http://doi.org/10.5281/zenodo.14721530

Optimización de hiperparámetros en algoritmos de *boosting*: aplicación en la biosorción de metales tóxicos

Juan Crescenciano Cruz-Victoria[1] iD, Ignacio Algredo Badillo[2] iD, Alma Rosa Netzahuatl-Muñoz[3]* iD

[1]Programa académico de Ingeniería Mecatrónica, Universidad Politécnica de Tlaxcala. Avenida Universidad Politécnica No. 1, San Pedro Xalcaltzinco. Tepeyanco, Tlaxcala, 90180, México.
[2]CONACYT-INAOE. Luis Enrique Erro #1, Sta María Tonanzintla, Puebla 72840, Puebla, México.
[3]Programa académico de Ingeniería en Biotecnología, Universidad Politécnica de Tlaxcala. Avenida Universidad Politécnica No. 1, San Pedro Xalcaltzinco. Tepeyanco, Tlaxcala, 90180, México.

Email de autor para correspondencia: *almarosa.netzahuatl@uptlax.edu.mx

Recibido: 25 agosto 2024. **Aceptado:** 18 noviembre 2024

RESUMEN

Este estudio evalúa y compara el rendimiento de cinco algoritmos de *boosting* (AdaBoost, Gradient Boosting, LightGBM, XGBoost y CatBoost) para predecir la capacidad de biosorción de cadmio divalente [Cd(II)] y níquel divalente [Ni(II)] por biomasa de la microalga *Chlorella vulgaris*. Se implementó una metodología de optimización de hiperparámetros utilizando Optuna, con 500 ensayos y validación cruzada de 5 pliegues para cada algoritmo, considerando como variables el pH, temperatura, concentraciones iniciales de ambos metales y concentración de biomasa. Durante la optimización, CatBoost mostró la mayor precisión (RMSE < 0.027 mmol g^{-1}, $R^2 > 0.97$) mientras LightGBM destacó por su equilibrio entre rendimiento y eficiencia computacional. En general, los modelos mostraron mayor variabilidad en RMSE que en R^2, y fueron más estables en la predicción de la biosorción de Ni(II). El análisis de importancia de características reveló que las concentraciones iniciales de Cd(II) y Ni(II) fueron las variables más influyentes, capturando la interacción competitiva entre ellos. Esta metodología proporciona una guía robusta para implementar algoritmos de *boosting* en el modelado de procesos complejos de biorremediación, con potencial para optimizar sistemas de tratamiento de aguas residuales.

Palabras clave: *Chlorella*; Cd(II); Ni(II); importancia de características; Optuna; validación cruzada.

ABSTRACT

This study evaluates and compares the performance of five *boosting* algorithms (AdaBoost, Gradient Boosting, LightGBM, XGBoost, and CatBoost) to predict the biosorption capacity of divalent cadmium [Cd(II)] and divalent nickel [Ni(II)] by *Chlorella vulgaris* microalgae biomass. A hyperparameter optimization methodology was implemented using Optuna, with 500 trials and 5-fold cross-validation for each algorithm, considering variables such as pH, temperature, initial concentrations of both metals, and biomass concentration. During optimization, CatBoost achieved the highest accuracy (RMSE < 0.027 mmol g^{-1}, $R^2 > 0.97$) while LightGBM stood out for its balance between performance and computational efficiency. Overall, the models showed higher variability in RMSE than in R^2, and were more stable in predicting Ni(II) biosorption. Feature importance analysis revealed that initial concentrations of Cd(II) and Ni(II) were the most influential variables, capturing the competitive interaction between them. This methodology provides a robust guide for implementing *boosting* algorithms in modeling complex bioremediation processes, with potential to optimize wastewater treatment systems.

Keywords: *Chlorella*, Cd(II), Ni(II), feature importance, Optuna, cross-validation.

INTRODUCCIÓN

Los ensambles son una técnica de aprendizaje automático que combina múltiples modelos para obtener un rendimiento superior al de los modelos individuales [1]. Esta estrategia aprovecha las fortalezas de diversos modelos y compensa sus debilidades mediante la ponderación y agregación de sus predicciones [2].

Entre los métodos de ensamble más destacados se encuentran *bagging (Boostrap Aggregating)*, que entrena modelos en subconjuntos aleatorios de datos [3], *stacking*, que entrena un modelo de nivel superior usando las predicciones de múltiples modelos base [4], y *boosting*, que construye un modelo combinado de forma iterativa, donde en cada iteración se entrena un nuevo modelo que se enfoca en corregir los errores cometidos por el modelo anterior, prestando especial atención a los ejemplos más difíciles de predecir [5, 6].

Algoritmos de *boosting* como *Adaptative Boosting* (AdaBoost), *Gradient Boosting*, *Extreme Gradient Boosting* (XGBoost), *Light Gradient Boosting Machine* (LightGBM) y *Categorical Boosting* (CatBoost) han demostrado un rendimiento superior en comparación con otros enfoques de clasificación y regresión [5, 7]. La eficacia de los algoritmos de *boosting* se ha evidenciado en diversos campos como el financiero, el ambiental y el biomédico [8−11]. Su capacidad para capturar relaciones complejas y no lineales entre variables [8, 12], junto con su robustez

frente a valores atípicos y su habilidad para manejar diferentes tipos de datos [1, 13], los ha posicionado como herramientas poderosas en el análisis predictivo.

Para obtener el máximo rendimiento de estos algoritmos, la optimización de hiperparámetros juega un papel crucial. Los hiperparámetros son configuraciones que determinan cómo el modelo aprenderá de los datos, controlando aspectos como la complejidad del modelo, la tasa de aprendizaje y los mecanismos de regularización, los cuales influyen significativamente en la precisión y generalización del modelo [14, 15]. Para abordar esta dificultad se han propuesto técnicas de búsqueda en cuadrícula, búsqueda aleatoria y optimización bayesiana [9, 15].

En el ámbito de las ciencias ambientales, la remoción de contaminantes por biosorción presenta desafíos particulares para el modelado debido a la compleja interacción de múltiples factores fisicoquímicos [16]. La biosorción es un proceso en el que materiales biológicos se utilizan para eliminar contaminantes de soluciones acuosas a través de mecanismos de adsorción física, microprecipitación, intercambio iónico, complejación, coordinación y quelatación [17, 18]. Las microalgas han demostrado ser biosorbentes eficaces debido a su alta relación superficie-volumen, rápido crecimiento y presencia de diversos grupos funcionales en sus paredes celulares [19]. El género *Chlorella* ha exhibido una notable capacidad para adsorber y acumular metales como el cadmio divalente [Cd(II)] y níquel divalente [Ni(II)] [16, 20]. La biosorción de metales pesados es de gran interés debido a su toxicidad y persistencia en el ambiente [21]. La eficacia de *Chlorella* en la biosorción de Cd(II) y Ni(II) se debe a la presencia de grupos funcionales carboxilos, hidroxilos y aminos en su pared celular, que interactúan con los iones metálicos y facilitan su adsorción [22]. El proceso depende fuertemente del pH, con un intervalo óptimo entre 4.0 y 5.0 para ambos metales. La cinética es rápida, alcanzando el equilibrio entre 15 y 60 minutos, y se ha modelado tradicionalmente con un enfoque univariante de pseudo-segundo orden. Los datos de equilibrio se han ajustado a modelos como las isotermas de Langmuir y Freundlich, con capacidades máximas de biosorción de 60-90 mg g^{-1} para Cd(II) y 50-65 mg g^{-1} para Ni(II). Termodinámicamente, el proceso es espontáneo para ambos metales, siendo endotérmico para Ni(II) y exotérmico para Cd(II). En sistemas binarios, se observa competencia entre Cd(II) y Ni(II), con mayor afinidad por Cd(II) [16, 23−25].

Esta complejidad presenta desafíos significativos para el modelado mecanístico multivariante [26, 27]. Los principales problemas incluyen la inadecuación de modelos convencionales para sistemas de adsorción multicomponente [28], la dificultad para modelar la heterogeneidad de los biosorbentes y los múltiples mecanismos de adsorción [29], y la limitada aplicabilidad de modelos simples en condiciones reales [30]. Estas limitaciones evidencian la necesidad de enfoques más avanzados, como los algoritmos de aprendizaje automático [31], que permitan abordar

eficazmente la naturaleza no lineal y multivariante de los procesos de biosorción.

El objetivo principal de este estudio fue desarrollar una metodología para la optimización y evaluación de algoritmos de *boosting* en la predicción de procesos complejos, utilizando como caso de estudio la biosorción multivariante de Cd(II) y Ni(II) por biomasa de la microalga *Chlorella vulgaris*. Se analizó el desempeño de cinco algoritmos de *boosting* (AdaBoost, Gradient Boosting, XGBoost, LightGBM y CatBoost), optimizando sus hiperparámetros y evaluando su precisión, estabilidad y eficiencia computacional. Este trabajo busca establecer una guía metodológica robusta para la implementación y comparación de algoritmos de *boosting* para el modelado eficaz de relaciones multivariantes complejas y no lineales en distintas áreas científicas.

METODOLOGÍA

Con el objetivo de desarrollar una metodología sistemática para la optimización y evaluación de algoritmos de *boosting* en la predicción de procesos complejos, se seleccionaron cinco algoritmos ampliamente utilizados y de alto rendimiento: AdaBoost, Gradient Boosting, XGBoost, LightGBM y CatBoost. La metodología propuesta se centró en la optimización de los hiperparámetros de estos algoritmos. Aunque la búsqueda en cuadrícula es un enfoque común para la optimización de hiperparámetros, puede resultar computacionalmente costosa. Por otro lado, técnicas más eficientes, como la búsqueda aleatoria y la optimización bayesiana, permiten una exploración más amplia del espacio de búsqueda, lo que posibilita encontrar soluciones óptimas en un menor número de iteraciones [32]. En este estudio, se propone el uso de Optuna, una biblioteca de optimización bayesiana altamente eficiente [33], con el objetivo de mejorar el rendimiento de los algoritmos y reducir el tiempo de cómputo necesario para la optimización de los hiperparámetros.

Para evaluar de manera integral el rendimiento de los algoritmos, se consideraron métricas de precisión, estabilidad y eficiencia computacional. Además, se aplicó la validación cruzada de k-fold para evaluar el rendimiento de los modelos durante la optimización de hiperparámetros, lo que proporciona estimaciones más robustas y reduce el riesgo de sobreajuste [34]. Por último, se realizó un análisis de importancia de características para cada modelo, utilizando métodos específicos de cada algoritmo, con el fin de identificar las variables predictoras más influyentes en el proceso del caso de estudio. En la figura 1 se resume la metodología empleada en este trabajo.

Datos de biosorción

Los datos experimentales sobre la biosorción de Cd(II) y Ni(II) en soluciones acuosas por la microalga *Chlorella vulgaris* se obtuvieron mediante una revisión y recopilación de la literatura científica. Específicamente, se

extrajeron datos de cuatro estudios previamente publicados [16, 23−25]. Estos datos cubren un amplio intervalo de condiciones fisicoquímicas incluyendo pH, concentración inicial de Ni(II), concentración inicial de Cd(II), temperatura y concentración de biomasa. Los estudios abarcan tanto sistemas de un solo componente como sistemas binarios Cd(II)-Ni(II). Los valores de concentración inicial de los metales reportados en mg L^{-1} fueron convertidos a mmol L^{-1} empleando los pesos moleculares del Cd (112.4 g mol^{-1}) y del Ni (58.7 g mol^{-1}). De manera similar, los valores de biosorción en el equilibrio (q_{eq}) reportados en mg g^{-1} fueron convertidos a mmol g^{-1}. En total se recopilaron 96 datos de q_{eq} obtenidas a diferentes condiciones (Tablas S1 y S2).

La Tabla 1 resume las variables utilizadas en este estudio, las variables se clasifican en dos categorías principales en el contexto del aprendizaje automático: a) variables de entrada (características) que son condiciones experimentales controlables para hacer predicciones y b) variables de salida (objetivos) que constituyen las variables a predecir con el modelo, en este caso, las capacidades de biosorción de Ni(II) y Cd(II). Este conjunto de datos conformó la base para el desarrollo de los modelos predictivos utilizando los algoritmos de *boosting* propuestos.

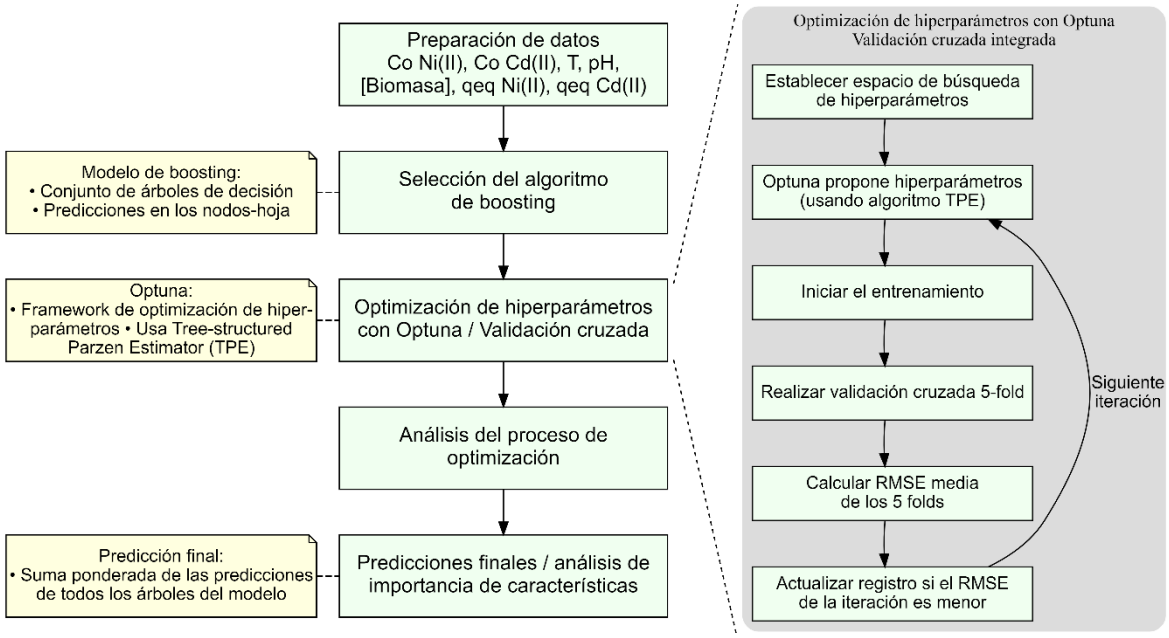

Figura 1. Metodología para el modelado multivariante de la biosorción de Cd(II) y Ni(II) con algoritmos de *boosting*.

Tabla 1. Variables para el modelado de biosorción de Ni(II) y Cd(II)[a]

Tipo de variable	Variable	Clave	Valor mínimo	Valor máximo
Característica	Concentración inicial de Ni(II) (mmol L^{-1})	Co Ni(II)	0	4.41
	Concentración inicial de Cd(II) (mmol L^{-1})	Co Cd(II)	0	1.81
	Temperatura (°C)	T	15.0	50.0
	pH	pH	2.0	6.0
	Concentración de biomasa (g L^{-1})	Biomasa	0.75	1.0
Objetivo	Capacidad de biosorción de Ni(II) (mmol g^{-1})	q_{eq} Ni(II)	0	1.025
	Capacidad de biosorción de Cd(II) (mmol g^{-1})	q_{eq} Cd(II)	0	0.759

[a]Resumen elaborado a partir de los datos reportados en [16, 23–25].

Algoritmos de *Boosting*

Un modelo de *boosting* consta de múltiples árboles de decisión poco profundos (Figura 2). Cada árbol divide los datos basándose en diferentes características y realiza predicciones en sus nodos hoja [35]. Un árbol poco profundo típicamente se refiere a árboles con una profundidad de 1 a 5 niveles, desde los "stumps" de un solo nivel (como en AdaBoost) hasta árboles más complejos, aunque esto puede variar de acuerdo con el algoritmo implementado y la complejidad del problema. La profundidad del árbol determina la complejidad del modelo: árboles más profundos pueden capturar relaciones más complejas, pero también son más propensos al sobreajuste [36]. Este tipo de modelos permite capturar relaciones no lineales complejas incluso con muestras pequeñas [37], lo cual es crucial en estudios experimentales donde la obtención de grandes volúmenes de datos puede consumir una gran cantidad de recursos.

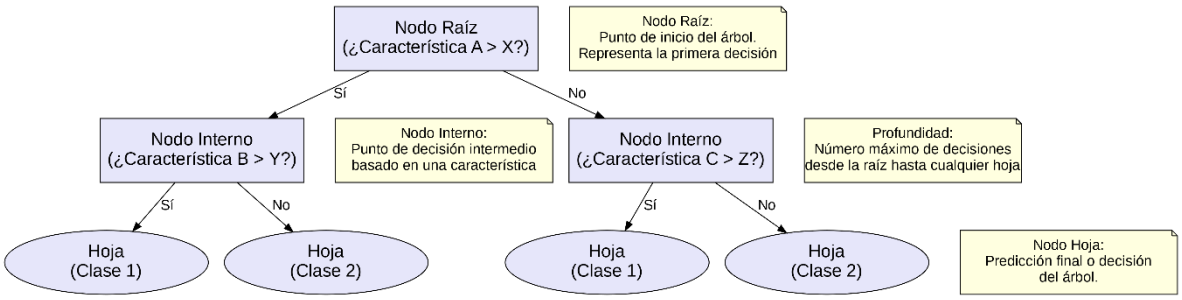

Figura 2. Estructura de árbol de decisión poco profundo

Los algoritmos de *boosting* se fundamentan en tres componentes principales: la función de pérdida, que guía el proceso de aprendizaje; la ponderación de errores, que ajusta la importancia de cada muestra; y el proceso de optimización, que define la estrategia de combinación de modelos base. La Tabla 2 detalla la implementación específica de estos componentes en cada algoritmo.

A continuación, se describen detalladamente los algoritmos implementados, analizando sus características distintivas y el funcionamiento específico de sus componentes:

a) Adaboost

Desarrollado originalmente como un algoritmo de clasificación por Freund y Schapire [43], AdaBoost fue el algoritmo pionero que estableció las bases del *boosting*. Su posterior adaptación para problemas de regresión, conocida como AdaBoost.R2 [38], mantiene el principio fundamental de entrenar iterativamente regresores débiles en versiones ponderadas del conjunto de datos, enfatizando las instancias mal predichas.

El algoritmo actualiza los pesos de las muestras ($w_i \rightarrow w_{i+1}$) mediante una función exponencial que considera la diferencia entre el valor experimental (y_i) y la predicción del modelo ($f(x_i)$), ponderada por un coeficiente de aprendizaje α. La optimización se realiza mediante un sumatorio que contabiliza predicciones que exceden un umbral de error θ, donde I actúa como función indicadora que evalúa si el error $|y_i - f(x_i)|$ supera dicho umbral.

b) Gradient *Boosting*

Optimiza funciones de pérdida diferenciables arbitrarias mediante el gradiente negativo de la función de pérdida. La predicción final suma las contribuciones ponderadas de los modelos débiles, donde una tasa de aprendizaje η controla su impacto para prevenir el sobreajuste [39].

Tabla 2. Comparación de las principales características de los algoritmos de *boosting*.

Algoritmo	Función de Pérdida	Ponderación de Errores	Optimización / *Boosting*	Referencia
AdaBoost	Lineal	$w_{i+1} = w_i \cdot e^{-\alpha_t \cdot \lvert y_i - f(x_i)\rvert}$	$\sum_{i=1}^{n} I(\lvert y_i - f(x_i)\rvert > \theta)$	[38]
GradientBoost	Cuadrática Media	$w_{i+1} = w_i - \eta \cdot \nabla f_p(y, f)$	$f_m(x) = f_{m-1}(x) + \eta \cdot h_m(x)$	[39]
CatBoost	Cuadrática Media	$w_i = \alpha \cdot L'(y_i, f(x_i))$	$\nabla L(y, f)$	[40]
LightGBM	Cuadrática Media	$w_{i+1} = w_i \cdot e^{-\alpha g_i}$	$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$	[41]
XGBoost	Cuadrática Media	$w_j = -\frac{\sum_{i=1}^{n} g_i}{\sum_{i=1}^{n} h_i + \lambda}$	$L = \sum_{i=1}^{n}\left[g_i f(x_i) + \frac{1}{2} h_i f(x_i)^2\right] + \Omega(f)$	[42]

El algoritmo actualiza las predicciones iterativamente, donde $f_m(x)$ se construye añadiendo al modelo anterior $f_{m-1}(x)$ una nueva predicción $h_m(x)$ ponderada por η. La dirección de actualización se determina por el gradiente negativo ∇f_p, que minimiza el error de predicción.

c) CatBoost

Se caracteriza por su manejo nativo de variables categóricas y la implementación de un método de ordenamiento simétrico y *boosting* ordenado para abordar el sesgo de predicción. Incorpora regularización basada en gradientes para un entrenamiento más estable [40]. El algoritmo ofrece implementaciones optimizadas tanto para CPU como para GPU [44], lo que facilita el procesamiento eficiente de grandes volúmenes de datos.

La ponderación de errores se realiza mediante dos elementos: un factor de escala α que controla la magnitud del ajuste, y el gradiente de la función de pérdida $L'(y_i,f(x_i))$ que indica la dirección de mejora para cada muestra. La optimización utiliza este gradiente $\nabla L(y,f)$ para minimizar iterativamente el error de predicción.

d) LightGBM

Diseñado para manejar grandes conjuntos de datos eficientemente, utiliza un enfoque de crecimiento de árbol por hoja (*leaf-wise*) en lugar del enfoque de crecimiento por nivel (*level-wise*) [45]. Esta estrategia, junto con técnicas como el muestreo basado en gradiente (GOSS) y el empaquetado de características exclusivas (EFB), permite mayor precisión manteniendo la misma cantidad de nodos hoja [41].

La ponderación de errores emplea una función exponencial que incorpora el gradiente de la función de pérdida (g_i), permitiendo actualizar los pesos según la dirección del error. La optimización se realiza mediante una función logarítmica que utiliza la tasa de error en cada iteración (ϵ_t), donde la convergencia se alcanza cuando la proporción entre aciertos y errores (representada por el término $\ln((1-\epsilon_t)/\epsilon_t)$) se estabiliza.

e) XGBoost

Implementación eficiente de *Gradient Boosting* que destaca por su enfoque en prevenir el sobreajuste mediante regularización basada en la complejidad del árbol y técnicas de submuestreo de filas y columnas. Incorpora poda de árboles basada en profundidad máxima y manejo automático de valores faltantes, lo que mejora su escalabilidad y rendimiento en diversos escenarios [42].

La ponderación de errores utiliza tanto el gradiente (g_i) como el hessiano (h_i) de la función de pérdida, permitiendo un ajuste más preciso al considerar tanto la pendiente como la curvatura del error. La optimización converge cuando la relación entre el gradiente negativo y el hessiano $(-(\sum g_i)/(\sum h_i + \lambda))$ se estabiliza, donde el término de regularización λ y la función de penalización $\Omega(f)$ aseguran que esta convergencia mantenga un balance entre ajuste y complejidad del modelo.

Implementación Computacional

a) Entorno de Desarrollo y Hardware Utilizado

Los experimentos se llevaron a cabo en una computadora con las siguientes especificaciones: procesador Intel(R) Core(TM) i5-8265U CPU, 12 GB de memoria RAM, tarjeta gráfica NVIDIA GeForce MX110. El entorno de desarrollo utilizado fue Visual Studio Code 1.89.1 (Microsoft Corporation) con Python 3.11.9 (Python Software Fundation).

b) Implementación en Python

La implementación de los algoritmos de *boosting* se realizó utilizando bibliotecas de Python especializadas en aprendizaje automático. Se empleó Scikit-learn para AdaBoost (*AdaBoostRegressor*) [46] y Gradient Boosting (*GradientBoostingRegressor*) [47], mientras que CatBoost (*CatBoostRegressor*) [48], LightGBM (*LGBMRegressor*) [49] y XGBoost (*XGBRegressor*) [50] se implementaron usando sus respectivas bibliotecas homónimas. Cada una de estas clases proporciona una implementación eficiente y altamente configurable de su algoritmo correspondiente, permitiendo el ajuste de hiperparámetros relevantes para cada modelo.

Optimización de hiperparámetros y validación cruzada

Para evaluar el rendimiento de los modelos y mitigar el riesgo de sobreajuste, se empleó una técnica de validación cruzada. Los métodos de validación cruzada de 5 y 10 pliegues son ampliamente reconocidos como prácticas estándar en aprendizaje automático. En este estudio, se optó por una validación cruzada de 5 pliegues (k-fold = 5-fold), debido a que ofrece un buen equilibrio entre el sesgo y la varianza en la estimación del error, y proporciona una eficiencia computacional óptima en comparación con métodos de más pliegues [51, 52].

La metodología de validación cruzada de 5 pliegues divide el conjunto de datos en cinco subconjuntos. En cada iteración, se utilizan cuatro subconjuntos para entrenamiento y uno para validación, rotando este proceso cinco veces [34]. Este enfoque resulta en una partición efectiva de 80% para entrenamiento y 20% para prueba en cada iteración, proporcionando una estimación robusta del rendimiento del modelo en datos no vistos.

Para la optimización de hiperparámetros, se empleó Optuna, una biblioteca de código abierto para Python que utiliza el algoritmo de optimización bayesiana Tree-structured Parzen Estimator (TPE) [33]. Este método guía la búsqueda de forma inteligente, construyendo un modelo probabilístico basado en evaluaciones previas para sugerir configuraciones de hiperparámetros con mayor potencial de mejora.

La Tabla 3 presenta el espacio de búsqueda de hiperparámetros para los algoritmos de *boosting* utilizados en el modelado de biosorción de Cd(II) y Ni(II). Este espacio se

definió considerando estudios reportados en la literatura [53, 54] y experiencia previa [55]. En la tabla S3 se puede consultar la descripción de los hiperparámetros que se optimizaron por algoritmo.

Durante el proceso de optimización se realizaron 500 pruebas por algoritmo de *boosting*. Optuna utilizó la media del Error Cuadrático Medio (RMSE) considerando los cinco conjuntos de prueba de la validación cruzada como criterio para sugerir la siguiente configuración de hiperparámetros. El RMSE de cada pliegue se calculó mediante la siguiente ecuación:

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

donde n es el número de muestras del conjunto de prueba, y_i es el valor real y $\hat{y}_i$ es el valor predicho.

Este enfoque iterativo de validación cruzada no solo permite una exploración eficiente del espacio de hiperparámetros, sino que también proporciona una evaluación más confiable para la selección de los modelos de *boosting* más adecuados, reduciendo así el riesgo de sobreajuste y garantizando un mejor rendimiento en datos no vistos.

Tabla 3. Espacio de búsqueda de hiperparámetros para algoritmos de *boosting* optimizados por Optuna.

Algoritmo	Hiperparámetro	Valor mínimo	Valor máximo
AdaBoost	n_estimators	10	400
	learning_rate	0.01	0.3
	max_depth	1	9
	min_samples_split	2	1
	min_samples_leaf	1	10
CatBoost	Iterations	10	400
	depth	1	9
	learning_rate	0.01	0.3
	random_strength	0	100
	bagging_temperature	0	10
Gradient Boost	n_estimators	10	400
	max_depth	1	9
	learning_rate	0.01	0.3
	min_samples_split	2	10
	min_samples_leaf	1	10
LightGBM	n_estimators	10	400
	num_leaves	2	20
	subsample	0.4	1.0
	colsample_bytree	0.4	1.0
	min_child_samples	1	10
	learning_rate	0.01	0.3
XGBoost	n_estimators	10	400
	lambda	0.01	100
	max_depth	1	9
	eta	0.01	1.0

Evaluación del rendimiento y estabilidad de los modelos

Para evaluar el rendimiento y la estabilidad de los modelos de *boosting* seleccionados se utilizaron dos métricas principales: el RMSE y el coeficiente de determinación (R^2). Se calculó el coeficiente de variación (CV) de ambas métricas considerando los valores obtenidos para RMSE y R^2 de los conjuntos de prueba en la validación cruzada.

El coeficiente de determinación (R^2) se calculó utilizando la siguiente ecuación:

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

donde $\bar{y}$ es la media de los valores experimentales.

El coeficiente de variación (CV) se calculó utilizando la siguiente fórmula:

$$CV = \frac{\sigma}{\mu} \times 100 \quad (3)$$

donde σ y μ son la desviación estándar y la media respectivamente.

El CV es una medida de la dispersión relativa de un conjunto de datos, expresada como porcentaje. Un CV bajo sugiere que los valores de RMSE o R^2 son consistentes entre los diferentes pliegues, indicando que el rendimiento del algoritmo es estable a lo largo de la validación cruzada. En contraste, un CV alto implica una mayor variabilidad en el rendimiento de los modelos entre los pliegues, señalando que el algoritmo podría ser menos estable [56]. Este análisis de CV es particularmente relevante para abordar las preocupaciones sobre el sesgo de selección y la estabilidad del modelo, especialmente en conjuntos de datos pequeños [57]. Al proporcionar una medida cuantitativa de la robustez de los resultados, esta metodología complementa la optimización de hiperparámetros realizada con Optuna, ofreciendo una visión más completa de la confiabilidad y estabilidad de los modelos desarrollados.

Evaluación de la importancia de las características

Se evaluó la importancia de las características para uno de los modelos de cada algoritmo de *boosting* con el fin de identificar las variables predictoras más influyentes en la capacidad de biosorción de Cd(II) y Ni(II) por *Chlorella*. Para AdaBoost y Gradient Boosting, se utilizó el enfoque de disminución media de la impureza (MDI) también conocida como importancia de Gini. En este método, la fracción de muestras a las que contribuye una característica se combina con la disminución de la impureza al dividirlas para crear una estimación normalizada del poder predictivo de esa característica [46, 47]. En XGBoost, la importancia se basa en la contribución fraccional de cada característica al modelo, calculada a partir de la ganancia total de las divisiones de esa característica, donde un mayor porcentaje indica una mayor importancia [50], mientras que LightGBM ofrece por defecto la importancia de las características basada en el conteo (*count-based feature importance*), donde se cuenta el número de veces que una característica se utiliza para

dividir los datos en los árboles del modelo [49]. CatBoost proporciona diferentes medidas de importancia de características; la medida que se utilizó fue la de *PredictionValuesChange*, que mide el cambio en las predicciones del modelo cuando se permutan aleatoriamente los valores de una variable [48].

RESULTADOS Y DISCUSIÓN

Optimización de hiperparámetros para el modelado de biosorción de Cd(II) y Ni(II)

La Tabla 4 muestra el tiempo de ejecución de 500 ensayos en la optimización de hiperparámetros para los algoritmos de *boosting* empleados en el modelado de la capacidad de biosorción de Cd(II) y Ni(II). Se observa que LightGBM exhibe los tiempos de ejecución más cortos, siendo de 58.82 y 67.00 s para el modelado de la capacidad de biosorción de Cd(II) y Ni(II), respectivamente. Esto es consistente con las características de LightGBM, que ha sido diseñado para ser altamente eficiente y manejar grandes conjuntos de datos con tiempos de entrenamiento reducidos. Por otro lado, AdaBoost y CatBoost presentan los tiempos de ejecución más prolongados, superando los 990 segundos en ambos casos. Estos resultados son congruentes con la naturaleza de estos algoritmos, que pueden requerir un mayor tiempo de computación debido a su complejidad y a las técnicas específicas que emplean. En el caso de CatBoost, su método de ordenamiento y la prevención del desplazamiento de predicción pueden contribuir a tiempos de ejecución más largos. AdaBoost, por su parte, puede ser computacionalmente intensivo debido a su naturaleza secuencial y la necesidad de recalcular los pesos de las instancias en cada iteración. Con tiempos de ejecución intermedios se encontraron Gradient Boosting y XGBoost.

En las Tablas 5 y 6 se muestran los valores de los hiperparámetros para tres modelos de cada uno de los algoritmos de *boosting* ensayados para Cd(II) y Ni(II), respectivamente. Estos modelos fueron los que presentaron valores más bajos de la función objetivo durante el proceso de optimización con Optuna.

Tabla 4. Tiempo de ejecución de la optimización de hiperparámetros de algoritmos de *Boosting*.

Algoritmo	**Tiempo de ejecución (s)**	
	Modelado q_{eq} Cd(II)	**Modelado q_{eq} Ni(II)**
AdaBoost	999.99	1021.31
Catboost	992.37	1719.07
Gradient Boosting	362.97	415.69
LightGBM	58.82	67.00
XGBoost	175.08	157.97

Tabla 5. Modelos de *boosting* que mostraron el mejor desempeño en la predicción de capacidad de biosorción de Cd(II).

Algoritmo de Boosting	**Hiperparámetro**	**Modelo**		
		1	**2**	**3**
AdaBoost	n_estimators	265	302	325
	learning_rate	0.0504	0.0277	0.0352
	max_depth	9	8	8
	min_samples_split	9	9	9
	min_samples_leaf	1	1	1
CatBoost	Iterations	321	389	343
	depth	4	4	4
	learning_rate	0.1264	0.1349	0.1467
	random_strength	27	70	35
	bagging_temperature	4.4238	8.5859	4.2887
Gradient Boosting	n_estimators	358	382	373
	max_depth	3	3	3
	learning_rate	0.2924	0.2923	0.2928
	min_samples_split	10	10	10
	min_samples_leaf	2	2	2
LightGBM	n_estimators	209	228	243
	num_leaves	4	4	4
	subsample	0.4259	0.4843	0.5344
	colsample_bytree	0.9066	0.9546	0.9513
	min_child_samples	3	3	3
	learning_rate	0.2995	0.2995	0.2850
XGBoost	n_estimators	319	332	313
	lambda	6.5180	4.7024	9.0538
	max_depth	5	5	5
	eta	0.3454	0.2202	0.2027

Tabla 6. Modelos de *boosting* con mejor desempeño en la predicción de capacidad de biosorción de Ni(II)

Algoritmo de *Boosting*	Hiperparámetros	Modelo		
		1	2	3
AdaBoost	n_estimators	250	266	282
	learning_rate	0.2222	0.2191	0.2140
	max_depth	8	9	8
	min_samples_split	3	4	4
	min_samples_leaf	1	1	1
CatBoost	Iterations	385	393	386
	depth	2	2	2
	learning_rate	0.1663	0.1708	0.1695
	random_strength	20	20	23
	bagging_temperature	2.7483	3.1013	3.5558
Gradient Boost	n_estimators	192	201	206
	max_depth	3	3	3
	learning_rate	0.2378	0.2378	0.2375
	min_samples_split	9	9	9
	min_samples_leaf	2	2	2
LightGBM	n_estimators	203	186	198
	num_leaves	7	7	7
	subsample	0.7059	0.7198	0.7097
	colsample_bytree	0.9817	0.9414	0.9275
	min_child_samples	6	6	6
	learning_rate	0.2815	0.2802	0.2802
XGBoost	n_estimators	235	249	204
	lambda	39.8942	38.3178	39.3045
	max_depth	2	2	2
	eta	0.9961	0.9924	0.9999

Es notable que las configuraciones óptimas de hiperparámetros difieren significativamente entre los mejores modelos para la biosorción de Cd(II) y Ni(II). Por ejemplo, en el caso de CatBoost, los modelos para Cd(II) muestran una profundidad constante de 4, mientras que para Ni(II) la profundidad óptima es de 2. De forma similar, XGBoost para Cd(II) utiliza una profundidad máxima de 5, en contraste con la profundidad de 2 para Ni(II). Es probable que esto refleje las diferencias de las interacciones entre el biomaterial y cada metal, así como y de la influencia de las diversas condiciones fisicoquímicas (como pH, temperatura, y concentración inicial) sobre la capacidad de biosorción de Cd(II) y Ni(II). Esta variación en las relaciones subyacentes podría explicar por qué los algoritmos requieren estructuras de modelo diferentes para capturar eficazmente los patrones de biosorción específicos de cada metal.

Al examinar los hiperparámetros de los mejores modelos obtenidos para la biosorción de Cd(II) por cada algoritmo, se observa una notable

similitud entre estos hiperparámetros. Por ejemplo: en AdaBoost, los tres mejores modelos alcanzaron un número de estimadores (*n_estimators*) entre 265 y 325, una tasa de aprendizaje (*learning*_rate) entre 0.0277 y 0.0504, y una profundidad máxima (*max_depth*) de 8 o 9. De manera similar, para CatBoost, los mejores modelos presentaron una profundidad (*depth*) de 4, (*learning*_rate) entre 0.1264 y 0.1467 y un número de iteraciones (*iterations*) entre 321 y 389. Para la biosorción de Ni(II) se observó una selección de hiperparámetros semejante, de esta forma los mejores modelos de CatBoost alcanzaron una profundidad (*depth*) de 2 y un número de iteraciones (*iterations*) entre 385 y 393. Mientras que XGBoost en la biosorción de Ni(II) los modelos mantienen una profundidad máxima de 2 y valores de lambda muy cercanos (entre 38.3178 y 39.8942).

La convergencia hacia valores similares de hiperparámetros es una característica del funcionamiento de Optuna y sugiere que el algoritmo ha identificado regiones prometedoras en el espacio de búsqueda para este problema específico de modelado de la capacidad de biosorción de Cd(II) y Ni(II). El algoritmo TPE de Optuna construye y actualiza secuencialmente un modelo probabilístico del espacio de hiperparámetros basado en evaluaciones previas, utilizando estimadores de densidad de Parzen [33]. Esta estrategia permite concentrar la búsqueda en áreas de alto rendimiento. La consistencia en los resultados proporciona confianza en la calidad de los modelos seleccionados, demostrando la eficacia del proceso de optimización en la identificación de configuraciones de hiperparámetros que producen un buen rendimiento para cada algoritmo de *boosting*.

Predicciones de biosorción de Cd(II) y Ni(II) por modelos de *boosting* e importancia de las características

A) Adaboost

Se desarrollaron modelos independientes de AdaBoost para predecir la capacidad de biosorción de Cd(II) y Ni(II) por *Chlorella vulgaris*. La Tabla 7 presenta las métricas de desempeño medias de los conjuntos de prueba de los pliegues de la validación cruzada de los tres mejores modelos para cada metal.

Tabla 7. Métricas de desempeño de modelos de Adaboost seleccionados en la predicción de capacidad de biosorción de Cd(II) y Ni(II).

Metal	Modelo	RMSE media (mmol g^{-1})	R^2 media	CV RMSE (%)	CV R^2 (%)
Cd(II)	1	0.05158	0.9309	24.21	4.38
	2	0.05137	0.9302	26.90	4.75
	3	0.05175	0.9294	26.49	4.79
Ni(II)	1	0.05867	0.9652	12.21	0.57
	2	0.06136	0.9608	15.90	1.36
	3	0.05973	0.9331	10.77	0.88

Para Cd(II), los modelos mostraron valores de RMSE media entre 0.05137 y 0.05175 mmol g^{-1}, con R^2 ligeramente superiores a 0.929. Los modelos de Ni(II) presentaron valores de RMSE media entre 0.05867 y 0.06136 mmol g^{-1}, con R^2 entre 0.9331 y 0.9652. Aunque los modelos de Ni(II) alcanzaron R^2 ligeramente superiores, también mostraron valores de RMSE más altos, indicando mayor variabilidad en las predicciones.

La estabilidad de los modelos, evaluada mediante el coeficiente de variación (CV) del RMSE y R^2, mostró diferencias entre los metales. Considerando un CV < 10% como indicador de buena estabilidad y > 20% de variabilidad considerable [56], los modelos de Cd(II) mostraron estabilidad limitada en RMSE (CV entre 24.21% y 26.90%) y moderada en R^2 (CV entre 4.38% y 4.79%). Los modelos de Ni(II) exhibieron excelente estabilidad en R^2 (CV entre 0.57% y 1.36%) y variabilidad moderada en RMSE (CV entre 10.77% y 15.90%).

La Figura 3 ilustra las predicciones de q_{eq} y la importancia relativa de las características para ambos metales. Para Cd(II), se observa un aumento no lineal de q_{eq} con la concentración inicial (Figura 3A), característico de procesos de adsorción que siguen modelos como el de Langmuir. Este comportamiento refleja la transición desde un régimen de adsorción lineal a uno de saturación.

De acuerdo con el análisis de características (figura 3B), la concentración inicial de Cd(II) es, por mucho, la variable más influyente en el modelo, seguida por la temperatura y el pH. Esta jerarquía de importancia de las variables es coherente con el comportamiento típico de los procesos de biosorción, donde la concentración inicial del adsorbato juega un papel crucial en la determinación de la capacidad de adsorción en equilibrio.

La gráfica 3C revela una disminución en la q_{eq} de Ni(II) al aumentar la concentración inicial de Cd(II), lo que sugiere una competencia entre estos iones por los sitios de unión en la biomasa. Si bien la concentración inicial de Ni(II) es la variable más importante para su biosorción (figura 3D), la concentración inicial de Cd(II) tiene gran influencia en el modelo. Esto confirma la relevancia de la interacción competitiva entre ambos metales en el proceso de biosorción.

Las diferencias observadas en el rendimiento y estabilidad de los modelos podrían atribuirse a factores como la distribución de los datos experimentales y la posible presencia de relaciones más lineales en los datos de Ni(II).

B) Catboost

Los modelos de CatBoost para la biosorción de Cd(II) y Ni(II) por *Chlorella vulgaris* exhibieron una mejora sustancial en la precisión predictiva en comparación con AdaBoost (Tabla 8). El RMSE disminuyó a valores entre 0.03172 y 0.03185 mmol g^{-1} para Cd(II), y entre 0.02632 y 0.02722 mmol g^{-1} para Ni(II). Esta mejora en precisión se reflejó también en los coeficientes de determinación (R^2), que superaron 0.97 para Cd(II) y 0.99 para Ni(II).

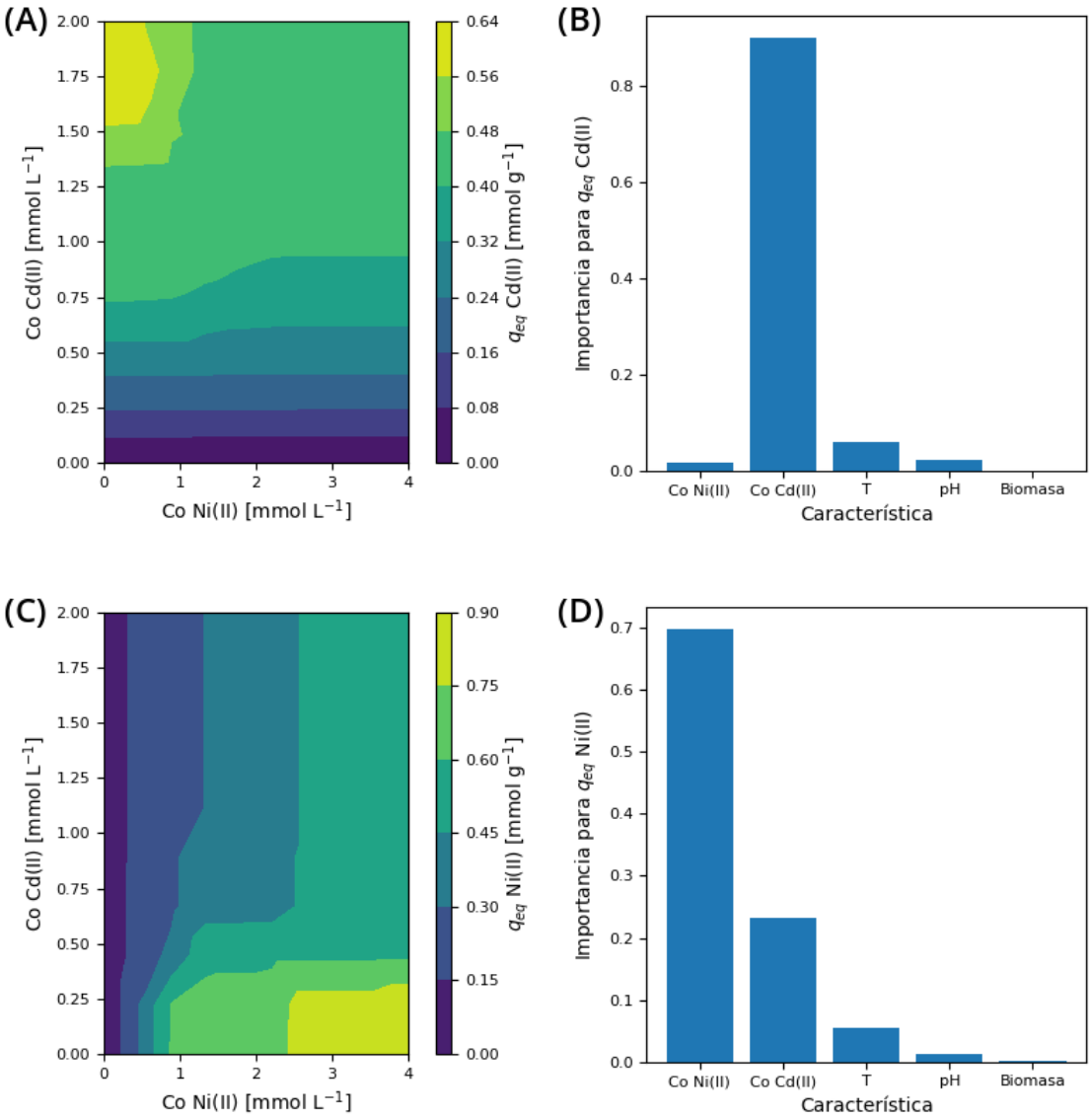

Figura 3. Modelado de la biosorción de Cd(II) y Ni(II) por *Chlorella vulgaris* mediante AdaBoost. A) Predicciones de q_{eq} y B) importancia relativa de las características en la biosorción de Cd(II). C) Predicciones de q_{eq} y D) importancia relativa de las características para la biosorción de Ni(II). T=30 °C, pH=5.0, Biomasa=1 g L^{-1}.

Tabla 8. Métricas de desempeño de modelos de Catboost seleccionados en la predicción de capacidad de biosorción de Cd(II) y Ni(II).

Metal	Modelo	RMSE media (mmol g^{-1})	R^2 media	CV RMSE (%)	CV R^2 (%)
Cd(II)	1	0.03185	0.9732	25.09	1.77
	2	0.03180	0.9737	22.43	1.66
	3	0.03172	0.9734	24.46	1.76
Ni(II)	1	0.02643	0.9924	24.59	0.36
	2	0.02632	0.9922	29.93	0.46
	3	0.02722	0.9917	25.97	0.50

Una característica notable de CatBoost es la estabilidad de sus predicciones en términos de R^2. A pesar de mantener una alta variabilidad en el RMSE (CV > 20%), similar a la observada en AdaBoost, CatBoost logró una estabilidad excepcional en R^2 para ambos metales. Esto se evidencia en los bajos valores de CV para R^2: entre 1.66% y 1.77% para Cd(II), y entre 0.36% y 0.50% para Ni(II). Esta característica sugiere que CatBoost puede ofrecer una mayor consistencia en la capacidad de ajuste del modelo a través de diferentes conjuntos de datos.

Un aspecto significativo de los resultados de CatBoost es la coexistencia de un alto CV en RMSE (>20%) con un bajo CV en R^2 (<2% para Cd(II) y <0.5% para Ni(II)). Esta aparente discrepancia puede atribuirse a las propiedades inherentes de estas métricas. El alto CV del RMSE indica que la magnitud absoluta del error varía considerablemente entre diferentes subconjuntos de datos, posiblemente debido a la heterogeneidad en las condiciones experimentales de biosorción. En contraste, el bajo CV del R^2 sugiere que el modelo es consistente a través de estos subconjuntos. Esta característica es particularmente relevante en sistemas complejos como la biosorción multicomponente, donde las condiciones experimentales pueden tener un impacto sustancial en los valores absolutos, pero las tendencias generales se mantienen relativamente constantes. La capacidad de CatBoost para modelar interacciones complejas se ha observado también en otros estudios ambientales, como en la predicción de la adsorción de selenita en sistemas continuos usando zeolita modificada químicamente [11], y en la predicción del índice de calidad del aire en Visakhapatnam, India [58].

La Figura 4 corrobora las observaciones realizadas con AdaBoost sobre el comportamiento de biosorción, pero con mayor precisión en las predicciones. CatBoost captura con más detalle la no linealidad en la biosorción de Cd(II) y la interacción competitiva entre Cd(II) y Ni(II)

En cuanto al análisis de importancia de las características, CatBoost ofrece una perspectiva más matizada. Aunque confirma la preponderancia de las concentraciones iniciales de los metales, resalta con mayor énfasis la influencia de la concentración de biomasa en la biosorción de Cd(II) y la interacción Cd-Ni en la biosorción de Ni(II).

C) Gradient Boosting

Los modelos de GradientBoost para la biosorción de Cd(II) y Ni(II) por *Chlorella vulgaris* mostraron un rendimiento intermedio en comparación con otros algoritmos de *boosting* evaluados. La Tabla 9 presenta las métricas de desempeño de los tres mejores modelos para cada metal.

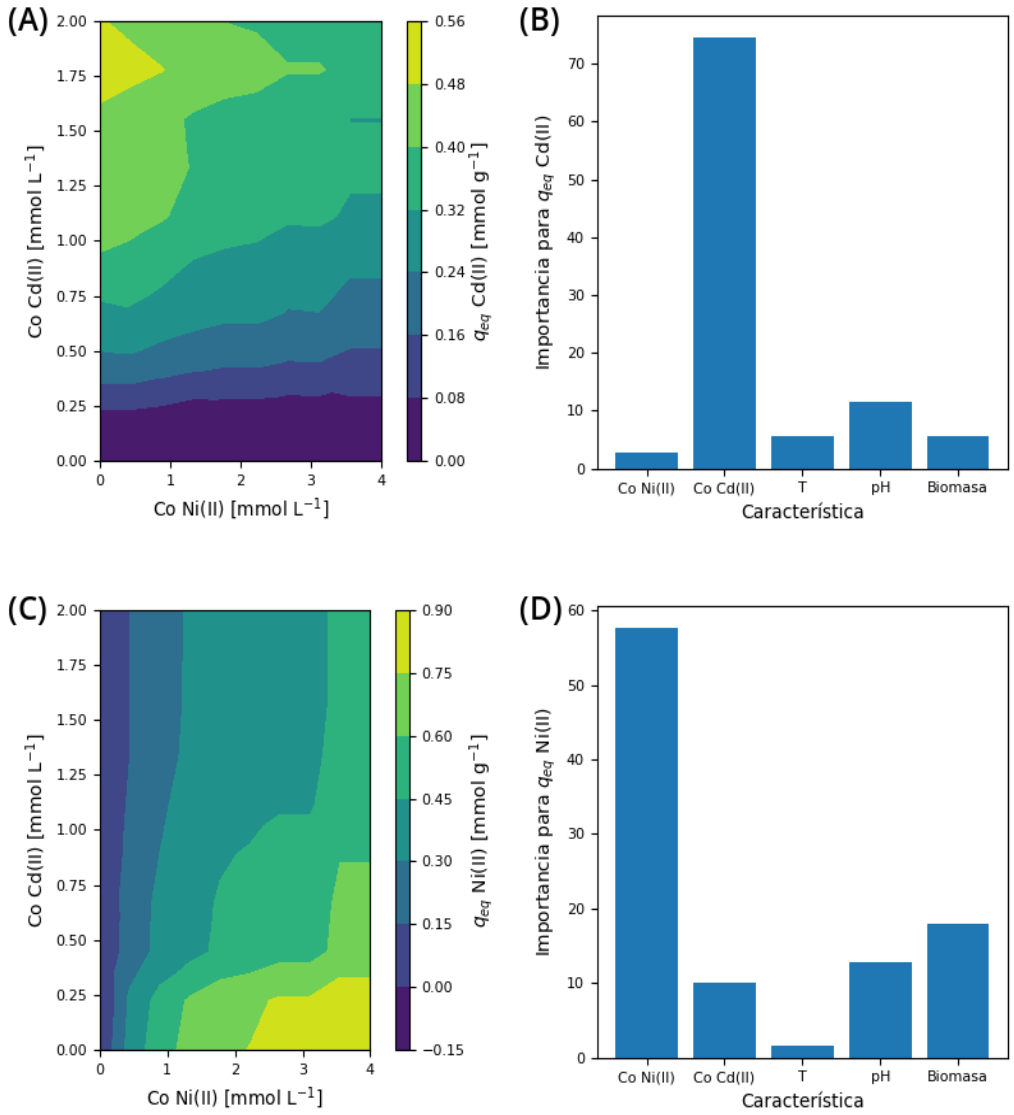

Figura 4. Modelado de la biosorción de Cd(II) y Ni(II) por *Chlorella vulgaris* mediante Catboost. A) Predicciones de q_{eq} y B) importancia relativa de las características en la biosorción de Cd(II). C) Predicciones de q_{eq} y D) importancia relativa de las características para la biosorción de Ni(II).Condiciones: T=30 °C, pH=5.0, Biomasa=1 g L^{-1}.

Tabla 9. Métricas de desempeño de modelos de GradientBoost seleccionados en la predicción de capacidad de biosorción de Cd(II) y Ni(II).

Metal	Modelo	RMSE media (mmol g^{-1})	R^2 media	CV RMSE (%)	CV R^2 (%)
Cd(II)	1	0.04201	0.9492	39.77	4.39
	2	0.04216	0.9490	39.36	4.39
	3	0.04230	0.9487	39.35	4.41
Ni(II)	1	0.03650	0.9843	21.66	1.04
	2	0.03651	0.9843	21.59	1.03
	3	0.03656	0.9843	21.45	1.01

Para la biosorción de Cd(II), los modelos de Gradient Boosting alcanzaron valores de RMSE media entre 0.04201 y 0.04230 mmol g^{-1}, con R^2 ligeramente superiores a 0.948. Estos resultados indican una mejora en la precisión predictiva en comparación con AdaBoost, pero no alcanzan el nivel de rendimiento observado en CatBoost.

En el caso de la biosorción de Ni(II), Gradient Boosting mostró un desempeño notablemente mejor, con valores de RMSE media entre 0.03650 y 0.03656 mmol g^{-1}, y R^2 consistentemente alrededor de 0.9843. Este rendimiento superior para Ni(II) sugiere que GradientBoost pudo capturar mejor las relaciones subyacentes en los datos de biosorción de níquel en comparación con los de cadmio.

En términos de estabilidad, los modelos de Gradient Boosting mostraron una variabilidad considerable, especialmente para Cd(II). Los coeficientes de variación (CV) del RMSE para Cd(II) fueron alrededor del 39%, indicando una alta variabilidad en las predicciones a través de los diferentes pliegues de validación cruzada. Para Ni(II), la variabilidad fue menor pero aún significativa, con CV de RMSE alrededor del 21%. Los CV del R^2 fueron más bajos, alrededor del 4.4% para Cd(II) y 1% para Ni(II), sugiriendo una mayor estabilidad en términos de la proporción de varianza explicada por los modelos.

La Figura 5 ilustra las predicciones de q_{eq} y la importancia relativa de las características para ambos metales utilizando Gradient Boosting. Para Cd(II), se observa un aumento de q_{eq} con la concentración inicial (Figura 5A), con una tendencia clara hacia valores más altos de q_{eq} a medida que aumenta la concentración inicial de Cd(II).

En cuanto a la biosorción de Ni(II), GradientBoost predice un patrón más complejo (Figura 5C). Se observan áreas de mayor q_{eq} tanto a concentraciones bajas como altas de Ni(II), con una zona intermedia de menor biosorción. Además, se aprecia un efecto de la concentración de Cd(II) en la biosorción de Ni(II), sugiriendo una interacción competitiva entre ambos metales.

El análisis de importancia de las características (Figuras 5B y 5D) revela diferencias significativas entre la biosorción de Cd(II) y Ni(II). Para Cd(II), la concentración inicial de Cd(II) domina completamente el modelo, con una importancia relativa cercana a 1.0. Las demás variables, incluyendo la concentración inicial de Ni(II), el pH y la concentración de biomasa, muestran una importancia prácticamente nula.

En contraste, para la biosorción de Ni(II), aunque la concentración inicial de Ni(II) sigue siendo la variable más importante, su dominancia es menos pronunciada. La concentración inicial de Cd(II) muestra una importancia considerable, reforzando la idea de una interacción competitiva entre ambos metales. El pH y la concentración de biomasa, aunque con menor importancia, también parecen tener cierta influencia en el modelo de biosorción de Ni(II).

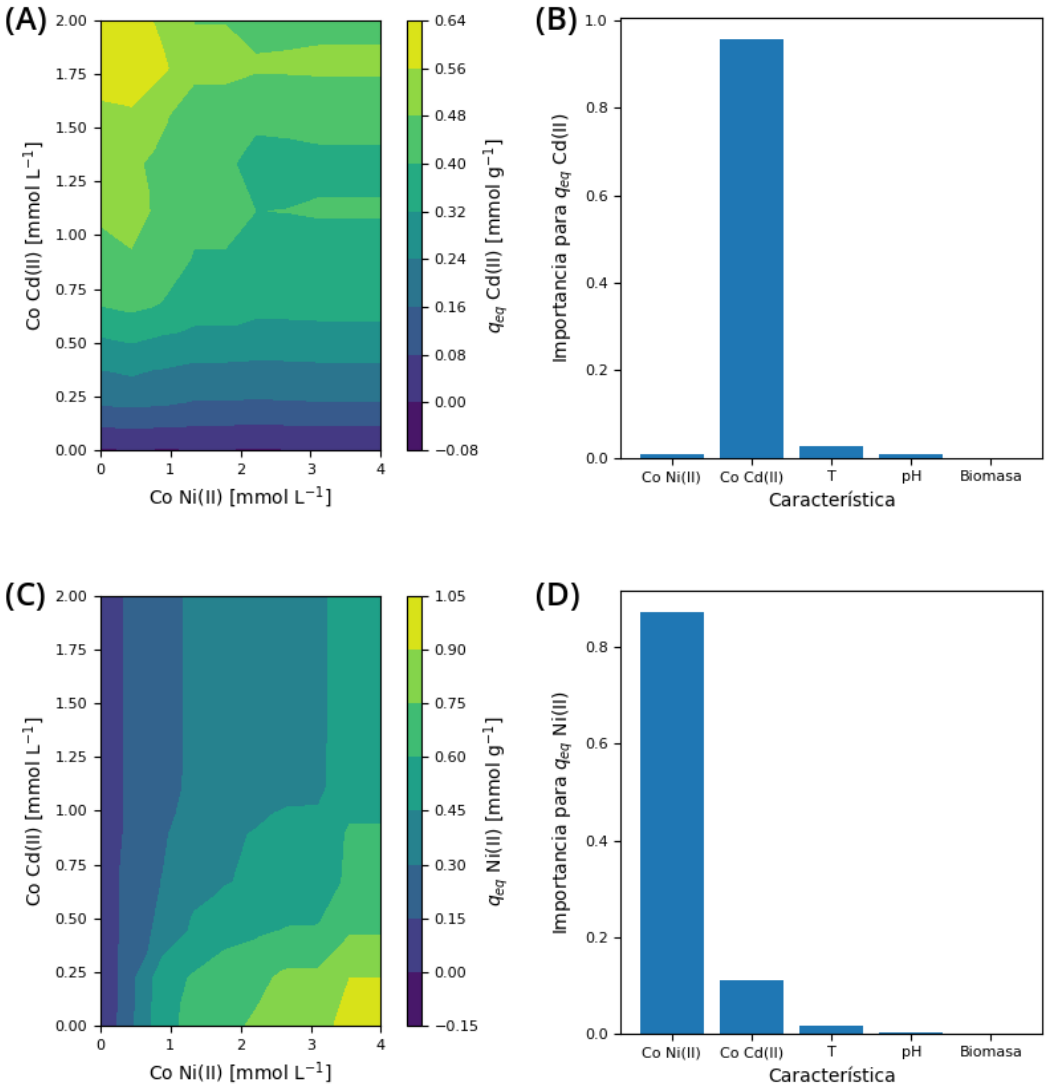

Figura 5. Modelado de la biosorción de Cd(II) y Ni(II) por *Chlorella vulgaris* mediante Gradient Boosting. A) Predicciones de q_{eq} y B) importancia relativa de las características en la biosorción de Cd(II). C) Predicciones de q_{eq} y D) importancia relativa de las características para la biosorción de Ni(II).Condiciones: T=30 °C, pH=5.0, Biomasa=1 g L^{-1}.

Estas diferencias en la importancia de las características sugieren que Gradient Boosting ha capturado dinámicas distintas para la biosorción de Cd(II) y Ni(II). La biosorción de Cd(II) parece ser un proceso más simple, controlado casi exclusivamente por su concentración inicial. En cambio, la biosorción de Ni(II) se presenta como un fenómeno más complejo, influenciado por múltiples factores, incluyendo la presencia competitiva de Cd(II).

La capacidad de Gradient Boosting para modelar estas diferencias podría explicar su mejor rendimiento en la predicción de la biosorción de Ni(II) en comparación con Cd(II). El algoritmo parece ser más efectivo en capturar y representar las interacciones

complejas presentes en la biosorción de Ni(II), mientras que la relación más directa en la biosorción de Cd(II) podría no requerir la complejidad ofrecida por Gradient Boosting.

D) LightGBM

Los modelos de LightGBM para la biosorción de Cd(II) y Ni(II) por *Chlorella vulgaris* mostraron un rendimiento destacable, posicionándose entre los algoritmos con mejores métricas (Tabla 10). Para la biosorción de Cd(II), los modelos alcanzaron valores de RMSE media entre 0.03840 y 0.03856 mmol g^{-1}, con R^2 superiores a 0.961. En el caso de Ni(II), LightGBM mostró un buen desempeño, con valores de RMSE media entre 0.04270 y 0.04332 mmol g^{-1}, y R^2 consistentemente superiores a 0.980. Estos resultados indican una precisión predictiva muy alta, comparable a la de CatBoost y superior a la de GradientBoost y AdaBoost.

En términos de estabilidad, los modelos de LightGBM mostraron una variabilidad moderada. Los coeficientes de variación (CV) del RMSE para Cd(II) fueron entre 20.65% y 22.84%, mientras que para Ni(II) la variabilidad fue aún menor, con CV de RMSE entre 11.09% y 12.09%. Los CV del R^2 fueron notablemente bajos, alrededor del 2.4% para Cd(II) y por debajo del 0.6% para Ni(II), lo que sugiere una alta estabilidad en términos de la proporción de varianza explicada por los modelos.

En términos de estabilidad, los modelos de LightGBM mostraron una variabilidad moderada. Los coeficientes de variación (CV) del RMSE para Cd(II) fueron entre 20.65% y 22.84%, mientras que para Ni(II) la variabilidad fue aún menor, con CV de RMSE entre 11.09% y 12.09%. Los CV del R^2 fueron notablemente bajos, alrededor del 2.4% para Cd(II) y por debajo del 0.6% para Ni(II), lo que sugiere una alta estabilidad en términos de la proporción de varianza explicada por los modelos.

La Figura 6 ilustra las predicciones de q_{eq} y la importancia relativa de las características para ambos metales. Para Cd(II), se observa un aumento claro de q_{eq} con la concentración inicial (Figura 6A), mostrando una relación no lineal característica de los procesos de biosorción. En cuanto a la biosorción de Ni(II), LightGBM predice un patrón más complejo (Figura 6C), donde se observa una influencia significativa tanto de la concentración inicial de Ni(II) como de Cd(II) en la capacidad de biosorción de Ni(II).

El análisis de importancia de las características revela patrones interesantes y distintos para cada metal. Para Cd(II), la concentración inicial de Cd(II) es la variable más importante, seguida por la concentración de biomasa y el pH. La concentración inicial de Ni(II) muestra una importancia relativamente baja. En el caso de Ni(II), aunque su concentración inicial sigue siendo la variable más importante, la concentración inicial de Cd(II) también muestra una influencia significativa, seguida por el pH. La concentración de biomasa tiene una importancia menor en comparación con su rol en la biosorción de Cd(II).

Tabla 10. Métricas de desempeño de modelos de LightGBM seleccionados en la predicción de capacidad de biosorción de Cd(II) y Ni(II).

Metal	Modelo	RMSE media (mmol g^{-1})	R^2 media	CV RMSE (%)	CV R^2 (%)
Cd(II)	1	0.03840	0.9624	20.65	2.40
	2	0.03855	0.9615	21.99	2.38
	3	0.03856	0.9617	22.84	2.43
Ni(II)	1	0.04270	0.9808	11.50	0.55
	2	0.04305	0.9803	12.09	0.60
	3	0.04332	0.9801	11.09	0.58

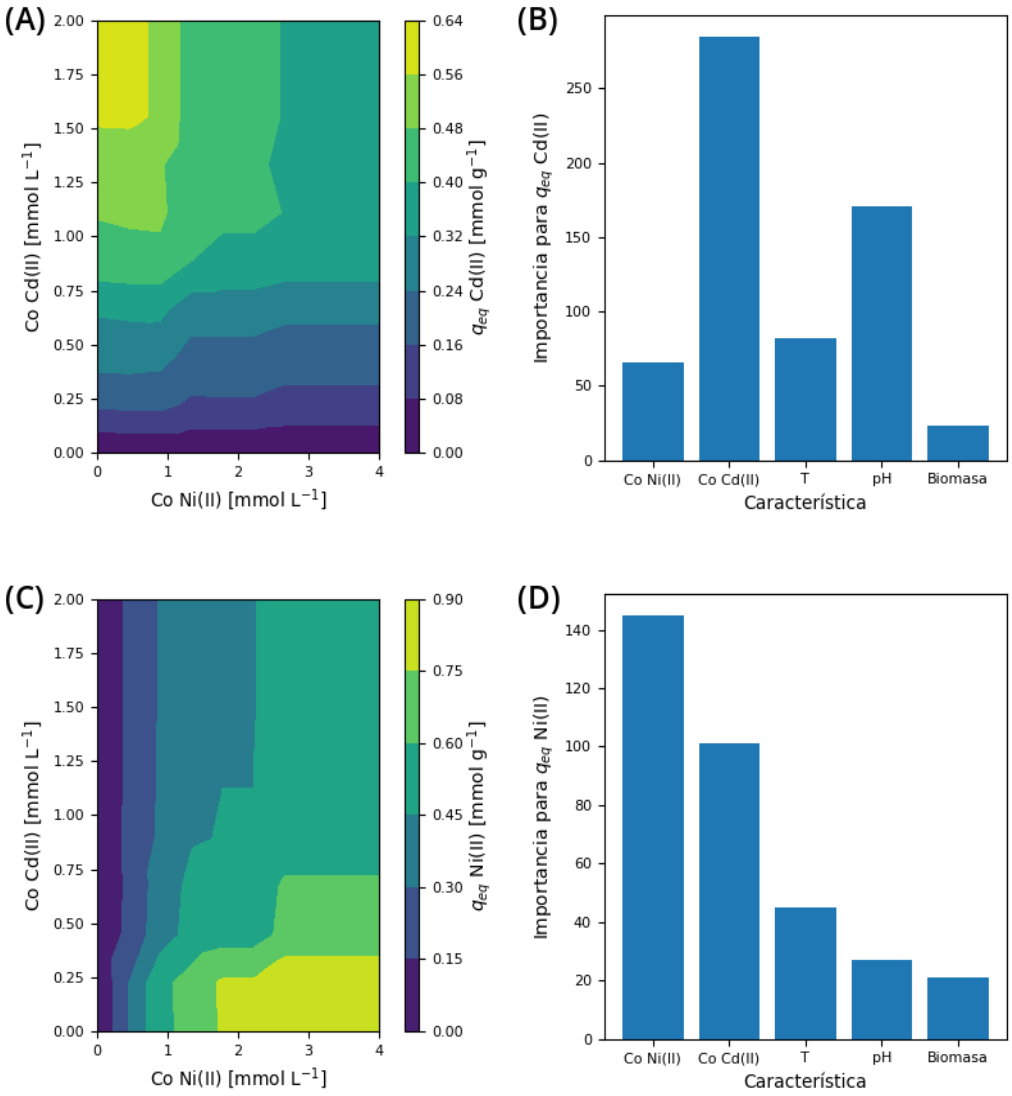

Figura 6. Modelado de la biosorción de Cd(II) y Ni(II) por *Chlorella vulgaris* mediante LightGBM. A) Predicciones de q_{eq} y B) importancia relativa de las características en la biosorción de Cd(II). C) Predicciones de q_{eq} y D) importancia relativa de las características para la biosorción de Ni(II).Condiciones: T=30 °C, pH=5.0, Biomasa=1 g L^{-1}.

Estas diferencias en la importancia de las características sugieren que LightGBM ha capturado dinámicas distintas para la biosorción de Cd(II) y Ni(II). La biosorción de Cd(II) parece estar controlada principalmente por su concentración inicial, el pH y la temperatura, mientras que la biosorción de Ni(II) se ve afectada por todos los factores, incluyendo una interacción competitiva significativa con Cd(II).

La capacidad de LightGBM para modelar estas diferencias y capturar las relaciones no lineales en los datos se refleja en su alto rendimiento predictivo para ambos metales. El modelo logra representar la complejidad de un sistema de biosorción multicomponente, donde las interacciones entre los metales y las variables del proceso juegan un papel crucial.

E) XGBoost

Los modelos de XGBoost para la biosorción de Cd(II) y Ni(II) por *Chlorella vulgaris* mostraron un rendimiento aceptable, aunque no tan destacado como CatBoost o LightGBM. Como se presenta en la tabla 11, para la biosorción de Cd(II), los modelos alcanzaron valores de RMSE media entre 0.04634 y 0.04641 mmol g^{-1}, con R^2 alrededor de 0.947. En el caso de Ni(II), XGBoost mostró un mejor desempeño, con valores de RMSE media entre 0.04738 y 0.04844 mmol g^{-1}, y R^2 superiores a 0.974.

En términos de estabilidad, los modelos de XGBoost mostraron para Cd(II) valores de CV del RMSE relativamente bajos, entre 14.68% y 17.20%, indicando una buena consistencia en las predicciones. Sin embargo, para Ni(II), la variabilidad fue considerablemente mayor, con CV de RMSE entre 39.67% y 42.08%. Los CV del R^2 fueron bajos para ambos metales, alrededor del 2% para Cd(II) y entre 1.63% y 1.76% para Ni(II), sugiriendo una buena estabilidad en términos de la proporción de varianza explicada.

Tabla 11. Métricas de desempeño de modelos de XG Boost seleccionados en la predicción de capacidad de biosorción de Cd(II) y Ni(II).

Metal	Modelo	RMSE media (mmol g^{-1})	R^2 media	CV RMSE (%)	CV R^2 (%)
Cd(II)	1	0.04634	0.9471	17.20	2.37
	2	0.04637	0.9474	14.68	2.13
	3	0.04641	0.9474	15.57	2.07
Ni(II)	1	0.04738	0.9756	42.08	1.76
	2	0.04835	0.9750	39.77	1.63
	3	0.04844	0.9749	39.67	1.68

La Figura 7 ilustra las predicciones de q_{eq} y la importancia relativa de las características para ambos metales. Para Cd(II), se observa un aumento claro de q_{eq} con la concentración inicial (Figura 7A), mostrando una relación no lineal similar a la observada con otros algoritmos. En cuanto a la biosorción de Ni(II), XGBoost predice un patrón más complejo (Figura 7C), donde se aprecia una influencia tanto de la concentración inicial de Ni(II) como de Cd(II).

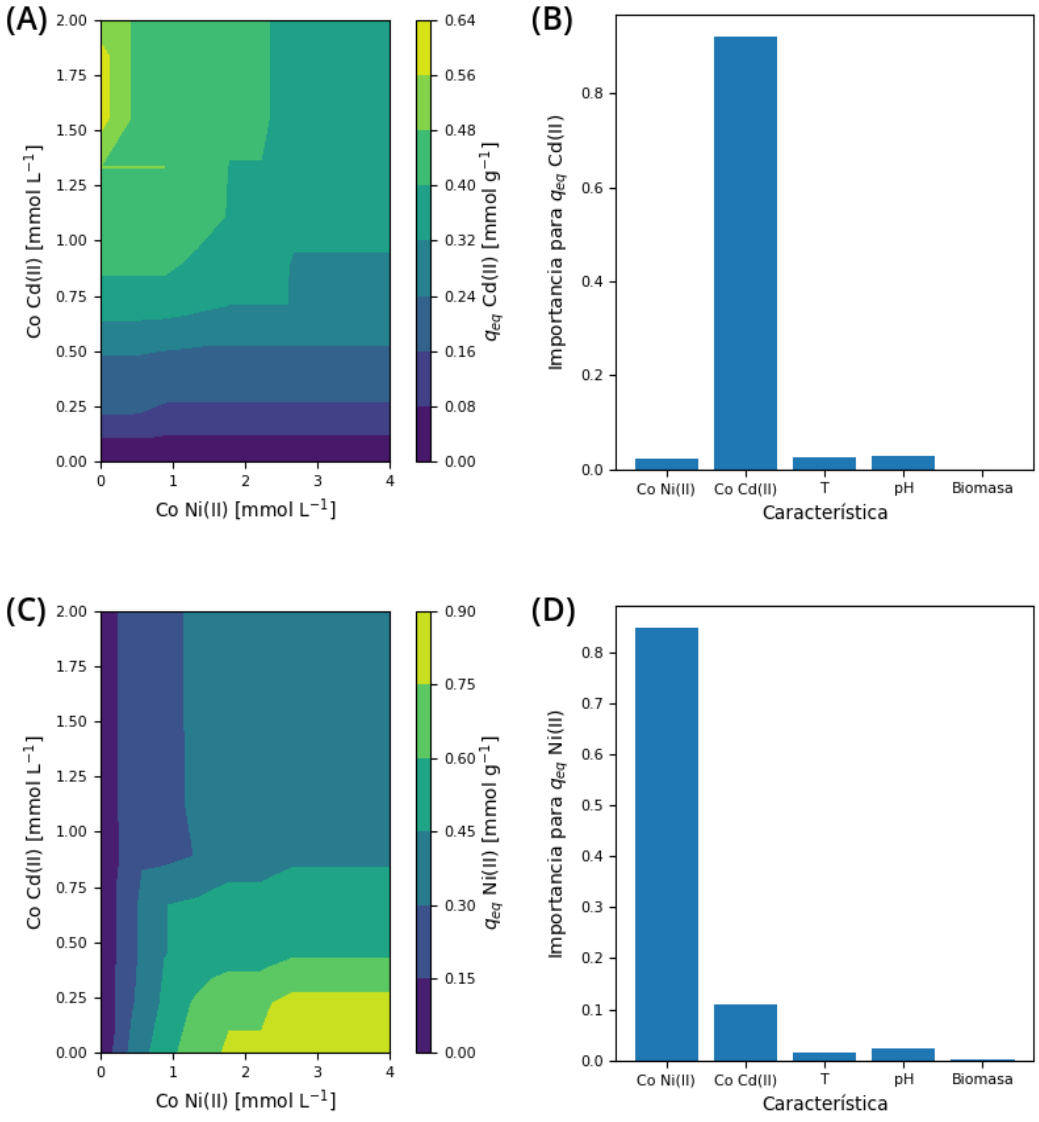

Figura 7. Modelado de la biosorción de Cd(II) y Ni(II) por *Chlorella vulgaris* mediante XG Boost. A) Predicciones de q_{eq} y B) importancia relativa de las características en la biosorción de Cd(II). C) Predicciones de q_{eq} y D) importancia relativa de las características para la biosorción de Ni(II).Condiciones: T=30 °C, pH=5.0, Biomasa=1 g L^{-1}.

El análisis de importancia de las características revela patrones distintos para cada metal. Para Cd(II), la concentración inicial de Cd(II) es, con mucho, la variable más importante, con una importancia relativa cercana a 0.95. Las demás variables muestran una importancia muy baja, por debajo de 0.05. En el caso de Ni(II), aunque la concentración inicial de Ni(II) sigue siendo la variable más importante (importancia relativa de aproximadamente 0.85), la concentración inicial de Cd(II) también muestra una influencia significativa (alrededor de 0.13). Esto sugiere una interacción competitiva entre ambos metales en la biosorción de Ni(II).

Estas diferencias en la importancia de las características indican que XGBoost ha capturado dinámicas distintas para la biosorción de Cd(II) y Ni(II). La biosorción de Cd(II) parece estar controlada casi exclusivamente por su concentración inicial, mientras que la biosorción de Ni(II) se ve afectada tanto por su propia concentración como por la presencia competitiva de Cd(II).

Evaluación comparativa y aplicabilidad de algoritmos de *boosting* en el modelado de procesos de biosorción

La selección inicial de modelos, mediante validación cruzada de 5 pliegues, reveló diferencias significativas entre los algoritmos de *boosting*. Se consideraron umbrales de RMSE (0.038 mmol g^{-1} para Cd(II) y 0.051 mmol g^{-1} para Ni(II)) que representan el 5% de los valores máximos experimentales observados (0.759 mmol g^{-1} para Cd(II) y 1.025 mmol g^{-1} para Ni(II)), y para R^2 se establecieron valores superiores a 0.95. Los modelos CatBoost, GradientBoost, LightGBM y XGBoost se destacaron al superar los criterios establecidos para el modelado de q_{eq} Ni(II). En contraste, CatBoost fue el único modelo que cumplió estrictamente con los criterios para q_{eq} Cd(II), mientras que LightGBM obtuvo valores de RMSE cercanos al límite, superando el 0.95 para R^2.

Tras la evaluación inicial de los algoritmos de *boosting*, se procedió a realizar predicciones en el total de las instancias experimentales utilizando el mejor modelo obtenido para cada algoritmo y cada ion metálico. Los resultados detallados se presentan en el material suplementario (Tablas S1 y S2, Figuras S1 y S2). El análisis revela una buena concordancia general entre los valores experimentales y predichos de q_{eq} para ambos iones metálicos. Se observaron algunas desviaciones moderadas en la predicción de capacidades de biosorción, como por ejemplo en temperaturas superiores a 40 °C y en condiciones de pH bajo (pH=2.0), entre otros casos. Esta limitación se atribuye principalmente a la escasez de datos experimentales en estas condiciones específicas. Es importante señalar que la mayoría de los estudios de biosorción emplean la técnica de una variable a la vez, lo que resulta en una distribución de datos no óptima para el desarrollo de modelos de aprendizaje automático. No obstante, una ventaja significativa de este enfoque es la posibilidad de añadir datos y refinar el modelo de manera iterativa. La incorporación de datos más

diversos podría mejorar sustancialmente la robustez y precisión de los modelos, especialmente en estas situaciones límite [59].

En la Tabla 12 se presentan las métricas de desempeño de los modelos seleccionados, revelando un sobresaliente rendimiento para ambos metales, Cd(II) y Ni(II). Los valores de RMSE se mantuvieron consistentemente por debajo de 0.03 mmol g^{-1}, mientras que los R^2 superaron 0.98 en todos los casos. Notablemente, estas métricas globales muestran una mejora respecto a los resultados obtenidos durante la validación cruzada de 5 pliegues utilizada para la selección de modelos. Esta mejora se atribuye principalmente al uso del conjunto completo de datos en la evaluación global. Es importante destacar que la mejora se observa de manera uniforme en todos los modelos y para ambos metales, lo que sugiere un patrón sistemático más que un sobreajuste aleatorio. Esta consistencia en el alto rendimiento valida la eficacia de la metodología de selección y optimización, indicando la robustez y estabilidad de los modelos seleccionados.

La capacidad de todos los algoritmos para predecir con precisión la biosorción en diversas condiciones experimentales valida la robustez del enfoque adoptado y sugiere su potencial aplicabilidad en el modelado de procesos de biosorción de metales pesados.

Al comparar estos resultados con otros estudios de modelado multivariante de biosorción, se observa que los modelos de *boosting* optimizados en este trabajo ofrecen un rendimiento equivalente o ligeramente superior en términos de precisión y capacidad predictiva. Los modelos desarrollados alcanzaron valores de R^2 mínimos de 0.9819 para Cd(II) y 0.9899 para Ni(II), que son comparables a los resultados reportados en

Tabla 12. Métricas de desempeño de los modelos de *boosting* optimizados en la predicción de la capacidad de biosorción de Cd(II) y Ni(II) para todas las instancias experimentales (n=96).

Metal	Modelo	RMSE (mmol g^{-1})	R^2
Cd(II)	Adaboost	0.02907	0.9819
	Catboost	0.02035	0.9911
	GradientBoost	0.02135	0.9902
	LGBMBoost	0.02648	0.9850
	XGBoost	0.02592	0.9856
Ni(II)	Adaboost	0.02720	0.9930
	Catboost	0.01953	0.9964
	GradientBoost	0.01456	0.9980
	LGBMBoost	0.02246	0.9952
	XGBoost	0.03257	0.9899

estudios previos. Por ejemplo, Chen *et al*. [60] emplearon seis algoritmos de aprendizaje automático diferentes para predecir la capacidad de adsorción de Cd(II) en biocarbón, obteniendo un R^2 de 0.971 para el mejor modelo (CatBoost). En otro estudio multivariante, Hafsa *et al*. [61] aplicaron diferentes modelos a la biosorción de Cd(II) en biomasa de *Spirulina*, logrando un R^2 de 0.98 con el modelo *Bayesian Additive Regression Tree* (BART) para un conjunto de datos, y un R^2 de 0.97 con el modelo *Support Vector Regression* con kernel de Función de Base Radial (SVR-RBF) para otro conjunto.

La obtención de modelos que capturen efectivamente el fenómeno de biosorción tiene implicaciones significativas para el tratamiento de efluentes industriales. Si bien los modelos de *boosting* demuestran una alta precisión predictiva para la biosorción competitiva de múltiples metales, presentan tanto desafíos prácticos como limitaciones teóricas. En el aspecto práctico, la variabilidad en la composición de efluentes industriales reales y la presencia de matrices complejas requeriría una actualización continua de los modelos, mientras que su integración con sistemas de control en tiempo real necesitaría considerar la velocidad de procesamiento y la capacidad de adaptación a cambios en las condiciones de operación [62, 63]. En cuanto a las limitaciones teóricas, la complejidad inherente de estos modelos, que involucran múltiples árboles de decisión y complejas interacciones entre variables, dificulta la traducción de sus predicciones a ecuaciones físicas simples. Además, al estar optimizados principalmente para la precisión predictiva, estos modelos no necesariamente capturan los mecanismos subyacentes del proceso de biosorción de manera explícita.

No obstante, el análisis de importancia de características realizado en este estudio proporciona información valiosa sobre las variables que más influyen en el proceso de biosorción. Esta información, aunque parcial, puede guiar futuras investigaciones hacia aspectos clave del fenómeno, permitiendo un enfoque más dirigido en la exploración de los mecanismos físicos y químicos involucrados. Por ejemplo, la identificación de la concentración inicial del metal como la variable más influyente en todos los modelos es congruente con los principios termodinámicos de la adsorción. Además, la mayoría de los algoritmos capturan la interacción competitiva entre Cd(II) y Ni(II), aunque con diferentes grados de énfasis, CatBoost y XGBoost tienden a resaltar más esta interacción, especialmente en la biosorción de Ni(II).

La evaluación sistemática de algoritmos de *boosting* ha evidenciado que cada algoritmo tiene fortalezas y limitaciones que pueden influir en su desempeño en el modelado de problemas específicos [64]. Sin embargo, un proceso riguroso de selección y optimización permite obtener modelos confiables para la predicción y comprensión de sistemas complejos. La consistencia en los resultados obtenidos valida este enfoque metodológico como una herramienta robusta para el estudio de procesos de biosorción multivariante.

CONCLUSIONES

Se desarrolló y evaluó una metodología sistemática para la optimización y comparación de algoritmos de *boosting* en la predicción de la capacidad de biosorción de Cd(II) y Ni(II) por *Chlorella vulgaris*. La optimización bayesiana de hiperparámetros con Optuna, combinada con validación cruzada de 5 pliegues, permitió una exploración eficiente del espacio de búsqueda y una evaluación robusta del rendimiento de los modelos. La evaluación mediante múltiples métricas (RMSE, R^2 y CV) fue crucial para seleccionar modelos que equilibran precisión y estabilidad.

CatBoost y LightGBM destacaron durante la etapa de optimización de hiperparámetros, CatBoost por su alta precisión predictiva, y LightGBM por su equilibrio entre rendimiento, estabilidad y eficiencia computacional.

Los modelos optimizados de los cinco algoritmos estudiados (Adaboost, Catboost, GradientBoot, LightGBM y XGBoost) alcanzaron valores de R^2 mínimos de 0.9811 para la predicción de q_{eq} Cd(II) y 0.9899 para q_{eq} Ni(II), con valores de RMSE consistentemente por debajo de 0.03 mmol g^{-1} para ambos metales.

Se identificaron limitaciones en la predicción de capacidades de biosorción en valores de características límite, entre ellas temperaturas superiores a 40 °C y pH bajo (especialmente pH=2.0), sugiriendo áreas para futura investigación y mejora de los modelos.

El análisis de importancia de las características aportó interpretabilidad a los modelos, proporcionando información valiosa sobre la influencia relativa de las variables en el proceso de biosorción, destacando la concentración inicial del metal como la variable más influyente.

Este enfoque metodológico proporciona una guía robusta para la implementación y comparación de algoritmos de *boosting* en diversos campos de la ingeniería ambiental. La implementación y refinamiento de estos modelos promete revolucionar los procesos de biorremediación, contribuyendo a una gestión más eficiente y sostenible de los recursos hídricos.

CONFLICTO DE INTERESES

Los autores declaran que no existe conflicto de interés.

AGRADECIMIENTOS

J.C.-V. y A.R.N.-M agradecen a la dirección de Investigación y Posgrado de Universidad Politécnica de Tlaxcala por el apoyo necesario para llevar a cabo esta investigación.

REFERENCIAS

[1]. Sagi O, Rokach L. Ensemble learning: A survey. WIREs Data Min Knowl Discovery 2018; 8(4): e1249. Disponible en: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1249

[2]. Dietterich TG. Ensemble Methods in Machine Learning. In: Multiple Classifier

Systems. Berlin, Heidelberg: Springer; 2000. p. 1-15. Disponible en: https://link.springer.com/chapter/10.1007/3-540-45014-9_1

[3]. Breiman L. Bagging predictors. Mach Learn 1996; 24(2): 123-40. Disponible en: https://link.springer.com/article/10.1023/A:1018054314350

[4]. Clarke B, Fokoue E, Zhang HH. Principles and Theory for Data Mining and Machine Learning. New York, NY: Springer New York; 2009. Disponible en: https://link.springer.com/book/10.1007/978-0-387-98135-2

[5]. Ferreira AJ, Figueiredo MAT. Boosting Algorithms: A Review of Methods, Theory, and Applications. In: Zhang C, Ma Y, Eds. Ensemble Machine Learning: Methods and Applications. New York, NY: Springer; 2012. p. 35-85. Disponible en: https://link.springer.com/chapter/10.1007/978-1-4419-9326-7_2

[6]. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). Ann Stat 2000; 28(2): 337-407. Disponible en: https://projecteuclid.org/journals/annals-of-statistics/volume-28/issue-2/Additive-logistic-regression--a-statistical-view-of-boosting-With/10.1214/aos/1016218223.full

[7]. Zhang Y, Haghani A. A gradient boosting method to improve travel time prediction. Transp Res Part C Emerging Technol 2015; 58: 308-24. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0968090X15000741

[8]. Ren Y, Zhang L, Suganthan PN. Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article]. IEEE Comput Intell Mag 2016; 11(1): 41-53. Disponible en: https://ieeexplore.ieee.org/document/7379058

[9]. Asif D, Bibi M, Arif MS, Mukheimer A. Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization. Algorithms 2023; 16(6): 308. Disponible en: https://www.mdpi.com/1999-4893/16/6/308

[10]. Li X, Wang L, Sung E. AdaBoost with SVM-based component classifiers. Eng Appl Artif Intell 2008; 21(5): 785-95. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0952197607000978

[11]. Halalsheh N, Alshboul O, Shehadeh A, Al Mamlook RE, Al-Othman A, Tawalbeh M, *et al*. Breakthrough Curves Prediction of Selenite Adsorption on Chemically Modified Zeolite Using Boosted Decision Tree Algorithms for Water Treatment Applications. Water 2022; 14(16): 2519. Disponible en: https://www.mdpi.com/2073-4441/14/16/2519

[12]. Bühlmann P, Hothorn T. Boosting Algorithms: Regularization, Prediction and Model Fitting. Stat Sci 2007; 22(4): 477-505. Disponible en: https://projecteuclid.org/journals/statistical-science/volume-22/issue-4/Boosting-Algorithms-Regularization-Prediction-and-Model-Fitting/10.1214/07-STS242.full

[13]. Mayr A, Binder H, Gefeller O, Schmid M. The Evolution of Boosting Algorithms. From Machine Learning to Statistical Modelling. Methods Inf Med 2014; 53(6): 419-27. Disponible en: https://www.thieme-connect.com/products/ejournals/abstract/10.3414/ME13-01-0122

[14]. Akinpelu DA, Adekoya OA, Oladoye PO, Ogbaga CC, Okolie JA. Machine learning applications in biomass pyrolysis: From biorefinery to end-of-life product management. Digital Chem Eng 2023; 8: 100103. Disponible en: https://www.sciencedirect.com/science/article/pii/S2772508123000212

[15]. Duong-Trung N, Born S, Kim JW, Schermeyer MT, Paulick K, Borisyak M, *et al*. When Bioprocess Engineering Meets Machine Learning: A Survey from the Perspective of Automated Bioprocess Development. Biochem Eng J 2023; 190: 108764. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S1369703X22004338

[16]. Aksu Z, Dönmez G. Binary biosorption of cadmium(II) and nickel(II) onto dried *Chlorella vulgaris*: Co-ion effect on mono-component isotherm parameters. Process Biochem 2006; 41(4): 860-8. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S1359511305004411

[17]. Fomina M, Gadd GM. Biosorption: current perspectives on concept, definition and application. Bioresour Technol 2014; 160: 3-14. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0960852413019421

[18]. Volesky B. Detoxification of metal-bearing effluents: biosorption for the next century. Hydrometallurgy 2001; 59(2): 203-16. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0304386X00001602

[19]. Zeraatkar AK, Ahmadzadeh H, Talebi AF, Moheimani NR, McHenry MP. Potential use of algae for heavy metal bioremediation, a critical review. J Environ Manage 2016; 181: 817-31. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S030147971630425X

[20]. Cheng J, Yin W, Chang Z, Lundholm N, Jiang Z. Biosorption capacity and kinetics of cadmium(II) on live and dead *Chlorella vulgaris*. J Appl Phycol 2017; 29(1): 211-21. Disponible en: https://link.springer.com/article/10.1007/s10811-016-0916-2

[21]. Nathan RJ, Jain AK, Rosengren RJ. Biosorption of heavy metals from water: Mechanism, critical evaluation and translatability of methodology. Environ Technol Rev 2022; 11: 91-117. Disponible en: https://www.tandfonline.com/doi/full/10.1080/21622515.2022.2078232

[22]. Monteiro CM, Castro PML, Malcata FX. Metal uptake by microalgae: underlying mechanisms and practical applications. Biotechnol Prog 2012; 28(2): 299-311. Disponible en: https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.1504

[23]. Aksu Z. Equilibrium and kinetic modelling of cadmium(II) biosorption by *C. vulgaris* in a batch system: effect of temperature. Sep Pur Technol 2001; 21(3): 285-94. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S1383586600002124

[24]. Aksu Z. Determination of the equilibrium, kinetic and thermodynamic parameters of the batch biosorption of nickel(II) ions onto *Chlorella vulgaris*. Process Biochem 2002; 38(1): 89-99. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0032959202000511

[25]. Çetinkaya Dönmez G, Aksu Z, Öztürk A, Kutsal T. A comparative study on heavy metal biosorption characteristics of some algae. Process Biochem 1999; 34(9): 885-92. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0032959299000059

[26]. Witek-Krowiak A, Chojnacka K, Podstawczyk D, Dawiec A, Bubała K. Application of response surface methodology and artificial neural network methods in modelling and optimization of biosorption process. Bioresour Technol 2014; 160: 150-60. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S096085241400039X

[27]. Fertu DI, Bulgariu L, Gavrilescu M. Modeling and Optimization of Heavy Metals Biosorption by Low-Cost Sorbents Using Response Surface Methodology. Processes 2022; 10(3): 523. Disponible en: https://www.mdpi.com/2227-9717/10/3/523

[28]. Foo KY, Hameed BH. Insights into the modeling of adsorption isotherm systems. Chem Eng J 2010; 156(1): 2-10. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S1385894709006147

[29]. Vijayaraghavan K, Yun YS. Bacterial biosorbents and biosorption. Biotechnol Adv 2008; 26(3): 266-91. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0734975008000177

[30]. Zhu X, Wang X, Ok YS. The application of machine learning methods for prediction of metal sorption onto biochars. J Hazard Mater 2019; 378: 120727. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0304389419306703

[31]. Yaseen ZM. An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals: Review, challenges and solutions. Chemosphere 2021; 277: 130126. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0045653521005956

[32]. Imani M, Arabnia HR. Hyperparameter optimization and combined data sampling techniques in machine learning for customer churn prediction: a comparative analysis. Technologies 2023; 11(6): 167. Disponible en: https://www.mdpi.com/2227-7080/11/6/167

[33]. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge

Discovery & Data Mining. New York, NY, USA: Association for Computing Machinery; 2019. p. 2623-31.

Disponible en: https://arxiv.org/abs/1907.10902

[34]. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1995. p. 1137-43. Disponible en: https://dl.acm.org/doi/10.5555/1643031.1643047

[35]. Coadou Y. Boosted Decision Trees and Applications. EPJ Web Conf 2013; 55: 02004. Disponible en: https://www.researchgate.net/publication/260633214_Boosted_Decision_Trees_and_Applications

[36]. Bühlmann P, Yu B. Boosting. WIREs Comput Stat 2010; 2(1): 69-74. Disponible en: https://wires.onlinelibrary.wiley.com/doi/10.1002/wics.55

[37]. Mienye ID, Sun Y. A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. IEEE Access 2022; 10: 99129-49. Disponible en: https://ieeexplore.ieee.org/document/9893798

[38]. Drucker H. Improving regressors using boosting techniques. In: Kaufmann M, ed. ICML '97 Proc. Fourteenth International Conference on Machine Learning. Lille; 1997. p. 107-15. Disponible en: https://www.researchgate.net/publication/2424244_Improving_Regressors_Using_Boosting_Techniques

[39]. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat 2001; 29(5): 1189-232. Disponible en: https://doi.org/10.1214/aos/1013203451

[40]. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2018. p. 6639-49. Disponible en: https://arxiv.org/abs/1706.09516

[41]. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, *et al*. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 3149-57. Disponible en: https://dl.acm.org/doi/10.5555/3294996.3295074

[42]. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery; 2016. p. 785-94. Disponible en: https://arxiv.org/abs/1603.02754

[43]. Freund Y, Schapire RE. A Decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 1997; 55(1): 119-39. Disponible en:

https://www.sciencedirect.com/science/article/pii/S002200009791504X

[44]. Ershov V. High performance Insights from GPU version CatBoost. Comput Tools Educ 2022; 2: 59-73. Disponible en: https://www.researchgate.net/publication/366418837_High_performance_Insights_from_GPU_version_CatBoost

[45]. Zhang Y, Zhu C, Wang Q. LightGBM-based model for metro passenger volume forecasting. IET Intel Transport Syst 2020; 14(13): 1815-23. Disponible en: https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/iet-its.2020.0396

[46]. AdaBoostRegressor. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html

[47]. GradientBoostingRegressor. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

[48]. Overview - CatBoostRegressor. Disponible en: https://catboost.ai/en/docs/concepts/python-reference_catboostregressor

[49]. lightgbm.LGBMRegressor — LightGBM 4.0.0 documentation. Disponible en: https://lightgbm.readthedocs.io/en/stable/pythonapi/lightgbm.LGBMRegressor.html

[50]. XGBoost Documentation — xgboost 2.0.3 documentation. Disponible en: https://xgboost.readthedocs.io/en/stable/index.html

[51]. Normawati D, Ismi DP. K-Fold Cross Validation for Selection of Cardiovascular Disease Diagnosis Features by Applying Rule-Based Datamining. Sig Img Proc Lett 2019; 1(2): 23-35. Disponible en: https://simple.ascee.org/index.php/simple/article/view/3

[52]. Soper DS. Greed Is Good: Rapid Hyperparameter Optimization and Model Selection Using Greedy k-Fold Cross Validation. Electronics 2021; 10(16): 1973. Disponible en: https://www.mdpi.com/2079-9292/10/16/1973

[53]. Alhakeem ZM, Jebur YM, Henedy SN, Imran H, Bernardo LFA, Hussein HM. Prediction of Ecofriendly Concrete Compressive Strength Using Gradient Boosting Regression Tree Combined with GridSearchCV Hyperparameter-Optimization Techniques. Materials 2022; 15(21): 7432. Disponible en: https://www.mdpi.com/1996-1944/15/21/7432

[54]. Shahhosseini M, Hu G, Pham H. Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. Mach Learn Appl 2022; 7: 100251. Disponible en: https://www.sciencedirect.com/science/article/pii/S2666827022000020

[55]. Torres-Alvarado A, Morales-Rosales LA, Algredo-Badillo I, López-Huerta F, Lobato-Baez M, López-Pimentel JC. Trade-Off Analysis of Hardware Architectures for Channel-Quality Classification Models. Sensors 2022; 22(7): 2497. Disponible en:

https://www.mdpi.com/1424-8220/22/7/2497

[56]. He T, Niu D, Chen G, Wu F, Chen Y. Exploring Key Components of Municipal Solid Waste in Prediction of Moisture Content in Different Functional Areas Using Artificial Neural Network. Sustainability 2022; 14(23): 15544. Disponible en: https://www.mdpi.com/2071-1050/14/23/15544

[57]. Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 2010; 11(Jul): 2079-107. Disponible en: https://dl.acm.org/doi/10.5555/1756006.1859921

[58]. Ravindiran G, Hayder G, Kanagarathinam K, Alagumalai A, Sonne C. Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam. Chemosphere 2023; 338: 139518. Disponible en: https://www.sciencedirect.com/science/article/pii/S004565352301785X

[59]. Cruz-Victoria JC, Netzahuatl-Muñoz AR, Cristiani-Urbina E. Long Short-Term Memory and Bidirectional Long Short-Term Memory Modeling and Prediction of Hexavalent and Total Chromium Removal Capacity Kinetics of Cupressus lusitanica Bark. Sustainability 2024; 16(7): 2874. Disponible en: https://www.mdpi.com/2071-1050/16/7/2874

[60]. Chen L, Hu J, Wang H, He Y, Deng Q, Wu F. Predicting Cd(II) adsorption capacity of biochar materials using typical machine learning models for effective remediation of aquatic environments. Sci Total Environ 2024; 944: 173955. Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S0048969724041032

[61]. Hafsa N, Rushd S, Al-Yaari M, Rahman M. A Generalized Method for Modeling the Adsorption of Heavy Metals with Machine Learning Algorithms. Water 2020; 12(12): 3490. Disponible en: https://www.mdpi.com/2073-4441/12/12/3490

[62]. Imen S, Croll HC, McLellan NL, Bartlett M, Lehman G, Jacangelo JG. Application of machine learning at wastewater treatment facilities: a review of the science, challenges and barriers by level of implementation. Environ Technol Rev 2023; 12(1): 493-516. Disponible en: https://www.tandfonline.com/doi/full/10.1080/21622515.2023.2242015

[63]. Singh BJ, Chakraborty A, Sehgal R. A systematic review of industrial wastewater management: Evaluating challenges and enablers. J Environ Manage 2023; 348: 119230. Disponible en: https://www.sciencedirect.com/science/article/pii/S0301479723020182

[64]. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev 2021; 54(3): 1937-67. Disponible en: https://link.springer.com/article/10.1007/s10462-020-09896-5