

Predicción del riesgo de Cianobacterias en el agua

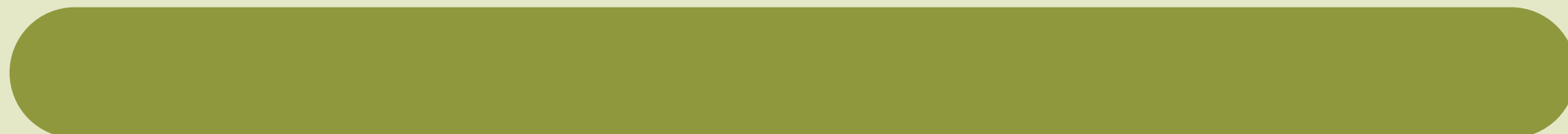
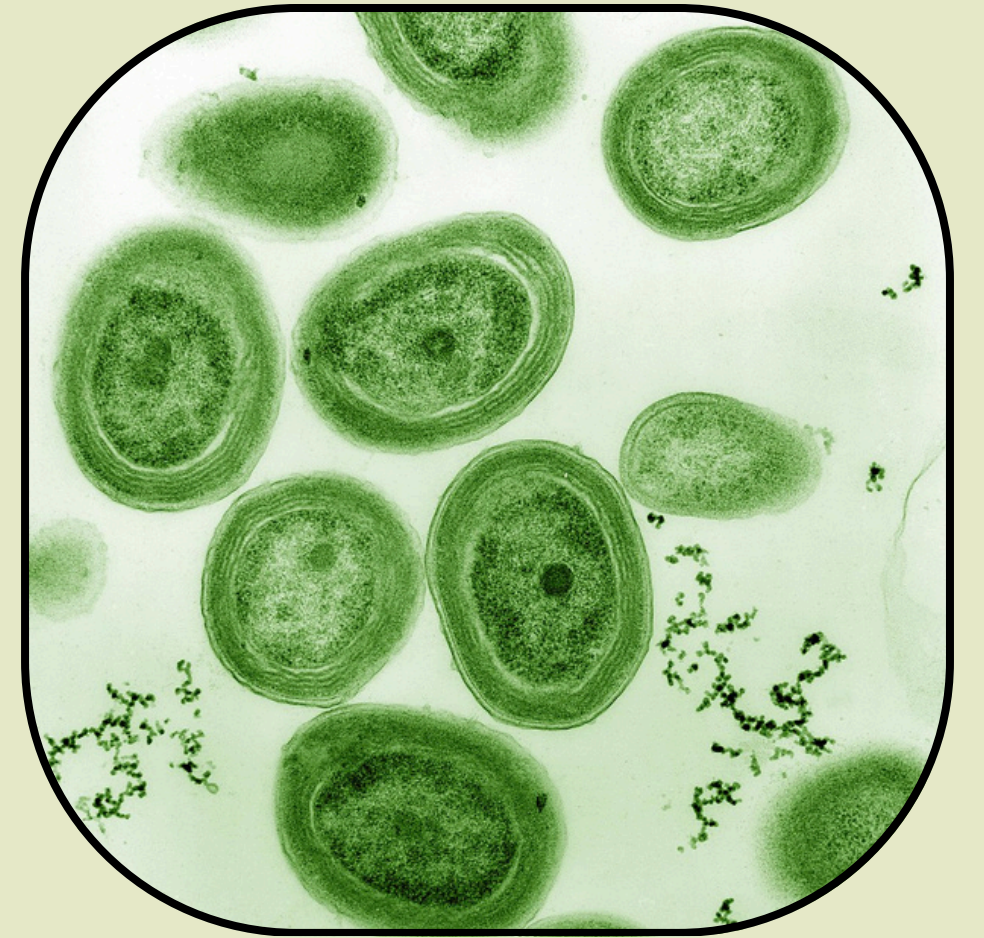
Autor: Tahiel Maccario

tahiel53@gmail.com



Introducción

Las cianobacterias son microorganismos que habitan en el agua y son esenciales en los ambientes naturales, ya que aumentan la disponibilidad de oxígeno. Sin embargo, algunas especies pueden producir compuestos tóxicos para los humanos, conocidos como cianotoxinas. Según la OMS, en cuerpos de agua destinados a fines recreativos y para el consumo, una concentración de cianobacterias mayor a 20000 células por mililitro puede presentar un riesgo para la salud debido a la producción de estas toxinas. Ciertas condiciones como la temperatura y el exceso de nutrientes pueden producir un crecimiento de cianobacterias desmedido, conocido como "floraciones". Por esta razón, el monitoreo de las floraciones en fuentes de agua naturales y el estudio de las variables que las afectan es fundamental.



Objetivo y audiencia

A través del análisis de datos ambientales, este proyecto busca identificar patrones y factores que contribuyan al crecimiento de cianobacterias y generar un modelo de clasificación que pueda predecir el riesgo de que ocurra un florecimiento peligroso (que haya más de 20000 células por mililitro) bajo determinadas condiciones ambientales. Esto puede ser de gran utilidad para organismos reguladores en el monitoreo y la gestión de la calidad del agua en el desarrollo de estrategias de gestión para mitigar el impacto en la salud pública y la calidad del agua.



Preguntas



Preguntas principales

¿Cómo influye la disponibilidad de nutrientes en los florecimientos de cianobacterias?

¿La temperatura tiene un impacto significativo en el crecimiento de las cianobacterias?

Preguntas secundarias

¿Cuál es la tendencia durante los años de aparición de floraciones?

¿Cómo cambia la concentración de cianobacterias entre distintos lugares y tipos de ambiente?

¿Qué nutrientes tienen mayor impacto en la aparición de floraciones?

¿Es más relevante la temperatura bajo el agua que en la superficie para el crecimiento de cianobacterias?

Metadata

Se partió de un dataset con 640 muestreos en 20 embalses distintos de Estados Unidos. Tres embalses fueron suprimidos por falta de datos. El resto de datos fueron procesados quedando una tabla de 361 datos y 19 columnas.

Reservoir abbreviation	Reservoir	Type	Stratification
	Name		
BHR	Buckhorn Lake	Forest	Strong
BRR	Barren River Lake	Agriculture	Strong
BVR	Brookville Lake	Agriculture	Strong
CBR	C. J. Brown Lake	Agriculture	None
CCK	Caesar Creek Lake	Agriculture	Strong
CFK	Carr Creek Lake	Forest	Strong
CHL	C. M. Harden Lake	Agriculture	Weak
CMR	Cagles Mill Lake	Agriculture	Weak
CRR	Cave Run Lake	Agriculture	Strong
EFR	East Fork Lake	Agriculture	Strong
GRR	Green River Lake	Agriculture	Strong
HTR	J. E. Roush Lake	Agriculture	None
MNR	Monroe Lake	Agriculture	Strong
MSR	Mississinewa Lake	Agriculture	None
NRR	Nolin Lake	Agriculture	Weak
PRR	Patoka Lake	Forest	Strong
RRR	Rough River Lake	Agriculture	Strong
SRR	Salamonie Lake	Agriculture	None
TAR	Taylorsville Lake	Agriculture	Strong

La estratificación nos dice si hay muchas capas de vegetación en la zona

Nutrientes

Variables de estudio

- **Reservoir:** Iniciales del nombre del embalse
- **Reservoir_type:** Tipo de ambiente según la vegetación
 - 1 = boscoso con vegetación estratificada
 - 2 = agrícola con vegetacion estratificada
 - 3 = estratificación debil o no estratificado
- **Year:** Año del muestreo
- **Cyanobacteria_Max_cells/ml:** **concentración de cianobacterias (células/ml) -> variable de interés**
- **Summer_precip_inches** = Precipitaciones totales durante el verano
- **ST_Celsius** = Temperatura (°C) en la superficie (promedio del verano)
- **DT_Celsius** = Temperatura (°C) en la profundidad (promedio del verano)
- **P_dissolved_ppb** = Fosforo disuleto (ppb)
- **TKN_ppm** = Nitrógeno total de Kjeldahl (ppm)
- **NH3_ppm** = Amoníaco total (ppm)
- **Alcalinity_ppm** = Alcalinidad (ppm)
- **DO_mg/l** = Oxígeno disuelto en la profundidad (mg/l)

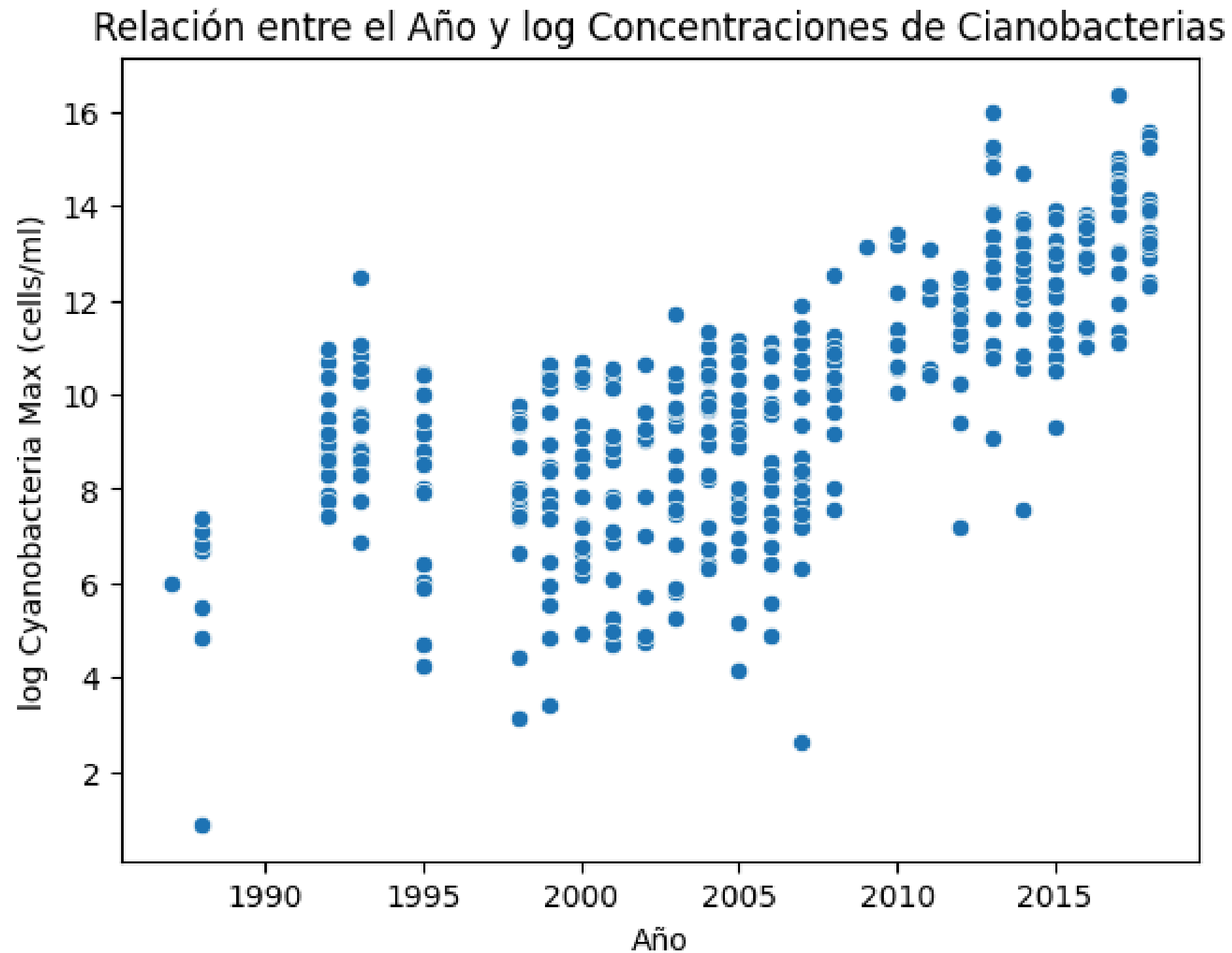
Las siguientes variables estan medidas en los embalses y en los afluentes al embalse (aclarado como “6to8inflow” en los afluentes):

- **NOx_ppm** = Nitritos y nitratos totales (ppm)
- **TKN_ppm** = Nitrógeno total de Kjeldahl (ppm)
- **NH3_ppm** = Amoníaco total (ppm)
- **TP_ppb** = Fósforo total (ppm)
- **TOC_ppm** = Carbono orgánico total (ppm)

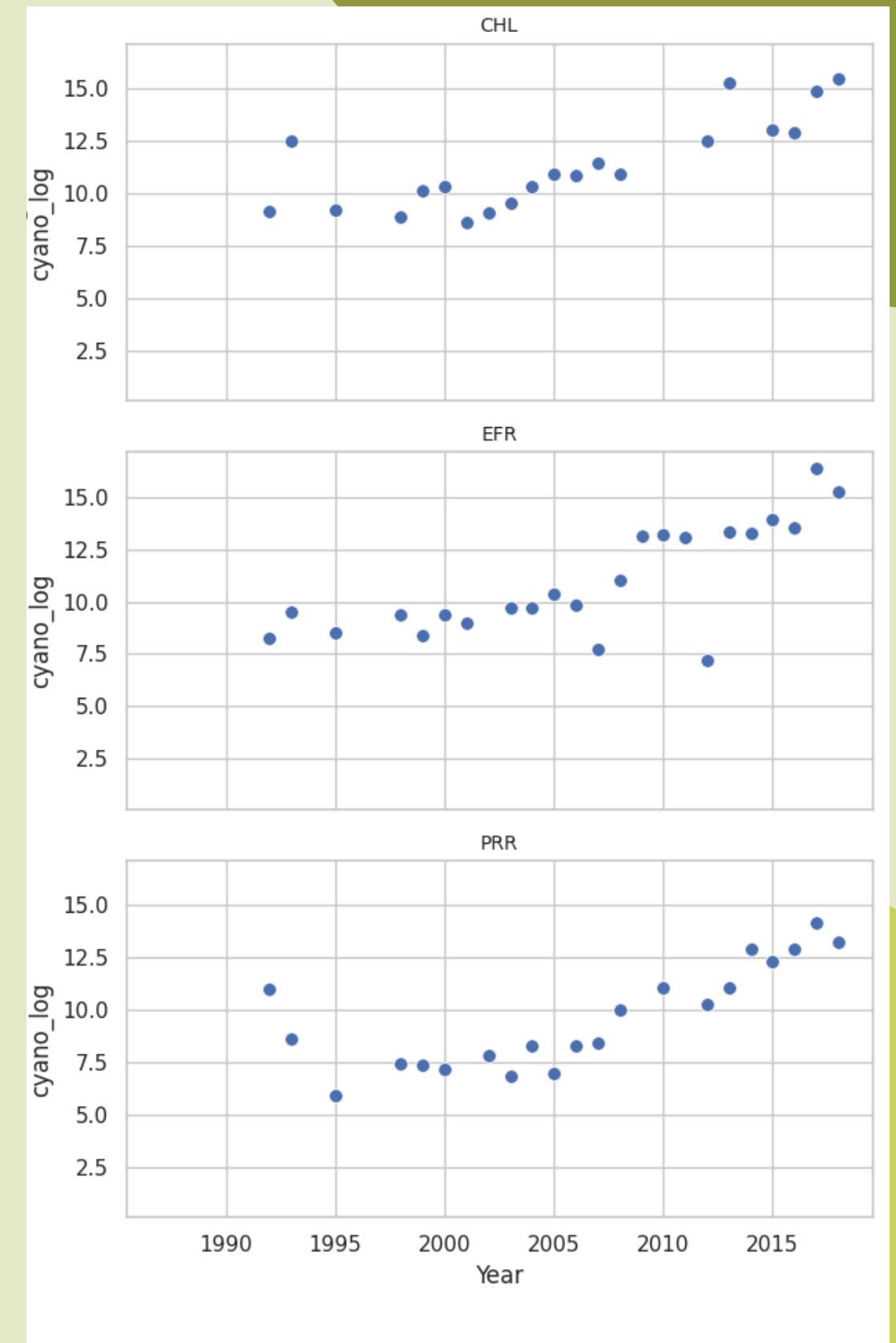
Análisis Exploratorio

¿Cual es la tendencia en el tiempo de la aparición de floraciones?

Globalmente se ve un aumento de las cianobacterias durante los años



La tendencia también se observa viendo cada embalse por separado (tres ejemplos)



¿Cómo cambia la concentración de cianobacterias entre distintos lugares y tipos de ambiente?

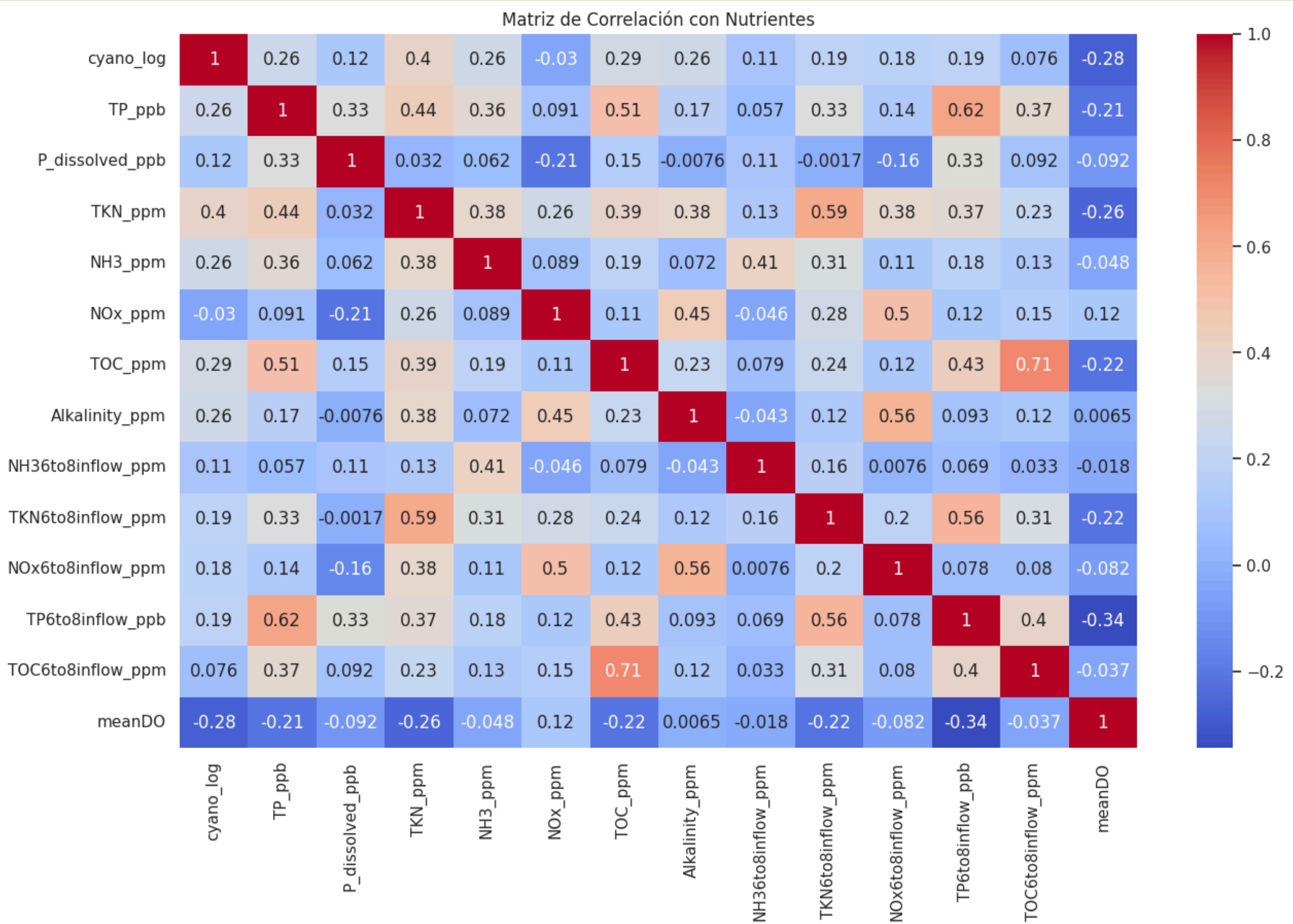
Menos
cianobacterias

Tipo de embalse	Embalse	Concentracion de cianobacterias (células/ml)
Boscoso con vegetación estratificada	BHR	135.528
	CFK	83.440
	CRR	55.463
	MNR	103.604
	PRR	154.168
Agrícola con vegetacion estratificada	BRR	518.669
	BVR	328.703
	CCK	312.232
	EFR	944.370
	GRR	170.631
	RRR	141.070
	TAR	291.089
Estratificación debil o no estratificado	CBR	1079.754
	CHL	714.175.
	CMR	283.843
	MSR	176.059
	NRR	185.421

La concentración de cianobacterias varía entre embalses

(Es necesario hacer tests de significancia)

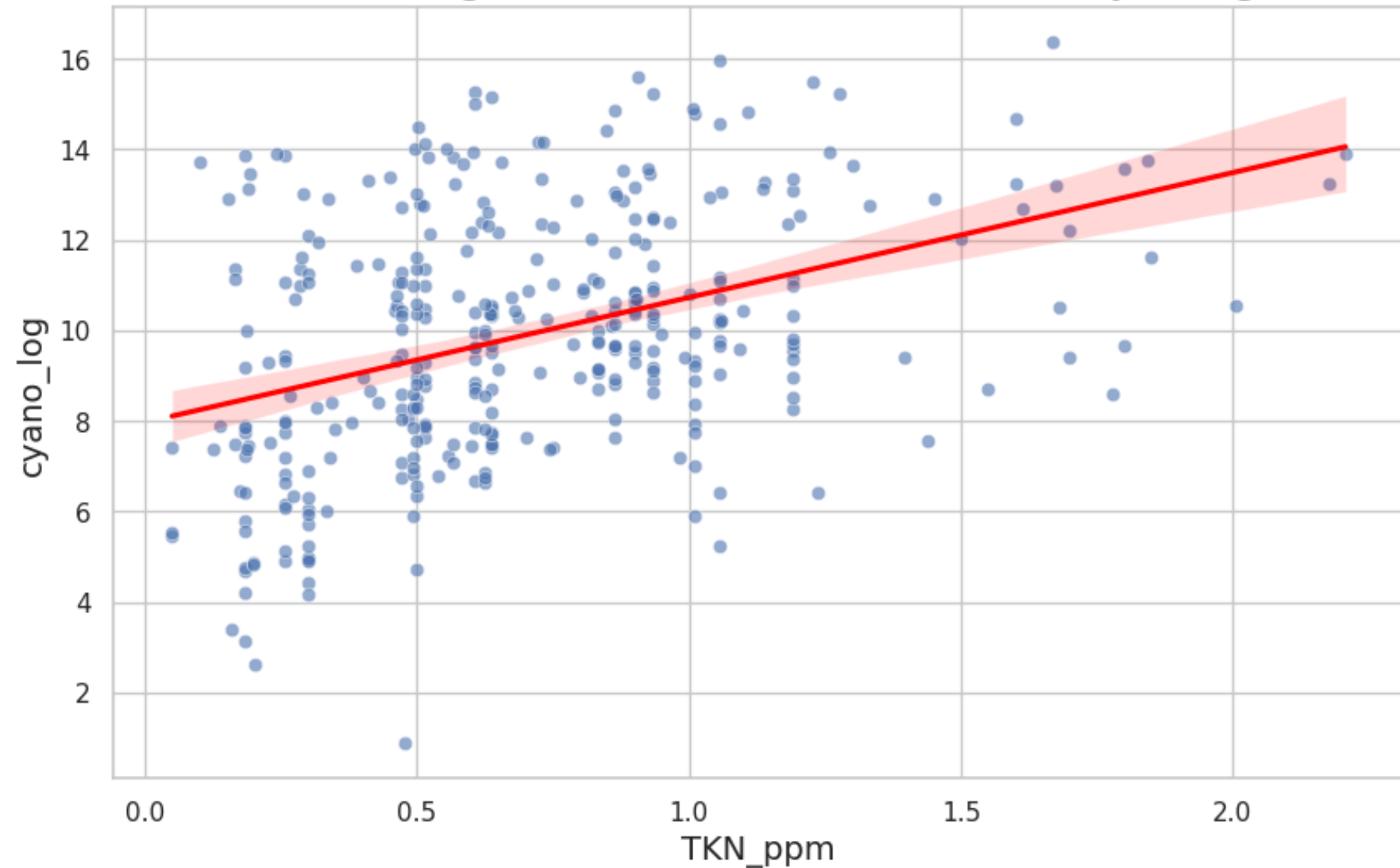
¿Qué nutrientes tienen mayor impacto en la aparición de floraciones?



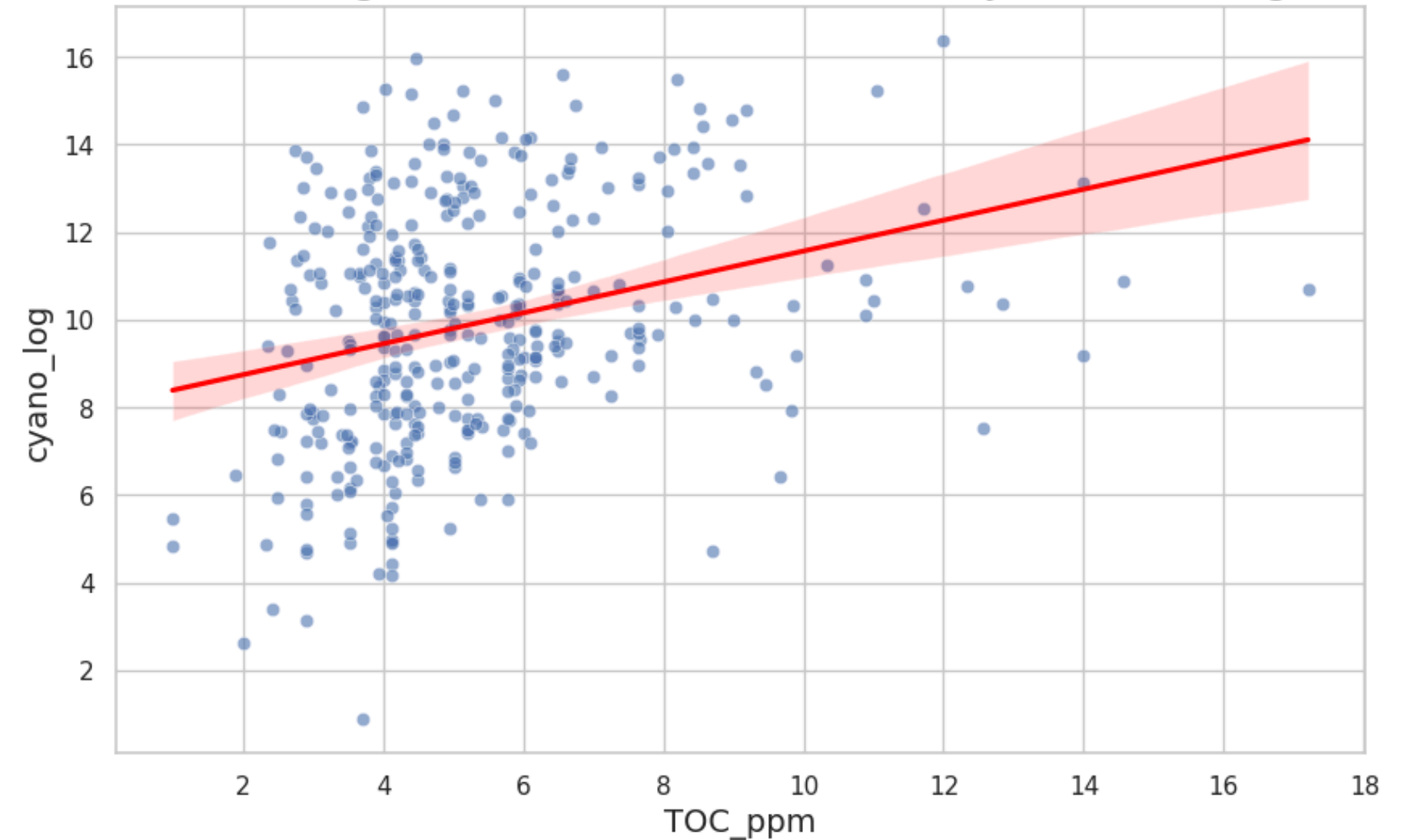
Las variables que tienen más correlación con la concentración de cianobacterias (en escala logarítmica) son el Nitrógeno Total (TKN), el Carbono Orgánico Total (TOC) y en menor medida el Fósforo Total (TP), la Alcalinidad y el Amoníaco Total (NH3).

¿Qué nutrientes tienen mayor impacto en la aparición de floraciones?

Correlación entre log concentración de cianobacterias y Nitrogeno total

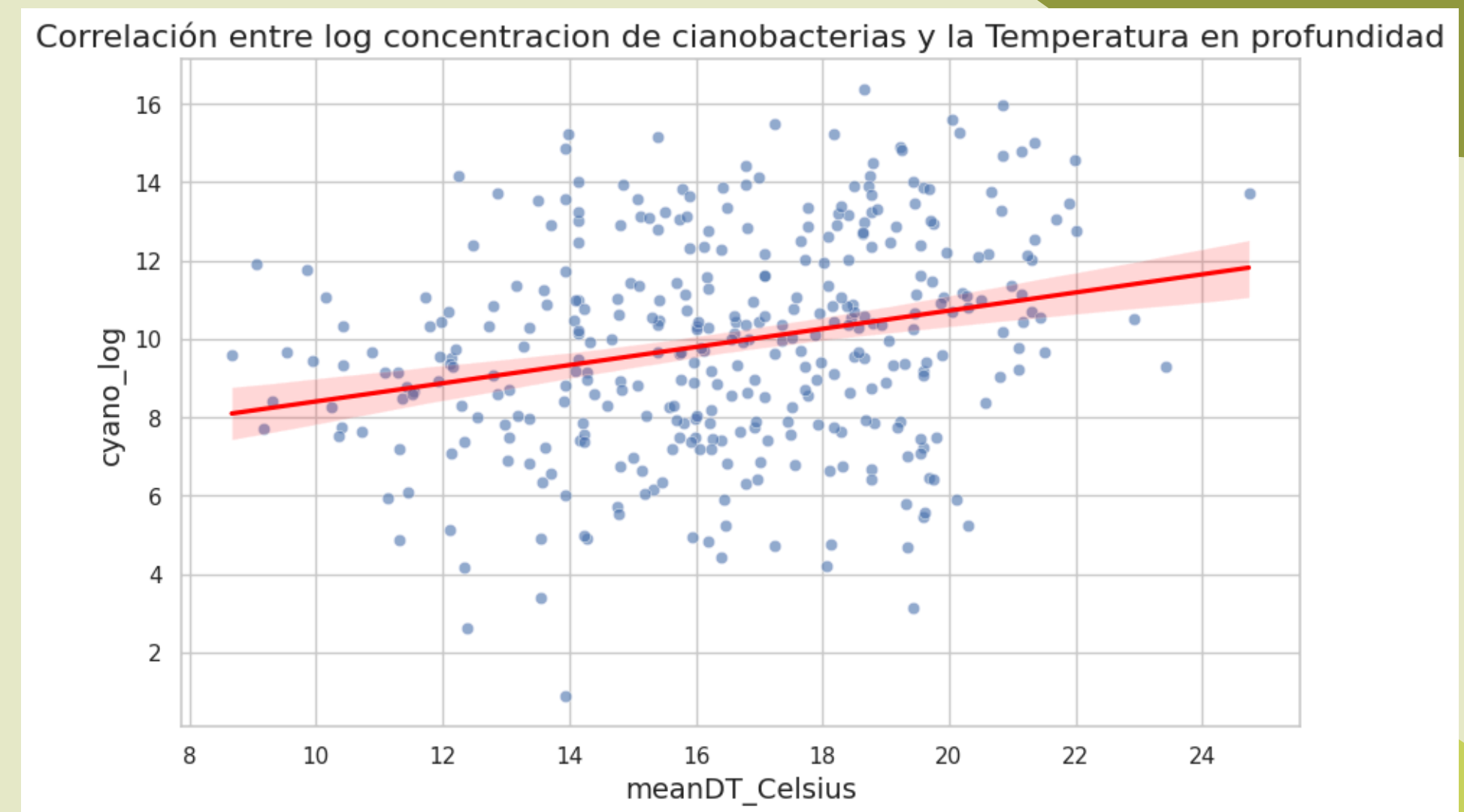
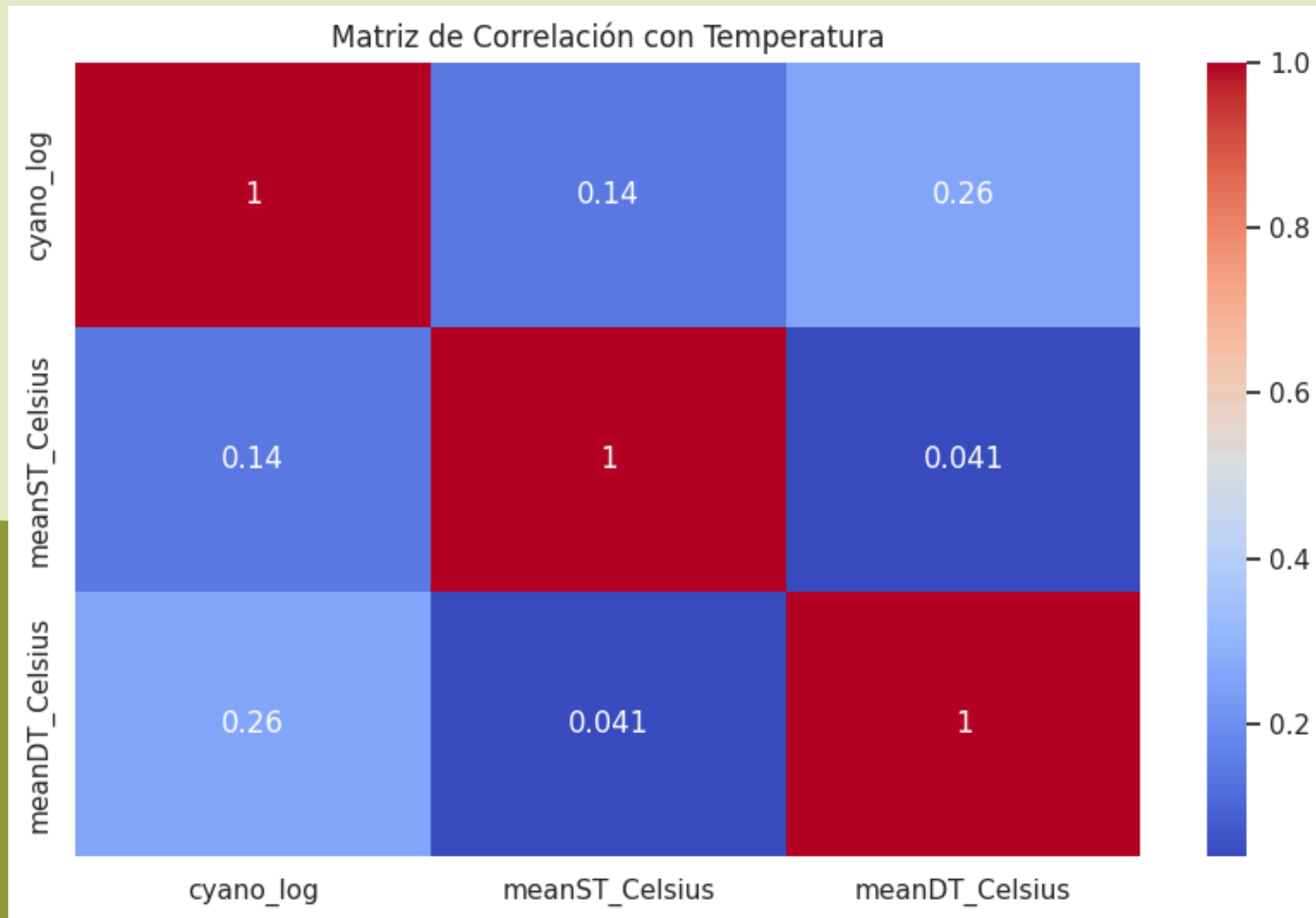


Correlación entre log concentración de cianobacterias y el Carbono Orgánico Total



El Nitrógeno Total (TKN) y el Carbono Orgánico Total (TOC) son los nutrientes que más se correlacionan con la cantidad de cianobacterias en el agua

¿Es más relevante la temperatura bajo el agua que en la superficie para el crecimiento de cianobacterias?



Se observa mayor correlación con la temperatura bajo el agua (DT) que en superficie (ST)

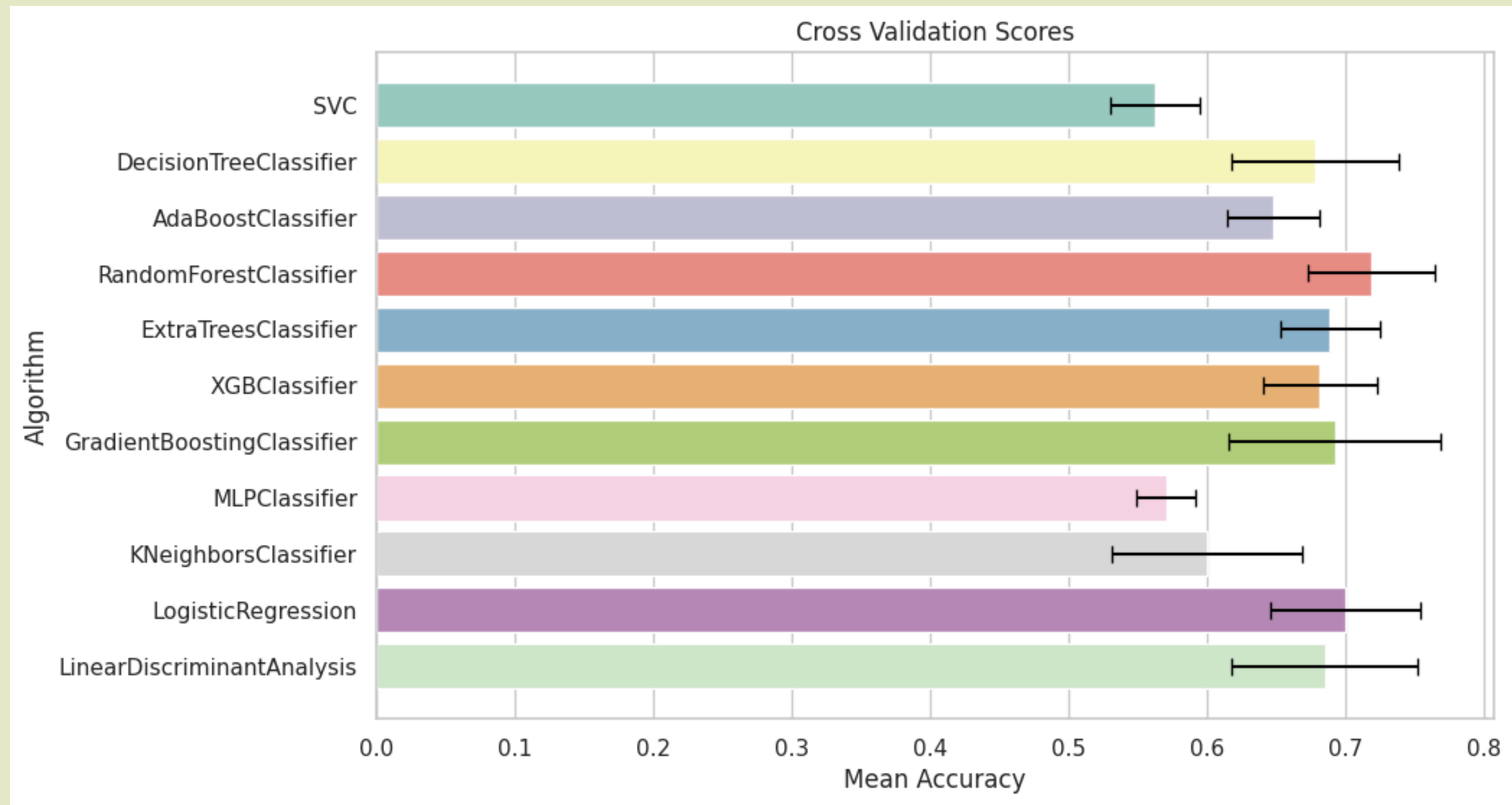
Insights

- Se observa un aumento de la concentración de cianobacterias en el agua a lo largo de los años.
- Los ambientes boscosos y con vegetación estratificada tienen una menor concentración de cianobacterias. Esto podría deberse a que la vegetación está atenuando la aparición de floraciones o que, al ser una región menos intervenida por el hombre, esta menos afectada por disturbios (como el cambio en concentración de nutrientes).
- Se observa una correlación positiva entre la concentración de cianobacterias y la de nutrientes. Los nutrientes que mostraron mayor correlación son el Nitrógeno Total (TKN) y el Carbono Orgánico Total (TOC). Esto puede deberse a que sean nutrientes limitantes en el ambiente, es decir, que están en baja proporción.
- La temperatura bajo el agua tiene mayor relevancia en el crecimiento de las cianobacterias.

Modelo de clasificación

Entrenamiento y crossvalidation

Se entrenaron distintos algoritmos de clasificación para determinar cuál ofrece el mejor rendimiento en términos de precisión. Se aplicó validación cruzada K-Fold para evaluar cada modelo.



RandomForestClassifier, GradientBoostingClassifier y LogisticRegression muestran una precisión promedio más alta.

Hypertunning

Se realizó la optimización de hiperparámetros utilizando GridSearchCV para los algoritmos de RandomForestClassifier, GradientBoostingClassifier y XGBClassifier

La puntuación de accuracy en los tres algoritmos fue similar, todas superiores a 0.7. El modelo con mayor score fue Gradient Boosting (0.752). Esto sugiere un poder predictivo limitado pero aceptable dentro de lo esperable, ya que estos resultados provienen de un conjunto de datos limitado y de datos biológicos, los cuales suelen presentar una alta variabilidad.

También se probó un modelo de ensamblaje entre los mejores modelos de RandomForest, Gradient Boosting y XGBoost. Aunque no se obtuvo un resultado mejor

Evaluación del rendimiento

Para verificar qué tan bien funcionan los modelos, se evaluaron utilizando el conjunto de datos de prueba que no se utilizaron durante el entrenamiento y la validación cruzada.

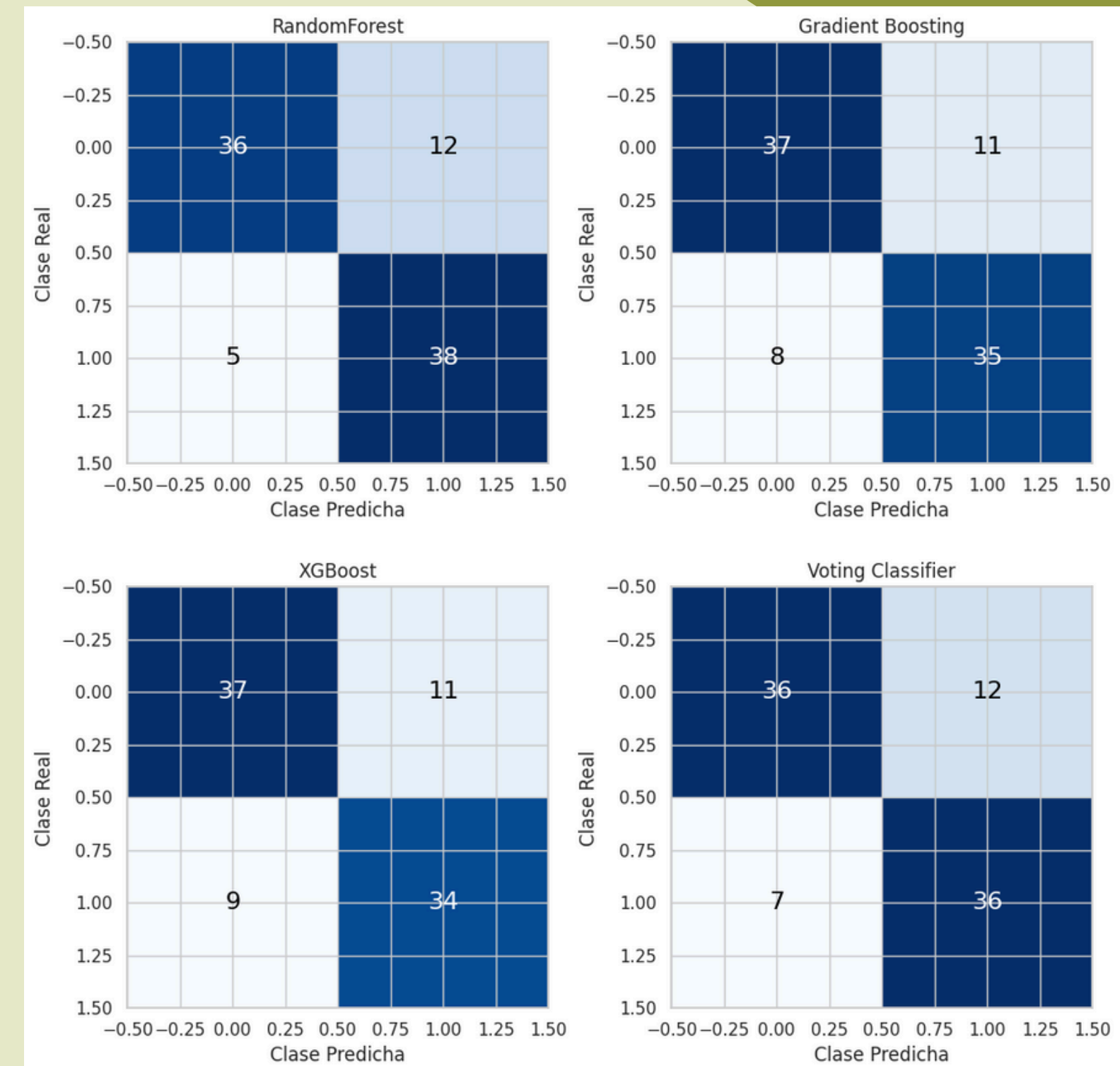
El modelo con mayor precisión al evaluarlo fue Random Forest (accuracy = 0,81).

Matriz de confusión

Los modelos tienen buenos resultados en términos de precisión para ambas clases (0 = <20000 cel/ml y 1 = >20000 cel/ml).

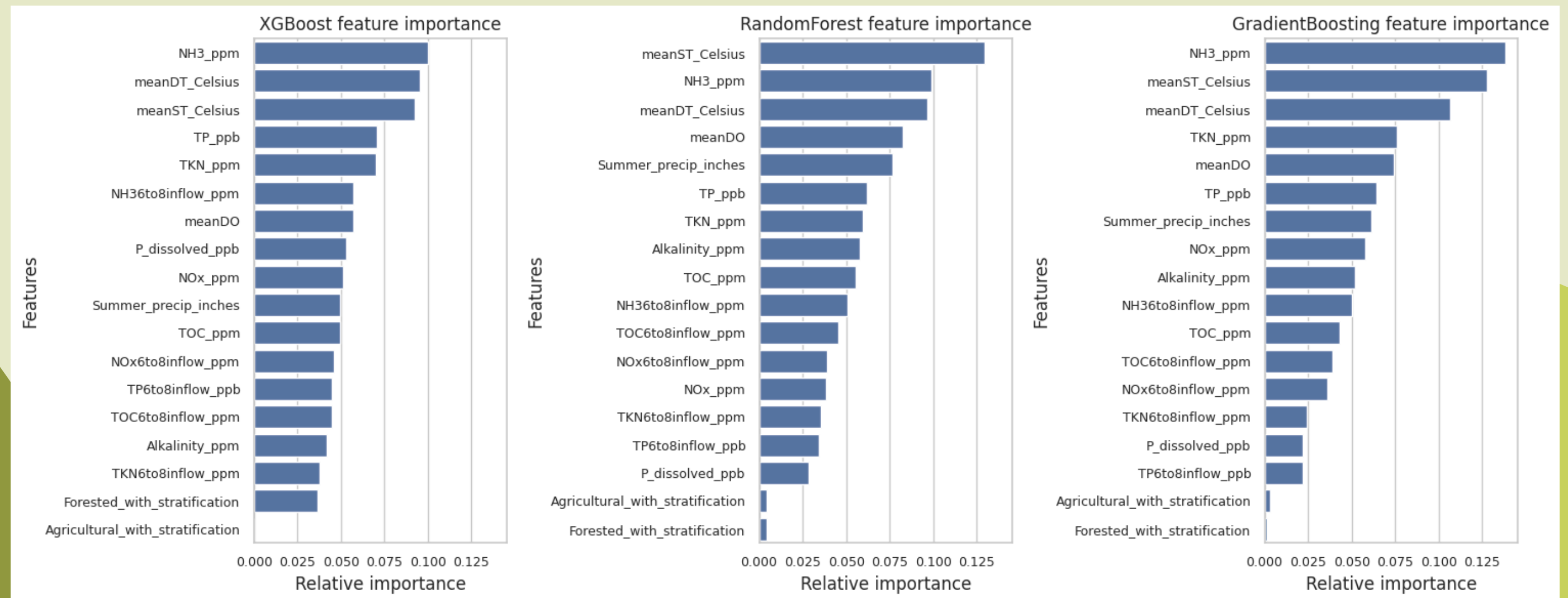
La diagonal principal (0,0 y 1,1) representa las predicciones correctas y son más altas, lo que indica que el modelo tiene un buen desempeño.


No obstante, existen algunos errores. La cantidad de falsos positivos (0,1) y falsos negativos (1,0) muestra el margen de mejora posible.



Explicabilidad

En todos los modelos, la concentración de amoníaco (NH3) y la temperatura del agua (tanto en superficie como en profundidad) son de las variables que más peso tienen en la determinación de una concentración de cianobacterias peligrosa en el agua. Esta relación coincide con investigaciones científicas previas, las cuales indican que la disponibilidad de nutrientes, como el amoníaco, y las condiciones térmicas del medio acuático, son factores determinantes para el crecimiento y proliferación de las cianobacterias





¡Gracias por su
atención!