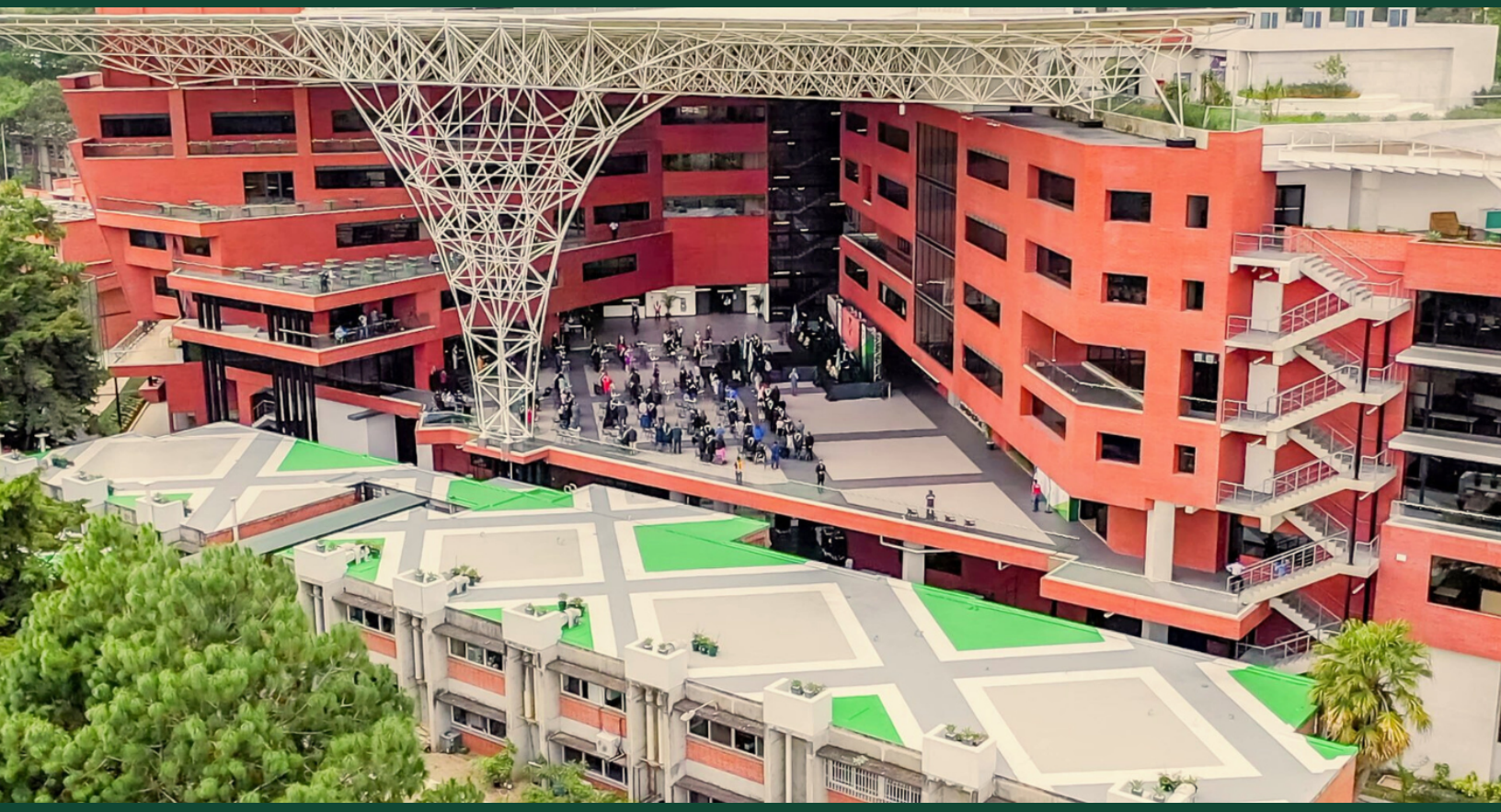
Diseño de un sistema de detección de cianobacteria en cuerpos de agua por medio de aprendizaje automático

Paola Andrea Ayala Pineda



UVG | UNIVERSIDAD DEL VALLE DE GUATEMALA

UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería

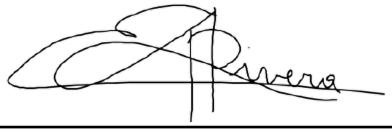
Diseño de un sistema de detección de cianobacteria en cuerpos de agua por medio de aprendizaje automático

Trabajo de graduación presentado por Paola Andrea Ayala Pineda para optar al grado académico de Licenciada en Ingeniería Mecatrónica

Guatemala,

2025

Vo.Bo.:



(f) _____

Dr. Luis Alberto Rivera Estrada

(f) _____

M.Sc. Carlos Esquit Hernández

Fecha de aprobación: Guatemala, 20 de noviembre de 2025.

Prefacio

El presente trabajo nace del profundo interés por contribuir a la recuperación del lago de Amatitlán, un ecosistema vital que enfrenta una problemática de gran magnitud: la contaminación. Desde el inicio de la carrera, se fijó como meta personal aprovechar cada aprendizaje adquirido, ya sea en investigación, experimentación o uso de herramientas tecnológicas, para generar un aporte significativo y planificado a la búsqueda de soluciones sostenibles.

Más allá de un ejercicio académico, este proyecto se concibe como un esfuerzo consciente por sumar un "granito de arena" al cambio, demostrando que desde el ámbito de la ingeniería y la ciencia se pueden desarrollar alternativas modernas que apoyen la protección y conservación de los cuerpos de agua. Se espera que este trabajo no solo refleje la dedicación invertida, sino que también inspire futuras iniciativas que fortalezcan el compromiso hacia la preservación ambiental.

Este trabajo no hubiera sido posible sin el apoyo de quienes me acompañaron en el camino. En primer lugar, agradezco a Dios por la bendición de permitirme llegar hasta aquí; a mis padres, René y Glenda, por darme el regalo más preciado: el estudio, que hoy comienza a dar frutos; a mis hermanas, Pamela y Ana, por alentarme en todo momento; y a mis abuelos, Leticia, Amparo, Jorge y René, quienes han sido fuente de inspiración durante este proceso.

Extiendo también mi gratitud a las instituciones encargadas de la investigación en cuerpos de agua, como Ibagua, Amsa y Cea, así como al Departamento de Biología por su tiempo y apoyo en la recolección de muestras. También agradezco a Miguel Zea y Luis Rivera, por su paciencia y valiosa retroalimentación durante la elaboración de este proyecto. Y sobretodo a Carlos Búcaro, mi pareja, quien ha sido un pilar fundamental, brindándome todo su apoyo, motivación y amor incondicional. A todos ustedes, mi más sincero agradecimiento pues, sin su acompañamiento, este proyecto no habría sido posible.

Índice general

Índice de figuras

Índice de cuadros

Resumen

El trabajo de graduación propuesto consistió en el diseño y la implementación de un sistema de sensores que, junto con técnicas de aprendizaje automático, permitió estimar indirectamente la presencia de cianobacterias en cuerpos de agua. Para ello, se llevó a cabo una serie de mediciones *in situ* en el estanque del jardín botánico de la Universidad del Valle de Guatemala, donde se midieron parámetros fisicoquímicos como turbidez, temperatura, conductividad eléctrica, oxígeno disuelto y pH. Paralelamente, se emplearon datos históricos del lago de Amatitlán, obtenidos con autorización de instituciones como Ibagua, AMSA y CEA, los cuales incluyen registros de clorofila, utilizada como variable indicadora de proliferación de cianobacterias.

La metodología integró técnicas de procesamiento de datos, selección de características relevantes y entrenamiento de modelos de aprendizaje automático, tales como regresión, redes neuronales, k-vecinos más cercanos y máquina de vectores de soporte. El modelo final se seleccionó con base en métricas de desempeño y validación cruzada como matrices de confusión y lógica difusa, evaluando su capacidad para predecir niveles de contaminación en el estanque a partir de los datos obtenidos por los sensores. Este enfoque buscó contribuir al monitoreo ambiental de cuerpos de agua con herramientas accesibles, replicables y basadas en inteligencia artificial, especialmente en contextos donde el monitoreo directo de microalgas resulta costoso o inaccesible.

Palabras clave: aprendizaje automático, cianobacteria, estado trófico, redes neuronales, regresión, matrices de confusión, lógica difusa.

Abstract

The proposed graduation project consisted of the design and implementation of a sensor system that, together with machine learning techniques, allowed for the indirect estimation of cyanobacteria presence in water bodies. To achieve this, a series of *in situ* measurements were carried out in the pond of the Botanical Garden at Universidad del Valle de Guatemala, where physicochemical parameters such as turbidity, temperature, electrical conductivity, dissolved oxygen, and pH were measured. In parallel, historical data from Lake Amatitlán were used, obtained with authorization from institutions such as IBAGUA, AMSA, and CEA, which include chlorophyll records used as an indicator variable of cyanobacterial proliferation.

The methodology integrated data processing techniques, selection of relevant features, and training of machine learning models such as regression, neural networks, k-nearest neighbors, and support vector machines. The final model was selected based on performance metrics and cross-validation, including confusion matrices and fuzzy logic, assessing its ability to predict pollution levels in the pond from the data obtained by the sensors. This approach aimed to contribute to environmental monitoring of water bodies using accessible, replicable, and artificial intelligence based tools, especially in contexts where direct monitoring of microalgae is costly or inaccessible.

Keywords: Machine learning, cyanobacteria, trophic state, neural networks, regression, confusion matrix, fuzzy logic.

CAPÍTULO 1

Introducción

La calidad del agua es un tema prioritario a nivel global, ya que está relacionada con la salud pública, el equilibrio ecológico y el aprovechamiento sostenible de los recursos hídricos. En cuerpos de agua continentales, la proliferación de cianobacterias y microalgas constituye uno de los principales indicadores de deterioro ambiental, afectando tanto la disponibilidad como la calidad de este recurso. En este contexto, la clorofila-a se ha consolidado como un parámetro de referencia para evaluar la biomasa fitoplanctónica y, en consecuencia, para determinar el estado trófico de los ecosistemas acuáticos. Clasificaciones estandarizadas, como las propuestas por organismos internacionales, establecen rangos de concentración que permiten categorizar un lago o estanque en estados ultraoligotrófico, oligotrófico, mesotrófico, eutrófico e hipereutrófico. Sin embargo, la medición directa de clorofila-a implica métodos de laboratorio costosos, lentos y de difícil implementación en sistemas de monitoreo continuo.

Ante este desafío, surge la necesidad de desarrollar metodologías alternativas que permitan estimar la concentración de clorofila-a a partir de variables físico-químicas fácilmente medibles en campo. Entre ellas, destacan la conductividad eléctrica, el pH, el oxígeno disuelto, la turbidez y la temperatura, parámetros que en conjunto brindan información indirecta sobre la dinámica de nutrientes, la actividad fotosintética y las condiciones de hábitat acuático.

Para su detección, es posible implementar un sistema de sensores electrónicos de bajo costo, integrados a un microcontrolador que actúa como unidad de procesamiento y adquisición de datos. Este enfoque no sólo optimiza recursos, sino que también habilita la creación de prototipos funcionales para la investigación aplicada y la gestión ambiental. El avance en técnicas de aprendizaje automático ha potenciado la posibilidad de utilizar estos datos multivariados para predecir concentraciones de clorofila-a con niveles de exactitud competitivos frente a los métodos tradicionales.

Algoritmos como redes neuronales (NN), máquinas de vectores soporte (SVM) y k-vecinos más cercanos (k-NN) permiten identificar relaciones no lineales entre los parámetros fisicoquímicos y la biomasa fitoplanctónica, proporcionando herramientas más robustas y adaptables. Asimismo, la implementación de lógica difusa en la interpretación de resultados

contribuye a manejar la incertidumbre en observaciones cercanas a los límites de clasificación, evitando errores de etiquetado rígido y favoreciendo una representación más realista de la naturaleza continua del fenómeno ecológico.

Este proyecto integra de manera experimental tanto la dimensión tecnológica (diseño e impresión 3D de soportes para sensores, integración electrónica y validación de prototipos en el estanque del jardín botánico de la Universidad del Valle de Guatemala) como la dimensión analítica (entrenamiento de modelos de predicción mediante los conjuntos de datos de AMSA, Ibagua, CEA y pruebas experimentales).

Para ello, se plantearon tres fases principales: diseño y construcción de sistemas de adquisición de datos, recopilación y procesamiento de parámetros de calidad de agua, y entrenamiento, validación y comparación de modelos de predicción de clorofila-a. En paralelo, se llevaron a cabo pruebas piloto que permitieron ajustar el diseño experimental, calibrar sensores y evaluar la viabilidad de las mediciones en condiciones reales.

De esta manera, el presente trabajo no solo busca demostrar la pertinencia del uso de sensores de bajo costo y técnicas de aprendizaje automático en la estimación indirecta de clorofila-a, sino también establecer una base metodológica para futuros desarrollos en el monitoreo autónomo de cuerpos de agua. En particular, se pretende ofrecer un sistema flexible, replicable y de bajo presupuesto que contribuya al control temprano de floraciones algales y a la gestión sustentable de los recursos hídricos, ayudando a la toma de decisiones tanto en ámbitos académicos como institucionales.

CAPÍTULO 2

Antecedentes

La eutrofización, causada principalmente por la contaminación con nutrientes como nitrógeno y fósforo, junto con la acumulación de materia orgánica en cuerpos de agua dulce, ha provocado un incremento en los eventos de floración de cianobacterias a nivel mundial. Estas floraciones pueden liberar toxinas perjudiciales para los seres humanos, la fauna acuática y los ecosistemas en su conjunto, representando un riesgo tanto ambiental como sanitario [1].

En el contexto guatemalteco, el lago de Amatitlán representa un caso crítico de deterioro ambiental. Los reportes anuales de la Autoridad para el Manejo Sustentable de la Cuenca del Lago de Amatitlán (AMSA) [2] documentan altos niveles de nutrientes, disminución del oxígeno disuelto y presencia recurrente de biomasa algal, todos ellos indicadores típicos de un ecosistema eutrofizado.

Estudios realizados por Rodas y Vásquez [3], en colaboración con la Asociación para la Investigación e Innovación Biotecnológica por el Agua (IBAGUA) [4], evaluaron la composición, abundancia y distribución del fitoplancton en el lago de Amatitlán, estableciendo una relación clara entre la calidad del agua y la presencia de cianobacterias.

A partir del análisis de parámetros fisicoquímicos y biológicos, se identificó al género *Microcystis* como el dominante, y se observó un aumento de diversidad algal durante la época seca, atribuido a la mayor radiación solar y acumulación de nutrientes. Como parte de las acciones para mitigar el deterioro, IBAGUA ha implementado tecnologías de tratamiento de aguas residuales con nanoburbujas, con el objetivo de reducir la carga de contaminantes y frenar las floraciones nocivas.

En un esfuerzo por mejorar el monitoreo ambiental, Cano [5] desarrolló una boya multisensorial de bajo costo para la vigilancia de parámetros físicos y químicos en el lago de Atitlán. El sistema integra sensores de pH, turbidez, oxígeno disuelto, temperatura y velocidad del viento, junto con módulos GPS y GSM para la geolocalización y transmisión remota de datos. Si bien el prototipo demostró ser funcional para la recolección de información en campo, no se logró implementar un sistema de análisis automático de datos, limitando así su capacidad predictiva.

Esta situación evidencia la necesidad de integrar enfoques computacionales más avanzados que permitan interpretar grandes volúmenes de datos y generar alertas tempranas de riesgo ecológico.

2.1. Aplicaciones de aprendizaje automático *(machine learning)*

Rodríguez-Rangel et al. [6] aplicaron cinco modelos de aprendizaje automático para predecir la acumulación de carbohidratos en biomasa de cianobacterias cultivadas en aguas residuales. Entre los enfoques utilizados se incluyen redes neuronales artificiales (ANN), redes convolucionales unidimensionales (CNN-1D), redes de memoria a largo plazo (LSTM), así como los algoritmos *k-nearest neighbors* (kNN) y *Random Forest*. Estos modelos fueron entrenados con un conjunto de datos compuesto por 18 variables relacionadas con la calidad del agua, la composición microbiana y las condiciones operativas. El modelo CNN-1D obtuvo el mejor desempeño, demostrando una alta precisión en la predicción del porcentaje de carbohidratos, lo que destaca el potencial del aprendizaje automático para identificar las carácterísticas principales que corresponden a la composición de cianobacteria.

En el estudio realizado por Zhang et al. [7] utilizaron los enfoques LASSO y *Random Forest* para predecir la abundancia de cianobacterias en 331 estanques de acuicultura en el centro de China. Su objetivo fue identificar los factores ambientales clave que favorecen la proliferación de cianobacterias, con el fin de optimizar la gestión del agua en sistemas acuícolas. Los resultados destacaron al carbono orgánico total (TOC) y a la demanda química de oxígeno (COD) como los principales predictores, y evidenciaron la utilidad de estos modelos incluso en entornos con limitaciones de datos.

Finalmente, Zolfaghari et al. [8] examinaron la influencia de la resolución espectral en la estimación de ficocianina (PC), un pigmento representativo de cianobacterias, mediante el uso de sensores remotos y modelos de aprendizaje automático. A partir de 905 muestras con mediciones *in situ* y espectros de reflectancia, se comparó el desempeño de los modelos PLSR, SVR, XGBoost y MLP con datos hiperespectrales y multiespectrales simulados. El modelo MLP, una red neuronal, obtuvo los mejores resultados, particularmente al trabajar con datos del sensor hiperespectral HICO. Este estudio reafirma que una mayor resolución espectral combinada con algoritmos de aprendizaje automático puede mejorar significativamente la detección remota de cianobacterias, superando ampliamente los enfoques empíricos tradicionales.

CAPÍTULO 3

Justificación

Actualmente, la detección de cianobacterias en cuerpos de agua como el lago de Amatitlán se realiza principalmente mediante análisis de laboratorio. Si bien estos métodos ofrecen alta precisión, requieren personal especializado, equipos costosos y tiempos prolongados para la obtención de resultados. Estas limitaciones dificultan su aplicación en esquemas de monitoreo continuo o en tiempo real, lo cual es esencial para responder de manera oportuna ante eventos de floraciones algales tóxicas que ponen en riesgo la salud pública, la biodiversidad y el uso recreativo y productivo del recurso hídrico.

En este contexto, el uso de técnicas de aprendizaje automático en el monitoreo ambiental representa una alternativa innovadora y eficiente, capaz de procesar grandes volúmenes de datos y detectar patrones complejos que permitan predecir condiciones propicias para la proliferación de cianobacterias. Estas herramientas pueden generar alertas tempranas y facilitar la toma de decisiones por parte de autoridades ambientales, reduciendo los impactos negativos de las floraciones algales sobre los ecosistemas acuáticos y las comunidades humanas que dependen de ellos.

No obstante, la implementación efectiva de este tipo de modelos requiere contar con conjuntos de datos suficientemente amplios y representativos, lo que ha sido una limitación en experiencias previas de monitoreo automatizado. En este trabajo se propone el desarrollo de modelos de aprendizaje automático para la predicción de concentraciones de cianobacterias, utilizando técnicas como *Random Forest*, LASSO y redes neuronales, entrenadas con una mayor cantidad de datos ambientales y de calidad del agua. Con ello, se busca contribuir al diseño de un sistema de monitoreo más ágil, económico y escalable, que complemente las técnicas tradicionales y fortalezca la gestión integrada del recurso hídrico.

CAPÍTULO 4

Objetivos

4.1. Objetivo general

Evaluar y aplicar métodos de aprendizaje automático para estimar concentraciones de cianobacterias en el lago de Amatitlán según parámetros fisicoquímicos y ambientales.

4.2. Objetivos específicos

- Gestionar la obtención y validación de datos fisicoquímicos del lago de Amatitlán proveniente de instituciones que estudian cuerpos de agua.
- Analizar y depurar los datos recolectados, identificando las características más relevantes para la estimación de la concentración de cianobacterias.
- Diseñar un sistema de sensores para la recolección de datos ambientales en un entorno experimental controlado.
- Evaluar métodos de aprendizaje automático y aplicarlos a los datos obtenidos para la estimación de la concentración de cianobacterias.

CAPÍTULO 5

Definición del problema

El presente proyecto aborda la problemática del monitoreo de cianobacterias en cuerpos de agua, un aspecto crítico debido a los impactos ambientales, económicos y de salud pública asociados con sus floraciones. En Guatemala, lagos como Atitlán y Amatitlán presentan condiciones propicias para el crecimiento de cianobacterias, lo que genera la necesidad de contar con herramientas accesibles y confiables para su detección y seguimiento. Sin embargo, los métodos tradicionales de análisis resultan costosos, tardados y requieren de infraestructura especializada, lo que limita su aplicación frecuente en contextos locales.

Ante esta situación, se plantea como objetivo el desarrollo varios modelos de aprendizaje automático capaces de predecir concentraciones de cianobacterias a partir de parámetros fisicoquímicos medidos con sensores analógicos. Como variables objetivo se consideran la clorofila-a y la ficocianina, pigmentos característicos de las cianobacterias que funcionan como indicadores robustos de su proliferación.

El alcance del proyecto contempla el entrenamiento y validación del modelo utilizando datos fisicoquímicos provenientes de los lagos de Atitlán y Amatitlán, complementados con mediciones experimentales realizadas en un estanque controlado del Jardín Botánico de la universidad, empleado como cuerpo de agua experimental.

Finalmente, se busca que este trabajo siente las bases para el desarrollo de un sistema de bajo costo que permita obtener mediciones rápidas y confiables de floraciones de cianobacterias, con el fin de facilitar su futura implementación en boyas multisensoriales destinadas al monitoreo en tiempo real de cuerpos de agua.

CAPÍTULO 6

Marco teórico

6.1. Calidad del agua

Un cuerpo de agua se compone por el conjunto de características físicas, químicas y biológicas, estas pueden determinar su calidad dependiendo el uso que el usuario le brinde [9]. Actualmente, la causa principal que afecta la calidad del agua es la eutrofización, esto se define como la proliferación de algas que deteriora la columna de agua y ecosistemas como estanques, lagunas y lagos. Según el nivel de eutrofización, la calidad de agua se reduce y así también perjudica su consumo y efectividad [10].

6.1.1. Características fisicoquímicos

- **Turbidez:** determina la claridad en un cuerpo de agua por medio de las partículas que estén dispersas, suspendidas o que absorben luz. Algunas partículas como la materia orgánica, microorganismos y lodo pueden perjudicar la transparencia en el cuerpo de agua [11].
- **Oxígeno disuelto:** se determina por la cantidad de oxígeno gaseoso que está disuelto en el agua, este es un factor importante para la vida acuática debido a que la gran mayoría de microorganismos dependen de este parámetro para sobrevivir [12].
- **Conductividad:** se refiere a la capacidad de conducir la corriente eléctrica en un cuerpo de agua, está directamente relacionada con minerales y sales que disocian iones los cuales también pueden transportar una cierta cantidad de carga eléctrica [13].
- **Temperatura:** es la magnitud física que expresa la ausencia o el incremento de calor que hay en un entorno en específico [14].
- **pH:** es una medida de alcalinidad o acidez de una muestra que indica la concentración de iones de hidrógeno presentes y su escala es algorítmica con valores entre 0 y 14 [15].

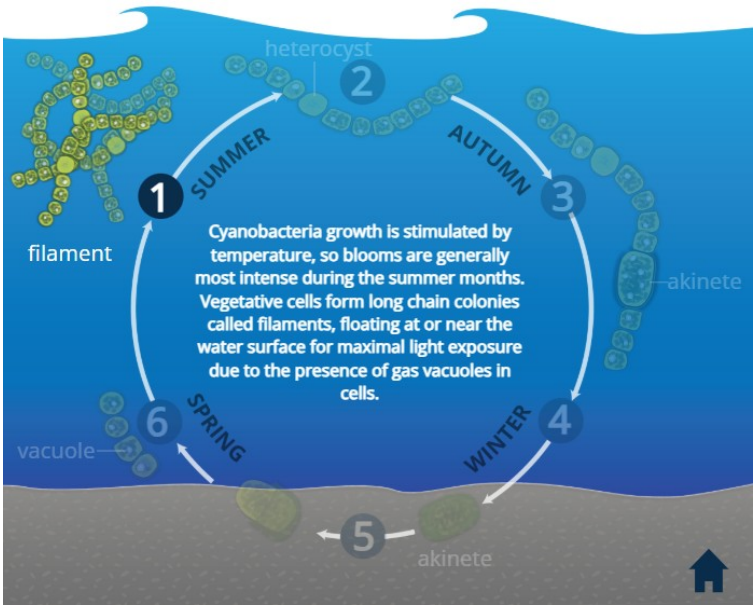
6.1.2. Características biológicas

- **Clorofila-a:** es un pigmento vegetal que absorbe la luz en el espectro azul y rojo, este desempeña una función crucial en la fotosíntesis y puede encontrarse en la Membrana tilacoide (p. 9): [16] de los Cloroplastos (p. 9): [17] ya sea en algas o plantas verdes [18].
- **Ficocianina:** es un pigmento que se define como Ficobiliproteína (p. 9): [19] azul que funciona como una unidad de almacenamiento y actúa en algas verde azuladas [20].

6.1.3. Cianobacteria

Conocida como alga verde azulada, es una bacteria foto sintética, es parte de los microorganismos más importantes del planeta y es, hasta la fecha, motivo de estudios evolutivos, ecológicos y biológicos [21]. Las cianobacterias suelen encontrarse entre el fitoplancton y aunque no son las únicas bacterias Fotótrofas (p. 9): [22], si son dominantes en las zonas oxigenadas de los lagos. Son capaces de fijar nitrógeno y, por lo mismo, pueden desempeñar una función importante en los ciclos de carbono y nitrógeno dependiendo el entorno. Sin embargo, producen toxinas y forman floraciones nocivas en sistemas eutróficos, como se describe en la plataforma dinámica mostrada en la Figura 1 [23].

Figura 1. Ciclo estacional de las cianobacterias.



Nota. La imagen muestra el ciclo anual de las cianobacterias, donde se observa cómo las condiciones ambientales influyen en su crecimiento, formación de filamentos, heterocistos y acinetos. Imagen obtenida de [24].

6.1.4. Estado Trófico

Es un concepto que describe la productividad biológica característica de un ecosistema dentro de las ciencias acuáticas. La productividad puede ser difícil de calcular ya que las Heterogeneidades (p. 10): [25] espaciales y temporales dentro de un ecosistema pueden cambiar sustancialmente en la productividad general de un cuerpo de agua. Variables como la clorofila, la profundidad del Disco de Secchi (p. 10): [26] y los nutrientes son empleadas como *proxies* para evaluar un estado trófico dentro de indicadores y formulaciones que se basan en el color verdadero, estimaciones de biomasa, carbono orgánico, entre otros [27].

No obstante, según la disponibilidad de datos de un cuerpo de agua, el estado trófico se puede expresar como un grupo discreto, índice continuo o como una probabilidad. En este caso, cada formulación indica la incertidumbre en la clasificación [28] como se muestra en el Cuadro 1.

Cuadro 1. Síntesis de esquemas representativos de clasificación del estado trófico.

Grupo de estado trófico	Chl ($\mu g/L$)	Max Chl ($\mu g/L$)
Ultraoligotrófico	<1	<2.5
Oligotrófico	1–2.5	2.5–8
Mesotrófico	2.5–8	8–25
Eutrófico	8–25	25–75
Hipereutrófico	>25	>75

Nota. Índices de estado trófico para la clasificación de lagos y reservas alrededor del mundo. Abreviaciones: Chl es clorofila, Max Chl es máximo valor de clorofila. Referencias de los estados y su definición: Ultraoligotrófico (p. 10): [29], Oligotrófico (p. 10): [29], Mesotrófico (p. 10): [29], Eutrófico (p. 10): [29] y Hipereutrófico (p. 10): [29]. Elaboración de *ECOSPHERE* [27].

6.2. Sistema de sensores

En sistemas de control y monitoreo de calidad de agua como se aplica en la acuicultura, hidroponía, tratamiento de agua, control ambiental, se suele medir una variedad de parámetros físico-químicos que indican el estado del agua. Algunos de los parámetros más relevantes incluyen: conductividad eléctrica (EC), pH, oxígeno disuelto (DO), turbidez y temperatura. Cada uno de estos parámetros aporta información sobre la composición iónica, acidez/alcalinidad, disponibilidad de oxígeno para organismos acuáticos, limpidez o presencia de partículas suspendidas, y condiciones térmicas del medio, respectivamente. Para medirlos se utilizan sensores específicos, que convierten una variable física o química (por ejemplo, concentración de iones, intensidad lumínica, temperatura) en una señal eléctrica (analógica o digital) que puede ser procesada por un microcontrolador (Arduino, ESP32, etc.).

6.2.1. Sensor de tubidez

En la Figura 2 se muestra el sensor que mide la transmisión de luz (o la cantidad de luz que atraviesa el agua) y/o la dispersión de luz: las partículas en la masa de agua disminuyen la cantidad de luz que llega al fotodetector. Ofrece salida analógica (0–4.5 V) y también una salida digital (con umbral ajustable mediante un potenciómetro) para detección binaria de turbidez elevada. Opera típicamente a 5 V con consumo máximo de 40 mA y tiempo de respuesta menor a 500 ms. La salida analógica puede relacionarse con una unidad de NTU *(Nephelometric Turbidity Units)* o similar mediante calibración (aunque la relación voltaje-NTU depende del sensor y el medio) [30].

Figura 2. Sensor de turbidez.

Nota. Representación visual del sensor. Imagen obtenida de DFROBOT [30].

6.2.2. Sensor de oxígeno disuelto

Este sensor utiliza un electrodo galvánico, lo que significa que genera una corriente proporcional a la concentración de oxígeno sin requerir polarización (o con mínima polarización), lo cual facilita su uso inmediato. Como se muestra en la Figura 3, se incluye un módulo conversor que acepta alimentación entre 3.3 V y 5.5 V y produce una salida analógica de 0 a 3.0 V que representa la concentración de oxígeno. El rango de diseño es de 0 a 20 mg/L de oxígeno disuelto. El electrodo requiere mantenimiento: su membrana permeable es reemplazable, se necesita rellenar solución electrolítica (generalmente NaOH en el interior de la membrana), y la membrana debe cambiarse periódicamente dependiendo de la calidad del agua (cada 1 a varios meses). El tiempo de respuesta (responder al 98 por ciento del valor) es típicamente hasta 90 segundos a 25 °C en condiciones ideales [31].

Figura 3. Sensor de oxígeno disuelto.

Nota. Representación visual del sensor. Imagen obtenida de DFROBOT [31].

6.2.3. Sensor de conductividad

Este sensor mide la conductividad eléctrica de una solución, es decir, su capacidad para conducir corriente eléctrica en función de los iones disueltos. Cuanto más alta es la concentración de iones, mayor la conductividad. Se utiliza una excitación de señal AC para reducir el efecto de polarización en los electrodos, lo que mejora la estabilidad de la medición y prolonga la vida útil del sensor. El módulo de señal convierte la medición de corriente/resistencia en una salida analógica (voltaje) compatible con microcontroladores de 3.0 a 5.0 V, con filtrado de *hardware* para reducir el *jitter* en la señal (variación temporal de la señal) [32]. El rango de medición típico soportado es de 0 a 20 mS/cm, con un rango recomendado de 1 a 15 mS/cm para mayor precisión [33]. La Figura 4 muestra que el sensor cuenta con sus concentraciones para su calibración.

Figura 4. Sensor de conductividad.

Nota. Representación visual del sensor. Imagen obtenida de DFROBOT [33].

6.2.4. Sensor de temperatura

Es un sensor de temperatura digital ampliamente usado que proporciona una salida digital (bus único *"1-wire"*) en lugar de una señal analógica (Figura 5). Ofrece una resolución de hasta 12 bits (0.0625 °C por bit) y puede operar en un rango de aproximadamente de –55 °C a +125 °C.

En conjunto con sensores de conductividad, pH, DO, etc., permite aplicar compensaciones de temperatura (por ejemplo, para corrección de conductividad) y monitorear condiciones térmicas del medio [34].

Figura 5. Sensor de temperatura.

Nota. Representación visual del sensor. Imagen obtenida de Tettsa [34].

6.2.5. Sensor de potencial de hidrógeno

El sensor consta de un electrodo de vidrio (generalmente de tipo *"laboratory grade"*) y un módulo conversor que adapta la señal del electrodo a un voltaje analógico legible para el microcontrolador. El módulo conversor admite una fuente de alimentación entre 3.3 V y 5.5 V, y genera una salida analógica (0 a 3.0 V). Para mejorar la precisión, este módulo incluye filtrado de señal y utiliza calibración de dos puntos (usualmente con las soluciones estándar pH 4.0 y 7.0), identificándolas automáticamente si es necesario [35]. Ver Figura 6.

Figura 6. Sensor de pH.

Nota. Representación visual del sensor. Imagen obtenida de DFROBOT [35].

6.3. Aprendizaje automático

El aprendizaje automático (machine learning) es una rama fundamental dentro del campo de la inteligencia artificial (IA), cuyo propósito principal es entrenar a los sistemas computacionales de la capacidad de aprender automáticamente a partir de datos, mejorando progresivamente su desempeño sin necesidad de ser programados explícitamente para cada tarea específica. En otras palabras, se busca que las máquinas sean capaces de identificar patrones, inferir relaciones y tomar decisiones basándose en información empírica [36].

6.3.1. Clasificación de los tipos de aprendizaje automático

El aprendizaje automático se puede categorizar principalmente en tres grandes enfoques aprendizaje supervisado, no supervisado y por refuerzo. Cada uno de estos tipos responde a una metodología distinta en cuanto al uso de los datos y al objetivo que se busca alcanzar [37]. A continuación, se describen detalladamente.

6.3.2. Aprendizaje supervisado

El aprendizaje supervisado se basa en el uso de un conjunto de datos etiquetados, es decir, que contiene ejemplos de entrada junto con sus salidas deseadas o respuestas esperadas. El objetivo del algoritmo es aprender una función de mapeo que permita predecir la salida correspondiente a nuevas entradas no vistas previamente [37].

Tipos de problemas supervisados:

- **Clasificación:** se refiere a aquellos casos donde la variable objetivo es categórica. Ejemplos incluyen el reconocimiento de correos *spam*, diagnóstico de enfermedades, o clasificación de imágenes [38].
- **Regresión:** en estos casos, la salida es una variable continua. Ejemplos comunes son la predicción del precio de una vivienda, la estimación de temperatura o la demanda energética [39].

Modelos comunes en aprendizaje supervisado:

- **Regresión lineal:** modelo matemático que representa la relación lineal entre una variable dependiente continua y una o más variables independientes [40].
- **Regresión logística:** utilizado para clasificación binaria. Modela la probabilidad de pertenencia a una clase mediante la función sigmoide [40].
- **Árboles de decisión *(Random trees)*:** algoritmos que segmentan iterativamente el espacio de características para tomar decisiones jerárquicas basadas en reglas [41].

- **Bosques aleatorios *(Random forests)*:** conjunto de árboles de decisión que operan en paralelo y cuyas predicciones se agregan por votación (clasificación) o promedio (regresión) [41].
- **Máquinas de vectores soporte *(Support Vector Machines, SVM)*:** encuentran el hiperplano óptimo que separa clases en un espacio de características, maximizando el margen [42].
- **k-Vecinos más cercanos *(k-Nearest Neighbors, k-NN)*:** clasifica una observación basada en las clases de sus k vecinos más cercanos en el espacio de entrada [43].
- **Redes neuronales artificiales *(Artificial Neuronal Network, ANN)*:** modelos inspirados en la estructura del cerebro humano, compuestos por capas de neuronas artificiales que aprenden representaciones complejas de los datos [44].

6.3.3. Aprendizaje no nupervisado

A diferencia del aprendizaje supervisado, en este enfoque no se dispone de etiquetas o salidas conocidas. El objetivo principal es explorar la estructura subyacente de los datos, descubrir patrones ocultos y representar la información de forma más eficiente [37].

Tipos de problemas no supervisados:

- **Agrupamiento *(Clustering)*:** consiste en dividir un conjunto de datos en grupos o clústeres de manera que los elementos dentro de un mismo grupo sean más similares entre sí que con los de otros grupos [45].
- **Reducción de dimensionalidad:** tiene como finalidad disminuir el número de variables conservando la mayor cantidad posible de información relevante. Es especialmente útil para visualización de datos y eliminación de ruido [46].
- **Detección de anomalías:** consiste en la identificación de observaciones que se desvían significativamente de la mayoría de los datos [47].
- **Representaciones:** tiene como objetivo encontrar formas compactas y útiles de describir los datos para que otros modelos puedan explorarlas [48].

Modelos comunes en aprendizaje no supervisado:

- **K-medias *(k-means)*:** minimiza la distancia entre puntos y los centroides de cada grupo, ya que el algoritmo divide los datos en "k" grupos [49].
- **Componente principal de análisis *(Principal Component Analysis)*:** transforma los datos en un nuevo sistema de coordenadas, amplia la visualización y elimina el ruido [50].
- ***Autoencoders*:** son redes neuronales que están entrenadas para comprimir y reconstruir datos con el objetivo de detectar anomalías [51].

- **Redes generativas antagónicas:** consisten en dos redes neuronales que compiten entre sí, la primera es el generador la cual produce ejemplos sintéticos de datos y la segunda es el discriminador el cual intenta distinguir los datos si son reales o no [52].

6.4. Lógica difusa

Se refiere a una familia de lógicas multivaluadas, donde los valores de verdad se interpretan como grados de verdad. En este caso, el valor de verdad de una proposición lógicamente compuesta como por ejemplo "persona 1 es alto y persona 2 es adinerado", se determina por el valor de verdad de sus componentes. En otras palabras, la lógica difusa se impone la funcionalidad de verdad al igual que en la lógica clásica [53].

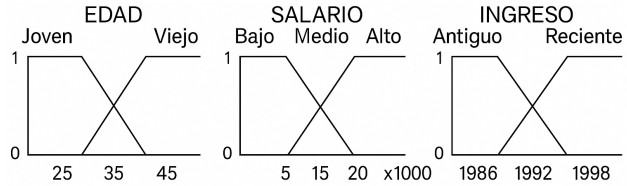
Otra descripción de lógica difusa puede ser un conjunto de procedimientos para manejar la información precisa. En sistemas de bases de datos se tiene como propósito la organización de la información, a continuación de muestra un ejemplo al respecto donde en el Cuadro 2 hay una pequeña base de datos y en la Figura 7 una gráfica de ejemplo que brinda la interpretación de "Cuáles son los nombres de las personas jóvenes o recientemente empleadas, pero con sueldo alto" [54].

Cuadro 2. Base de datos tradicional de ejemplo.

Nombre	Edad	Salario	Año de Ingreso
Anderson	30	20.000	1995
Brown	30	15.000	1995
Long	25	40.000	1993
Nelson	55	20.000	1980
Smith	25	23.000	1996

Nota. Resultados de clasificación para dos clases, positivo y negativo. Cuadro obtenido de la Universidad Nacional de Colombia [54].

Figura 7. Variables lingüísticas para el ejemplo.



Nota. Interpretación visual de cada parámetro. Imagen obtenida de la Universidad Nacional de Colombia [54].

6.5. Matrices de confusión

Una matriz de confusión resume directamente el rendimiento de clasificación de un clasificador con respecto a un conjunto de datos de prueba. Un caso especial de la matriz de confusión se utiliza a menudo con dos clases, una designada como la clase positiva y la otra como la clase negativa [55]. En este contexto, las cuatro celdas de la matriz se designan como verdaderos positivos (VP), falsos positivos (FP), verdaderos negativos (VN) y falsos negativos (FN), como se indica en el Cuadro 3.

Cuadro 3. Matriz de confusión.

	Clase asignada Positiva	**Clase asignada Negativa**
Clase actual Positiva	VP	VN
Clase actual Negativa	FP	FN

Nota. Resultados de clasificación para dos clases, positivo y negativo. Cuadro obtenido de *Springer Nature* [55].

CAPÍTULO 7

Diseño experimental, entrenamiento y validación

7.1. Definición del problema y variables de estudio

El diseño experimental partió de la identificación del estado trófico como variable de respuesta principal, expresada a través de la concentración de clorofila-a. Las variables independientes fueron los parámetros fisicoquímicos medidos mediante un sistema de sensores de:

- Conductividad eléctrica
- Potencial de hidrógeno
- Oxígeno disuelto
- Turbidez
- Temperatura

Cada parámetro se consideró un factor experimental, dado que su variación influye directamente en la respuesta del sistema. La temperatura funcionó como variable de referencia para la compensación de las demás mediciones. Este enfoque permitió estudiar las correlaciones multivariadas entre las condiciones fisicoquímicas y la concentración de clorofila-a.

7.2. Diseño experimental en espacio controlado

El estanque del jardín botánico de la Universidad del Valle de Guatemala se utilizó como espacio experimental controlado. El diseño se estructuró en dos fases principales:

- **Fase piloto:** se efectuaron pruebas iniciales con un prototipo del sistema de sensores para verificar su calibración, estabilidad y tiempo de respuesta.
- **Fase experimental formal:** se implementó una estructura impresa en 3D para sostener los sensores durante las mediciones. Se establecieron condiciones de muestreo controladas, incluyendo volumen de muestra, tiempo de estabilización, número de réplicas y profundidad de inmersión.

Se definieron rangos de muestreo temporales (días y horas) con el fin de capturar la variabilidad natural del estanque y garantizar la representatividad de los datos. Este diseño permitió obtener una base experimental coherente para la posterior validación del modelo de estimación.

7.3. Estrategia de recolección de datos

La recolección de datos se llevó a cabo mediante gestiones formales con instituciones que monitorean la calidad del agua de distintos lagos de Guatemala. La primera institución que compartió información fue la Asociación para la Investigación e Innovación Biotecnológica por el Agua (IBAGUA), a través de su presidenta, Evelyn Rodas, quien facilitó registros del año 2017 utilizados en su publicación científica *Evaluación anual del fitoplancton y su respuesta a la calidad de agua en el lago de Amatitlán, Guatemala*.

Por otra parte, la Autoridad para el Manejo Sustentable de la Cuenca del Lago de Amatitlán (AMSA) publica informes mensuales de calidad del agua. Sin embargo, estos contienen información resumida. Dado que el desarrollo de un modelo de aprendizaje automático requiere un volumen considerable de datos (al menos 150 registros por conjunto de variables para este proyecto que tiene 5 características), se gestionó una solicitud formal de acceso a la información pública para obtener las series completas del período 2017–2024.

Asimismo, se estableció comunicación con el Centro de Estudios de Atitlán (CEA), cuyo jefe de laboratorio, Moisés López, proporcionó registros de 2020 hasta la fecha, con un total superior a 10,000 observaciones. La integración de ambos conjuntos de datos permitió construir un marco comparativo entre diferentes ecosistemas acuáticos.

7.3.1. Preprocesamiento y depuración de datos

Los registros obtenidos se sometieron a un tratamiento previo que incluyó:

- Eliminación de valores extremos y no aplicados (NA) o no registrados (NR).
- Normalización y estructuración de las variables para garantizar la coherencia dimensional.
- Agrupación de muestras según concentraciones de clorofila y estados tróficos (Ultraoligotrófico, Oligotrófico, Mesotrófico, Eutrófico e Hipereutrófico).

Este proceso aseguró la comparabilidad de los datos y su idoneidad para el análisis multivariante y el entrenamiento de modelos de aprendizaje automático.

7.3.2. Enfoque de análisis de datos

El grueso de los análisis realizados y los algoritmos aplicados se basaron en los conjuntos de datos proporcionados por las instituciones AMSA y CEA. Estos registros, por su volumen y variabilidad, permitieron entrenar y validar modelos con suficiente representatividad y robustez estadística.

Los datos recolectados en el estanque del jardín botánico se emplearon únicamente para pruebas piloto y validaciones experimentales del sistema sensorial, orientadas a comprobar la estabilidad del hardware y la coherencia entre las mediciones locales y las predicciones del modelo.

7.4. Modelado con aprendizaje automático como herramienta analítica

En el diseño experimental, se incorporaron modelos de aprendizaje supervisado (regresión, *Support Vector Machines* (SVM), *k-Nearest Neighbors* (k-NN) y *Deep Neural Network* (DNN)) como herramientas analíticas avanzadas. Estos modelos se entrenaron para identificar las relaciones entre los factores fisicoquímicos (pH, conductividad, oxígeno disuelto, turbidez y temperatura) y la variable de respuesta (clorofila-a).

Se aplicaron métricas de evaluación como R^2, RMSE y MAE, como también clasificadores tipo matrices de confusión junto con lógica difusa para validar la clasificación de los rangos tróficos. Adicionalmente, se empleó lógica difusa para manejar la incertidumbre en observaciones cercanas a los límites de cada categoría, lo que incrementó la capacidad interpretativa del modelo y su aplicabilidad en entornos reales.

7.5. Análisis de resultados y validación

El análisis de resultados se centró en:

- Determinar la importancia de cada parámetro fisicoquímico sobre la predicción de clorofila-a, identificando a la turbidez y la conductividad como variables dominantes.
- Implementar validación cruzada como alternativa a las pruebas de hipótesis tradicionales, asegurando la estabilidad y generalización del modelo.
- Clasificar el estado trófico del estanque mediante la comparación entre valores observados y predichos, evaluando el desempeño del sistema de sensores.

7.6. Análisis y selección de características

El análisis de características se basó en identificar qué variables fisicoquímicas están más asociadas a la variabilidad de clorofila-a. Para ello, se emplearon técnicas de correlación, *boxplots* comparativos y análisis de importancia de variables en modelos. La selección se redujo a cinco parámetros clave: pH, temperatura, conductividad, oxígeno disuelto y turbidez, que mostraron mayor estabilidad y relevancia estadística.

Se implementó un *pipeline* de limpieza y transformación de datos que incluyó: reemplazo de valores NR, normalización de variables con *StandardScaler* y exclusión de parámetros irrelevantes. Posteriormente, se evaluó la importancia relativa de cada característica. El análisis reveló que la turbidez y la conductividad tienen mayor peso en la predicción de clorofila-a, mientras que parámetros secundarios no mejoraban de forma significativa el rendimiento del modelo.

En la práctica, la selección de características también respondió a la disponibilidad de sensores confiables y calibrables en campo. Por ello, el modelo se ajustó a las variables que los sensores de la boya multisensorial de Cano [5], pudiesen medir con precisión. Esta decisión asegura coherencia entre la teoría y la implementación real del dispositivo, garantizando que el modelo funcione de forma continua en entornos como el lago Amatitlán o Atitlán.

7.7. Implementación de algoritmos de aprendizaje automático

El sistema de predicción se implementó en *Python*, integrando bibliotecas como *scikit-learn, TensorFlow* y *pandas*. Se aplicaron diferentes enfoques de aprendizaje automático para comparar su desempeño: algoritmos clásicos de clasificación (SVM y KNN), modelos basados en árboles *(Random Forest)* y redes neuronales. Este diseño experimental buscó balancear precisión, interpretabilidad y facilidad de despliegue.

El *pipeline* de implementación incluyó: división de datos en entrenamiento y validación, normalización de entradas, ajuste de hiperparámetros y evaluación mediante métricas de clasificación. En el caso de redes neuronales, se aplicó *early stopping* para evitar sobreajuste y se visualizaron las curvas de pérdida y precisión. El uso de matrices de confusión permitió evaluar la efectividad en la clasificación binaria ($>$40 µg/L y $<$40 µg/L de clorofila-a).

Previo a la presentación de los resultados de los modelos seleccionados, es importante aclarar que, a lo largo del proyecto, la expresión "redes neuronales" se refiere específicamente a redes neuronales profundas de tipo *Feedforward* o "de avance directo", implementadas mediante un *Multilayer Perceptron (MLP)*, conocido en español como "Perceptrón multicapa". Este tipo de red está conformado por tres tipos de capas: una capa de entrada, una o varias capas ocultas y una capa de salida.

La capa de entrada recibe los valores de los parámetros fisicoquímicos proporcionados por las instituciones. Las capas ocultas están formadas por perceptrones (neuronas artificiales) conectados mediante operaciones lineales y funciones de activación no lineales, responsables del proceso de aprendizaje del modelo. Cabe señalar que algunas transformaciones, como la transformación logarítmica aplicada a la variable objetivo (clorofila-a), se realizaron antes

del entrenamiento para mejorar la estabilidad del proceso de optimización, no dentro de las capas ocultas. Finalmente, la capa de salida está compuesta por una única neurona, encargada de generar la predicción de la concentración de clorofila-a.

7.7.1. Modelo de detección con datos de Ibagua

Primer entrenamiento

Durante el primer entrenamiento realizado con el conjunto de datos de Ibagua, que constaba de 50 datos en total, se utilizó una división en la cual el 80 % de las muestras (40 datos) se destinaron al entrenamiento y el 20 % restante (10 datos) se reservaron para validación. Para este primer entrenamiento, se usaron únicamente redes neuronales profundas que fueron entrenadas a lo largo de 300 épocas. La Figura 8 ilustra la evolución del error, donde se representaron de forma comparativa, la pérdida de entrenamiento y la pérdida de validación.

Figura 8. Evolución de error del primer entrenamiento.

Nota. La imagen muestra el comportamiento de los datos de Ibagua mediante redes neuronales. Elaboración propia.

Se observó una disminución sostenida del error cuadrático medio en el entrenamiento de las redes neuronales profundas aplicadas a los datos de Ibagua, pasando de valores iniciales cercanos a 1.1×10^5 hasta acercarse a cero a medida que avanzaban las épocas. Este comportamiento indicó que el modelo fue capaz de ajustarse progresivamente a los datos de entrenamiento, reduciendo de manera continua la discrepancia entre las predicciones y los valores reales del conjunto utilizado para aprender. La tendencia descendente y estable de la curva azul fue evidencia de que la red poseía la capacidad de memorizar los patrones en el subconjunto de entrenamiento.

Por otro lado, la curva de validación (Figura 8) mostró un comportamiento marcadamente distinto. El error inicial en la validación alcanzó valores del orden de 5.6×10^5, lo que reflejó una discrepancia considerable entre las predicciones iniciales y los valores reales de las

muestras no vistas durante el entrenamiento. En las primeras 50 a 80 épocas, la pérdida de validación descendió rápidamente, llegando a un mínimo cercano a 2.5×10^5. Sin embargo, a partir de ese punto, la tendencia cambió: el error de validación comenzó a incrementarse de manera sostenida y alcanzó valores superiores a los registrados en el mínimo temprano, llegando aproximadamente a 3.4×10^5 hacia el final del entrenamiento.

El patrón observado en ambas curvas fue característico de un fenómeno de sobreajuste. El modelo continuó mejorando su desempeño en el conjunto de entrenamiento, como lo evidenció la reducción progresiva del error azul, pero perdió capacidad de generalización en el conjunto de validación. Esto se manifestó en la divergencia entre las curvas: mientras que la de entrenamiento se reducía continuamente, la de validación presentaba un comportamiento en U, con un mínimo temprano seguido de un deterioro progresivo.

La explicación de este comportamiento estuvo directamente relacionada con las condiciones del conjunto de datos y con las propiedades estadísticas de las variables medidas. En primer lugar, el tamaño de la muestra fue extremadamente reducido. El hecho de contar con únicamente 10 muestras en validación implicó que cada dato tuviera un peso muy elevado en el cálculo del error cuadrático medio. De esta forma, un solo valor con una discrepancia significativa entre la predicción y el valor real podía alterar de manera notable el valor global de la pérdida de validación, generando una alta variabilidad en la curva naranja.

En segundo lugar, la distribución de la variable de salida la concentración de clorofila-a presentó una marcada asimetría y tendencia a valores extremos. Este tipo de distribución produjo que la función de pérdida (MSE) se viera dominada por pocos puntos con valores altos de clorofila-a, en los cuales el error absoluto era mayor y, al elevarse al cuadrado incrementaba de forma desproporcionada el error total. Así, cuando las muestras con valores extremos quedaron ubicadas en el subconjunto de validación, la curva reflejó una magnitud de error superior y más volátil.

Otro factor relevante fue la heterogeneidad propia de los datos ambientales. El error de predicción fue mayor en rangos Moderado de clorofila-a que en valores Bajo, debido a la mayor variabilidad natural en concentraciones elevadas. Dado que la validación contenía pocos puntos y algunos de ellos correspondían a concentraciones altas, la pérdida de validación tendió a ser mayor y menos estable que la de entrenamiento.

Además, la selección aleatoria de las muestras para validación pudo haber generado un desbalance entre los conjuntos, con posibles diferencias en la distribución de las condiciones ambientales representadas en cada subconjunto. En particular, fue probable que las observaciones de entrenamiento se concentraran en rangos medios o bajos de clorofila, mientras que en la validación quedaron almacenados casos con valores atípicos o más extremos. Esta desalineación entre las distribuciones de entrenamiento y validación reforzó la divergencia observada entre las curvas.

Segundo entrenamiento

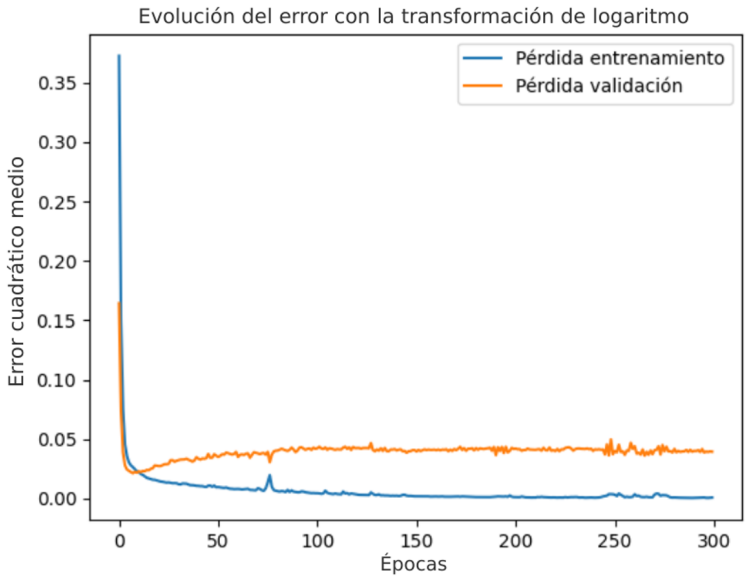
En el segundo entrenamiento con redes neuronales profundas (ver Figura 9), se aplicó una transformación logarítmica a la variable de salida mediante la expresión mostrada en la ecuación (1), con el fin de modificar la escala de la concentración de clorofila-a. Esta técnica

permitió reducir la influencia de valores extremos presentes en la base de datos de Ibagua, favoreciendo un ajuste más estable durante el proceso de entrenamiento del modelo.

$$\log 1p(y) = \log(1 + y) \tag{1}$$

Su aplicación tuvo como efecto principal la reducción de la asimetría y la heterogeneidad de la variable objetivo, generando un comportamiento más estable durante el proceso de optimización.

Figura 9. Evolución de error del segundo entrenamiento.



Nota. La imagen muestra el cambio de comportamiento durante el entrenamiento y validación cuando se aplicó la transformación de logaritmo. Elaboración propia.

La gráfica de evolución del error (Figura 9) mostró que tanto la pérdida de entrenamiento como la de validación presentaron una disminución marcada en las primeras épocas. Inicialmente, los errores se encontraban en valores relativamente elevados (0.35 para entrenamiento y 0.15 para validación), pero ambos descendieron de manera rápida durante las primeras 10–20 épocas, hasta alcanzar un régimen más estable. Este comportamiento fue indicativo de que el modelo de redes neuronales profundas logró aprender patrones generales desde etapas tempranas, aprovechando la homogeneización introducida por la transformación de logaritmo.

Posteriormente, la curva de entrenamiento (Figura 9) continuó descendiendo de manera sostenida hasta acercarse a valores cercanos a cero. La tendencia descendente del error azul fue evidencia de que el modelo se ajustó progresivamente al conjunto de entrenamiento, reduciendo de manera significativa las discrepancias en la escala logarítmica.

A diferencia del primer experimento sin transformación, el modelo no se limitó a memorizar de manera evidente, sino que mostró un aprendizaje más progresivo y menos abrupto.En

el caso de la curva de validación (Figura 9), se observó que el error disminuyó de forma paralela a la de entrenamiento durante las primeras épocas y luego se estabilizó con respecto a valores cercanos a 0.04–0.05. A lo largo de las 300 épocas, la pérdida de validación presentó ligeras oscilaciones, pero sin la tendencia creciente marcada que se había observado anteriormente. Este comportamiento fue señal de que la generalización del modelo mejoró con respecto al primer entrenamiento, logrando un desempeño más consistente al evaluar sobre datos no vistos.

El hecho de que ambas curvas decrecieran de manera conjunta y alcanzaran un estado estacionario sin divergencia pronunciada, reflejó que la transformación de logaritmo haya mitigado el impacto de las muestras atípicas, que en el entrenamiento previo habían distorsionado la métrica de error. En la práctica, esto permitió utilizar el conjunto completo de datos de Ibagua sin descartar muestras con valores extremos de clorofila-a, los cuales anteriormente dominaban la función de pérdida debido al cálculo cuadrático. Con la transformación, esos valores pasaron a una escala comprimida, lo que equilibró la importancia relativa de todas las muestras.

El resultado obtenido evidenció que el modelo fue capaz de aprender representaciones más robustas y generalizables. La estabilidad de la curva de validación en niveles bajos de error, aun cuando el modelo continuó ajustándose en el conjunto de entrenamiento, fue un signo de que la red logró capturar patrones subyacentes sin sobreajustarse de manera evidente. En términos prácticos, esto indicó que el preprocesamiento mediante la transformación de logaritmo, permitió que el entrenamiento se beneficiara de la totalidad del *dataset* de Ibagua, en lugar de verse limitado por los efectos de pocos valores atípicos de gran magnitud.

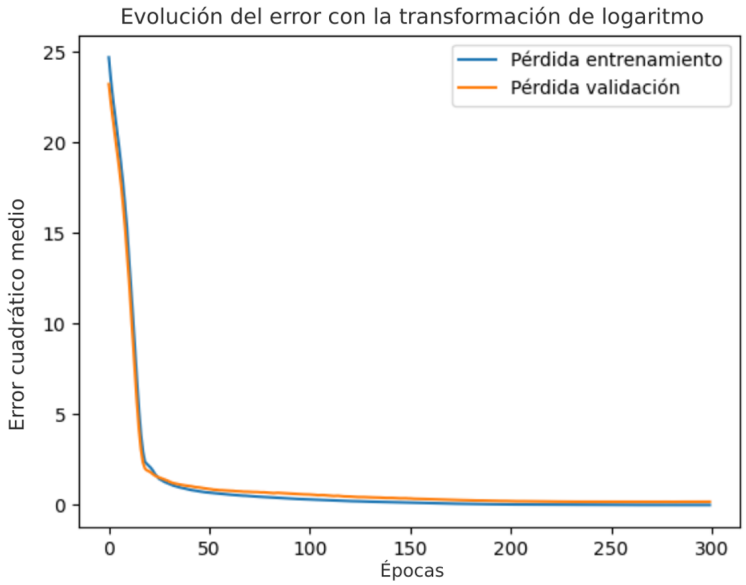
Tercer entrenamiento

En el tercer entrenamiento se implementó un proceso de depuración de la base de datos de Ibagua. Este consistió en la eliminación de registros incompletos en los parámetros de entrada pH, temperatura, conductividad, oxígeno disuelto y turbidez, así como en la exclusión de valores de clorofila-a que aparecían registrados como cero o como NR. La decisión de realizar esta limpieza respondió a la observación de que dichos valores distorsionaban el comportamiento del modelo de redes neuronales profundas, ya que eran interpretados como muestras válidas, pero en realidad no correspondían a mediciones representativas del fenómeno.

La gráfica de evolución del error reflejó un cambio sustancial en el desempeño del modelo tras la depuración (ver Figura 10). Inicialmente, tanto la pérdida de entrenamiento como la de validación comenzaron en valores relativamente altos (25 y 23 respectivamente), pero ambas descendieron de manera pronunciada durante las primeras 20–30 épocas, alcanzando niveles cercanos a 2. A partir de ese punto, la disminución continuó de forma más gradual, y hacia la época 300 ambas curvas convergieron hacia valores cercanos a cero, con un comportamiento prácticamente superpuesto entre entrenamiento y validación.

Este resultado evidenció que la limpieza de datos eliminó una de las principales fuentes de ruido y variabilidad observada en los entrenamientos previos. Los registros de clorofila-a en cero o sin valor habían actuado como valores atípicos, generando fluctuaciones abruptas en la función de pérdida y aumentando la brecha entre entrenamiento y validación.

Figura 10. Evolución de error del tercer entrenamiento.



Nota. La imagen muestra el cambio de comportamiento durante el entrenamiento y validación cuando se aplica la transformación de logaritmo y se discriminan los datos tipo NR o iguales a cero para clorofila-a. Elaboración propia.

Al ser eliminados, el modelo dejó de verse forzado a ajustar casos irreales o inconsistentes, lo que se tradujo en una reducción significativa de la inestabilidad.

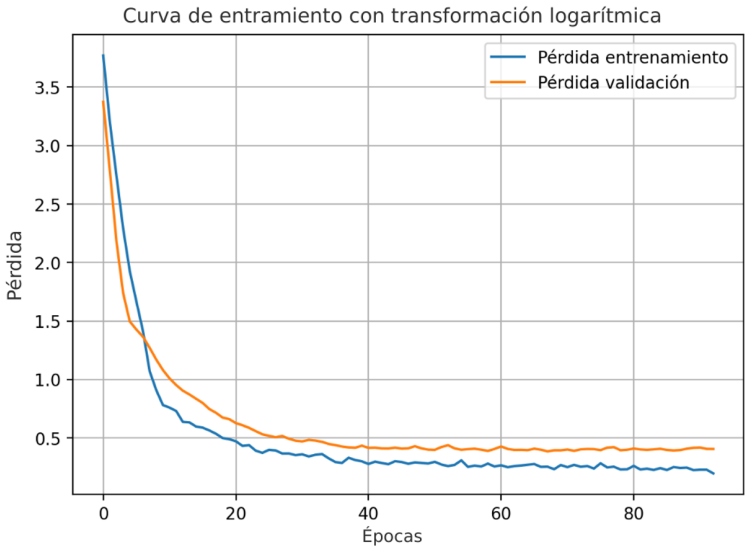
El paralelismo casi exacto entre las curvas azul y naranja mostró que el modelo logró un ajuste balanceado, sin indicios de sobreajuste. La validación acompañó de cerca a la curva de entrenamiento en todo el rango de épocas, manteniendo un error cuadrático medio muy bajo y decreciente (Figura 10). Esto fue un signo de que las muestras que permanecieron tras la limpieza poseían una mayor coherencia estadística y representaban de manera más homogénea la relación entre los parámetros físico-químicos y la concentración de clorofila-a.

No obstante, como consecuencia de la depuración, el número de muestras disponibles se redujo. Esta disminución en el tamaño del conjunto de datos implicó que, aunque el modelo alcanzó un mejor comportamiento en términos de error y estabilidad, la capacidad de generalización futura quedó limitada por la menor cantidad de información con la que fue entrenado. La reducción de la base de datos significó que el modelo trabajó sobre un subconjunto más consistente, pero menos representativo de la variabilidad completa del lago.

7.7.2. Modelos de detección con datos de AMSA

Primer entrenamiento

Figura 11. Entrenamiento de datos de AMSA.



Nota. La imagen muestra el comportamiento del entrenamiento y validación de los datos compartidos por AMSA. Elaboración propia.

El primer entrenamiento con el conjunto de datos de AMSA (con 330 muestras por variable) se usó un modelo de redes neuronales profundas. Para estabilizar la optimización frente a valores extremos, la variable objetivo (clorofila-a) se transformó con la transformación de logaritmo, y la función de pérdida utilizada fue *Huber*, que penalizó de forma cuadrática los errores pequeños y de manera lineal los errores grandes, reduciendo la sensibilidad a atípicos. El objetivo fue aprender una regresión de clorofila-a a partir de pH, temperatura, oxígeno disuelto, conductividad y turbidez, y posteriormente mapear las predicciones continuas a categorías tróficas con umbrales predefinidos (0–2, 2–7, 7–40 y ≥ 40 µg/L).

La curva de entrenamiento mostró (Figura 11) un descenso pronunciado de la pérdida *Huber* desde 3.8 hasta valores próximos a 0.3 durante las primeras 30 épocas, seguido de un régimen de disminución más lenta hasta el final del entrenamiento. La curva de validación siguió una trayectoria paralela: partió alrededor de 3.4, cayó con rapidez hasta 0.5 en las 25–30 primeras épocas y posteriormente se estabilizó alrededor de 0.40–0.45, con oscilaciones de baja amplitud. Se observó una brecha moderada entre ambas curvas (validación $>$ entrenamiento), que fue estable a lo largo de las épocas. Esto se interpretó como un ajuste adecuado al conjunto de AMSA con sobreajuste leve y controlado.

La ausencia de divergencia y la proximidad de las curvas indicaron que la capacidad del modelo estuvo en proporción con el tamaño del conjunto y que la combinación *Huber* más la transformación logarítmica, produjo una convergencia suave, robusta y reproducible. Las predicciones continuas se discretizaron en cuatro clases: Muy bajo (0–2), Bajo (2–7),

Moderado (7–40) y Muy alto ($\geq 40\,\mu g/L$). La matriz de confusión de validación evidenció que el conjunto evaluado prácticamente no contuvo ejemplos de las clases Muy bajo y Bajo como se puede observar en la Figura 41 (ver en Anexos) las celdas con 0 y un solo caso de Bajo, clasificado como Muy alto. En contraste, las clases Moderado y Muy alto concentraron la mayor parte del soporte:

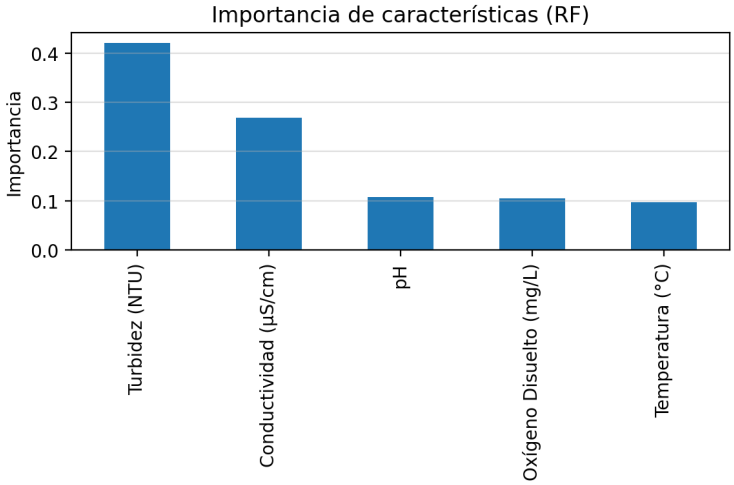
A partir de esta matriz de confusión se calcularon métricas sobre las clases con soporte:

- Fila Muy alto (≥ 40): precisión 0.83 (40/48) y sensibilidad 0.87 (40/46).
- Fila Moderado (7–40): precisión 0.67 (12/18) y sensibilidad 0.63 (12/19)

El acierto global fue 0.79 (52/66). La ausencia de ejemplos en Muy bajo y Bajo impidió estimar métricas significativas para esos rangos y explicó que el modelo tendiera a asignar casos limítrofes hacia las clases superiores, especialmente desde Moderado a Muy alto.

La distribución de clases en validación estuvo desbalanceada hacia Moderado y Muy alto, por lo que el modelo discriminó con mayor confianza en esos rangos y resultó poco informativo para concentraciones bajas. El patrón de errores (confusiones entre Moderado y Muy alto) fue coherente con la proximidad de los umbrales en el espacio de predicción y con la incertidumbre propia de la medición en torno a 40 µg/L.

Figura 12. Gráfica de barras para la clasificación de parámetros con los datos de AMSA.



Nota. La imagen muestra jerarquía de importancia de cada parámetro dentro del modelo entrenado. Elaboración propia.

La importancia de variables obtenida con regresión de bosques aleatorios mostró una jerarquía encabezada por Turbidez y Conductividad (Figura 12). El valor representado en el eje y fue la reducción media de impureza *(Mean Decrease in Impurity, MDI)* normalizada. Cada barra cuantificó la contribución relativa de la variable a la disminución del error de los árboles, acumulada a lo largo de todo el bosque y escalada para que la suma de importancias

sea 1. Por ello, las alturas oscilaron entre 0 y 0.42 (máximo observado) y las barras, en conjunto, sumaron aproximadamente la unidad (las pequeñas discrepancias se debieron al redondeo).

Bajo esta métrica:

- Turbidez (0.42) se posicionó como el predictor dominante, consistente con su papel como *proxy* de sólidos suspendidos y, en múltiples escenarios, como indicador indirecto de biomasa fitoplanctónica.
- Conductividad (0.27) ocupó el segundo lugar, alineada con gradientes iónicos y aportación de nutrientes que modulaban la productividad.
- pH y Oxígeno disuelto (0.10–0.11 cada uno) aportaron señal complementaria, coherente con procesos de fotosíntesis y el estado ácido-base del sistema.
- Temperatura (0.10) contribuyó en menor medida, aunque mantuvo relevancia en tanto controlador cinético del crecimiento de algas.

La coincidencia entre esta jerarquía y el conocimiento limnológico apoyó la validez ecológica del modelo: los dos pilares de la predicción fueron variables con relación directa (o estrechamente asociada) a la concentración de clorofila-a.

El análisis de la matriz de confusión, junto con las curvas de entrenamiento y validación, evidenció que el conjunto de datos de AMSA utilizado en este primer entrenamiento presenta una fuerte concentración de muestras en los rangos Moderado y Muy alto, mientras que los rangos Muy Bajo y Bajo se encuentran subrepresentados. Este desbalance ocasionó que el modelo aprendiera una frontera de decisión desplazada hacia las clases superiores, particularmente alrededor del umbral de 40 µg/L , correspondiente al límite definido por la Organización Mundial de la Salud (OMS) entre las categorías Moderado y Muy alto.

En consecuencia, dicho umbral se ajustó para evitar que pequeñas variaciones dentro del intervalo alto fueran clasificadas erróneamente como valores Moderados. No obstante, la identificación de concentraciones por debajo de 7 µg/L permaneció limitada debido a la escasa cantidad de ejemplos disponibles en esos rangos, reduciendo la capacidad del modelo para delimitar adecuadamente las fronteras inferiores.

Segundo entrenamiento

En esta etapa se buscó mejorar la detección categórica de clorofila-a con los datos de AMSA. En el primer entrenamiento, la discretización rígida por umbrales generó errores en observaciones cercanas a los límites entre clases tróficas (por ejemplo, alrededor de 7 y 40 µg/L), donde una pequeña variación numérica cambiaba la etiqueta de manera abrupta. Para mitigar ese efecto de borde, se incorporó una formulación difusa en la evaluación y se entrenaron clasificadores KNN y SVM con salidas probabilísticas.

La lógica difusa se utilizó para relajar la asignación binaria de clases. En vez de etiquetar cada muestra con una única categoría, cada muestra tuvo grados de pertenencia $\mu \in [0, 1]$

[54] a los cuatro rangos tróficos (Muy bajo 0–2, Bajo 2–7, Moderado 7–40, Muy alto ≥ 40 µg/L), construidos con funciones triangulares/trapezoidales centradas en los intervalos de la OMS [56].

En las etiquetas de validación, una muestra próxima al límite entre 7 y 40 µg/L contribuyó parcialmente a Moderado y parcialmente a Muy alto. Mientras que en la predicción, se usaron las probabilidades del clasificador como grados de pertenencia predichos. En KNN, proporción de votos normalizada y en SVM, probabilidades calibradas.

La matriz de confusión difusa se construyó acumulando, para cada par (clase-verdad, clase-predicha), la suma de productos de pertenencias $\sum \mu_{\text{verdad}}(c_i)\,\mu_{\text{pred}}(c_j)$. Por eso, las celdas mostraron valores decimales (no enteros): fueron recuentos ponderados que reflejaron la ambigüedad cerca de los umbrales.

Se eligieron KNN y SVM porque aprendieron con supuestos completamente distintos y funcionaron bien con el tamaño y dimensionalidad del set (330 muestras, 5 variables):

- KNN (no paramétrico, basado en vecindarios) modeló fronteras no lineales complejas aprovechando la estructura local de los datos. Sus probabilidades suaves derivaron de la proporción de etiquetas en el vecindario.
- SVM (margen máximo, con kernel RBF) priorizó fronteras con gran margen, lo que redujo la varianza y favoreció generalización en regiones con menor densidad de muestras. La salida probabilística provino de la calibración del puntaje, útil para la agregación difusa.

De este modo, KNN aportó flexibilidad local, y SVM aportó regularidad global, ambos generaron probabilidades que encajaron naturalmente con la evaluación difusa. Las Figuras 42 y 43 (ver en Anexos) mostraron patrones muy similares entre KNN y SVM:

- Fila Muy bajo (0–2): prácticamente sin masa en todas las columnas, lo que indicó ausencia (o casi ausencia) de soporte real para ese rango en validación.
- Fila Bajo (2–7): masa pequeña repartida sobre Moderado y Muy alto, con leve mejoría en SVM (celda Bajo 0.04 frente a 0.00 en KNN), consistente con el escaso soporte de esta clase y la proximidad al límite 7 µg/L.
- Fila Moderado (7–40): la diagonal concentró la mayor parte de la masa (9.5 en KNN, 7.1 en SVM), con confusiones apreciables en Muy alto (8.9 KNN, 11.4 SVM), reflejando la zona de transición alrededor de 40 µg/L.
- Fila Muy alto (≥ 40): la diagonal fue dominante (38.4 KNN, 38.3 SVM) y el traslado principal ocurrió hacia Moderado (7.3–7.3), lo que evidenció consistencia para detectar estados tróficos elevados y ambigüedad en casos cercanos al umbral.

La lógica difusa capturó explícitamente la incertidumbre en los bordes: en lugar de fallar por un cambio de centésimas alrededor de 40 µg/L, las observaciones contribuyeron parcialmente a ambas clases. Por eso, las diagonales conservaron la mayor parte de la masa

(buena concordancia), y las celdas adyacentes a la diagonal absorbieron la ambigüedad legítima del problema. Las diferencias finas entre KNN y SVM se encuentran entre las clases Moderado a Muy alto, en SVM fueron coherentes con su sesgo. El método de máquinas de vectores de soporte tendió a márgenes más precisos, lo que en presencia de desbalance hacia concentraciones altas, desplazó algunos casos hacia Muy alto. Se obtuvo también su aproximación lineal como se muestra en la ecuación (5) mediante el método de redes neuronales profundas.

7.7.3. Modelo de detección con datos del CEA

Primer entrenamiento

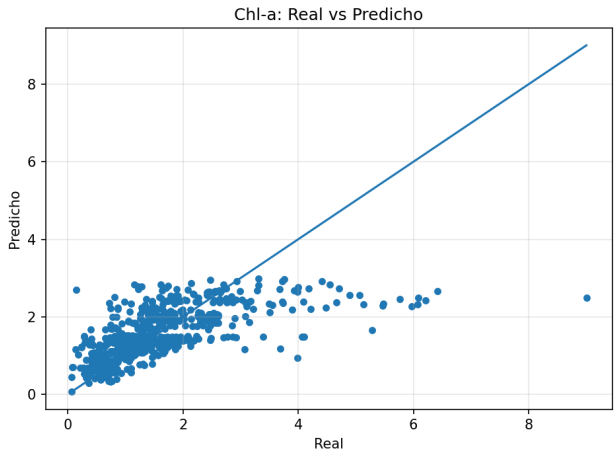
El primer entrenamiento realizado con el conjunto de datos del CEA tuvo como objetivo evaluar la capacidad de distintos modelos de clasificación (KNN, SVM y redes neuronales profundas) para discriminar el estado trófico a partir de variables fisicoquímicas como pH, temperatura, oxígeno disuelto, conductividad y turbidez. A diferencia del conjunto de AMSA, los datos provenientes del CEA presentaron una mayor dispersión en los valores de clorofila-a, así como una variabilidad más amplia en la distribución de clases, lo cual permitió analizar el comportamiento de los algoritmos bajo condiciones más heterogéneas.

El proceso consistió en entrenar modelos de regresión y clasificación y, posteriormente, obtener predicciones discretizadas según los umbrales definidos por la OMS. Las Figuras 13 y 46 (ver en Anexos) resumen el comportamiento de cada modelo en términos de ajuste, precisión y tipo de errores. La Figura 13 muestra la relación entre las concentraciones reales y las predichas por el modelo. El patrón general evidencia una tendencia creciente coherente ya que las predicciones aumentan cuando los valores reales son mayores. Sin embargo, los puntos presentan una dispersión significativa alrededor de la línea de referencia.

Este comportamiento sugiere una subestimación sistemática en valores altos. Para clorofila-a superior a 4–5 µg/L, los valores predichos se situaron mayoritariamente por debajo de la línea 1:1. Esto indica que el modelo tiende a subestimar los niveles elevados, posiblemente debido a la menor frecuencia de valores altos en el conjunto CEA, la estructura no lineal del fenómeno o la suavización introducida por la regularización de los modelos de regresión empleados.

Con mayor precisión relativa en rangos bajos y medios, la zona entre 0 y 3 µg/L, los puntos se concentraron alrededor de la diagonal, con menor variabilidad. Esto implica que el modelo captura adecuadamente las variaciones de baja magnitud, reproduce mejor las condiciones fisicoquímicas típicas del lago Atitlán bajo estados tróficos habituales, evidencia menos influencia de valores atípicos comparado con AMSA. La dispersión se incrementa conforme crecen los valores reales, lo cual es típico en procesos ecológicos donde la biomasa fitoplanctónica exhibe una mayor variabilidad en estados eutróficos [57].

Figura 13. Relación entre valores reales y predichos de clorofila-a en el primer entrenamiento con datos del CEA.



Nota. La línea diagonal representa el ajuste perfecto. Elaboración propia.

En conjunto, la Figura 13 muestra que el modelo reproduce adecuadamente las tendencias generales, pero presenta incertidumbre creciente hacia los valores altos, lo cual se refleja en las matrices de confusión. El desempeño del modelo KNN (ver Anexos, Figura 44) mostró un comportamiento estable en los extremos de la clasificación. Excelente sensibilidad en la clase Alta con 212 casos bien identificados y nula confusión con la clase Baja. En la clase Baja, hubo un elevado número de aciertos (226), con 12 casos confundidos como Moderada.

El principal reto se observó en la clase Moderada, en donde 37 casos fueron clasificados como Alta y 20 como Baja. Este patrón es coherente con la posición intermedia de la clase, que actúa como frontera entre rangos ecológicos colindantes. La distribución de los errores sugiere fronteras difusas en los parámetros limnológicos del CEA, especialmente en escenarios donde la variabilidad natural del lago genera solapamientos entre las zonas de transición entre estados mesotróficos y eutróficos.

En la Figura 45 (ver Anexos), se muestra el modelo SVM con mejor desempeño para la clase Baja, con 236 casos correctamente identificados y solo 2 clasificados como Moderada. Una clasificación robusta para la clase Alta (212 acierto) sin confusiones con Baja. La clase Moderada, en cambio, presentó el mayor grado de incertidumbre con 62 casos clasificados como Baja y 49, como Alta.

Este comportamiento es típico de modelos SVM: cuando las clases no son linealmente separables, los puntos se cubren en el espacio de atributos o existe una especie de asimetría en la densidad de muestras por clase. El margen permitió manejar la dispersión inherente, pero la separación resultó menos precisa en el rango para la clase Moderada.

La red neuronal presentó un desempeño intermedio entre KNN y SVM debido a que clasificó correctamente 216 casos de Baja y 214 de Alta, lo que indica estabilidad en los extremos. La clase Moderada obtuvo 163 aciertos, con 32 casos que descendieron a Baja y 40 que se desplazaron a Alta. Este patrón refleja una buena capacidad del modelo para

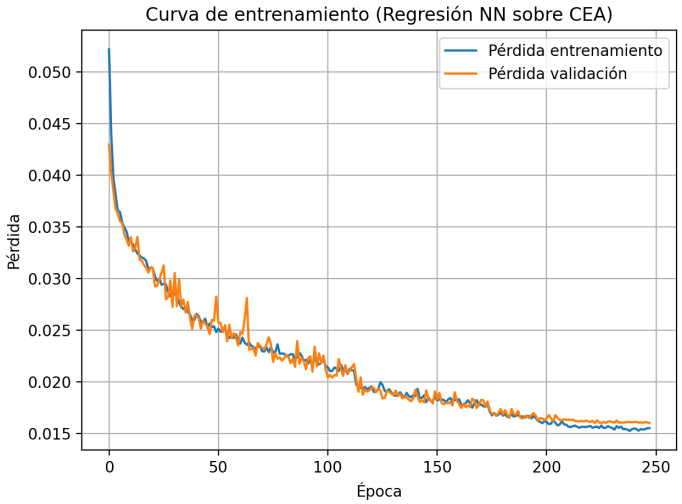
capturar relaciones no lineales, pero persistencia de los rangos limnológicos medios del lago. La red neuronal suavizó en mayor medida la frontera entre clases, lo que redujo los errores extremos, aunque mantuvo desviaciones sistemáticas en la zona de transición.

Segundo entrenamiento

Con el fin de mejorar la capacidad de los modelos para distinguir categorías, y particularmente para capturar de mejor manera las transiciones dentro del rango mesotrófico y eutrófico, se implementó un segundo entrenamiento basado en dos estrategias complementarias. La primera estrategia fue profundizar el modelado continuo mediante una red neuronal profunda de regresión, evaluando posteriormente las predicciones a través de una matriz de confusión difusa para evitar pérdidas de información al discretizar.

En la segunda estrategia, se aplicó a las matrices de confusión lógica difusa para KNN, SVM y redes neuronales profundas, con el objetivo de interpretar la pertenencia parcial de cada muestra a múltiples clases y obtener una visión más matizada del comportamiento del modelo en condiciones de frontera. Este segundo entrenamiento permitió evaluar de manera más precisa el desempeño del sistema, especialmente en escenarios donde la predicción discreta puede ocultar transiciones entre clases.

Figura 14. Curva de entrenamiento de la red neuronal de regresión con datos del CEA.



Nota. La figura muestra la evolución de la pérdida para entrenamiento y validación. Elaboración propia.

La Figura 14 muestra un comportamiento altamente estable del entrenamiento en convergencia, suave y progresivo. La pérdida inicia alrededor de 0.05 y cae rápidamente durante las primeras 20 épocas, estabilizándose posteriormente en un rango entre 0.015 y 0.020. Este descenso sostenido indica un buen ajuste al tamaño y variabilidad del conjunto CEA, así como una función de pérdida adecuada para capturar la dinámica continua de la clorofila-a y la ausencia de oscilaciones abruptas que indiquen divergencias numéricas. En el compor-

tamiento paralelo entre entrenamiento y validación, no hubo sobreajuste significativo, ya que la validación nunca diverge de manera sostenida debido a la variabilidad del lago. Este comportamiento confirma que en la aproximación continua se visualizó una ventaja frente a un enfoque puramente categórico que captura patrones más finos y reduce errores abruptos en zonas de transición.

La matriz de confusión con lógica difusa del SVM (ver en Anexos la Figura 47) aporta una interpretación más amplia que la matriz clásica ya que el dominio absoluto de la clase Muy bajo (0–2 µg/L) se obtuvo porque modelo asignó una pertenencia acumulada de 1723.41 a esta categoría, lo cual indica que la mayor parte del conjunto CEA se concentra efectivamente en este rango y que el SVM identifica con alta consistencia los patrones fisicoquímicos asociados a condiciones oligotróficas.

En las pertenencias parciales hacia Bajo, el modelo conservó una estructura coherente con 129.91, en la clase Muy bajo apenas 1.34, en la clase y prácticamente 0 en Muy alto. Con esto se obtuvo que las fronteras SVM son rígidas y fueron capaces de separar correctamente los extremos, pero permiten cierta flexibilidad para reconocer transiciones leves entre 0–7 µg/L. La ausencia total de pertenencias en la clase Muy alto es consistente con la etiqueta real del *dataset*. Además, el CEA rara vez reporta valores $\geq 40\,\mu$g/L, las condiciones del Lago Atitlán suelen mantenerse en rangos bajos y el SVM identifica esto como una estructura determinística del conjunto.

En la matriz de confusión con lógica difusa con redes neuronales profundas se muestra (ver en Anexos la Figura 48) una estructura más flexible con mayor capacidad para reconocer la clase Bajo (2–7 µg/L) y con pertenencias de 41.98 desde Muy bajo, 90.62 dentro de Bajo y 3.47 desde Moderado. En este caso la red neuronal suavizó la transición entre categorías y captó las variaciones graduales del fitoplancton. Aunque las pertenencias acumuladas son bajas, esto refleja que existe escasez de muestras en la clase Moderado del dataset, generando un reconocimiento parcial de dicha clase.

El método de KNN conserva un comportamiento intermedio entre SVM y NN ya que domina claramente la clase Muy bajo (1769.56), permitió una transición moderada hacia Bajo (83.98) y conservó valores mínimos hacia Moderado. La matriz de confusión difusa obtenida con KNN (ver en Anexos la Figura 49) refleja adecuadamente las características del lago Atitlán, donde la estabilidad fisicoquímica del agua, la homogeneidad en los gradientes y los niveles generalmente bajos de clorofila favorecen que el modelo respete las estructuras locales presentes en los datos. Se obtuvo también su aproximación lineal como se muestra en la ecuación (4) mediante el método de redes neuronales profundas.

7.7.4. Modelo de detección con datos combinados del CEA y AMSA

Primer entrenamiento

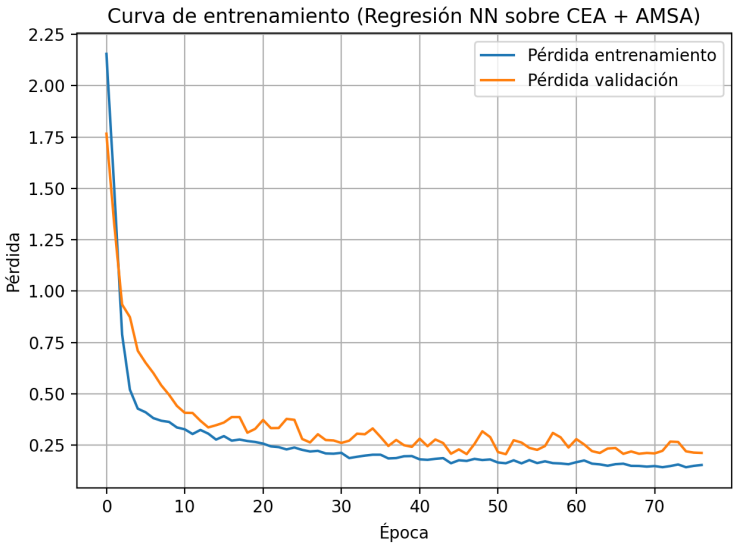
Con el fin de incrementar la cobertura de estados tróficos y mejorar la capacidad de generalización de los modelos, se integraron en un mismo conjunto los datos procedentes del CEA y de AMSA. Esta combinación permitió cubrir desde condiciones oligotróficas típicas de Atitlán hasta escenarios claramente eutróficos e hipereutróficos observados con mayor frecuencia en Amatitlán.

De este modo, el modelo de detección se entrenó y evaluó sobre un espectro más amplio de concentraciones de clorofila-a y de configuraciones fisicoquímicas. Al igual que en los casos anteriores, se implementó una red neuronal profunda de regresión para predecir clorofila-a de forma continua y, posteriormente, se construyeron matrices de confusión con lógica difusa para KNN, SVM y redes neuronales.Con el objetivo de analizar cómo se distribuye la pertenencia de cada muestra a las categorías Muy bajo (0–2 µg/L), Bajo (2–7 µg/L), Moderado (7–40 µg/L) y Muy alto ($\geq$ 40 µg/L) cuando el modelo se entrena con información proveniente de ambos lagos.

En la Figura 15 se observa que la pérdida de entrenamiento inicia con un valor cercano a 2.1 y disminuye de forma pronunciada durante las primeras 5–10 épocas, hasta aproximadamente 0.4. A partir de ese punto, la curva entra en una fase de descenso más lento y se estabiliza en torno a 0.18–0.20 hacia el final del entrenamiento. La curva de validación sigue una trayectoria paralela, pero ligeramente desplazada hacia arriba, con valores iniciales próximos a 1.8 y una estabilización alrededor de 0.22–0.25. Esta diferencia moderada entre entrenamiento y validación indica un ligero sobreajuste, esperable al aumentar la complejidad del problema al combinar dos sistemas limnológicos distintos, pero sin llegar a comprometer la capacidad de generalización del modelo.

En conjunto, la curva sugiere que la red neuronal logra capturar patrones comunes de la relación entre variables fisicoquímicas y clorofila-a en ambos lagos, aunque el error residual es mayor que en el entrenamiento exclusivo con el CEA. Esto es coherente con el hecho de que, al mezclar datos de dos cuerpos de agua con dinámicas ecológicas diferentes, se incrementa la diversidad de patrones que el modelo debe representar y, por tanto, la pérdida mínima alcanzable aumenta ligeramente.

Figura 15. Curva de entrenamiento de la red neuronal de regresión con datos combinados del CEA y AMSA.



Nota. La figura muestra la evolución de la pérdida para entrenamiento y validación. Elaboración propia.

La matriz difusa del modelo KNN (ver en Anexos la Figura 50) muestra que la diagonal principal concentra la mayor parte de la pertenencia en todas las clases, lo que indica una buena capacidad del modelo para identificar correctamente el estado trófico cuando se combinan datos de ambos lagos. En particular, se observan los siguientes patrones:

Para la clase Muy bajo, la pertenencia acumulada en la celda correcta alcanza 30.89, mientras que las contribuciones hacia Bajo, Moderado y Muy alto son muy reducidas (3.45, 0.00 y 0.50, respectivamente). Esto sugiere que el modelo es robusto al distinguir condiciones oligotróficas, incluso cuando se incorporan datos eutróficos de Amatitlán. En la clase Bajo, la diagonal acumula 23.06 unidades de pertenencia frente a valores menores hacia otras clases (4.05 hacia Muy bajo, 0.06 hacia Moderado y 0.10 hacia Muy alto). La mayor parte de los datos con lógica difusa se mantiene en la categoría correcta, pero el modelo permite un pequeño solapamiento hacia rangos inferiores, lo que refleja la continuidad natural entre los estados Muy bajo y Bajo.

La clase Moderado muestra una distribución más compleja: 6.81 de pertenencia se concentra en la celda correcta, pero 7.37 se desplazan a Muy alto, y pequeñas fracciones (1.00 y 0.53) se asignan a Muy bajo y Bajo. Este patrón revela que los casos moderados se encuentran en una zona de frontera, especialmente respecto al umbral de 40 µg/L, por lo que el KNN tiende a repartir su pertenencia entre los rangos Moderado y Muy alto cuando las condiciones fisicoquímicas se asemejan a episodios de alto proceso fotosintético.

En la clase Muy alto, la mayor parte de la pertenencia se concentra en la celda correspondiente (36.61), con un desvío principal hacia la categoría Moderado (5.14). Esta confusión hacia abajo es coherente con la definición de los umbrales y con la presencia de eventos en el borde del límite de 40 µg/L, donde pequeñas variaciones de clorofila pueden mover una muestra entre un estado eutrófico alto y uno moderado.

En conjunto, el modelo KNN conserva una estructura de clasificación coherente y relativamente nítida, especialmente en los extremos Muy bajo y Muy alto, mientras que la clase Moderado se comporta como una zona de transición, compartiendo pertenencia con los rangos laterales.

La matriz de la red neuronal (ver en Anexos la Figura 51) presenta un comportamiento similar al de KNN, pero con transiciones aún más suaves:

La clase Muy bajo conserva la mayor parte de su pertenencia en la diagonal (29.18), aunque se observa un incremento de la masa difusa hacia Bajo (5.41) en comparación con KNN. Esto indica que la red tiende a reconocer gradientes más finos entre ambos rangos, lo que es consistente con su capacidad para aproximar relaciones no lineales. En la clase Bajo, la pertenencia dominante (20.93) se mantiene sobre la diagonal, con desviaciones hacia Muy bajo (6.19) y prácticamente nulas hacia Moderado. Este patrón refleja una transición dominante hacia valores inferiores, probablemente motivada por la influencia de los registros de Atitlán, donde prevalecen concentraciones más bajas.

La clase Moderado mostró una estructura similar a la observada con KNN: 6.56 de pertenencia correcta y 8.11 asignada a Muy alto. De nuevo, esto confirma que las muestras moderadas, especialmente en la franja superior del intervalo de 7–40 µg/L, presentan características limnológicas similares a eventos eutróficos, de modo que la red reparte la pertenencia entre ambas categorías.

En la clase Muy alto, la red neuronal consigue la asignación más limpia: 39.14 unidades de pertenencia en la diagonal y 3.04 hacia Moderado, sin prácticamente contribuciones hacia rangos bajos. Esto es consistente con el hecho de que las condiciones de alta clorofila asociadas a Amatitlán generan firmas fisicoquímicas claramente diferenciables del resto de los estados tróficos.

En síntesis, la red neuronal reprodujo un patrón de clasificación muy similar al de KNN, pero con transiciones más diferenciadas entre Muy bajo, Bajo y una separación particularmente clara del estado Muy alto. Por lo tanto, con el conjunto combinado, la red capta adecuadamente los extremos del gradiente trófico y matiza las zonas de frontera mediante asignaciones parciales de pertenencia.

La matriz SVM (ver en Anexos la Figura 52) muestra un comportamiento intermedio, con fronteras algo más rígidas que las de la red neuronal y KNN, pero sin perder la capacidad de generalización. En Muy bajo, 25.10 unidades se mantienen en la diagonal, con desviaciones hacia Bajo (8.86) y contribuciones menores a Moderado y Muy alto. El modelo tiende a separar bien el estado oligotrófico, aunque permite cierta permeabilidad hacia el rango 2–7 µg/L, lo que se interpreta como reconocimiento de gradientes suaves en la base del sistema.

En Bajo, la diagonal acumula 20.07, con 6.69 hacia Muy bajo y aportes mínimos a Moderado y Muy alto. De nuevo se observó una simetría respecto al comportamiento de la clase Muy bajo, con una ligera tendencia a clasificar hacia el extremo inferior, pero manteniendo coherencia con la definición de los umbrales. La clase Moderado presentó 5.96 de pertenencia correcta y 8.19 en Muy alto, además de pequeñas fracciones en los rangos inferiores.

Este patrón, similar al de KNN y NN, confirmó la dificultad sistemática para separar con nitidez el rango intermedio, especialmente cuando se integran datos de Amatitlán, donde las condiciones altamente productivas pueden acercar los perfiles fisicoquímicos de muestras moderadas a aquellos claramente eutróficos. En Muy alto, la mayor parte de la pertenencia se mantuvo en la diagonal (36.25), con 5.27 unidades asignadas a Moderado. Aunque la confusión es algo mayor que en la red neuronal, la estructura global siguió siendo consistente y adecuada para propósitos de alerta y clasificación de riesgo. Se obtuvo también su aproximación lineal como se muestra en la ecuación (3) mediante el método de redes neuronales profundas.

CAPÍTULO 8

Diseño e integración del sistema de sensores

El diseño del sistema de sensores se rigió por criterios de seguridad eléctrica en campo, repetibilidad geométrica en la inmersión de las sondas, y portabilidad e higiene en el manejo de muestras. En todas las iteraciones se procuró:

1. Separar físicamente la electrónica de potencia/señal del entorno húmedo.
2. Garantizar una geometría de guía para insertar las sondas a una profundidad constante y con alineamiento vertical.
3. Minimizar contaminaciones cruzadas entre vasos.
4. Facilitar montaje, limpieza y transporte durante la toma de datos.

La evolución del diseño se documentó en tres etapas (prototipo inicial, boceto CAD y diseño definitivo).

8.1. Primer diseño

La Figura 16 corresponde al primer prototipo, utilizado durante la primera prueba piloto en el estanque del jardín botánico UVG. Se implementó un montaje rápido con materiales de disponibilidad inmediata: caja de cartón desarmada a modo de plataforma y barrera mecánica entre la banca y los componentes electrónicos, vasos desechables limpios para contener las muestras, agua pura para enjuagues entre mediciones, papel mayordomo para secado y cintas de aislar y bolsas de basura para fijación y control de salpicaduras.

Este arreglo proveyó una zona seca para el computador y los módulos electrónicos, mantuvo los cables en superficie y permitió alinear las sondas sobre recipientes independientes.

Si bien su naturaleza era provisional, el montaje resolvió con efectividad la humedad del entorno, habilitó secuencias de enjuague y ofreció accesibilidad para reconexión de sensores y verificación de señales en sitio. La experiencia de uso evidenció la necesidad de incorporar geometrías guía que limitaran el bamboleo de las sondas y aseguraran profundidad constante, además de una estructura más rígida y modular para repetibilidad.

Figura 16. Primer diseño para el sistema de sensores.

Nota. Prototipo inicial para la medición y obtención de muestras de agua del estanque del jardín botánico. Elaboración propia.

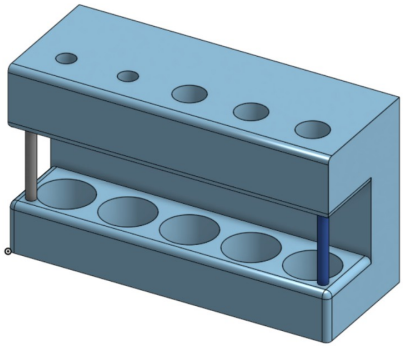
8.2. Segundo diseño

Con base en las lecciones del prototipo 1, se desarrolló un boceto paramétrico en *Onshape* (Figura 17). El concepto adoptó una arquitectura en dos planos: una pieza superior con perforaciones guía para las sondas y una pieza inferior con alojamientos circulares para recipientes, ambas unidas por soportes cilíndricos.

Esta configuración buscó un control geométrico de la inmersión (eje y profundidad homogéneos entre sensores), un trazado claro de cableado y alivio de tensión en la pieza superior, y un modulador para el intercambio de recipientes sin intervenir la parte eléctrica.

Durante la fase de estimación de fabricación se determinó que la versión monolítica inicial implicaba tiempos de impresión prolongados. Por ello, el boceto se utilizó como antecesor conceptual y se revisó la estrategia de manufactura buscando segmentación de piezas, radios de filete para resistencia y espesores compatibles con impresiones extensas sin deformación.

Figura 17. Segundo diseño para el sistema de sensores.



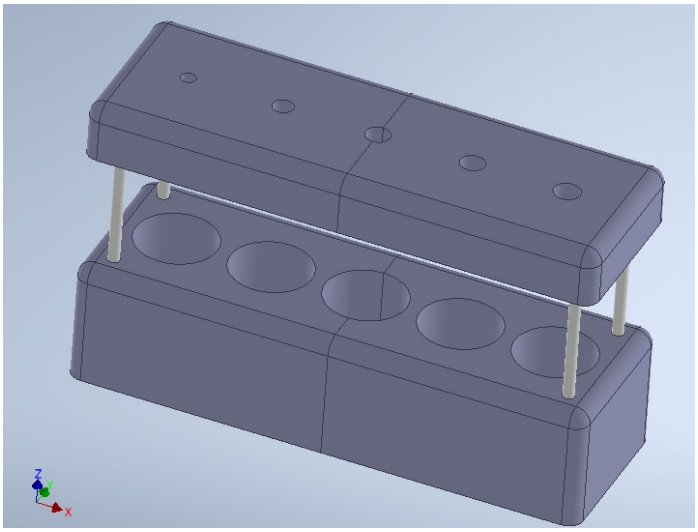
Nota. Uso de CAD mediante la herramienta para el diseño 3D en *Onshape*, este fue el primer boceto. Elaboración propia.

8.3. Tercer diseño

El diseño definitivo (Figura 18) conservó la lógica de doble nivel y se materializó en módulos acoplables entre sí. La pieza superior incorporó orificios guía dimensionados específicamente para cada tipo de sonda (pH, conductividad, oxígeno disuelto, turbidez y temperatura), con holguras funcionales que permitieron una inserción suave y redujeron el juego lateral para mantener el eje vertical de medición. La bandeja inferior alojó cavidades circulares destinadas a recipientes individuales, lo que mitigó contaminaciones cruzadas y facilitó protocolos de enjuague entre mediciones. Ambas partes se unieron mediante cuatro soportes cilíndricos, que aseguraron coplanari-dad y rigidez del conjunto, además de permitir el desarme para limpieza y transporte. El diseño incluyó chaflanes y filetes para evitar concentraciones de tensión y mejorar el acabado superficial en las interfaces mano-pieza. El conjunto se fabricó por impresión 3D con filamento PLA de 1.75 mm, color blanco nieve. El tiempo total de impresión fue de 52 horas (incluyendo los soportes cilíndricos), distribuido en las dos piezas principales y elementos de unión.

El uso de PLA respondió a su buena rigidez específica, estabilidad dimensional y acabado adecuados para geometrías guía, resultando conveniente para un dispositivo de laboratorio o campo con exposición intermitente al ambiente. La combinación de espesores de pared y porcentajes de relleno se eligió para equilibrar peso, tiempo de fabricación y resistencia de las columnas. El montaje final se utilizó en la segunda prueba piloto de mediciones en el estanque. En operación, la pieza superior actuó como plantilla, asegurando que cada sonda alcanzara una profundidad comparable en su respectivo recipiente. La pieza inferior confinó los vasos, evitando desplazamientos por vibración o manipulación. Y la separación vertical entre planos brindó un corredor limpio para cableado, reduciendo tracción en conectores y minimizando artefactos por movimiento de cable. El conjunto permitió repetibilidad posicional, lavado y secado rápidos de superficies. Las piezas que componen este diseño se encuentra en la sección de Anexos: Parte inferior izquierda (Figura 33), parte inferior derecha (Figura 34), parte superior izquierda (Figura 35) y parte superior derecha (Figura 36).

Figura 18. Tercer diseño para el sistema de sensores.



Nota. Uso de CAD mediante la herramienta para el diseño 3D en *Inventor*, este fue el segundo boceto. Elaboración propia.

8.4. Plataforma de control y adquisición

La integración de sensores se implementó utilizando una placa Arduino Mega 2560, seleccionada por su capacidad ampliada de pines digitales y analógicos, así como por su mayor memoria de programa en comparación con modelos básicos como el Arduino Uno. Esta elección permitió conectar simultáneamente múltiples sensores, asegurar tasas de muestreo consistentes y disponer de un margen suficiente para el manejo de bibliotecas de software específicas y rutinas de calibración.

El sistema se diseñó para la adquisición de los parámetros fisicoquímicos considerados relevantes para la estimación indirecta de la concentración de clorofila-a y la proliferación de cianobacterias. Cada sensor requirió condiciones de alimentación, calibración y compensación diferentes, por lo que la integración requirió un análisis cuidadoso de compatibilidad y manejo de bibliotecas.

8.5. Estructura del código y manejo de sensores

El proceso de integración se documentó en dos rutinas principales en lenguaje Arduino (C++):

- Archivo sensores.ino [58]: incluyó la lógica para la adquisición de los valores de pH, oxígeno disuelto, turbidez y temperatura. Todos estos sensores pudieron trabajar de forma conjunta sin generar interferencias significativas en las líneas analógicas ni en la transmisión de datos.

- Archivo conduc_temp.ino [59]: se implementó de manera separada debido a que, al integrarse el sensor de conductividad en el mismo código con los demás dispositivos, se producían errores de lectura y, en ocasiones se obtenían valores nulos o cercanos a cero. Para garantizar la confiabilidad de las mediciones, se decidió aislar este sensor en un programa específico, manteniendo siempre la referencia de temperatura como parámetro de compensación.

Los programas se encuentran disponibles en el enlace del repositorio correspondiente (Anexo 13.2). Ambos códigos compartieron estructuras comunes como la selección de pines analógicos y digitales así como una configuración de la comunicación serial para la visualización de datos en tiempo real junto con una implementación de funciones de lectura, filtrado y conversión de valores crudos a unidades físicas (°C, mg/L, µS/cm, NTU, etc.). Además, mostraron una compensación por temperatura, indispensable para la correcta interpretación de los datos de pH, oxígeno disuelto y conductividad.

8.6. Consideraciones sobre la integración

La necesidad de utilizar dos códigos separados reflejó una limitación propia de la integración directa de sensores heterogéneos en una misma placa, especialmente cuando los módulos empleaban bibliotecas diferentes o compartían rutinas de temporización que podían interferir entre sí. Esta estrategia aseguró lo siguiente:

- Estabilidad en la adquisición de datos, evitando lecturas erráticas en el sensor de conductividad.
- Consistencia en el uso de la temperatura como variable de referencia transversal.
- Escalabilidad del sistema, permitiendo que el desarrollo futuro contemple la integración de los códigos en una sola rutina optimizada, o el uso de microcontroladores con buses de comunicación más robustos.

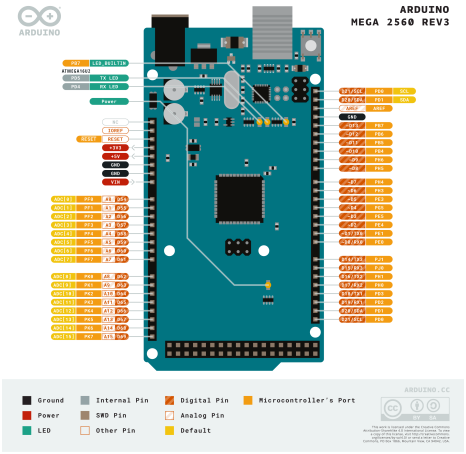
8.7. Arduino Mega

El Arduino Mega 2560 fue la plataforma seleccionada para la integración del sistema de sensores debido a sus características técnicas superiores en comparación con otros modelos de la familia Arduino. Esta placa (Figura 19) incorporó un microcontrolador ATmega2560 con 54 pines digitales de entrada/salida (15 de ellos con salida PWM), 16 entradas analógicas, 4 puertos UART, un reloj a 16 MHz y 256 KB de memoria de programa.

Estas especificaciones permitieron conectar múltiples sensores sin limitaciones de pines, manejar diferentes bibliotecas de forma paralela y garantizar una ejecución estable durante períodos prolongados de adquisición de datos. En este proyecto, el Arduino Mega se utilizó como unidad central de adquisición, cuyas funciones fueron las siguientes:

- Leer señales analógicas y digitales provenientes de los sensores de pH, temperatura, oxígeno disuelto, conductividad y turbidez.
- Sincronizar el muestreo asegurando consistencia temporal entre parámetros.
- Transmitir los datos vía comunicación serial al computador para su almacenamiento, visualización y posterior procesamiento con los modelos de aprendizaje automático.

Figura 19. Microcontrolador Arduino Mega 2560.



Nota. En la imagen se muestran las entradas y salidas del microcontrolador. Adaptada de Arduino [60].

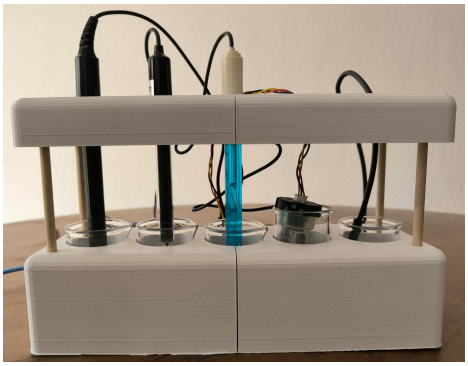
8.8. Integración final del sistema de sensores

La Figura 20 presenta la integración física final del sistema multisensorial utilizado para la adquisición simultánea de parámetros fisicoquímicos del agua. El diseño consistió en una estructura impresa en 3D que permitió alojar y posicionar de manera estable cada uno de los sensores: pH, oxígeno disuelto (OD), turbidez, temperatura y conductividad eléctrica. Esta configuración aseguró que las puntas de medición permanecieran a una profundidad constante dentro de las cámaras de muestreo, minimizando las variaciones asociadas al movimiento, burbujas o diferencias en el volumen de agua. La distribución espacial también se planificó para evitar interferencias entre sensores, especialmente en aquellos con componentes ópticos o electrónicos sensibles.

La carcasa superior cumplió una doble función: proporcionar soporte mecánico y organizar los cables de alimentación y señal, reduciendo las tensiones sobre los conectores y mejorando la estabilidad del montaje durante las pruebas. Asimismo, las cámaras transparentes facilitaron la observación directa del contacto entre los electrodos y la muestra, permitiendo verificar su correcta inmersión durante los procesos de calibración y adquisición. Este ensamblaje final permitió ejecutar mediciones controladas y garantizar condiciones reproducibles durante la captura de datos utilizados posteriormente en la fase de modelado mediante aprendizaje automático.

Su diseño modular facilita la futura incorporación de nuevos sensores o modificaciones en la geometría interna para adaptarse a distintos volúmenes de muestra o requerimientos experimentales.

Figura 20. Sistema de sensores completo para la adquisición de parámetros fisicoquímicos.

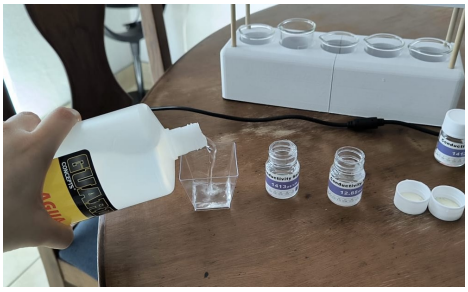


Nota. Estructura impresa en 3D que permite sostener simultáneamente los sensores de pH, temperatura, oxígeno disuelto, conductividad y turbidez en posiciones fijas y uniformes. Elaboración propia.

8.9. Calibración de sensores

Dado que se utilizaron los sensores correspondientes a la boya descrita por Cano [5], todos los dispositivos fueron limpiados superficialmente con agua desmineralizada (Figura 21) antes de cada calibración. Este procedimiento permitió eliminar residuos de pruebas anteriores y asegurar una mayor precisión en las mediciones, especialmente en aquellos sensores que emplean soluciones de referencia específicas.

Figura 21. Limpieza previa a la calibración de los sensores.



Nota. Uso de agua desmineralizada para la limpieza de recipientes y sensores. Elaboración propia.

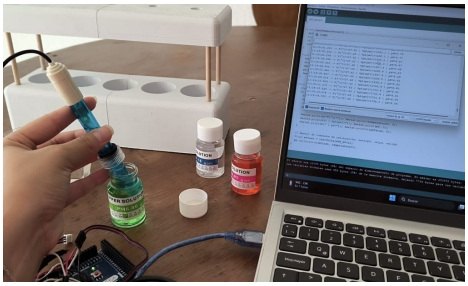
8.9.1. Sensor analógico de pH

Para la primera calibración del sensor de pH se utilizaron tres recipientes de 35 mL, cada uno con sustancias de distinto nivel de acidez y alcalinidad. El primero contenía 30 ml de jugo de limón (pH ácido entre 2 y 3), el segundo 30 mL de agua pura (pH cercano a la neutralidad entre 6 y 7) y el tercero una mezcla de 10 mL de agua con 20 g de soda cáustica (pH alcalino entre 13 y 14).

Dado que estas sustancias eran experimentales, se implementó en el código una regresión que permitiera aproximar el pH en función de los valores analógicos obtenidos por el sensor. Esta aproximación buscó establecer una relación inicial útil para las mediciones.

Con el fin de validar dicha aproximación, se realizó una segunda calibración utilizando soluciones patrón de pH 4, 6 y 9. Los valores registrados coincidieron satisfactoriamente con los valores de referencia, lo cual confirmó la precisión del proceso de calibración (Figura 22). Esta segunda fase permitió estandarizar la respuesta del sensor bajo condiciones más controladas.

Figura 22. Calibración del sensor de pH con soluciones patrón.



Nota. Segunda calibración con soluciones de pH 4, 6 y 9. Elaboración propia.

8.9.2. Sensor analógico de temperatura

El sensor de temperatura (Figura 23) mostró un comportamiento estable y preciso en ambas calibraciones realizadas. Debido a su versatilidad, el dispositivo registró de manera confiable la temperatura del agua tanto en condiciones frías como moderadamente calientes. No se observaron errores significativos ni desviaciones durante el proceso, lo cual lo posicionó como uno de los sensores con respuesta más consistente dentro del sistema.

Figura 23. Calibración del sensor de temperatura.

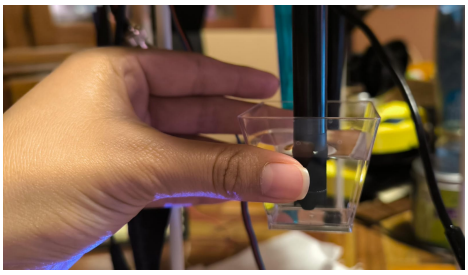
Nota. Calibración del sensor de temperatura con agua a temperatura ambiente. Elaboración propia.

8.9.3. Sensor analógico de oxígeno disuelto

La calibración del sensor de oxígeno disuelto (Figura 24) requirió una preparación química específica. Para activar el sensor, se mezclaron pequeños trozos de soda cáustica con 10 mL de agua. Seguidamente, se retiró la tapa del sensor y se aplicaron diez gotas de esta solución en la membrana interna utilizando un gotero. Posteriormente, la tapa fue reinstalada y el sensor se sumergió en un recipiente con agua.

La verificación del funcionamiento se realizó generando burbujas dentro del recipiente mediante el gotero. En el monitor serial de Arduino se observaron variaciones claras en la señal del sensor tanto en presencia como en ausencia de burbujas, lo que confirmó su correcta respuesta. Dado que la mezcla de soda cáustica tiende a secarse con el tiempo, se aplicó en dos ocasiones durante el proceso de calibración. Asimismo, por motivos de seguridad, el manejo de este compuesto se realizó utilizando guantes.

Figura 24. Calibración del sensor de oxígeno disuelto.



Nota. Calibración del sensor de oxígeno disuelto con mezcla de soda cáustica. Elaboración propia.

8.9.4. Sensor analógico de conductividad

La calibración del sensor de conductividad (Figura 25) se realizó utilizando el código proporcionado por el fabricante [59], lo cual permitió agilizar el proceso. Dicho código permite

ingresar directamente en el monitor serial la conductividad de la solución patrón empleada. Al sumergir el sensor en soluciones de referencia de 1413 µS/cm y 12.88 mS/cm, las mediciones coincidieron adecuadamente con los valores esperados.

El sensor mostró un desempeño estable durante las calibraciones y en las mediciones posteriores. Debido a su sensibilidad, se evitó mantenerlo sumergido por periodos prolongados, reduciendo así el riesgo de deterioro prematuro.

Figura 25. Calibración del sensor de conductividad.

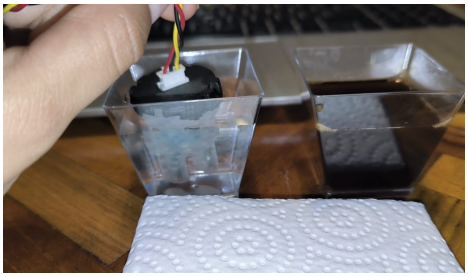
Nota. Segunda calibración del sensor de conductividad. Elaboración propia.

8.9.5. Sensor analógico de turbidez

Para calibrar el sensor de turbidez (Figura 26) se utilizaron tres recipientes de 35 mL con diferentes niveles de concentración de partículas. El primero contenía únicamente agua pura, el segundo, una mezcla de 5 g de café molido en 30 mL de agua, y el tercero, una mezcla más concentrada con 20 g de café en 10 mL de agua. Aunque las diferencias visuales entre los recipientes eran claras, el sensor no registró un valor intermedio en la segunda mezcla, como sería teóricamente esperado.

Ante esta limitación, fue necesario implementar en el código de Arduino [58] una ecuación de ajuste que permitiera aproximar un valor de NTU a partir de la señal analógica del sensor. Este ajuste permitió establecer un criterio de clasificación más coherente con las observaciones cualitativas realizadas durante la calibración.

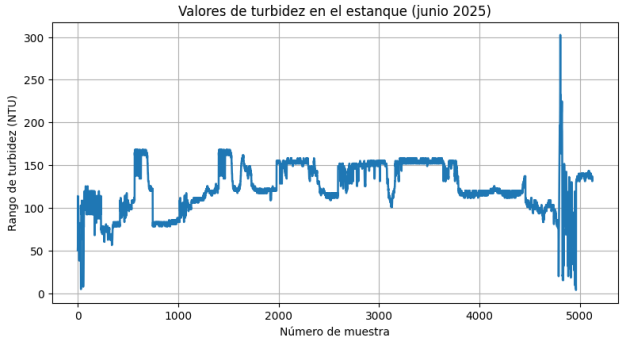
Figura 26. Calibración del sensor de turbidez.



Nota. Primera calibración del sensor de turbidez. Elaboración propia.

Para validar la calibración de dichos sensores, en la Figura 27 se muestra la evolución de los valores de turbidez registrados en el estanque durante la primera prueba piloto, realizada para verificar el funcionamiento y la precisión del sensor de turbidez. Las series temporales correspondientes al resto de parámetros fisicoquímicos se presentan en los Anexos.

Figura 27. Comportamiento de los datos de turbidez.



Nota. Datos de turbidez obtenidos en la primera prueba piloto. Elaboración propia.

CAPÍTULO 9

Resultados de las predicciones de clorofila-a

El modelo de redes neuronales profundas que se desarrolló integra datos experimentales del lago Amatitlán, del lago de Atitlán y del estanque controlado para predecir concentraciones de clorofila-a como *proxy* de cianobacterias. Se llevaron a cabo otros algoritmos de clasificación como SVM y KNN, con un ochenta por ciento de los datos y evaluar en el veinte por ciento restante. Los resultados se interpretaron a través de métodos de razonamiento y evaluación de rendimiento como las matrices de confusión y lógica difusa. El sistema permite identificar, según la categoría de clorofila-a, las condiciones del estado trófico en cuerpos de agua, proporcionando información importante para entidades como AMSA, Ibagua y CEA.

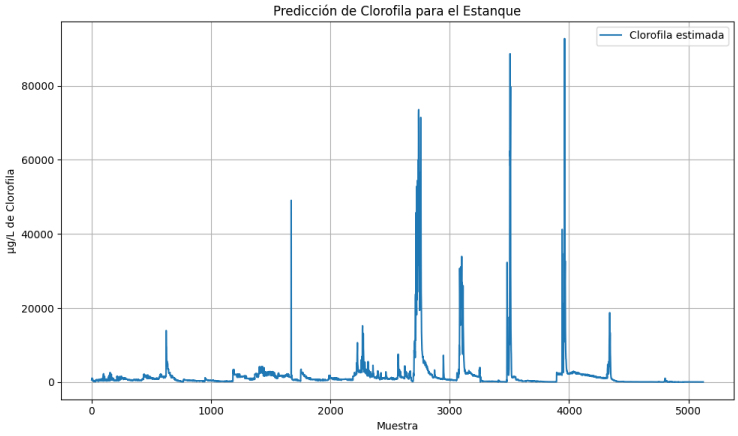
9.1. Primera predicción usando datos de Ibagua aplicados a los datos recolectados del estanque

La primera predicción de clorofila se realizó utilizando el modelo entrenado con el conjunto de datos de Ibagua. Este modelo fue aplicado posteriormente a las mediciones recolectadas en el estanque, que comprendían más de 5000 datos de parámetros fisicoquímicos. El resultado se representó en la gráfica de la Figura 28 de predicción, donde se observó la evolución de la clorofila estimada a lo largo de todas las muestras.

En la gráfica (Figura 28) se evidenció que el modelo arrojó valores de clorofila extremadamente altos para un entorno controlado como lo es un estanque.

En varios puntos, las concentraciones predichas superaron los 60 000 y hasta 80 000 µg/L, cifras que exceden considerablemente los rangos esperados para cuerpos de agua con condiciones de manejo artificial. Este comportamiento reflejó que el modelo, aunque había sido entrenado con los datos históricos de Ibagua, no logró trasladar de manera adecuada el aprendizaje a un contexto con características diferentes.

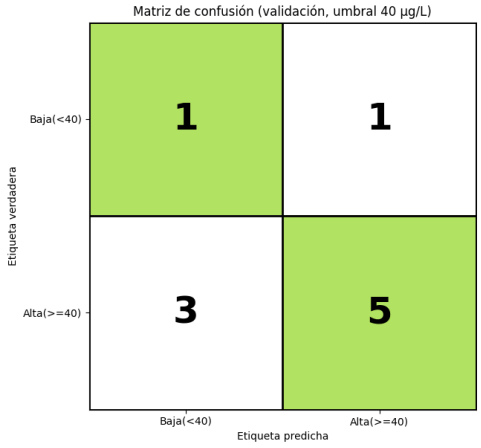
Figura 28. Predicción de clorofila en el estanque del jardín botánico UVG.



Nota. La imagen muestra los valores predichos de clorofila usando el modelo de los datos de Ibagua. Elaboración propia.

Es importante señalar que el estanque en cuestión recibió un mantenimiento particular, ya que se le añadió agua fresca semanalmente. Este proceso de recarga favoreció la oxigenación del sistema, lo cual contribuyó a mantener un equilibrio más estable en las variables fisicoquímicas, y consecuentemente, en la dinámica real de la clorofila. El contraste entre las predicciones del modelo y las condiciones observadas en el estanque puso en evidencia una discrepancia entre los patrones de entrenamiento y la realidad experimental del entorno controlado.

Figura 29. Matriz de confusión con los datos de validación del modelo de Ibagua.



Nota. La imagen muestra la diagonal de valores clasificados correctamente a pesar de que no haya sido el resultado esperado. Elaboración propia.

Con el objetivo de evaluar el desempeño del modelo en términos de clasificación, se aplicó una matriz de confusión (Figura 29) considerando únicamente los datos del conjunto de validación.

Para ello, se estableció un umbral de 40 $\mu g/L$, valor que permitió diferenciar entre presencia Baja ($< 40\,\mu g/L$) y presencia Alta ($\geq 40\,\mu g/L$) de clorofila. Los resultados mostraron que el modelo clasificó correctamente 6 de las 10 muestras de validación, alcanzando una precisión limitada. Dentro de estas predicciones, únicamente una muestra correspondió a la categoría de Baja concentración de clorofila, mientras que la mayoría de las predicciones se concentraron en la categoría de Alta concentración.

Este comportamiento se explicó principalmente por el desbalance en la base de datos de entrenamiento. Los datos de Ibagua utilizados en el modelo presentaban una predominancia de valores elevados de clorofila, lo cual condujo a que la red neuronal aprendiera con mayor fuerza patrones asociados a escenarios de alta concentración. Como consecuencia, la capacidad del modelo para reconocer escenarios con bajas concentraciones resultó limitada.

Además, la comparación entre el tamaño del conjunto de entrenamiento y la magnitud de los datos provenientes del estanque mostró una desproporción significativa. Mientras que el modelo se entrenó con apenas 50 muestras de Ibagua, la serie temporal del estanque incluyó más de 5000 mediciones. Esta diferencia implicó que el modelo resultara insuficiente para capturar toda la variabilidad y complejidad presente en el nuevo escenario. Por ello, el sistema de predicción se quedó corto al ser aplicado a un conjunto de datos mucho más amplio, revelando la necesidad de contar con bases de datos de entrenamiento más extensas y representativas.

9.2. Segunda predicción usando datos de AMSA aplicados a los datos de Ibagua (sin clorofila-a)

La segunda predicción de clorofila se efectuó mediante una interfaz pública (Figura 30) desarrollada en *Streamlit* [61] e integrada con *GitHub* [62], la cual funcionó como un tablero digital interactivo para la ejecución y visualización de modelos de aprendizaje automático.

En esta plataforma se habilitó la carga de archivos con los parámetros fisicoquímicos de entrada pH, temperatura, oxígeno disuelto, conductividad y turbidez, con el propósito de estimar clorofila-a y, posteriormente, clasificar el estado trófico con base en el valor predicho. Para el entrenamiento de los modelos empleados en la interfaz se utilizaron los datos provistos por AMSA, los cuales fueron limpiados, seleccionados y depurados.

Tras este proceso, se contó con 330 muestras por cada variable, incluyendo la salida (clorofila-a). Con el fin de evaluar la capacidad de generalización del modelo entrenado con AMSA, se aplicó el modelo de redes neuronales profundas al conjunto de datos de Ibagua en el que se omitió intencionalmente la columna de clorofila.

Este procedimiento permitió predecir clorofila a partir de las cinco variables de entrada y, luego, contrastar los resultados con los valores reales de Ibagua disponibles de forma externa al archivo de predicción, evitando cualquier posibilidad de fuga de información durante el proceso. El resultado inicial de esta aplicación se presentó como un histograma (Figura 31) de la distribución de clorofila predicha, donde se observó la concentración de valores dentro de un rango acotado y plausible para aguas continentales eutróficas e hipereutróficas.

Figura 30. Interfaz inicial para la visualización de predicciones.

Nota. En la imagen, se muestra la última sección del tablero digital interactivo en donde está la opción de subir archivos para detectar clorofila. Está fue la primera iteración del tablero. Elaboración propia.

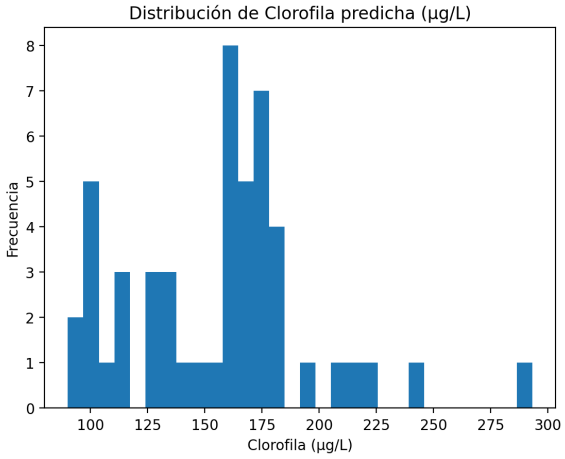
La distribución de clorofila estimada mostró un comportamiento de una sola moda con concentración de datos en el intervalo aproximado 140–180 $\mu g/L$, acompañado de una cola derecha que alcanzó valores en torno a 200–300 $\mu g/L$.

A diferencia del primer intento de predicción (el cual había arrojado magnitudes no realistas en el orden de decenas de miles de $\mu g/L$ cuando se aplicó un modelo no adecuado al dominio), en esta segunda aproximación la escala de las predicciones se mantuvo consistente con lo esperado para cuerpos de agua con elevada productividad, sin la presencia de picos desproporcionados. De forma complementaria, la interfaz permitió exportar una tabla de resultados (Cuadro 4) con las predicciones puntuales de clorofila para cada muestra de Ibagua. Esta tabla se contrastó con la tabla original de Ibagua mostrada en el 5), lo que habilitó una verificación directa caso por caso.

En dicha comparación, se constató que los órdenes de magnitud de las predicciones fueron congruentes con los valores de referencia y que las variaciones locales (incrementos o descensos relativos entre muestras consecutivas) se reprodujeron en la mayoría de los tramos, lo cual indicó que el modelo respondió coherentemente a la dinámica inducida por las variables de entrada. En términos de interpretación ecológica, la concentración de los datos de probabilidad en el rango 140–180 $\mu g/L$ con colas hacia valores superiores implicó que, bajo los umbrales utilizados en este trabajo (clasificación por 40 $\mu g/L$ para Baja vs. Alta presencia), la mayor parte de las muestras habría sido clasificada como Alta.

Esto concordó con la naturaleza histórica de los datos de AMSA empleados para el entrenamiento, en los que existía una proporción relevante de episodios de alta clorofila, de modo que el modelo aprendió relaciones que priorizaron escenarios de productividad elevada.

Figura 31. Histograma de la predicción de clorofila aplicado a datos de Ibagua.



Nota. La imagen muestra la frecuencia de los valores de clorofila en un intervalo de 0 a 300 $\mu g/L$. Elaboración propia.

En la predicción sobre Ibagua, esta propensión se tradujo en estimaciones consistentemente altas, aunque dentro de rangos físicamente verosímiles, lo que diferenció este resultado del primer experimento.

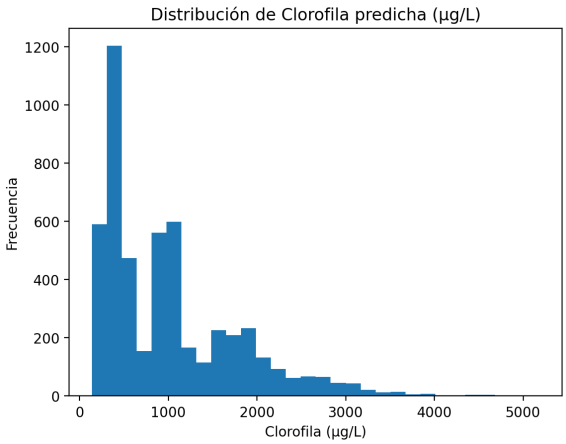
9.3. Tercera predicción usando datos de AMSA aplicados a los datos recolectados del estanque

En el tercer entrenamiento se aplicó el modelo de redes neuronales profundas entrenado con AMSA al conjunto de más de cinco mil mediciones obtenidas en el estanque experimental. El *pipeline* de inferencia replicó exactamente el ajuste de normalización de entradas con los estadísticos del entrenamiento de AMSA, transformación logarítmica sobre la variable objetivo durante el aprendizaje y la transformación inversa para reportar las predicciones finales de clorofila-a en $\mu g/L$. La Figura 32 mostró el histograma de dichas predicciones.

La distribución resultante presentó una asimetría positiva marcada (cola derecha), con una alta concentración de frecuencias en el entorno de centenas de $\mu g/L$ y una disminución progresiva hacia valores del orden de los miles de $\mu g/L$. Visualmente, el intervalo de datos con mayor frecuencia se ubicó en el rango 300–700 $\mu g/L$, se observó un bloque secundario de densidad entre 800–1 200 $\mu g/L$, y una cola larga que extendió el conjunto de datos hasta 5 000 $\mu g/L$, aunque con muy baja frecuencia en los extremos. Este patrón implicó que, según el modelo, la mayor parte del período monitorizado en el estanque habría correspondido a concentraciones altas de clorofila, con episodios esporádicos de valores muy elevados.

Siguiendo los umbrales operativos utilizados en el estudio (Muy bajo 0–2, Bajo 2–7, Moderado 7–40, Muy alto $\geq 40\,\mu g/L$), la distribución estimada situó prácticamente la totalidad de las observaciones por encima de 40 $\mu g/L$, es decir, dentro del estado hipereutrófico.

Figura 32. Histograma de la predicción de clorofila-a en la primera prueba piloto en el estanque.



Nota. La imagen muestra la frecuencia de los valores de clorofila-a en un intervalo de 0 a 300 µg/L mediante el uso de un modelo de redes neuronales profundas entrenado con datos de AMSA. Elaboración propia.

La presencia de una cola derecha amplia indicó la ocurrencia de picos (pocas muestras con valores extremadamente altos), mientras que el cuerpo de la distribución se concentró en niveles persistentemente elevados. En términos operativos, el modelo describió un estanque dominantemente productivo, que se manifestó en la heterogeneidad de la parte media de la distribución y en la aparición de eventos de alta pigmentación (cola).

El resultado fue coherente con la estructura aprendida en el entrenamiento de AMSA del capítulo 7 donde la jerarquía de variables encabezada por turbidez y conductividad (identificada con *Random Forest* en el primer entrenamiento) favoreció predicciones altas cuando ambos indicadores se situaron en rangos elevados. Incluso, la escasez de ejemplos de concentraciones Muy Bajo y Bajo en los datos de entrenamiento condicionó al conjunto de datos predictivos hacia los rangos moderados y altos. En consecuencia, al proyectar el modelo sobre el estanque, la combinación de patrones fisicoquímicos propios de ese sistema y el sesgo de soporte del entrenamiento, se tradujo en una distribución desplazada hacia valores altos.

9.4. Cuarta predicción usando datos de AMSA aplicados a las pruebas piloto en el estanque

Con el propósito de evaluar el comportamiento de los modelos en un entorno experimental controlado, se realizó una cuarta fase de predicción en la que se aplicaron los modelos entrenados con AMSA a los datos fisicoquímicos medidos en el estanque. A diferencia de la tercera predicción donde se utilizaron todas las mediciones disponibles para construir un histograma global, en esta etapa se trabajó con subconjuntos específicos asociados a dos pruebas piloto.

De la primera prueba piloto se tomaron 120 registros obtenidos al seleccionar de forma aleatoria seis mediciones por día durante veinte días de monitoreo continuo. Y de la segunda prueba piloto se tomaron 416 registros recolectados en una segunda prueba piloto, con mayor resolución temporal y espacial dentro del estanque.

En ambos casos, las variables de entrada fueron pH, temperatura, oxígeno disuelto, conductividad y turbidez, procesadas con el mismo *pipeline* de normalización y transformación utilizado en el entrenamiento con AMSA. A partir de estas entradas, los modelos predictivos estimaron la concentración de clorofila-a y, posteriormente, se construyeron matrices de confusión con lógica difusa para los clasificadores SVM, KNN y redes neuronales profundas. Las matrices representan el grado de pertenencia acumulado de cada muestra a las categorías Muy bajo (0–2), Bajo (2–7), Moderado (7–40) y Muy alto ($\geq$ 40 µg/L).

9.4.1. Resultados de la primera prueba piloto

Las Figuras 65–67 (ver en Anexos) muestran las matrices de confusión difusas para SVM, KNN y redes neuronales profundas aplicadas a los 120 registros de la primera prueba piloto. En las tres matrices se observa un patrón común: la única fila con valores distintos de cero corresponde a la categoría Muy alto ($\geq$ 40 µg/L), mientras que las filas asociadas a Muy bajo, Bajo y Moderado son nulas. Esto indica que, bajo los umbrales definidos, las 120 muestras seleccionadas del estanque se localizaron íntegramente en el rango hipereutrófico. Dicho resultado es coherente con el histograma de la Figura 32 (ver en Anexos), donde la distribución de clorofila se encontraba fuertemente desplazada hacia valores superiores a 40 µg/L.

Dentro de la fila Muy alto, las diferencias entre modelos se manifiestan en cómo se reparte la pertenencia entre las clases predichas:

- En el modelo SVM (ver en Anexos la Figura 65), la categoría correcta acumula 95.83 unidades de pertenencia en la diagonal, mientras que 22.35 se asignan a Moderado y 0.82 a Bajo. Esto sugiere que, aunque el clasificador reconoce mayoritariamente el estado Muy alto, mantiene una proporción no despreciable de pertenencia difusa hacia el rango inmediatamente inferior, lo cual refleja la existencia de muestras cercanas al umbral de 40 µg/L o con perfiles fisicoquímicos compatibles tanto con estados eutróficos altos como moderados.
- En el modelo KNN (ver en Anexos la Figura 66), la pertenencia en la diagonal asciende a 99.83, con 19.17 unidades repartidas en Moderado y nula asignación a las clases bajas. Este patrón es similar al del SVM, pero con mayor concentración en la categoría correcta y una dispersión ligeramente menor hacia Moderado, lo que indica una frontera de decisión algo más ajustada en torno al estado hipereutrófico.
- En la red neuronal profunda (ver en Anexos la Figura 67), la matriz es prácticamente degenera: las 119.00 unidades de pertenencia se concentran íntegramente en la celda Muy alto–Muy alto, sin asignación a otras categorías. Esto indica que, para este subconjunto de datos, la red interpreta todas las muestras como inequívocamente hipereutróficas, sin registrar ambigüedad en la pertenencia a otros estados tróficos.

En términos ecológicos, estos resultados refuerzan la conclusión de que, durante la primera prueba piloto, el estanque se mantuvo en condiciones dominadas por altas concentraciones de clorofila, sin episodios detectables de mejora hacia estados Moderados o inferiores. La lógica difusa evidencia que, aunque SVM y KNN reconocen algunos casos en la frontera con la categoría inmediatamente inferior, la dinámica general del sistema corresponde a un estado persistentemente Muy alto.

9.4.2. Resultados de la segunda prueba piloto

En la segunda prueba piloto se incrementó el número de muestras a 416, manteniendo el mismo protocolo de medición y el mismo *pipeline* de inferencia. Las matrices de confusión difusas para SVM, KNN y redes neuronales se muestran en las Figuras 62–64 (ver en Anexos).

De nuevo, las tres matrices comparten una característica clave: todas las filas asociadas a Muy bajo, Bajo y Moderado son nulas, mientras que la fila Muy alto concentra el 100 % de la pertenencia acumulada. Esto indica que, al igual que en la primera prueba piloto, las 416 muestras de la segunda prueba piloto se ubicaron por encima del umbral de 40 µg/L. Es decir, el estanque continuó en estado hipereutrófico durante este periodo ampliado de monitoreo.

Al analizar el reparto de pertenencia en la fila Muy alto se observa:

- En el modelo SVM, la diagonal Muy alto–Muy alto acumula 334.81 unidades, con 78.31 asignadas a Moderado y 2.88 a Bajo. En comparación con la primera prueba piloto, aumenta tanto el número total de muestras como la magnitud absoluta de la pertenencia difusa hacia Moderado. Esto sugiere que, en la segunda prueba piloto, el estanque experimentó una mayor variabilidad interna dentro del rango hipereutrófico, con un subconjunto de mediciones que el modelo considera próximas a la frontera con el estado Moderado. No obstante, la predominancia de la diagonal confirma que la interpretación global sigue siendo Muy alto.
- En el modelo KNN, la celda correcta registra 323.06 unidades, frente a 92.94 en Moderado y 0.00 en las clases bajas. El patrón es similar al de SVM, pero con mayor énfasis en la transición hacia la categoría inmediatamente inferior. Este comportamiento es coherente con la filosofía de KNN, que responde de manera sensible a las estructuras locales de los datos, reflejando la existencia de mediciones con condiciones fisicoquímicas intermedias dentro del propio rango de alta clorofila.
- En la red neuronal profunda, la matriz vuelve a ser degenerada, con 416.00 unidades de pertenencia concentradas exclusivamente en la celda Muy alto–Muy alto. Este resultado refuerza la idea de que la red neuronal, entrenada únicamente con datos de AMSA, interpreta todos los patrones observados en el estanque como inequívocamente hipereutróficos, sin introducir gradientes de pertenencia hacia estados tróficos menores.

9.5. Quinta predicción usando datos del CEA aplicados a las pruebas piloto en el estanque

Con el objetivo de contrastar el comportamiento observado en la cuarta predicción, donde los modelos entrenados con AMSA clasificaron el estanque casi exclusivamente en el estado Muy alto, se realizó una quinta predicción utilizando ahora los modelos entrenados con el conjunto de datos del CEA. La lógica de este experimento fue evaluar cómo se modificaba la clasificación trófica del estanque cuando el modelo de referencia provenía de un sistema predominantemente oligotrófico (Lago de Atitlán), manteniendo invariable el conjunto de mediciones locales del estanque.

Al igual que en el caso anterior, se emplearon dos subconjuntos de datos del estanque: (i) 120 registros correspondientes a la primera prueba piloto (seis mediciones aleatorias por día durante veinte días) y (ii) 416 registros recolectados en la segunda prueba piloto. Las variables de entrada fueron pH, temperatura, oxígeno disuelto, conductividad y turbidez, normalizadas con los parámetros del entrenamiento del CEA. A partir de estas entradas se obtuvieron predicciones de clorofila-a que luego se discretizaron en las cuatro clases tróficas definidas en este trabajo. Para cada subconjunto se construyeron matrices de confusión con lógica difusa para los clasificadores SVM, KNN y redes neuronales profundas.

9.5.1. Resultados de la primera prueba

Las Figuras 56–58 (ver en Anexos) muestran las matrices de confusión difusas para SVM, KNN y redes neuronales profundas cuando se aplican los modelos entrenados con CEA a los 120 registros de la primera prueba piloto. En las tres matrices se observa un patrón inverso al obtenido con los modelos entrenados con AMSA. La única fila con valores distintos de cero es la correspondiente a la categoría Muy bajo (0–2 µg/L), mientras que las filas asociadas a Bajo, Moderado y Muy alto son nulas. Esto indica que, según los modelos calibrados con CEA, las 120 muestras del estanque se ubican íntegramente en el rango oligotrófico. Es decir, el mismo conjunto de mediciones fisicoquímicas que bajo el modelo de AMSA se interpretaba como hipereutrófico, ahora es interpretado como de muy baja concentración de clorofila.

Dentro de la fila Muy bajo, las diferencias entre modelos se reflejan en la distribución de la pertenencia entre clases predichas:

- El modelo SVM (ver en Anexos la Figura 56) concentra 117.94 unidades de pertenencia en la celda Muy bajo–Muy bajo, con una pequeña fracción (1.06) asignada a la categoría Bajo. Ello sugiere que, aunque la gran mayoría de las muestras se clasifican de manera inequívoca como oligotróficas, el SVM reconoce un subconjunto de observaciones cercanas al umbral superior de 2 µg/L, para las cuales la pertenencia se reparte parcialmente entre los rangos Muy bajo y Bajo.
- El modelo KNN (ver en Anexos la Figura 57) presenta un patrón ligeramente más difuso: 102.01 unidades se sitúan en la diagonal Muy bajo–Muy bajo, mientras que 16.99 se asignan a la categoría Bajo. El carácter local de KNN hace que algunos puntos con condiciones fisicoquímicas algo más extremas (por ejemplo, valores relativamente

mayores de turbidez o conductividad) se interpreten como candidatos a pertenecer parcialmente al rango 2–7 µg/L, aunque sin dejar de ser dominados por el estado Muy bajo.

- En la red neuronal profunda (ver en Anexos la Figura 58), la matriz es prácticamente degenerada: las 119.00 unidades de pertenencia se concentran íntegramente en la celda Muy bajo–Muy bajo, sin asignaciones a otras categorías. En este caso, la red interpreta todas las muestras de la primera prueba piloto como inequívocamente oligotróficas, sin evidencias de transición hacia estados superiores.

La comparación con los resultados de la cuarta predicción pone de manifiesto un contraste importante: mientras que los modelos basados en AMSA ubicaban el estanque en el estado Muy alto, los modelos basados en CEA lo sitúan claramente en Muy bajo. Esta discrepancia cuantitativa sugiere que la interpretación trófica del estanque depende fuertemente del dominio de entrenamiento utilizado como referencia (eutrófico–hipereutrófico en AMSA vs. oligotrófico en CEA) y evidencia un fenómeno de desajuste de dominio entre los sistemas limnológicos de origen y el estanque experimental.

9.5.2. Resultados de la segunda prueba piloto

Las matrices de confusión difusas obtenidas al aplicar los mismos modelos del CEA a los 416 registros de la segunda prueba piloto se presentan en las Figuras 61, 59 y 60 (ver en Anexos).

Al igual que en la primera prueba piloto, todas las filas correspondientes a las categorías Bajo, Moderado y Muy alto son nulas, de modo que la totalidad de la pertenencia se concentra en la fila Muy bajo. Esto implica que, bajo el criterio de los modelos entrenados con CEA, las 416 muestras de la segunda prueba piloto también se encuentran dentro del rango 0–2 µg/L. No obstante, la distribución interna de la pertenencia en la fila Muy bajo presenta matices específicos en cada modelo:

- En el SVM (ver en Anexos la Figura 59), la diagonal Muy bajo-Muy bajo acumula 411.47 unidades, mientras que 4.52 se asignan a la categoría Bajo. La presencia de esta pequeña fracción refuerza la idea de que existe un subconjunto de mediciones cuya posición en el espacio de atributos se sitúa cerca del umbral entre Muy bajo y Bajo, aunque la clasificación dominante sigue siendo oligotrófica.
- En el KNN (ver en Anexos la Figura 60), la pertenencia se concentra de forma totalmente nítida en la celda Muy bajo–Muy bajo (416 unidades), sin asignaciones a otras celdas. El clasificador por vecinos interpreta todas las muestras de la segunda prueba piloto como inequívocamente ligadas al patrón de baja clorofila aprendido del CEA.
- La red neuronal profunda (ver en Anexos la Figura 61) reproduce el mismo patrón: 416 unidades de pertenencia en la diagonal y cero en el resto de celdas, confirmando una interpretación fuertemente oligotrófica del estanque bajo el modelo de Atitlán.

9.6. Sexta predicción usando datos combinados de AMSA y CEA aplicados a las pruebas piloto del estanque

Con el propósito de evaluar si la combinación de bases de datos provenientes de AMSA y CEA podía generar un modelo más robusto y generalizable, se desarrolló un conjunto de clasificadores entrenados con el conjunto de datos de AMSA y CEA. Esta integración incrementó tanto la diversidad de escenarios fisicoquímicos como la variabilidad de concentraciones de clorofila-a, lo cual permitió construir fronteras de decisión más amplias, capaces de representar condiciones desde estados mesotróficos hasta hipereutróficos. Posteriormente, estos modelos fueron aplicados a las pruebas pilotos medidas en el estanque del jardín botánico compuesta por 120 datos seleccionadas aleatoriamente para la primera prueba y con 416 datos obtenidos en la segunda prueba.

En ambos casos, la evaluación se realizó mediante matrices de confusión difusas, generadas a partir de los grados de pertenencia asignados a cada categoría trófica (*Muy bajo*, *Bajo*, *Moderado* y *Muy alto*). Este enfoque permitió capturar la incertidumbre inherente al proceso de predicción, especialmente en límites de decisión donde un valor de clorofila puede presentar afinidad simultánea por dos o más clases.

9.6.1. Resultados de la primera prueba piloto

- La Figura 53 (ver en Anexos) presenta los resultados obtenidos por el modelo SVM al ser aplicado a los 120 datos de la primera prueba piloto del estanque. Se observó que la totalidad de la pertenencia se concentró en la categoría Muy alto (≥ 40 μg/L), sin asignaciones significativas a las categorías inferiores. Esto sugiere que el clasificador interpretó los patrones fisicoquímicos del estanque como característicos de un ambiente altamente productivo, reforzando la marcada tendencia hacia valores elevados detectada en etapas anteriores del análisis.

- El modelo KNN, aplicado a la primera prueba piloto, mostró un patrón muy distinto respecto al SVM. En la Figura 54 (ver en Anexos) se evidencia que las predicciones se distribuyeron mayoritariamente entre Bajo, Moderado y Muy alto, lo cual refleja que el clasificador de vecinos más cercanos es sensible a pequeñas variaciones en las combinaciones de entrada. Su comportamiento sugiere que el estanque presentó intervalos donde la composición fisicoquímica pudo coincidir parcialmente con condiciones de menor productividad, aun cuando la mayoría de las observaciones correspondieron a estados elevados.

- El modelo de redes neuronales profundas (NN) aplicado a la primera prueba piloto mostró una asignación casi monolítica a la categoría Muy alto (ver en Anexos la Figura 55). Este patrón sugiere que la red aprendió representaciones altamente no lineales que priorizan los escenarios de alta productividad integrados en la base de entrenamiento combinada.

9.6.2. Resultados de la segunda prueba piloto

- El comportamiento del modelo SVM mostró una mayor dispersión (ver en Anexos la Figura 68). Aunque la categoría Muy alto continuó siendo dominante, aparecieron contribuciones relevantes en Moderado y, en menor proporción, en Bajo. Esto indica que el modelo reconoció una mayor heterogeneidad en las condiciones del estanque durante esta segunda prueba piloto, posiblemente debido a oscilaciones naturales en turbidez, oxígeno disuelto o variaciones de temperatura.
- El modelo KNN presentó un cambio notable (ver en Anexos la Figura 69). Se observó una separación marcada entre las categorías Moderado y Muy alto, sin contribuciones en Muy bajo o Bajo. Este comportamiento refleja un aumento significativo en la estabilidad interna del modelo: al contar con un mayor número de muestras en esta segunda prueba piloto, las predicciones se concentraron en rangos más consistentes y representativos del estado eutrófico del sistema.
- Con la red neuronal profunda se exhibió una división más balanceada entre Moderado y Muy alto (ver en Anexos la Figura 70), lo cual sugiere que el modelo percibió oscilaciones relevantes en las variables fisicoquímicas que matizaron la asignación exclusiva hacia la categoría más alta.

La integración de sensores, modelos de aprendizaje automático y visualizaciones interactivas permitió analizar de manera integral la dinámica de clorofila-a en distintos cuerpos de agua, así como su manifestación en el estanque experimental del jardín botánico de la Universidad del Valle de Guatemala.

$$\hat{y} = 565.7 + 2.463\,\mathrm{pH} - 29.1\,\mathrm{T} + 0.06368\,\mathrm{CO} - 0.3786\,\mathrm{OD} + 6.261\,\mathrm{TU}\ [\mu\mathrm{g/L}] \qquad (2)$$

La ecuación (3) corresponde a la aproximación lineal derivada del modelo de red neuronal profunda entrenado con los datos combinados de CEA y AMSA. Cada coeficiente representa la contribución marginal de una variable fisicoquímica sobre la estimación de la concentración de clorofila-a. En particular, un aumento en el pH, la conductividad (CO) y la turbidez (TU) incrementa la predicción de clorofila-a, mientras que la temperatura (T) y el oxígeno disuelto (OD) presentan efectos negativos en el valor estimado. Esta expresión resume la relación lineal aprendida por el modelo y permite interpretar de manera directa la magnitud y dirección de la influencia de cada parámetro sobre la variable objetivo.

CAPÍTULO 10

Conclusiones

El presente trabajo integró un diseño experimental en un espacio controlado con un sistema multisensor y técnicas de aprendizaje automático para la detección indirecta de clorofila-a, utilizada como variable *proxy* del estado trófico. A partir de la definición explícita de la variable respuesta (clorofila-a) y los factores experimentales (pH, conductividad, oxígeno disuelto, turbidez y temperatura), se estableció un flujo metodológico que abarcó la instrumentación, la recolección y depuración de datos, el modelado supervisado con estrategias de validación y la evaluación operativa mediante una interfaz de despliegue. Las principales conclusiones se resumen a continuación.

- **Predicción indirecta exitosa.** La integración de datos fisicoquímicos provenientes de distintos lagos (AMSA y CEA), junto con modelos de aprendizaje automático, permitió estimar de manera fiable la concentración de clorofila-a como indicador indirecto de proliferación de cianobacterias.
- **Modelos contrastantes según dominio de entrenamiento.** Los modelos entrenados con AMSA mostraron un sesgo hacia estados hipereutróficos (Muy alto), mientras que los entrenados con CEA favorecieron estados oligotróficos (Muy bajo), evidenciando un desajuste de dominio entre cuerpos de agua ecológicamente distintos. El conjunto combinado AMSA+CEA produjo el comportamiento más equilibrado y generalizable.
- **Desempeño diferencial entre algoritmos.** Las redes neuronales profundas mostraron mayor estabilidad y menor dispersión en las predicciones, aunque con fronteras más rígidas los modelos SVM y KNN capturaron transiciones difusas entre categorías, especialmente cerca de los umbrales tróficos, ofreciendo interpretaciones más matizadas del gradiente ecológico.
- **Relevancia de los parámetros.** La turbidez, la conductividad y el pH se confirmaron como las variables más influyentes en todos los modelos, coherentes con procesos limnológicos asociados a biomasa algal, sólidos suspendidos y actividad fotosintética. El oxígeno disuelto y la temperatura fueron indicadores de efectos negativos durante las pruebas piloto en el estanque.

- **Validación experimental en el estanque.** En ambas pruebas piloto, los modelos detectaron condiciones predominantemente eutróficas o hipereutróficas, consistentes con los valores elevados de turbidez y conductividad medidos *in situ*. Las matrices difusas mostraron pequeñas transiciones hacia la categoría Moderado, sin evidenciar episodios sostenidos de mejora ecológica.

- **Sistema multisensor y tablero interactivo como herramientas de monitoreo.** El sistema de sensores implementado entregó mediciones estables y reproducibles de pH, turbidez, oxígeno disuelto, conductividad y temperatura. La plataforma interactiva en *Streamlit* funcionó como un visor operativo para interpretar predicciones, facilitando usos futuros en vigilancia ambiental, alerta temprana y toma de decisiones institucionales.

- **Aporte de ingeniería del sistema multisensor.** El desarrollo del prototipo requirió tres iteraciones hasta obtener un diseño definitivo impreso en 3D que aseguró repetibilidad geométrica, separación entre zonas húmedas y secas, y facilidad de mantenimiento. La integración electrónica evidenció limitaciones prácticas, particularmente la necesidad de ejecutar el sensor de conductividad en un programa independiente debido a incompatibilidades de temporización y librerías.

- **Limitaciones de calidad y distribución de los datos.** El conjunto de Ibagua presentó valores nulos, etiquetas NR y presencia de extremos que afectaron la estabilidad del entrenamiento. Aunque la transformación logarítmica mitigó la asimetría, la reducción del tamaño muestral limita la capacidad de generalización. Estas restricciones no aparecen reflejadas en los modelos combinados.

- **Transferencia entre dominios.** El modelo entrenado con AMSA fue aplicado a los conjuntos de Inagua y del estanque experimental, produciendo distribuciones plausibles, pero altamente concentradas por encima de 40 μg/L. Esta tendencia se atribuyó al desbalance de clases del entrenamiento y diferencias de dominio entre fuentes, aspectos no capturados en los modelos entrenados con CEA o datos combinados.

- **Limitaciones generales del estudio.** Persistieron restricciones importantes como la escasa representación de niveles bajos de clorofila, diferencias metodológicas entre fuentes (AMSA, Ibagua y estanque), y la necesidad de separar el módulo de conductividad en un código aislado. Estas limitaciones recomiendan cautela al extrapolarlas fuera de rangos bien representados.

CAPÍTULO 11

Recomendaciones

A partir de la experiencia acumulada en el diseño experimental, la instrumentación multisensor y el modelado supervisado para la detección indirecta de clorofila-a, se formulan las siguientes recomendaciones técnicas y metodológicas. Su propósito es fortalecer la validez interna y externa del estudio, mejorar la trazabilidad metrológica y aumentar la utilidad operativa del sistema para el monitoreo del estado trófico.

- **Ampliar la base de entrenamiento con otros lagos.** Incorporar datos de calidad de agua de otros cuerpos de agua como el lago de Petén Itzá, lago de Izabal, lago Güija (Jutiapa) y otros para aumentar la variabilidad ecológica, mejorar la robustez estadística y reducir el desajuste de dominio entre modelos y aplicaciones reales.
- **Mejorar la representación de rangos bajos de clorofila.** Ejecutar las muestras dirigidas a registrar más observaciones en los rangos Muy bajo y Bajo (0–7 μg/L), con el fin de fortalecer la discriminación del modelo y evitar sesgos hacia estados eutróficos o hipereutróficos.
- **Optimizar los sensores ópticos.** Evaluar la sustitución del sensor de turbidez por tecnologías relacionadas con color aparente o absorbancia espectral, que ofrecen mayor sensibilidad a biomasa algal y sólidos disueltos, especialmente en concentraciones bajas.
- **Fortalecer los mecanismos de validación del modelo.** Implementar métricas interpretables como SHAP y AUC-ROC para evaluar la consistencia interna del modelo, detectar datos atípicos y verificar la calibración estadística de las redes neuronales profundas.
- **Mejorar la calidad metrológica del sistema multisensor.** Mantener protocolos reproducibles de calibración y control (QA/QC) para pH, conductividad, oxígeno disuelto y turbidez, registrando condiciones ambientales y verificaciones pre/post- mediciones para garantizar trazabilidad.
- **Fortalecer el diseño experimental mediante muestreo temporal y réplicas.** Estructurar las pruebas piloto con réplicas intra-día (por ejemplo, 8:00, 12:00 y 16:00)

y muestreos semanales, manteniendo al menos tres réplicas por punto de medición. Además, registrar metadatos ambientales relevantes (precipitación, insolación, viento y caudales de aporte) para mejorar el control de confundidores y reforzar la validez interna del estudio.

- **Documentar y aplicar compensaciones metrológicas por temperatura, junto con prácticas de mantenibilidad del sistema.** Conservar la temperatura como referencia transversal para pH, conductividad y oxígeno disuelto, documentando las ecuaciones de compensación utilizadas por cada sensor. Asimismo, mantener un diseño modular del sistema físico, incorporando alivio de tensión en el cableado y guías mecánicas que aseguren una inmersión constante y repetible durante las pruebas piloto.
- **Estandarizar la gestión y trazabilidad de datos mediante reglas de limpieza y versionamiento estructurado.** Implementar un esquema robusto de datos con versiones diferenciadas (*raw*, *clean* y *features*), acompañado de metadatos obligatorios que incluyan sensor, calibración, operador y ubicación. Definir criterios explícitos para el tratamiento de valores NR, ceros físicos y atípicos, documentando las transformaciones aplicadas la logarítmica a fin de garantizar reproducibilidad y mejorar la integridad del flujo analítico.

CAPÍTULO 12

Referencias

[1] F. García y V. Miranda, «Eutrofización: una amenaza para el recurso hídrico» en *Agenda pública para el desarrollo regional, la metropolización y la sostenibilidad*, Universidad Nacional Autónoma de México, 2018, págs. 35-367, http://ru.iiec.unam.mx/4269/1/2-Vol2_Parte1_Eje3_Cap5-177-Garc%C3%ADa-Miranda.pdf.

[2] Autoridad para el Manejo Sustentable de la Cuenca del Lago de Atitlán y su Entorno, «Proyectos para la recuperación y manejo del Lago de Atitlán» 2020, https://amsa.gob.gt/?page_id=287, visitado el 22 de marzo de 2025.

[3] E. Rodas-Pernillo y C. A. Vasquez-Moscoso, «Evaluación anual del fitoplancton y su respuesta a la calidad de agua en el lago de Amatitlán, Guatemala» *Ciencia, Tecnología y Salud*, vol. 7, n.º 2, págs. 170-188, 2020. https://revistas.usac.edu.gt/index.php/cytes/article/view/708.

[4] IBAGUA ONG, «Perfil institucional en LinkedIn» 2025, https://gt.linkedin.com/in/ibagua-ong-0a3271259, visitado el 29 de marzo de 2025.

[5] P. Cano Ruiz, «Boyas multisensoriales para el monitoreo de posibles áreas vulnerables a la proliferación de la cianobacteria en el Lago de Atitlán» Trabajo de graduación, Universidad del Valle de Guatemala, Guatemala, 2020.

[6] H. Rodríguez-Rangel, D. M. Arias, L. A. Morales-Rosales, V. Gonzalez-Huitron, M. Valenzuela Partida y J. García, «Machine learning methods modeling carbohydrate-enriched cyanobacteria biomass production in wastewater treatment systems» *Energies*, vol. 15, n.º 7, 2022. DOI: 10.3390/en15072500.

[7] M. Zhang, Y. Zhang, S. Yu, Y. Gao, J. Dong y W. Zhu, «Two machine learning approaches for predicting cyanobacteria abundance in aquaculture ponds» *Ecotoxicology and Environmental Safety*, vol. 258, 2023. DOI: 10.1016/j.ecoenv.2023.114944.

[8] K. Zolfaghari, N. Pahlevan, C. Binding, D. Gurlin, S. G. H. Simis y A. Ruiz Verdú, «Impact of spectral resolution on quantifying cyanobacteria in lakes and reservoirs: A machine-learning assessment» *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, págs. 1-25, 2022. DOI: 10.1109/TGRS.2021.3114635.

[9] C. L. Cardenas Garcia, C. Gonzalez Amarillo y M. Mendoza Moreno, *IoTMonitor-WQ: A Remote water-quality monitoring IoT platform*, ver. 1, Mendeley Data, 2021. DOI: 10.17632/gyryhnp52h.1. visitado el 30 de agosto de 2025. https://data.mendeley.com/datasets/gyryhnp52h/1.

[10] ScienceDirect Topics, «Water Quality Criteria — an overview» https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/water-quality-criteria, visitado el 30 de agosto de 2025.

[11] U.S. Environmental Protection Agency, «Turbidity Parameter Factsheet» https://www.epa.gov/awma/turbidity-parameter-factsheet, visitado el 30 de agosto de 2025.

[12] U.S. Environmental Protection Agency, «Dissolved Oxygen Parameter Factsheet» https://www.epa.gov/awma/dissolved-oxygen-parameter-factsheet, visitado el 30 de agosto de 2025.

[13] U.S. Environmental Protection Agency, «Indicators: Conductivity» https://www.epa.gov/national-aquatic-resource-surveys/indicators-conductivity, visitado el 30 de agosto de 2025.

[14] U.S. Environmental Protection Agency, «Temperature» https://www.epa.gov/caddis/temperature, visitado el 30 de agosto de 2025.

[15] U.S. Environmental Protection Agency, «pH Parameter Factsheet» https://www.epa.gov/awma/ph-parameter-factsheet, visitado el 30 de agosto de 2025.

[16] L. Taiz y E. Zeiger, *Fisiología Vegetal*, 3.ª ed. Barcelona, España: Editorial Omega, 2002, ISBN: 978-84-282-1202-6.

[17] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts y P. Walter, *Biología molecular de la célula*, 6.ª ed. Editorial Médica Panamericana, 2015, ISBN: 978-84-9835-720-8.

[18] ScienceDirect Topics, «Chlorophyll A — an overview» https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/chlorophyll-a, visitado el 22 de septiembre de 2025.

[19] G. D. Domínguez, E. Ortiz y C. Cruz, «Principales propiedades inmunomoduladoras y antiinflamatorias de la ficobiliproteína C-ficocianina» *Revista Cubana de Hematología, Inmunología y Hemoterapia*, vol. 32, n.º 4, págs. 399-412, 2016. https://www.medigraphic.com/pdfs/revcubheminmhem/rch-2016/rch164d.pdf.

[20] ScienceDirect Topics, «Phycocyanin — an overview» https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/phycocyanin, visitado el 22 de septiembre de 2025.

[21] F. Garcia-Pichel, J. P. Zehr, D. Bhattacharya y H. B. Pakrasi, «What's in a name? The case of cyanobacteria» *Journal of Phycology*, vol. 56, n.º 1, págs. 1-5, 2020. DOI: 10.1111/jpy.12934. visitado el 22 de septiembre de 2025. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7065140/.

[22] M. T. Madigan, J. M. Martinko y J. Parker, *Brock: Biología de los microorganismos*, 14.ª ed. Pearson Educación, 2014, ISBN: 978-84-205-4363-6.

[23] ScienceDirect Topics, «Cyanobacteria — an overview» https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/cyanobacteria, visitado el 22 de septiembre de 2025.

[24] Woods Hole Oceanographic Institution (WHOI), «Cyanobacteria — Species Life Cycle» n.d. https://hab.whoi.edu/species/species-life-cycle/cyanobacteria/, visitado el 25 de septiembre de 2025.

[25] M. Begon, C. R. Townsend y J. L. Harper, *Ecología: de individuos a ecosistemas*, 4.ª ed. Ediciones Omega, 2006, ISBN: 978-84-282-1481-5.

[26] E. Orquera y M. Cabrera, «Caracterización del estado trófico de la Laguna de Yambo mediante análisis de fósforo» *InfoANALÍTICA*, vol. 4, n.º 2, págs. 23-31, 2020. https://dialnet.unirioja.es/descarga/articulo/7407203.pdf.

[27] M. F. Meyer, B. M. Kraemer, C. C. Barbosa, D. G. F. Cunha, W. K. Dodds, S. E. Hampton, C. Ordóñez, R. M. Pilla, A. I. Pollard, J. A. Culpepper, A. K. Fremier, T. V. King, R. Ladwig, D. M. Leech, S. I. S. Matsuzaki, I. A. Oleksy, S. N. Topp, R. I. Woolway, L. S. Brighenti, K. C. Fickas, B. P. Lanouette, J. Ren, M. Werther y X. Yang, «Clarifying the trophic state concept to advance macroscale freshwater science and management» *Ecosphere*, 2025. DOI: 10.1002/ecs2.70392. visitado el 22 de septiembre de 2025. https://esajournals.onlinelibrary.wiley.com/doi/epdf/10.1002/ecs2.70392.

[28] F. Nojavan, B. J. Kreakie, J. W. Hollister y S. S. Qian, «Rethinking the lake trophic state index» *PeerJ*, 2019. DOI: 10.7717/peerj.7936. visitado el 22 de septiembre de 2025. https://peerj.com/articles/7936/.

[29] R. G. Wetzel, *Limnología: Lago y Río Ecosistemas*, 3.ª ed. Academic Press, 2001, ISBN: 978-0-12-744760-5.

[30] DFRobot, «Gravity: Analog Turbidity Sensor» n.d. https://www.dfrobot.com/product-1394.html, visitado el 25 de septiembre de 2025.

[31] DFRobot, «Gravity: Analog Dissolved Oxygen Sensor» n.d. https://www.dfrobot.com/product-1628.html?srsltid=AfmBOopn_VxWzOPeC2uK0Wv9ApHfPpxe-hlEnA6U0i2rmBZZSoM5nmyw, visitado el 25 de septiembre de 2025.

[32] J. Hancock, «Jitter — Understanding it, Measuring It, Eliminating It. Part 1: Jitter Fundamentals» 2004, https://www.highfrequencyelectronics.com/Apr04/HFE0404_Hancock.pdf, visitado el 12 de noviembre de 2025.

[33] DFRobot, «Gravity: Analog Electrical Conductivity Sensor Meter V2 (K=1)» n.d. https://www.dfrobot.com/product-1123.html, visitado el 25 de septiembre de 2025.

[34] Naylamp Mechatronics, «Sensor de temperatura digital DS18B20» n.d. https://naylampmechatronics.com/sensores-temperatura-y-humedad/16-sensor-de-temperatura-digital-ds18b20.html, visitado el 25 de septiembre de 2025.

[35] DFRobot, «Gravity: Analog pH Sensor Meter Kit V2» n.d. https://www.dfrobot.com/product-1782.html?tracking=62b57cd3364b3, visitado el 25 de septiembre de 2025.

[36] OpenAI ChatGPT, «Conversación en ChatGPT» 2025, https://chatgpt.com/c/68d77f56-6054-8329-b65e-45d5b14bb000, visitado el 25 de septiembre de 2025.

[37] E. Topol, S. R. Steinhubl y A. Torkamani, «Artificial intelligence in global health» *The Lancet Digital Health*, vol. 3, n.º 1, e1-e2, 2021. DOI: 10.1016/S2589-7500(21)00002-3. https://www.sciencedirect.com/science/article/pii/S2666285X21000042.

[38] *Springer Handbook of Engineering Statistics*. Springer, 2008. DOI: 10.1007/978-0-387-84858-7. https://link.springer.com/book/10.1007/978-0-387-84858-7.

[39] Purdue University, Department of Statistics, «Search results for: Applied Linear Regression» n.d. https://www.stat.purdue.edu/results.html?q=Applied+Linear+Regression+document#gsc.tab=0&gsc.q=Applied%20Linear%20Regression%20document&gsc.page=1, visitado el 25 de septiembre de 2025.

[40] S. Weisberg, *Applied Linear Regression*, 4.ª ed. Wiley, 2014. DOI: 10.1002/9781118548387. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387.

[41] L. Breiman, «Random Forests» *Machine Learning*, vol. 45, n.º 1, págs. 5-32, 2001. DOI: 10.1023/A:1010933404324. https://link.springer.com/article/10.1023/A%3A1010933404324.

[42] S. P. Lloyd, «Least squares quantization in PCM» *IEEE Transactions on Information Theory*, vol. 28, n.º 2, págs. 129-137, 1982. DOI: 10.1007/BF00994018. https://link.springer.com/article/10.1007/BF00994018.

[43] C. E. Shannon, «The zero error capacity of a noisy channel» *IEEE Transactions on Information Theory*, vol. 13, n.º 1, págs. 13-19, 1967. DOI: 10.1109/TIT.1967.1053964. https://dl.acm.org/doi/10.1109/TIT.1967.1053964.

[44] Y. LeCun, Y. Bengio y G. Hinton, «Deep learning» *Nature*, vol. 521, págs. 436-444, 2015. DOI: 10.1038/nature14539. https://www.nature.com/articles/nature14539.

[45] J. Kleinberg, «Navigation in a small world» en *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing (STOC)*, 2000, págs. 163-170. DOI: 10.1145/331499.331504. https://dl.acm.org/doi/10.1145/331499.331504.

[46] M. E. J. Newman, «The structure of scientific collaboration networks» *Science*, vol. 290, n.º 5500, págs. 2323-2326, 2001. DOI: 10.1126/science.290.5500.2323. https://www.science.org/doi/10.1126/science.290.5500.2323.

[47] V. Chandola, A. Banerjee y V. Kumar, «Anomaly detection: A survey» *ACM Computing Surveys*, vol. 41, n.º 3, págs. 1-58, 2009. DOI: 10.1145/1541880.1541882. https://dl.acm.org/doi/10.1145/1541880.1541882.

[48] A. Krizhevsky, I. Sutskever y G. E. Hinton, «ImageNet classification with deep convolutional neural networks» en *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Curran Associates Inc., 2012, págs. 1097-1105. DOI: 10.1145/3065386. https://ieeexplore.ieee.org/document/6472238.

[49] J. MacQueen, «Some methods for classification and analysis of multivariate observations» en *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, Berkeley, CA: University of California Press, 1967, págs. 281-297.

[50] L. Breiman, «Statistical modeling: The two cultures (with comments and a rejoinder by the author)» *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 373, n.º 2053, pág. 20 150 202, 2015. DOI: 10.1098/rsta.2015.0202. https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202.

[51] J. Schmidhuber, «Deep learning in neural networks: An overview» *Neural Networks*, vol. 61, págs. 85-117, 2015. DOI: 10.1145/2689746.2689747. https://dl.acm.org/doi/10.1145/2689746.2689747.

[52] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville e Y. Bengio, «Generative adversarial nets» en *Advances in Neural Information Processing Systems*, MIT Press, 2014, págs. 2672-2680. https://papers.nips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

[53] P. Cintula, C. G. Fermüller y C. Noguera, «Fuzzy logic,» 2016.

[54] O. G. Duarte V., «Aplicaciones de la lógica difusa» *Ingeniería e Investigación*, n.º 45, págs. 5-12, enero de 2000. DOI: 10.15446/ing.investig.n45.21308. https://revistas.unal.edu.co/index.php/ingeinv/article/view/21308.

[55] M. F. Goodchild, «GIScience» en *Encyclopedia of GIS*, Springer, 2008, págs. 403-407, https://link.springer.com/referenceworkentry/10.1007/978-0-387-30164-8_157.

[56] World Health Organization, *Guidelines for Safe Recreational Water Environments, Volume 1: Coastal and Fresh Waters*. Geneva: World Health Organization, 2003, Directrices para ambientes acuáticos recreativos seguros, ISBN: 9241545801. https://www.who.int/publications/i/item/9241545801.

[57] M. Zhang, X. Shi, Z. Yang, Y. Yu, L. Shi y B. Qin, «Long-term dynamics and drivers of phytoplankton biomass in eutrophic Lake Taihu» *Science of the Total Environment*, vol. 645, págs. 876-886, 2018. DOI: 10.1016/j.scitotenv.2018.07.220.

[58] P. Ayala Pineda, «sensores.ino — repositorio de código para sensor de cianobacteria» 2025, https://github.com/paolayalap/Dashboard_modelo_cianobacteria/blob/master/sensores.ino, visitado el 29 de septiembre de 2025.

[59] P. Ayala Pineda, «conduc_temp.ino — repositorio de código para conductividad y temperatura» 2025, https://github.com/paolayalap/Dashboard_modelo_cianobacteria/blob/master/conduc_temp.ino, visitado el 29 de septiembre de 2025.

[60] Arduino, «Arduino Mega 2560 — Documentación oficial de hardware» n.d. https://docs.arduino.cc/hardware/mega-2560/, visitado el 29 de septiembre de 2025.

[61] P. Ayala Pineda, «Dashboard modelo cianobacteria (app interactiva)» 2025, https://dashboard-cianobacteria.streamlit.app/, visitado el 29 de septiembre de 2025.

[62] P. Ayala Pineda, «Dashboard_modelo_cianobacteria — repositorio completo en GitHub» 2025, https://github.com/paolayalap/Dashboard_modelo_cianobacteria/tree/master, visitado el 29 de septiembre de 2025.

CAPÍTULO 13

Anexos

13.1. Tablas de datos

Esta sección corresponde a los datos utilizados para los entrenamientos, validaciones y modelos. A continuación, se muestran los datos iniciales.

Cuadro 4. Predicciones fisicoquímicas y clasificación de cianobacterias (AMSA + Ibagua).

pH	Temp. (°C)	Cond. (μS/cm)	O.D. (mg/L)	Turbidez (NTU)	Clorofila pred. (μg/L)
8.80	23.5	0.743	14.43	100	429.63
8.52	23.8	0.760	12.92	93	415.28
8.38	23.9	0.768	14.15	26	293.91
8.00	23.1	0.782	9.61	233	1169.05
9.09	23.5	0.582	16.71	25	265.03

Nota. Los valores de clorofila fueron estimados mediante modelos de aprendizaje automático y categorizados en clases de riesgo (NN, SVM, KNN). Elaboración propia.

Cuadro 5. Parámetros fisicoquímicos de muestras de agua (datos Ibagua).

Temp. (°C)	pH	O.D. (mg/L)	Turbidez (NTU)	Cond. (μS/cm)	Clorofila (μg/L)
23.5	8.80	14.43	100	0.743	1378.84
23.8	8.52	12.92	93	0.760	401.95
23.9	8.38	14.15	26	0.768	187.29
23.1	8.00	9.61	233	0.782	1948.05
23.5	9.09	16.71	25	0.582	234.83

Nota. Elaboración propia con base en datos de Ibagua.

13.2. Enlaces directos del proyecto

Repositorio con el código fuente y documentación: Repositorio en GitHub.

13.3. Aproximaciones lineales

Dado que las redes neuronales profundas capturan relaciones no lineales complejas, se derivaron aproximaciones lineales locales con el fin de interpretar cuál es la contribución marginal de cada variable fisicoquímica en la predicción final de clorofila-a. Las expresiones obtenidas para los tres conjuntos de entrenamiento (AMSA, CEA y AMSA+CEA) se presentan en las ecuaciones 3–5. Donde los parámetros están descritos por las siguientes variables: potencial de Hidrógeno (pH), temperatura (T), conductividad (CO), oxígeno disuelto (OD) y turbidez (TU).

13.3.1. Aproximación lineal NN entrenada con datos combinados

Estos datos contienen información del CEA y de AMSA.

$$\hat{y} = 565.7 + 2.463\,\mathrm{pH} - 29.1\,\mathrm{T} + 0.06368\,\mathrm{CO} - 0.3786\,\mathrm{OD} + 6.261\,\mathrm{TU}\ [\mu\mathrm{g/L}] \tag{3}$$

13.3.2. Aproximación lineal NN entrenada con datos del CEA

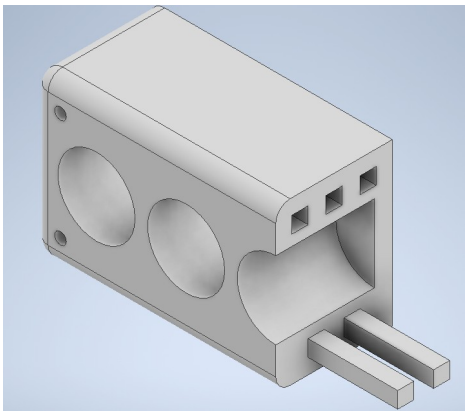
$$\hat{y} = -0.1174 + 0.1324\,\mathrm{pH} - 0.00366\,\mathrm{T} - 0.002064\,\mathrm{CO} + 0.2084\,\mathrm{OD} + 0.0004833\,\mathrm{TU}\ [\mu\mathrm{g/L}] \tag{4}$$

13.3.3. Aproximación lineal NN entrenada con datos del AMSA

$$\hat{y} = -558 + 53.8\,\mathrm{pH} + 1.109\,\mathrm{T} + 0.2245\,\mathrm{CO} - 0.0648\,\mathrm{OD} + 0.9882\,\mathrm{TU}\ [\mu\mathrm{g/L}] \tag{5}$$

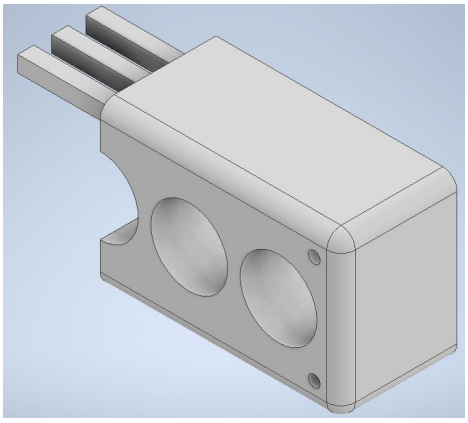
13.4. Estructura del sistema de sensores

Figura 33. Vista isométrica del diseño de la pieza con cavidades frontales para las sondas.



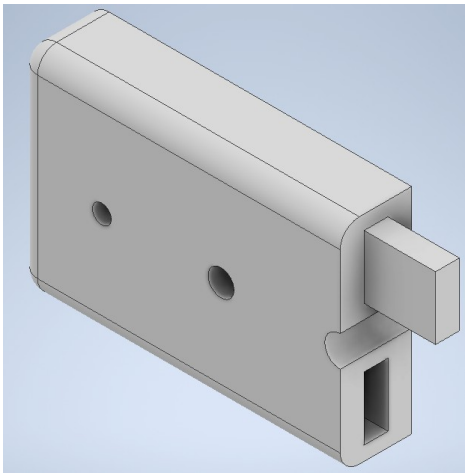
Nota. El diseño muestra los orificios principales de fijación y guía, así como las ranuras de entrada para cableado. Elaboración propia.

Figura 34. Vista posterior del diseño con perforaciones de guía.



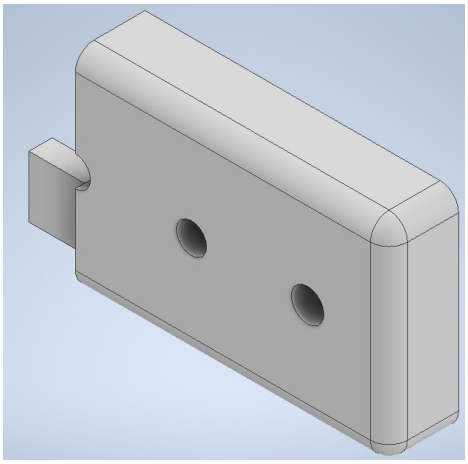
Nota. Se aprecian los alojamientos cilíndricos y el refuerzo estructural para los recipientes. Elaboración propia.

Figura 35. Vista lateral con puntos de fijación.



Nota. La imagen muestra los orificios de unión y el soporte estructural en la base de la pieza. Elaboración propia.

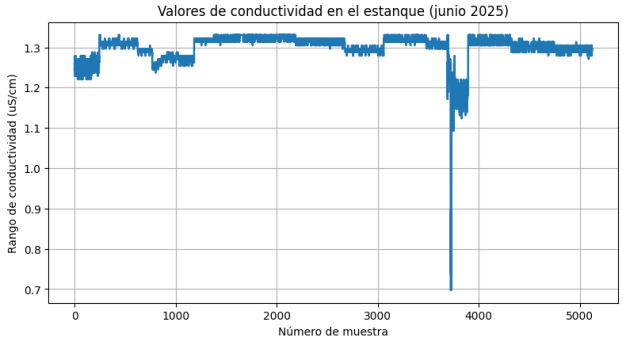
Figura 36. Vista trasera del módulo.



Nota. Se observan los bordes redondeados que mejoran la ergonomía y reducen concentraciones de esfuerzo. Elaboración propia.

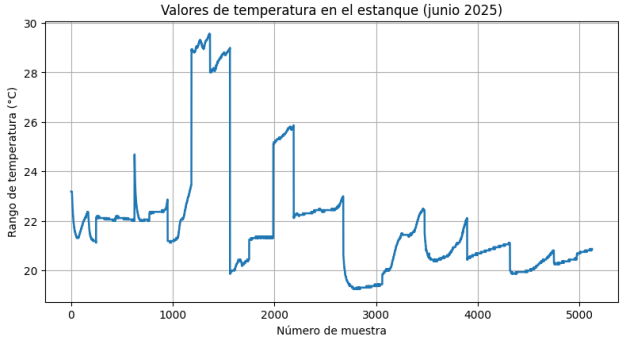
13.5. Comportamiento de los parámetros fisicoquímicos registrados en la primera prueba piloto del estanque

Figura 37. Comportamiento de los datos de conductividad.



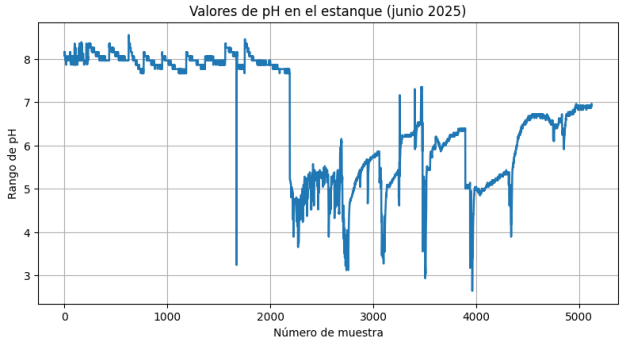
Nota. Datos de conductividad obtenidos en la primera prueba piloto. Elaboración propia.

Figura 38. Comportamiento de los datos de temperatura.



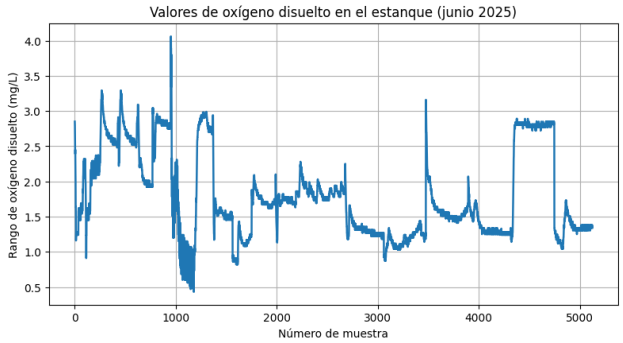
Nota. Datos de temperatura obtenidos en la primera prueba piloto. Elaboración propia.

Figura 39. Comportamiento de los datos de pH.



Nota. Datos de pH obtenidos en la primera prueba piloto. Elaboración propia.

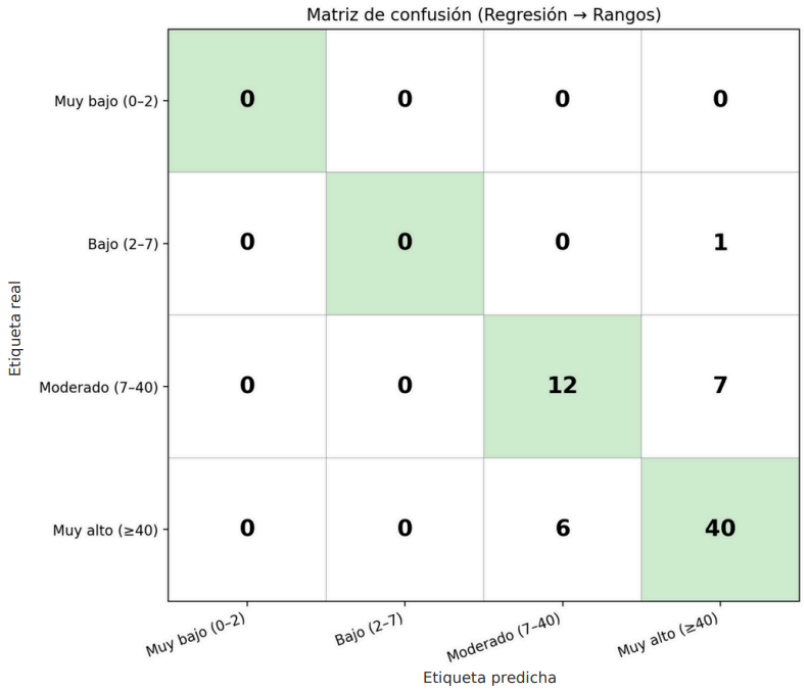
Figura 40. Comportamiento de los datos de oxígeno disuelto.



Nota. Datos de oxígeno disuelto obtenidos en la primera prueba piloto. Elaboración propia.

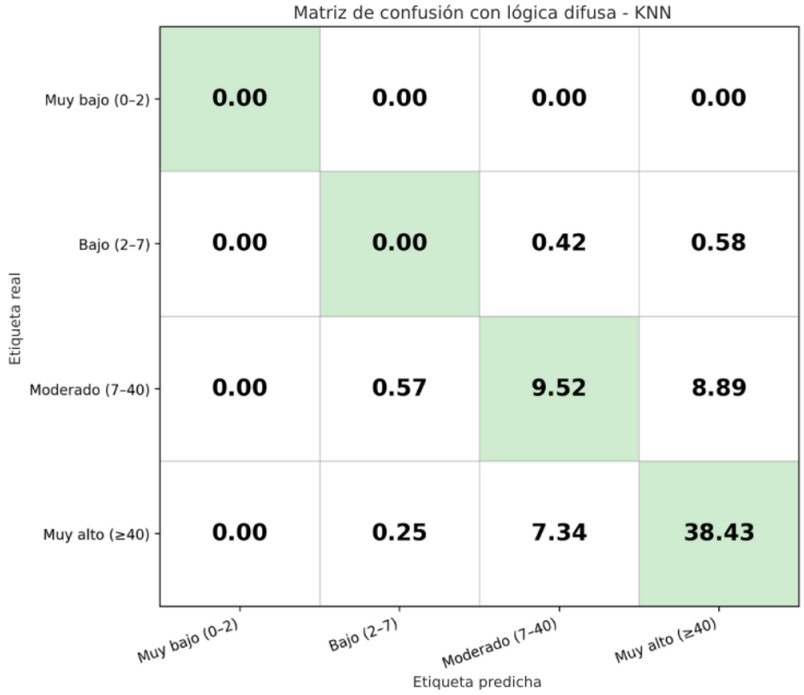
13.6. Clasificadores para las predicciones de clorofila-a

Figura 41. Matriz de confusión mediante regresión con datos de AMSA.



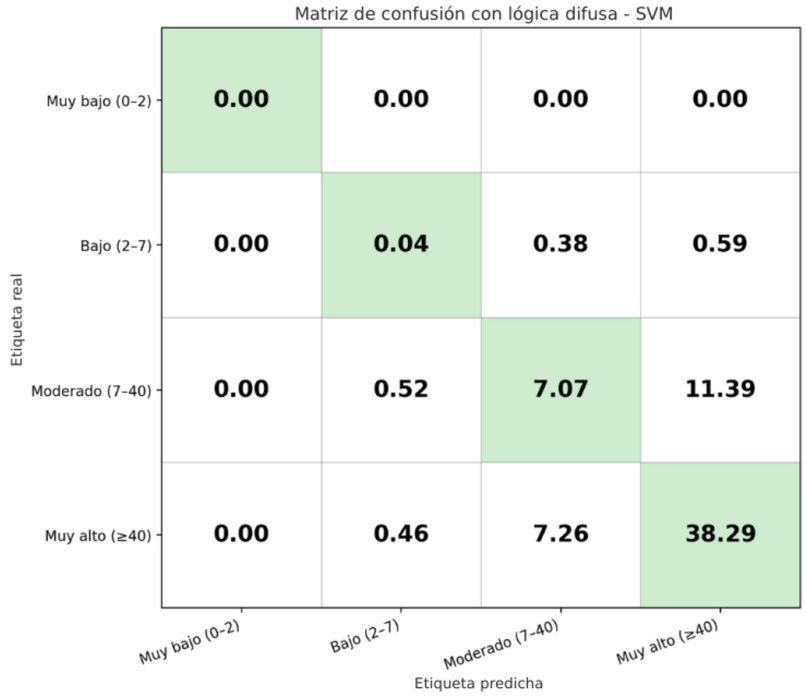
Nota. La imagen muestra la discretización de las cuatro clases definidas por la Organización Mundial de la Salud (OMS) [56] aplicado a los datos de entrenamiento de AMSA. Elaboración propia.

Figura 42. Matriz de confusión difusa con clasificador KNN para los datos de AMSA.



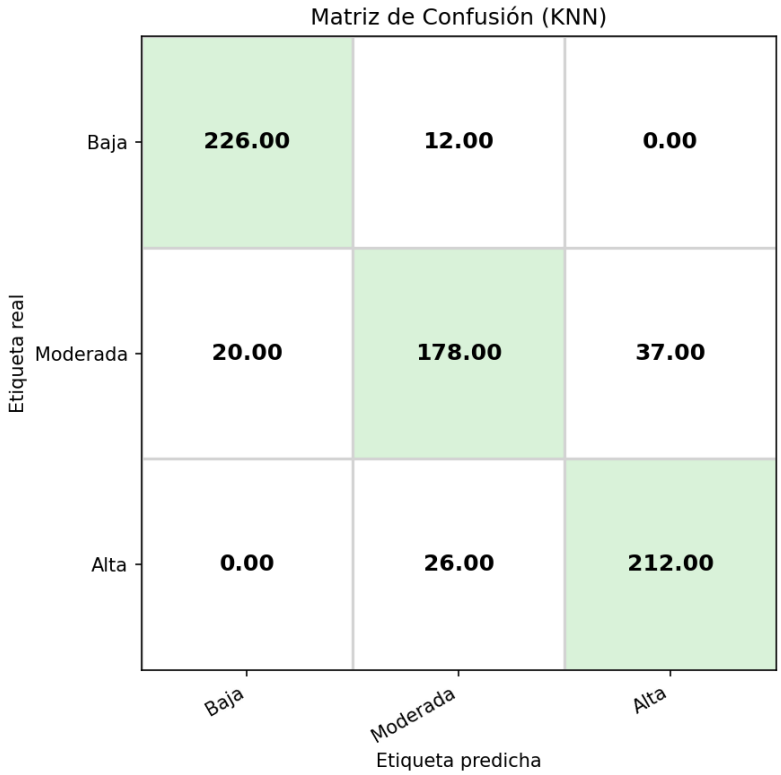
Nota. La imagen muestra jerarquía de importancia de cada parámetro dentro del modelo entrenado. Elaboración propia.

Figura 43. Matriz de confusión difusa con clasificador SVM para los datos de AMSA.



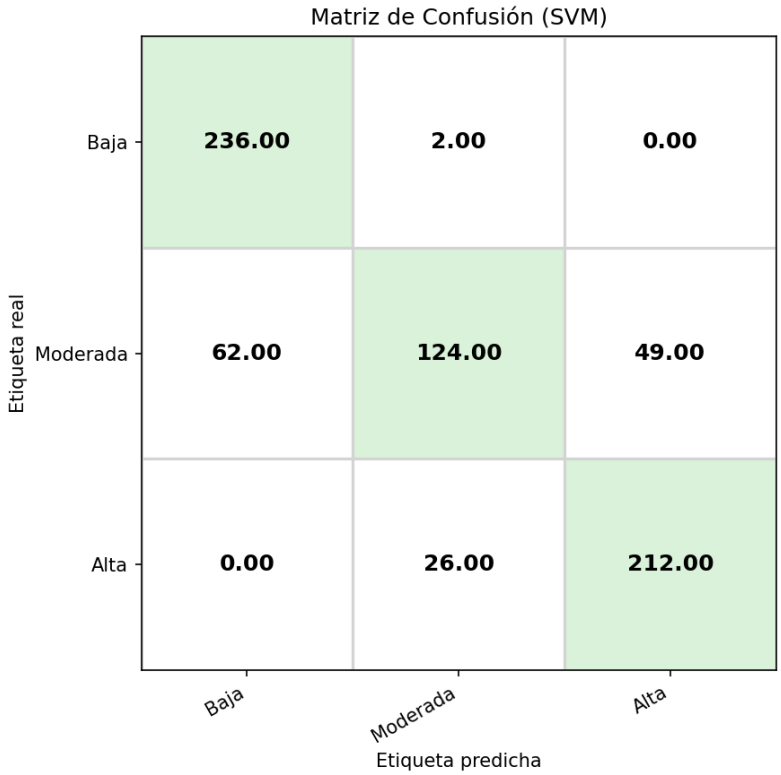
Nota. La imagen muestra jerarquía de importancia de cada parámetro dentro del modelo entrenado. Elaboración propia.

Figura 44. Matriz de confusión para el modelo KNN con datos del CEA.



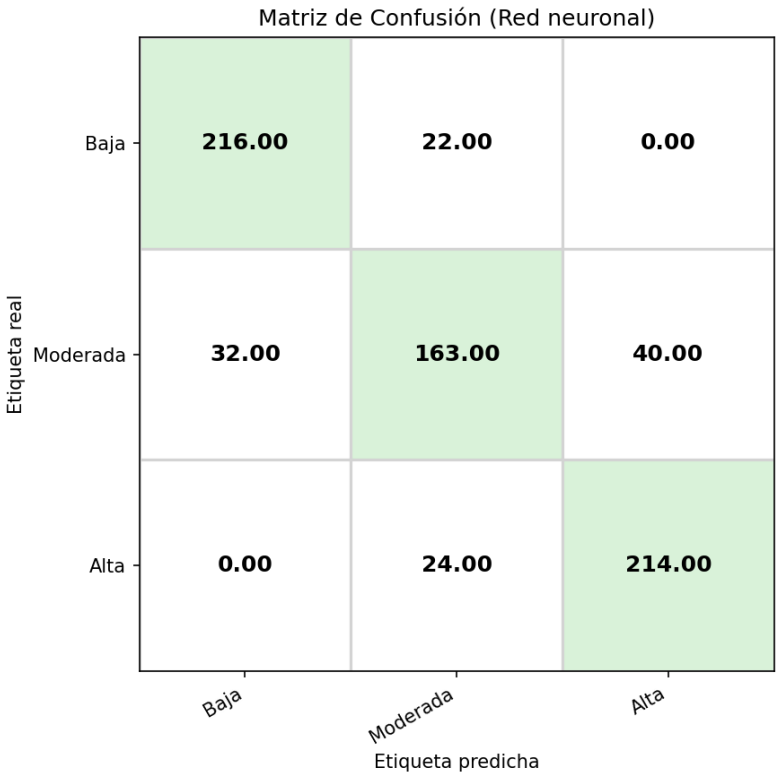
Nota. Clasificación en tres clases: Baja, Moderada y Alta. Elaboración propia.

Figura 45. Matriz de confusión para el modelo SVM con datos del CEA.



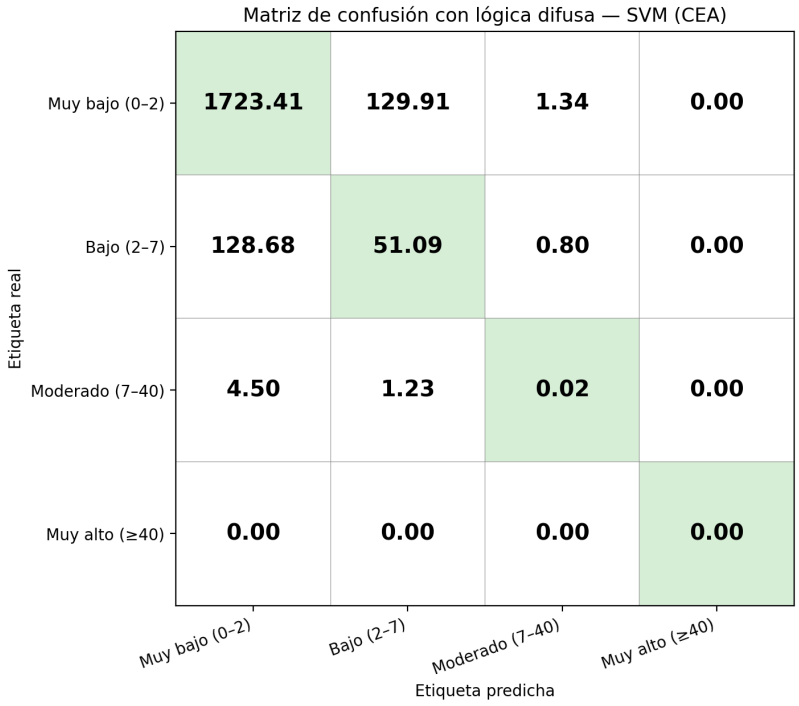
Nota. Clasificación en tres clases con un modelo SVM de margen blando. Elaboración propia.

Figura 46. Matriz de confusión para el modelo de red neuronal con datos del CEA.



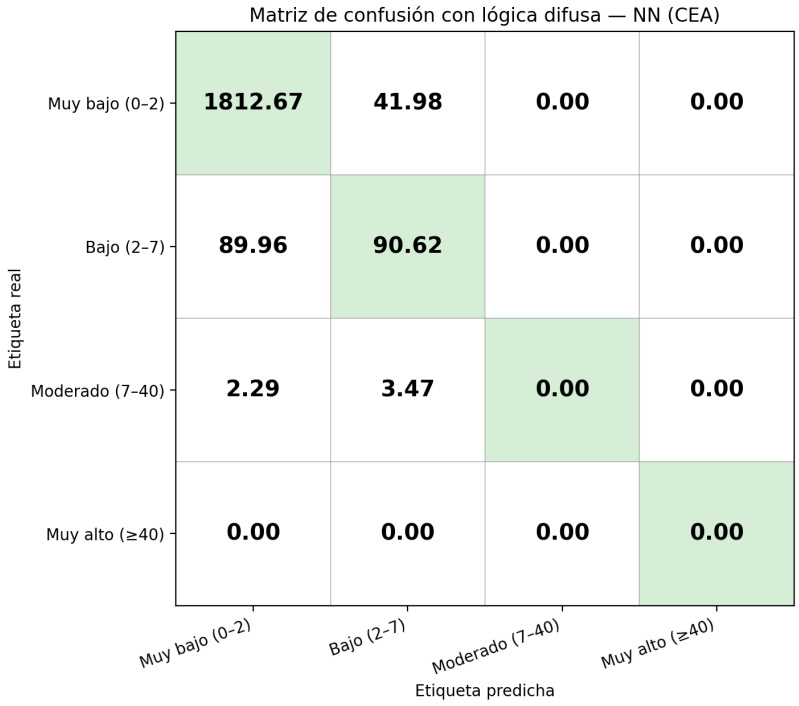
Nota. Clasificación en tres clases mediante un perceptrón multicapa. Elaboración propia.

Figura 47. Matriz de confusión con lógica difusa para el modelo SVM con datos del CEA.



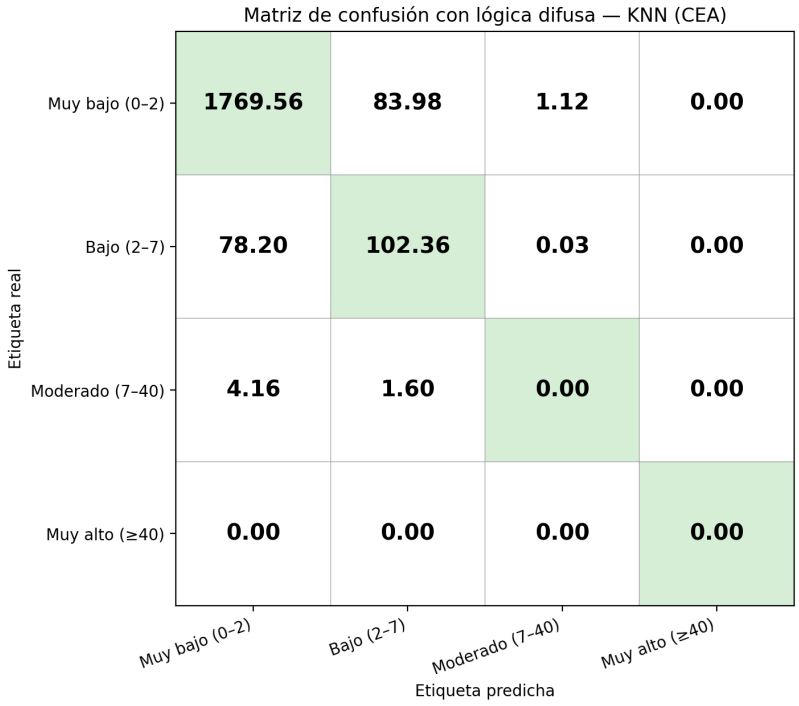
Nota. Los valores representan grados de pertenencia acumulados. Elaboración propia.

Figura 48. Matriz de confusión con lógica difusa para la red neuronal entrenada con datos del CEA.



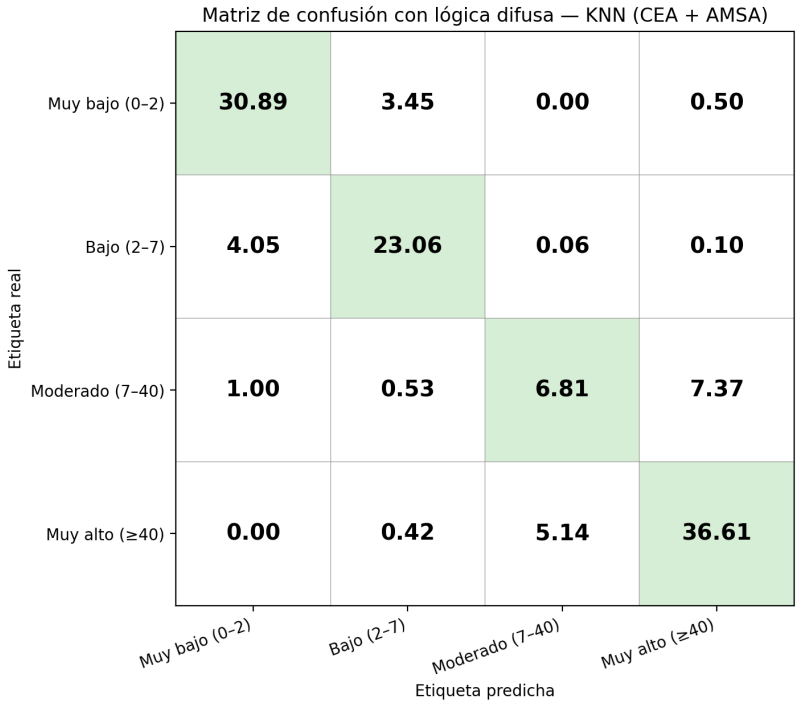
Nota. Se observan patrones más suaves en comparación con SVM. Elaboración propia.

Figura 49. Matriz de confusión con lógica difusa para el modelo KNN con datos del CEA.



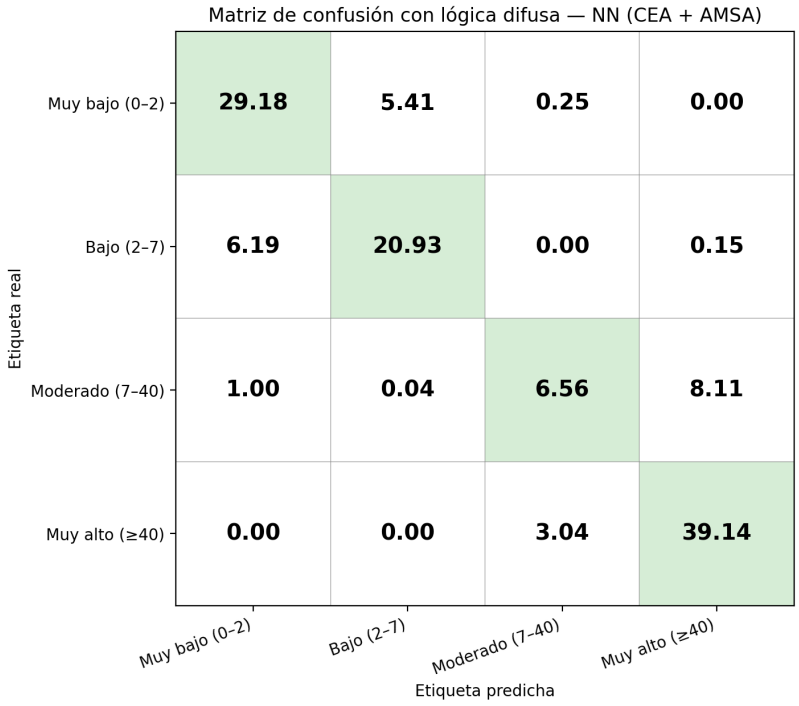
Nota. KNN mantiene patrones similares a la red neuronal, pero más rígidos. Elaboración propia.

Figura 50. Matriz de confusión con lógica difusa para el modelo KNN con datos combinados del CEA y AMSA.



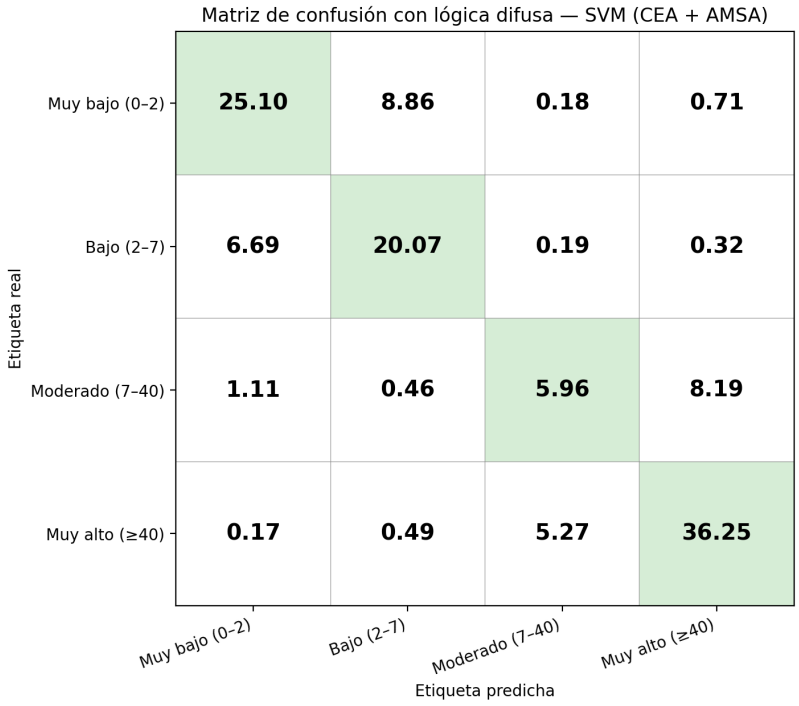
Nota. Los valores representan grados de pertenencia acumulados en cada categoría trófica. Elaboración propia.

Figura 51. Matriz de confusión con lógica difusa para la red neuronal profunda con datos combinados del CEA y AMSA.



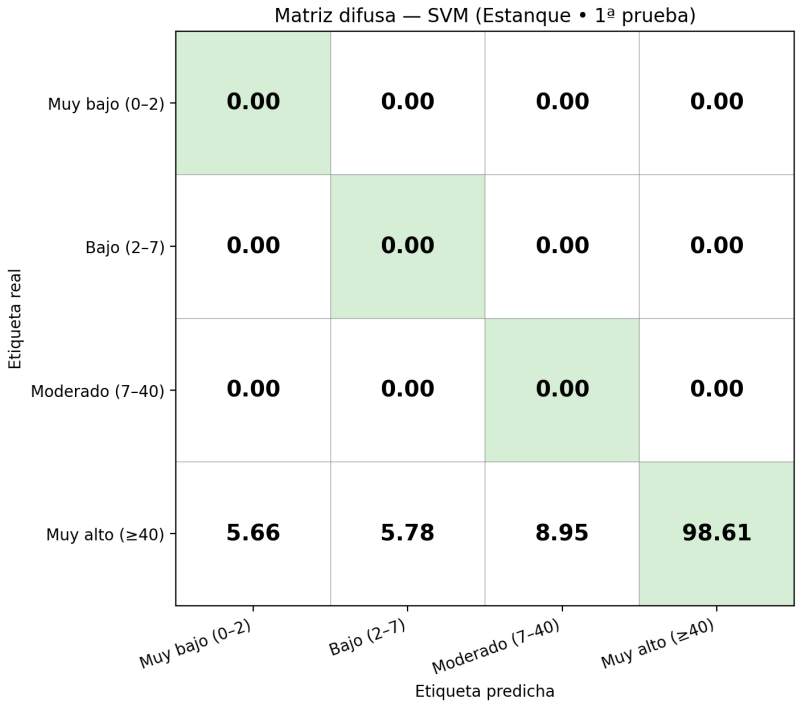
Nota. La red neuronal muestra transiciones más suaves entre clases. Elaboración propia.

Figura 52. Matriz de confusión con lógica difusa para el modelo SVM con datos combinados del CEA y AMSA.



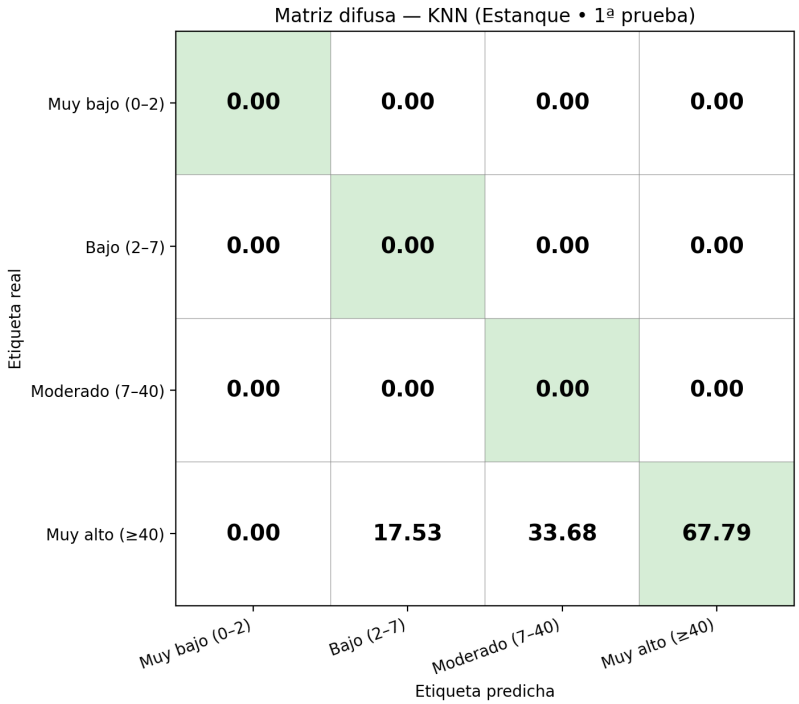
Nota. El modelo SVM mantiene fronteras más rígidas y simétricas entre clases. Elaboración propia.

Figura 53. Matriz de confusión con lógica difusa para el modelo SVM entrenado con AMSA y CEA, aplicado a la primera prueba piloto del estanque.



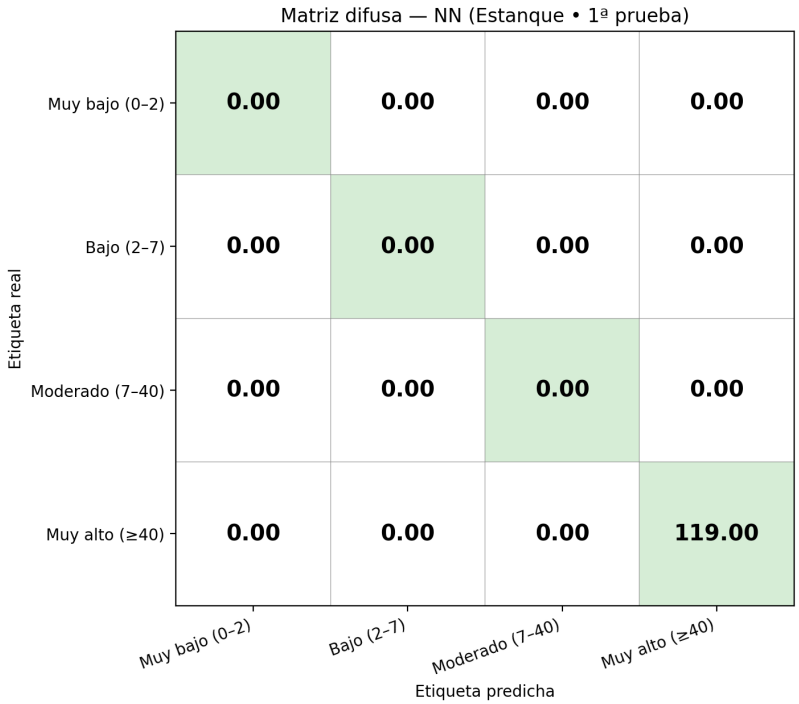
Nota. El SVM asigna prácticamente toda la pertenencia a la categoría Muy alto, lo que indica que percibe condiciones fisicoquímicas propias de un sistema persistentemente eutrófico o hipereutrófico. Elaboración propia.

Figura 54. Matriz de confusión con lógica difusa para el modelo KNN entrenado con AMSA y CEA, aplicado a la primera prueba piloto del estanque.



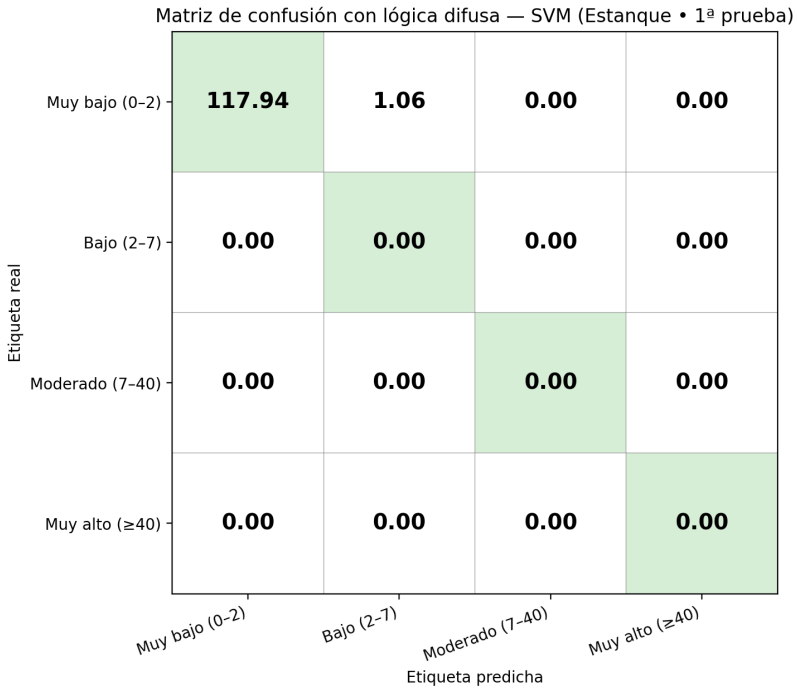
Nota. El KNN asigna pertenencias relevantes a categorías múltiples, indicando sensibilidad ante variaciones locales en las características fisicoquímicas. Elaboración propia.

Figura 55. Matriz de confusión con lógica difusa para el modelo de redes neuronales entrenado con AMSA y CEA, aplicado a la primera prueba piloto del estanque.



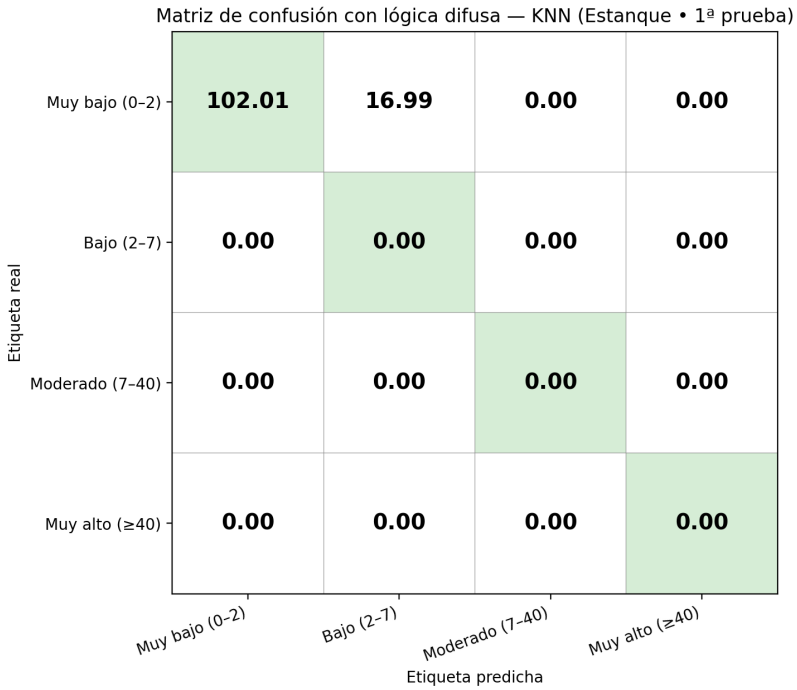
Nota. La red neuronal asignó prácticamente toda la pertenencia a Muy alto, indicando un patrón de respuesta estable y fuertemente condicionado por la presencia de valores elevados en el conjunto de entrenamiento combinado. Elaboración propia.

Figura 56. Matriz de confusión con lógica difusa para el modelo SVM entrenado con CEA, aplicado a la primera prueba piloto del estanque.



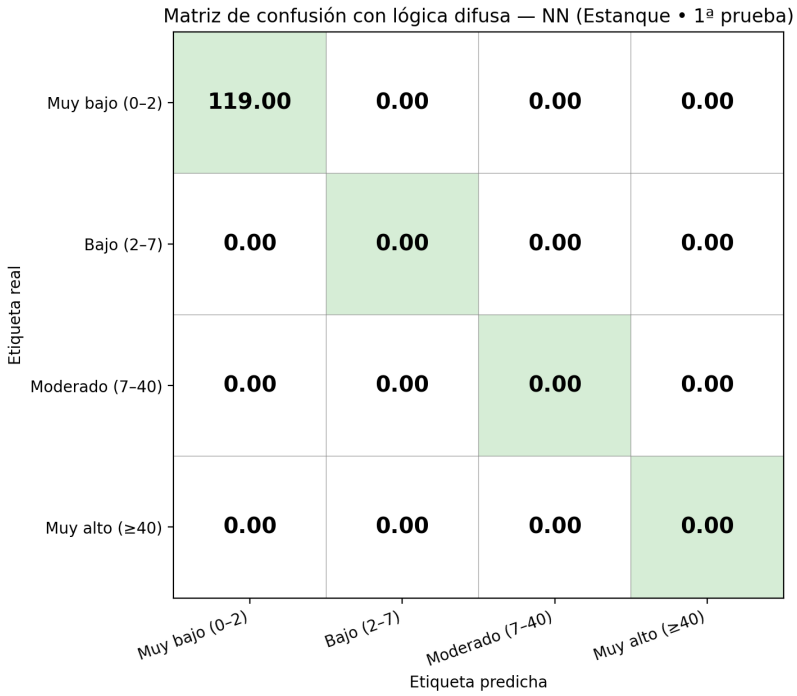
Nota. Los valores representan grados de pertenencia acumulados en cada categoría trófica. Elaboración propia.

Figura 57. Matriz de confusión con lógica difusa para el modelo KNN entrenado con CEA, aplicado a la primera prueba piloto del estanque.



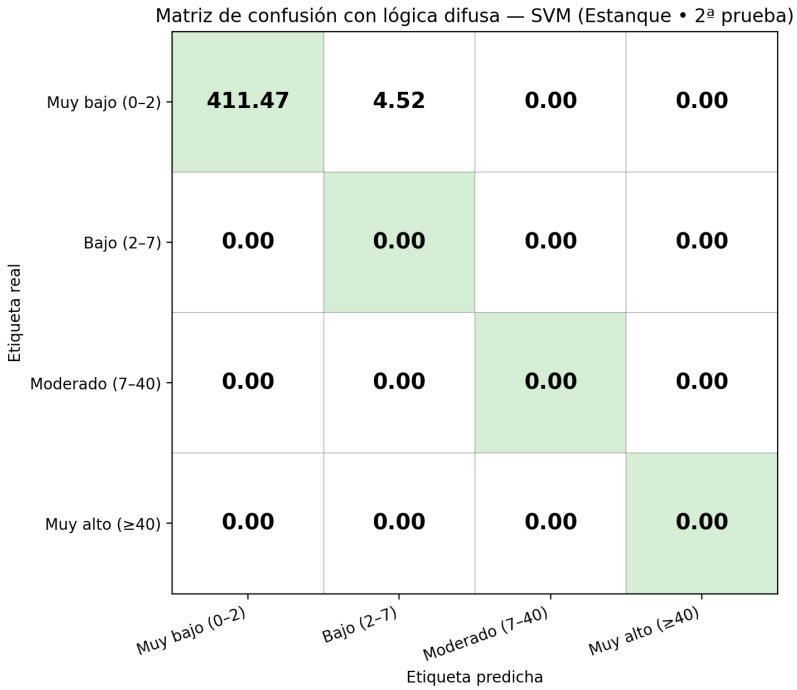
Nota. Clasificación difusa en cuatro categorías tróficas. Elaboración propia.

Figura 58. Matriz de confusión con lógica difusa para la red neuronal profunda entrenada con CEA, aplicada a la primera prueba piloto del estanque.



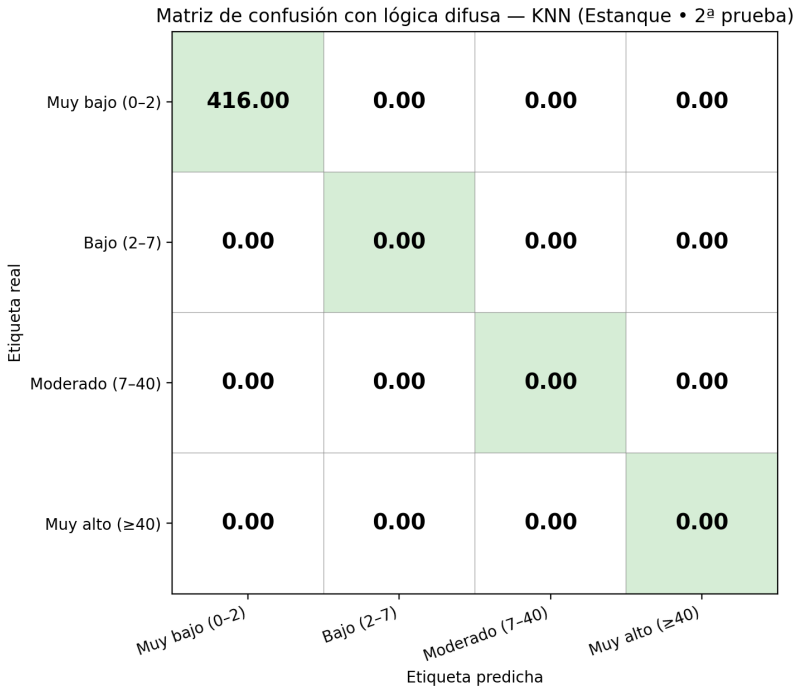
Nota. La red neuronal concentra prácticamente toda la pertenencia en la categoría Muy bajo. Elaboración propia.

Figura 59. Matriz de confusión con lógica difusa para el modelo SVM entrenado con CEA, aplicado a la segunda prueba piloto del estanque.



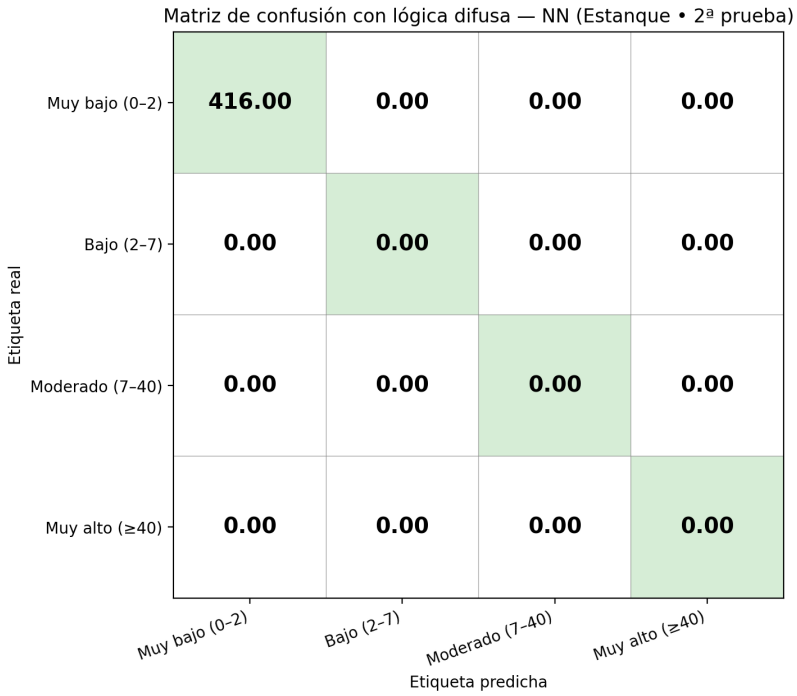
Nota. Con modelo SVM, se clasificó la mayoría de datos como Muy bajo. Elaboración propia.

Figura 60. Matriz de confusión con lógica difusa para el modelo KNN entrenado con CEA, aplicado a la segunda prueba piloto del estanque.



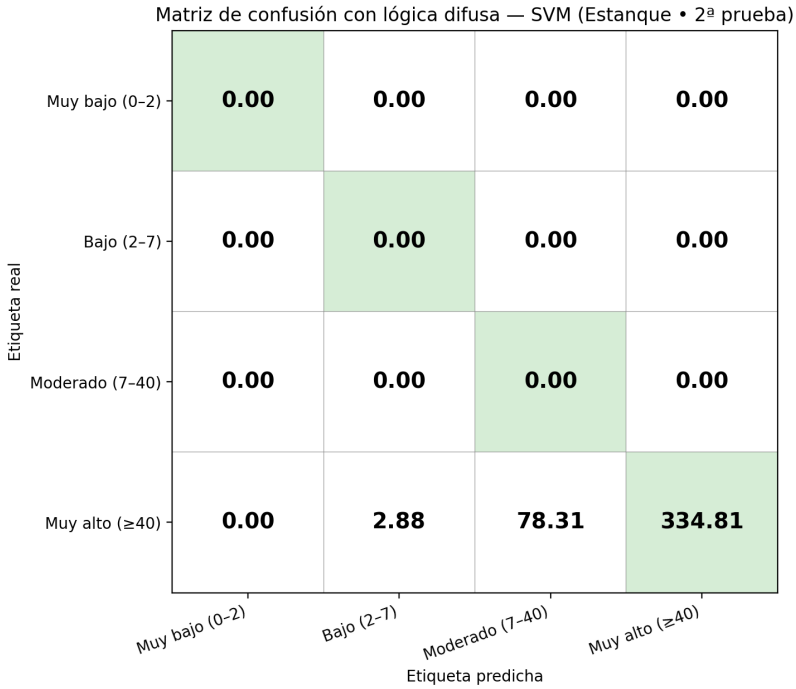
Nota. Con modelo KNN, se clasificó la mayoría de datos como Muy bajo.

Figura 61. Matriz de confusión con lógica difusa para la red neuronal profunda entrenada con CEA, aplicada a la segunda prueba piloto del estanque.



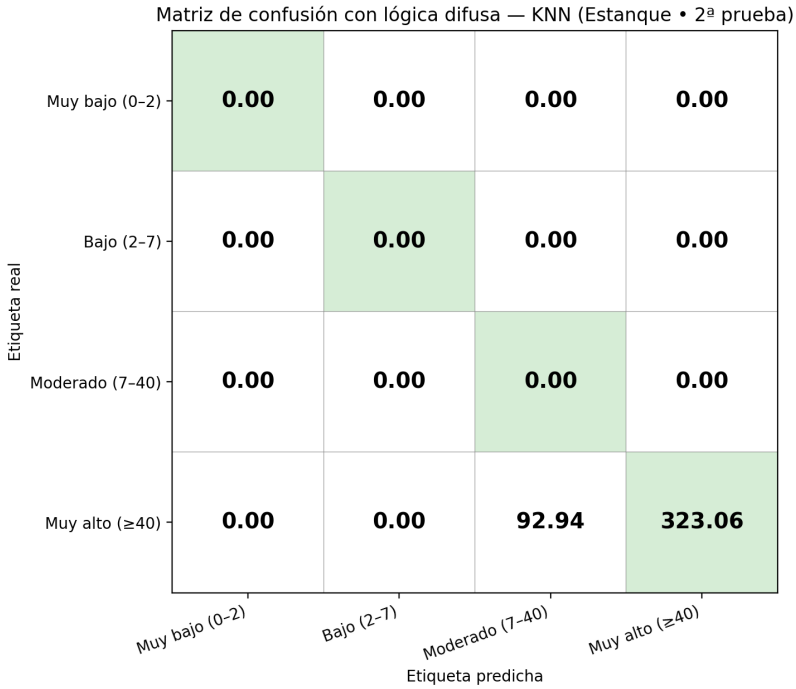
Nota. La red neuronal concentra prácticamente toda la pertenencia en la categoría Muy bajo. Elaboración propia.

Figura 62. Matriz de confusión con lógica difusa para el modelo SVM con datos del estanque en la segunda prueba piloto.



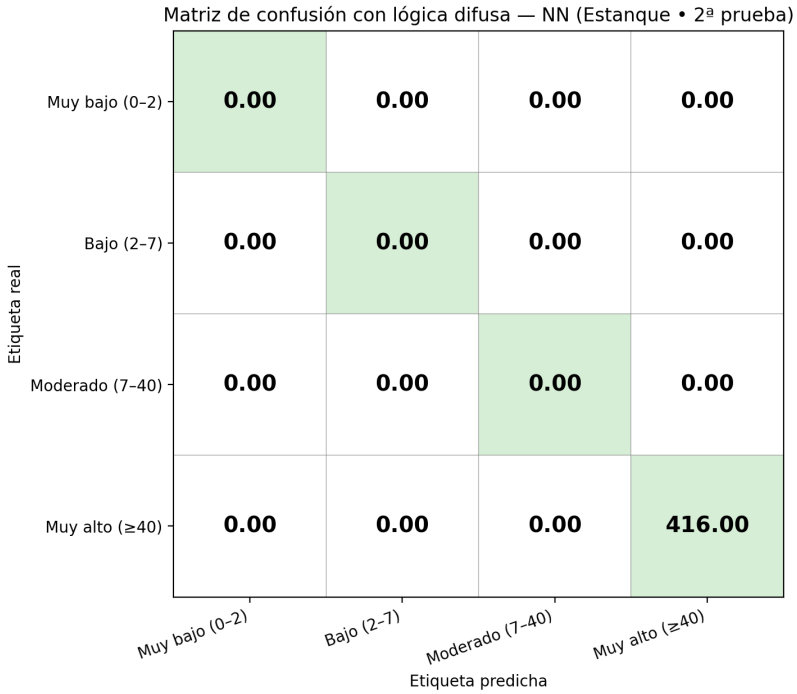
Nota. El modelo SVM clasificó la mayoría de datos como Muy alto. Elaboración propia.

Figura 63. Matriz de confusión con lógica difusa para el modelo KNN con datos del estanque en la segunda prueba piloto.



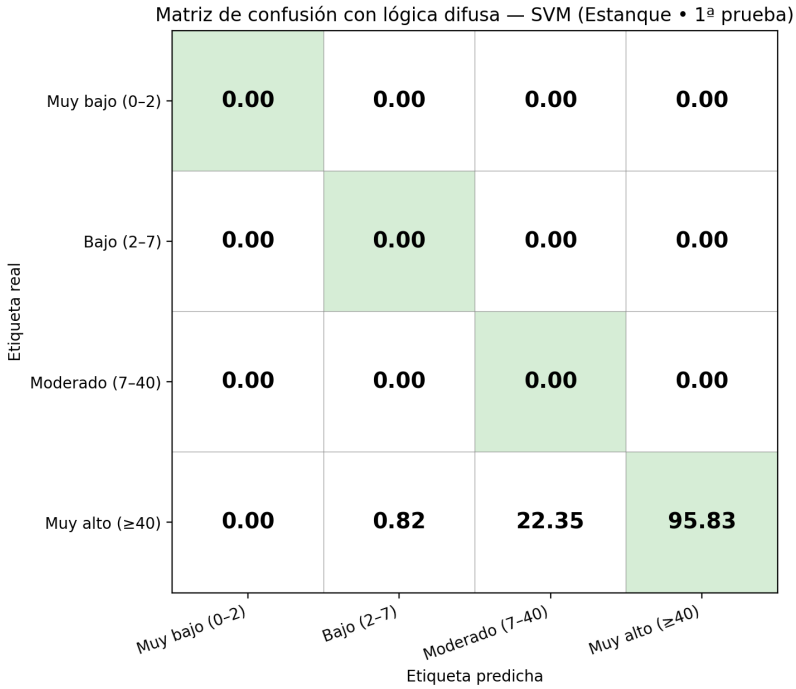
Nota. El modelo KNN mantiene una clasificación consistente en la categoría Muy alto. Elaboración propia.

Figura 64. Matriz de confusión con lógica difusa para la red neuronal profunda con datos del estanque en la segunda prueba piloto.



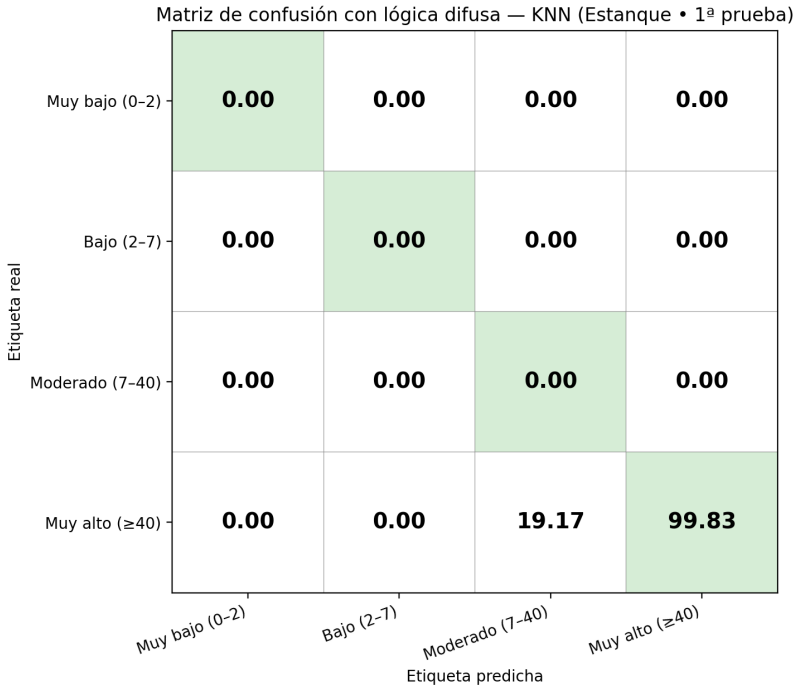
Nota. La red neuronal asignó la totalidad de la pertenencia a la categoría Muy alto. Elaboración propia.

Figura 65. Matriz de confusión con lógica difusa para el modelo SVM con datos del estanque en la primera prueba piloto.



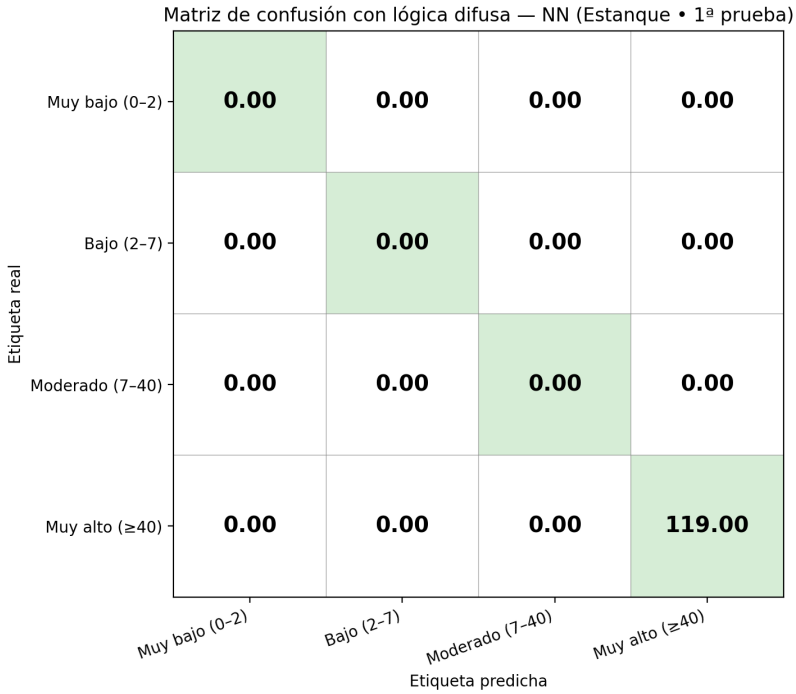
Nota. Con el modelo SVM, se clasificó la mayoría de datos como Muy alto. Elaboración propia.

Figura 66. Matriz de confusión con lógica difusa para el modelo KNN con datos del estanque en la primera prueba piloto.



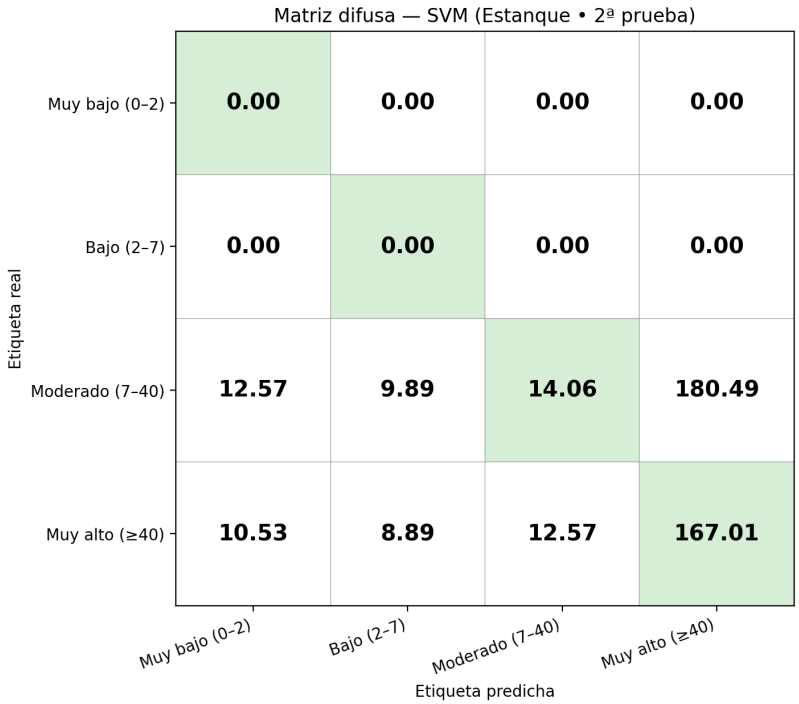
Nota. Con el modelo KNN, se clasificó la mayoría de datos como Muy alto. Elaboración propia.

Figura 67. Matriz de confusión con lógica difusa para la red neuronal profunda con datos del estanque en la primera prueba piloto.



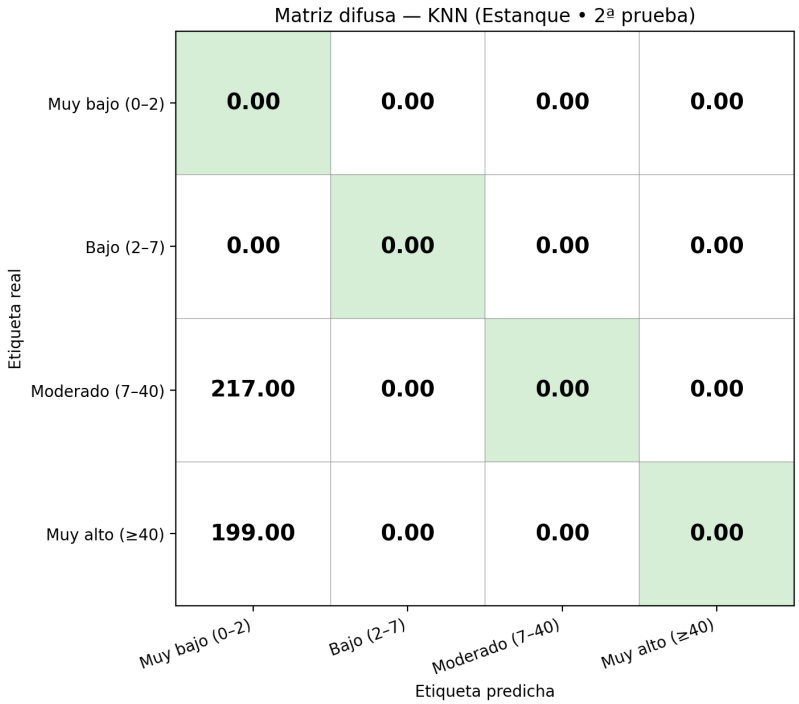
Nota. La red neuronal asignó prácticamente toda la pertenencia a la categoría Muy alto. Elaboración propia.

Figura 68. Matriz de confusión con lógica difusa para el modelo SVM entrenado con AMSA y CEA, aplicado a la segunda prueba piloto del estanque.



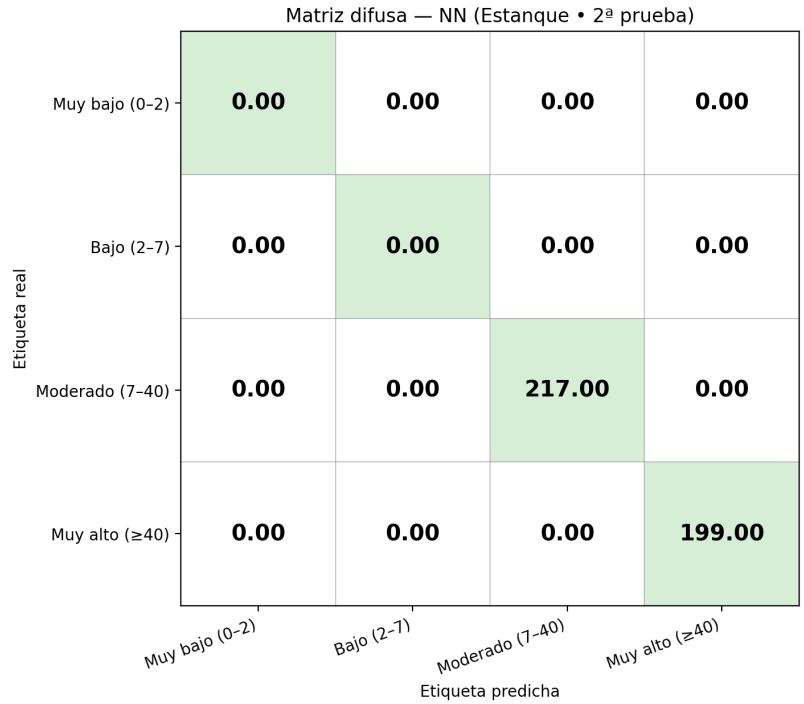
Nota. En la segunda prueba piloto, el SVM identifica patrones más variados, distribuyendo pertenencia entre Moderado y Muy alto, reflejando un ecosistema ligeramente más heterogéneo en sus condiciones fisicoquímicas. Elaboración propia.

Figura 69. Matriz de confusión con lógica difusa para el modelo KNN entrenado con AMSA y CEA, aplicado a la segunda prueba piloto del estanque.



Nota. Las pertenencias se concentran en Moderado y Muy alto, evidenciando que el KNN identificó transiciones sutiles sin clasificar observaciones en rangos bajos. Elaboración propia.

Figura 70. Matriz de confusión con lógica difusa para el modelo de redes neuronales profundas entrenado con AMSA y CEA, aplicado a la segunda prueba piloto del estanque.



Nota. La NN reconoce dos regímenes tróficos dominantes: Moderado y Muy alto, reflejando una transición más compleja en la segunda prueba piloto. Elaboración propia.

13.7. Tablero digital interactivo como herramienta de interpretación

Para facilitar el análisis de los modelos y permitir una exploración transparente de los resultados, se desarrolló un tablero digital interactivo basado en *Streamlit*. Este tablero integra las bases de datos, los modelos entrenados, las curvas de entrenamiento, las matrices de confusión difusas y las predicciones aplicadas a los datos del estanque. Su objetivo es ofrecer una plataforma accesible, reproducible y científicamente robusta para investigadores, gestores ambientales y personal técnico.

Figura 71. Tablero digital interactivo empleado para la visualización y análisis de los modelos de aprendizaje automático desarrollados en este trabajo.

Nota. El tablero permite explorar predicciones, matrices difusas, curvas de entrenamiento y comparaciones entre modelos entrenados con AMSA, CEA y AMSA+CEA. Elaboración propia.

CAPÍTULO 14

Glosario

Cloroplastos (p. 9): Orgánulos celulares presentes en células vegetales y algas, encargados de realizar la fotosíntesis. Contienen membranas tilacoides y pigmentos como la clorofila y carotenoides.

Disco de Secchi (p. 10): Instrumento circular utilizado en limnología para medir la transparencia del agua. Se sumerge hasta que deja de ser visible, para estimar así la turbidez y la penetración de la luz.

Eutrófico (p. 10): Cuerpos de agua ricos en nutrientes (nitrógeno y fósforo), con elevada productividad primaria. Suelen presentar proliferaciones de algas y menor transparencia.

Ficobiliproteína (p. 9): Pigmento accesorio presente en cianobacterias y algas rojas, que capta energía lumínica en longitudes de onda que la clorofila no absorbe eficientemente. Ejemplos: ficocianina y ficoeritrina.

Fotótrofas (p. 9): Microorganismos que utilizan la energía de la luz como fuente principal para su metabolismo. Pueden ser oxigénicas (liberan oxígeno, como las cianobacterias) o anoxigénicas (no producen oxígeno).

Heterogeneidades (p. 10): Diferencias o variaciones dentro de un sistema. En ecología acuática, se refiere a la variabilidad espacial o temporal de parámetros como nutrientes, temperatura o turbidez en un cuerpo de agua.

Hipereutrófico (p. 10): Estado de extrema riqueza en nutrientes en un ecosistema acuático, asociado a floraciones masivas de algas y cianobacterias, pérdida de oxígeno disuelto y riesgo de mortandad de fauna acuática.

Membrana tilacoide (p. 9): Estructura interna del cloroplasto formada por sacos aplanados (tilacoides) donde se localizan los pigmentos fotosintéticos como la clorofila. Es el sitio principal de las reacciones fotoquímicas de la fotosíntesis.

Mesotrófico (p. 10): Estado intermedio de nutrientes en un cuerpo de agua. Presentan una productividad moderada y cierta turbidez, con diversidad biológica equilibrada.

Oligotrófico (p. 10): Lagos o cuerpos de agua con baja concentración de nutrientes, alta transparencia y baja productividad biológica. Generalmente, presentan aguas limpias y buena calidad para consumo.

Ultraoligotrófico (p. 10): Estado trófico de cuerpos de agua con muy baja concentración de nutrientes y producción primaria mínima. Se caracterizan por aguas muy claras y pobres en fitoplancton.