FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA AMBIENTAL

Desarrollo de modelos de machine learning para la predicción de la calidad del agua utilizando datos históricos, Cuenca Azángaro – 2023

TESIS PARA OBTENER EL TÍTULO PROFESIONAL DE:

Ingeniera Ambiental

AUTORA:

Cazasola Cuno, Zhaida Yoshy (orcid.org/0009-0003-5939-1974)

ASESOR:

Dr. Sernaque Auccahuasi, Fernando Antonio (orcid.org/0000-0003-1485-5854)

LÍNEA DE INVESTIGACIÓN:

Calidad y Gestión de los Recursos Naturales

LÍNEA DE RESPONSABILIDAD SOCIAL UNIVERSITARIA:

Desarrollo sostenible y adaptación al cambio climático

LIMA – PERÚ

2024

Dedicatoria

A Dios y a mi familia.

Agradecimiento

A Dios, mi familia y a mi asesor, el

Dr. Fernando Antonio Sernaque Auccahuasi.

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA AMBIENTAL

Declaratoria de Autenticidad del Asesor

Yo, SERNAQUE AUCCAHUASI FERNANDO ANTONIO, docente de la FACULTAD DE INGENIERÍA Y ARQUITECTURA de la escuela profesional de INGENIERÍA AMBIENTAL de la UNIVERSIDAD CÉSAR VALLEJO SAC - LIMA NORTE, asesor de Tesis titulada: "Desarrollo de Modelos de Machine Learning para la Predicción de la Calidad del Agua
Utilizando Datos Históricos, Cuenca Azángaro – 2023", cuyo autor es CAZASOLA CUNO ZHAIDA YOSHY, constato que la investigación tiene un índice de similitud de 16.00%,
verificable en el reporte de originalidad del programa Turnitin, el cual ha sido realizado sin filtros, ni exclusiones.

He revisado dicho reporte y concluyo que cada una de las coincidencias detectadas no constituyen plagio. A mi leal saber y entender la Tesis cumple con todas las normas para el uso de citas y referencias establecidas por la Universidad César Vallejo.

En tal sentido, asumo la responsabilidad que corresponda ante cualquier falsedad, ocultamiento u omisión tanto de los documentos como de información aportada, por lo cual me someto a lo dispuesto en las normas académicas vigentes de la Universidad César Vallejo.

LIMA, 08 de Abril del 2024

Apellidos y Nombres del Asesor:	**Firma**
SERNAQUE AUCCAHUASI FERNANDO ANTONIO **DNI:** 07234567 **ORCID:** 0000-0003-1485-5854	Firmado electrónicamente por: FSERNAQUEA el 24- 04-2024 14:45:31

Código documento Trilce: TRI - 0742170

UNIVERSIDAD CÉSAR VALLEJO

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA AMBIENTAL

Declaratoria de Originalidad del Autor

Yo, CAZASOLA CUNO ZHAIDA YOSHY estudiante de la FACULTAD DE INGENIERÍA Y ARQUITECTURA de la escuela profesional de INGENIERÍA AMBIENTAL de la UNIVERSIDAD CÉSAR VALLEJO SAC - LIMA NORTE, declaro bajo juramento que todos los datos e información que acompañan la Tesis titulada: "Desarrollo de Modelos de Machine Learning para la Predicción de la Calidad del Agua Utilizando Datos Históricos, Cuenca Azángaro – 2023", es de mi autoría, por lo tanto, declaro que la Tesis:

1. No ha sido plagiada ni total, ni parcialmente.
2. He mencionado todas las fuentes empleadas, identificando correctamente toda cita textual o de paráfrasis proveniente de otras fuentes.
3. No ha sido publicada, ni presentada anteriormente para la obtención de otro grado académico o título profesional.
4. Los datos presentados en los resultados no han sido falseados, ni duplicados, ni copiados.

En tal sentido asumo la responsabilidad que corresponda ante cualquier falsedad, ocultamiento u omisión tanto de los documentos como de la información aportada, por lo cual me someto a lo dispuesto en las normas académicas vigentes de la Universidad César Vallejo.

Nombres y Apellidos	**Firma**
ZHAIDA YOSHY CAZASOLA CUNO **DNI:** 71477959 **ORCID**: 0009-0008-1493-9672	Firmado electrónicamente por: ZYCAZASOLA el 08-04-2024 17:15:14

Código documento Trilce: TRI – 0742169

INVESTIGA UCV

Índice de contenidos

Índice de gráficos

Resumen

En años recientes, la ausencia de sistemas automatizados en la predicción de la calidad del agua ha ocasionado retrasos significativos en la obtención de datos precisos, lo cual ha impactado la fiabilidad de los cálculos y ha elevado los costos asociados a todo el proceso. La investigación está enfocada en desarrollar modelos de aprendizaje automático para automatizar el sistema para predecir calidad del agua en la cuenca de Azángaro. Los datos se consultaron de la base de datos que cuenta la institución nacional SNIRH de Perú dentro de la temática calidad del agua, obteniendo un total de 136 muestras, donde la metodología empleada para el desarrollo del modelo fue, recolección de datos históricos, selección de parámetros, procesamiento y limpieza de datos, división de datos (prueba - entrenamiento), entrenamiento del modelo y finalmente la etapa de validación de cada modelo, en este punto es donde se evaluó el rendimiento de que tan bien puede predecir cada modelo la calidad del agua. Los resultados fueron, de los 5 modelos de predicción desarrollados, Random Forest (RF) seguido de Decisión Trees (DTs) lograron un buen rendimiento en métricas de evaluación, en el modelo Random Forest se obtuvo un Root Mean Squared Error (RMSE) de 3.354, Mean Squared Error (MSE) de 12.886, un Mean Absolute Error (MAE) de 2.563 y Coefficient of Determination (R2) de 0.613. Por ende, se concluye que el desarrollo de este modelo presenta un desempeño óptimo para la predicción de la calidad del agua.

Palabras clave: Random forest, métricas, python.

Abstract

In recent years, the absence of automated systems for water quality prediction has caused significant delays in obtaining accurate data, which has impacted the reliability of the calculations and increased the costs associated with the whole process. The research is focused on developing machine learning models to automate the system for predicting water quality in the Azángaro watershed. The data were consulted from the database of the national institution SNIRH of Peru within the thematic of water quality, obtaining a total of 136 samples, where the methodology used for the development of the model was, historical data collection, parameter selection, data processing and cleaning, data division (test - training), model training and finally the validation stage of each model, at this point is where the performance of how well each model can predict water quality was evaluated. The results were, of the 5 prediction models developed, Random Forest (RF) followed by Decision Trees (DTs) achieved good performance in evaluation metrics, in the Random Forest model a Root Mean Squared Error (RMSE) of 3. 354, Mean Squared Error (MSE) of 12.886, Mean Absolute Error (MAE) of 2.563 and Coefficient of Determination (R2) of 0.613. Therefore, it is concluded that the development of this model presents an optimal performance for the prediction of water quality.

Keywords: Random forest, metrics, python.

I. INTRODUCCIÓN

A nivel mundial, el ecosistema se ve significativamente afectado por el rápido crecimiento de la población (Jankowska et al. 2022). Desafortunadamente, las ciudades se están expandiendo más rápido de lo que se puede planificar, financiar e implementar la infraestructura necesaria en muchas regiones del mundo. En consecuencia, hay un aumento en la demanda de agua a nivel mundial (Wilderer 2011), ya que el agua tiene una cantidad limitada que se puede consumir porque es un recurso que todos los seres vivos necesitan (Nasir et al. 2022).

En China, en un Sistema Acuícola, el equipo necesario para el monitoreo directo de la calidad del agua resulta costoso. Además, el método de prueba no solo es lento en detectar los parámetros, sino que también produce contaminación durante el proceso (Wawrzyniak, Matas Serrato y Blanchoud 2021). En este sentido, la complejidad del tratamiento del agua se vio incrementada debido a la inclusión de seres humanos en el proceso y en los cálculos del modelo (Chen et al. 2022).

También en Vietnam específicamente en el río de La Buong la calidad de agua se evalúa y clasifica utilizando el índice de calidad del agua (ICA), que se ha utilizado ampliamente (Khoi et al. 2022). Para determinar el nivel de calidad de agua, el índice de Brown se calculaba utilizando las características fisicoquímicas del agua como (temperatura, pH, turbidez, oxígeno disuelto (OD), demanda bioquímica de oxígeno (DBO), y concentraciones de otros contaminantes), esto ofrecía a los planificadores y tomadores de decisiones, información cuantitativamente significativa (Brown et al. 1970). Sin embargo, los cálculos necesarios para crear un índice de calidad de agua requieren mucho tiempo y esfuerzo (Singha et al. 2021). Las fórmulas carecen de consistencia, ya que frecuentemente emplean varias ecuaciones (Bui et al. 2020a). Para abordar los problemas antes mencionados, es crucial contar con una estrategia diferente para el cálculo correcto y eficiente del índice de calidad de agua (Khoi et al. 2022).

Por otro lado, a nivel nacional, ANA (Autoridad Nacional del Agua) utiliza una metodología para determinar el Índice de Calidad de Agua de los Recursos Hídricos

Superficiales (ICA-PE) en numerosos sitios de monitoreo y en función de varios criterios. Esta metodología produce un resultado cualitativo que se representa en cinco escalas, de pésimo a excelente; pero, debido a que se realiza utilizando los parámetros que se están evaluando en ese momento, no se puede utilizar para hacer predicciones (ANA, 2018).

En los últimos años, el aprendizaje automático se ha consolidado como un área clave de la inteligencia artificial que permite a los sistemas aprender automáticamente de sus experiencias y avanzar sin necesidad de una programación consciente (Sun y Scanlon 2019). Como resultado, durante los últimos 20 años, el aprendizaje automático supervisado y otros enfoques de inteligencia artificial (IA) se han aplicado ampliamente para abordar una variedad de problemas ambientales(Aldrees et al. 2023a). ML (Machine Learning) ha demostrado un gran potencial para la toma de decisiones basada en datos(Sun y Scanlon 2019), además ML (Machine Learning) posibilita la creación de modelos que pueden analizar los datos de manera ágil y generar resultados de forma rápida(Shailaja, Seetharamulu y Jabbar 2018).

Por ello, se plantea como **pregunta general** ¿El desarrollo de modelos de ML tiene un desempeño óptimo para mejorar la predicción de la calidad del agua utilizando datos históricos?, y como **preguntas específicas** ¿Cuál es el valor promedio del error cuadrático medio (MSE) obtenido por los modelos ML en la predicción de la calidad del agua (PCA) utilizando datos históricos?, ¿Cuál es el valor promedio del error cuadrático medio (RMSE) obtenido por los modelos ML en la predicción de la calidad del agua (PCA) utilizando datos históricos?, ¿Cuál es el valor promedio del error absoluto medio (MAE) obtenido por los modelos ML en la PCA utilizando datos históricos?, ¿Cuál es el valor promedio del coeficiente de determinación (R2) obtenido por los modelos ML en la PCA utilizando datos históricos?

El objetivo de esta investigación es analizar si la implementación de modelos de ML puede alcanzar un rendimiento óptimo al predecir la calidad del agua. Se utilizarán datos históricos y métricas de problemas de regresión para medir el rendimiento del modelo.

Por ello el presente trabajo ubica una **justificación metodológica** debido a que este estudio busca desarrollar modelos ML que se enfoquen en la predicción de la calidad del agua, siendo este un campo en el que intervienen algoritmos que permiten a los sistemas informáticos deducir patrones a partir de datos en este caso datos históricos, tenemos como **justificación práctica** ya que al contar con un modelo de predicción, se pueden anticipar cambios en la calidad del agua, esto permite una mejor planificación de las acciones y medidas necesarias para garantizar la calidad del agua y su disponibilidad. Finalmente, como **justificación ambiental** el presente trabajo proporciona una herramienta que pretende contribuir en la detección temprana de problemas ambientales, la identificación de fuentes de contaminación y la adopción de medidas de mitigación, al mejorar la capacidad de predicción de la calidad del agua se puede aportar a tomar medidas preventivas y de conservación para mantener la salud y la integridad de los sistemas acuáticos.

Por ello se tiene como **objetivo general** desarrollar modelos de ML para la predicción de la calidad del agua utilizando datos históricos, como **objetivos específicos**, determinar el valor promedio de MSE obtenido por los modelos ML en la predicción de la calidad del agua utilizando datos históricos, determinar el valor promedio de RMSE obtenido por los modelos ML en la predicción de la calidad del agua utilizando datos históricos, determinar el valor promedio del MAE obtenido por los modelos ML en la predicción de la calidad del agua utilizando datos históricos, determinar el valor promedio del coeficiente de determinación R2 obtenido por los modelos en la predicción de la calidad del agua utilizando datos históricos.

De la misma manera, se plantea la **hipótesis general,** el desarrollo de modelos de ML basado en datos históricos permitirá obtener predicciones precisas y confiables de la calidad del agua. Por consiguiente, se formulan las siguientes **hipótesis específicas**. El valor promedio del MSE obtenido por los modelos al PCA utilizando datos históricos será aproximadamente cero, lo que denotará una alta precisión y ajuste del modelo. El valor de RMSE obtenido por los modelos PCA utilizando datos históricos será cercano cero, indicando una precisión y ajuste elevados del modelo. El valor de MAE obtenido por los modelos en la PCA utilizando datos históricos será mínimo, lo que sugiere una baja discrepancia entre las predicciones y los valores

reales de calidad de agua. El valor promedio del R2 obtenido por los modelos en la PCA utilizando datos históricos será cercano a uno, lo que demuestra un buen ajuste del modelo y una alta capacidad para explicar la variabilidad de los datos.

II. MARCO TEÓRICO

En los antecedentes, según (Lap et al. 2023) tuvieron como objetivo observar que tan bien se desempeñó el método basado en ML para calcular el índice de calidad del agua (ICA) Como metodología se propuso un método novedoso basado en ML que combina técnicas de selección de características con algoritmos para así calcular el WQI sin comprometer la precisión de los resultados, primero se identificaron los factores más influyentes en la calidad del agua (CA) dentro de los datos históricos recopilados y luego se aplicó y evaluó cada algoritmo para determinar su desempeño en la predicción precisa de los valores ICA. Los resultados mostraron que el modelo RF (Random Forest) ofreció la mejor precisión en el pronóstico de valores de ICA con un RMSE (un Root Mean Squared Error) (11.08) y MAE (Mean Absolute Error) (6.57) más bajos, de acuerdo a los parámetros con mayor influencia en calidad del agua. Concluyeron que el método basado en ML (Machine Learnig) puede llegar a calcular el ICA con predicciones precisas en consecuencia a la evaluación respectiva.

Según (Khoi et al. 2022b), su objetivo fue evaluar el rendimiento de algoritmos de ML (Machine Learning) en la estimación de la calidad del agua superficial del río La Buong en Vietnam. Como metodología se propusieron usar algoritmos basados en refuerzo, algoritmo basado en árboles de decisión y cuatro algoritmos basados en RNA (Red Neuronal artificial), las variables de entrada consideraron 10, antes de entrenar a cada modelo dividieron la data en dos partes 70% para la etapa de entrenamiento y 30% para la etapa de prueba. Los resultados mostraron que el modelo RF (Random Forest) ofreció una de las mejores predicciones con un RMSE 0.121 y R2 0.986 seguido DT (Decision Trees) con un RMSE 0.147 y R2 0.979. Concluyeron que ambos modelos reprodujeron bien el ICA, pero en cuanto a rendimiento RF tuvo la predicción más precisa.

Según (Bui et al. 2020b), tuvieron como objetivo principal predecir el ICA del río Talar en Irán. En su metodología recopilaron 6 años (2012-2018) de datos mensuales sobre la calidad del agua en dos estaciones de monitoreo de calidad del agua ubicadas estratégicamente dentro de la cuenca, el conjunto de datos se dividió en dos subconjuntos para entrenamiento y prueba (70:30). Usaron dos grupos principales de modelos de algoritmos de árbol de decisión, algoritmos meta clasificadores o híbridos, para la evaluación utilizaron cinco métricas para evaluar cuantitativamente los modelos. En los resultados se observó que RF obtuvo un RMSE 2.97 y un MAE 1.67. En conclusión, RF fue uno de los modelos que mejor rendimiento tuvo para predecir los datos.

Según (Haghiabi, Nasrolahi y Parsaie 2018) tuvieron como objetivo evaluar el rendimiento de técnicas de IA (Inteligencias Artificiales), incluidas GMDH (Group Method of Data Handling), SVM (Support Vector Machine) y ANN (Artificial Neural Network), para predecir los componentes de la calidad del agua del río Tireh (Irán). En la metodología, se propusieron utilizar métodos de IA como MLP, SVM y el método de grupo de manejo de datos (GMDH) basado en los valores umbrales de RMSE y R2. En este estudio. El resultado más resaltante se vio que el mejor rendimiento de RNA (Red Neuronal Artificial) estaba relacionado con la función tansig como la mejor función de transferencia entre las etapas de entrenamiento y prueba, con R2 (0,92 y 0,84) y MSE (0,238 y 0,295) que mostraban la mayor correlación entre ambas etapas. Debido a estos factores, concluyeron que la precisión del RNA podía considerarse aceptable a efectos prácticos gracias a los resultados de este estudio.

Según (Chou, Ho y Hoang 2018) tuvieron como objetivo evaluar la aplicabilidad de los modelos a las necesidades especiales de los profesionales, así como su precisión al hacer predicciones. En la metodología usaron cuatro paquetes de software de minería de datos. El conjunto de datos históricos que utilizaron se compiló durante un periodo

de diez años. El estudio empleó características del agua fácilmente disponibles para modelar las relaciones no lineales entre variables y el ICA. Sus resultados indicaron que el modelo SVR (Support Vector Regression)en IBM SPSS Modeler arrojó los mejores valores R2 (0.840) RMSE (5.035) y MAE (3,814). En conclusión, IBM SPSS Modeler es la mejor opción para realizar predicciones utilizando métodos de conjunto.

Según (Zheng et al. 2023), tuvieron como objetivo emplear algoritmos predictivos híbridos para analizar la calidad del agua a partir de varios parámetros. En la metodología, propusieron un modelo de predicción de series temporales llamado VARLST, diseñado para predecir datos multiparamétricos de calidad del agua en muestras de datos pequeñas. Los resultados que obtuvieron demostraron que este enfoque híbrido de predicción logró un bajo error de predicción RMSE de 0.01. En sus conclusiones, el modelo de predicción híbrido demostró un rendimiento integral destacado, y sus predicciones se asemejan de manera significativa a los indicadores obtenidos de datos reales.

Según (Uddin et al. 2023), tuvieron como objetivo mejorar el método de evaluación de la calidad del agua y desarrollar una herramienta práctica para los reguladores ambientales en Irlanda, con el propósito de reducir la contaminación en cuerpos de agua costeros y de transición. Para lograrlo, se basaron en cinco componentes idénticos que se utilizan para evaluar la calidad del agua. Para evaluar el rendimiento del modelo, utilizaron diversas métricas de rendimiento. En sus resultados, hallaron en un cuerpo de agua costera llamada Mulroy Bay se encontró el error de predicción del modelo más bajo (RMSE = 0,21, MSE = 0,045, MAE = 0,15). En conclusión, los resultados indican que el modelo IEWQI representa una técnica altamente eficiente y confiable para evaluar con mayor precisión la calidad del agua en entornos costeros.

Según (Uddin et al. 2022) su objetivo fue encontrar un algoritmo de ML robusto y optimizar sus hiperparámetros para lograr predicciones

precisas del ICA en cada sitio de monitoreo en el puerto de Cork, Irlanda. Se compararon 8 algoritmos ML (Machine Learning) en la metodología utilizando once variables de calidad del agua para calcular el ICA. Los resultados del estudio revelaron que dos algoritmos presentaron un destacado rendimiento de predicción. En primer lugar, el algoritmo XGBoost(XGB) obtuvo los errores de predicción más bajos durante el periodo de entrenamiento (RMSE = 3.3, MSE = 10.91, MAE = 1.67 y R2 =1.0). El segundo algoritmo con buen rendimiento fue RF, también con un error de predicción muy bajo. En conclusión, el algoritmo XGB demostró ser el más efectivo para la predicción del ICA, superando a los otros algoritmos evaluados en este estudio.

Según (Ahmed et al. 2019) tuvieron como objetivo buscar un método alternativo más rápido y económico para estimar el ICA. En la metodología, emplearon algoritmos de aprendizaje automático supervisado para realizar la estimación, utilizando cuatro parámetros de entrada en el modelo: temperatura, turbidez, pH y sólidos disueltos totales. En sus resultados obtenidos, encontraron que el algoritmo Gradient Boosting demostró ser el más eficiente para realizar la predicción del WQI. La evaluación de este modelo tuvo un MAE de 1.9642, MSE de 7.2011 y RMSE de 2.6835. En conclusión, los resultados indican que el algoritmo Gradient Boosting es altamente eficiente en la predicción del WQI.

Según (Aldrees et al. 2023), tuvieron como objetivo abordar los crecientes problemas de calidad del agua mediante la aplicación de modelos de ML evolutivos y en conjunto. En la metodología utilizaron 360 lecturas temporales de conductividad eléctrica y sólidos disueltos totales, junto con varias variables de entrada, para establecer dos modelos: modelo de programación de expresiones múltiples MEP, modelo de regresión de bosque aleatorio RF. Además, realizaron evaluaciones utilizando diversas métricas estadísticas para determinar la precisión de los modelos desarrollados. En sus resultados mostraron que el modelo

RF demostró un rendimiento relativamente mejor que el modelo MEP con un R2 de 20% RMSE de 17% MAE 14%. En conclusión, se observó que los modelos MEP y RF dieron resultados precisos, pero RF mostró un rendimiento excepcional.

El aprendizaje automático es una rama de la inteligencia artificial (IA) que se ocupa de cómo los ordenadores pueden aprender a partir de una gran cantidad de datos (Deo 2015). Se utiliza para lograr una estimación más precisa al alimentar una gran cantidad de conjuntos de datos a la computadora para lograr el mayor nivel de precisión posible (Abuodeh, Abdalla y Hawileh 2020). Los modelos basados en ML son más precisos, ya que se basan en enfoques basados en datos y a su vez pueden capturar relaciones y patrones más matizados en los respectivos datos (Sayed et al. 2023). Es así que este método se puede utilizar como método para la gestión avanzada de la calidad del agua (Lee et al. 2022)tiene dos categorías principales, algoritmos supervisados y no supervisados, el primero se usa cuando se conoce la entrada (características) y la salida (etiquetas) de un evento y el segundo se usa cuando se tiene solo un conjunto de datos con solo entradas con la finalidad de que el algoritmo de ML investigue el conjunto de datos e identifique el patrón específico mirando entre puntos de datos (Sayed et al. 2023).

El ML supervisado consiste en métodos para construir automáticamente una función predictiva, dado un conjunto de instancias de entrenamiento (Witten, Frank y Geller 2002). Además, ML supervisado se compone de datos anotados, mientras que el conjunto de prueba se compone de datos no anotados, es discreto y desordenado y se le conoce como etiquetas de clase; cuando se incluyen valores numéricos continuos se denomina variable objetivo (o de salida) continuas y se utilizan típicamente para representar instancias. Cada variable de este conjunto es conocida como características (o predictor), refleja una característica de una instancia. El objetivo del aprendizaje supervisado es educar una función que asigna una entrada a una salida basada en emparejamiento

de entrada-salida, la situación ideal permite que el algoritmo prediga con precisión la respuesta o el resultado en circunstancias desconocidas. Asimismo, pueden tener el propósito adicional de descubrir el conocimiento interpretable (Fabris, Magalhães y Freitas 2017).

Random Forest es una técnica de aprendizaje automático supervisado por conjuntos que surgió a principios del siglo XXI (Saxena et al. 2017) se basa en la idea de que muchos modelos de árboles de decisión independiente que trabajan juntos como un equipo funcionan mejor que cualquier otro modelo único (Zhou 2021), pertenece a la categoría del algoritmo de aprendizaje por conjuntos (Patel et al. 2015). Por lo general, se necesitan árboles para crear RF con el fin de mejorar la estabilidad del modelo y reducir los errores de predicción (Zhang y Wang 2009). El árbol de decisión funciona como un clasificador, por lo que habrá N árboles con N resultados de clasificación para una muestra de entrada dada y designada como el resultado final cuando el bosque aleatorio combina todos los resultados de la votación de clasificación (Pinheiro et al. 2021). Es así que RF es un conjunto de varios árboles de decisión independientes diferentes que trabajan juntos; el árbol individual proporciona un pronóstico o clasificación y la predicción que recibe la mayor cantidad de votos se considera predicción o clasificación del modelo de bosque aleatorio (Zhou 2021). Para evitar que un árbol de decisión sobre ajuste los datos, este enfoque elige aleatoriamente un subconjunto de características y divide el conjunto de datos de entrenamiento entre todos los árboles (Rahman, Arafin y Muntasir Billah 2023). Finalmente, el algoritmo de RF tiene ventajas significativas sobre otros métodos en términos de manejo de datos biológicos altamente no lineales, simplicidad de ajuste (en comparación con otros algoritmos de aprendizaje de conjunto) y capacidad para ejecutar procesamiento paralelo. También produce una de las mejores precisiones en muchas aplicaciones (Gajowniczek, Grzegorczyk y Ząbkowski 2019).

Linear Regression (RL) es una técnica estadística utilizada para analizar las interacciones entre dos o más variables. En este proceso se manejan dos categorías de variables: la variable respuesta (también llamada dependiente) y la variable predictora (también conocida como independiente). Cada una toma un valor específico para cada elemento en el conjunto de datos. Este modelo es de naturaleza simple y no se ocupa de la variabilidad y la incertidumbre presente en los datos. (Reyes et al. 2024). Cuando se enfrenta a una gran cantidad de datos o se experimenta una demora significativa en su gestión, es fundamental recurrir a hojas de cálculo o programas estadísticos. Esto garantiza una interpretación y análisis efectivos, oportunos y transparentes de los resultados. En la práctica, es poco frecuente encontrar una relación lineal efectiva entre dos variables. En ocasiones, el investigador se ve en la necesidad de realizar transformaciones en los datos para lograr que la relación sea lineal. Si se identifican observaciones inusuales o influyentes que no pueden pasarse por alto, y la línea de regresión no explica adecuadamente la variabilidad en la v-dependiente en función a los cambios en la variable independiente, es crucial considerar la posibilidad de abandonar el modelo lineal y explorar un enfoque no lineal (Fernando et al. 2013)

k-nearest neighbors (kNN), es un método de clasificación en el aprendizaje supervisado que destaca por su simplicidad y eficiencia. Este algoritmo es ampliamente reconocido como uno de los diez algoritmos clásicos más relevantes en el campo del aprendizaje automático (Wu et al. 2008). Una de las particularidades más interesantes de kNN es su enfoque no paramétrico, lo que significa que no crea un modelo específico durante la fase de entrenamiento. En lugar de ello, se basa en una regla de clasificación que se apoya en una función de similitud calculada entre las instancias de entrenamiento y la instancia que se está tratando de clasificar en ese momento. Esto implica que kNN toma decisiones de clasificación en función de la semejanza entre los datos disponibles y las características de la instancia en cuestión (Derrac et al.

2016) Debido a la efectividad del kNN, se ha convertido en uno de los algoritmos más relevantes (Hu et al. 2020)

Ridge Regression (RR) es una versión avanzada de la regresión lineal que soluciona el problema de la colinealidad entre las variables predictoras. Fue presentada por Hoerl y popularizada por Hastie. Este enfoque utiliza una forma de regresión de mínimos cuadrados regularizada, lo que permite abordar con éxito la multicolinealidad. RR se ha convertido en una herramienta esencial tanto en estadística como en el aprendizaje automático, y resulta especialmente útil cuando se trabaja con conjunto de datos complejos (Zhang et al. 2019).

Para evaluar el rendimiento de estos modelos de predicción se utilizan métricas, error absoluto (RMSE), error absoluto medio (MAE) y coeficiente de determinación (R2) (Patel et al. 2015). Es de gran importancia aplicar índices de medida de error al evaluar la capacidad de pronóstico de los modelos (Gajowniczek, Grzegorczyk y Ząbkowski 2019). RMSE mide el promedio de los cuadrados de los errores, por lo que cuanto más cerca esté el RMSE de 0, mejor. R2 mide que tan bien las predicciones de regresión hechas se aproximan a los puntos de datos reales, y si R2 es 1, indica que las predicciones se ajustan perfectamente a los datos dados. Es decir que cuanto más cerca R2 de 1, mejor será el modelo de predicción (Liang et al. 2018).

La calidad del agua es un conjunto de rasgos físico químicos, microbiológicos y físicos que tienen un impacto directo en el crecimiento adecuado de la biodiversidad (Facultad de Ciencias Ambientales de la Universidad Científica del Sur 2020). Además, es ampliamente reconocido que la calidad del agua es uno de los indicadores más importantes del desarrollo sostenible y es fundamentalmente una cuestión de salud ambiental (Villena Chávez, 2018). Para evaluar la calidad del agua, es fundamental utilizar enfoques innovadores que

incorporen más de dos indicadores. Estas metodologías a menudo implican una serie de parámetros fisicoquímicos, en algunos casos microbiológicos (Rodríguez Fernández Moraima y Lacaba Guardado, Rafael Miguel, 2021).

Los parámetros primordiales para predecir la calidad del agua son pH, turbiedad, total de sólidos disueltos (TDS), demanda Química de Oxígeno (DQO), Oxígeno disuelto (OD). El PH es uno de los parámetros más importantes, mide propensión a la acidez o la alcalinidad, está determinada principalmente por sustancias como el dióxido de carbono, el carbono y el bicarbonato (Jeyashanthi et al. 2023). El rango de pH es de 6.5 a 8.5, no es saludable para el cuerpo humano si el pH es inferior a 6.5. Quiere decir que los valores más altos de pH corresponden a agua más básica, mientras que los valores de pH más bajos se asocian con agua ácida (Jain, Yevatikar y Raxamwar 2022).

La turbiedad es un término utilizado para describir cuánta transparencia se ha perdido en el agua como resultado de partículas de agua suspendidas, es considerado como un indicador fiable de la calidad de agua (Gupta et al. 2011) provocada por material coloidal y suspendida, incluidos pequeños fragmentos de material orgánico e inorgánico, arcilla, limo y otras bacterias en el agua (Jain, Yevatikar y Raxamwar 2022). El TDS es una medida para determinar los constituyentes del agua.

La vida y el crecimiento de los seres vivos acuáticos tienden a perderse con concentraciones fuertes o leves de TDS. Si se tiene una concentración superior a 500 mg/L el agua no es apta para su consumo (Gupta et al. 2011). OD es un elemento crucial para la vida en el océano, sin DO ningún organismo puede existir, por lo tanto, esta medida captura los procesos biológicos que tienen lugar en los cuerpos de agua (Tyagi et al. 2003), se utiliza en la cuantificación indirecta de contaminantes orgánicos en aguas superficiales. Denota la masa de oxígeno necesaria para que los compuestos orgánicos que están unidos al agua se oxiden en amoniaco, dióxido de carbono y agua (Jain, Yevatikar y Raxamwar 2022).

III. METODOLOGÍA

3.1 Tipo y diseño de investigación

i. Tipo de investigación

El presente estudio tiene un enfoque aplicado, ya que su objetivo principal es resolver un problema práctico y específico que enfrentan las personas o la sociedad en general. Además, aborda cuestiones comunes que afectan la calidad de vida, el trabajo, la salud y el bienestar (Truijens et al. 2019). Asimismo, la investigación aplicada da prioridad a los problemas que afectan con frecuencia a las personas (Hameed et al. 2007), su principal objetivo de la Investigación Aplicada es resolver problemas del mundo real para contribuir a nuestra comprensión del fundamento de la evolución y sus consecuencias en diversos problemas del mundo real. En tal sentido, dicha información ejerce como referencia futura útil para problemas relacionados (Truijens et al. 2019), ya que buscar informar con la práctica (Park, 2007).

En este sentido, la investigación que se realizará es aplicada porque se aplicará conocimiento teórico y modelos de ML, para resolver un problema práctico y específico: predecir la calidad del agua utilizando datos históricos

ii. Diseño de investigación

El diseño de la investigación es experimental puro porque se controlan todos los factores que puedan afectar al proceso para conseguir el resultado deseado (Yan, Zhou y Shen 2023). Es decir, la variable independiente se refiere al factor que se considera como causa propuesta que influye en la relación entre variables (Collado R. Hernández Sampieri y P. Lucio L. 1997). En este contexto, la investigación sigue un diseño experimental puro, porque se manipularon activamente los parámetros de los modelos ML para predecir la calidad del agua. Asimismo, se

considerará la aleatorización de datos a los conjuntos de entrenamiento y prueba para reducir sesgos y obtener resultados más confiables.

3.2 Variables y operacionalización

En el presente estudio se tiene dos variables:

Variable independiente: Modelo Machine Learning.

Variable dependiente: Calidad del agua.

En el anexo 1 se encuentra la tabla de operacionalización.

3.3 Población y muestra

3.3.1 Población

(Arias-Gómez, Ángel Villasís-Keever y Guadalupe Miranda-Novales 2016) Indica que las poblaciones de estudio consisten en un conjunto de individuos fácilmente accesibles, limitados y capaces de proporcionar la base para el muestreo, y cada uno de los cuales cumple varios criterios predeterminados. Además, especificar la población de estudio al principio es muy importante, porque al final será posible generalizar o extrapolar los resultados obtenidos del estudio al resto de la población o universo en su conjunto.

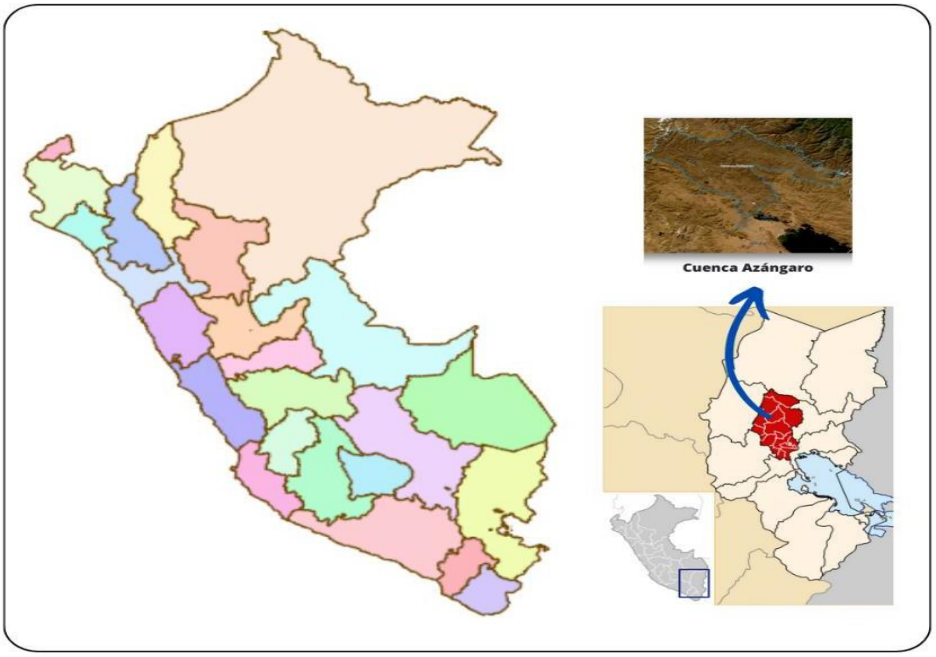
En el presente estudio, se seleccionó como población de interés la cuenca de Azángaro, ubicada en la sierra de Perú, con un área de 8754.00 km^2, considerada una de las cuencas más relevantes en el país. La selección de esta cuenca se basó en la disponibilidad de datos históricos de monitoreo de calidad del agua. Lo cual brindó una base sólida para la investigación.

El criterio de inclusión especifica los datos deben tener una duración mínima de alrededor de 11 años, 2011 a 2021. Esto garantizo que se dispóngase de una cantidad significativamente de datos históricos para realizar un análisis eficaz, además abarca

que los datos tenían intervalos regulares de un año, lo cual es adecuado para el análisis.

La cuenca de Azángaro se sitúa entre la cordillera oriental y el altiplano occidental, abarcando las laderas del lago Titicaca, y se encuentra políticamente dentro de la región de Puno (Carpio, Baclimer y Yanapa 2021).

Gráfico .1 Población Cuenca Azángaro.



Fuente: Elaboración propia

iii. Muestra

La muestra fue obtenida de la base de datos que cuenta la institución nacional SNIRH de Perú dentro de la temática calidad del agua, obteniendo un total de 136 muestras de la base de datos en un periodo de 10 años 2011 a 2021.

La selección de datos se llevó a cabo siguiendo criterios específicos que fueron desarrollados a partir de una revisión exhaustiva de la literatura científica relacionada con la calidad del agua en estudios similares. El método ICA – NSF, que constituye una de las variables clave de este estudio, fue la base metodológica seleccionada. Este método fue diseñado inicialmente por Brown, tomando como referencia una versión basada en el índice de calidad del agua elaborado por la fundación de Sanidad Nacional de EE.UU.

La elección de los parámetros se fundamenta en aquellos parámetros empleados por el ICA-NSF, ya que son considerados los más confiables para medir la naturaleza del agua dulce (Cutillas et al. 2019).

La muestra final se conformó asegurando que los datos fueran confiables y representativos, lo que permitirá realizar un análisis riguroso y obtener resultados significativos para el desarrollo del modelo.

iv. Muestreo

En el marco de esta investigación, el tipo de muestreo es no probabilístico porque para la selección de parámetros relevantes destinados al desarrollo del modelo ML no fueron seleccionados aleatoriamente. En cambio, se basó en criterios específicos que se consideraron relevantes de ciertos parámetros en predicción de calidad de agua, según experimentos de anteriores trabajos de investigación.

Por ejemplo, según (Bui et al. 2020a) llevaron a cabo varios experimentos para evaluar la QWI con base en un conjunto de datos que recopilaron de 2007 al 2020, para esto aplicaron distintos modelos de selección de características para seleccionar los parámetros clave o de más importancia que alimentan los modelos ML. Donde descubrieron que usar solo

cuatro parámetros, coliformes, DO, turbidez y TSS, es suficiente para que el modelo de RF calcule el WQI con precisión.

v. Unidad de análisis

La unidad de análisis de esta investigación está constituida por los datos históricos de monitoreo de calidad de agua en la cuenca Azángaro, abarcando un período de tiempo que va desde el año 2011 hasta el 2021. Estos datos permitirán una evaluación a lo largo del tiempo de la calidad del agua en la región, lo que brindará una visión amplia de las variaciones.

3.4 Técnicas e instrumentos de recolección de datos

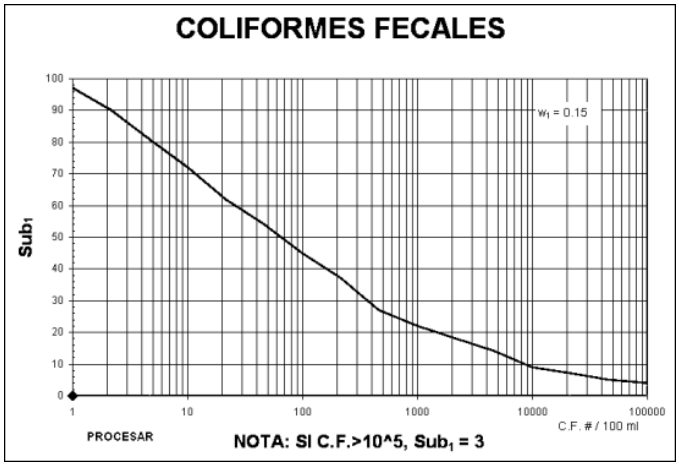
Para calcular el índice de Brown se usó una suma lineal ponderada de los subíndices:

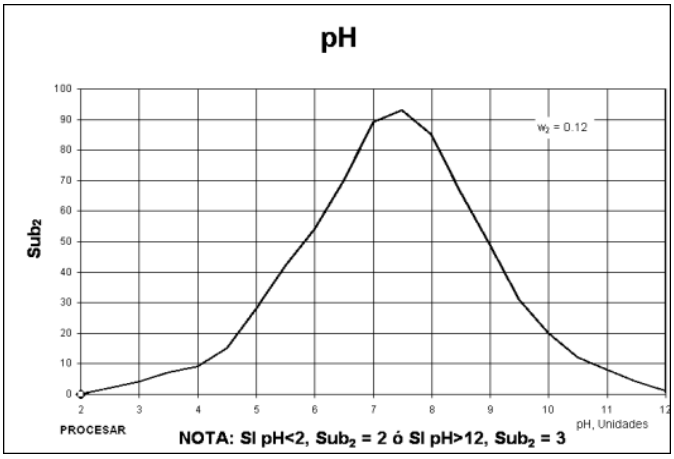
$$ICA = \sum_{i=1}^{9} (sub_i * W_i)$$

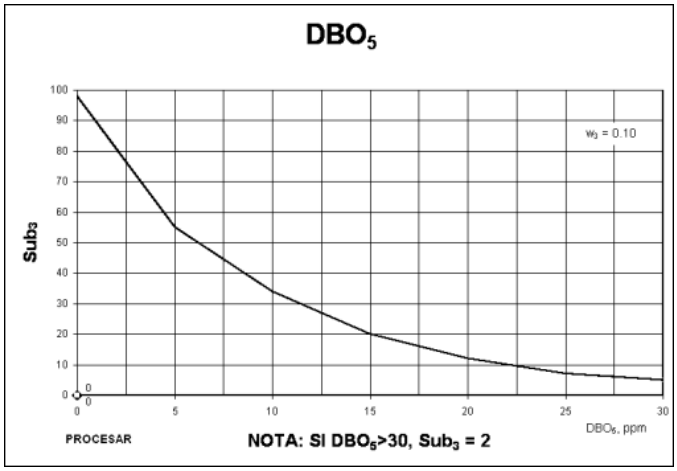
Dónde:

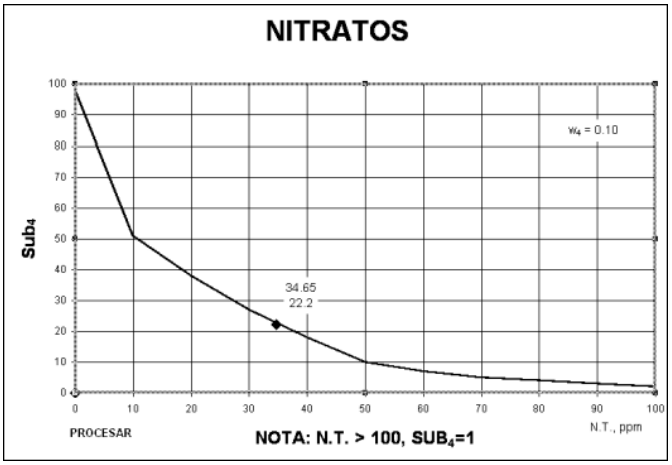
W_i: Pesos relativos asignados a cada parámetro sub_i, y ponderamos entre 0 y 1 de tal forma que la sumatoria sea igual a uno. Para determinar el valor del “ICA” es necesario sustituir los datos de la ecuación antes mencionada obteniendo así sub_i de distintas graficas que se detallara a continuación, dicho valor se multiplica por su respectivo W_i(Brown et al. 1970).

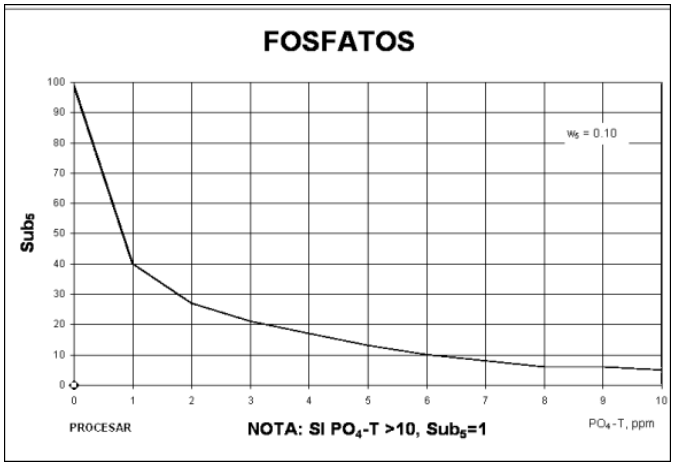
Gráfico .2 Parámetros para índice de calidad general de Brown

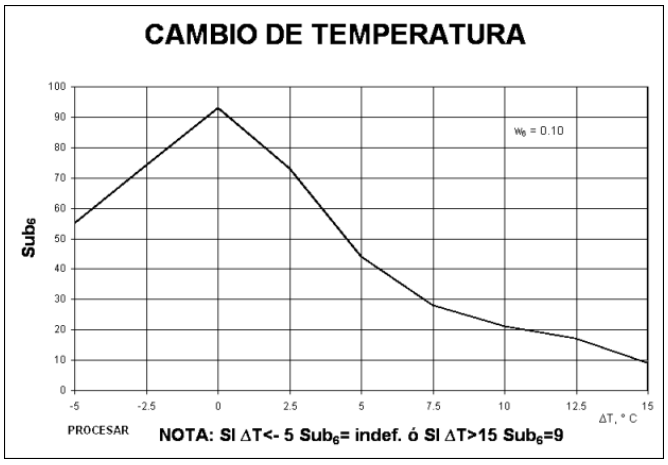


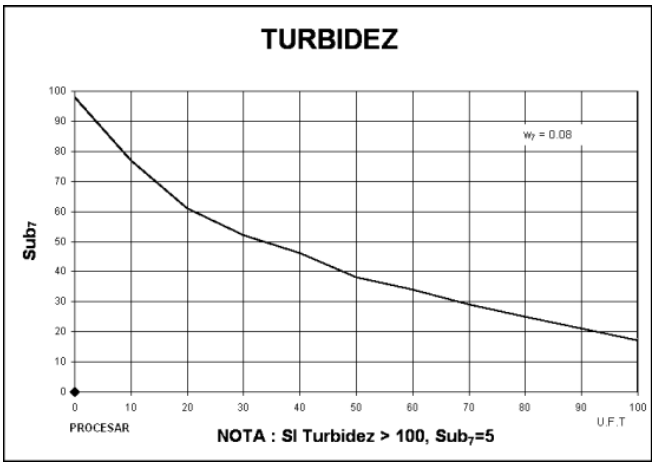


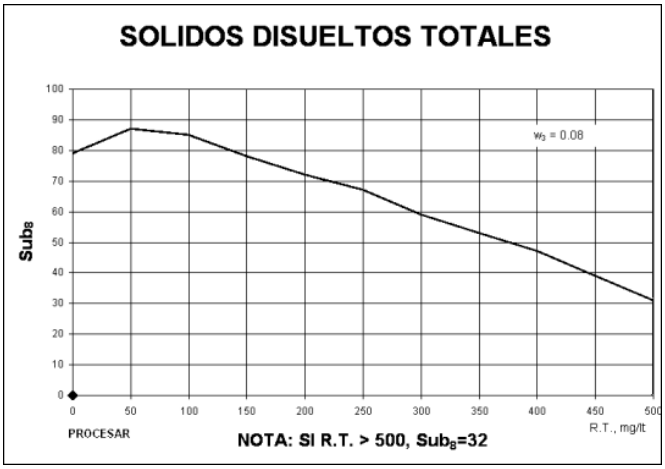


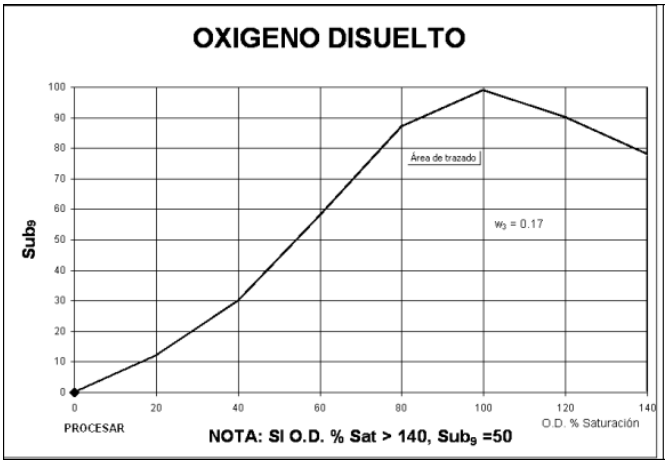












Fuente: SNET (Servicio Nacional de Estudios Territoriales)

Seguidamente se analizó si había datos faltantes dentro de la base de datos.

Gráfico .3 Análisis de datos faltantes en dataset

```
train_df.isnull().sum()
```

```
Periodo                                0
Conductividad                          8
Demanda Bioquimica de oxigeno (DBO5)   13
Oxigeno Disuelto                       8
PH                                     1
Temperatura                            1
Coliformes Fecales                     19
Nitratos                               27
Turbidez                               53
dtype: int64
```

Fuente: Elaboración propia

Como se pudo apreciar se encontró valores faltantes, para esta investigación se usó una técnica para la imputación de valores faltantes conocido como KNN significa (k-Nearest Neighbors) y hace referencia a un algoritmo de clasificación y regresión que se basa en encontrar los vecinos más cercanos a un punto dado.

Gráfico .4 KNN (k-Nearest Neighbors)

```
from sklearn.impute import KNNImputer

column_names = train_df.columns
imputer = KNNImputer(n_neighbors=5)
train_df = imputer.fit_transform(train_df)
```

Fuente: Elaboración propia

Software que se usara para el desarrollo de la investigación, cuentan con certificación internacional por su amplio uso en la inteligencia artificial y afines.

Python: Es un lenguaje de programación interpretado, de alto nivel y de propósito general (Virtanen et al. 2020), en el campo de la investigación permite a los científicos agregar operaciones de matriz rápida y álgebra lineal y de esta manera logran hacer su trabajo dentro de un solo lenguaje de programación (Harris et al. 2020).

Panda: Es una librería de código abierto, la cual surgió de la necesidad de disponer de una librería específica para analizar los datos usando Python como lenguaje de programación (Fabio Nelli 2018).

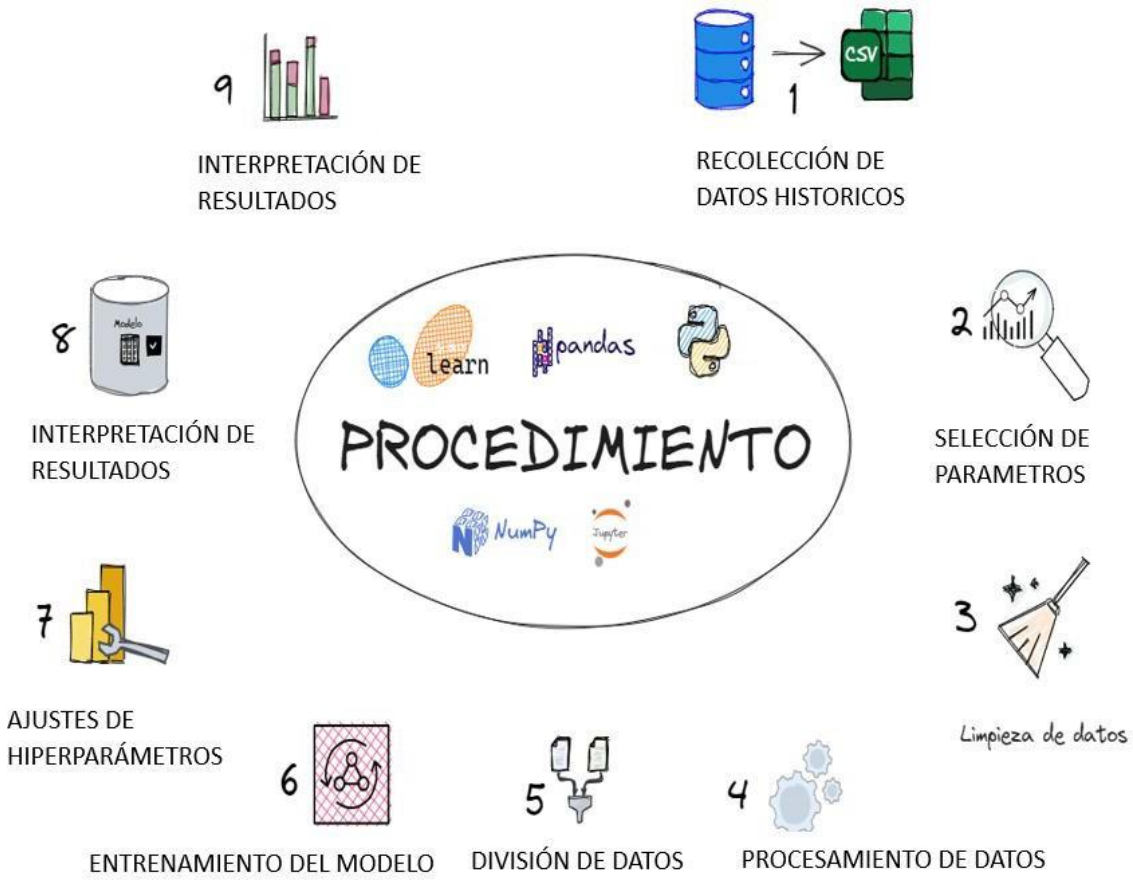
NumPy: También conocido como python numérico, es una biblioteca base en la computación científica, ya que proporciona estructuras de datos y también funciones de alto rendimiento. Sirve para desarrollar cálculos numéricos, cálculos de matrices multidimensionales y cálculos de matrices de gran tamaño (Fabio Nelli 2018).

Scikit-learn: Se trata de un paquete de Python que incorpora una diversidad de algoritmos de ML, abarcando tanto problemas supervisados como no supervisados. A pesar de su amplia funcionalidad, se caracteriza por mantener una interfaz sencilla y perfectamente integrada con el lenguaje Python. Además, cuenta con una licencia BSD (Pedregosa, FABIANPEDREGOSA et al. 2011).

3.5 Procedimiento

A continuación, se explican detalladamente los procedimientos que se ejecutarán para realizar este trabajo. Además, estos procedimientos han sido representados en un diagrama para una mejor visualización del proceso.

Gráfico .5 Diagrama de procedimientos



Fuente: Elaboración propia

- Etapa 1: Recolección de datos históricos

  La calidad del agua en la cuenca de Azángaro es monitoreada por ANA Perú. ANA cuenta con un sistema integral control/garantía de calidad para garantizar que los datos generados por el programa de monitoreo nacional

sean confiables y tengan suficiente exactitud y precisión. Asimismo, desde el año 2009 viene realizando monitoreos participativos de la calidad del agua. Por consiguiente, se recopilaron datos históricos de monitoreo de calidad de agua en la cuenca de Azángaro durante un periodo de 11 años (2011-2021), se obtendrá estos datos de la plataforma ANA | SNIRH (Sistema Nacional de Información de Recursos Hídricos).

Gráfico .6 Plataforma SNIRH – ANA

Fuente: Pagina web de SNIRH - ANA

- Etapa 2: Selección de parámetros relevantes

  Los modelos ICA tradicionales existentes utilizan una variedad de enfoques estadísticos para seleccionar parámetros cruciales de la calidad de agua. De toda la data de la cuenca se tomó el criterio de ICA-NSF. Este índice es ampliamente usado entre todos los índices de calidad de agua existentes para la determinación del ICA intervienen 9 parámetros, los cuales son:

  1. Temperatura(°C)
  2. PH (unidades de pH)
  3. Oxígeno Disuelto (OD % saturación)

4. Demanda Bioquímica de Oxígeno (mg/L)
5. Conductividad
6. Coliformes Fecales (NMP/100ml)
7. Nitratos (NO3 mg/L)
8. Turbidez (FAU)
9. Fosfatos (mg/L)

- Etapa 3: Procesamiento y limpieza de datos

  En esta etapa se limpiará y preproceso los datos recopilados previamente elegidos para eliminar valores faltantes, corregir ciertos errores y si se encuentra datos atípicos, eliminarlos. Este paso es sumamente crucial para asegurar la calidad de los datos y evitar sesgos en el modelo.

Gráfico .7 Data base de Cuenca Azángaro asignada a variable data

Cargar archivo Excel

```
[ ] data = pd.read_excel(r'parametros indice de brown.xlsx')
```

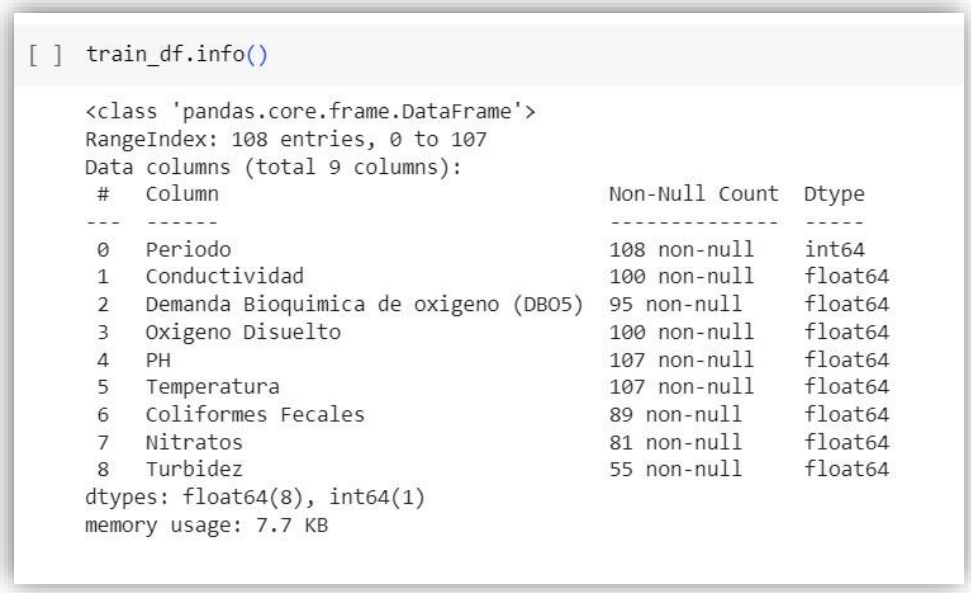
```
data.head()
```

	Periodo	Conductividad	Demanda Bioquimica de oxigeno (DBO5)	Oxigeno Disuelto	PH	Temperatura	Coliformes Fecales	Nitratos	Turbidez
0	2011	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2.0
1	2011	1046.0	NaN	5.36	8.03	11.9	NaN	NaN	3.0
2	2011	694.0	NaN	5.75	8.37	10.5	NaN	NaN	3.0
3	2011	62.1	NaN	5.43	7.38	13.1	NaN	NaN	3.0
4	2011	203.5	NaN	6.35	8.71	13.1	NaN	NaN	3.0

Fuente: Elaboración propia

Descripción 1: Se almacena en una variable nombrada data el enlace que contiene la base de datos de los monitoreos de calidad del agua durante un periodo de 11 años. Para lograrlo, se utiliza la biblioteca Pandas con la función de lectura de archivos Excel.

Gráfico .8 Análisis del tipo de dato.

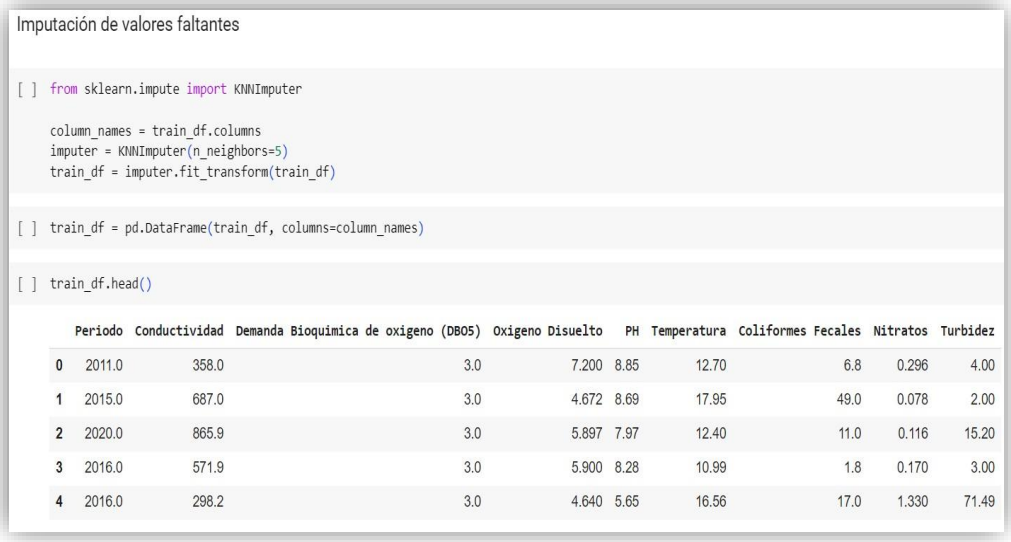


```
[ ]  train_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 108 entries, 0 to 107
Data columns (total 9 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Periodo                               108 non-null    int64
 1   Conductividad                         100 non-null    float64
 2   Demanda Bioquimica de oxigeno (DBO5)  95 non-null     float64
 3   Oxigeno Disuelto                      100 non-null    float64
 4   PH                                    107 non-null    float64
 5   Temperatura                           107 non-null    float64
 6   Coliformes Fecales                    89 non-null     float64
 7   Nitratos                              81 non-null     float64
 8   Turbidez                              55 non-null     float64
dtypes: float64(8), int64(1)
memory usage: 7.7 KB
```

Fuente: Elaboración propia.

Descripción 2: Se realiza un análisis del tipo de dato de los parámetros previamente leídos. Este análisis es esencial, ya que los modelos de machine learning operan con datos numéricos, ya sea en formato float(decimal) o int (enteros).

Gráfico .9 Imputación de valores faltantes



Imputación de valores faltantes

```
[ ]  from sklearn.impute import KNNImputer

     column_names = train_df.columns
     imputer = KNNImputer(n_neighbors=5)
     train_df = imputer.fit_transform(train_df)

[ ]  train_df = pd.DataFrame(train_df, columns=column_names)

[ ]  train_df.head()
```

	Periodo	Conductividad	Demanda Bioquimica de oxigeno (DBO5)	Oxigeno Disuelto	PH	Temperatura	Coliformes Fecales	Nitratos	Turbidez
0	2011.0	358.0	3.0	7.200	8.85	12.70	6.8	0.296	4.00
1	2015.0	687.0	3.0	4.672	8.69	17.95	49.0	0.078	2.00
2	2020.0	865.9	3.0	5.897	7.97	12.40	11.0	0.116	15.20
3	2016.0	571.9	3.0	5.900	8.28	10.99	1.8	0.170	3.00
4	2016.0	298.2	3.0	4.640	5.65	16.56	17.0	1.330	71.49

Fuente: Elaboración propia.

Descripción 3: Se lleva a cabo la imputación de valores faltantes. En el archivo de datos, es común encontrar valores no registrados en ciertas fechas. Antes de entrenar el modelo, es crucial realizar una limpieza y análisis de los datos. Para esta operación, se optó por utilizar el método KNNinputer con un parámetro de vecino igual a 5 (n_neighbors=5).

Gráfico .10 Renombramiento de parámetros

```
column_names_abbrev = {'Conductividad': 'co',
                       'Demanda Bioquimica de oxigeno (DBO5)': 'dbo',
                       'Demanda Quimica de oxigeno (DQO)' : 'dqo',
                       'Oxigeno Disuelto' : 'od',
                       'PH' : 'ph',
                       'Temperatura' : 'temp',
                       'Coliformes Fecales' : 'ct',
                       'Nitratos' : 'ni',
                       'Turbidez' : 'turb'}
```

```
[ ]  train_df.rename(columns=column_names_abbrev, inplace=True)
```

```
[ ]  train_df.head()
```

	Periodo	co	dbo	od	ph	temp	ct	ni	turb
0	2011.0	358.0	3.0	7.200	8.85	12.70	6.8	0.296	4.00
1	2015.0	687.0	3.0	4.672	8.69	17.95	49.0	0.078	2.00
2	2020.0	865.9	3.0	5.897	7.97	12.40	11.0	0.116	15.20
3	2016.0	571.9	3.0	5.900	8.28	10.99	1.8	0.170	3.00
4	2016.0	298.2	3.0	4.640	5.65	16.56	17.0	1.330	71.49

Fuente: Elaboración propia.

Descripción 4: Se llevó a cabo la tarea de renombrar cada parámetro con el fin de facilitar su análisis.

- Etapa 4: División de los datos

  En esta etapa se dividirá el conjunto de datos en dos partes, el conjunto de entrenamiento y el conjunto de prueba. El conjunto de entrenamiento se usará para entrenar el modelo ML, mientras que el conjunto de prueba se utilizará para evaluar el rendimiento del modelo ML.

Gráfico .11 División de datos

Fuente: Elaboración propia.

Descripción 5: Se realizó la partición de los datos, asignando una parte para el entrenamiento y la otra para las pruebas.

- Etapa 5: Entrenamiento del modelo

  En esta etapa se implementará los algoritmos Linear Regression, Ridge Regression, K-Neighbors Regression, Decision Tree, Random Forest, dichos modelos se extraerán de una biblioteca de ML como scikit-learn en Python. Se utilizará el conjunto de entrenamiento para entrenar al modelo, lo que

implica ajustar los parámetros internos del algoritmo para que se adapten mejor a los datos.

Gráfico .12 Importación de modelos

```
[ ] import model_selection
    from model_selection import *
    from sklearn.metrics import mean_squared_error, mean_absolute_error
    from sklearn.metrics import r2_score, mean_absolute_percentage_error
    from sklearn.ensemble import RandomForestRegressor
    from sklearn.neighbors import KNeighborsRegressor
    from sklearn.linear_model import LinearRegression
    from sklearn.linear_model import Ridge
```

Fuente: Elaboración propia.

Gráfico .13 Modelo Linear Regression

```
models = {
    'Linear Regression': LinearRegression()
}
```

```
cv_results = cross_validate(models, X, y, scores, feature_transformer, n_splits=5)
print_table_results(cv_results, cv=True)
#print_results(cv_results, cv=True)
```

Fuente: Elaboración propia.

Gráfico .14 Modelo Ridge Regression

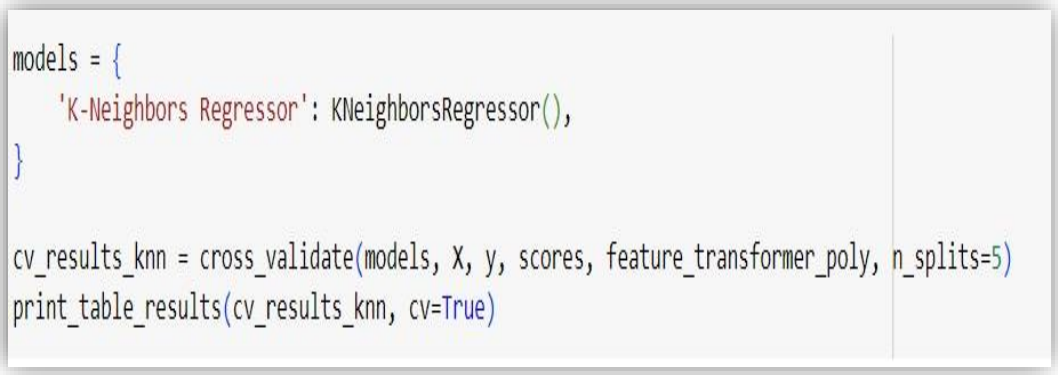
```
from sklearn.linear_model import Ridge

models = {
    'Linear Regression (poly)': LinearRegression(),
    'Ridge Regression (poly)': Ridge()
}
```

+ Código + Texto

```
cv_results_ridge = cross_validate(models, X, y, scores, feature_transformer_poly, n_splits=5)
print_table_results(cv_results_ridge, cv=True)
#print_results(cv_results, cv=True)
```

Fuente: Elaboración propia.

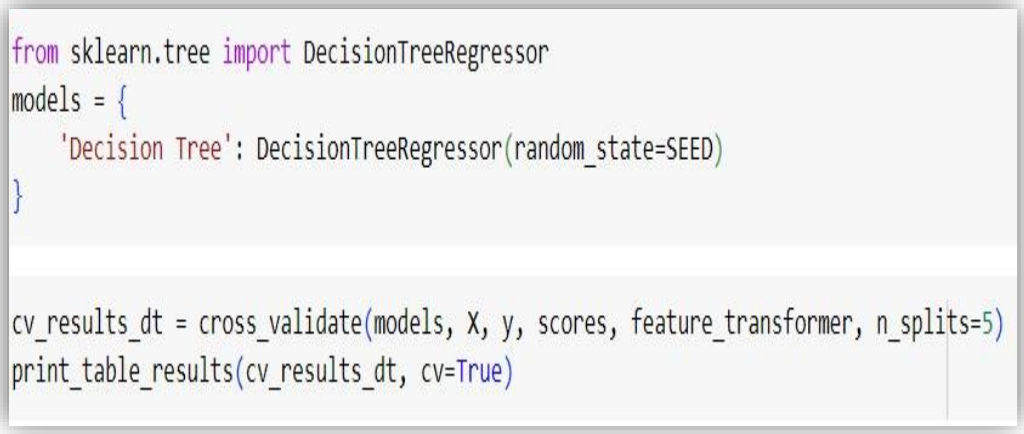
Gráfico .15 K-Neighbors Regressor



```
models = {
    'K-Neighbors Regressor': KNeighborsRegressor(),
}

cv_results_knn = cross_validate(models, X, y, scores, feature_transformer_poly, n_splits=5)
print_table_results(cv_results_knn, cv=True)
```

Fuente: Elaboración propia.

Gráfico .16 Decision Tree



```
from sklearn.tree import DecisionTreeRegressor
models = {
    'Decision Tree': DecisionTreeRegressor(random_state=SEED)
}

cv_results_dt = cross_validate(models, X, y, scores, feature_transformer, n_splits=5)
print_table_results(cv_results_dt, cv=True)
```

Fuente: Elaboración propia.

Gráfico .17 Random Forest

```
models = {
    'Random Forest': RandomForestRegressor(random_state=SEED)
}

cv_results_rf = cross_validate(models, X, y, scores, feature_transformer, n_splits=5)
print_table_results(cv_results_rf, cv=True)
```

Fuente: Elaboración propia.

- Etapa 6: Validación a modelos

  En esta etapa se evaluará el rendimiento del modelo ML, para esto se utilizará métricas de evaluación de regresión, MSE, RMSE, MAE y el R2. Estas métricas proporcionarán una medida de que tan bien se ajusta el modelo a los datos de prueba y que tan bien puede predecir la calidad del agua.

  Gráfico .18 Validación de modelo usando métricas.

```
[ ]  def root_mean_squared_error(y_true, y_pred):
         return mean_squared_error(y_true, y_pred, squared=False)

     scores = {'mse': mean_squared_error,
               'rmse': root_mean_squared_error,
               'mae': mean_absolute_error,
               'r2': r2_score,
               'mape': mean_absolute_percentage_error
     }
```

Fuente: Elaboración propia.

Descripción 6 Se utilizaron métricas para evaluar el rendimiento de cada modelo en la predicción de la calidad del agua.:

3.6 Método de análisis de datos

En la fase de análisis de datos de esta investigación, se llevarán a cabo varias etapas clave. En primer lugar, se realizará una limpieza y preprocesamiento exhaustivo de los datos históricos de la calidad del agua obtenidos de la cuenca de Azángaro. Posteriormente, se dividirán los datos en conjuntos de entrenamiento y prueba para entrenar el modelo de RF utilizando la biblioteca Scikit-learn en Python. A través de una búsqueda de hiperparámetros mediante validación cruzada, se afinarán los ajustes internos del modelo para lograr un mejor rendimiento en la predicción. Finalmente, se evaluará el modelo utilizando métricas de regresión como el MSE, RMSE, MAE y R2, y se interpretarán los resultados para comprender la relevancia de cada parámetro en la predicción de calidad del agua.

3.7 Aspectos éticos

El trabajo de investigación se fundamenta en principios esenciales de integridad y legitimidad, garantizando la veracidad y exactitud de los datos sin ninguna alteración. Se han respetado rigurosamente los derechos de autor, citando apropiadamente a los autores en cada párrafo. Asimismo, se ha asegurado que el trabajo sea original y libre de plagio mediante la supervisión con el programa anti plagio Turnitin.

IV. RESULTADOS

En el marco de este estudio se ha desarrollado y evaluado cinco modelos de aprendizaje ML, buscando determinar de manera eficiente cuál de ellos presenta el mejor rendimiento promedio. Cabe mencionar que los resultados obtenidos se han desglosado y presentado de manera individual para cada modelo.

Regresión Lineal, en este primer modelo, indica que los resultados obtenidos en la fase de validación revelaron un desafío significativo en la capacidad predictiva de este modelo. Específicamente, el Mean Squared Error (MSE) en el conjunto de validación alcanzó un valor elevado de 116.681. Este hallazgo tiene implicaciones importantes para la investigación, ya que apunta a limitaciones en la capacidad del modelo para capturar con precisión las relaciones subyacentes entre las variables de entrada y la variable de salida, que en este caso es la calidad del agua. La alta magnitud del MSE indica que las predicciones del modelo estuvieron lejos de los valores reales, lo que dificulta su capacidad para generalizar y realizar predicciones precisas en un contexto más amplio. El RMSE en el conjunto de validación alcanzó un valor de 4.512. Indicando que el modelo lineal tiene limitaciones para generalizar de manera efectiva a nuevos datos. Este incremento en el RMSE entre conjuntos indico cierta falta de capacidad del modelo para adaptarse a la variabilidad en los datos de validación. El R^2(Coeficiente de determinación) proporciona información sobre la proporción de la varianza en la variable de salida que el modelo es capaz de explicar. En términos sencillos, un valor de R^2 cercano a 1 indica que el modelo tiene la capacidad de explicar una gran parte de la variabilidad presente en los datos. Por otro lado, un valor R^2 cercano a 0 o incluso negativo indica que el modelo tiene dificultades para ofrecer una explicación adecuada de la variabilidad observada. Como parte de esta investigación, observamos que el valor de R^2 en el conjunto de validación fue negativo (-2.534). Esto indicó que el modelo no pudo proporcionar una explicación adecuada para la variabilidad presente en los datos de validación. MAE

(Error Absoluto Medio), en validación (5.708) es relativamente alto, esto indica que las predicciones del modelo tienen un error absoluto medio significativo, lo que implica que el modelo no se ajusta bien a los datos.

Ridge Regression, el modelo muestra un MSE de 6.128 en el conjunto de entrenamiento y 14.140 en el conjunto de validación. Esto indica que el modelo tuvo un buen equilibrio entre el ajuste a los datos de entrenamiento y la capacidad de generalización, ya que el MSE en validación es significativamente menor que el modelo de regresión lineal. El RMSE en entrenamiento es 2.473 y 3.674 en el conjunto de validación. La regularización de Ridge contribuyó a un mejor rendimiento en el conjunto de validación al controlar el sobreajuste. Aunque hay una diferencia entre las métricas de entrenamiento y validación, esta diferencia es menor en comparación con la regresión lineal. El MAE en entrenamiento es 1.914, indica que las predicciones en los datos de entrenamiento tienen un error absoluto medio bajo. En validación, el MAE es 2.906, lo que indica que el modelo es capaz de realizar predicciones con un error absoluto medio relativamente bajo en datos no vistos. El R^2 en entrenamiento es 0.804, lo que indica que el modelo explica una cantidad razonable de la variabilidad en los datos de entrenamiento. En validación, el R^2 es 0.494, lo que indica una capacidad moderada de generalización. En este modelo presentó un MSE significativamente menor en el conjunto de validación a comparación de LR, lo que indica una mejora en la capacidad de generalización. El R^2 en validación del modelo de RR (0.494) es positivo y considerablemente mejor que el valor negativo del modelo de LR, lo que indica una mejor capacidad de explicar la variabilidad en los datos de validación.

K-neighbors Regressor, el modelo mostró un MSE extremadamente bajo en el conjunto de entrenamiento (0.000) y un MSE significativamente más alto en el conjunto de validación (21.213). Este contraste entre el MSE de entrenamiento y el MSE de validación tiene un problema de sobre ajuste, ya que el modelo parece ajustarse perfectamente a los datos de

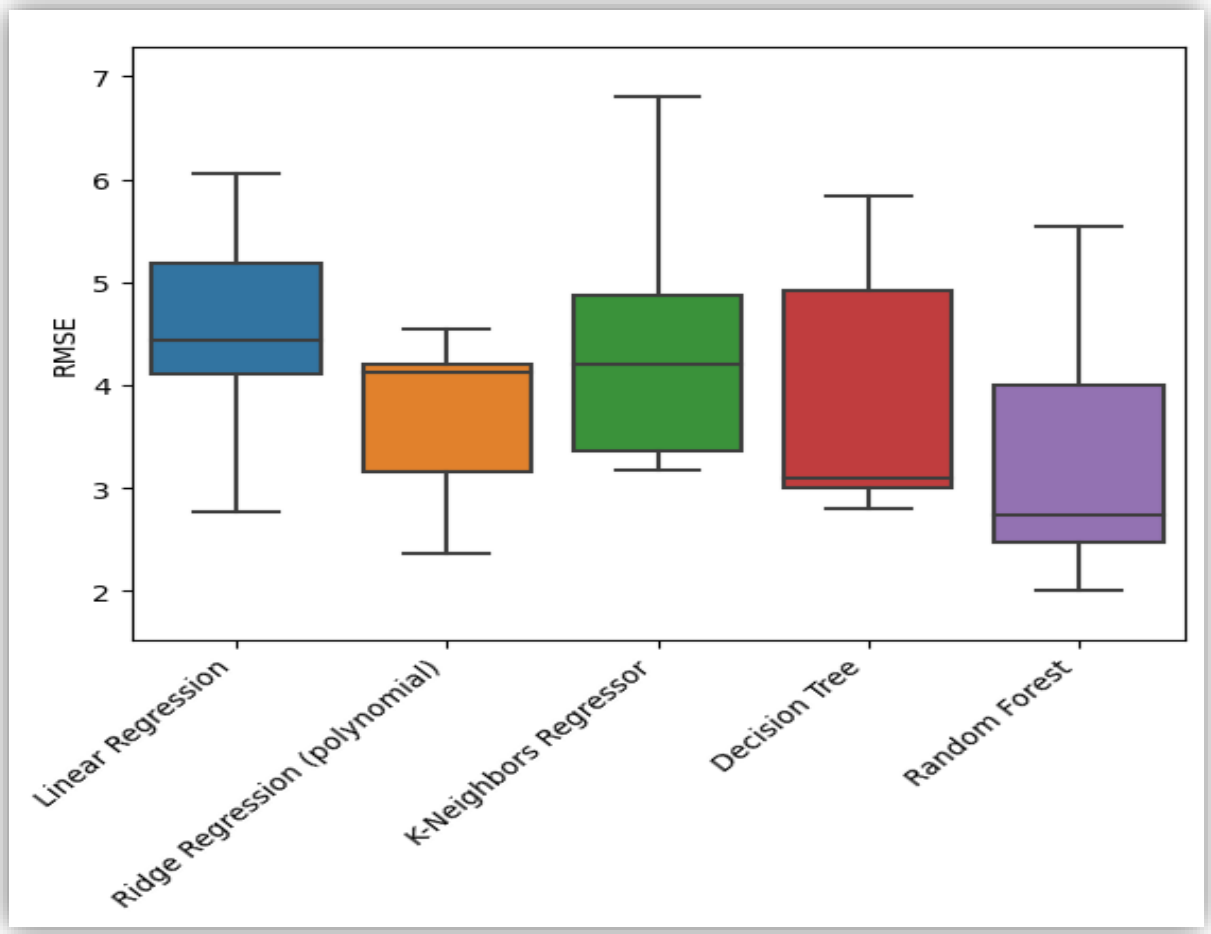
entrenamiento, pero no generaliza bien nuevos datos. El RMSE de 0.0 00 en el conjunto de entrenamiento, es inusual y podría indicar un ajuste perfecto a los datos de entrenamiento. Sin embargo, en el conjunto de validación, el RMSE es de 4.485, lo que indicó que el modelo no generaliza bien a nuevos datos. Este resultado indica que el modelo es demasiado complejo y se ajusta demasiado a los datos de entrenamiento. El MAE en entrenamiento es 0.000, lo que indica que el modelo no comete errores en los datos de entrenamiento, pero el MAE en validación es 3.429, lo que indica que el modelo comete errores significativos en los datos no vistos. El R^2 en entrenamiento es 1.000, lo que indica que el modelo explica la totalidad de la variabilidad en los datos de entrenamiento, lo que es un claro signo de sobre ajuste. En validación, el R^2 es relativamente bajo (0.296). Lo que indica que el modelo no puede explicar bien la variabilidad de los datos.

Dada esta situación, se consideró pertinente explorar alternativas y evaluar otros modelos para determinar si alguno de ellos podría ofrecer un desempeño más favorable.

Decision Tree, este modelo muestra un MSE de 16.910 en el conjunto de validación, lo que indica una capacidad de generalización razonable en comparación con los modelos iniciales. El RMSE de 1.817 en el conjunto de entrenamiento y 3.927 en el conjunto de validación. Aunque el modelo ajusta bien a los datos de entrenamiento, la brecha entre el RMSE de entrenamiento y validación indica que podría haber sobre ajuste. Esto significa que el modelo puede estar capturando demasiado los detalles específicos de los datos de entrenamiento y no generalizando bien a nuevos datos. El MAE en validación es 2.946, lo que implica un error absoluto medio moderado en las predicciones de datos de validación. En el R^2 en validación es 0.464, lo que indica que el modelo es capaz de explicar una parte significativa de la variabilidad.

Random Forest, este modelo presentó un MSE en validación de 12.886, que es inferior al de regresión lineal, ridge, k-neighbors Regressor y Decision Tree, Esto indica mejora en la capacidad de generalización y una reducción en el error cuadrático medio. El RMSE fue de 1.364 en el conjunto de entrenamiento y 3.354 en el conjunto de validación. Estos valores indican que el modelo generaliza bien a nuevos datos, ya que tienen un bajo error en ambos conjuntos. La brecha entre las métricas es relativamente baja. El MAE en validación es 2.563, lo que indica un error absoluto medio moderado. El R^2 en validación es 0.613, esto indica que el modelo RF es capaz de explicar una parte significativa de la variabilidad en los datos de validación, indicando una mayor capacidad de generalización en comparación con los demás modelos.

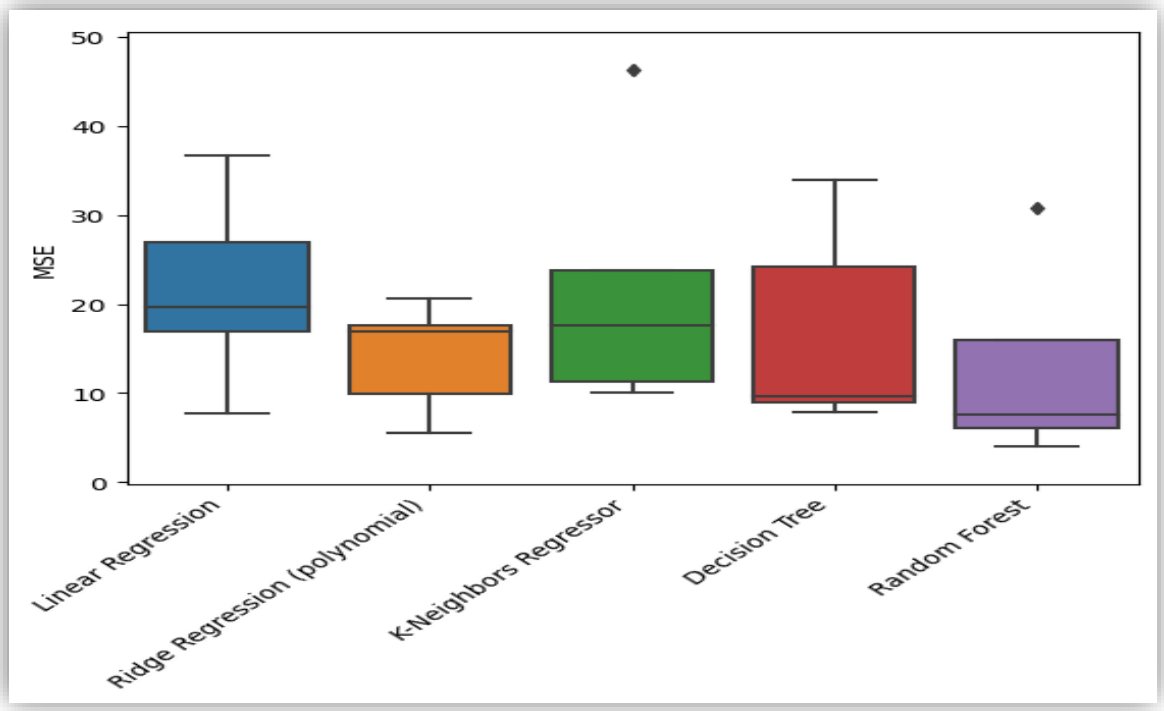
Gráfico .19 Métrica Raíz del error cuadrático medio (RMSE)



Fuente: Elaboración propia.

Descripción 7: Comparación de modelos con RMSE, en la evaluación de varios modelos, se destaca el Random Forest como el modelo con mejor resultado, exhibiendo un RMSE de 1.364 en entrenamiento y 3.354 en validación, lo que indica una consistencia y precisión notables en la generación a nuestros datos. Por otro lado, Linear Regression muestra el peor desempeño, evidenciado por un RMSE más alto de 4.512 en el conjunto de validación, indicando una mejor capacidad para ajustarse eficazmente a estos datos específicos.

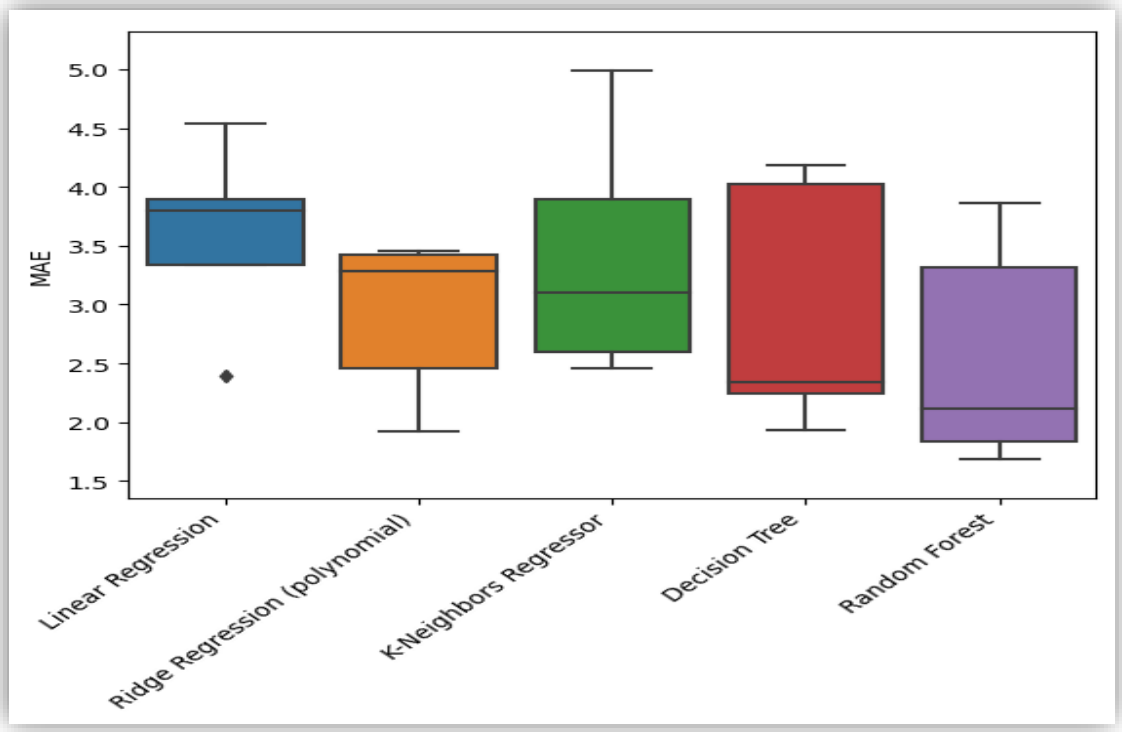
Gráfico .20 Métrica Error Cuadrático Medio (MSE)



Fuente: Elaboración propia.

Descripción 8: Comparación de modelos acorde a MSE, el modelo más destacado resultó ser Random Forest, evidenciado por un MSE en validación de 12.886, inferior a los demás. Este resultado. Por otro lado, el peor rendimiento lo exhibe Linaear Regression, con un MSE elevado de 116.681 en el conjunto de validación.

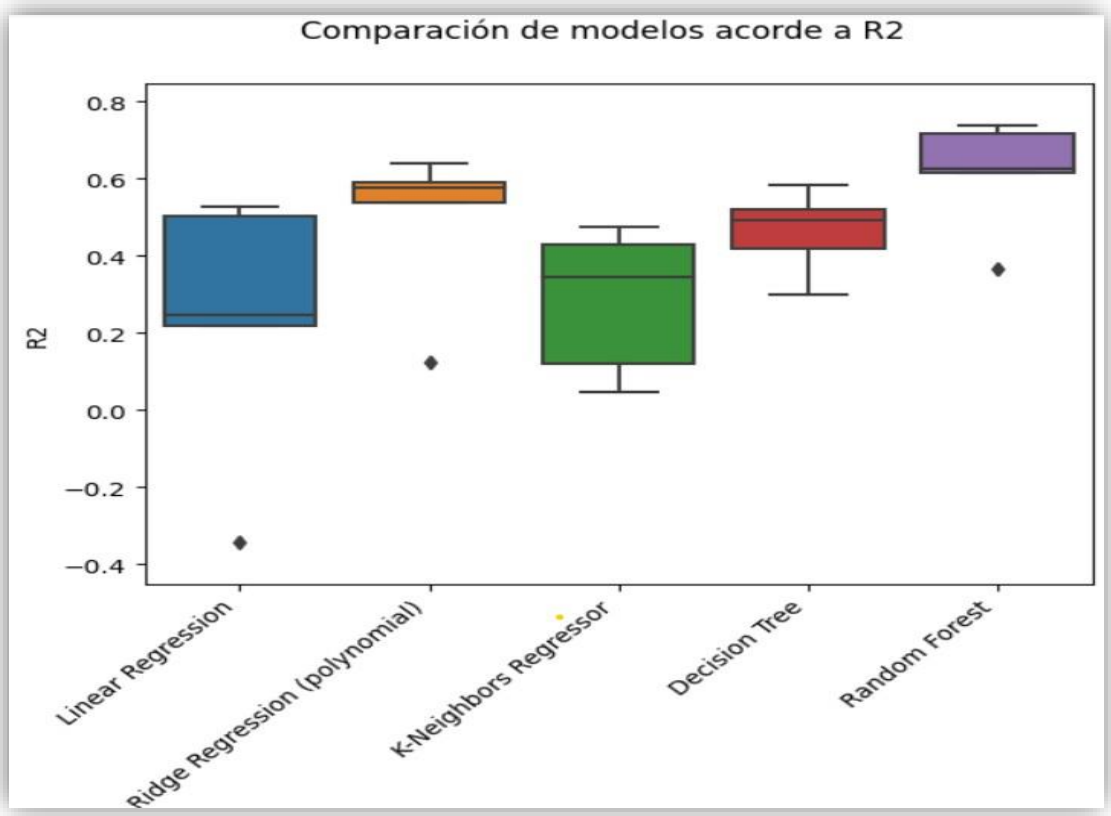
Gráfico .21 Métrica Error Absoluto Medio (MAE)



Fuente: Elaboración propia.

Descripción 8: Comparación de modelos con la métrica MAE, el modelo más efectivo resultó ser Random Forest, con un MAE en validación de 2.563, indicando un error absoluto medio moderado y una sólida capacidad para generalizar a nuevos datos. En contraste, el rendimiento menos favorable lo presenta K-neighbors Regression, evidenciado por un MAE en validación de 3.429, señalando dificultades significativas para adaptar a datos no vistos.

Gráfico .22 Métrica Coeficiente de Determinación (R2)



Fuente: Elaboración propia.

Descripción 9: Comparación de modelos acorde a R2, el modelo más efectivo resulta ser Ridge Regression, con un R2 positivo en validación de 0.494, considerablemente mejor que el R2 negativo de Linear Regression (-2.534). En contraste, el rendimiento menos favorable corresponde a Linear Regression con un R2 negativo. Además, Decision Tree un R2 en validación de 0.464

V. DISCUSIÓN

En el trabajo de (Lap et al. 2023), dividieron el dataset en dos partes: un conjunto de entrenamiento (67%) y otro de prueba (33%). En los experimentos que realizaron, especialmente centrados en modelos basados en árboles, se reveló un sólido desempeño. En particular, el algoritmo RF sobresalió, exhibiendo el mejor rendimiento con valores más bajos de Error Cuadrático Medio (RMSE) a 11,03 y Error Absoluto Medio (MAE) a 6,78. Le siguió el algoritmo DT con un RMSE de 12,60 y MAE de 8,99.

Al comparar estos resultados con el estudio previo, se evidencia que, al igual que en investigaciones anteriores, el modelo RF demostró ser el más eficiente. No obstante, en este estudio, se logró un rendimiento superior con un RMSE más bajo (3.354), indicando predicciones más precisas al minimizar los errores cuadráticos. Además, se observó un MAE más bajo, resaltando la precisión de las predicciones al minimizar las diferencias absolutas entre las predicciones y valores reales.

Mientras (Khoi et al. 2022), llevó a cabo la recopilación de datos bimensuales de calidad del agua a lo largo de un periodo de ocho años, abarcan desde 2010 hasta 2017. Sus datos lo obtuvieron de manera sistemática en cuatro estaciones de control de calidad que ubicaron estratégicamente a lo largo del río La Buong fueron recopilados por el departamento de Recursos Naturales y Medio Ambiente de Dong Nai.

La información de calidad del agua del río La Boung la sometieron a un proceso de partición, dividiéndola en dos conjuntos de 70% destinado al proceso de entrenamiento y el restante 30% para proceso de prueba. Durante el análisis, se identificó que uno de los modelos destacó por su rendimiento superior. Específicamente, el modelo RF sobresalió con un RMSE de 0.121, posicionándose como el más efectivo en términos de rendimiento de predicción. Seguidamente, se ubicó DT con un RMSE de 0,147 y con un R2 de 0.979.

Al realizar una comparación minuciosa de estos resultados, se evidenció que, de manera similar a los hallazgos de esta investigación, el modelo que exhibe un rendimiento óptimo en precisión es el RF, demostrando ser más eficiente. Es relevante destacar que, a diferencia de este estudio, el antecedente logró un RMSE aún más bajo, indicando que su modelo realiza predicciones con una presión adicional.

En concordancia con la investigación previa (Bui et al. 2020), observamos una variación en la recopilación de datos, con una diferencia temporal de 3 años. Utilizaron datos históricos, dividieron su conjunto en dos subconjuntos para el entrenamiento y la prueba del modelo, con una proporción de (70-30). En nuestro caso, optamos por una división diferente, con un (80-20), empleando la técnica de validación cruzada en ambos trabajos.

En nuestro estudio, el RF alcanzó un RMSE de 2.97 y MAE de 1.67. Sin embargo, al comparar los resultados métricos con el trabajo de (Bui et al. 2020), se observan valores más bajos de RMSE y MAE.

Según los hallazgos de (Aldrees et al. 2023), existe una diferencia notable en la recopilación de datos, abarcan un periodo de 30 años desde 1975 hasta 2005. En su estudio, optaron por asignar el 70% de los datos recopilados al conjunto de entrenamiento, mientras que el 15% se destinó al conjunto de prueba y el 15% restante se reservó para el conjunto de validación.

En su estudio (Aldrees et al. 2023)también identificó al modelo RF como uno que demostró un rendimiento superior en la predicción, logrando un RMSE de 13.54 y un MAE de 9.73. Al comparar estos resultados con los obtenidos en este estudio, se observó que, de manera similar a este trabajo, el modelo RF se destaca por su eficiencia en la precisión. Sin embargo, es crucial señalar que en este estudio se logró valores de precisión mayores con un RMSE de 3.354 y un MAE de 2.563.

Estas discrepancias en los valores de las métricas indican que, a pesar de la diferencia temporal en la recopilación de datos, este estudio logro predicciones más precisas en comparación con el trabajo anterior.

En esta investigación, el conjunto de datos de cuenca – Azángaro se dividió en dos subconjuntos, conjunto de entrenamiento (representa un 80%) y el conjunto de prueba (toma el resto 20%). De todos los modelos de ML desarrollados, especialmente el modelo RF tuvo resultados sólidos en términos de métricas, que tienen por objetivo ver el desempeño de los modelos para la predicción de calidad de agua; empezando con RMSE en el conjunto de validación, el modelo RF obtuvo el RMSE más bajo (3.354). Esto significa que la predicción es de RF dando la raíz cuadrada del error cuadrático medio, indicando un ajuste superior a los datos de validación. Siguiendo con MSE en el conjunto de validación, el moldeo RF obtuvo un MSE (12.886), que es el más bajo de todos los modelos evaluados. Esto indica que las predicciones de RF son más cercanas a los valores reales en comparación con los otros modelos. Con respecto a la siguiente métrica R2 en el conjunto de validación, RF tiene un R2 más alto (0.613). En donde un valor de 1 sería un ajuste perfecto, y el valor de (0.6134) obtenido por modelo indica que RF es capaz de explicar el 61.3% de la variabilidad en los datos. Finalmente, en MAE en el conjunto de validación tiene un MAE de (2.563). Esto indica que en promedio las predicciones de RF tienen un error absoluto medio. Un MAE más bajo indica que las predicciones del modelo son en promedio más cercanas a los valores reales. En este caso, el valor de 2.563 para RF es el más bajo en comparación con los otros modelos evaluados, lo que implica que RF tiene un desempeño superior al predecir los valores de la variable objetivo en el conjunto de validación.

VI. CONCLUSIONES

Las conclusiones establecen una conexión precisa con los objetivos específicos delineados en el marco del trabajo de investigación. Para una descripción más detallada, cada modelo se analizó de manera individual, destacando de manera específica como cada uno responde de manera eficiente a los objetivos predefinidos.

La regresión lineal presenta un rendimiento general sólido, demostrando capacidad para generalizar datos no vistos. Sin embargo, se observó una necesidad de mejorar la robustez del modelo, especialmente en el conjunto de validación, indicada por un R^2 más bajo de 0.613. Aunque el MSE y RMSE son aceptables, la atención se centra en la optimización para mejorar la capacidad predictiva.

Ridge Regression presenta un rendimiento mejorado en comparación con la regresión lineal, evidenciando una capacidad más robusta para la predicción. Los valores de MSE 12.886, RMSE, MAE, R2 0.613 indican una mejora significativa en la capacidad de generalización en ambos conjuntos. Este modelo destaca como una opción prometedora para la predicción de la calidad del agua, mostrando un equilibrio entre ajuste y generalización.

El modelo K-Neigbords Regression muestra un rendimiento aceptable, aunque con un mayor error en comparación con la de Ridge Regression. Aunque el R^2 índica una explicación moderada de la variabilidad en los datos.

El modelo Decision Tree logró un ajuste perfecto en el conjunto de entrenamiento, pero presenta una menor capacidad de generalización en el conjunto de validación. Aunque el modelo se adapta excepcionalmente a los datos de entrenamiento, se evidenció la necesidad de mitigar el sobreajuste para mejorar la utilidad del modelo en la predicción de datos no vistos.

El modelo Random Forest destacó como un rendimiento sólido, mostró una buena capacidad predictiva en ambos conjuntos. En el conjunto de validación, este modelo logró un RMSE más bajo 3.354, un MSE mínimo 12.886, un R2 más alto 0.613, indicando una capacidad del 61.3% para

explicar la variabilidad, y un MAE inferior 2.563. Con dichos valores bajos de MSE, RMSE, MAE y un alto $R^{2,}$ este modelo demuestra ser robusto y eficaz para prever la calidad del agua.

VII. RECOMENDACIONES

Basándonos en los resultados obtenidos de los modelos evaluados para la predicción de la calidad del agua en la cuenca de Azángaro, se proponen las siguientes recomendaciones:

- Considerando los resultados, se sugiere utilizar el modelo Bosques Aleatorios (Random Forest) para las predicciones de calidad del agua. Este modelo demostró un rendimiento sólido en todas las métricas evaluadas en el conjunto de validación, indicando una capacidad superior para generalizar a nuevos datos.
- Dada la posible evolución de las condiciones en la cuenca de Azángaro, se propone establecer un plan de actualización periódica del modelo. Incorporar nuevos datos y ajustar el modelo según sea necesario garantizará su relevancia y precisión a lo largo del tiempo.
- Debido a la restricción identificada en la cantidad de datos recolectados, se recomienda ampliar de manera significativa los conjuntos de datos. La adquisición de datos adicionales, que incluyan diversas condiciones y abarquen varios periodos temporales, posibilitará una representación más completa de las complejidades relacionadas con la calidad del agua en la cuenca de Azángaro.

REFERENCIAS

1. ABUODEH, O.R., ABDALLA, J.A. y HAWILEH, R.A., 2020. Prediction of shear strength and behavior of RC beams strengthened with externally bonded FRP sheets using machine learning techniques. *Composite Structures*, vol. 234, ISSN 0263-8223. DOI 10.1016/J.COMPSTRUCT.2019.111698.
2. AHMED, U., MUMTAZ, R., ANWAR, H., SHAH, A.A., IRFAN, R. y GARCÍA-NIETO, J., 2019. Efficient Water Quality Prediction Using Supervised Machine Learning. *Water 2019, Vol. 11, Page 2210* [en línea], vol. 11, no. 11, [consulta: 19 julio 2023]. ISSN 2073-4441. DOI 10.3390/W11112210. Disponible en: https://www.mdpi.com/2073-4441/11/11/2210/htm.
3. ALDREES, A., JAVED, M.F., BAKHEIT TAHA, A.T., MUSTAFA MOHAMED, A., JASIŃSKI, M. y GONO, M., 2023a. Evolutionary and ensemble machine learning predictive models for evaluation of water quality. *Journal of Hydrology: Regional Studies*, vol. 46, ISSN 2214-5818. DOI 10.1016/J.EJRH.2023.101331.
4. ALDREES, A., JAVED, M.F., BAKHEIT TAHA, A.T., MUSTAFA MOHAMED, A., JASIŃSKI, M. y GONO, M., 2023b. Evolutionary and ensemble machine learning predictive models for evaluation of water quality. *Journal of Hydrology: Regional Studies*, vol. 46, ISSN 2214-5818. DOI 10.1016/J.EJRH.2023.101331.
5. ARIAS-GÓMEZ, J., ÁNGEL VILLASÍS-KEEVER, M. y GUADALUPE MIRANDA-NOVALES, M., 2016. El protocolo de investigación III: la población de estudio. [en línea], [consulta: 20 julio 2023]. Disponible en: www.nietoeditores.com.mx.
6. AUTORIDAD NACIONAL DEL AGUA. DIRECCIÓN DE CALIDAD Y EVALUACIÓN DE RECURSOS HÍDRICOS, 2018. Metodología para la determinación del índice de calidad de agua Ica-PE, aplicado a los cuerpos de agua continentales superficiales. *Autoridad Nacional del Agua* [en línea], [consulta: 2 julio 2023]. Disponible en: https://repositorio.ana.gob.pe/handle/20.500.12543/2440.
7. BROWN, R.M., et al., 1970. A water quality index-do we dare. Water and sewage works. , vol. 117,
8. BUI, D.T., KHOSRAVI, K., TIEFENBACHER, J., NGUYEN, H. y KAZAKIS, N., 2020a. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of The Total Environment*, vol. 721, ISSN 0048-9697. DOI 10.1016/J.SCITOTENV.2020.137612.

9. BUI, D.T., KHOSRAVI, K., TIEFENBACHER, J., NGUYEN, H. y KAZAKIS, N., 2020b. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of The Total Environment*, vol. 721, ISSN 0048-9697. DOI 10.1016/J.SCITOTENV.2020.137612.
10. CARPIO, J., BACLIMER, F. y YANAPA, Q., 2021. Hidrogeología de la cuenca del río Azángaro (019), región Puno - [Boletín H 13]. [en línea]. S.l.: Disponible en: https://repositorio.ingemmet.gob.pe.
11. CHEN, C., LIAO, Z., JU, Y., HE, C., YU, K. y WAN, S., 2022. Hierarchical Domain-Based Multicontroller Deployment Strategy in SDN-Enabled Space-Air-Ground Integrated Network. *IEEE Transactions on Aerospace and Electronic Systems*, vol. 58, no. 6, ISSN 15579603. DOI 10.1109/TAES.2022.3199191.
12. CHOU, J.S., HO, C.C. y HOANG, H.S., 2018. Determining quality of water in reservoir using machine learning. *Ecological Informatics*, vol. 44, ISSN 1574-9541. DOI 10.1016/J.ECOINF.2018.01.005.
13. COLLADO R. HERNANDEZ SAMPIERI y P. LUCIO L., 1997. Diseños experimentales de investigación: preexperimentos, experimentos “verdaderos” y cuasiexperimentos. . S.l.:
14. CUTILLAS, P.P., ÁLVAREZ, J.P.A., ORTEGA, E.F.S., GARCÍA, C.C. y CABAÑERO, J.J.A., 2019. La degradación ambiental y sus efectos en la contaminación de las aguas superficiales en la cuenca del río Conchos (Chihuahua - México). *Cuadernos Geográficos* [en línea], vol. 58, no. 1, [consulta: 18 febrero 2024]. ISSN 2340-0129. DOI 10.30827/CUADGEO.V58I1.6636. Disponible en: https://revistaseug.ugr.es/index.php/cuadgeo/article/view/6636.
15. DEO, R.C., 2015. Machine Learning in Medicine. *Circulation* [en línea], vol. 132, no. 20, [consulta: 19 julio 2023]. ISSN 15244539. DOI 10.1161/CIRCULATIONAHA.115.001593. Disponible en: https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.115.001593.
16. DERRAC, J., CHICLANA, F., GARCÍA, S. y HERRERA, F., 2016. Evolutionary fuzzy k-nearest neighbors algorithm using interval-valued fuzzy sets. *Information Sciences*, vol. 329, ISSN 0020-0255. DOI 10.1016/J.INS.2015.09.007.
17. FABIO NELLI, 2018. Python Data Analytics With Pandas, NumPy, and Matplotlib-Second Edition-. [en línea], [consulta: 20 julio 2023]. DOI 10.1007/978-1-4842-3913-1. Disponible en: https://doi.org/10.1007/978-1-4842-3913-1.

18. FABRIS, F., MAGALHÃES, J.P. de y FREITAS, A.A., 2017. A review of supervised machine learning applied to ageing research. *Biogerontology* [en línea], vol. 18, no. 2, [consulta: 19 julio 2023]. ISSN 15736768. DOI 10.1007/S10522-017-9683-Y/TABLES/1. Disponible en: https://link.springer.com/article/10.1007/s10522-017-9683-y.
19. FACULTAD DE CIENCIAS AMBIENTALES DE LA UNIVERSIDAD CIENTÍFICA DEL SUR, L.P., 2020. South Sustainability. *South Sustainability*, ISSN 27087077. DOI 10.21142/SS.
20. FERNANDO, D., MADARIAGA, C., GONZÁLEZ, J.L., MILLER, R., LOZANO, R. y CÁRDENAS VALLEJO, E., 2013. Inferencia estadística Módulo de regresión lineal simple. *Universidad del Rosario* [en línea], [consulta: 3 diciembre 2023]. Disponible en: http://editorial.urosario.edu.co.
21. GAJOWNICZEK, K., GRZEGORCZYK, I. y ZĄBKOWSKI, T., 2019. Reducing False Arrhythmia Alarms Using Different Methods of Probability and Class Assignment in Random Forest Learning Methods. *Sensors 2019, Vol. 19, Page 1588* [en línea], vol. 19, no. 7, [consulta: 19 julio 2023]. ISSN 1424-8220. DOI 10.3390/S19071588. Disponible en: https://www.mdpi.com/1424-8220/19/7/1588/htm.
22. GUPTA, N., NAFEES, S.M., JAIN, M.K. y KALPANA, S., 2011. Physico-chemical assessment of water quality of river Chambal in Kota city area of Rajasthan state (India). *Rasayan Journal of Chemistry*, vol. 4, no. 3, ISSN 09741496.
23. HAGHIABI, A.H., NASROLAHI, A.H. y PARSAIE, A., 2018. Water quality prediction using machine learning methods. *Water Quality Research Journal* [en línea], vol. 53, no. 1, [consulta: 19 julio 2023]. ISSN 1201-3080. DOI 10.2166/WQRJ.2018.025. Disponible en: http://iwaponline.com/wqrj/article-pdf/53/1/3/224144/wqrjc0530003.pdf.
24. HAMEED, A., KHAN, P., QURASHI, M.M. y HAYEE, M.I., 2007. Commission on Science and Technology for Sustainable Development in the South BASIC OR APPLIED RESEARCH Dilemma of Developing Countries. ,
25. HARRIS ET AL., 2020. Array programming with NumPy. *Nature 2020 585:7825* [en línea], vol. 585, no. 7825, [consulta: 20 julio 2023]. ISSN 1476-4687. DOI 10.1038/s41586-020-2649-2. Disponible en: https://www.nature.com/articles/s41586-020-2649-2.

26. HU, J., PENG, H., WANG, J. y YU, W., 2020. kNN-P: A kNN classifier optimized by P systems. *Theoretical Computer Science*, vol. 817, ISSN 0304-3975. DOI 10.1016/J.TCS.2020.01.001.
27. JAIN, N., YEVATIKAR, R. y RAXAMWAR, T.S., 2022. Comparative study of physico-chemical parameters and water quality index of river. *Materials Today: Proceedings*, vol. 60, ISSN 2214-7853. DOI 10.1016/J.MATPR.2021.09.508.
28. JANKOWSKA, A., EJSMONT, A., GALARDA, A. y GOSCIANSKA, J., 2022. The outcome of human exposure to environmental contaminants. Importance of water and air purification processes. *Sustainable Materials for Sensing and Remediation of Noxious Pollutants* [en línea], [consulta: 1 julio 2023]. DOI 10.1016/B978-0-323-99425-5.00003-7. Disponible en: https://www.sciencedirect.com/science/article/pii/B9780323994255000037.
29. JEYASHANTHI, J., BARSANA BANU, J., PANDI MAHARAJAN, M. y RAMUVEL, M., 2023. Assessment of physical and chemical water quality parameters using naive bayes control algorithm. *Materials Today: Proceedings*, vol. 80, ISSN 2214-7853. DOI 10.1016/J.MATPR.2022.11.319.
30. KHOI, D.N., QUAN, N.T., LINH, D.Q., NHI, P.T.T. y THUY, N.T.D., 2022a. Using Machine Learning Models for Predicting the Water Quality Index in the La Buong River, Vietnam. *Water 2022, Vol. 14, Page 1552* [en línea], vol. 14, no. 10, [consulta: 2 julio 2023]. ISSN 2073-4441. DOI 10.3390/W14101552. Disponible en: https://www.mdpi.com/2073-4441/14/10/1552/htm.
31. KHOI, D.N., QUAN, N.T., LINH, D.Q., NHI, P.T.T. y THUY, N.T.D., 2022b. Using Machine Learning Models for Predicting the Water Quality Index in the La Buong River, Vietnam. *Water 2022, Vol. 14, Page 1552* [en línea], vol. 14, no. 10, [consulta: 6 diciembre 2023]. ISSN 2073-4441. DOI 10.3390/W14101552. Disponible en: https://www.mdpi.com/2073-4441/14/10/1552/htm.
32. LAP, B.Q., PHAN, T.T.H., NGUYEN, H. Du, QUANG, L.X., HANG, P.T., PHI, N.Q., HOANG, V.T., LINH, P.G. y HANG, B.T.T., 2023. Predicting Water Quality Index (WQI) by feature selection and machine learning: A case study of An Kim Hai irrigation system. *Ecological Informatics*, vol. 74, ISSN 1574-9541. DOI 10.1016/J.ECOINF.2023.101991.
33. LEE, J.H., LEE, J.Y., LEE, M.H., LEE, M.Y., KIM, Y.W., HYUNG, J.S., KIM, K.B., CHA, Y.K. y KOO, J.Y., 2022. Development of a short-term water quality prediction

model for urban rivers using real-time water quality data. *Water Supply* [en línea], vol. 22, no. 4, [consulta: 19 julio 2023]. ISSN 1606-9749. DOI 10.2166/WS.2022.038. Disponible en: http://iwaponline.com/ws/article-pdf/22/4/4082/1041167/ws022044082.pdfbyguestGRAPHICALABSTRACT.

34. LIANG, C., LI, H., LEI, M. y DU, Q., 2018. Dongting Lake Water Level Forecast and Its Relationship with the Three Gorges Dam Based on a Long Short-Term Memory Network. *Water 2018, Vol. 10, Page 1389* [en línea], vol. 10, no. 10, [consulta: 19 julio 2023]. ISSN 2073-4441. DOI 10.3390/W10101389. Disponible en: https://www.mdpi.com/2073-4441/10/10/1389/htm.
35. NASIR, N., KANSAL, A., ALSHALTONE, O., BARNEIH, F., SAMEER, M., SHANABLEH, A. y AL-SHAMMA'A, A., 2022. Water quality classification using machine learning algorithms. *Journal of Water Process Engineering* [en línea], vol. 48, [consulta: 1 julio 2023]. ISSN 2214-7144. DOI 10.1016/J.JWPE.2022.102920. Disponible en: https://www.sciencedirect.com/science/article/pii/S2214714422003646.
36. PARK, C., 2007. Redefining the doctorate. ,
37. PATEL, J., SHAH, S., THAKKAR, P. y KOTECHA, K., 2015. Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications*, vol. 42, no. 4, ISSN 0957-4174. DOI 10.1016/J.ESWA.2014.10.031.
38. PEDREGOSA FABIANPEDREGOSA ET AL., 2011. Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. *Journal of Machine Learning Research* [en línea], vol. 12, [consulta: 20 julio 2023]. Disponible en: http://scikit-learn.sourceforge.net.
39. PINHEIRO, J.F., DE ALMEIDA, J.D.S., TEIXEIRA, J.A.M., JUNIOR, G.B., DE PAIVA, A.C., SILVA, A.C. y VERAS, R. de M.S., 2021. Automatic ocular version evaluation in images using random forest. *Expert Systems with Applications*, vol. 176, ISSN 0957-4174. DOI 10.1016/J.ESWA.2021.114847.
40. RAHMAN, J., ARAFIN, P. y MUNTASIR BILLAH, A.H.M., 2023. Machine learning models for predicting concrete beams shear strength externally bonded with FRP. *Structures*, vol. 53, ISSN 2352-0124. DOI 10.1016/J.ISTRUC.2023.04.069.
41. REYES, D.M.A., SOUZA, L.C., DE SOUZA, R.M.C.R. y DE OLIVEIRA, A.L.I., 2024. Parametrized linear regression for boxplot-multivalued data applied to the Brazilian

Electric Sector. *Information Sciences*, vol. 652, ISSN 0020-0255. DOI 10.1016/J.INS.2023.119758.

42. RODRÍGUEZ FERNÁNDEZ MORAIMA y LACABA GUARDADO RAFAEL MIGUEL, 2021. Evaluación del Índice de Calidad del Agua (ICAsup) en el río Cabaña, Moa-Cuba. [en línea]. [consulta: 19 julio 2023]. Disponible en: http://scielo.sld.cu/scielo.php?pid=S1993-80122021000100105&script=sci_arttext.
43. SAXENA, A., PRASAD, M., GUPTA, A., BHARILL, N., PATEL, O.P., TIWARI, A., ER, M.J., DING, W. y LIN, C.T., 2017. A review of clustering techniques and developments. *Neurocomputing*, vol. 267, ISSN 0925-2312. DOI 10.1016/J.NEUCOM.2017.06.053.
44. SAYED, Y.A.K., IBRAHIM, A.A., TAMRAZYAN, A.G. y FAHMY, M.F.M., 2023. Machine-learning-based models versus design-oriented models for predicting the axial compressive load of FRP-confined rectangular RC columns. *Engineering Structures*, vol. 285, ISSN 0141-0296. DOI 10.1016/J.ENGSTRUCT.2023.116030.
45. SHAILAJA, K., SEETHARAMULU, B. y JABBAR, M.A., 2018. Machine Learning in Healthcare: A Review. *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, DOI 10.1109/ICECA.2018.8474918.
46. SINGHA, S., PASUPULETI, S., SINGHA, S.S., SINGH, R. y KUMAR, S., 2021. Prediction of groundwater quality using efficient machine learning technique. *Chemosphere* [en línea], vol. 276, [consulta: 2 julio 2023]. ISSN 0045-6535. DOI 10.1016/J.CHEMOSPHERE.2021.130265. Disponible en: https://www.sciencedirect.com/science/article/pii/S0045653521007347.
47. SUN, A.Y. y SCANLON, B.R., 2019. How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. *Environmental Research Letters* [en línea], vol. 14, no. 7, ISSN 1748-9326. DOI 10.1088/1748-9326/ab1b7d. Disponible en: https://iopscience.iop.org/article/10.1088/1748-9326/ab1b7d.
48. TRUIJENS, F.L., CORNELIS, S., DESMET, M., DE SMET, M.M. y MEGANCK, R., 2019. Validity beyond measurement: Why psychometric validity is insufficient for valid psychotherapy research. *Frontiers in Psychology* [en línea], vol. 10, no. MAR, [consulta: 19 julio 2023]. ISSN 16641078. DOI 10.3389/FPSYG.2019.00532. Disponible en: https://www.verywellmind.com/what-is-applied-research-2794820.

49. TYAGI, P., BUDDHI, D., SAWHNEY, R.L. y KOTHARI, R., 2003. A correlation among physico-chemical parameters of ground water in and around Pithampur industrial area. *Indian Journal of Environmental Protection*, vol. 23, no. 11, ISSN 02537141.
50. UDDIN, M.G., NASH, S., MAHAMMAD DIGANTA, M.T., RAHMAN, A. y OLBERT, A.I., 2022. Robust machine learning algorithms for predicting coastal water quality index. *Journal of Environmental Management*, vol. 321, ISSN 0301-4797. DOI 10.1016/J.JENVMAN.2022.115923.
51. UDDIN, M.G., NASH, S., RAHMAN, A. y OLBERT, A.I., 2023. A sophisticated model for rating water quality. *Science of The Total Environment*, vol. 868, ISSN 0048-9697. DOI 10.1016/J.SCITOTENV.2023.161614.
52. VILLENA CHÁVEZ, J.A., 2018. Calidad del agua y desarrollo sostenible. *Revista Peruana de Medicina Experimental y Salud Pública*, vol. 35, no. 2, ISSN 1726-4634. DOI 10.17843/RPMESP.2018.352.3719.
53. VIRTANEN ET AL., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods 2020 17:3* [en línea], vol. 17, no. 3, [consulta: 20 julio 2023]. ISSN 1548-7105. DOI 10.1038/s41592-019-0686-2. Disponible en: https://www.nature.com/articles/s41592-019-0686-2.
54. WAWRZYNIAK, M.K., MATAS SERRATO, L.A. y BLANCHOUD, S., 2021. Long-term monitoring data logs of a recirculating artificial seawater based colonial ascidian aquaculture. *Data in Brief*, vol. 38, ISSN 2352-3409. DOI 10.1016/J.DIB.2021.107372.
55. WILDERER, P., 2011. The Importance of Water Science in a World of Rapid Change: A Preface to the Treatise on Water Science. *Treatise on Water Science* [en línea], vol. 1, [consulta: 1 julio 2023]. DOI 10.1016/B978-0-444-53199-5.09003-5. Disponible en: https://www.sciencedirect.com/science/article/pii/B9780444531995090035.
56. WITTEN, I.H., FRANK, E. y GELLER, J., 2002. Data mining. *ACM SIGMOD Record* [en línea], vol. 31, no. 1, [consulta: 19 julio 2023]. ISSN 01635808. DOI 10.1145/507338.507355. Disponible en: https://dl.acm.org/doi/10.1145/507338.507355.
57. WU, X., KUMAR, V., ROSS, Q.J., GHOSH, J., YANG, Q., MOTODA, H., MCLACHLAN, G.J., NG, A., LIU, B., YU, P.S., ZHOU, Z.H., STEINBACH, M.,

HAND, D.J. y STEINBERG, D., 2008. Top 10 algorithms in data mining. *Knowledge and Information Systems* [en línea], vol. 14, no. 1, [consulta: 3 diciembre 2023]. ISSN 02193116. DOI 10.1007/S10115-007-0114-2/METRICS. Disponible en: https://link.springer.com/article/10.1007/s10115-007-0114-2.

58. YAN, T., ZHOU, A. y SHEN, S.L., 2023. Prediction of long-term water quality using machine learning enhanced by Bayesian optimisation. *Environmental Pollution*, vol. 318, ISSN 0269-7491. DOI 10.1016/J.ENVPOL.2022.120870.
59. ZHANG, H. y WANG, M., 2009. Search for the smallest random forest. *Statistics and its interface* [en línea], vol. 2, no. 3, [consulta: 19 julio 2023]. ISSN 19387989. DOI 10.4310/SII.2009.V2.N3.A11. Disponible en: /pmc/articles/PMC2822360/.
60. ZHANG, S., ZHOU, T., SUN, L. y LIU, C., 2019. Kernel Ridge Regression Model Based on Beta-Noise and Its Application in Short-Term Wind Speed Forecasting. *Symmetry 2019, Vol. 11, Page 282* [en línea], vol. 11, no. 2, [consulta: 3 diciembre 2023]. ISSN 2073-8994. DOI 10.3390/SYM11020282. Disponible en: https://www.mdpi.com/2073-8994/11/2/282/htm.
61. ZHENG, Z., DING, H., WENG, Z. y WANG, L., 2023. Research on a multiparameter water quality prediction method based on a hybrid model. *Ecological Informatics*, vol. 76, ISSN 1574-9541. DOI 10.1016/J.ECOINF.2023.102125.
62. ZHOU, Z.-H., 2021. Ensemble Learning. *Machine Learning* [en línea], [consulta: 19 julio 2023]. DOI 10.1007/978-981-15-1967-3_8. Disponible en: https://link.springer.com/chapter/10.1007/978-981-15-1967-3_8.

ANEXOS

DESARROLLO DE UN MODELO DE MACHINE LEARNING PARA LA PREDICCIÓN DE LA CALIDAD DEL AGUA UTILIZANDO DATOS HISTÓRICOS								
PROBLEMA	OBJETIVO	HIPÓTESIS	VARIABLE	DEFINICIÓN CONCEPTUAL	DEFINICIÓN OPERACIONAL	DIMENSIONES	INDICADORES	ESCALA
¿El desarrollo de modelos de machine learning tiene un desempeño óptimo para mejorar la precisión de la calidad del agua utilizando datos históricos?	Desarrollar modelos de Machine Learning para la predicción de la calidad del agua utilizando datos históricos.	El desarrollo de modelos de Machine Learning basado en datos históricos permitirá obtener predicciones precisas y confiables de la calidad de agua	INDEPENDIENTE MACHINE LEARNIG	Machine learning es un modelo utilizado para la predicción para lo cual hace uso de ciertos **algoritmos** subyacentes para deducir relaciones matemáticas a partir de datos de entrenamiento (Kapitanova, Son y Kang 2012)	Modelos Machine Learning son algoritmo de aprendizaje automatico que combinan multiples árboles de decisión para obtener prediciones más precisas.	Modelos Machine Learning	Error cuadrático medio de raíz (MSE)	Unidad de concentración al cuadrado
							Error cuadrático medio de raíz (RMSE)	Unidad de concentración al cuadrado
							Error Absoluto Medio (MAE)	Unidad de concentración al cuadrado
							Coeficiente de Determinación (R2)	Medida adimensional
PROBLEMAS ESPECÍFICOS	OBJETIVOS ESPECÍFICOS	HIPÓTESIS ESPECÍFICAS	DEPENDIENTE					
¿Cuál es el valor promedio del MSE obtenido por los modelos en la predicción de la calidad del agua utilizando datos históricos?	Determinar el valor promedio del MSE obtenido por los modelos en la predicción de la calidad de agua utilizando datos históricos	El valor promedio del MSE obtenido por los modelos en la predicción de la WQ utilizando datos históricos será mínimo, indicando un bajo error cuadrático medio y una alta precisión en las predicciones del modelo.	DEPENDIENTE CALIDAD DEL AGUA	La calidad del agua es un conjunto de características fisicoquímicas, microbiológicas y físicas que influyen directamente en el correcto desarrollo de la biodiversidad. ((Pulido Capurro 2018))	La calidad del agua se define operacionalmente como las propiedades físicas, químicas y biológicas del agua que determina su aptitud para diferentes usos.	Parámetros fisicoquímicos	Temperatura	°C
¿Cuál es el valor promedio del RMSE obtenido por los modelos en la predicción de la calidad del	Determinar el valor promedio del RMSE obtenido por los modelos en la predicción de la	El valor promedio del RMSE obtenido por los modelo en la predicción de la WQ utilizando					PH	Unidad de PH

agua utilizando datos históricos?	calidad de agua utilizando datos históricos	datos históricos será mínimo, indicando una baja raíz del error cuadrático medio y una alta precisión en las predicciones del modelo.					Oxígeno Disuelto	OD % saturación
¿Cuál es el valor promedio del MAE obtenido por el modelo en la predicción de la calidad del agua utilizando datos históricos?	Determinar el valor promedio del MAE obtenido por los modelos en la predicción de la calidad de agua utilizando datos históricos	El valor de MAE obtenido por los modelos en la WQP utilizando datos históricos será mínimo, indicando una baja discrepancia entre las predicciones y los valores reales de calidad de agua.					Demanda Bioquímica de oxígeno (DBO5)	mg/L
							Conductividad	S/m
							Turbidez	(FAU)
							Fosfatos	mg/L
¿Cuál es el valor promedio del R2 obtenido por los modelos en la predicción de la calidad del agua utilizando datos históricos?	Determinar el valor promedio del coeficiente de determinación (R2) obtenido por los modelos en la predicción de la calidad de agua utilizando datos históricos	El valor promedio del R2 obtenido por los modelos en la WQP utilizando datos históricos será cercano a 1, lo que indica un buen ajuste del modelo y una alta capacidad para explicar la variabilidad de lis datos.					Nitratos	mg/L
						Parámetros microbiológicos	Coliformes Fecales	NMP/100ml

ANEXO 2: PROTOTIPO EN FIGMA WEB Y PHONE