. 2022, [En línea]. Disponible en: http://arxiv.org/abs/2103.10948

[31] C. Molnar, «Interpretable Machine Learning A Guide for Making Black Box Models Explainable», 2019. [En línea]. Disponible en: http://leanpub.com/interpretable-machine-learning

[32] S. Lundberg y S.-I. Lee, «A Unified Approach to Interpreting Model Predictions», nov. 2017, [En línea]. Disponible en: http://arxiv.org/abs/1705.07874

[33] A. Altmann, L. Toloşi, O. Sander, y T. Lengauer, «Permutation importance: A corrected feature importance measure», *Bioinformatics*, vol. 26, n.º 10, pp. 1340-1347, abr. 2010, doi: 10.1093/bioinformatics/btq134.

[34] K. A. Bueno-Zabala, P. Torres-Lozada, y L. G. Delgado-Cabrera, «MONITOREO Y MEDICIÓN DEL AJUSTE DEL pH DEL AGUA TRATADA DEL RÍO CAUCA MEDIANTE ÍNDICES DE ESTABILIZACIÓN», 2014.

[35] I. D. López, A. Figueroa, y J. C. Corrales, «Un mapeo sistemático sobre predicción de calidad del agua mediante técnicas de inteligencia computacional», *Rev. Ing. Univ. Medellín*, vol. 15, n.º 28, pp. 35-52, 2016, doi: 10.22395/rium.v15n28a2.

[36] M. Zhu *et al.*, «A review of the application of machine learning in water quality evaluation», *Eco-Environ. Health*, vol. 1, n.º 2, p. 107, jun. 2022, doi: 10.1016/J.EEHL.2022.06.001.

[37] A. O. Al-Sulttani, S. K. Ali, A. A. Abdulhameed, y D. T. Jassim, «Artificial Neural Network Assessment of Groundwater Quality for Agricultural Use in Babylon City: An Evaluation of Salinity and Ionic Composition», *Int. J. Des. Nat. Ecodynamics*, vol. 19, n.º 1, pp. 329-336, feb. 2024, doi: 10.18280/ijdne.190136.

[38] M. J. Islam *et al.*, «Machine Learning-Driven Water Quality Index Prediction: Enhancing Accuracy with Gradient Boosting and Explainable AI for Sustainable Water Monitoring», *Appl. Agric. Sci.*, vol. 2, n.º 1, pp. 1-7, feb. 2024, doi: 10.25163/agriculture.2110031.

[39] J. M. Z. Hoque, N. A. Nor, S. Alelyani, M. Mohana, y M. Hosain, «Improving Water Quality Index Prediction Using Regression Learning Models», *Int. J. Environ. Res. Public. Health*, vol. 19, n.º 20, oct. 2022, doi: 10.3390/ijerph192013702.

[40] I. D. L. Gómez, «Anexo B Predicción Adaptativa de la Calidad del Agua Mediante Técnicas de Inteligencia Computacional», Universidad del Cauca; Universidad del Valle, oct. 2015.

[41] R. B. Baird, E. W. Rice, y S. Posavec, *Standard Methods for the Examination of Water and Wastewater*, 23rd ed. Washington, DC, USA: APHA Press, 2017. doi: 10.2105/SMWW.2882.216.

[42] F. Suzanne, «QUALITY ASSURANCE PROJECT PLAN FOR OARS' Water Quality and Quantity Monitoring Program», dic. 2018.

ANEXO A. Diccionario de variables para proyectos similares

Tabla A.1. Glosario y descripción técnica de variables y términos de modelación

SIGLA / TÉRMINO	SIGNIFICADO	DESCRIPCIÓN TÉCNICA Y APLICACIÓN EN EL MODELO
ANN	Artificial Neural Network (Red Neuronal Artificial)	Conjunto de modelos de aprendizaje profundo inspirados en el cerebro humano. Permiten capturar patrones no lineales y complejos en grandes volúmenes de datos.
Boosting	—	Estrategia de aprendizaje que combina múltiples modelos débiles para crear un modelo fuerte. Cada iteración corrige los errores de la anterior, mejorando la predicción final.
Ca	Calcio	Ion metálico que contribuye a la dureza del agua. Es clave en la predicción de índices de estabilización y balance iónico.
Ca^{2+}	Ion Calcio	Ion bivalente determinante en la dureza total y en la estabilidad química. Se usa en índices de equilibrio y salinidad.
DBO	Demanda Bioquímica de Oxígeno	Representa la cantidad de oxígeno requerida por los microorganismos para descomponer materia orgánica. Se usa para evaluar contaminación y como variable dependiente en modelos de predicción de calidad.
DT	Decision Tree (Árbol de Decisión)	Modelo jerárquico que clasifica o predice valores dividiendo el conjunto de datos en nodos basados en criterios de impureza. Interpretación sencilla, pero sensible al sobreajuste.
EC	Conductividad Eléctrica (Electrical Conductivity)	Mide la capacidad del agua para conducir electricidad, directamente relacionada con la cantidad de iones disueltos. Es una de las variables predictoras más empleadas.

SIGLA / TÉRMINO	SIGNIFICADO	DESCRIPCIÓN TÉCNICA Y APLICACIÓN EN EL MODELO
Hiperparámetro	—	Parámetro ajustado antes del entrenamiento del modelo (como learning rate o depth). Determina la capacidad del modelo para generalizar sin sobreajustar los datos.
K	Potasio	Catión natural asociado con fertilización o descargas orgánicas. Aporta información sobre procesos de eutrofización o contaminación difusa.
KNN	K-Nearest Neighbors	Algoritmo supervisado que predice valores en función de la similitud con sus vecinos más cercanos en el espacio de características. Útil como modelo base comparativo.
MAR	Relación de Magnesio (Magnesium Adsorption Ratio)	Índice complementario al SAR, empleado para medir la influencia del magnesio sobre la alcalinidad y la permeabilidad del suelo.
Mg^{2+}	Ion Magnesio	Contribuye a la dureza del agua y afecta la salinidad total. Se usa en modelos de predicción de índices iónicos y calidad para riego.
MLP	Multilayer Perceptron (Perceptrón Multicapa)	Red neuronal con varias capas ocultas que aprenden representaciones complejas mediante retropropagación del error. Adecuada para conjuntos grandes y no lineales.
Na	Sodio	Ion alcalino importante para evaluar la salinidad. Su concentración afecta la aptitud del agua para riego agrícola y la estabilidad química del recurso.
Na^{+}	Ion Sodio	Ion monovalente que incrementa la salinidad del agua. Valores altos pueden afectar la estructura del suelo y la infiltración del agua.

SIGLA / TÉRMINO	SIGNIFICADO	DESCRIPCIÓN TÉCNICA Y APLICACIÓN EN EL MODELO
OD	Oxígeno Disuelto	Cantidad de oxígeno presente en el agua. Es un indicador de la capacidad del agua para sustentar vida acuática y un predictor relevante en modelos de calidad.
pH	Potencial de Hidrógeno	Indica la acidez o alcalinidad del agua. En modelos predictivos, se utiliza como variable objetivo o de entrada clave, ya que influye directamente en los procesos químicos y biológicos del agua.
R^2	Coeficiente de Determinación	Métrica estadística que indica qué proporción de la variabilidad observada en los datos es explicada por el modelo. Valores cercanos a 1 indican mejor ajuste.
Regularización (L1/L2)	—	Métodos para penalizar la complejidad del modelo, evitando el sobreajuste al limitar la magnitud de los coeficientes o pesos internos.
RF	Random Forest	Ensamble de múltiples árboles de decisión que combina predicciones promediadas. Reduce la varianza y mejora la estabilidad frente al ruido.
RLM	Regresión Lineal Múltiple	Modelo estadístico que asume una relación lineal entre las variables predictoras y la respuesta. Se usa como modelo de referencia por su interpretabilidad.
RMSE	Raíz del Error Cuadrático Medio (Root Mean Square Error)	Indicador de precisión del modelo; expresa el promedio de las desviaciones entre valores observados y predichos. Cuanto menor el RMSE, mayor la exactitud del modelo.
SAR	Relación de Adsorción de Sodio (Sodium Adsorption	Índice agronómico que evalúa el efecto del sodio sobre la estructura del suelo. En modelos predictivos, ayuda a determinar la aptitud del

SIGLA / TÉRMINO	SIGNIFICADO	DESCRIPCIÓN TÉCNICA Y APLICACIÓN EN EL MODELO
	Ratio)	agua para riego.
SHAP	SHapley Additive exPlanations	Técnica para interpretar modelos de machine learning que asigna y muestra la importancia cuantitativa de cada variable en las predicciones, ayudando a identificar las más influyentes y a mejorar la transparencia del modelo.
SO_4^{2-}	Ion Sulfato	Anión derivado de la disolución de sales minerales. Afecta la conductividad eléctrica y puede contribuir a la acidez del agua en exceso.
Sobreajuste (Overfitting)	—	Ocurre cuando el modelo se ajusta demasiado a los datos de entrenamiento y pierde capacidad para generalizar. Es evaluado mediante validación cruzada.
Stacking	—	Técnica de ensamble que combina diferentes algoritmos (por ejemplo, RF, XGBoost y SVR) apilando sus salidas para generar una predicción más precisa.
SVM	Support Vector Machine	Método que busca hiperplanos óptimos para separar o ajustar datos. En su versión de regresión (SVR) se adapta a valores continuos.
SVR	Support Vector Regression	Extiende el SVM para predecir valores numéricos. Utiliza funciones núcleo (kernels) para modelar relaciones no lineales en espacios de alta dimensión.
T	Temperatura	Parámetro físico que afecta la solubilidad del oxígeno y la actividad biológica. Suele ser una variable independiente crítica en la modelación

SIGLA / TÉRMINO	SIGNIFICADO	DESCRIPCIÓN TÉCNICA Y APLICACIÓN EN EL MODELO
		de sistemas hídricos.
TDS	Sólidos Disueltos Totales (Total Dissolved Solids)	Mide la concentración total de sales, minerales y materia disuelta. Se utiliza para estimar salinidad y conductividad, influyendo en el WQI.
Validación Cruzada (Cross-Validation)	—	Procedimiento para evaluar el rendimiento de un modelo usando múltiples particiones de datos de entrenamiento y prueba, garantizando estabilidad en las métricas.
WQI	Índice de Calidad del Agua (Water Quality Index)	Valor numérico integrado que resume varios parámetros (pH, DBO, OD, TDS, etc.) para expresar la calidad general del agua. En modelos predictivos, puede ser la variable objetivo o una medida de salida agregada.
XGBoost	Extreme Gradient Boosting	Método de ensamble basado en boosting secuencial. Corrige errores iterativamente y aplica regularización L1/L2 para evitar sobreajuste. Destacado por su alta precisión y eficiencia computacional.
CatBoost	Categorical Boosting	Variante del gradient boosting optimizada para variables categóricas. Utiliza un esquema de ordenamiento estadístico que evita la sobrecodificación (target leakage) y mejora la estabilidad del entrenamiento. Implementa un enfoque simétrico en la construcción de árboles que acelera la convergencia.
GBM	Gradient Boosting Machine	Método de ensamble que combina múltiples árboles débiles (de poca profundidad) entrenados secuencialmente. Cada nuevo árbol se ajusta al gradiente negativo de la función de pérdida del modelo anterior, minimizando los

SIGLA / TÉRMINO	SIGNIFICADO	DESCRIPCIÓN TÉCNICA Y APLICACIÓN EN EL MODELO
		errores progresivamente.
Extra Trees	Extremely Randomized Trees	Ensamble de árboles de decisión donde la división de nodos se realiza de forma aleatoria tanto en las variables como en los umbrales. A diferencia de Random Forest, no utiliza bagging, sino que entrena con todo el conjunto de datos.
AdaBoost	Adaptive Boosting	Método secuencial de boosting que ajusta los pesos de las observaciones en función de los errores de predicción. Los clasificadores posteriores se centran en los casos mal predichos. Utiliza una combinación ponderada de modelos débiles (habitualmente árboles simples) para construir un modelo final más robusto.

ANEXO B. Estadísticas de las variables fisicoquímicas del Rio Cauca

Tabla B.1. Estadísticos básicos de tendencia central y dispersión de las variables fisicoquímicas del Río Cauca

VARIABLE	Tipo de dato	N.Valid	Pct.Valid	Mean	Std.Dev	Min	Q1	Median
pH	float64	2208	97.96	7.04	0.41	4.10	6.80	7.07
Temperatura (°C)	float64	1956	86.78	22.02	3.60	0	21	22.40
Color (UPC)	float64	1882	83.50	135.45	244.58	0	38.50	70
Turbiedad (UNT)	float64	2090	92.72	130.73	171.03	1	30	65
Solidos Totales (mg SST/l)	float64	2194	97.34	275.19	227.62	0	142	206
Solidos Suspendidos Totales (mg SS/l)	float64	2188	97.07	165.20	202.91	1.20	41	92.65
Solidos Disueltos (mg SD/l)	float64	2190	97.16	117.51	63.64	0	83	108
Demanda Bioquímica de Oxigeno (mg O2/l)	float64	2092	92.81	5.22	16.25	0.10	2.08	3.10
Demanda Química de Oxigeno (mg O2/l)	float64	2066	91.66	29.89	37.29	1.46	14.10	20.73
Oxígeno Disuelto (mg O2/l)	float64	2179	96.67	4.10	2.96	0	2.34	3.84
Dureza Total (mg CaCO3/l)	float64	2167	96.14	49.35	21.40	0	35.20	46
Dureza Cálcica (mg CaCO3/l)	float64	2166	96.10	27.37	11.50	0	19.20	26

VARIABLE	Tipo de dato	N.Valid	Pct.Valid	Mean	Std.Dev	Min	Q1	Median
Dureza Magnésica (mg CaCO3/l)	float64	2167	96.14	23.81	17.51	0	15.40	20.70
Calcio (mg Ca/l)	float64	2166	96.10	11.89	10.45	1.60	7.80	10.40
Magnesio (mg Mg/l)	float64	2167	96.14	6.32	8.88	0	3.70	5
Alcalinidad Total (mg CaCO3/l)	float64	2159	95.79	37.53	38.45	0	22.10	35
Bicarbonatos (mg CaCO3/l)	float64	2158	95.74	42.93	24.26	0	25.60	40.20
Conductividad Eléctrica (µS/cm)	float64	2196	97.43	124.34	107.31	0	86.90	116.25
Hierro Total (mg Fe/l)	float64	2062	91.48	8.23	15.77	0.01	1.76	3.67
Manganeso Total (mg Mn/l)	float64	1594	70.72	0.50	1.15	0	0.12	0.21
Sodio Total (mg Na/l)	float64	1945	86.29	6.77	6.35	0	4.23	5.98
Potasio Total (mg K/l)	float64	1984	88.02	1.98	1.30	0	1.30	1.79
Cobre Total (mg Cu/l)	float64	537	23.82	0.06	0.10	0	0.02	0.04
Zinc Total (mg Zn/l)	float64	1039	46.10	0.43	1.36	0	0.05	0.09
Nitrógeno Total (mg N/l)	float64	1322	58.65	2.28	1.18	0	1.54	2.17
Nitrógeno Amoniacal (mg N-NH3/l)	float64	1288	57.14	0.98	0.91	0	0.35	0.91
Nitritos (mg N-NO2/l)	float64	1943	86.20	1.2×10^{5}	3.60×10^{6}	0	0.01	0.02

VARIABLE	Tipo de dato	N.Valid	Pct.Valid	Mean	Std.Dev	Min	Q1	Median
Nitratos (mg N-NO3/l)	float64	1958	86.87	1.15×10^{5}	3.58×10^{6}	0	0.19	0.38
Cloruros (mg Cl/l)	float64	2083	92.41	7.24	20.58	0.17	4.25	5.99
Fosforo Total (mg P/l)	float64	1797	79.72	0.30	0.89	0	0.11	0.18
Fosfatos (mg PO4/l)	float64	1113	49.38	0.10	0.63	0	0.04	0.07
Sulfatos (mg SO4/l)	float64	1932	85.71	18.76	17.88	0	13.84	17.43
Cadmio Total (mg Cd/l)	float64	100	4.44	3.40	7.26	0	0.00	0.02
Cromo Total (mg Cr/l)	float64	195	8.65	0.12	0.16	0	0.01	0.06
Níquel Total (mg Ni/l)	float64	206	9.14	0.12	0.55	0	0	0.04
Plomo Total (mg Pb/l)	float64	172	7.63	0.28	0.28	0	0.01	0.22
Coliformes Totales (NMP/100 ml)	float64	1814	80.48	5.6×10^{9}	1.1×10^{11}	0	24000	210000
Coliformes Fecales (NMP/100 ml)	float64	1870	82.96	1.2×10^{13}	5.5×10^{14}	0	4850	42000
Caudal (m3/s)	float64	222	9.85	1.2×10^{5}	6.4×10^{5}	1	156	233.50

Tabla B.2. Estadísticos complementarios de forma y variabilidad de las variables fisicoquímicas del Río Cauca

VARIABLE	Q3	MAX	IQR	MAD	CV	SKEWNESS	SE.SKEWNESS	KURTOSIS
pH	7.3	9.7	0.48	0.29	0.06	-0.10	0.05	5.23
Temperatura (°C)	24.1	32.7	3.10	2.35	0.16	-2.34	0.05	8.93
Color (UPC)	134.7	2956	96.2	117.34	1.81	6.16	0.05	49.53
Turbiedad (UNT)	156.1	1900	126.1	114.8	1.31	2.86	0.05	12.12
Solidos Totales (mg SST/l)	324	2361	182	149.99	0.83	3.14	0.05	15.69
Solidos Suspendidos Totales (mg SS/l)	216.2	2112	175.2	136	1.23	3.06	0.05	14.35
Solidos Disueltos (mg SD/l)	136	864.4	53	39	0.54	3.89	0.05	28.74
Demanda Bioquímica de Oxigeno (mg O2/l)	4.80	427	2.72	3.9	3.11	16.25	0.05	333.46
Demanda Química de Oxigeno (mg O2/l)	32	706	18	18.67	1.25	7.24	0.05	87.94
Oxígeno Disuelto (mg O2/l)	5.7	51.5	3.3	1.85	0.72	5.94	0.05	68.07
Dureza Total (mg CaCO3/l)	60	220	24.8	15.7	0.43	1.66	0.05	7.28
Dureza Cálcica (mg CaCO3/l)	33.6	126.5	14.4	8.67	0.42	1.54	0.05	6.29
Dureza Magnésica (mg CaCO3/l)	28.25	243.50	12.85	10.13	0.74	4.86	0.05	40.78
Calcio (mg Ca/l)	13.60	189.60	5.80	4.46	0.88	9.79	0.05	131.68

VARIABLE	Q3	MAX	IQR	MAD	CV	SKEWNESS	SE.SKEWNESS	KURTOSIS
Magnesio (mg Mg/l)	6.86	168.80	3.16	3.19	1.40	10.89	0.05	155.51
Alcalinidad Total (mg CaCO3/l)	47.40	1502	25.30	15.36	1.02	26.49	0.05	980.28
Bicarbonatos (mg CaCO3/l)	55.63	343	30.03	16.71	0.57	4.06	0.05	37.59
Conductividad Eléctrica (µS/cm)	151	4259	64.10	38.48	0.86	28.11	0.05	1030.2
Hierro Total (mg Fe/l)	8.94	186	7.18	7.59	1.92	6.80	0.05	59.58
Manganeso Total (mg Mn/l)	0.37	12.95	0.25	0.50	2.30	5.48	0.05	35.13
Sodio Total (mg Na/l)	8.20	198	3.97	2.75	0.94	17.32	0.05	459.41
Potasio Total (mg K/l)	2.34	24.52	1.04	0.75	0.66	6.57	0.05	83.59
Cobre Total (mg Cu/l)	0.08	1.71	0.06	0.05	1.61	9.93	0.05	147.86
Zinc Total (mg Zn/l)	0.17	21.52	0.12	0.57	3.18	7.87	0.05	90.47
Nitrógeno Total (mg N/l)	2.80	20.83	1.26	0.79	0.52	4.27	0.05	51.69
Nitrógeno Amoniacal (mg N-NH3/l)	1.48	15.70	1.13	0.66	0.93	4.07	0.05	55.24
Nitritos (mg N-NO2/l)	0.07	1.5×10^{8}	0.06	2.4×10^{5}	29.6	38.46	0.05	1560.9
Nitratos (mg N-NO3/l)	0.68	1.5×10^{8}	0.49	2.3×10^{5}	31.1	38.86	0.05	1587.2
Cloruros (mg Cl/l)	8.22	738	3.97	3.39	2.84	29.46	0.05	944.95

VARIABLE	Q3	MAX	IQR	MAD	CV	SKEWNESS	SE.SKEWNESS	KURTOSIS
Fosforo Total (mg P/l)	0.27	14.18	0.16	0.24	2.94	9.51	0.05	97.39
Fosfatos (mg PO4/l)	0.10	20.10	0.07	0.08	5.97	29.95	0.05	943.65
Sulfatos (mg SO4/l)	21.60	696.80	7.76	5.78	0.95	29.17	0.05	1079.9
Cadmio Total (mg Cd/l)	0.04	24.40	0.04	5.48	2.14	1.80	0.05	1.51
Cromo Total (mg Cr/l)	0.14	1.00	0.13	0.12	1.37	2.29	0.05	6.36
Níquel Total (mg Ni/l)	0.11	7.70	0.11	0.13	4.72	12.92	0.05	176.44
Plomo Total (mg Pb/l)	0.48	1.54	0.47	0.24	1.01	0.97	0.05	1.42
Coliformes Totales (NMP/100 ml)	2.1×10^{6}	2.4×10^{12}	2.1×10^{6}	1.1×10^{10}	19.9	21.09	0.05	445.09
Coliformes Fecales (NMP/100 ml)	2.4×10^{5}	2.4×10^{16}	2.3×10^{5}	2.6×10^{13}	43.2	43.24	0.05	1870
Caudal (m3/s)	401.7	6.6×10^{6}	2.5×10^{2}	2.2×10^{5}	5.5	7.05	0.05	57.67

ANEXO C. Graficas de histogramas con asimetría

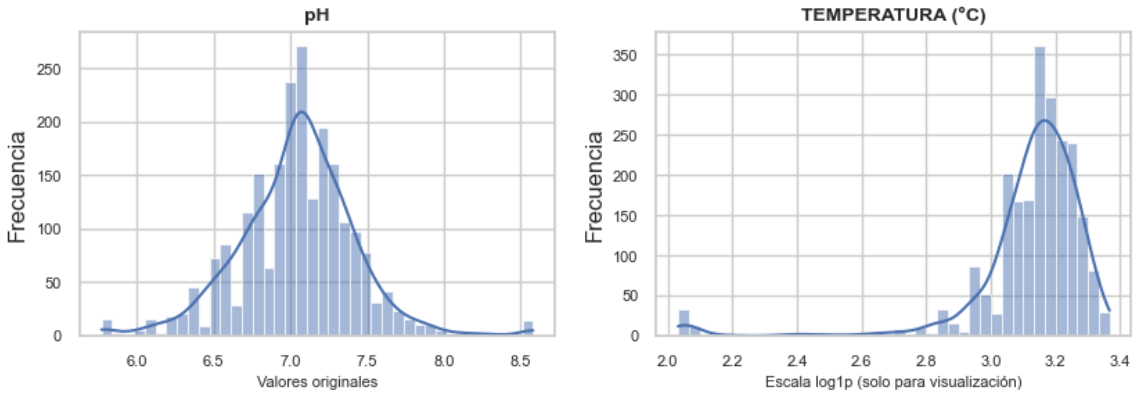

Figura C.1. Análisis de asimetría mediante histogramas de frecuencia para las variables pH y temperatura

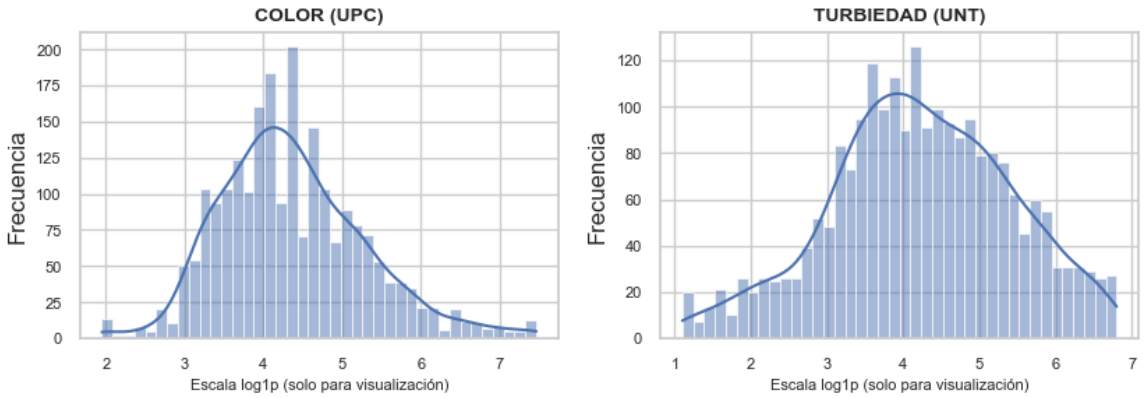

Figura C.2. Análisis de asimetría mediante histogramas de frecuencia para las variables color y turbiedad

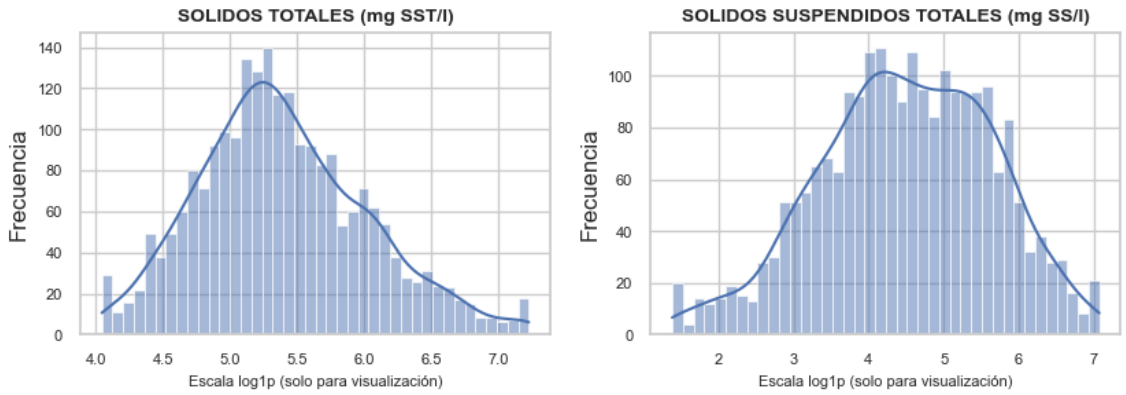

Figura C.3. Análisis de asimetría mediante histogramas de frecuencia para las variables sólidos totales y sólidos suspendidos totales

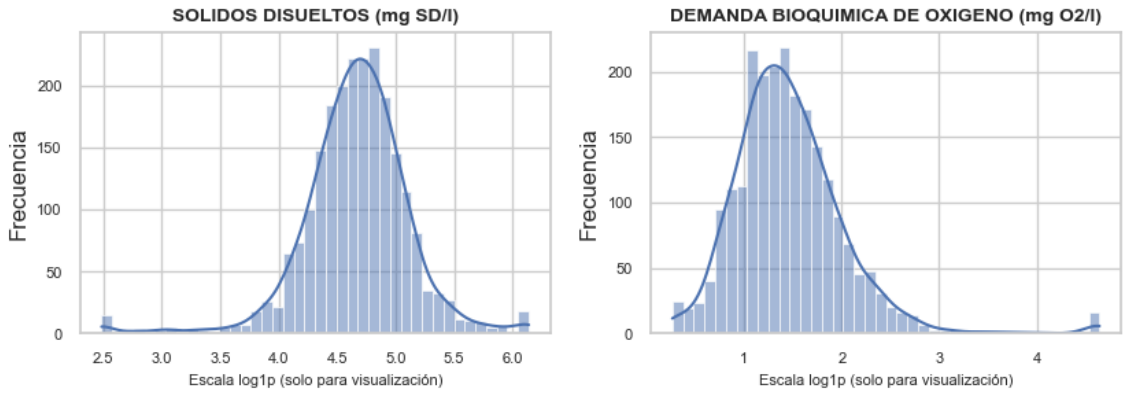

Figura C.4. Análisis de asimetría mediante histogramas de frecuencia para las variables sólidos disueltos y DBO

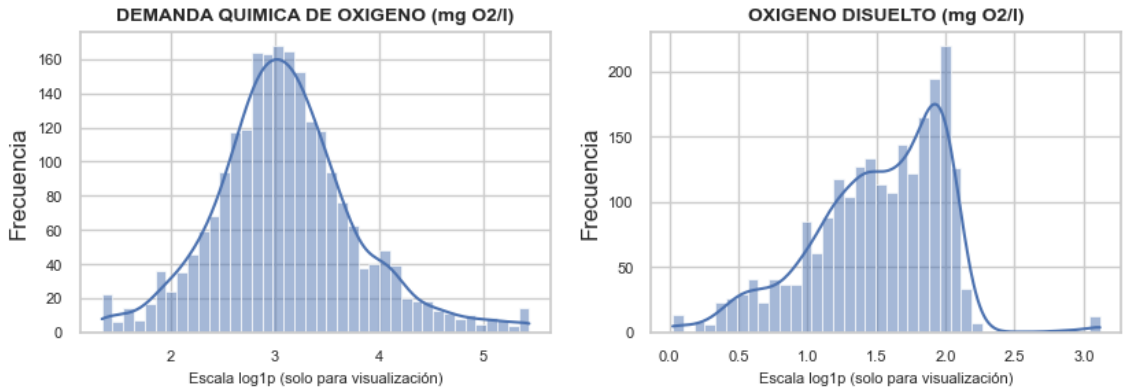

Figura C.5. Análisis de asimetría mediante histogramas de frecuencia para las variables DQO y oxígeno disuelto

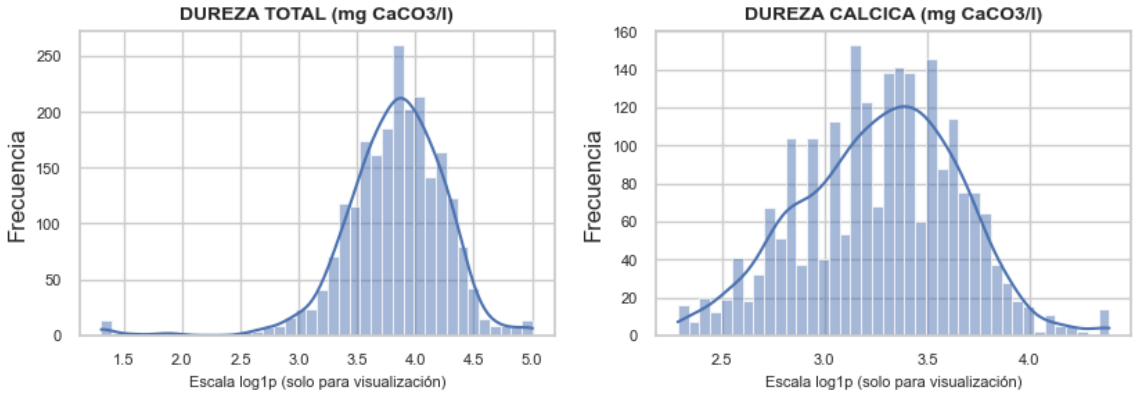

Figura C.6. Análisis de asimetría mediante histogramas de frecuencia para las variables dureza total y dureza cálcica

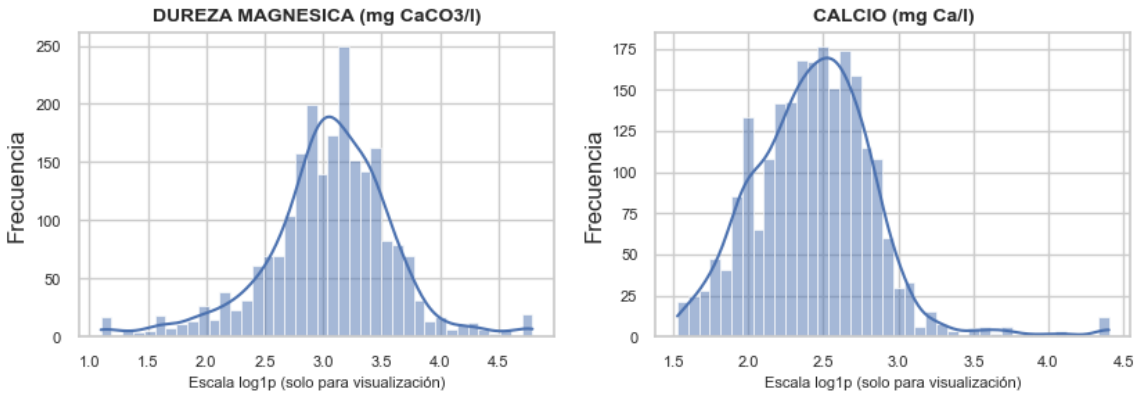

Figura C.7. Análisis de asimetría mediante histogramas de frecuencia para las variables dureza magnésica y calcio

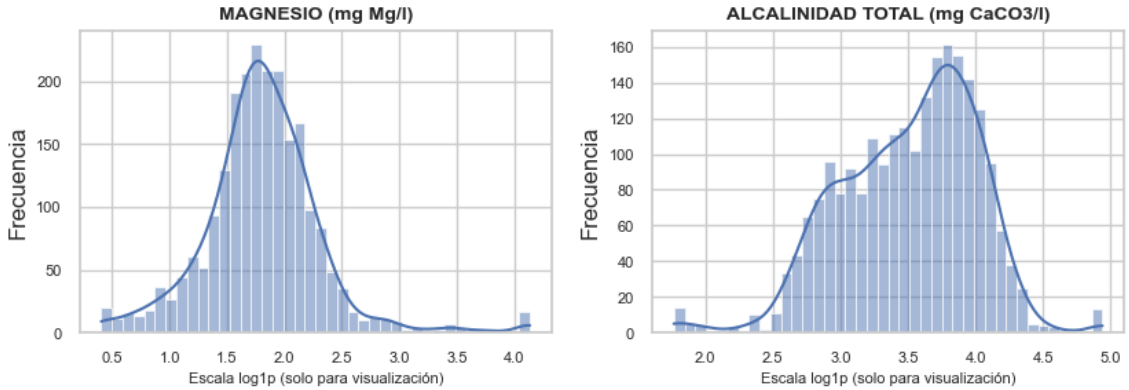

Figura C.8. Análisis de asimetría mediante histogramas de frecuencia para las variables magnesio y alcalinidad total

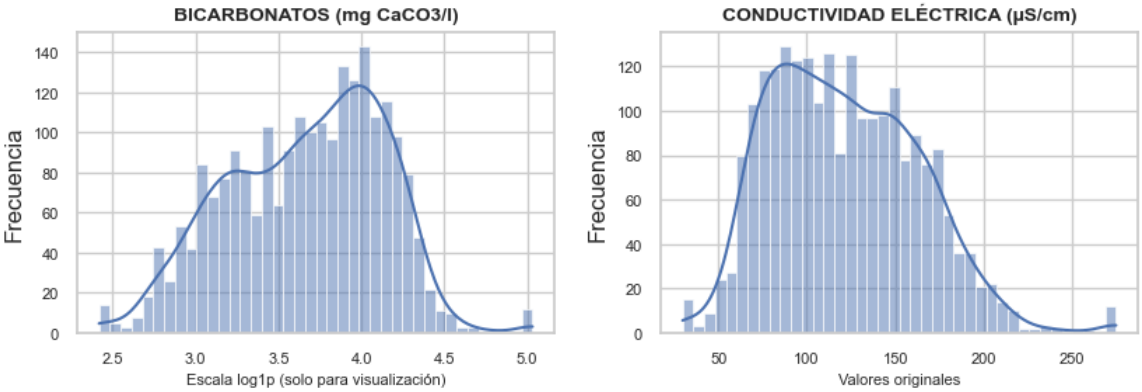

Figura C.9. Análisis de asimetría mediante histogramas de frecuencia para las variables bicarbonatos y conductividad eléctrica

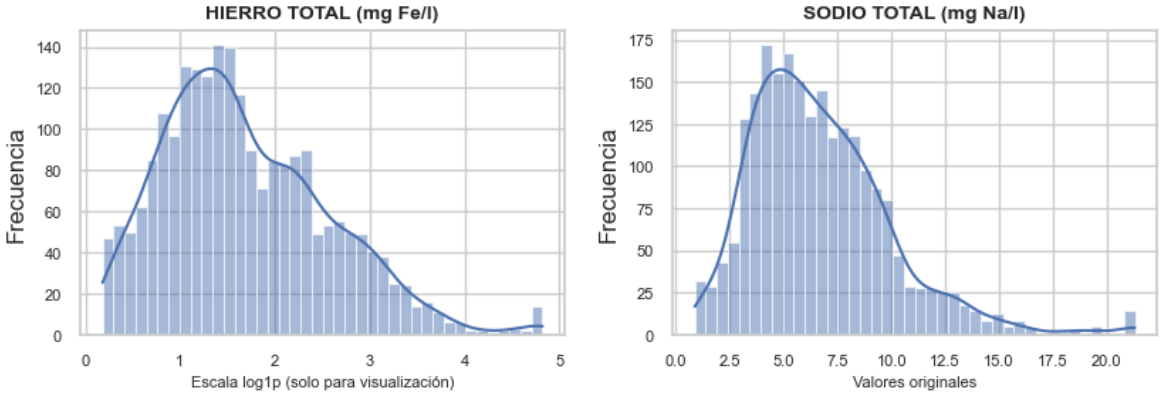

Figura C.10. Análisis de asimetría mediante histogramas de frecuencia para las variables hierro total y sodio total

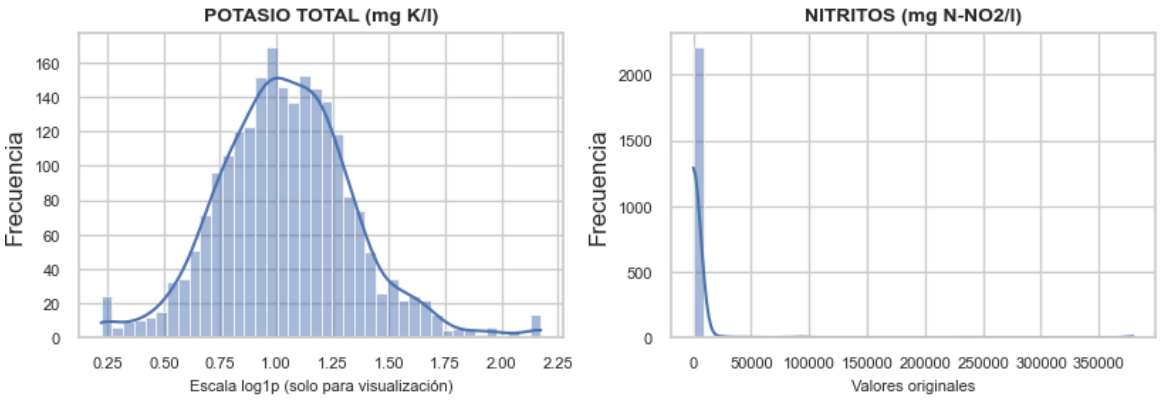

Figura C.11. Análisis de asimetría mediante histogramas de frecuencia para las variables potasio total y nitratos

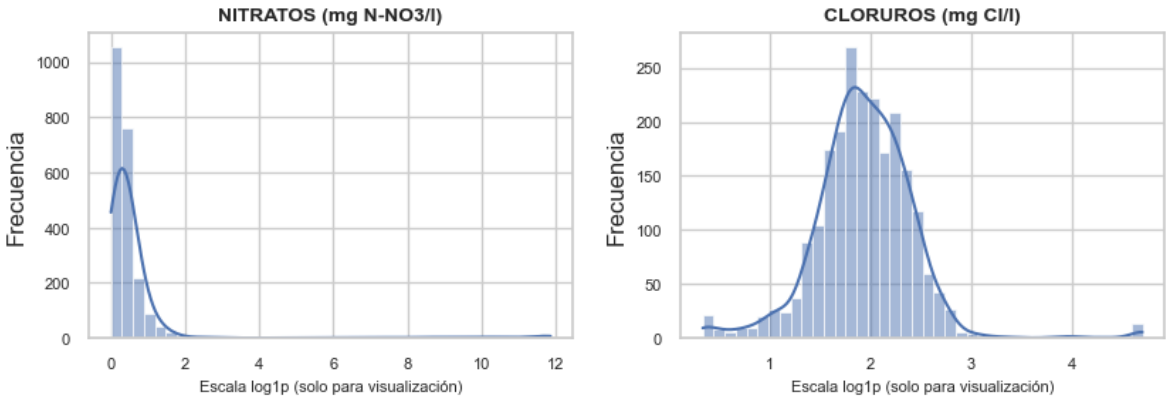

Figura C.12. Análisis de asimetría mediante histogramas de frecuencia para las variables nitratos y cloruros

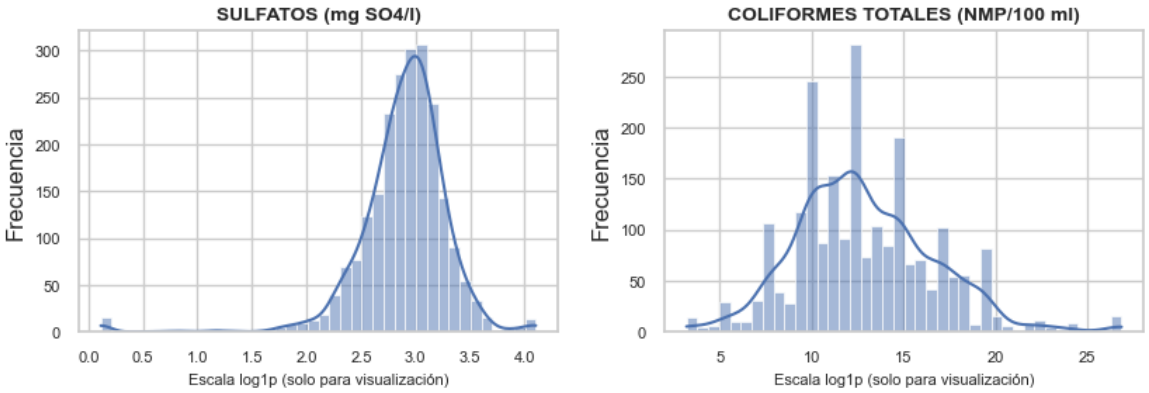

Figura C.13. Análisis de asimetría mediante histogramas de frecuencia para las variables sulfatos y coliformes totales

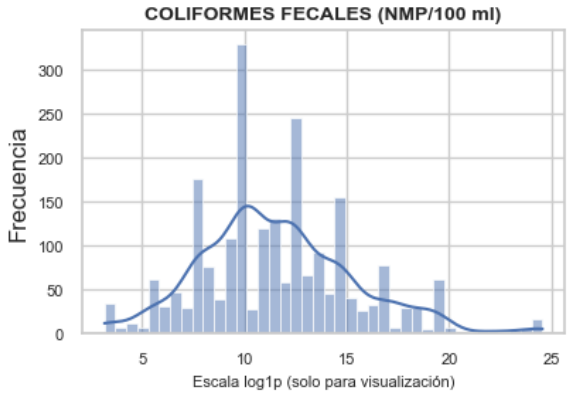

Figura C.14. Análisis de asimetría mediante histogramas de frecuencia para la variable coliformes fecales

ANEXO D. Diagramas de cajas por variable

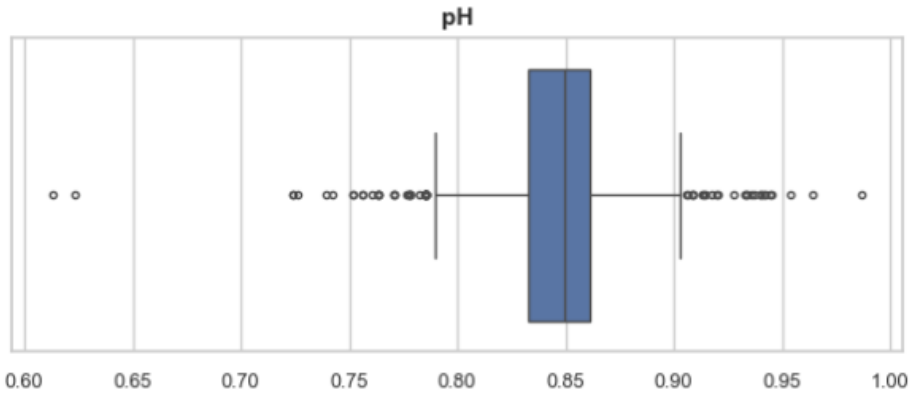

Figura D.1. Diagrama de cajas y análisis de distribución estadística para la variable pH

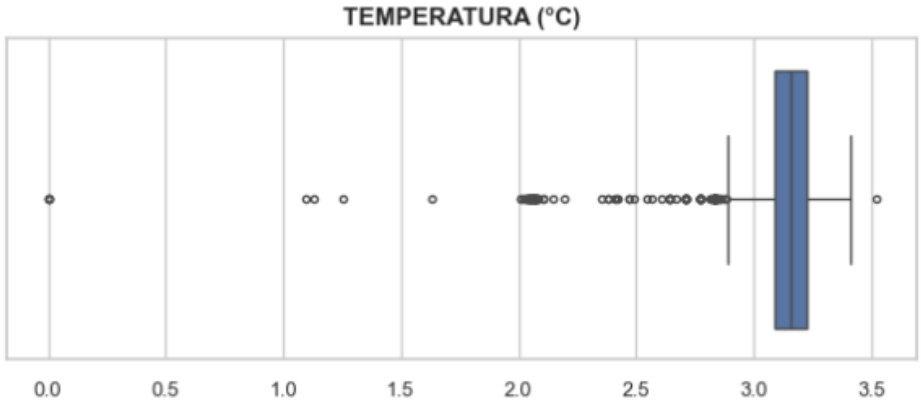

Figura D.2. Diagrama de cajas y análisis de distribución estadística para la variable temperatura

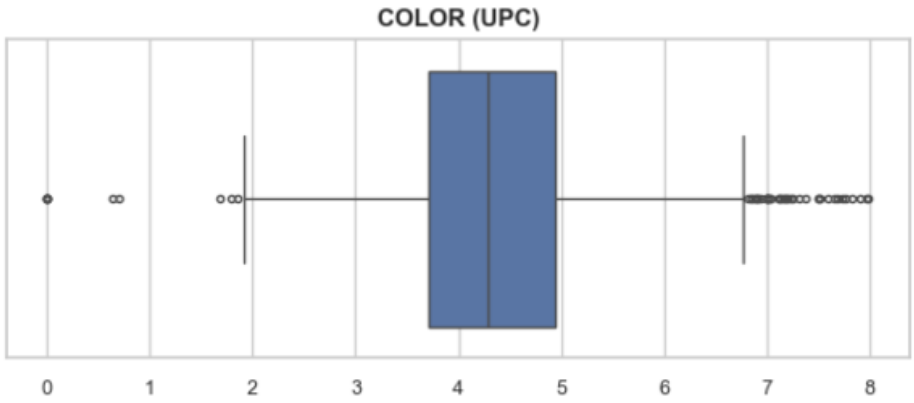

Figura D.3. Diagrama de cajas y análisis de distribución estadística para la variable color

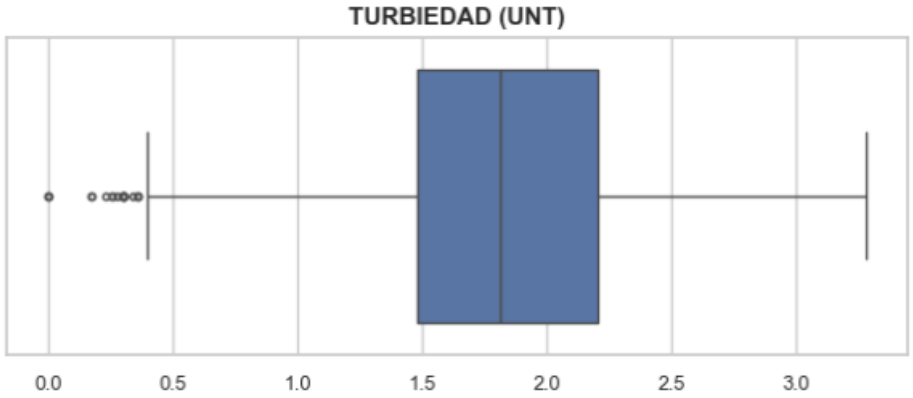

Figura D.4. Diagrama de cajas y análisis de distribución estadística para la variable turbiedad

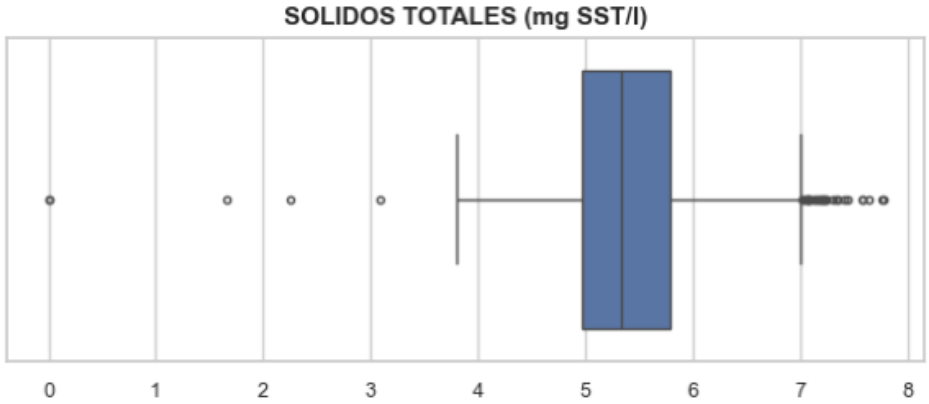

Figura D.5. Diagrama de cajas y análisis de distribución estadística para la variable sólidos totales

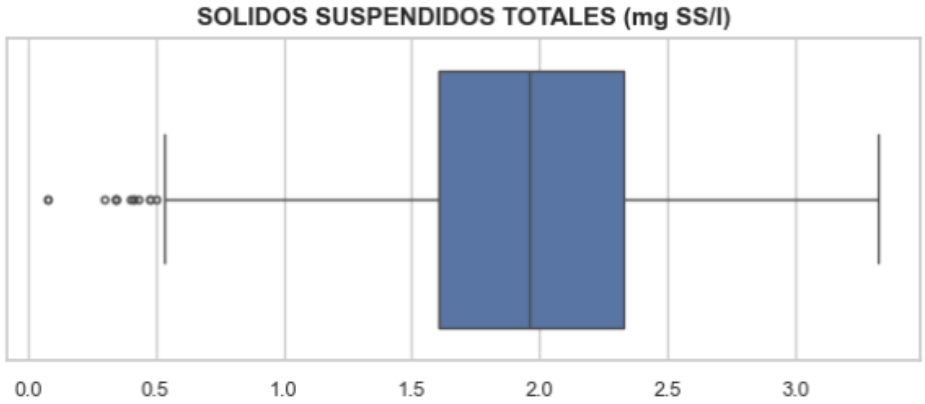

Figura D.6. Diagrama de cajas y análisis de distribución estadística para la variable sólidos suspendidos totales

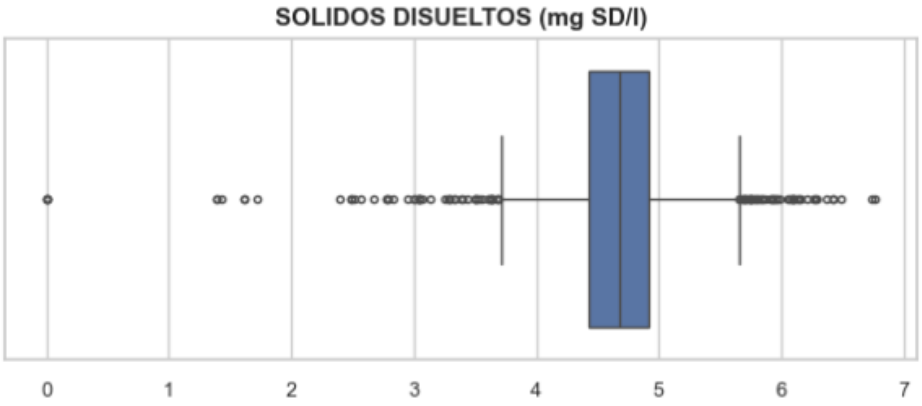

Figura D.7. Diagrama de cajas y análisis de distribución estadística para la variable sólidos disueltos

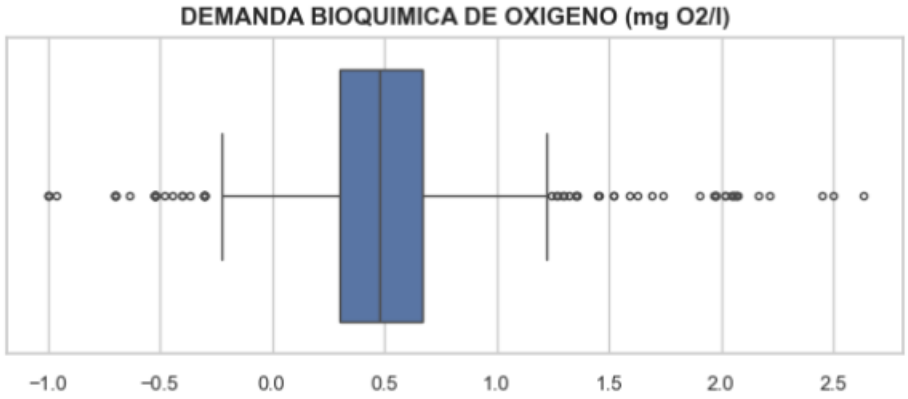

Figura D.8. Diagrama de cajas y análisis de distribución estadística para la variable DBO

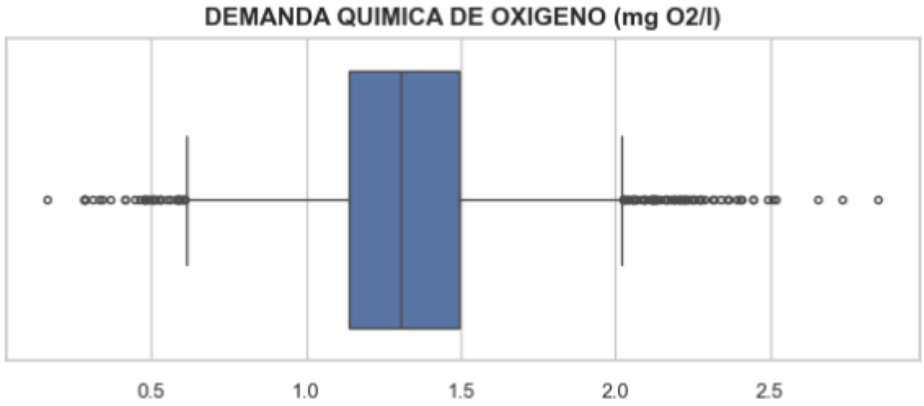

Figura D.9. Diagrama de cajas y análisis de distribución estadística para la variable DQO

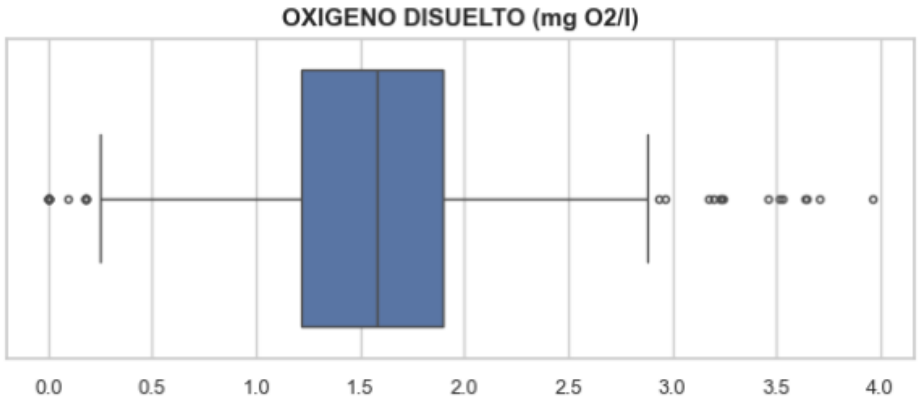

Figura D.10. Diagrama de cajas y análisis de distribución estadística para la variable oxígeno disuelto

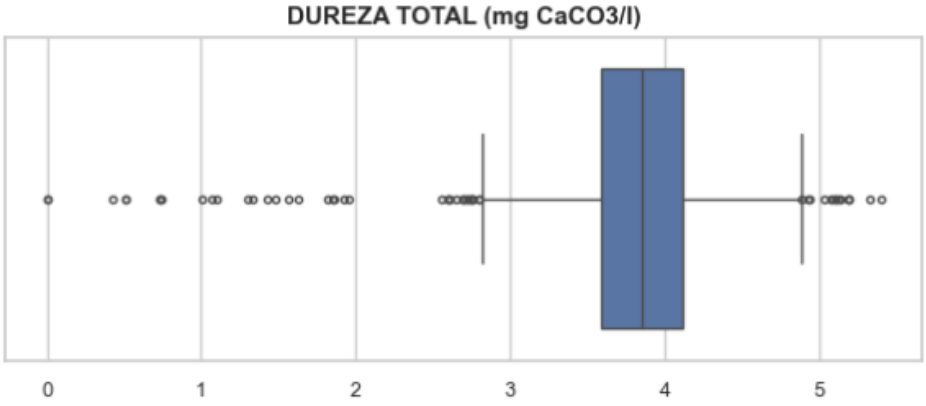

Figura D.11. Diagrama de cajas y análisis de distribución estadística para la variable dureza total

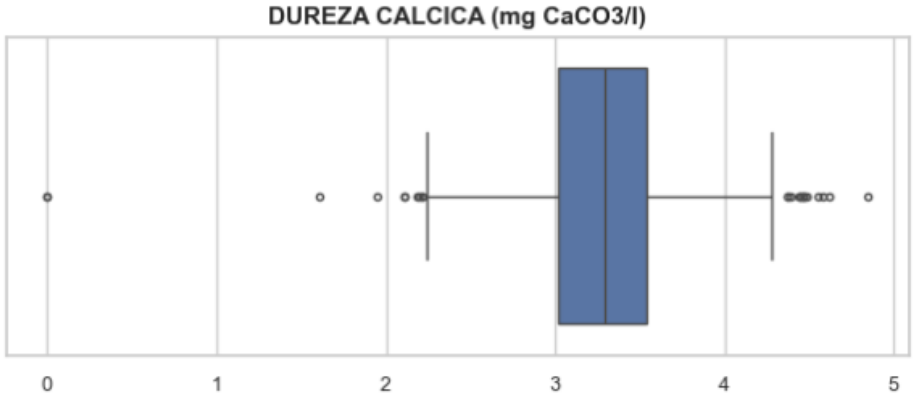

Figura D.12. Diagrama de cajas y análisis de distribución estadística para la variable dureza cálcica

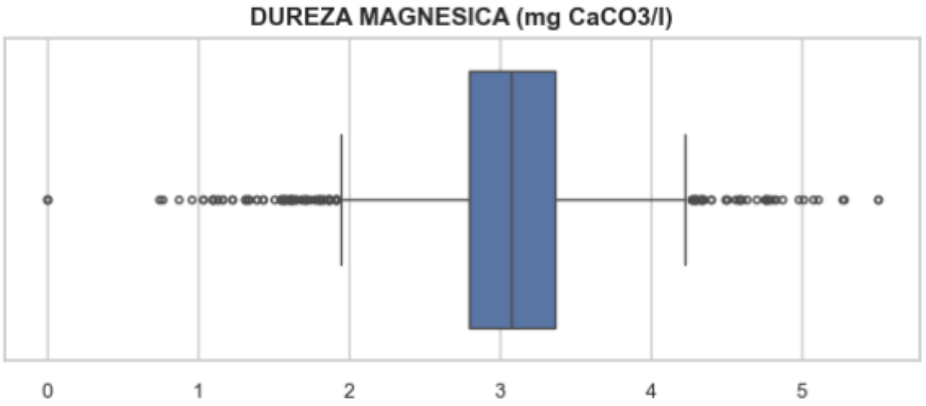

Figura D.13. Diagrama de cajas y análisis de distribución estadística para la variable dureza magnésica

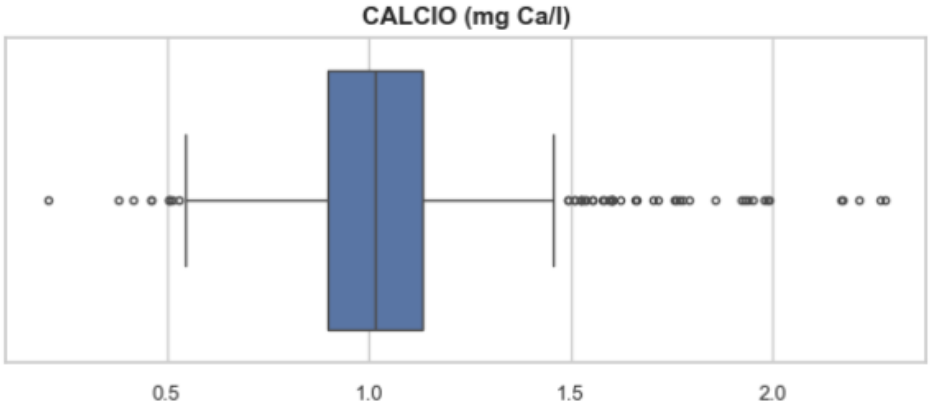

Figura D.14. Diagrama de cajas y análisis de distribución estadística para la variable calcio

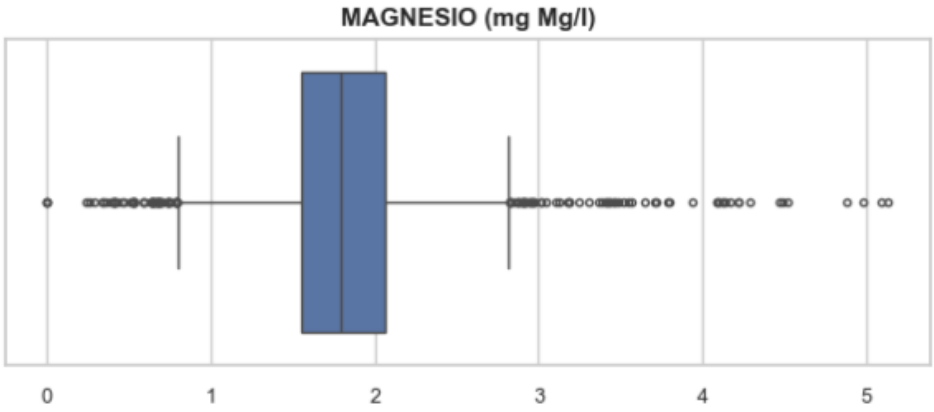

Figura D.15. Diagrama de cajas y análisis de distribución estadística para la variable magnesio

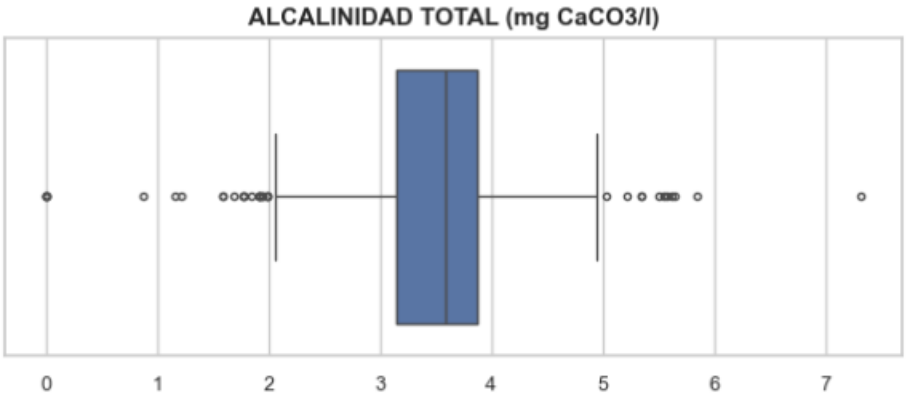

Figura D.16. Diagrama de cajas y análisis de distribución estadística para la variable alcalinidad total

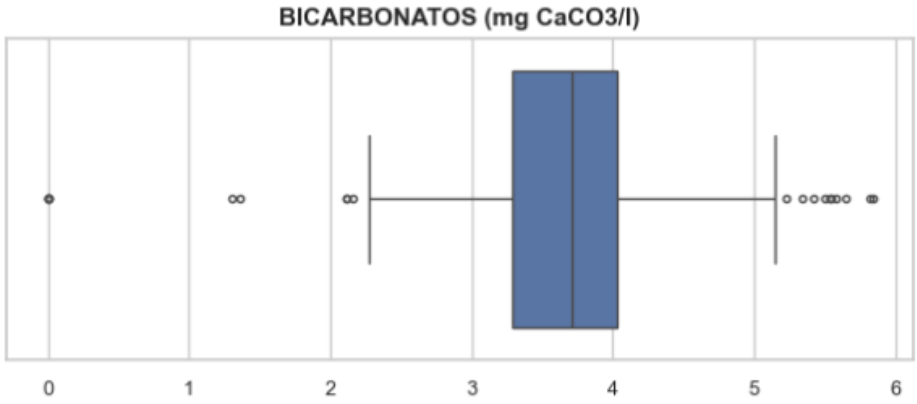

Figura D.17. Diagrama de cajas y análisis de distribución estadística para la variable bicarbonatos

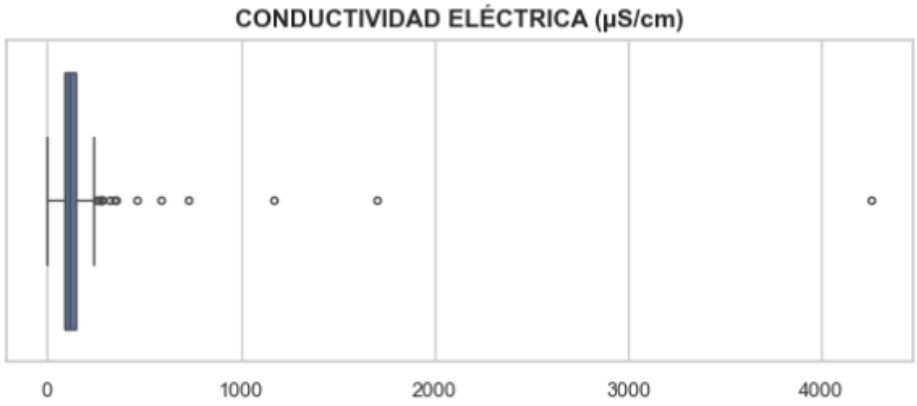

Figura D.18. Diagrama de cajas y análisis de distribución estadística para la variable conductividad eléctrica

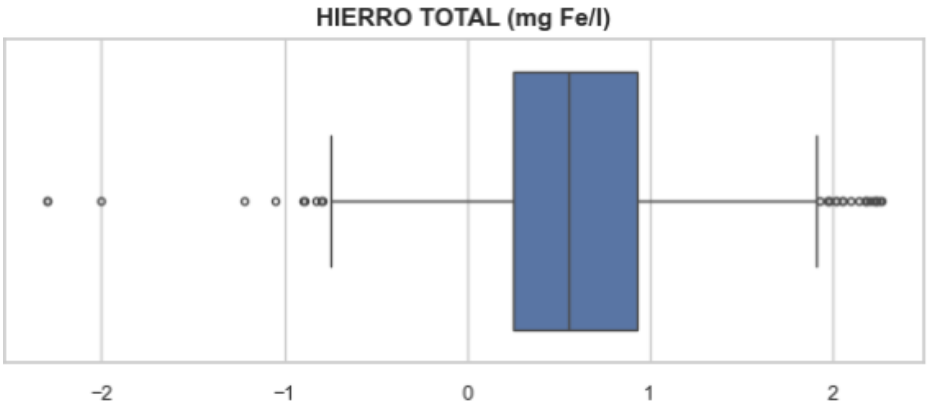

Figura D.19. Diagrama de cajas y análisis de distribución estadística para la variable hierro total

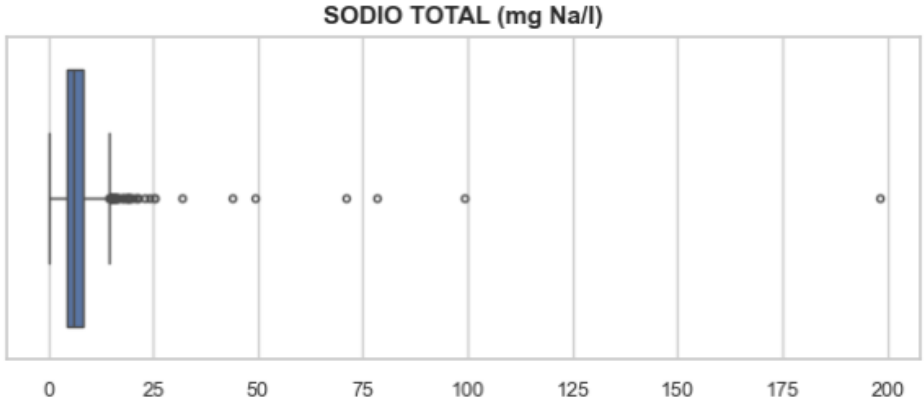

Figura D.20. Diagrama de cajas y análisis de distribución estadística para la variable sodio total

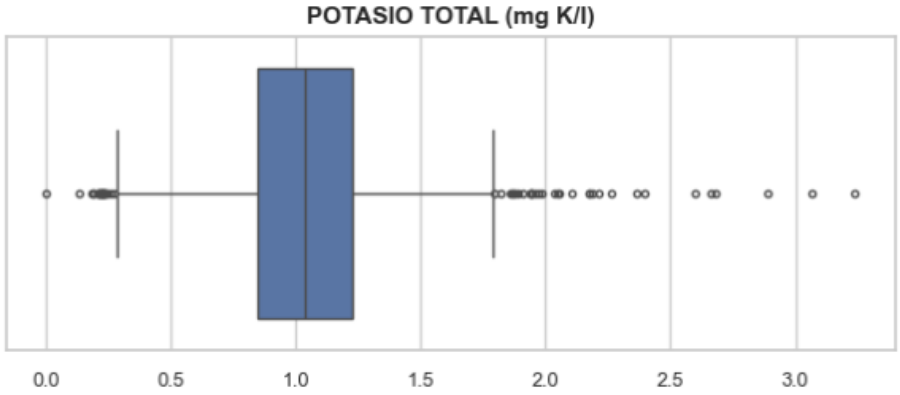

Figura D.21. Diagrama de cajas y análisis de distribución estadística para la variable potasio total

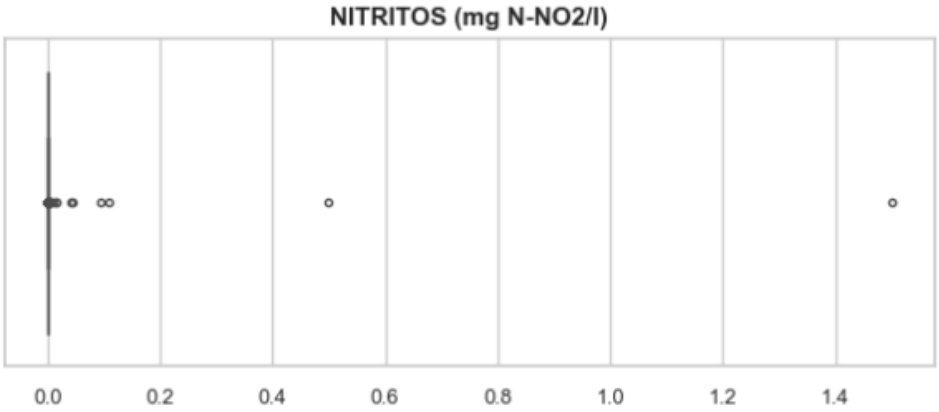

Figura D.22. Diagrama de cajas y análisis de distribución estadística para la variable nitritos

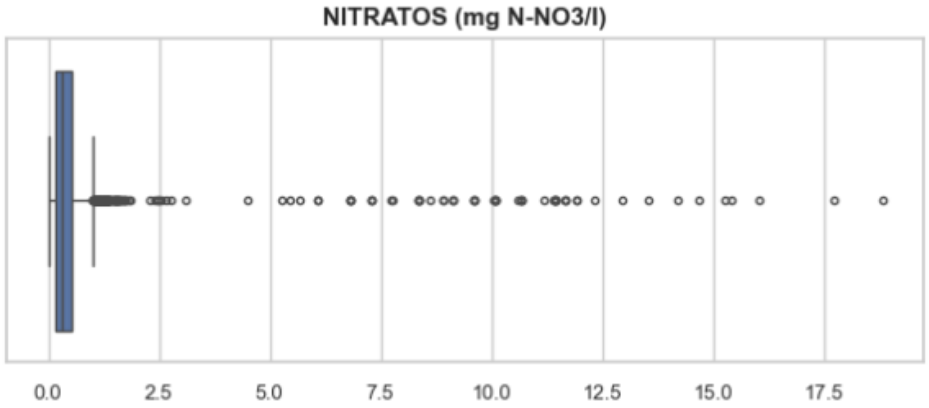

Figura D.23. Diagrama de cajas y análisis de distribución estadística para la variable nitratos

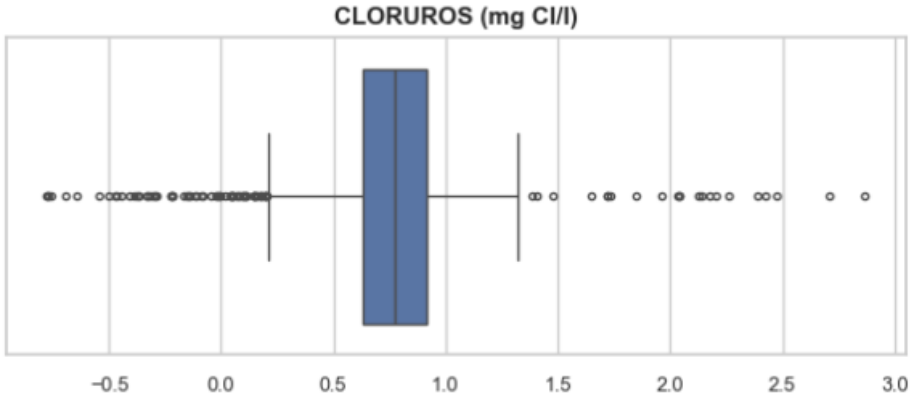

Figura D.24. Diagrama de cajas y análisis de distribución estadística para la variable cloruros

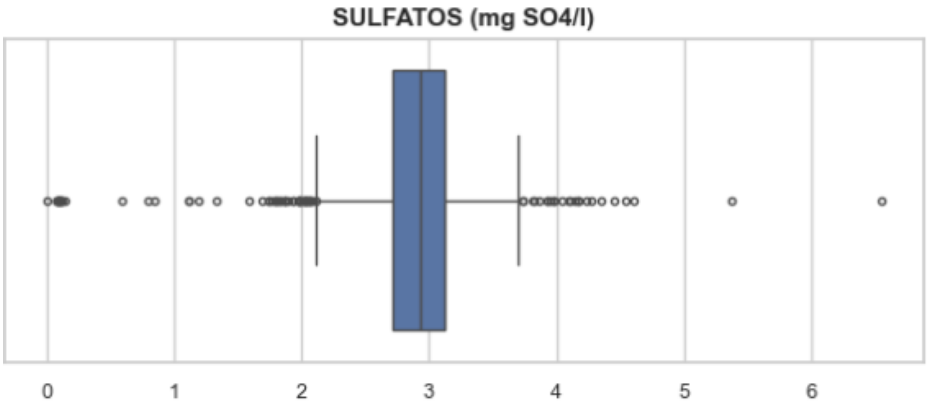

Figura D.25. Diagrama de cajas y análisis de distribución estadística para la variable sulfatos

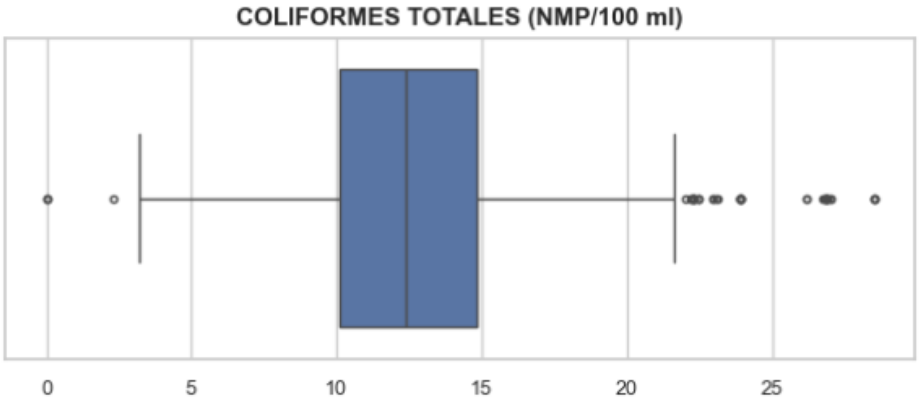

Figura D.26. Diagrama de cajas y análisis de distribución estadística para la variable coliformes totales

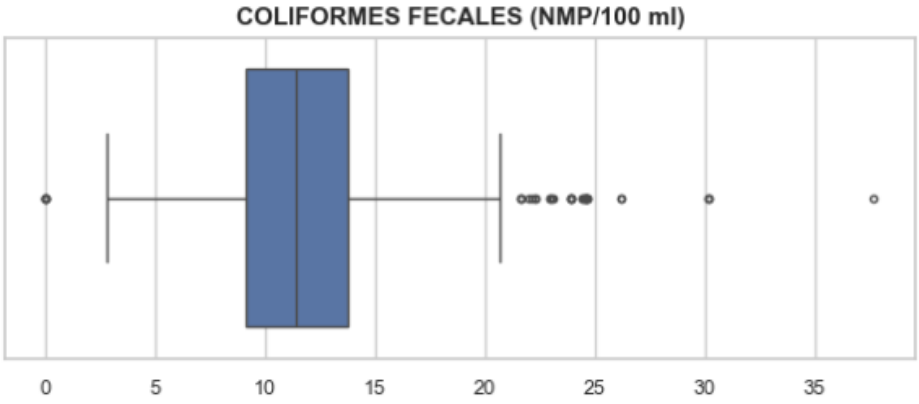

Figura D.27. Diagrama de cajas y análisis de distribución estadística para la variable coliformes fecales

ANEXO E. Comparativa de precisión entre algoritmos de aprendizaje automático para la estimación de la calidad del agua (DBO)

Tabla E.1. Resultados de modelamiento para Demanda Bioquímica de Oxigeno

ESCENARIO	MODELO	R^2	RMSE
VIF	LR	0.533	7,237
VIF	CART	-0.651	13,599
VIF	RF	0.149	9,766
VIF	SVR (RBF)	0.07	10,209
VIF	MLP	0.779	4,970
VIF	XGBoost	0.281	8,973
RFE-DBO	LR	0.538	7,198
RFE-DBO	CART	-2,312	19,261
RFE-DBO	RF	-0.448	12,738
RFE-DBO	SVR (RBF)	0.049	10,319
RFE-DBO	MLP	0.698	5,816
RFE-DBO	XGBoost	-0.554	13,194
RFE-pH	LR	0.027	10,440
RFE-pH	CART	-2,052	18,491
RFE-pH	RF	0.825	4,429
RFE-pH	SVR (RBF)	0.067	10,221
RFE-pH	MLP	0.369	8,406
RFE-pH	XGBoost	0.903	3,301
RFE-Unión	LR	0.53	7,258
RFE-Unión	CART	-0.653	13,607
RFE-Unión	RF	0.105	10,015
RFE-Unión	SVR (RBF)	0.071	10,202
RFE-Unión	MLP	0.859	3,969
RFE-Unión	XGBoost	0.151	9,754

ANEXO F. Comparativa de precisión entre algoritmos de aprendizaje automático para la estimación de la calidad del agua (pH)

Tabla F.1. Resultados de modelamiento para pH

ESCENARIO	MODELO	R^2	RMSE
VIF	LR	0.189	0.348
VIF	CART	-0.436	0.464
VIF	RF	0.341	0.314
VIF	SVR (RBF)	0.308	0.322
VIF	MLP	-0.134	0.412
VIF	XGBoost	0.414	0.296
RFE-DBO	LR	0.027	0.382
RFE-DBO	CART	-1,486	0.61
RFE-DBO	RF	0.065	0.374
RFE-DBO	SVR (RBF)	0.081	0.371
RFE-DBO	MLP	-0.117	0.409
RFE-DBO	XGBoost	-0.029	0.392
RFE-pH	LR	0.133	0.36
RFE-pH	CART	-0.635	0.495
RFE-pH	RF	0.254	0.334
RFE-pH	SVR (RBF)	0.22	0.342
RFE-pH	MLP	0.034	0.38
RFE-pH	XGBoost	0.151	0.357
RFE-Unión	LR	0.143	0.358
RFE-Unión	CART	-0.579	0.486
RFE-Unión	RF	0.308	0.322
RFE-Unión	SVR (RBF)	0.224	0.341
RFE-Unión	MLP	0.047	0.378
RFE-Unión	XGBoost	0.304	0.323

ANEXO G. Comparación de modelos y escenarios de predicción para la Demanda Bioquímica de Oxígeno (DBO) después de Grid Search y Optuna

Tabla G.1. Métricas de rendimiento y comparación de modelos predictivos por escenario para DBO

ESCENARIO	MODELO	TÉCNICA	R^2	RMSE	MAE
VIF	LR	Grid Search	0.533	7.237	3.753
VIF	CART	Grid Search	0.452	7.839	2.27
VIF	RF	Grid Search	0.719	5.615	1.981
VIF	SVR (RBF)	Grid Search	0.85	4.1	1.74
VIF	MLP	Grid Search	0.785	4.913	2.305
VIF	XGBoost	Optuna	0.8797	3.6707	1.6924
RFE-DBO	LR	Grid Search	0.538	7.198	3.774
RFE-DBO	CART	Grid Search	0.338	8.613	2.518
RFE-DBO	RF	Grid Search	0.634	6.403	2.321
RFE-DBO	SVR (RBF)	Grid Search	0.693	5.862	2.076
RFE-DBO	MLP	Grid Search	0.806	4.661	2.448
RFE-DBO	XGBoost	Optuna	0.6662	6.1153	2.3329
RFE-pH	LR	Grid Search	0.027	10.44	4.707
RFE-pH	CART	Grid Search	0.865	3.886	1.928
RFE-pH	RF	Grid Search	0.898	3.381	1.898
RFE-pH	SVR (RBF)	Grid Search	0.668	6.094	2.131
RFE-pH	MLP	Grid Search	0.805	4.669	2.233
RFE-pH	XGBoost	Optuna	0.6616	6.1573	2.8059
RFE-Unión	LR	Grid Search	0.53	7.258	3.82
RFE-Unión	CART	Grid Search	0.66	6.171	2.212
RFE-Unión	RF	Grid Search	0.837	4.267	1.928
RFE-Unión	SVR (RBF)	Grid Search	0.765	5.131	1.93
RFE-Unión	MLP	Grid Search	0.86	3.961	2.253
RFE-Unión	XGBoost	Optuna	0.7821	4.9407	1.9543

ANEXO H. Comparación de modelos y escenarios de predicción para el pH después de Grid Search y Optuna

Tabla H.1. Métricas de rendimiento y comparación de modelos predictivos por escenario para pH

ESCENARIO	MODELO	TÉCNICA	R^2	RMSE	MAE
VIF	LR	Grid Search	0.189	0.348	0.263
VIF	CART	Grid Search	0.037	0.38	0.282
VIF	RF	Grid Search	0.373	0.306	0.219
VIF	SVR (RBF)	Grid Search	0.304	0.323	0.236
VIF	MLP	Grid Search	0.007	0.386	0.29
VIF	XGBoost	Optuna	0.4115	0.2968	0.2097
RFE-DBO	LR	Grid Search	0.027	0.382	0.285
RFE-DBO	CART	Grid Search	0.079	0.371	0.279
RFE-DBO	RF	Grid Search	0.106	0.366	0.275
RFE-DBO	SVR (RBF)	Grid Search	0.081	0.371	0.276
RFE-DBO	MLP	Grid Search	0.048	0.378	0.285
RFE-DBO	XGBoost	Optuna	0.1081	0.3654	0.2742
RFE-pH	LR	Grid Search	0.133	0.36	0.274
RFE-pH	CART	Grid Search	0.126	0.362	0.273
RFE-pH	RF	Grid Search	0.28	0.328	0.245
RFE-pH	SVR (RBF)	Grid Search	0.21	0.344	0.258
RFE-pH	MLP	Grid Search	-0.133	0.412	0.315
RFE-pH	XGBoost	Optuna	0.2338	0.3387	0.2542
RFE-Unión	LR	Grid Search	0.143	0.358	0.272
RFE-Unión	CART	Grid Search	0.126	0.362	0.273
RFE-Unión	RF	Grid Search	0.336	0.315	0.231
RFE-Unión	SVR (RBF)	Grid Search	0.224	0.341	0.255
RFE-Unión	MLP	Grid Search	0.14	0.359	0.274
RFE-Unión	XGBoost	Optuna	0.3233	0.3183	0.2352

ANEXO I. Funcionamiento básico del código

1. FASE 1

Instalación y preparación del entorno

Como paso previo a la ejecución del análisis, el notebook instala de forma explícita las dependencias necesarias mediante instrucciones pip. Este bloque garantiza que el entorno de ejecución disponga de las versiones requeridas de las librerías utilizadas, independientemente de la plataforma donde se ejecute el proyecto. La instalación silenciosa evita salidas excesivas en el log y permite identificar rápidamente errores de compatibilidad antes de iniciar el flujo analítico:

```
%pip install optuna -q
%pip install xgboost -q
%pip install pandas -q
%pip install scikit-learn -q
%pip install shap -q
%pip install statsmodels -q
```

De manera complementaria, se verifica la versión activa del intérprete Python y se actualiza pip, lo que asegura que la gestión de paquetes opere sobre una base estable. Esta verificación confirma la coherencia entre el entorno virtual y las dependencias instaladas, evitando conflictos silenciosos durante la ejecución del pipeline.

Carga de dependencias y variables globales de control

Una vez asegurado el entorno, todas las librerías se importan en un único bloque consolidado, lo que favorece la legibilidad y el diagnóstico de errores. La carga organizada define el soporte técnico para la depuración de datos, el análisis exploratorio, la imputación multivariada, la verificación estadística de colinealidad y asociaciones, el entrenamiento comparativo de modelos y la interpretabilidad de resultados aplicados a la calidad del agua del río Cauca.

Como base operativa se cargan librerías estándar del lenguaje para control del sistema, manejo de tiempos de ejecución y utilidades matemáticas:

```
import os, json, joblib, sys, platform, time, warnings
from datetime import datetime
from math import sqrt
```

El entorno de trabajo en Jupyter se habilita con componentes de visualización e interacción, que permiten organizar resultados y navegar el análisis por variable sin modificar el dataset fuente:

```
from IPython.display import display
import ipywidgets as widgets
```

```
from ipywidgets import interact, Dropdown, HTML, Text, VBox, HBox, Layout
```

En el componente analítico se integran las librerías centrales para manipulación de datos y visualización. pandas y numpy sostienen la construcción del DataFrame maestro, transformaciones y agregaciones; matplotlib y seaborn soportan la visualización exploratoria y diagnóstica necesaria para reconocer patrones, dispersión y comportamiento distribucional previo al modelado:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Para la fase de modelado y validación se incorporan herramientas estadísticas y de aprendizaje supervisado. Respecto a la versión previa del pipeline, se añaden RepeatedKFold para una validación cruzada más robusta por repetición, y clone para instanciar modelos frescos durante la búsqueda de hiperparámetros sin arrastrar estados previos de entrenamiento:

```
from scipy.stats import spearmanr
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.dummy import DummyRegressor
from sklearn.linear_model import LinearRegression, Ridge
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR
from sklearn.neural_network import MLPRegressor
import xgboost as xgb
from sklearn.model_selection import (
    train_test_split, GridSearchCV,
    cross_val_score, KFold, RepeatedKFold, learning_curve
)
from sklearn.base import clone
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
```

El módulo de preprocesamiento, selección de variables e interpretabilidad completa el stack técnico. Se incorpora explícitamente permutation_importance desde sklearn.inspection, y los samplers y pruners de Optuna se importan de forma directa para un control preciso de la estrategia de búsqueda bayesiana:

```
from sklearn.impute import KNNImputer
```

```
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from sklearn.pipeline import Pipeline
from sklearn.feature_selection import RFE
from sklearn.inspection import permutation_importance
import optuna
from optuna.samplers import TPESampler
from optuna.pruners import MedianPruner
import shap
import joblib
```

Como cierre del bloque de importaciones se definen dos variables globales de control que estructuran todo el flujo posterior. RANDOM_STATE centraliza la semilla aleatoria utilizada en todos los procesos que requieren reproducibilidad, mientras que artefactos actúa como repositorio central de objetos intermedios y finales del pipeline, evitando redundancias de cómputo entre secciones:

```
RANDOM_STATE = 42
artefactos   = {}
```

La combinación de semilla fija y repositorio centralizado garantiza trazabilidad completa: cualquier resultado puede reproducirse o recuperarse sin necesidad de reejecutar etapas costosas como el entrenamiento o la optimización bayesiana.

Carga del diccionario de variables

El diccionario de variables se carga desde un módulo local, el cual centraliza la definición conceptual, técnica y normativa de cada parámetro analizado. La recarga explícita mediante importlib garantiza que el módulo no opere sobre versiones en caché y que sus objetos estén sincronizados con la ejecución actual:

```
import importlib, diccionario
importlib.reload(diccionario)
print('¿Está diccionario_variables?:', hasattr(diccionario, 'diccionario_variables'))
```

Esta verificación actúa como control de integridad previo: confirma la disponibilidad de las estructuras necesarias antes de vincularlas funcionalmente al flujo principal. Una vez validado, el diccionario habilita un visor interactivo basado en ipywidgets que permite filtrar variables por nombre y visualizar de forma organizada su definición, relación con la contaminación, referencia y rangos asociados, sin alterar los datos originales:

```
dropdown = Dropdown(options=opciones, description='Variable:', layout=Layout(width='60%'))
```

El visor responde dinámicamente a la selección del usuario y presenta la información del diccionario en formato tabular HTML, lo que facilita la inspección técnica y metodológica sin interrumpir el flujo de análisis.

Carga del conjunto de datos y preparación para el análisis exploratorio

En esta etapa se realiza la carga inicial del conjunto de datos de calidad del agua del río Cauca y su adecuación para el análisis exploratorio. El objetivo principal es construir una base numérica coherente, depurada y comparable, que permita evaluar la estructura, completitud y comportamiento estadístico de las variables antes de avanzar hacia procesos de imputación y modelado predictivo.

La lectura del archivo se realiza directamente desde formato CSV con separador punto y coma, conservando la totalidad de columnas disponibles para una inspección preliminar:

```
df = pd.read_csv('Calidad_del_agua_del_Rio_Cauca.csv', sep=';')
```

Una vez cargado el dataset, se separan las variables descriptivas —fecha de muestreo y nombre de la estación— para concentrar el análisis en las variables cuantitativas de calidad del agua:

```
df_sin_columnas = df.drop(columns=['FECHA DE MUESTREO', 'ESTACIONES'])
```

Normalización y limpieza de variables numéricas

Con el fin de garantizar consistencia en los cálculos posteriores, se crea una copia de trabajo del DataFrame y se procede a la conversión sistemática de todas las columnas a formato numérico. Este proceso incluye la normalización de separadores decimales y la coerción controlada de valores no convertibles a nulos, evitando errores silenciosos durante el análisis:

```
df_eda[columna] = pd.to_numeric(
    df_eda[columna].astype(str).str.replace(',', '.'),
    errors='coerce'
)
```

Posteriormente se eliminan columnas completamente vacías, asegurando que el conjunto de variables conserve únicamente dimensiones con contenido informativo real.

Análisis descriptivo ampliado (EDA)

Sobre la base depurada se calcula un conjunto extendido de estadísticas descriptivas que trasciende las métricas básicas para incorporar indicadores de dispersión, forma y completitud. A diferencia de la aproximación anterior, las métricas se calculan variable por variable de forma explícita, lo que otorga mayor control sobre cada indicador y facilita su inclusión directa en tablas de resultados o anexos. Entre los indicadores calculados se incluyen el rango intercuartílico (IQR), la desviación absoluta media (MAD), el coeficiente de variación (CV), los coeficientes de asimetría y curtosis, el error estándar de la asimetría y el porcentaje de datos válidos:

```
estadisticas_completas['Pct.Valid'] = (estadisticas_completas['count'] / n) * 100
```

Evaluación de completitud y filtrado de variables

Con el propósito de establecer un criterio objetivo de calidad de datos, se define un umbral mínimo de completitud del 80 %. A partir de este criterio se identifican las variables con porcentaje de datos válidos inferior al umbral, las cuales requieren atención especial en etapas posteriores de imputación o, en su defecto, exclusión del análisis predictivo:

```
vars_baja_completitud = estadisticas_completas[estadisticas_completas['Pct.Valid'] < umbral]
```

Este análisis permite diferenciar variables con información suficiente de aquellas con alta tasa de ausencia. Las variables que cumplen el umbral conforman el subconjunto analítico robusto utilizado en todas las etapas posteriores:

```
df_filtrado = df_eda[columnas_mantener].copy()
```

Definición de variables objetivo y construcción del conjunto analítico

Con base en criterios técnicos y ambientales se definen como variables objetivo la Demanda Bioquímica de Oxígeno (DBO) y el pH. Ambas se declaran explícitamente como constantes con nombre canónico, lo que garantiza coherencia y evita errores por variaciones tipográficas a lo largo del pipeline:

```
OBJ_DBO = 'DEMANDA BIOQUIMICA DE OXIGENO (mg O2/l)'
OBJ_PH  = 'pH'
```

Para cada variable objetivo se implementa la función get_X_y_para, que construye de forma sistemática el conjunto de predictores (X) y la variable respuesta (y). El procedimiento incorpora controles clave: verificación de existencia de la variable objetivo, eliminación de observaciones sin valor objetivo, exclusión del otro objetivo del espacio de predictores, selección de columnas numéricas con variabilidad efectiva y alineación estricta de índices:

```
def get_X_y_para(objetivo):
    """Retorna X (predictores) e y (objetivo), alineados, sin NaN en ninguno."""
```

La validación dimensional confirma que ambos objetivos disponen de un número suficiente de observaciones y predictores para proceder con imputación, evaluación de colinealidad y entrenamiento supervisado:

```
DBO → X: (2092, 26), y: (2092,)
pH  → X: (2208, 26), y: (2208,)
```

Estas diferencias en el tamaño muestral reflejan la disponibilidad diferencial de datos por variable y se tienen en cuenta en las etapas posteriores de imputación y validación.

2. FASE 2

Esta fase profundiza el análisis exploratorio incorporando un tratamiento sistemático de valores faltantes y una evaluación detallada de la distribución de las variables, con el objetivo de preparar una base de datos estadísticamente consistente para el modelado predictivo. La estrategia adoptada combina diagnóstico previo de ausencia de datos, imputación multivariada, verificación posterior del comportamiento distribucional y depuración iterativa de multicolinealidad.

Diagnóstico de valores nulos previo a la imputación

Antes de aplicar cualquier técnica de imputación, se realiza un análisis visual de los valores faltantes por variable. Se construye un mapa de calor (Figura 1) que permite inspeccionar la distribución espacial de los valores nulos a lo largo de todos los registros, facilitando la detección de patrones de ausencia estructural o concentraciones temporales:

```
sns.heatmap(df_filtrado.isnull(), cbar=False, cmap='viridis', yticklabels=False)
```

Este diagnóstico asegura que la imputación posterior no se aplique de manera ciega, sino informada por la estructura real del conjunto de datos.

Imputación mediante K-Nearest Neighbors (KNN)

La imputación de valores faltantes se realiza utilizando el método KNN, que estima valores ausentes a partir de observaciones similares en el espacio multivariado. Dado que este método es sensible a la escala de las variables, se implementa un flujo en tres etapas: escalamiento, imputación y retorno a la escala original. Primero, las variables se estandarizan con StandardScaler para garantizar comparabilidad entre magnitudes heterogéneas. Posteriormente, se aplica KNNImputer ponderado por distancia, lo que otorga mayor peso a observaciones más cercanas. Finalmente, los datos imputados se transforman de vuelta a su escala original, preservando la interpretación física de cada variable:

```
scaler_imp = StandardScaler()
imputer    = KNNImputer(n_neighbors=5, weights='distance')
```

Como verificación visual, se reconstruye el mapa de calor de valores nulos tras la imputación (Figura 2), confirmando la eliminación efectiva de valores faltantes sin introducir discontinuidades evidentes en la estructura del dataset:

```
print(f'Nulos restantes: {df_imputed.isnull().sum().sum()}')  # debe ser 0
```

Análisis distribucional y detección de valores atípicos

Con la base imputada, se analiza el comportamiento distribucional de todas las variables mediante dos visualizaciones complementarias. En primer lugar, se generan boxplots para cada variable (Figura 3), aplicando transformaciones logarítmicas seguras únicamente cuando la naturaleza de los datos lo permite, lo que evita transformaciones forzadas en presencia de valores negativos:

```
def log_transform_safe(series):
    if (s > 0).all():  return np.log10(s)
    elif (s >= 0).all(): return np.log1p(s)
    else: return s
```

Los boxplots mantienen visibles los valores atípicos como parte del comportamiento real del sistema, sin eliminarlos arbitrariamente. En segundo lugar, se generan histogramas de todas las variables (Figura 5) utilizando una función de transformación robusta que combina logaritmo y estandarización basada en mediana e IQR, aplicando winsorización suave en los percentiles extremos para limitar la influencia de valores extremos durante la visualización, sin modificar los datos utilizados para el modelado:

```
def transform_robusta_para_plot(s, p_low=0.005, p_high=0.995):
    if (s >= 0).all():
        s_t = np.log10(s + eps)
    else:
        s_t = (s - med) / (iqr + eps)
    return s_t.clip(lo, hi)
```

Análisis focalizado de variables problemáticas

Para variables con alta asimetría o valores extremos particularmente pronunciados —como nitritos y conductividad eléctrica— se realiza un análisis diagnóstico específico que incluye el cálculo de estadísticos robustos (cuartiles, IQR, asimetría, curtosis) y la cuantificación explícita del número y porcentaje de outliers según el criterio intercuartílico:

```
outliers = ((s < Q1 - 1.5*IQR) | (s > Q3 + 1.5*IQR)).sum()
```

Los resultados se visualizan mediante boxplots con transformación robusta (Figura 4), lo que permite comparar la distribución de estas variables de forma más informativa y homogénea, reforzando el control de calidad previo al modelado.

Análisis de correlaciones entre predictores (excluyendo variables objetivo)

Una vez consolidada la base imputada, se analiza la relación interna entre variables predictoras con el fin de identificar dependencias fuertes que pudieran introducir redundancia estructural en los modelos posteriores. Para este análisis se excluyen explícitamente las variables objetivo, garantizando que la evaluación de colinealidad se realice únicamente entre predictores. El análisis se desarrolla de manera paralela utilizando los coeficientes de Pearson y Spearman (Figuras 10 y 11), permitiendo contrastar relaciones lineales y monotónicas respectivamente:

```
corr_p = df_pred.corr(method='pearson')
corr_s = df_pred.corr(method='spearman')
```

Con el propósito de focalizar el análisis en relaciones potencialmente problemáticas, se extraen pares de variables con correlaciones absolutas superiores a un umbral alto ($|\rho| \geq 0.80$), eliminando duplicidades mediante el uso del triángulo superior de la matriz de correlación. De forma complementaria, se calcula la correlación Spearman de cada predictor con cada variable objetivo, lo que sirve como insumo para la etapa de filtrado y selección de variables:

```
for col in cols_pred_corr:
    r, _ = spearmanr(tmp[col], tmp[obj])
    corrs[col] = round(r, 4)
```

Filtrado por colinealidad basado en Spearman

A partir de las correlaciones monotónicas se implementa un filtro sistemático de colinealidad utilizando un umbral conservador ($|\rho| \geq 0.85$). La decisión de trabajar con Spearman responde a su mayor robustez frente a asimetrías y valores extremos previamente identificados en varias variables:

```
UMBRAL_SPEARMAN = 0.85
```

La eliminación de variables no se realiza de forma automática ni arbitraria. Se aplica una heurística jerárquica con los siguientes criterios: preservación de variables objetivo, prioridad a la variable con mayor asociación absoluta con la(s) variable(s) objetivo cuando estas están definidas, y en caso de empate, conservación de la variable con mayor correlación media absoluta con el resto del conjunto. Este enfoque reduce la dimensionalidad manteniendo coherencia conceptual y relevancia predictiva.

Evaluación iterativa de multicolinealidad (VIF)

Sobre la base filtrada por correlación, se aplica un procedimiento iterativo de cálculo del Variance Inflation Factor (VIF) con el objetivo de eliminar colinealidad residual no capturada por el análisis bivariado. Se implementa una función auxiliar calcular_vif que recalcula el VIF completo en cada iteración, permitiendo registrar el historial del proceso de depuración:

```
def calcular_vif(df_vars):
    """Calcula VIF para cada columna del DataFrame."""
UMBRAL_VIF = 10.0
while True:
    vifs = calcular_vif(df_vif_work)
    if vifs.iloc[0] <= UMBRAL_VIF: break
    df_vif_work = df_vif_work.drop(columns=[vifs.index[0]])
```

En cada iteración se elimina la variable con mayor VIF, siempre que no corresponda a una variable objetivo. El registro histórico (historial_vif) documenta cada variable eliminada y su VIF en el

momento de extracción, garantizando transparencia y reproducibilidad del proceso. El resultado es una base numérica depurada con predictores libres de multicolinealidad estructural, lista para la selección supervisada de variables.

Selección de variables mediante RFE

Una vez controlada la colinealidad estructural, se aplica Recursive Feature Elimination (RFE) como mecanismo de selección supervisada, incorporando explícitamente la relación entre predictores y variables objetivo. El procedimiento se implementa mediante la función rfe_ensemble_para, diseñada para ser robusta y replicable.

Como primer paso, se aplica una preselección top-k basada en correlación Spearman con el objetivo, reduciendo el espacio de búsqueda a las variables más informativas antes del proceso recursivo. El número final de variables a seleccionar se determina dinámicamente como el máximo entre 5 y la raíz cuadrada de la dimensión del espacio de entrada:

```
X_r = _preselect_topk_spearman(X_r, y_r, k=top_n_spearman)
k_final = max(5, int(sqrt(p)))
```

La estrategia multi-modelo ejecuta el RFE sobre tres estimadores lineales con propiedades complementarias: Regresión Lineal (sensibilidad directa a relaciones lineales), Ridge (penalización L2, estabilidad frente a multicolinealidad residual) y SVR lineal (margen máximo, robustez frente a outliers). Cada estimador opera sobre una versión escalada del dataset, y las variables se escalan manualmente antes de pasarlas al RFE para evitar el conflicto de acceso al atributo coef_ dentro de un Pipeline:

```
estimadores = [
   ('RFE-Linear', StandardScaler(), LinearRegression()),
   ('RFE-Ridge',  StandardScaler(), Ridge(alpha=1.0)),
   ('RFE-SVRlin', MinMaxScaler(),   SVR(kernel='linear', C=2.0))
]
rfe = RFE(estimator=modelo_rfe, n_features_to_select=k_final, step=2)
```

La selección final se define por mayoría simple entre los tres modelos, evitando dependencias espurias de un único criterio inductivo. Para cada objetivo se obtiene una tabla de votos por variable, una lista de variables seleccionadas y el número final de predictores retenidos. Dado que el modelado posterior considera ambos objetivos, se construye una selección RFE-unión como intersección entre las variables seleccionadas para DBO y pH, con un fallback a la unión cuando la intersección resulta demasiado restrictiva:

```
rfe_comun = intersect(DBO, pH) if len(intersect) >= 5 else union(DBO, pH)
```

El resultado es el diccionario ESCENARIOS_VARS con cuatro conjuntos de variables: vif (todos los predictores tras VIF), rfe_dbo (selección específica para DBO), rfe_ph (selección específica para

pH) y rfe_comun (selección consolidada para ambos objetivos).

```
ESCENARIOS_VARS = {
  'vif':     [variables tras VIF],
  'rfe_dbo':  rfe_resultados[OBJ_DBO]['seleccion_mayoria'],
  'rfe_ph':   rfe_resultados[OBJ_PH]['seleccion_mayoria'],
  'rfe_comun': rfe_comun
}
```

3. FASE 3

Una vez estabilizado el espacio de variables mediante VIF y RFE, se procede al entrenamiento comparativo de modelos de regresión bajo cuatro escenarios estrictamente definidos, garantizando trazabilidad total entre selección de variables, objetivo y desempeño predictivo. El objetivo de esta fase no es solo optimizar hiperparámetros, sino evaluar la capacidad explicativa y predictiva real de distintos enfoques de modelado sobre conjuntos de variables coherentes y conservar todos los artefactos generados para su uso en etapas de interpretabilidad.

Configuración del entrenamiento base

La configuración global del entrenamiento centraliza los parámetros comunes a todos los experimentos. Se utiliza RepeatedKFold con 5 particiones y 3 repeticiones como esquema de validación cruzada, lo que proporciona estimaciones más estables del desempeño al promediar múltiples realizaciones de la partición aleatoria:

```
CV = RepeatedKFold(n_splits=5, n_repeats=3, random_state=42)
TEST_SIZE = 0.20
```

Los seis modelos se instancian mediante la función fábrica get_modelos_base, que retorna copias frescas en cada llamada evitando contaminación de estado entre iteraciones. La función auxiliar _metricas centraliza el cálculo de R^2, RMSE y MAE a partir de los vectores de valores reales y predichos:

```
def get_modelos_base():
  return {
    'LR':       Pipeline([...LinearRegression()]),
    'CART':     DecisionTreeRegressor(random_state=42),
    'RF':       RandomForestRegressor(random_state=42, n_jobs=-1),
    'SVR (RBF)': Pipeline([...SVR(kernel='rbf')]),
    'MLP':      Pipeline([...MLPRegressor(random_state=42, max_iter=600)]),
    'XGBoost':   xgb.XGBRegressor(random_state=42, tree_method='hist', ...)
```

```
}
```

Entrenamiento base y registro de resultados

El entrenamiento base evalúa 48 combinaciones (6 modelos × 4 escenarios × 2 objetivos) bajo el mismo esquema de partición 80/20 con semilla fija. Para garantizar comparabilidad directa, cada combinación objetivo × escenario genera una partición independiente que se almacena en el diccionario splits_base, permitiendo su reutilización exacta en las fases de optimización y evaluación sin riesgo de data leakage:

```
X_tr, X_te, y_tr, y_te = train_test_split(
    X_e, y_e, test_size=TEST_SIZE, random_state=42, shuffle=True
)
splits_base[(obj, esc)] = (X_tr, X_te, y_tr, y_te)
```

Para cada combinación se registran R^2, RMSE, MAE, número de variables y tiempo de ejecución. Todos los resultados se consolidan en la tabla resumen artefactos['resumen_global_4esc'], que permite análisis transversal inmediato. Las Figuras 13–16 presentan el top 5 de modelos por R^2 y el comparativo de RMSE/MAE para DBO y pH respectivamente.

Grid Search: optimización de hiperparámetros (LR, CART, RF, SVR, MLP)

El proceso de optimización por Grid Search se aplica a cinco de los seis modelos, mientras XGBoost se reserva para Optuna. Los espacios de búsqueda se definen acorde a la naturaleza de cada estimador, priorizando hiperparámetros que controlan complejidad, regularización y capacidad de generalización. Para Random Forest se amplía la búsqueda incorporando el parámetro max_features, que controla el número de predictores considerados en cada división:

```
param_grids = {
    'LR':      {'m__fit_intercept': [True, False]},
    'CART':    {'max_depth': [3,5,8,10,None], 'min_samples_split': [2,5,10], ...},
    'RF':              {'n_estimators': [100,300,500], 'max_depth': [...], 'max_features':
['sqrt','log2',0.5], ...},
    'SVR (RBF)': {'m__C': [0.1,1,10,100], 'm__epsilon': [...], 'm__gamma': [...]},
    'MLP':     {'m__hidden_layer_sizes': [...], 'm__activation': [...], 'm__alpha': [...], ...}
}
```

Una particularidad del proceso es el manejo de la convergencia en MLP: dado que las advertencias de ConvergenceWarning no se propagan correctamente a los workers paralelos, el Grid Search para este modelo se ejecuta con n_jobs=1, evitando salidas de error silenciosas en entornos multiproceso. En todos los casos se emplea la misma partición de entrenamiento almacenada en splits_base y el esquema de validación cruzada RepeatedKFold, maximizando la comparabilidad con los resultados base. Los modelos optimizados se almacenan en artefactos para su uso

posterior. Las Figuras 17–20 presentan los resultados post Grid Search para DBO y pH.

Optuna: optimización bayesiana para XGBoost

Para XGBoost se utiliza Optuna con el muestreador TPE (Tree-structured Parzen Estimator) y el podador MedianPruner. Este enfoque es más adecuado que una búsqueda exhaustiva dado el gran espacio de hiperparámetros del modelo. A diferencia de la versión previa del pipeline, el espacio de búsqueda se amplía para incluir parámetros de regularización adicionales (reg_alpha y reg_lambda para L1 y L2), control del crecimiento del árbol (min_child_weight y gamma) y rangos más amplios para n_estimators y max_depth:

```
def objective_xgb(trial, X_train, y_train, cv):
  params = {
    'n_estimators':     trial.suggest_int('n_estimators', 100, 2000),
    'max_depth':        trial.suggest_int('max_depth', 3, 12),
    'learning_rate':    trial.suggest_float('learning_rate', 0.01, 0.3, log=True),
    'subsample':        trial.suggest_float('subsample', 0.5, 1.0),
    'colsample_bytree': trial.suggest_float('colsample_bytree', 0.5, 1.0),
    'reg_alpha':        trial.suggest_float('reg_alpha', 1e-3, 10.0, log=True),
    'reg_lambda':       trial.suggest_float('reg_lambda', 1e-3, 10.0, log=True),
    'min_child_weight': trial.suggest_int('min_child_weight', 1, 8),
    'gamma':            trial.suggest_float('gamma', 0.0, 2.0),
    'random_state': 42
  }
```

La función objetivo minimiza el error cuadrático medio calculado mediante validación cruzada (cross_val_score), a diferencia de la versión anterior que evaluaba directamente sobre el conjunto de prueba. Esto proporciona una estimación más robusta del desempeño generalizable durante la búsqueda. Cada estudio se crea con TPESampler(seed=42) para garantizar reproducibilidad de los trials:

```
study = optuna.create_study(
  direction='minimize',
  sampler=TPESampler(seed=42),
  pruner=MedianPruner()
)
study.optimize(lambda trial: objective_xgb(trial, X_tr, y_tr, CV), n_trials=60)
```

El proceso se ejecuta sobre las cuatro combinaciones escenario × objetivo manteniendo las mismas particiones de splits_base, lo que garantiza comparabilidad directa con los resultados de

Grid Search. Los mejores modelos y sus hiperparámetros se almacenan en artefactos.

Selección de los mejores modelos finales

La selección de los modelos finales para cada objetivo se realiza de forma explícita y documentada, fijando las combinaciones que corresponden a los resultados reportados en el estudio: Random Forest con Grid Search en el escenario rfe_ph para DBO, y XGBoost con Optuna en el escenario vif para pH. Esta decisión manual garantiza reproducibilidad total con el documento académico:

```
df_dbo = df_tuning[(df_tuning['objetivo'] == 'DBO') &
                   (df_tuning['modelo']   == 'RF') &
                   (df_tuning['search']   == 'grid') &
                   (df_tuning['escenario']== 'rfe_ph')]
```

Se incorpora un mecanismo de fallback que, en caso de no encontrar la combinación exacta en los artefactos almacenados, selecciona automáticamente el mejor modelo disponible para cada objetivo. Esto garantiza la continuidad del flujo incluso en entornos donde el tuning se ejecutó de forma parcial.

Interpretabilidad DBO: Permutation Importance y SHAP

Una vez identificado el mejor modelo para DBO, se calcula la importancia de las variables mediante dos técnicas complementarias. Permutation Importance (Figura 21a) evalúa la degradación del desempeño del modelo cuando se altera aleatoriamente cada predictor, manteniendo intacta la estructura del modelo entrenado. La evaluación se realiza con 30 repeticiones sobre el conjunto de prueba:

```
result_pi_dbo = permutation_importance(
   mejor_modelo_dbo, X_te_dbo, y_te_dbo,
   n_repeats=30, random_state=42, n_jobs=-1
)
```

Para la interpretabilidad SHAP se implementan tres niveles de fallback que garantizan la obtención de valores SHAP independientemente del tipo de modelo o versión de las librerías. El primer intento utiliza TreeExplainer directamente sobre el objeto sklearn; si falla, se intenta con el booster nativo de XGBoost; y como último recurso, se emplea el Explainer genérico con la función de predicción como callable. Esta estrategia robusta evita interrupciones del flujo por incompatibilidades de versión:

```
try:
   explainer_dbo   = shap.TreeExplainer(mejor_modelo_dbo)  # intento 1
   shap_values_dbo = explainer_dbo.shap_values(X_shap_dbo)
```

except: ... # intento 2: Explainer genérico

Los resultados SHAP se presentan como ranking global de importancia media en valor absoluto (Figura 21) y como diagrama beeswarm que muestra dirección e intensidad del efecto de cada predictor sobre las predicciones (Figura 22). Ambas figuras se almacenan también como archivos PNG. Todos los resultados se guardan en artefactos para su reutilización sin recómputo.

Interpretabilidad pH: Permutation Importance y SHAP

El mismo flujo de interpretabilidad se aplica al mejor modelo para pH, con los ajustes correspondientes a su naturaleza (XGBoost entrenado sobre el escenario vif). Los tres intentos de SHAP se ejecutan secuencialmente, con indicación en log del método que resultó exitoso. Los resultados se presentan en las Figuras 24 (ranking global SHAP), 25 (beeswarm pH) y los artefactos se almacenan en artefactos['compacto_shap_ph'] para uso en el resumen final.

Curvas de aprendizaje

Como complemento diagnóstico, se generan curvas de aprendizaje para los mejores modelos de DBO (Figura 23) y pH (Figura 26). Estas curvas permiten observar la evolución del RMSE a medida que aumenta el tamaño del conjunto de entrenamiento, ayudando a identificar posibles problemas de sesgo o varianza. El cálculo se realiza con validación cruzada de 5 pliegues y la métrica neg_root_mean_squared_error, manteniendo consistencia con las métricas empleadas en las fases anteriores:

```
train_sizes, train_scores, val_scores = learning_curve(
    mejor_modelo_dbo, X_dbo_full, y_dbo_full,
    cv=5, scoring='neg_root_mean_squared_error',
    train_sizes=np.linspace(0.1, 1.0, 10), random_state=42
)
```

Los resultados incluyen bandas de incertidumbre (media ± desviación estándar) para los conjuntos de entrenamiento y validación, visualizando de forma simultánea el comportamiento del error en ambos conjuntos. Los datos de las curvas se almacenan como artefactos para posible visualización posterior bajo demanda.

Resumen final del pipeline

Al finalizar todas las fases de evaluación, la celda de resumen consolida en un único bloque los indicadores clave del pipeline: dimensiones del dataset original, número de variables retenidas tras completitud y VIF, número total de combinaciones evaluadas y métricas de los modelos seleccionados. De forma complementaria, se generan gráficos de predicho vs. real para DBO y pH, que permiten visualizar la distribución del error y detectar patrones sistemáticos de subpredicción o sobrepredicción. Todos los modelos, métricas y resultados quedan almacenados en el diccionario artefactos, garantizando trazabilidad, reproducibilidad y posibilidad de extensión del

estudio sin necesidad de reentrenar.

4. SÍNTESIS METODOLÓGICA

El estudio se desarrolló bajo un enfoque cuantitativo y predictivo, utilizando técnicas de aprendizaje supervisado para modelar variables de calidad del agua a partir de un conjunto estructurado de variables explicativas. El proceso metodológico se diseñó como un flujo reproducible que integra selección de variables, entrenamiento de modelos, ajuste de hiperparámetros, evaluación comparativa e interpretabilidad, con trazabilidad garantizada mediante la semilla fija RANDOM_STATE = 42 y el repositorio centralizado artefactos.

Previo al modelado, los datos fueron sometidos a procesos de depuración, imputación y filtrado. La imputación KNN ponderada por distancia (k=5) operó sobre los datos estandarizados para garantizar comparabilidad entre magnitudes heterogéneas. El umbral de completitud del 80 % permitió retener las variables con cobertura temporal suficiente, mientras el filtro Spearman ($|\rho| \geq 0.85$) y el VIF iterativo (umbral=10) depuraron la multicolinealidad de forma secuencial y transparente. A partir de este conjunto consolidado se construyeron particiones independientes de entrenamiento y prueba, almacenadas en splits_base para garantizar coherencia entre escenarios.

La selección de variables se abordó mediante RFE con votación multi-modelo (Regresión Lineal, Ridge y SVR lineal), ejecutada con preselección top-k por Spearman y dimensión dinámica k_final = max(5, $\sqrt{p}$), generando cuatro escenarios de predictores: VIF, RFE-DBO, RFE-pH y RFE-Unión.

Sobre cada escenario y variable objetivo se entrenaron seis algoritmos de regresión: Regresión Lineal, CART, Random Forest, SVR (RBF), MLP y XGBoost, bajo el mismo esquema de partición 80/20. El ajuste de hiperparámetros combinó Grid Search con validación cruzada RepeatedKFold (5 pliegues, 3 repeticiones) para cinco modelos, y búsqueda bayesiana con Optuna (60 trials, TPESampler con semilla 42, MedianPruner) para XGBoost. El espacio de búsqueda de XGBoost incluyó parámetros de regularización L1 y L2 (reg_alpha, reg_lambda), control del crecimiento del árbol (min_child_weight, gamma) y configuración del muestreo por instancia y columna.

La evaluación del desempeño se realizó sobre los conjuntos de prueba empleando R^2, RMSE y MAE, permitiendo una comparación homogénea entre modelos, escenarios y variables objetivo. Los mejores modelos resultantes fueron Random Forest (Grid Search, escenario rfe_ph) para DBO con R^2=0.898, y XGBoost (Optuna, escenario vif) para pH con R^2=0.411.

Finalmente, se incorporaron técnicas de interpretabilidad para analizar la contribución de las variables predictoras: Permutation Importance (30 repeticiones) para ambos objetivos, y SHAP con TreeExplainer para los modelos tipo árbol, con estrategia de tres niveles de fallback para garantizar robustez ante incompatibilidades de versión. Todos los modelos, métricas y resultados fueron almacenados como artefactos estructurados del proceso, garantizando trazabilidad, reproducibilidad y posibilidad de extensión del estudio.