
**Universidad de Buenos Aires**

**Facultad de Ciencias Exactas y Naturales**

**Maestría en Explotación de Datos y Descubrimiento de Conocimiento**

# Proyección de la Calidad del Agua en el Embalse de Salto Grande

Esp. Joaquín Sebastian Tschopp

**Año 2025**

# Índice general

## Resumen

Este trabajo aborda el desafío de proyectar la calidad del agua en el embalse de Salto Grande, con foco en la playa Las Palmeras, a fin de apoyar decisiones sobre su uso recreativo. Dada la creciente presión sobre los recursos hídricos y la ocurrencia de floraciones de cianobacterias, se buscó anticipar el estado del agua con un horizonte de siete días, integrando datos obtenidos in situ, satelitales, meteorológicos e hidrológicos.

Se aplicaron modelos de aprendizaje automático multiclase, con foco en Random Forest y LightGBM, evaluados mediante validación walk-forward. Se desarrolló un preprocesamiento específico para interpolar variables críticas de baja frecuencia y se diseñó un esquema de cálculo del target basado en los umbrales de la OMS para riesgo sanitario.

Los modelos no lineales entrenados mostraron alto desempeño, alcanzando Random Forest el mejor resultado (F1-score macro de 0,9491; exactitud de 0,9333) y lograron identificar con alta precisión las clases más críticas. Las variables más relevantes fueron de tipo biológico y fisicoquímico (cianobacterias, clorofila, fósforo total, nitrógeno total), en línea con lo observado en otros estudios sobre el papel de los nutrientes en la aparición de floraciones. El estudio demuestra que es factible integrar modelos predictivos en esquemas de monitoreo operativo, aunque se evidencian limitaciones en la cobertura y frecuencia de los datos que deben ser abordadas en trabajos futuros.



**Palabras clave:** calidad de agua, cianobacterias, predicción semanal, machine learning, embalse Salto Grande.

# Capítulo 1

# Introducción

## 1.1. Contexto y motivación científica

La calidad del agua es esencial para el consumo humano y para actividades recreativas, especialmente cuando se requiere evaluar la factibilidad de tratamiento y uso del recurso. En un contexto en el que la demanda de recursos hídricos es creciente, determinar la viabilidad de utilizar el agua en el corto plazo resulta crítico para la gestión y la prevención de riesgos.1.1a La aplicación de técnicas de Machine Learning, con datos provenientes de diversas fuentes (satelitales, in situ, meteorológicos, hídricos), puede aportar una herramienta de alerta temprana para optimizar la toma de decisiones. La respuesta a esta pregunta contribuirá a la implementación de estrategias de monitoreo y a la mejora en la gestión del recurso hídrico en diferentes puntos del Lago de Salto Grande.

(a) Ejemplo de fauna. (b) Otro ejemplo de fauna.

Figura 1.1: Imágenes de fauna en áreas con algas. Tomadas por Juan Menoni.

## 1.2. Objetivo y Pregunta

### 1.2.1. Objetivo

Desarrollar un modelo de proyección semanal que apoye la toma de decisiones sobre el uso del agua en un horizonte de 7 días, destinado a actividades recreativas (como bañarse). El modelo integrará información satelital, mediciones in situ, y datos meteorológicos e hidrológicos, aplicado a la playa Las Palmeras.

### 1.2.2. Pregunta de Investigación

¿Es factible proyectar el estado de la calidad del agua en el embalse de Salto Grande, a partir de los datos disponibles y mediante herramientas de predicción basadas en Machine Learning, considerando como referencia las recomendaciones de la Organización Mundial de la Salud sobre calidad de agua (World Health Organization, 2021)?

## 1.3. Estructura del documento

El presente trabajo se encuentra organizado en siete capítulos, los cuales abordan de manera progresiva el desarrollo del estudio:

- El **Capítulo 1** introduce el problema de investigación, presenta la motivación científica, define los objetivos del trabajo y describe la organización general del documento.
- El **Capítulo 2** recopila y analiza los principales antecedentes y fundamentos teóricos relevantes para el problema abordado. Se describen trabajos previos, marcos normativos y conceptos de ciencia de datos utilizados.
- El **Capítulo 3** describe detalladamente la metodología empleada: fuentes y estructura de los datos, técnicas de preprocesamiento, interpolación, definición del objetivo de predicción (*target*), y los modelos de Machine Learning seleccionados, junto con sus criterios de evaluación.
- El **Capítulo 4** presenta los resultados obtenidos a partir del entrenamiento y evaluación de los modelos predictivos. Se incluyen análisis comparativos entre algoritmos y se discuten las limitaciones y posibles mejoras del enfoque adoptado.
- El **Capítulo 5** resume los principales hallazgos, extrae conclusiones generales en relación con los objetivos planteados y sugiere posibles líneas futuras de investigación o mejora.
- El **Capítulo 6** incluye la bibliografía utilizada y citada a lo largo del trabajo, siguiendo el estilo APA.
- Finalmente, el **Capítulo 7** incorpora anexos complementarios que contiene referencias al código fuente utilizado en el desarrollo del proyecto.

# Capítulo 2

# Marco teórico

## 2.1. Cianobacterias y calidad de agua

Las cianobacterias son microorganismos fotosintéticos que prosperan en aguas cálidas y ricas en nutrientes; bajo ciertas condiciones forman floraciones masivas (*blooms*) capaces de liberar toxinas nocivas. Su proliferación depende, sobre todo, de la disponibilidad de nitrógeno (N) y fósforo (P) y de la estabilidad de la columna de agua.

**Factores ambientales.** Igwaran y cols. (2024) presentan una revisión sistemática de más de 200 investigaciones alrededor del mundo, en el cual identifican los aportes de nitrógeno y fósforo, especialmente provenientes de la agricultura y descargas residuales, como el motor principal de los blooms de cianobacterias, mientras que la temperatura actúa como un factor amplificador, extendiendo la duración de estas proliferaciones. De manera análoga, Rigosi y cols. (2014) analizan datos de una extensa evaluación hídrica de más de 1000 lagos en Estados Unidos y concluyen que, en general, los nutrientes tienen un peso mucho mayor que la temperatura en explicar la variación del biovolumen de cianobacterias. Además, su revisión resalta que el estado trófico modula esta relación: los nutrientes dominan en lagos oligotróficos, mientras que el efecto de la temperatura se vuelve relativo en sistemas mesotróficos y eutróficos, donde la interacción nutrientes–temperatura llega a ser relevante.

**Implicancias para la salud.** Las toxinas liberadas (microcistinas, cilindrospermopsina, anatoxinas) pueden provocar desde irritaciones dérmicas hasta hepatotoxicidad y fallos neuromusculares. Un reciente análisis conjunto de casos clínicos (Backer y Landsberg, 2024) identifica más de 355 episodios de enfermedad humana asociados a recreación en aguas contaminadas, destacando que la inhalación de aerosoles y la ingestión accidental son las vías de exposición más comunes.

Estos hallazgos sostienen la importancia de vigilar nutrientes, temperatura y condiciones hidrodinámicas para predecir y mitigar los blooms, eje central de nuestro modelo de proyección semanal.

## 2.2. Antecedentes de implementación de modelos predictivos de calidad de agua

Diversos estudios han abordado la problemática de la calidad del agua mediante herramientas de aprendizaje automático y observaciones satelitales. Schaeffer y cols. (2024), por ejemplo, presenta una revisión exhaustiva de modelos de predicción aplicados a imágenes de sensores remotos, mostrando el potencial de esta fuente para estimar variables como clorofila-*a*, turbidez y biovolumen de fitoplancton.

En el estudio de Rodríguez-López y cols. (2023), se evaluó el desempeño de distintos algoritmos de aprendizaje automático para la estimación de parámetros de calidad de agua, en particular clorofila-*a* en el Lago Llanquihue, en el sur de Chile. El trabajo integró 31 años de datos históricos *in situ*, para entrenar modelos con muy buen desempeño, como XGBoost, LightGBM y AdaBoost.

Sin embargo, no todos los trabajos analizados integran datos ambientales correspondientes al momento exacto de las mediciones, ni incorporan variables *in situ* que reflejen condiciones hidrobiológicas o químicas locales. Tampoco se consideran pronósticos meteorológicos de corto plazo como insumo en los modelos predictivos.

## 2.3. Indicadores de calidad en cuerpos de agua con fines recreativos

Los sistemas de alerta para aguas de recreo se basan, por lo general, en tres indicadores fáciles de medir (World Health Organization, 2021):

- **Concentración de cianobacterias** ($mm^3/mL$)
- **Clorofila-*a*** (µg/L),
- **Transparencia disco Secchi** (m) como indicador turbidez.

**Muestreo in-situ vs. teledetección.** Mientras que el muestreo de campo ofrece precisión puntual, la teledetección multiplica la cobertura espacial y temporal. Stumpf y cols. (2016) muestran que algoritmos satelitales (MERIS, actualizados a Sentinel-3) detectan concentraciones de cianobacterias con una sensibilidad del 84 %, permitiendo disparar alertas con varios días de antelación.

**Sistemas comparativos.** Hunter y cols. (2022) revisan la implementación de marcos de alerta en doce países, concluyendo que el esquema OMS de tres niveles es el más extendido y el que mejor balancea coste–beneficio para la gestión recreativa en lagos y embalses.

Estos antecedentes sustentan el uso combinado de mediciones *in situ* y satelitales en nuestro modelo de proyección semanal.

# Capítulo 3

# Metodología

## 3.1. Presentación y descripción de los datos utilizados

### 3.1.1. Conjunto de Datos

**Fuentes de obtención**

- **Comisión Técnica Mixta de Salto Grande:** Datos *in situ* (mediciones directas en el lugar de estudio y resultados de laboratorio de muestras tomadas en ese sitio), junto con registros hidrometeorológicos.
- **Imágenes satelitales:** Sentinel-2 y Landsat 8/9.3.1
- **Estaciones meteorológicas:** Lluvia, temperatura, viento.
- **Modelos meteorológicos:** Pronóstico de precipitaciones y temperatura, a 1–6 días.

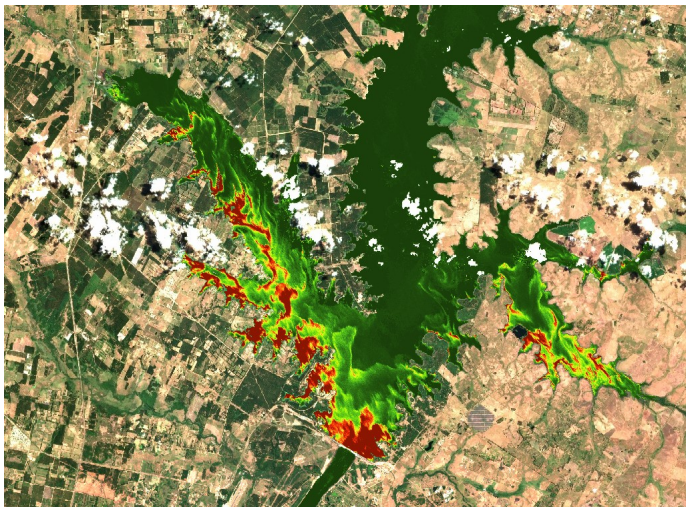


Figura 3.1: Imagen satelital que muestra la presencia de cianobacterias en el embalse. Las zonas en rojo indican áreas con alta concentración de floraciones (*blooms*).

**Estructura de los datos**

A continuación se describen brevemente los atributos del dataset, sus tipos de dato y su significado.

Cuadro 3.1: Descripción de los atributos y sus tipos en el dataset.

| **Atributo** | **Tipo** | **Descripción** |
|---|---|---|
| time | `datetime` | Identificación del registro (fecha y hora) |
| **Estación Meteorológica** | | |
| T.obs | `float64` | Temp. media observada (se toma la media) |
| Hr.obs | `float64` | Humedad relativa (media) |
| pr.obs | `float64` | Presión atmosférica (media) |
| Rad.obs | `float64` | Radiación solar (media) |
| u.obs | `float64` | Velocidad del viento (media) |
| u_gr.obs | `float64` | Dirección del viento (grados: 0=Norte, 90=Este, 180=Sur, 270=Oeste) |
| Td.obs | `float64` | Temperatura de punto de rocío |
| EPenman.obs | `float64` | Evapotranspiración (método Penman) |
| pp_obs | `float64` | Precipitación observada diaria (mm) |
| **Pronóstico meteorológico** | | |
| prono_PP_dia_[0–6] | `float64` | Pronóstico de precipitación diaria (mm) |
| prono_Tm_dia_[0–7] | `float64` | Pronóstico de temperatura media diaria (°C) |
| **Variables hidrológica** | | |
| Nivel_lago | `float64` | Nivel del embalse (m) |
| Caudal_lago | `float64` | Caudal de aporte ($m^3/s$) |
| **Mediciones in-situ** | | |
| Clorof_A | `float64` | Concentración de clorofila-a ($\mu$g/L) |
| Enteroc | `float64` | Recuento de enterococos (NMP/100 mL) |
| NO2_lab | `float64` | Nitrito (mg/L) |
| NO3_lab | `float64` | Nitrato (mg/L) |
| NT | `float64` | Nitrógeno total (mg/L) |
| Ptot_lab | `float64` | Fósforo total (mg/L) |
| SS_105 | `float64` | Sólidos suspendidos medidos a 105°C (mg/L) |
| ODsupC | `float64` | Oxígeno disuelto en superficie corregido (mg/L) |
| TempAmb | `float64` | Temperatura ambiente (°C) |
| TurSupC | `float64` | Turbidez en superficie corregida (NTU) |
| PO4_lab | `float64` | Fosfato reactivo medido en laboratorio (mg/L) |
| SS_550 | `float64` | Sólidos suspendidos medidos a 550°C (mg/L) |
| NH_lab | `float64` | Amonio medido en laboratorio (mg/L) |
| MCys_spp | `float64` | Abundancia de Microcystis spp. (org/mL) |

*Continúa en la siguiente página*

| Atributo | Tipo | Descripción |
|---|---|---|
| VientoI | `float64` | Intensidad del viento (km/h) |
| Cianobac | `float64` | Recuento total de cianobacterias (org/mL) |
| T_sup | `float64` | Temperatura del agua en superficie (°C) |
| CondsupC | `float64` | Conductividad eléctrica sup. corregida (µS/cm) |
| MicrtEli | `float64` | Total Microcystins ELISA µg/L |
| EscColi | `float64` | Recuento de Escherichia coli (NMP/100 mL) |
| PROFSITI | `float64` | Profundidad de sitio de muestreo (m) |
| SatSupC | `float64` | Saturación de oxígeno en sup. corregida ( %) |
| fitotot | `float64` | Biomasa total de fitoplancton (µg/L) |
| pHsupC | `float64` | pH en superficie corregido (adimensional) |
| Secc_sup | `float64` | Profundidad Secchi en superficie (m) |
| Target | `float64` | Nivel de alerta OMS: 0=Sin Riesgo, 1=Vigilancia, 2=Alerta 1, 3=Alerta 2 |

**Relevancia ecológica de las variables**

La selección de variables predictoras responde a criterios tanto metodológicos como ecológicos, en relación con los procesos que determinan la calidad del agua en sistemas lénticos. A continuación se destacan los aportes de cada grupo:

- **Estación meteorológica:** La temperatura, humedad relativa, presión atmosférica, radiación y viento influyen en la dinámica de estratificación térmica, la tasa de evaporación y la disponibilidad de oxígeno en el agua. En particular, el viento no solo afecta la mezcla de la columna de agua, sino que también favorece la concentración de floraciones en zonas costeras.
- **Pronóstico meteorológico:** Los valores proyectados de precipitación y temperatura permiten anticipar cambios en el balance hídrico y en la carga de nutrientes, lo que puede desencadenar o intensificar episodios de floraciones de cianobacterias en los días siguientes.
- **Variables hidrológicas:** El nivel del embalse y el caudal de aporte determinan la renovación y circulación del agua, influyendo en la concentración de nutrientes y organismos. Estos parámetros pueden ayudar a anticipar variaciones en el corto plazo que afecten la permanencia o dispersión de floraciones.
- **Mediciones in situ:** Parámetros como clorofila-a, abundancia de cianobacterias, microcistinas y biomasa fitoplanctónica reflejan directamente el estado de las floraciones. Otros, como fósforo total, nitrógeno total y compuestos disueltos, aportan información sobre la disponibilidad de nutrientes que sostienen su desarrollo. Finalmente, variables físico-químicas como oxígeno disuelto, turbidez, conductividad o pH

describen el estado actual del cuerpo de agua, que puede favorecer tanto el inicio como la persistencia de un bloom si ya está presente.

En conjunto, estos grupos de variables constituyen un marco integral de predicción, al abarcar los nutrientes que alimentan las floraciones, el estado actual de las mismas y las condiciones ambientales que pueden favorecer su aparición, concentración o persistencia.

## 3.2. Preprocesamiento y limpieza de los datos

### 3.2.1. Integración de datos satelitales y observaciones in-situ

En la integración de datos satelitales e in-situ se priorizó la fidelidad de la medición directa sin renunciar a la cobertura espacio-temporal adicional que ofrecen los productos remotos. Para cada fecha se buscó la coincidencia más cercana entre ambas fuentes; cuando la diferencia temporal fue <= 1 día se seleccionó la observación in-situ y el valor satelital quedó sólo como respaldo. En este proceso, los atributos considerados de manera conjunta fueron clorofila, anomalía de clorofila, cianobacterias, temperatura superficial del agua (SST) y turbidez, que constituyen la totalidad de los disponibles en registros satelitales y que, a su vez, también cuentan con mediciones in-situ. La Fig. 3.2 muestra que los datos satelitales (azul) acompañan la tendencia general de `Cianobac`, `Clorof_A` y `T_sup`, pero son más dispersos y, en varios picos, sobrestiman los valores medidos en campo (rojo). De haberse empleado exclusivamente la fuente satelital, se habrían aceptado estimaciones razonables pero con varianzas hasta un orden de magnitud superiores, lo que podría introducir ruido en la aumentación por interpolación. Por ello, los datos satélitales se utilizaron únicamente para rellenar ventanas sin registro directo, salvaguardando la coherencia de la serie temporal y minimizando la propagación de errores en las variables claves.

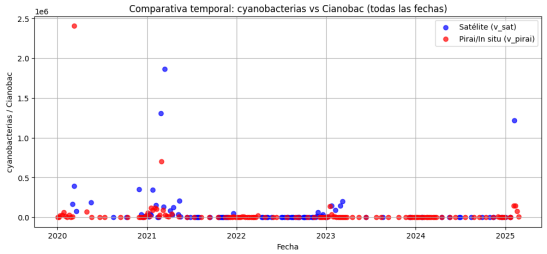
### 3.2.2. Aumentación por interpolación

El conjunto de datos utilizado presenta distintas frecuencias de medición según la variable considerada. Para poder entrenar modelos de aprendizaje automático con entradas diarias consistentes, fue necesario completar ciertas series temporales mediante técnicas de interpolación, preservando al mismo tiempo la estructura original de la información. A continuación se detallan los métodos aplicados según la naturaleza de los datos.

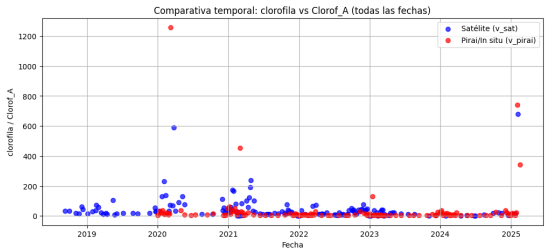
#### Interpolación funcional en variables biológicas

Las variables biológicas claves (biovolumen de cianobacterias, clorofila-*a*, fitoplancton total, etc.) fueron medidas con baja frecuencia (una vez por semana en verano y una vez al mes en invierno), lo que imposibilita un entrenamiento robusto. Para resolver esta limitación, se aplicó una estrategia de interpolación funcional basada en el trabajo de (Oh y cols., 2020), con el siguiente procedimiento:

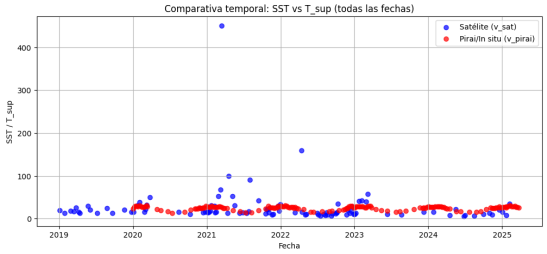
1. **Segmentación temporal**: la serie se divide en ventanas fijas.




(a) Cyanobacterias vs `Cianobac`




(b) Clorofila vs `Clorof_A`




(c) SST vs `T_sup`

Figura 3.2: Boxplots comparativos entre mediciones satelitales (azul) e in situ/laboratorio (rojo) para cuatro variables clave. Las divergencias confirman que, cuando existe registro directo cercano en el tiempo, éste resulta más representativo que la estimación satelital.

2. **Ajuste de funciones suaves**: en cada ventana se ajusta una spline cúbica natural que minimiza la curvatura entre puntos reales, forzando a que la curva pase exactamente por los valores observados.
3. **Muestreo denso**: la spline se evalúa a paso diario, generando valores sintéticos que respetan la tendencia local y preservan continuidad.
4. **Unión de ventanas**: se superponen días en los bordes para garantizar suavidad entre segmentos.
5. **Truncamiento de valores**: al finalizar el proceso, los valores generados que exceden el máximo observado o descienden por debajo de cero son truncados, evitando así la aparición de outliers superiores a los reales y de valores negativos artificiales.

El resultado es una serie diaria suave que conserva la forma global del fenómeno observado, introduce variación controlada útil para el aprendizaje supervisado, y garantiza que los valores sintéticos sean físicamente plausibles.

### Interpolación lineal en variables hidrológicas

Las variables hidrológicas, `Nivel_lago` y `Caudal_lago`, se encuentran disponibles con frecuencia semanal, representando valores promedio. Para completar las fechas intermedias entre mediciones y armonizar la frecuencia con el resto del dataset, se aplicó interpolación lineal entre puntos consecutivos, siguiendo la implementación disponible en `scipy.interpolate` (Virtanen y cols., 2020).

Dado que los valores originales ya representan medias semanales, la interpolación lineal entre ellos resulta una estrategia adecuada para generar estimaciones coherentes.

### 3.2.3. Cálculo del *target*

Los umbrales se toman de la OMS (World Health Organization, 2021), ajustados a las variables disponibles en nuestro registro (no se tuvieron en cuenta las mediciones de toxinas).

- **Sin Riesgo (0)**
  - Biovolumen $\leq 1\,\text{mm}^3/\text{L}$, **o**
  - Clorofila-$a \leq 3\,\mu\text{g/L}$ con dominancia de cianobacterias, **o**
  - Transparencia Secchi $> 2$ m.
- **Vigilancia (1)**
  - Biovolumen $\in [1,\ 4]\,\text{mm}^3/\text{L}$, **o**
  - Clorofila-$a \in [3,\ 12]\,\mu\text{ug/L}$ con dominancia de cianobacterias, **o**
  - Transparencia Secchi $\in [1,\ 2]$ m.
- **Alerta1 (2)**
  - Biovolumen $\in [4,\ 8]\,\text{mm}^3/\text{L}$, **o**
  - Clorofila-$a \in [12,\ 24]\,\mu\text{ug/L}$ con dominancia de cianobacterias, **o**
  - Transparencia Secchi $\in [0{,}5,\ 1]$ m.
- **Alerta2 (3)**
  - Presencia de *scum* de cianobacterias **o**
  - Clorofila-$a > 24\,\mu\text{ug/L}$, **o**
  - Transparencia Secchi $< 0{,}5$ m.

**Conversión de unidades mediante regresión lineal.** La variable correspondiente a la cianobacteria fue transformada a $\text{mm}^3/\text{L}$ utilizando un modelo de regresión lineal entrenado con datos de años en que se registraron ambas magnitudes. La ecuación resultante fue:

$$\text{BVCIANsp} = 0{,}0001 \times \text{Cianobac} + 0{,}2006$$

La decisión de emplearlo fue tomada conjuntamente con el personal encargado de la recolección de muestras y de la interpretación de los atributos.

Esta transformación garantiza la compatibilidad de las unidades derivadas con los umbrales de la OMS para biovolumen de cianobacterias ($\text{mm}^3/\text{L}$) y preserva la calidad interpretativa de los datos.

**Asignación del target a futuro.** Para modelar la predicción de la calidad del agua con una anticipación de 7 días, se asignó a cada registro el valor del target correspondiente a 7 días después. Así, al entrenar el modelo con los datos disponibles hasta una fecha determinada, se busca predecir el estado de la calidad del agua una semana más adelante

(DataCamp, 2022). Por ejemplo, si se realiza una predicción utilizando los datos del 01-01-2025, el modelo estará estimando el estado correspondiente al 08-01-2025. Esta estrategia permite que el modelo aprenda a anticipar condiciones futuras basándose únicamente en información pasada, respetando la secuencia temporal de los datos.

Sin embargo, debido a la frecuencia irregular de las observaciones reales, no fue posible construir un target uniforme con un horizonte fijo de 7 días sin recurrir a interpolación. En muchos casos, la distancia entre una observación y la siguiente superaba los 7 días (por ejemplo, 10 o 30 días), lo que impedía definir con precisión el estado futuro deseado. Esta situación hacía inviable el entrenamiento del modelo exclusivamente con datos reales, ya que se perdería una porción sustancial del conjunto de entrenamiento. Por este motivo, se optó por aplicar interpolaciones controladas para completar la secuencia temporal y permitir la construcción de un target coherente, aún reconociendo los riesgos que esto implica en términos de sobreajuste y fidelidad de los patrones aprendidos.

### 3.2.4. Filtrado de registros con datos faltantes

Durante el preprocesamiento de los datos, se eliminaron los registros iniciales y finales que presentaban valores faltantes en las variables hidrológicas, meteorológicas y biológicas. Esta decisión se tomó para asegurar que los conjuntos de entrenamiento y prueba contuvieran únicamente datos reales de calidad del agua, evitando así la influencia de valores interpolados en los extremos de la serie temporal.

## 3.3. Descripción de las técnicas de análisis y modelado

### 3.3.1. Validación y búsqueda de hiperparámetros con Walk-Forward

Con el objetivo de seleccionar los mejores hiperparámetros para los modelos predictivos utilizados, se implementó una estrategia de validación basada en la técnica Walk-Forward, la cual permite una evaluación realista del desempeño del modelo en un contexto temporal.

**Búsqueda de hiperparámetros.**

Para cada combinación de hiperparámetros, se realizó una serie de 60 entrenamientos consecutivos. En cada iteración:

Se utilizó una ventana de entrenamiento desde el inicio del data set hasta el inicio del conjunto de validación.

Se entrenó el modelo y se realizó la predicción del día siguiente, el primero de los 60 reservados para validación.

El día predicho se incorporó al conjunto de entrenamiento, desplazando la ventana una unidad hacia adelante (rolling window).

Este procedimiento se repitió hasta completar la predicción de 60 días consecutivos. La métrica utilizada para comparar combinaciones de hiperparámetros fue el accuracy obtenido en esas 60 predicciones.

**Evaluación final con conjunto de prueba.**

Una vez seleccionada la mejor configuración de hiperparámetros, se aplicó la misma técnica de Walk-Forward sobre un conjunto de 30 días reservados para prueba. Para ello: Se utilizaron todos los registros anteriores como conjunto de entrenamiento inicial. Se aplicó nuevamente el esquema secuencial de entrenamiento y prueba diario, conservando el orden temporal.

Esta evaluación final permitió estimar el desempeño real del modelo con los hiperparámetros seleccionados, evitando cualquier tipo de fuga de información.

Esta metodología asegura una validación robusta y temporalmente coherente, adecuada para tareas de predicción sobre series cronológicas, como lo es la calidad del agua.(Brownlee, 2018)

## 3.4. Descripción de la selección de características

**Transformación de la variable temporal.**

La variable temporal original, representada como fechas, fue transformada a valores enteros utilizando como referencia el 1 de enero de 1900. Esta conversión garantiza la continuidad y secuencia de los datos, facilitando su procesamiento por los modelos de aprendizaje automático. Es importante destacar que la división entre los conjuntos de entrenamiento y prueba se realizó antes de esta transformación.

### 3.4.1. Selección de características

Para el modelo de **Random Forest**, se consideraron inicialmente todas las características disponibles, aprovechando su capacidad para manejar conjuntos de datos con múltiples variables y evaluar la importancia de cada una en la predicción. A partir de los resultados de importancia, se seleccionó un subconjunto de variables cuya contribución superaba el **2 %** en la métrica de importancia relativa. Este conjunto reducido fue utilizado en las pruebas con **Regresión Logística Multinomial (RL)** y **Máquinas de Vectores de Soporte (SVM)**, con el fin de reducir la dimensionalidad y mejorar la eficiencia computacional.

En el caso del modelo **LightGBM**, se volvió a utilizar el conjunto completo de variables, dada su eficiencia para procesar grandes volúmenes de atributos y realizar selección interna de características durante el entrenamiento.

## 3.5. Descripción de las métricas de evaluación

### 3.5.1. Métricas de evaluación

Los modelos se compararán utilizando métricas recomendadas para problemas multiclase y con fuerte desbalance en la variable objetivo (Powers, 2011):

- **F1−score macro**: media armónica entre precisión y recall por clase, otorga igual peso a cada clase independientemente de su frecuencia.

- **Recall por clase** : permite evaluar qué tan bien el modelo identifica correctamente las instancias verdaderas de cada categoría, lo cual es clave en contextos donde los estados críticos están subrepresentados.

Se enfatiza el análisis de las **matrices de confusión** para test, lo cual permite observar explícitamente los aciertos y errores en la clasificación de cada clase.

## 3.6. Descripción de los métodos de Machine Learning

Se evaluarán cuatro algoritmos supervisados multiclase:

1. **Regresión logística multinomial** : modelo lineal que estima las probabilidades de pertenencia a cada nivel mediante la función softmax. Ofrece interpretabilidad de coeficientes.
2. **Máquinas de Soporte Vectorial (MSV)** : se utilizaran dos nucleos Linear y RBF para capturar relaciones no lineales.
3. **Random Forest** : ensamble de árboles de decisión, robusto a ruido y que maneja interacciones variable–variable de forma automática.
4. **Light Gradient Boosting Machine (LGBM)** : algoritmo boosting basado en histogramas, adecuado para datos heterogéneos y de gran tamaño.

# Capítulo 4

# Resultados y discusión

## 4.1. Transformaciones preliminares

### 4.1.1. Conversión de unidades mediante regresión lineal

Con el fin de homogeneizar las unidades de análisis, se transformó la variable *cianobacteria* a $mm^3/L$ mediante un modelo de regresión lineal. El ajuste obtenido arrojó un coeficiente de determinación $R^2 = 0{,}994$ y un error cuadrático medio (RMSE) de 0,4720.

Estos resultados (implementados con `LinearRegression()`: $R^2 = 0{,}9942645$, RMSE = 0,472014) evidencian un desempeño altamente preciso del estimador, garantizando la confiabilidad de la conversión de unidades para las etapas posteriores de entrenamiento del modelo predictivo.

## 4.2. Análisis exploratorio de datos (EDA)

Se realizó un EDA agrupando las variables por su naturaleza.

### 4.2.1. Resumen estadístico por grupo

Se analizó la distribución básica de los valores por grupo de variables. A continuación se resumen los principales hallazgos:

**Variables meteorológicas.** Las temperaturas observadas (`T.obs`) presentan una media de 20.2°C, con un rango que va de 5.8 a 34.3°C. La humedad relativa (`Hr.obs`) tiene una amplia dispersión, con valores entre 34 % y casi 100 %. La radiación solar (`Rad.obs`) y la presión atmosférica (`pr.obs`) exhiben también alta variabilidad. Se observa una anomalía en `u.obs` (velocidad del viento) con un valor mínimo de –1039, lo que sugiere posibles errores de medición o registro.Los valores negativos detectados en esta variable fueron corregidos a cero para evitar distorsiones en el análisis.

**Variables hidrológicas.** El caudal de aporte (`Aporte_m3s`) muestra alta dispersión (std $\approx$ 4884), con valores que oscilan entre 410 y más de 25000 $m^3/s$. El nivel del embalse (`Nivel_emb`) oscila entre 30.87 y 35.42 m, con media de 33.7m.

**Variables biológicas.** Las variables biológicas como `Cianbac` y `Clorof_A` presentan fuerte asimetría. Por ejemplo, `Clorof_A` varía entre 0.3 y más de 1250 µg/L (media: 52.2), mientras que `Cianobac` tiene una media de 75,934 org/mL y un máximo superior a 2.4 millones.

**Variables químicas.** El nitrógeno total (`NT`) y el fósforo total (`Ptot_lab`) presentan valores dentro de rangos esperados, pero con leve sesgo positivo. La variable `SS_105` también muestra dispersión significativa (hasta 42.7 mg/L).

**Otras variables.** La conductividad eléctrica en superficie (`CondsupC`) presenta alta variabilidad (hasta 1575 µS/cm). La temperatura ambiente (`TempAmb`) oscila entre 13 y 35°C, mientras que el viento (`VientoI`) alcanza máximos de 24km/h.

En todos los grupos se observaron valores extremos (outliers) y distribuciones sesgadas, lo que motiva el uso de transformaciones, normalización y modelos robustos para el análisis posterior.

### 4.2.2. Distribuciones y valores atípicos

Se analizaron las distribuciones de seis variables relevantes, en función del nivel de alerta definido por la OMS y de trabajos de investigación que resaltan su importancia. En la Figura 4.1 se presentan boxplots para cada variable, lo que permite identificar diferencias entre las clases del target y observar la presencia de valores extremos.

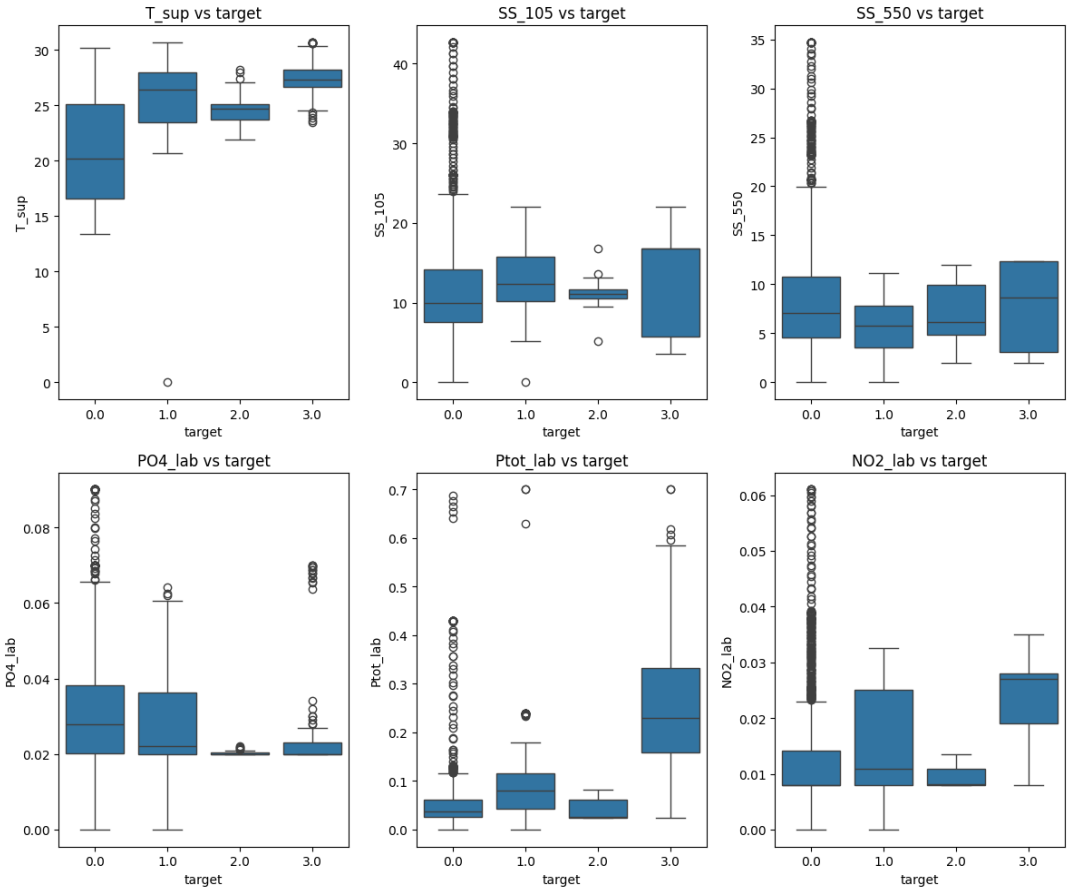



Figura 4.1: Distribución de variables seleccionadas respecto al nivel de alerta.

### 4.2.3. Correlaciones entre variables

Se construyó una matriz de correlación de Pearson entre todas las variables numéricas del conjunto de datos, limitada a coeficientes de correlación mayores a 0.5 para facilitar la lectura (ver Figura 4.2).

Se observan altas correlaciones entre variables relacionadas con sólidos en suspensión (`SS_105` y `SS_550`), y entre estas y `Aporte_m3s` y `EnterocPO4_lab` aunque mas leve. Además, las variables de temperatura muestran cierta correlación con la radiación y la conductividad. Estas relaciones sugieren que existen grupos de variables que responden a procesos ambientales similares o diferentes formas de medir atributos similares, lo que puede ser relevante al momento de seleccionar variables para los modelos predictivos y evitar problemas de multicolinealidad.

Se observan altas correlaciones entre variables relacionadas con sólidos en suspensión (`SS_105` y `SS_550`), y entre estas y `Aporte_m3s`; también se identifica correlación moderada con `Enteroc`. Además, las variables de temperatura muestran cierta correlación con la radiación y la conductividad. Estas relaciones sugieren que existen grupos de variables que responden a procesos ambientales similares, o bien distintas formas de medir atributos relacionados, lo que puede ser relevante al momento de seleccionar variables para los

modelos predictivos y evitar problemas de multicolinealidad.

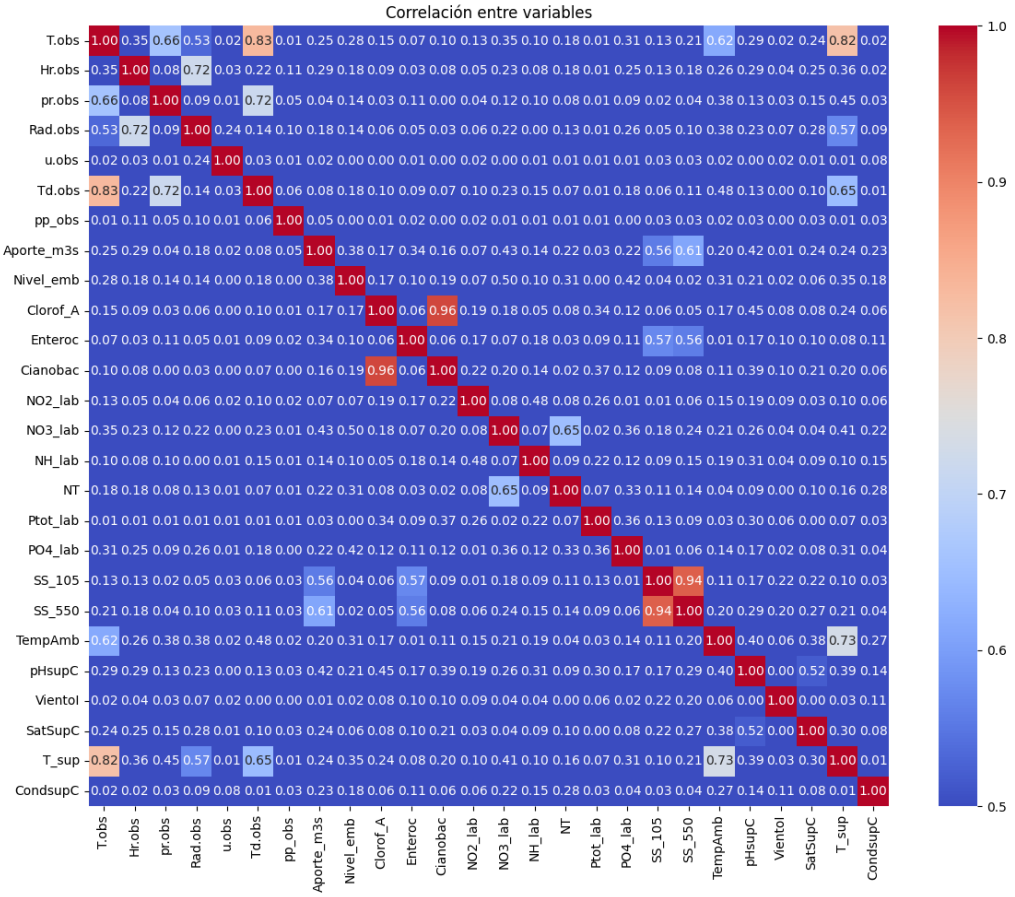



Figura 4.2: Matriz de correlación entre variables (coeficientes >0.5).

### 4.2.4. Relación de variables con el nivel de alerta

Se evaluó la dependencia entre cada variable y el nivel de alerta (`Target`) utilizando la métrica de información mutua, que permite capturar tanto relaciones lineales como no lineales entre variables.

Los resultados (ver Figura 4.3) indican que las variables más informativas para la clasificación del nivel de alerta son:

- **Cianobac** y **Clorof_A**: indicadores directos de la biomasa fitoplanctónica y la presencia de cianobacterias, fuertemente asociados a episodios de riesgo.
- **NO3_lab**, **Ptot_lab** y **SS_105**: representan condiciones de eutrofización (exceso de nutrientes), que favorecen el desarrollo de floraciones algales.
- **NO2_lab**, **NH_lab** y **SS_550**: reflejan procesos de descomposición y transformación del nitrógeno, así como la carga particulada en capas más profundas del cuerpo de agua.

Las variables meteorológicas (`Rad.obs`, `pr.obs`, `Hr.obs`, etc.) aportan menor información individual, aunque pueden tener valor complementario al combinarse con variables biológicas e hidrológicas.

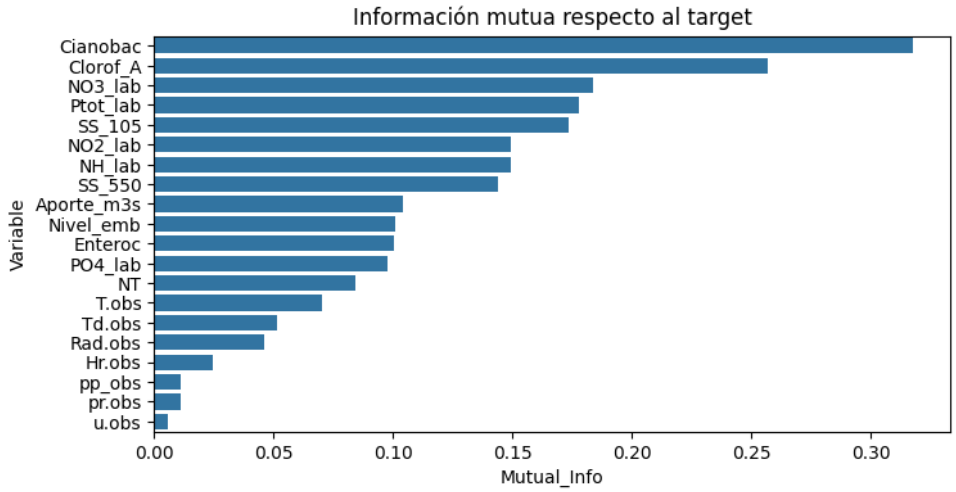



Figura 4.3: Ranking de variables según su información mutua con respecto al nivel de alerta.

### 4.2.5. Patrones temporales

Se graficaron las series temporales suavizadas (media móvil de 7 días) de cuatro variables claves: `Cianobac`, `Clorof_A`, `Aporte_m3s` y `T.obs`. La Figura 4.4 muestra su evolución a lo largo del período 2020–2025.

El caudal de aporte (`Aporte_m3s`) exhibe un patrón estacional con bajas precipitaciones en verano donde la temperatura del agua (`T.obs`) sigue un ciclo anual regular, con máximos en verano y mínimos en invierno, patrón consistente con la estacionalidad esperada en la región.

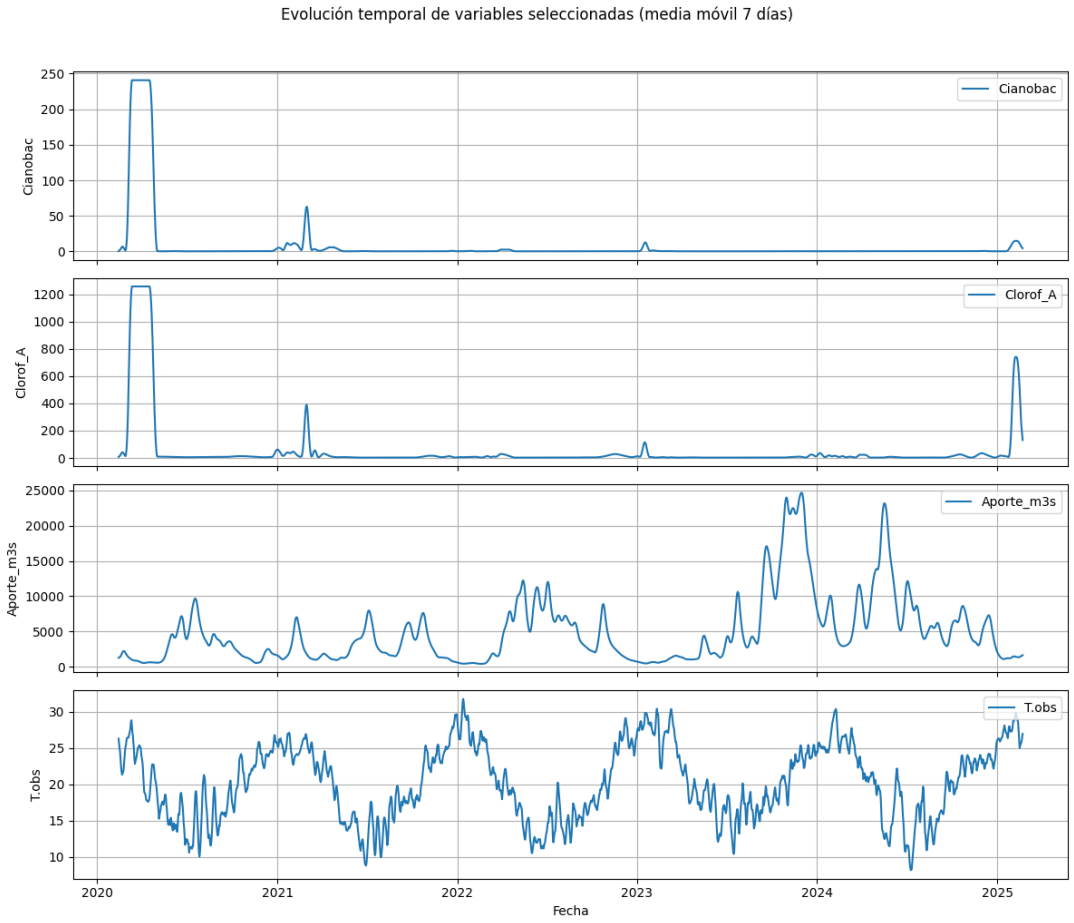



Figura 4.4: Evolución temporal de variables seleccionadas (media móvil de 7 días). Se observan eventos de floración mas marcados en los veranos de 2020, 2021, 2023 y 2025.

### 4.2.6. Distribución de clases en la variable objetivo

La Figura 4.5 muestra la distribución de clases en la variable objetivo utilizada para la predicción del nivel de alerta. Se observa un fuerte desbalance de clases, con una gran mayoría de observaciones correspondientes a la clase **0.0** (estado sin riesgo), que representa aproximadamente el 80 % del total. Las clases 1.0, 2.0 y 3.0, que indican niveles crecientes de alerta sanitaria, están considerablemente subrepresentadas, especialmente la clase 2.0.

Este desbalance plantea un desafío importante para los modelos de clasificación, ya que puede llevar a un sesgo hacia la clase mayoritaria y una baja sensibilidad frente a situaciones críticas.

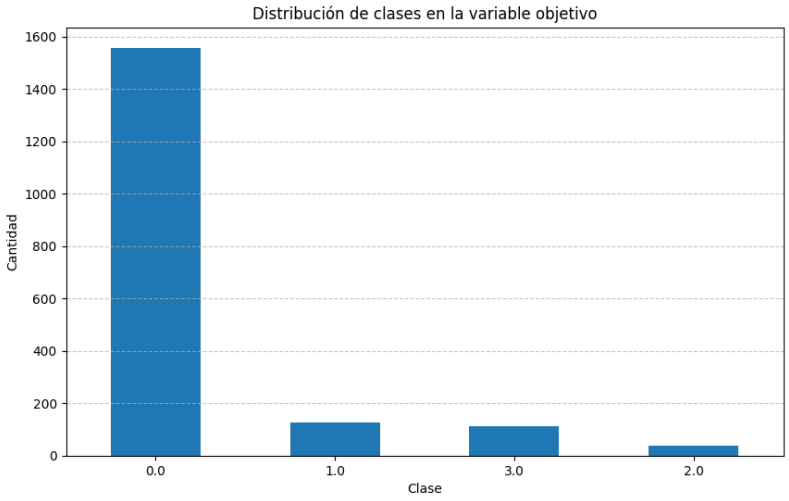



Figura 4.5: Distribución de clases en la variable objetivo

## 4.3. Presentación y análisis de resultados obtenidos

En esta sección se presentan los resultados obtenidos tras el entrenamiento de los modelos de Machine Learning propuestos. Cada subsección detalla el desempeño individual del modelo, sus principales métricas, y observaciones sobre su comportamiento frente al conjunto de test.

### 4.3.1. Random Forest

El modelo de *Random Forest* fue entrenado utilizando una búsqueda en malla (*Grid Search*) sobre los siguientes hiperparámetros:

- `n_estimators`: {50, 100, 200, 300, 400}
- `max_depth`: {2, 3, 5, 10, None}
- `min_samples_split`: {2, 3, 5}
- `min_samples_leaf`: {1, 2, 3}

Los mejores hiperparámetros encontrados fueron:

```
{'n_estimators': 200, 'max_depth': None, 'min_samples_split': 2,
                    'min_samples_leaf': 1}
```

El modelo alcanzó una exactitud del **0.9167** sobre el conjunto de validación. En cuanto a la evaluación en test logró una exactitud del **0.9333** y F1-score (macro): 0.9491. A continuación se presenta la matriz de confusión correspondiente al conjunto de test:

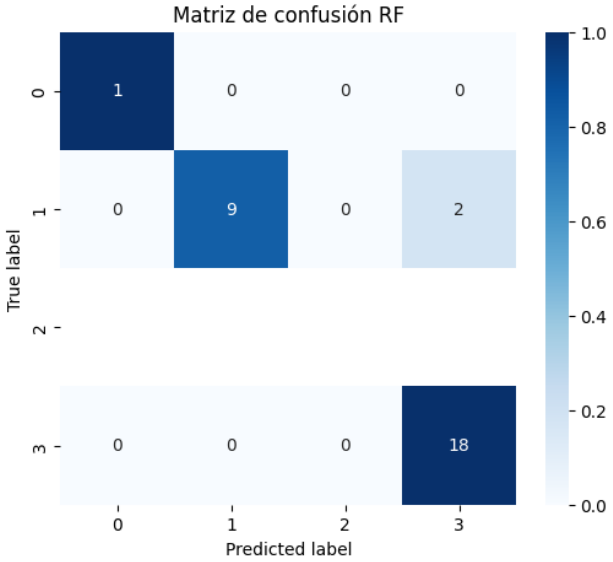



Figura 4.6: Matriz de confusión para el modelo Random Forest.

Cuadro 4.1: Recall por clase en el modelo Random Forest

| **Clase** | **Descripción** | **Recall** |
|---|---|---|
| 0 | Sin alerta | 1.00 |
| 1 | Alerta 0 | 0.82 |
| 2 | Alerta 1 | – |
| 3 | Alerta 2 (crítica) | 1.00 |

La matriz de confusión (ver Figura 4.6) muestra que el modelo logra identificar con alta precisión las clases extremas: la clase 0 (sin alerta) y la clase 3 (Alerta 2), ambas con un recall de 1.0. La clase 1 presenta un rendimiento levemente inferior (recall $\approx 0{,}82$), con algunos errores de clasificación hacia la clase 3. No se registran observaciones reales de la clase 2 en este conjunto de test, por lo que no se puede calcular su recall. Los valores de *recall* por clase se resumen en la Tabla 4.1. El comportamiento del modelo sugiere buena capacidad para distinguir entre estados contrastantes, aunque podrían requerirse más observaciones en clases intermedias para fortalecer la generalización en situaciones transicionales.

**Importancia de variables en Random Forest.**

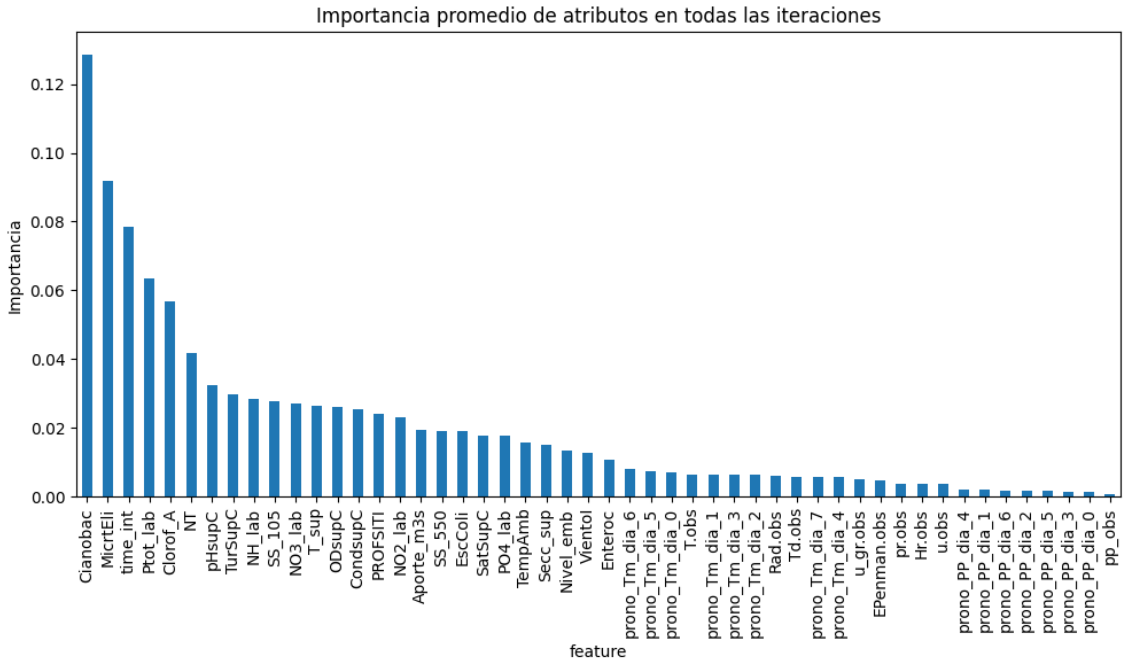



Figura 4.7: Importancia promedio de atributos de todas las iteraciones del modelo Random Forest.

Como se observa en la Figura 4.7, las variables `Cianobac`, `MicrtEli`, `time_int` y `Clorof_A` resultaron ser las más relevantes en la predicción del estado de alerta, seguidas por parámetros relacionados con nutrientes como `Ptot_lab`, `NH_lab` y otras con niveles significantes. También se destacan variables físico-químicas como `pHsupC`, `TursupC` y `SS_105` lo cual refuerza la influencia de condiciones biológicas y de turbidez en la clasificación. Las variables meteorológicas pronosticadas, en cambio, aportan menor valor predictivo en este modelo.

### 4.3.2. LightGBM

El modelo *LightGBM* fue entrenado utilizando una optimización bayesiana con *Optuna*, explorando el siguiente espacio de hiperparámetros:

- `learning_rate`: {0.01, 0.05, 0.1}
- `n_estimators`: {100, 200}
- `max_depth`: {3, 5, 7}
- `num_leaves`: {31, 63, 127}
- `min_child_samples`: {20, 50}
- `subsample`, `colsample_bytree`: {0.8, 1.0}
- `reg_alpha`, `reg_lambda`: {0, 0.1}

Los mejores hiperparámetros encontrados fueron:

```
{objective=multiclass, metric=multi_logloss, learning_rate=0.0445,
                        n_estimators=118,
```

`max_depth=7, num_leaves=64, min_child_samples=67, subsample=0.671, colsample_bytree=0.607, reg_alpha=7.18e-4, reg_lambda=3.24e-4}`

El modelo alcanzó una exactitud promedio de **0.9333** durante la validación. En el conjunto de test se obtuvo una exactitud de **0.9000** y un F1-score macro de **0.8285**. A continuación se presenta la matriz de confusión correspondiente:

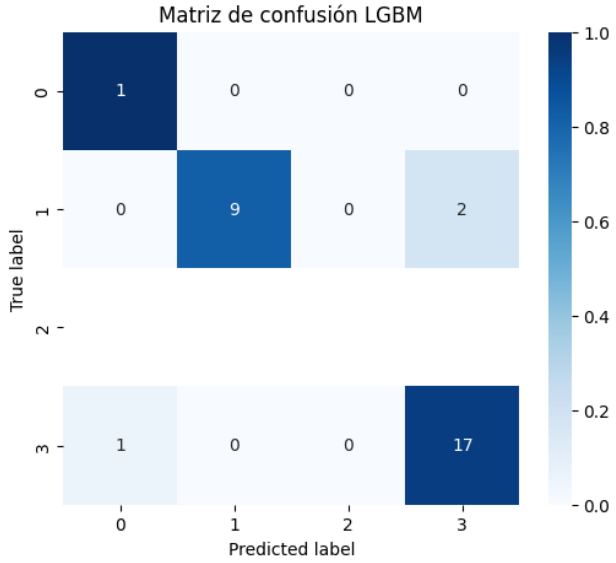



Figura 4.8: Matriz de confusión para el modelo LightGBM.

Cuadro 4.2: Recall por clase en el modelo LightGBM

| **Clase** | **Descripción** | **Recall** |
|---|---|---|
| 0 | Sin alerta | 1.00 |
| 1 | Alerta 0 | 0.82 |
| 2 | Alerta 1 | – |
| 3 | Alerta 2 (crítica) | 0.94 |

La matriz de confusión (ver Figura 4.8) muestra que el modelo logra clasificar correctamente la mayoría de las instancias de clase 3 (Alerta 2), aunque con un leve descenso en el recall respecto a Random Forest (recall = 0.94). También se observa una clasificación correcta de la clase 0 (sin alerta), mientras que la clase 1 mantiene un rendimiento similar (recall $\approx$ 0,82). No se registraron observaciones de la clase 2 en este conjunto de test. Los valores de *recall* por clase se resumen en la Tabla 4.2. En conjunto, el modelo demuestra buena capacidad discriminativa, con un balance razonable entre precisión y sensibilidad, aunque con una leve pérdida de robustez en los extremos críticos.

### 4.3.3. Otros modelos evaluados

Además de los modelos presentados, se realizaron pruebas con **Regresión Logística Multinomial** y **Máquinas de Soporte Vectorial (MSV)** utilizando el conjunto reducido de variables seleccionadas por importancia. Sin embargo, ambos modelos presentaron un desempeño muy pobre en comparación con Random Forest y LightGBM.

- **Regresión Logística:** Exactitud: 0.3000, F1-score macro: 0.2059
- **SVM:** Exactitud: 0.2000, F1-score macro: 0.2000

Estos resultados indican una baja capacidad de generalización, posiblemente asociada a la naturaleza del problema, el desbalance de clases y la limitada expresividad de estos modelos en este contexto.

### 4.3.4. Discusión de los resultados y su relevancia

Los resultados obtenidos muestran un desempeño destacado de los modelos *Random Forest* y *LightGBM*, que alcanzaron valores de exactitud superiores al 90 % y F1-score macro por encima de 0.82 en los conjuntos de test. Estas métricas superan ampliamente a las obtenidas por modelos lineales como la regresión logística o SVM, confirmando la superioridad de enfoques no lineales para este tipo de tareas. Este comportamiento es coherente con hallazgos previos en la literatura, donde modelos de árboles también se destacaron en la predicción de variables como clorofila-*a* o calidad de agua en lagos del hemisferio sur (Rodríguez-López y cols., 2023), así como en revisiones generales sobre el uso de aprendizaje automático en sistemas acuáticos (Schaeffer y cols., 2024).

Desde una perspectiva ambiental, estos modelos demostraron capacidad para identificar correctamente los estados más críticos de calidad del agua (clase 3), lo cual representa una contribución relevante para sistemas de alerta temprana. La alta tasa de aciertos en estas clases sugiere que los modelos pueden ser herramientas útiles en la gestión del riesgo sanitario y recreativo en cuerpos de agua como el embalse de Salto Grande.

En cuanto a las variables más relevantes para la predicción, los resultados reflejan una clara preponderancia de indicadores biológicos y fisicoquímicos `Cianobac` (densidad de células de cianobacterias), `MicrtEli` (microcistinas totales determinadas por ELISA-ADDA), `Ptot_lab` (fósforo total), `NT` (nitrógeno total) y `Clorof_A` (clorofila *a*), por sobre las variables meteorológicas o pronosticadas.

Esta jerarquía es coherente con la dinámica ecológica observada en cuerpos de agua continentales templados: los macronutrientes `Ptot_lab` y `NT` actúan como insumos primarios que estimulan la biomasa algal (reflejada en `Clorof_A`) y favorecen la proliferación de cianobacterias (`Cianobac`). A su vez, la variable `MicrtEli` representa la fracción tóxica que ciertas especies generan cuando alcanzan altas densidades y condiciones de estrés leve (p. ej., exceso de luz o limitación de N:P), por lo que suele mostrar una respuesta no lineal pero fuertemente asociada al incremento simultáneo de `Cianobac`, `Ptot_lab` y `Clorof_A`. Esta observación respalda estudios que destacan el rol dominante de los nutrientes como

impulsores del crecimiento y la toxicidad de cianobacterias, en comparación con factores climáticos como la temperatura (Rigosi y cols., 2014).

Finalmente, la implementación de modelos de aprendizaje automático no lineales permitió abordar la complejidad inherente a la dinámica ecológica del embalse, compensando parcialmente las limitaciones asociadas a la escasez y dispersión temporal de los datos disponibles. Esta evidencia refuerza el potencial de estas herramientas para su integración en esquemas de monitoreo ambiental y evaluación predictiva localizada.

### 4.3.5. Limitaciones y posibles mejoras

A pesar de los resultados prometedores obtenidos, el estudio presenta una serie de limitaciones tanto a nivel de datos como de enfoque metodológico, que deben ser consideradas al interpretar los hallazgos y planificar desarrollos futuros.

En primer lugar, el desbalance en la variable objetivo constituye un desafío relevante: la clase 0 (sin alerta) domina ampliamente el conjunto de datos, mientras que las clases correspondientes a niveles de riesgo medio o alto están subrepresentadas. Esto limita la capacidad del modelo para aprender patrones asociados a eventos críticos. Además, la cantidad limitada de observaciones reales restringe el entrenamiento de modelos más complejos o sensibles a la varianza. La frecuencia irregular de muestreo, semanal en verano y mensual en invierno, también dificulta la modelización temporal. Por otro lado, debido a la partición de test generada mediante la técnica *walk-forward*, ciertas clases no están representadas, lo que impide calcular métricas como el *recall* para cada clase.

A esto se suma que, debido a la falta de continuidad en las observaciones, no fue posible construir un conjunto de entrenamiento totalmente real que permitiera comparar de forma controlada el desempeño del modelo con y sin interpolación. Esta limitación impide cuantificar el impacto exacto de los datos sintéticos y representa un aspecto clave a explorar en trabajos futuros.

Desde el punto de vista metodológico, no se incorporaron enfoques explícitos de modelado de series temporales, como *Long Short-Term Memory (LSTM)* o *Prophet*, que podrían capturar patrones secuenciales o estacionales relevantes en la dinámica ecológica del sistema.

Si bien la interpolación de variables clave fue necesaria para compensar la escasez y dispersión temporal de los datos, en especial para atributos biológicos, esta estrategia introduce un riesgo potencial de sobreajuste. El modelo podría aprender patrones artificiales generados por la interpolación, en lugar de relaciones efectivamente observadas en el sistema. No obstante, esta técnica resultó indispensable para mantener un volumen suficiente de datos de entrenamiento. Sería deseable en futuros estudios evaluar formalmente el impacto de los datos sintéticos, comparando el desempeño del modelo con y sin interpolación, preferentemente sobre conjuntos de test construidos exclusivamente a partir de registros reales.

Como líneas de mejora, se propone ampliar el dataset mediante la incorporación de

campañas históricas y datos de estaciones ubicadas en otros puntos del lago. También se recomienda evaluar el desempeño de modelos probabilísticos o híbridos, que integren la predicción con la estimación de incertidumbre. Finalmente, sería conveniente implementar esquemas de validación empleando años completamente excluidos del entrenamiento, ya que esto permitiría evaluar con mayor realismo la capacidad de generalización de los modelos desarrollados.

# Capítulo 5

# Conclusión

## 5.1. Resumen de los hallazgos principales

Se desarrollaron modelos de predicción de calidad del agua con horizonte semanal, aplicados al embalse de Salto Grande, integrando datos satelitales, in situ, meteorológicos e hidrológicos. Entre los algoritmos evaluados, **Random Forest** y **LightGBM** demostraron un desempeño robusto, alcanzando exactitudes superiores al 90 % y F1-score macro por encima de 0.82. Estos modelos lograron identificar correctamente los estados críticos de alerta, con un alto recall en la clase de mayor riesgo. Las variables biológicas y fisicoquímicas, en especial cianobacterias, clorofila y nutrientes, resultaron ser las más relevantes para la predicción.

## 5.2. Conclusiones generales y relación con los objetivos

El estudio confirma la factibilidad de proyectar el estado de calidad del agua con hasta 7 días de anticipación en condiciones reales de monitoreo, utilizando técnicas de Machine Learning. Los resultados permiten afirmar que estas herramientas pueden ser integradas a esquemas de alerta temprana, fortaleciendo la toma de decisiones para el uso recreativo y sanitario del recurso en playas específicas como Las Palmeras. Se logró así responder positivamente a la pregunta planteada y cumplir con el objetivo general del trabajo.

## 5.3. Recomendaciones para futuros trabajos

Se recomienda ampliar el volumen y la cobertura temporal del conjunto de datos, así como incorporar nuevas fuentes satelitales, posiblemente mediante la contratación de servicios complementarios a los actualmente utilizados, con el objetivo de mejorar la resolución y frecuencia de observación. También se sugiere implementar modelos secuenciales que capten mejor la dinámica temporal de las variables, y evaluar la transferencia del enfoque desarrollado a otras estaciones del embalse, mediante un entrenamiento conjunto o comparativo entre sitios.

Una línea de trabajo futura relevante consiste en explorar horizontes de predicción más amplios, por ejemplo, extendiendo la proyección a 14 días. La elección del horizonte de 7 días en este estudio respondió a una necesidad operativa concreta: generar informes semanales de calidad del agua, solicitados por el organismo responsable durante la temporada de mayor uso recreativo (noviembre a marzo). Esta ventana representa un compromiso entre utilidad práctica y estabilidad predictiva, ya que la variabilidad de nutrientes y la incertidumbre en los pronósticos meteorológicos aumentan considerablemente a medida que se extiende el plazo.

No obstante, evaluar horizontes más largos permitiría anticipar eventos críticos con mayor antelación, lo cual resulta especialmente valioso en contextos de alerta sanitaria o planificación preventiva. Asimismo, se propone analizar más detalladamente los días intermedios entre el estado actual y la predicción a 7 días, particularmente en aquellos casos donde se produce un cambio de clase. Esta estrategia permitiría identificar el momento probable del quiebre y profundizar en el análisis de las variables que lo anticipan, aportando evidencia adicional sobre los atributos más determinantes en la dinámica transicional de la calidad del agua.

# Referencias


Backer, L. C., y Landsberg, J. H. (2024). Epidemiologic and clinical features of cyanobacteria harmful algal bloom–associated illnesses, 2007–2022. *Environmental Health*, *23*(1), 15. Descargado de `https://doi.org/10.1186/s12940-024-01121-y` doi: 10.1186/s12940-024-01121-y

Brownlee, J. (2018). *How to backtest machine learning models for time series forecasting.* Descargado de `https://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting/` (Consultado el 5 de junio de 2025)

DataCamp. (2022). *Time series forecasting tutorial.* Descargado de `https://www.datacamp.com/tutorial/tutorial-time-series-forecasting` (Recuperado el 5 de junio de 2025)

Hunter, P. D., Rudd, K. M., y Tyler, A. N. (2022). Global approaches to recreational-water bloom alert frameworks: a critical review. *Harmful Algae*, *118*, 102238. Descargado de `https://doi.org/10.1016/j.hal.2022.102238` doi: 10.1016/j.hal.2022.102238

Igwaran, A., Kayode, A. J., Moloantoa, K. M., Magadla, N. P., Adelagun, A. A., Ojekunle, S. A., ... Abiola, O. S. (2024). Cyanobacteria harmful algae blooms: Causes, impacts, and risk management. *Water, Air, & Soil Pollution*, *235*, 71. Descargado de `https://doi.org/10.1007/s11270-023-06782-y` doi: 10.1007/s11270-023-06782-y

Oh, C., Han, S., y Jeong, J. (2020). Time-series data augmentation based on interpolation. *Procedia Computer Science*, *175*, 64–71.

Powers, D. M. W. (2011). Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Rigosi, A., Carey, C. C., Ibelings, B. W., y Brookes, J. D. (2014). The interaction between climate warming and eutrophication to promote cyanobacteria is dependent on trophic state and varies among taxa. *Limnology and Oceanography*, *59*(1), 99–114. Descargado de `https://aslopubs.onlinelibrary.wiley.com/doi/10.4319/lo.2014.59.1.0099` doi: 10.4319/lo.2014.59.1.0099

Rodríguez-López, J. M., Torres, R., Mardones, J., y Cordero, R. R. (2023). Machine learning algorithms for the estimation of water quality parameters in lake llanquihue in southern chile. *Water*, *15*(11), 1994. Descargado de https://www.mdpi.com/2073-4441/15/11/1994 doi: 10.3390/w15111994

Schaeffer, B. A., Reynolds, N., Ferriby, H., Salls, W., Smith, D., Johnston, J. M., y Myer, M. (2024). Forecasting freshwater cyanobacterial harmful algal blooms for sentinel-3 satellite resolved u.s. lakes and reservoirs. *Journal of Environmental Management*, *349*, 119518. Descargado de https://doi.org/10.1016/j.jenvman.2023.119518 doi: 10.1016/j.jenvman.2023.119518

Stumpf, R. P., Davis, T. W., Wynne, T. T., y Graham, J. L. (2016). Challenges for mapping cyanotoxin patterns from remote sensing of cyanobacteria. *Harmful Algae*, *54*, 160–173. Descargado de https://doi.org/10.1016/j.hal.2016.01.005 doi: 10.1016/j.hal.2016.01.005

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... van der Walt, S. J. (2020). Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, *17*(3), 261–272. doi: 10.1038/s41592-019-0686-2

World Health Organization. (2021). *Toxic cyanobacteria in water: A guide to their public health consequences, monitoring and management* (2.ª ed.). Geneva: World Health Organization. Descargado de https://iris.who.int/bitstream/handle/10665/342625/9789240031302-eng.pdf

# Apéndice A

# Anexos

## A.1. Repositorio en GitHub

https://github.com/JoacoTschopp/proyeccion-calidad-agua-ml