Universidad de Costa Rica

Clasificación de la calidad del agua y predicción de componentes químicos de cuencas en Corcovado.

Andrey Prado, Cristhofer Urrutia, Joseph Romero, Gabriel Valverde, Diego Vega

Escuela de Matemática, Facultad de Ciencias, Universidad de Costa Rica, Sede Rodrigo Facio, San Pedro, Montes de Oca, San José, Costa Rica
joseandrey.prado@ucr.ac.cr cristhofer.urrutia@ucr.ac.cr joseph.romero@ucr.ac.cr
gabriel.valverdeguzman@ucr.ac.cr diego.vegaviquez@ucr.ac.cr,

Resumen

Este estudio tuvo como objetivo clasificar la calidad del agua y predecir componentes químicos en las cuencas de Corcovado, Costa Rica, utilizando indicadores clave como la Demanda Bioquímica de Oxígeno (DBO) y la Demanda Química de Oxígeno (DQO). Se implementó un modelo predictivo basado en el algoritmo XGBoost para analizar datos históricos y evaluar la contaminación en cuerpos de agua dulce y salada. Los resultados mostraron que variables como la temperatura del aire y del agua, así como la ubicación geográfica, fueron determinantes en los niveles de DQO. Además, se identificaron diferencias significativas en la contaminación entre épocas secas y lluviosas, con valores más altos de DBO en verano. El modelo demostró ser útil para predecir la calidad del agua, aunque se observaron limitaciones en la precisión para ciertos meses, como febrero. Estos hallazgos resaltan la importancia del monitoreo continuo y la implementación de estrategias diferenciadas para la gestión de recursos hídricos en la región.

Palabras Clave: Calidad del agua, DBO, DQO, XGBoost, contaminación, predicción, Corcovado.

1. Introducción

El tratamiento de aguas residuales es esencial para preservar el bienestar de las comunidades y proteger el medio ambiente. A medida que crecen las poblaciones y las actividades humanas, la correcta disposición de los desechos líquidos domésticos, comerciales e industriales se ha convertido en un desafío. Las fuentes naturales muchas veces no pueden absorber la carga contaminante, por lo que es necesario tratar las aguas residuales antes de su descarga, asegurando que su retorno al entorno no cause deterioro ambiental.

En Costa Rica, el tratamiento de aguas residuales enfrenta importantes desafíos técnicos, económicos e institucionales. La mayoría de los sistemas utiliza lodos activados y realiza vertidos directos a cuerpos receptores, mientras que los sistemas anaeróbicos presentan limitaciones por falta de capacitación y confianza en su operación. Además, existen

barreras económicas como la baja sostenibilidad financiera y la escasa capacidad de pago de los usuarios. A nivel institucional, se requiere una mejor coordinación del sector y una normativa clara que fomente la innovación. Para avanzar hacia una economía circular, es necesario un cambio cultural que reconozca el valor de las aguas residuales más allá de su descarte.

"El tratamiento de las aguas residuales puede definirse como la descontaminación o remoción de contaminantes del agua, tras ser usada en alguna actividad humana, hasta alcanzar una calidad compatible con su descarga en el ambiente"(Centeno Mora, Cruz Zúñiga, y Vidal Rivera, 2024).

Para un análisis efectivo sobre los cuerpos acuíferos y su contaminación asociada, es importante esclarecer el concepto de año hidrológico. Este viene a ser el periodo en el que inician las lluvias hasta el fin de la época seca. En Costa Rica, este periodo hidrológico va desde mayo hasta octubre -finaliza el periodo lluvioso- y luego de diciembre a marzo -comienza el periodo lluvioso-; los meses de noviembre y abril se consideran de transición (Meléndez Carranza, 2018).

Ahora bien, tener esta noción en los periodos de precipitación es útil en caso de querer evaluar contaminación en algún cuerpo acuífero, pues según Solano Arce (2011) en el verano es cuando más aumentan los valores de DBO y DQO; mientras que en el invierno se hallaron más metales suspendidos. Más adelante, se estudiará como otro tipo de factores pueden revertir los niveles de DB y DQO para distintas épocas del año hidrológico.

Se define la DQO como la *Demanda Química de Oxígeno* y representa la cantidad de oxígeno que se requiere para oxidar el total de la materia orgánica presente en una muestra de agua residual hasta dar dióxido de carbono y agua como productos finales bajo condiciones controladas (Hernandez, 2016). En tanto la DBO se define como la *Demanda Bioquímica de Oxígeno* y cuantifica el oxígeno que consumen los microorganismos como las bacterias y hongos durante la degradación de sustancias en orgánicas (Induanalisis, 2019). En ambos casos, mayores niveles de cuaqluiera indican una mayor contaminación. En Costa Rica, según el Sistema Costarricense de Información Jurídica , los niveles permitidos para los parámetros universales de agua residuales vertidas en un cuerpo receptor deben ser menores de $50mg/L$ para el DBO y menor a $150mg/L$ para el DQO.

Teniendo el conocimiento sobre DBO, DQO y qué son los metales pesados, se pueden analizar estudios dentro del país cuyas principales herramientas fueron la DBO y DQO.

- Microcuenca del río Damas (2011): Esta cuenca se ubica dentro del cantón de Desamparados. Cerca de la zona, se ha presentado una problemática en relación al hacinamiento cerca de la cuenca. Esto ha provocado que quienes en dichas condiciones, arrojen los desechos a los ríos cercanos. En el estudio, se hizo una comparación entre los niveles de DBO y DQO entre el verano y el invierno, los resultados dictaminaron mayores niveles en el verano que en el invierno. No se

encontraron mayores cantidades de metales pesados. Se sugiere que esto es debido a que en invierno y por las corrientes fluviales, los desechos del verano se desplazan por toda la cuenca generando una uniformidad casi impercetible de contaminación en los puntos analizados. Estudio realizado por Solano Arce (2011).

- Aguas residuales en el cantón de San Pablo (2014-2018): Se identificó una alta contaminación en los drenajes de metales pesados. Se le atribuye por residuos provenientes de farmacias (que forman parte importante de la economía del cantón). En cuanto a los indicadores DBO y DQO, demostraron ser altos debido a la cultura cantonal de depositar los desechos vertiéndolos en quiebras y ríos. Se identificó a a Quebrada Getrudis como la más afectada por desechos orgánicos. Estudio realizado por Sánchez-Gutiérrez, Pérez-Salazar, y Alfaro-Chinchilla (2021).
- Microcuenca río Ocloro (2020): En el estudio se concluye que en esta microcuenca la mayoría de ríos están gravemente contaminados. Se utilizaron parámetros DBO y DQO además de ciertos niveles de metales pesados que resultaron estar en niveles altos. La mayor actividad comercial de la zona es la de restaurantes y sodas, seguidas de talleres automotrices. Se determinó que la causa de la contaminación fue por desechos orgánicos y de hidrocarburos; siendo esta última medular en materia sanitaria pues hubo casos de personas que al entrar en contacto con algún río de la cuenca sufrían irritaciones cutáneas. Estudio realizado por Chaves-Villalobos y cols. (2023).

Para contextualizar este trabajo y analizar posibles causalidades, se necesita dilucidar el contexto industrial de estas regiones.

La península de Osa se caracteriza por ser una de las zonas del país con mayor afloro de turistas tanto nacionales como internacionales. Entre sus atractivos están el Parque Nacional de Corcovado, el Golfo Dulce y la Isla del Caño. Las condiciones climatólogicas de la zona se aprovechan en diversas actividades agropecuarias como la siembra y recolección de maíz, cacao, frijoles y tubérculos; así como la explotación de zonas acuíferas con propósitos pesqueros. La ganadería y la avicultura también son parte del volumen comercial de la región (Jiménez, 2024).

La región de Corcovado comprende el Parque Nacional de Corcovado y sus alrededores. Como se explicó anteriormente, esta región basa mayormente sus actividades en el turismo.

Corcovado se enfrenta a serios problemas de contaminación, no solamente de manera local (a raíz del turismo) sino internacional; residuos intercontinentales que viajan aproximadamente 12.000 kilómetros llegan a los mantos acuíferos de la zona (Chacón, 2025).

Uno de los ríos afectados por la contaminación en Corcovado es el río Claro. Se encontró niveles altos de metales pesados y niveles de DBO y DQO que sugieren una contaminación elevada (Salguero, 2016)

Este caso motiva los objetivos principales del proyecto: analizar y predecir la contami-

nación subyacente en los ríos que se pueden encontrar en la Península de Osa con fuerte énfasis en la zona de Corcovado; utilizando como indicadores y variables principales a las descritas anteriormente.

El monitoreo y la predicción de la calidad del agua se han convertido en una prioridad mundial ante el aumento sostenido de la contaminación y la necesidad de garantizar un uso eficiente de los recursos en procesos de tratamiento (Márquez Alvarado, 2022). Dentro de los parámetros más representativos para evaluar la carga contaminante en cuerpos de agua naturales y residuales destacan la Demanda Bioquímica de Oxígeno (DBO) y la Demanda Química de Oxígeno (DQO). Estos indicadores permiten estimar de forma indirecta la concentración de materia orgánica biodegradable y de compuestos oxidables presentes, proporcionando información crítica para la toma de decisiones (Aguilar y Díaz, 2020; Márquez Alvarado, 2022).

Sin embargo, la medición directa y continua de la DBO y la DQO en tiempo real representa un reto técnico y económico, ya que requiere procedimientos analíticos complejos, laboratorios especializados y recursos significativos, lo cual es un limitante, especialmente en regiones con recursos restringidos (Aguilar y Díaz, 2020). Frente a esta problemática, diversas investigaciones recientes han propuesto la utilización de modelos predictivos basados en técnicas estadísticas avanzadas y aprendizaje automático, que permiten estimar estos indicadores de manera eficiente a partir de variables fácilmente medibles.

Por ejemplo, el estudio de Aguilar Aguilar y Obando-Díaz (Aguilar y Díaz, 2020) destaca el uso de modelos como las redes neuronales artificiales (RNA), los sistemas de inferencia neurodifusa (ANFIS) y las máquinas de vectores de soporte (SVM), los cuales han alcanzado niveles de precisión superiores al 89 % al predecir los valores de DBO y DQO en escenarios reales. Estas aproximaciones no solo optimizan los tiempos de respuesta, sino que además reducen los costos asociados al monitoreo.

De manera complementaria, Márquez Alvarado (Márquez Alvarado, 2022) resalta la relevancia de los parámetros DBO y DQO en la gestión de aguas residuales industriales, al demostrar que la carga orgánica de los efluentes varía considerablemente según su origen. En particular, los residuos de industrias alimenticias tienden a presentar valores de DBO significativamente más altos, mientras que otros sectores, como los sanitarios o químicos, influyen de forma más notable en la DQO. Esta variabilidad constituye un desafío para los sistemas de tratamiento, dado que influye directamente en el tipo de tecnología, la duración de los procesos y el costo por metro cúbico tratado. Por lo que el uso de modelo predictivos ha demostrado ser una herramienta eficaz para anticipar los costos y optimizar la gestión de los recursos.

Por otro lado, los avances tecnológicos han permitido el desarrollo de sistemas integrados que combinan sensores IoT, análisis estadístico y predicción de la calidad del agua en tiempo real. Tal es el caso de la investigación de El Aatik Chouari (El Aatik Chouari y cols., 2024), donde se implementó un sistema de monitoreo en estaciones depuradoras de aguas residuales (EDARs) que, mediante sensores portátiles y técnicas como el Análisis de Componentes Principales (ACP), logró predecir los valores de DBO y DQO con altos

niveles de precisión, alcanzando coeficientes de determinación superiores al 99 %.

Por otra parte, Arias Araya (Araya, 2020) documenta que "la demanda bioquímica de oxígeno en cada uno de los canales muestreados en en el Distrito de Riego Arenal Tempisque se encuentra por debajo del límite máximo permisible en la legislación nacional (50 mg/L)" (p. 6), sugiriendo una carga orgánica relativamente moderada. Sin embargo, la autora advierte que la DQO superó los valores máximos permitidos en varios canales analizados. Específicamente, "la demanda química de oxígeno en los canales de Tamarindo, Bagatzí y La Guaria alcanzaron valores por encima del máximo permitido (150 mg/L)" (Araya, 2020, p. 6). En términos ecológicos, la autora destaca que "estos materiales pueden permanecer largo período en las aguas y en algunos casos bioacumularse en los seres vivos" (Araya, 2020, p. 6), identificándolos como compuestos orgánicos persistentes con implicaciones directas sobre la salud de los ecosistemas. Además, mediante el Índice Holandés, se identificó que en ciertos periodos críticos, como abril, "únicamente el canal de Tamarindo pasa a contaminación severa" (p. 8), lo cual evidencia la sensibilidad temporal de los sistemas de drenaje ante las prácticas agroindustriales.

En síntesis, la predicción de la DBO y la DQO mediante técnicas estadísticas y de aprendizaje automático se ha consolidado como una alternativa viable y eficiente para superar las limitaciones de la medición directa. Asimismo, la incorporación de herramientas tecnológicas, como sensores IoT y modelos multivariados, ha potenciado la capacidad de anticipar la calidad del agua en distintos contextos, ya sea en entornos naturales, industriales o urbanos.

Finalmente, es importante destacar que la pregunta de investigación que orienta el presente proyecto es la siguiente:

¿Qué tan predecibles son los niveles de DBO y DQO para las pruebas de tratamiento de aguas residuales en Corcovado?

A partir de esta interrogante, se plantearon los siguientes objetivos principales:

- Estimar el nivel de contaminación en los ríos de la región de Corcovado mediante un modelo de predicción basado en los indicadores de Demanda Bioquímica de Oxígeno (DBO) y Demanda Química de Oxígeno (DQO).
- Comparar los valores empíricos obtenidos a partir de las pruebas de campo y el modelo predictivo con los valores teóricos establecidos para los indicadores DBO y DQO, con el fin de evaluar la precisión del modelo y su aplicabilidad en contextos reales de monitoreo ambiental.

2. Métodos

Se realizó un estudio de tipo cuantitativo, predictivo y no experimental, orientado a estimar la calidad del agua en cuerpos hídricos de la Península de Osa mediante indicadores como la Demanda Química de Oxígeno (DQO). Esta aproximación se basó

en el uso de técnicas de aprendizaje automático, específicamente el algoritmo XGBoost, reconocido por su robustez en tareas de regresión con datos heterogéneos (Chen y Guestrin, 2016).

Los datos empleados provienen de registros históricos, previamente filtrados para excluir observaciones con valores faltantes en la variable dependiente. Las variables categóricas fueron transformadas mediante codificación *one-hot* y se aplicaron controles de normalidad sobre las cuantitativas.

Para la modelación se implementó XGBoost en R. Esta ha sido una de las herramientas más influyentes en competencias de ciencia de datos y aplicaciones industriales, destacando por su eficiencia y precisión . Adicionalmente se aplicó optimización de hiperparámetros mediante búsqueda en malla con el paquete `caret`.

El desempeño del modelo se evaluó mediante métricas como el error cuadrático medio (RMSE), error absoluto medio (MAE), coeficiente de determinación (R^2), error porcentual absoluto medio (MAPE) y error absoluto relativo (RAE). Adicionalmente, se estimaron intervalos de confianza sobre los errores y se utilizaron pruebas t para comparar variantes de modelo. El nivel de significancia adoptado fue $\alpha = 0{,}05$.
El algoritmo de XGBoost se basa en árboles de decisión, sin embargo, a diferencia de otros métodos más tradicionales como Random Forest, XGBoost es reconocido por su precisión y rapidez computacional además de evitar sobreajuste con pocos datos en comparación con Random Forest (Mendoza Vega, 2019). Anterior al modelo predictivo se realizó un análisis descriptivo de los datos para poder tomar pruebas de hipótesis que justificaran las hipótesis visuales.

Como parte del análisis del modelo XGBoost, se estimó la importancia relativa de cada variable predictora utilizando el criterio `Gain`, que mide la contribución de cada variable a la reducción del error cuadrático durante la construcción de los árboles. En la Tabla 1 se presentan las 10 variables más influyentes para la predicción de la demanda química de oxígeno (DQO) durante el mes de julio.

Cuadro 1: Top 10 variables más importantes según el modelo XGBoost

Variable	**Ganancia (Gain)**	**Cobertura (Cover)**	**Frecuencia (Freq.)**
`tempairej`	0.362	0.121	0.081
`tempaguaj`	0.159	0.084	0.060
`sitio.Río Sierpe`	0.098	0.033	0.029
`latitud.8.79175`	0.070	0.064	0.047
`latitud.8.86125`	0.065	0.052	0.041
`sat_oxigen`	0.044	0.020	0.015
`horarecolectaj`	0.039	0.014	0.011
`oxigeno`	0.020	0.017	0.015
`latitud.8.78405`	0.017	0.026	0.050
`longitud.-83.56314`	0.017	0.016	0.012

Los resultados indican que las variables más relevantes fueron la `tempairej` (temperatura del aire en julio) y la `tempaguaj` (temperatura del agua en julio), lo cual es coherente con la literatura que destaca el rol de la temperatura en la solubilidad del oxígeno y en la actividad de los contaminantes orgánicos. La inclusión del `sitio.Río Sierpe` como una variable importante sugiere que esta ubicación presenta condiciones particulares en los niveles de DQO, posiblemente debido al arrastre de plaguicidas previamente documentado en esta cuenca.

También destacan coordenadas geográficas específicas (latitud y longitud), lo que indica una dependencia espacial de los niveles de contaminación. Variables como la `saturación de oxígeno` (`sat_oxigen`) y el `oxígeno disuelto` refuerzan la relación esperada entre la calidad química del agua y sus propiedades fisicoquímicas.

En conjunto, estos hallazgos respaldan la utilidad del modelo no solo para predecir DQO, sino también para identificar variables clave que podrían ser monitoreadas de forma más frecuente en campañas de evaluación de la calidad del agua.

3. Resultados

3.1. Análisis Exploratorio

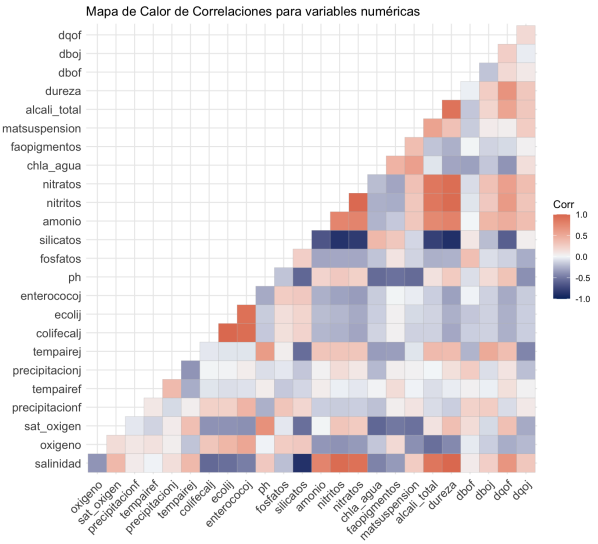


Figura 1: Matriz de Correlación

La figura 1 permite visualizar la fuerza y dirección de las relaciones lineales entre pares de variables mediante la correlación de Pearson, donde los valores cercanos a $+1$ indican una correlación positiva fuerte, los cercanos a -1 una correlación negativa fuerte, y los cercanos a 0 indican ausencia de relación lineal.

Dicha figura muestra patrones coherentes con la literatura, especialmente en la forma en que el amonio tóxico se relaciona con variables clave como pH, salinidad, nitritos, TAN y alcalinidad. Estas relaciones reflejan los procesos bioquímicos y ecológicos del sistema ecológico, donde un desbalance iónico o aumento de temperatura/pH puede detonar toxicidad por amonio y reducir la supervivencia, Lema Navarro (2023).

Por otro lado, la correlación negativa entre silicatos y nitritos puede interpretarse como una señal indirecta de presión antrópica: en áreas donde predominan las cargas nitrogenadas −producto de descargas residuales o fertilización−, suele observarse una menor concentración de elementos como el silicio, probablemente debido a diferencias en el origen de los aportes o al desplazamiento de fuentes naturales por actividades humanas. En diversos estudios se ha documentado que las concentraciones elevadas de nitritos y otros compuestos nitrogenados en cuerpos de agua están asociadas a presiones antrópicas, especialmente por fertilización y escorrentía agrícola (Solis, Bonetto, Marrochi, Paracampo, y Mugni, 2018; Mugni, 2008). Aunque no se reporta directamente una correlación con el silicio, es razonable suponer que en sistemas con carga nitrogenada predominante —como los observados en zonas agrícolas de la provincia de Entre Ríos—

los aportes naturales de elementos como el silicio pueden diluirse o alterarse, lo que puede reflejarse en correlaciones negativas entre ambos tipos de compuestos (Primost, 2019).

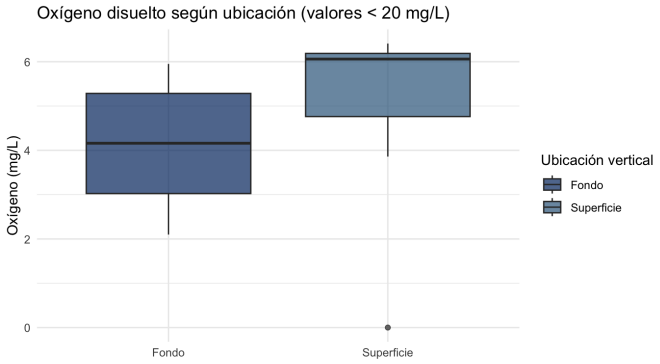


Figura 2: Variación del oxígeno disuelto entre la superficie y el fondo del cuerpo de agua

El análisis de oxígeno disuelto según la ubicación de muestreo (Figura 2) revela un patrón claro: las muestras tomadas en el fondo presentan concentraciones significativamente más bajas que aquellas recolectadas en la superficie. Mientras que en la superficie los niveles de oxígeno oscilan alrededor de los 6 mg/L, en el fondo se observan valores medianos en torno a los 4 mg/L, con algunos registros cercanos a 2 mg/L que indican posibles condiciones de hipoxia. Este comportamiento es coherente con la dinámica de cuerpos de agua estratificados, donde el intercambio gaseoso, la fotosíntesis y la mezcla turbulenta favorecen la oxigenación superficial (Niño, López, Pirard, Hillmer, y Gracía, 2015), mientras que la acumulación de materia orgánica y la falta de renovación vertical provocan un empobrecimiento del oxígeno en las capas profundas (Niño y cols., 2015).

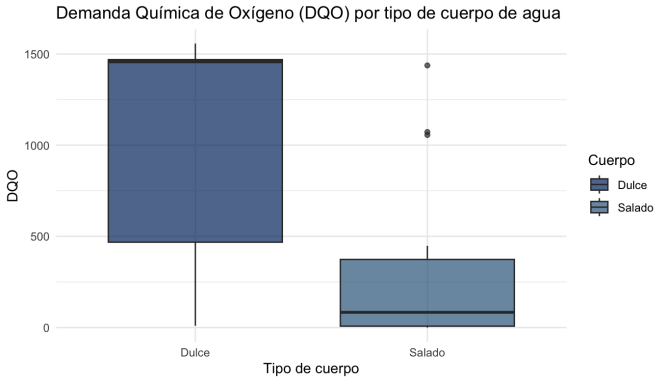


Figura 3: Comparación de la DQO entre cuerpos de agua dulce y salada

La comparación de la Demanda Química de Oxígeno (DQO) entre cuerpos de agua dulce

y salada muestra diferencias marcadas en los niveles de materia orgánica presente. En los cuerpos de agua dulce, la DQO alcanza valores considerablemente más altos, con una mediana cercana a 1200 mg/L, lo cual de acuerdo con Romero-Aguilar, Colín-Cruz, Sánchez-Salinas, y Ortiz-Hernández (2009) puede demostrar la presencia de materia orgánica proveniente de aguas residuales sanitarias, como urea, detergentes y diversos compuestos orgánicos, el arrastre de raíces o el desprendimiento de biopelículas dentro del sistema.

Por el contrario, los cuerpos de agua salada presentan valores de DQO mucho más bajos y estables, Déniz Quintana (2010) indica que los niveles bajos de DQO en aguas salinas pueden influir en la interpretación de parámetros como el ensuciamiento, y que las características químicas del agua de mar (como su baja materia orgánica biodegradable). Estos resultados podrían sugerir un mayor grado de deterioro en los cuerpos de agua dulce, lo cual podría refuerzar la necesidad de monitoreo y control más estricto en ecosistemas, pero también ponen de manifiesto que la composición química del agua —particularmente la salinidad y la carga orgánica— juega un papel determinante en los procesos de tratamiento y en la dinámica de la calidad del agua. Así, mientras que en ambientes salinos la baja DQO podría reflejar una menor presión antropogénica directa o una mayor eficiencia en la autodepuración, en cuerpos de agua dulce los valores elevados podrían estar asociados a descargas no tratadas, escasa renovación, o acumulación de materia orgánica, lo que subraya la importancia de establecer estrategias diferenciadas de gestión según el tipo de ecosistema acuático.

3.2. Preliminares

Como métodos preliminares al modelo predictivo se realizó un análisis descriptivo de los datos, en el cual se identificaron visualmente algunas relaciones relevantes. Inicialmente, se graficó un mapa con la ubicación de los puntos de muestreo, lo que permitió explorar ciertos supuestos espaciales, como la proximidad a zonas urbanas o áreas densamente pobladas. En estos casos, es común observar niveles más altos de ciertos indicadores de contaminación.

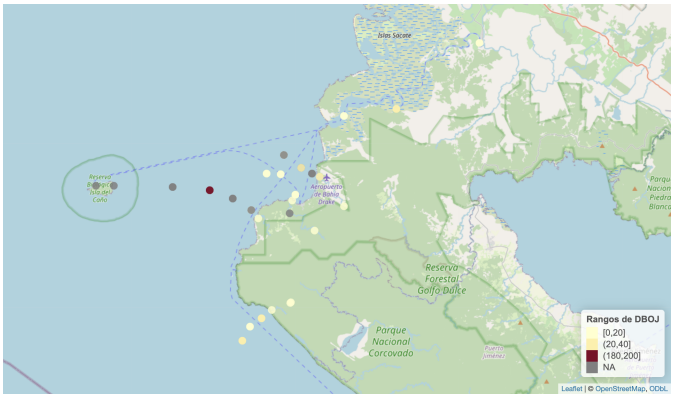


Figura 4: Mapa de puntos de muestreo

Las pruebas de hipótesis constituyen una herramienta estadística fundamental en el análisis de datos, y en este estudio se aplicaron con el propósito de validar la estabilidad y las relaciones entre distintas variables involucradas en la evaluación de la calidad del agua. Estas pruebas no solo respaldan la pertinencia del uso de modelos predictivos, sino que también permiten identificar diferencias significativas entre periodos temporales (por ejemplo, febrero y julio), así como correlaciones entre variables fisicoquímicas y biológicas.

Para evaluar la distribución de los datos, se aplicó la prueba de normalidad de Shapiro-Wilk. De todas las variables numéricas analizadas, únicamente `tempairef` y `tempairej` presentaron una distribución normal ($p > 0{,}05$), mientras que el resto de las variables rechazaron la hipótesis nula de normalidad.

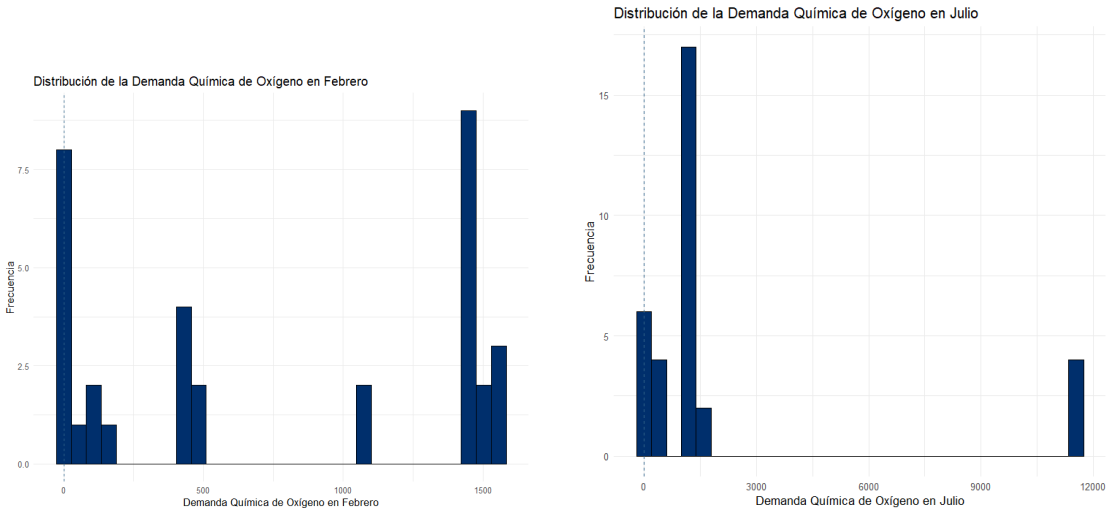
Dado que la mayoría de las variables no siguen una distribución normal, se empleó la prueba de Wilcoxon para comparar condiciones emparejadas entre los dos periodos. Los resultados más relevantes fueron los siguientes: la demanda biológica de oxígeno (DBO) mostró una diferencia significativa entre febrero y julio ($p = 0{,}0173$), mientras que la demanda química de oxígeno (DQO) no mostró diferencias significativas ($p = 0{,}1140$). La precipitación presentó una diferencia altamente significativa ($p < 0{,}0001$), en tanto que la temperatura del aire no mostró diferencias entre los meses evaluados ($p = 0{,}5420$).

Adicionalmente, se exploraron correlaciones de Spearman entre pares de variables. Se encontró una correlación alta y significativa entre el oxígeno disuelto y la saturación de oxígeno ($\rho = 0{,}84$, $p < 0{,}0001$), lo cual es coherente con su relación físico-química. En contraste, la correlación entre fosfatos y clorofila-a fue débil y no significativa ($\rho = -0{,}03$, $p = 0{,}84$). Asimismo, no se detectó una correlación significativa entre nitratos y coliformes fecales ($\rho = -0{,}15$, $p = 0{,}34$). La relación entre dureza y carbonatos fue moderada pero no significativa ($\rho = 0{,}25$, $p = 0{,}22$), mientras que la salinidad mostró una correlación positiva moderada y significativa con los sulfatos ($\rho = 0{,}53$, $p = 0{,}044$).

Estos resultados preliminares permiten tener una visión más clara del comportamiento de las variables medidas y fundamentan tanto las decisiones analíticas posteriores como la construcción de modelos predictivos confiables.

3.3. Modelo Predictivo.

Como se mencionó anteriormente se realizó un modelo XGBoost debido a su uso con pocos datos y su rapidez computacional, para la aplicación del modelo se usó un 60 % de los datos en entrenamiento y un 40 % en validación, se aplicaron métodos como `GridSearchCV` para la optimización de hiperparámetros aplicando a su vez una validación cruzada, esto con tal de obtener una mejor predicción y evitar el sobreajuste.



(a) DQO de Febrero

(b) DQO de Julio

Figura 5: Distribución de la demanda química de oxígeno para los distintos meses de muestreo

Es importante considerar una diferencia significativa entre la distribución del DQO de ambos meses, como se muestra en la figura 5a y 5b para el mes de Julio se concentra la mayor parte de los datos en un rango menor a los 3 mil a excepción de algunos valores cercanos a 12 mil, esto en unidades de mg/L, estos valores sumamente altos corresponden a la zona del Río Sierpe en el Térraba, hecho que parece curioso ya que según (Soto, 2025) gran parte de los plaguicidas utilizados en zonas relativamente cercanas a este río fueron llevados por causas naturales a este provocando una contaminación excesiva. Como se mencionó anteriormente con los datos suministrados se desea programar un modelo capaz de predecir el índice de DQO de los cuerpos de agua. A su vez se desean evaluar distintas métricas para ver si se está aprendiendo el comportamiento del modelo y como afecta la regularidad de los datos.

3.3.1. Predicción del DQO de Febrero

Para este mes se obtuvieron las distintas métricas

Métrica	Entrenamiento	Validación
RMSE	4.63	641.72
MAE	2.77	189.19
MAPE	6.25	37.90
RAE	0.0043	0.79
R^2	0.9678	0.5781

Cuadro 2: Métricas del modelo XGBoost para la predicción de la Demanda Química del Oxígeno para Febrero

Note como algunas de estas métricas para la validación son bastante malas como es el caso del RMSE (Root Mean Squared Error) el cual dice que los valores reales se alejan bastante de los calculados, con esto se mostrará un gráfico el cual presenta la comparación entre cada par (x_{pred}, x_{real}) del modelo, para mejorar la visualización de este se agregó al gráfico la recta $y = x$ el cual se puede interpretar como que entre más cercano está el punto a esta recta mejor fue dicha aproximación.

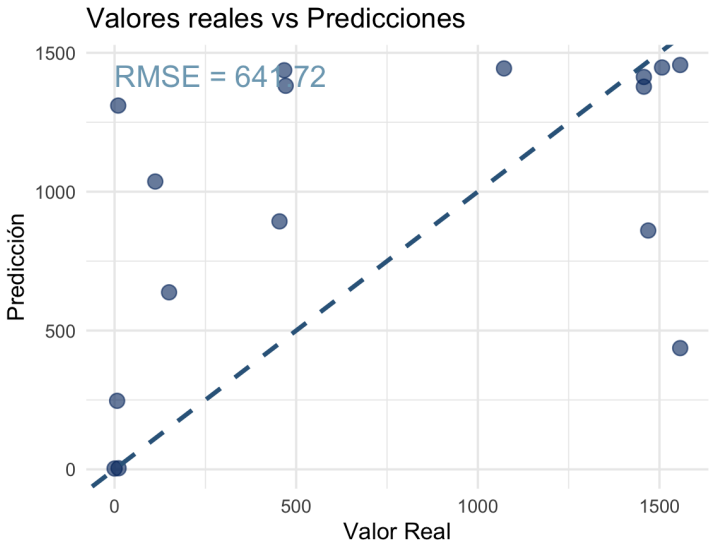


Figura 6: Comparación entre los valores reales y las predicciónes para el mes de Febrero

Posteriormente se mostrará el resultado del modelo para el mes de Julio donde se esperan mejores resultados a no ser tan diverso como en febrero. Pero para culminar la sección se analizarán las importancias de las variables en el modelo, las cuales se presentan en el siguiente gráfico

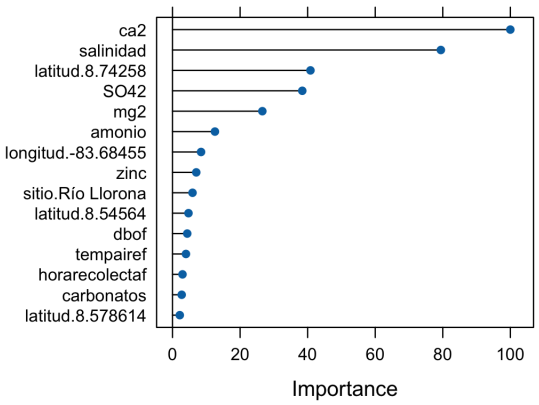


Figura 7: Importancia de las variables del modelo XGBoost aplicado en la predicción del DQO de Febrero.

Con respecto a la figura 7 se puede observar como variables como la presencia de algunos compuestos como el calcio, zinc, amonio, magnesio y los iones de sulfato fueron bastante importantes para determinar la predicción del modelo, así como algunos sitios y el dbo. Otro aspecto acerca del modelo que resulta importante apreciar es la distancia real que existe entre el valor teórico y el valor de la predicción ya que la figura 6 muestra como se distribuyen los puntos mas sin embargo resulta complicado ver dichas distancias, para ello considere el gráfico 8 el cual muestra las verdaderas distancias.

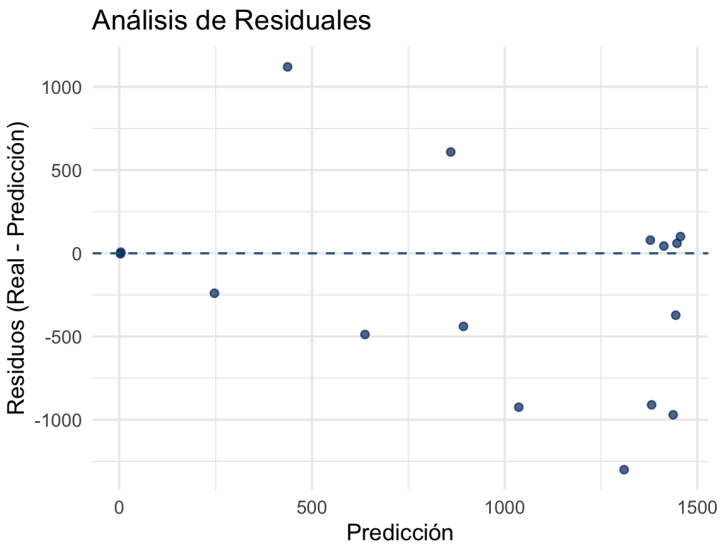


Figura 8: Distribución tipo "scatter plot"del residuo de las predicciones del DQO de Febrero.

Note que en este se puede apreciar como hay ciertos valores cuya predicción si estuvo relativamente cercana como lo son aquellos que se concentran al rededor de los $1200mg/L$ sin embargo resulta importante notar los dos residuos más grandes quienes son mayores a 1000 y menores a -1500 cerca de los $500mg/L$ y $1000mg/L$
El modelo planteado para el mes de Febrero tiene algunos detalles los cuales se pueden mejorar sin embargo dichas dificultades y mejoras se mencionan en la sección **Conclusiones y Recomendaciones**. Por ahora se procederá a analizar los resultados obtenidos en el mes de Julio.

3.3.2. Predicción del DQO de Julio

La presente sección será muy similar a la anterior en cuanto a estructura, pues se basa en el mismo modelo, para esta variable se obtuvieron las siguientes métricas.

Métrica	Entrenamiento	Validación
RMSE	45.37	1256.1
MAE	20.34	475.5
MAPE	5.85	15.46
RAE	0.0081	0.2034
R^2	0.9349	0.7691

Cuadro 3: Métricas del modelo XGBoost para la predicción de la Demanda Química del Oxígeno para Julio

Comparando primeramente con las obtenidas en Febrero se puede notar que en cuanto al comportamiento del coeficiente de determinación R^2 para el mes de Julio es mejor, es decir que la varianza de las predicciones son más similares a las reales en comparación con Febrero, sin embargo otras métricas como el RMSE y el MAE están muy por encima lo cual hace pensar que como tal no está prediciendo bien los valores y requiere de un ajuste, sin embargo si está prediciendo el comportamiento.
Para este caso también se analizaron la distribución de los pares ordenados, los residuos y la importancia de las variables ingresadas al modelo, para ello considere el siguiente gráfico

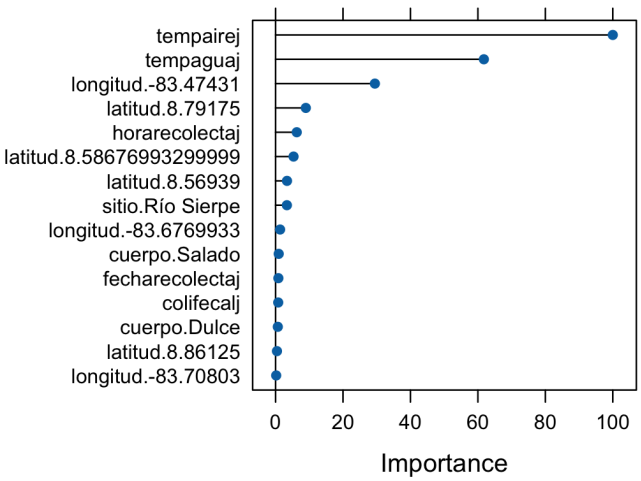


Figura 9: Importancia de las variables del modelo XGBoost aplicado en la predicción del DQO de Julio.

Para este caso se puede observar como la mayor importancia de variables se encuentra en la temperatura del aire y del agua, lo cual resulta sospechoso pues son aquellas variables que siguen una distribución más uniforme, en particular, según las pruebas de hipótesis anteriormente mencionadas la temperatura del aire es la única variable que sigue una distribución normal. Esto deja en duda si el modelo se encuentra sobre ajustado, pues esto y las métricas dan una señal de que esto está sucediendo.
Considere ahora el gráfico de comparación de los pares ordenados dado en la figura 10

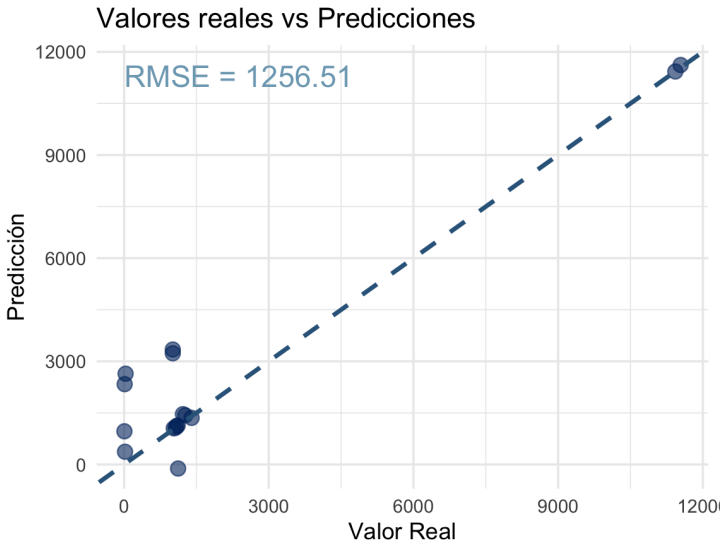


Figura 10: Comparación entre los valores reales y las predicciones para el mes de Julio

En esta note como si bien visualmente los puntos se encuentran más cerca de la recta $y = x$ que en el mes de Febrero, las distintas métricas dicen que algo está sucediendo con estas predicciones. Para poder visualizar mejor estos puntos considere el gráfico de residuos

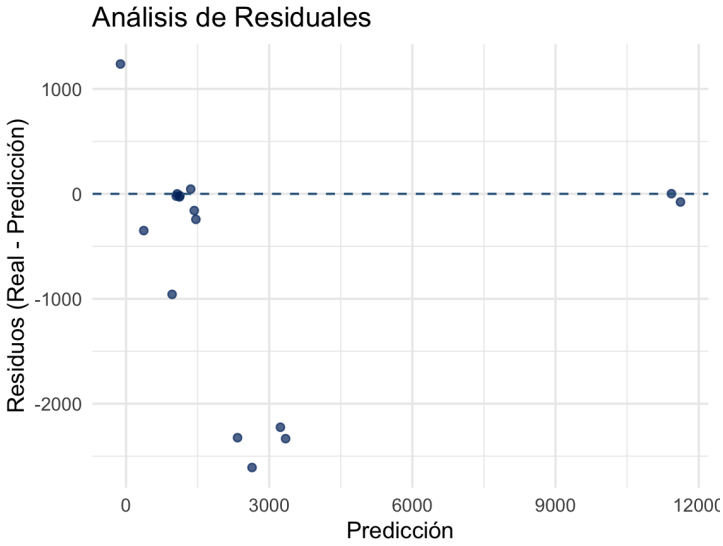


Figura 11: Distribución tipo "scatter plot"del resdiduo de las predicciones del DQO de Julio

Ahora note como en la figura 11 se pueden observar algunos valores que sobrepasan el error de 2000 en magnitud, esto es señal de que aun que hayan muchas de las predicciones que están por debajo de 500 de error, el modelo no está prediciendo bien esta variable. Una posible causa de esto sean los 2 valores del Río Sierpe que se encuentran por encima de los $10000 mg/L$ en DQO para Julio, queda como posible mejora la eliminación de estas variables para analizar los cambios sobre los gráficos y métricas del modelo.

4. Conclusión y Recomendaciones

El presente estudio permitió evaluar la calidad del agua en las cuencas de Corcovado mediante el análisis de parámetros clave como la DBO y DQO, utilizando técnicas de aprendizaje automático. Los resultados obtenidos confirman la utilidad del modelo XGBoost para predecir los niveles de contaminación, aunque con ciertas limitaciones en precisión durante algunos periodos. Se evidenció que factores como la temperatura ambiental, la ubicación geográfica y la época del año influyen significativamente en la calidad del agua, mostrando patrones diferenciados entre cuerpos de agua dulce y salada. Además, se identificaron casos puntuales de alta contaminación, como en el Río Sierpe, asociados a actividades humanas. Estos hallazgos resaltan la complejidad de los sistemas acuáticos y la necesidad de abordar su estudio con enfoques integrales que consideren tanto variables naturales como antropogénicas.
Con base en los resultados, se recomienda implementar estrategias de monitoreo más frecuentes, especialmente en zonas críticas identificadas, utilizando tecnologías como sensores IoT para obtener datos en tiempo real. Asimismo, sería valioso complementar

el modelo predictivo con información adicional sobre fuentes de contaminación local, como descargas industriales o agrícolas, para mejorar su precisión. Otra línea de acción importante sería promover políticas de gestión adaptativa que consideren las variaciones estacionales en la calidad del agua, así como campañas de educación ambiental dirigidas a las comunidades locales y actores económicos. Finalmente, se sugiere explorar en futuras investigaciones la integración de otros algoritmos de machine learning y variables adicionales para optimizar las predicciones y ampliar el entendimiento de los procesos contaminantes en la región. Estas medidas, en conjunto, podrían contribuir a una gestión más efectiva y sostenible de los recursos hídricos en Corcovado.

Referencias

Aguilar, A. C. A., y Díaz, F. F. O. (2020). Aprendizaje automático para la predicción de calidad de agua potable. *Ingeniare*(28), 47–62.

Araya, M. F. A. (2020). Bioindicadores de contaminación en aguas residuales de sistemas agropecuarios en el distrito de riego arenal tempisque, guanacaste, costa rica. *Oriolus*, *1*(1), 1–13.

Centeno Mora, E., Cruz Zúñiga, N., y Vidal Rivera, P. (2024). Tratamiento de aguas residuales ordinarias en costa rica: perfil tecnológico y perspectivas de sostenibilidad. *Ingeniería*, *34*(1), 7–22. Descargado 2025-06-29, de `http://www.scielo.sa.cr/scielo.php?script=sci_arttext&pid=S2215-26522024000100007&lng=es&nrm=iso` doi: 10.15517/ri.v34i1.55222

Chacón, M. N. (2025). *Basura que llega al parque nacional corcovado viaja más de 12.000 kilómetros desde otros continentes.* Publicado por el Semanario Universidad. (Consultado el 26 de junio del 2025)

Chaves-Villalobos, M., Quirós-Vega, J., Cordero-Cordero, S., Villalobos-Sequeira, J., Anchía-Leitón, D., Loría-Barquero, A., ... Paniagua-Pantoja, M. (2023). Evaluación de la salud ambiental del río ocloro, utilizando una metodología mixta. *Tecnología en Marcha*, *36*(4), 148–159. Descargado 2025-06-29, de `http://www.scielo.sa.cr/scielo.php?script=sci_arttext&pid=S0379-39822023000400148&lng=en&nrm=iso` doi: 10.18845/tm.v36i4.6392

Chen, T., y Guestrin, C. (2016). Xgboost: A scalable tree boosting system. En *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). Association for Computing Machinery. doi: 10.1145/2939672.2939785

de Información Jurídica, S. C. (s.f.). Artículo n.º21, Sistema Costarricense de Información Jurídica (PGR). Descargado de `http://www.pgrweb.go.cr/scij/Busqueda/Normativa/Normas/nrm_articulo.aspx?param1=NRA&nValor1=1&nValor2=59524&nValor3=83250&nValor5=21` (Consultado: 4 de julio de 2025)

Déniz Quintana, F. A. (2010). *Análisis estadístico de los parámetros dqo, dbo5 y ss de las aguas residuales urbanas en el ensuciamiento de las membranas de ósmosis inversa* (Tesis doctoral, Universidad de Las Palmas de Gran Canaria, Escuela de Ingenierías Industriales y Civiles). Descargado de `https://accedacris.ulpgc.es/handle/`

10553/4858 (Programa de doctorado: Tecnología Industrial. Consultado el 1 de julio de 2025)

El Aatik Chouari, A., y cols. (2024). Evaluación y predicción temporal de calidad de aguas residuales mediante tecnologías iot y estadísticas multivariantes.

Hernandez, E. R. B. (2016). *Dqo y dbo.* Trabajo con fines informativos. (Consultado el 26 de junio del 2025)

Induanalisis. (2019). *Dbo y dqo.* Trabajo con fines informativo. (Consultado el 26 de junio del 2025)

Jiménez, H. L. (2024). *Plan de desarrollo rural territorial 2024 – 2030.* Publicado por el Instituto Nacional de Desarrollo Rural. (Consultado el 26 de junio del 2025)

Lema Navarro, J. P. (2023). *Relación entre calidad de agua y amonio tóxico en un periodo de producción del cultivo de camarón Litopenaeus vannamei en santa elena-santa elena* (Tesis de licenciatura, Universidad Estatal Península de Santa Elena, La Libertad, Ecuador). Descargado 2025-07-01, de https://repositorio.upse.edu.ec/handle/46000/10101 (Facultad de Ciencias del Mar, Carrera de Biología. Dirigida por Mery Ramírez Muñoz)

Meléndez Carranza, A. (2018). Climatología. Descargado de https://ulibros.com/index.php/climatologia-s0if5.html (200 páginas. Libro electrónico. Texto en español)

Mendoza Vega, J. B. (2019). *Xgboost en r.* R Pubs. Descargado de https://rpubs.com/jboscomendoza/xgboost_en_r (Última actualización hace más de 5 años)

Mugni, H. D. (2008). *Concentración de nutrientes y toxicidad de pesticidas en aguas superficiales de cuencas rurales* (Tesis de doctorado, Universidad Nacional de La Plata, Facultad de Ciencias Naturales y Museo, La Plata, Argentina). (Consultado el 1 de julio de 2025) doi: 10.35537/10915/4410

Márquez Alvarado, A. B. (2022). *Diseño de investigación para la construcción de un modelo de regresión para la predicción del costo del tratamiento de aguas residuales y desechos sólidos industriales de una planta en guatemala* (Tesis Doctoral no publicada). Universidad de San Carlos de Guatemala.

Niño, Y., López, F., Pirard, C., Hillmer, I., y Gracía, M. H. (2015, diciembre). Modelación numérica de procesos de mezcla turbulentos inducidos por el viento en cuerpos de agua estratificados. *Tecnología y ciencias del agua*, *15*(1), 13–25. Descargado de https://www.revistatyca.org.mx/index.php/tyca/article/view/847

Primost, J. E. (2019). *Dinámica de nutrientes en aguas superficiales del delta del paraná: impactos del desarrollo productivo regional en la sustentabilidad del ecosistema* (Tesis de doctorado, Universidad Nacional de La Plata, Facultad de Ciencias Exactas, La Plata, Argentina). (Consultado el 1 de julio de 2025) doi: 10.35537/10915/96796

Romero-Aguilar, M., Colín-Cruz, A., Sánchez-Salinas, E., y Ortiz-Hernández, L. (2009). Tratamiento de aguas residuales por un sistema piloto de humedales artificiales: evaluación de la remoción de la carga orgánica. *Revista internacional de contaminación ambiental*, *25*(3), 157–167.

Salguero, M. E. A. (2016). *Estudio hidrogeolÓgico del Área de influencia del botadero de basura de la municipalidad de golfito: Ubicado en la fila manigordo en la esperanza de río claro, golfito, puntarenas.* Publicado por el MINAE. (Consultado el 26 de junio del 2025)

Solano Arce, M. d. M. (2011). *Impacto ambiental por aguas residuales y residuos sólidos en la calidad del agua de la parte media-alta de la microcuenca del río damas y propuesta de manejo* (Trabajo de graduación de licenciatura, Costa Rica). Descargado de `https://www.aya.go.cr/centroDocumetacion/catalogoGeneral/Impacto%20ambiental%20por%20aguas%20residuales%20y%20residuos%20s%C3%B3lidos%20en%20la%20calidad%20del%20agua.pdf` (Proyecto para optar por el grado de Licenciatura en Manejo de Recursos Hídricos)

Solis, M., Bonetto, C., Marrochi, N., Paracampo, A., y Mugni, H. (2018). Aquatic macroinvertebrate assemblages are affected by insecticide applications on the Argentine Pampas. *Ecotoxicology and Environmental Safety*, *148*, 11–16. Descargado de `https://www.sciencedirect.com/science/article/pii/S0147651317306887` doi: https://doi.org/10.1016/j.ecoenv.2017.10.017

Soto, V. C. (2025, 25 de junio). Plaguicidas de la piña viajan 70 kilómetros y contaminan delta del humedal térraba-sierpe. *Semanario Universidad*. Descargado de `https://semanariouniversidad.com/pais/plaguicidas-de-la-pina-viajan-70-kilometros-y-contaminan-delta-del-humedal-terraba-sierpe/`

Sánchez-Gutiérrez, R., Pérez-Salazar, R., y Alfaro-Chinchilla, C. (2021). *Aguas residuales generadas en el cantón san pablo durante el periodo 2014–2018 en plantas de tratamiento de aguas residuales* (Informe técnico). Universidad Nacional (Costa Rica). Descargado de `https://repositorio.una.ac.cr/items/ac4ff942-0358-4fd7-bc51-8c6148fd4a53` (Proyecto SIA 0244-17: Participación en los procesos de gestión ambiental del Cantón de San Pablo en la provincia de Heredia)