Tecnológico Nacional de México Campus Culiacán

SISTEMA DE CLASIFICACIÓN DE LA CALIDAD DEL AGUA PARA ARROYOS MEDIANTE ALGORITMO DE IA

Tópicos de Inteligencia Artificial

Profesor:
Zuriel Dathan Mora Felix

Integrantes:
Aguilar Recio Jesús Octavio
Echeagaray Aceves Astrid Monserrath

Repositorio GitHub:
https://github.com/OctavioAR/TOPICOS-IA.git

Índice

Índice de Imágenes

1. Introducción

La calidad del agua siendo un factor determinante para garantizar idoneidad para su uso especifico como el consumo humano, la agricultura o la industria, se ve afectada por el incremento de la contaminación en los cuerpos de agua como los son ríos, lagos y manantiales. Este daño no es ocasionado por la naturaleza, sino más bien por los mismos seres humanos que necesitamos de ella para sobrevivir y a causa del crecimiento poblacional, industrial y las malas practicas de manejo de residuos, representan claramente una amenaza ante la salud pública, también la estabilidad de los ecosistemas y el desarrollo económico sostenible. Gracias a los métodos tradicionales de monitoreo y análisis de la calidad humano, no tenemos problemas inmediatos en nuestra salud, ya que con estos análisis se puede detectar cualquier anomalía que pueda ser nociva para la salud, no obstante hay una probabilidad de error humano presente, que podría ser mínimo pero existe esa probabilidad de riesgo, también que dichos métodos suelen ser algo lentos, costosos y de un alcance limitado dificultando una respuesta temprana y efectiva ante emergencias.

Para evitar este tipo de problemas con los métodos tradicionales, la inteligencia artificial (IA) fácilmente podría solucionarlo, siendo una herramienta que en la actualidad ayuda a realizar múltiples tareas y no es raro pensar que la IA pueda ayudar a resolver problemas importantes como lo es la calidad del agua. Con la ayuda del aprendizaje automático, a través de técnicas de clasificación supervisada, brinda la capacidad de analizar grandes cantidades de datos históricos y en tiempo real para predecir el estado de contaminación de los cuerpos de agua de forma rápida, económica y escalable. Algoritmos de clasificación como XGBoost, Random Forest y K-Nearest Neighbors nos permiten identificar los patrones complejos necesarios para llevar acabo un análisis de la calidad del agua y poder así clasificarlo como “potable” o “no potable”.

En la presente investigación se dio a la tarea de comprobar si es factible utilizar la IA para resolver dicho problema del agua. La investigación propone utilizando un dataset robusto en el que se incluya la recolección y preprocesamiento de los datos, la implementación o desarrollo de los algoritmos seleccionados y llevar a cabo una validación exhaustiva de su desempeño. Demostrando así que gracias a soluciones tecnológicas accesibles se pueda mitigar los daños al impacto ambiental en relación a los cuerpos de agua.

2. Objetivos

2.1. Objetivo general

Elaborar un sistema de predicción utilizando un modelo de aprendizaje automático para medir calidad del agua en arroyos, utilizando parámetros recopilados de datos históricos.

2.2. Objetivos específicos

- Obtener un dataset que contenga los parámetros necesarios para el análisis.
- Realizar un análisis exploratorio del dataset para comprender mejor los patrones dentro de los datos.
- Entrenar y evaluar diferentes algoritmos de clasificación para poder así clasificar la calidad del agua.
- Seleccionar el modelo con el mejor resultado y proponer su uso en un prototipo de software.

3. Justificación

La calidad del agua es un tema bastante serio, el cual siempre se ha estado combatiendo para reducir los riesgos en la salud de los seres vivos. Conforme pasan los años, es una realidad que incrementa cada vez más la población y la contaminación en el mundo, por consecuencia ocurre una mayor demanda de agua.

En México el balance hídrico (medición de las entradas y salidas de agua) como se muestra en la figura 1 "recibe por precipitación un volumen anual promedio de 1,449 kilómetros cúbicos de agua, de los cuales el 71.5 % regresa a la atmósfera por evaporación. Además del agua de lluvia, se le suman aproximadamente 48 kilómetros cúbicos por importaciones de los ríos de las fronteras norte y sur" [13], comprendiendo así que México cuenta con una reserva de agua medianamente sostenible; sin embargo, esto no implica que se este aprovechando toda de manera eficiente. Por ejemplo, una gran parte se ve afectada a causa de las descargas de aguas residuales municipales, provenientes de las viviendas, edificios públicos y escorrentías, las cuales desembocan en el drenaje y posteriormente en cuerpos de agua como arroyos, afectando directamente en su calidad.

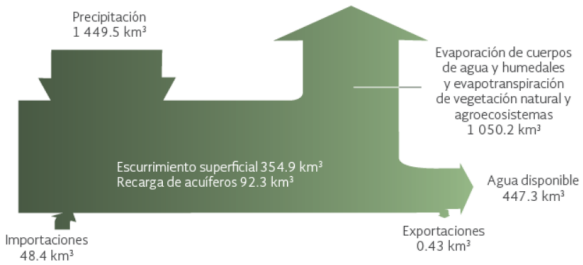

Figura 1: Balance de agua en México

Los principales contaminantes presentes en las aguas residuales son el nitrógeno, fósforo, compuestos orgánicos, bacterias coliformes fecales y materia orgánica, entre otros. A nivel nacional "entre los años 2000 y 2005 el volumen de aguas residuales municipales descargadas aumentó cerca del 6 % (pasando de 250 a 265.6 m3/s). Para 2014, el volumen proveniente de las descargas municipales fue de aproximadamente 7.2 miles de hectómetros cúbicos al año" [13]. A pesar de que se ha observado una disminución de aguas residuales de los municipios entre esos años, no se asegura que en la actualidad sigamos así. Ya que existe el riesgo de que el volumen aumente a causa del incremento de la población y la falta de conciencia sobre el daño que estamos causando al agua potable de nuestro planeta.

En un contexto cercano, en el año 2018 se realizó un análisis, "en los ríos Humaya, Tamazula y Culiacán situados en el valle de Culiacán, en la región noroeste de México, donde se analizaron la prevalencia de 111 cepas de Salmonella

recuperadas de dichos ríos" [4] (Pág. 3), bacteria que puede causar infección intestinal, llamada salmonelosis.

Gracias a la existencia de métodos físicos, químicos y biológicos los expertos pueden analizar la calidad del agua. Por ejemplo, con la espectrometría es posible medir la cantidad de luz que absorbe una muestra de agua para determinar la concentración de diversos analitos y poder evaluar su calidad. De esta manera, se pueden medir parámetros como el pH, oxígeno disuelto, nutrientes, metales pesados y compuestos orgánicos.

Los métodos tradicionales de análisis del agua son manuales, porque ocupan la intervención directa de un técnico, dando como resultado un proceso de análisis lento y costoso, además con un margen de error humano. Ahora bien, en la actualidad el uso de la Inteligencia Artificial (IA) es muy común en muchas de las actividades cotidianas del ser humano, ya que ha ayudado a resolver problemas en un menor tiempo y con una mejor respuesta que el mismo humano. Por lo tanto, el uso de IA para el análisis de la calidad del agua es una excelente opción. Esta permite una predicción en tiempo casi real, permitiendo así, anticipar riesgos de contaminación, ya que es posible detectar parámetros como pH, turbidez, nitratos, etc. Otra mejoría seria la optimización de recursos, reduciendo así, la frecuencia de muestreos físicos altamente costosos.

Es importante entender que la IA llegó para ayudar al ser humano, por eso mismo debemos de utilizarla para mejorar los procesos, métodos y técnicas ya conocidas para obtener mejores resultados.

4. Alcance

Esta investigación se enfoca en analizar cómo es que la Inteligencia Artificial puede contribuir a la solución de problemas de contaminación, enfocándolo en el análisis de la calidad del agua destinada para agricultura, insumos y otros usos. De esta manera, obtener resultados de análisis más certeros, precisos y rápidos en la clasificación de la calidad del agua. Para ello, se estudiarán algoritmos de clasificación de modelos de ML con el fin de identificar los más apropiados para este caso y de así, poder plantear cuál es el mejor para la elaboración de un sistema que incorpore dicho modelo de clasificación del agua.

Primero, es necesario comprender la importancia del uso de la IA en la actualidad y revisar resultados o proyectos de investigaciones similares, para poder contar con un marco de referencia. Posteriormente, buscaremos obtener un dataset público con los parámetros necesarios para el análisis, ya que el proyecto se limitará a la obtención de dichos datos, debido a que no se cuenta con el presupuesto y el apoyo de profesionales en el análisis de la calidad del agua de forma manual, tomando en cuenta que, hacer un dataset propio tomaría demasiado tiempo, y un dataset público con parámetros recopilados hace años, podría reducirlo.

Finalmente, se llevará a cabo el estudio de los distintos algoritmos de clasificación aplicados al análisis de la calidad del agua, y se evaluarán los resultados de su rendimiento considerando métricas que permitan identificar el modelo más adecuado. Y, con base en los resultados, escoger uno para la elaboración de un prototipo de software capaz de clasificar la calidad del agua.

5. Desarrollo

La calidad del agua es un tema al que, en muchas ocasiones, no se le da la atención necesaria, lo que indirectamente contribuye a la contaminación de ese recurso vital. Aunque hay acciones cotidianas en las que el uso del agua es indispensable, como por ejemplo, en el lavado de ropa, la higiene personal o el uso sanitario, en la mayoría de los casos es posible evitar su desperdicio y optimizar su consumo. El problema se intensifica cuando éstas descargas de aguas residuales terminan en los mismos ríos de los que posteriormente extraemos agua para nuestro consumo. Por lo tanto, es fundamental que las empresas e instituciones que se encargan de purificar el agua, cuenten con el equipo y metodología para medir la calidad del agua y tener resultados certeros. En este contexto, el uso de la Inteligencia Artificial (IA) surge como una solucion a este desafío. Se ha demostrado que "el uso de algoritmos de aprendizaje automático, como Random Forest (RF), redes neuronales (NN), regresión lineal múltiple (MLR), máquinas de soporte vectorial (SVM) y árboles potenciados (BTM), permite categorizar de manera eficiente conjuntos de datos relacionados con la calidad del agua, considerando parámetros como oxígeno disuelto (OD), coliformes totales (CT), demanda biológica de oxígeno (DBO), nitratos, pH y conductividad eléctrica (CE)" [6]. Estos modelos, mediante procesos de preprocesamiento, normalización, gestión de datos faltantes y clasificación, han mostrado un alto potencial para predecir el índice de calidad del agua (ICA), convirtiéndose en un indicador clave para su correcta gestión [6]. Además de las metodologías utilizadas para analizar la calidad del agua, es necesario considerar que hay normas y estándares específicos que regulan dicho aspecto en diferentes contexto. Ya que "existen estándares de calidad específicos para indicar la calidad de diferentes fuentes de agua, como aguas subterráneas, manantiales, ríos, lagos y arroyos, y existen criterios específicos de calidad del agua para usos agrícolas, industriales, humanos u otros"[12]. Por ejemplo, El agua destinada al consumo humano debe ser fresca y libre de contaminantes, mientras que el agua utilizada en la agricultura requiere cumplir con criterios específicos de salinidad y toxicidad. Por su parte, los estándares para el uso industrial varían de acuerdo con las características de cada proceso productivo [14].

5.1. Uso de la IA para resolver problemas de CA

El presente proyecto de investigación es en sí, el desarrollo de un modelo de Machine Learning (ML) para una tarea de clasificación. El ML "es una aplicación de la inteligencia artificial (IA) que proporciona a los sistemas la capacidad de aprender y mejorar automáticamente a partir de la experiencia sin estar programados explícitamente" [1] (Pág. 1). Además, en la actualidad es muy común estar en contacto o saber que ciertas tecnologías usan modelos de ML para facilitar tareas y dar mejores resultados, como, por ejemplo: Sistemas de recomendación utilizados por plataformas como Netflix o Amazon, usadas para seguir productos o contenido, otro ejemplo sería los asistentes virtuales, los cuales permiten a dispositivos entender y responder al lenguaje humano,

como Alexa, Google Assistant, etc.

Entonces los modelos de ML de clasificación se definen como "una técnica de aprendizaje supervisado (SL) que implica categorizar datos en clases distintas. Es un proceso recursivo que reconoce y agrupa objetos de datos en categorías o etiquetas predefinidas. Esta técnica se utiliza para predecir el resultado de un problema determinado basándose en las características de entrada" [2] (Pág. 3). Es por ello que la clasificación es ideal para este tipo de proyecto, por ejemplo, si se cuenta con un dataset con los parámetros necesarios para medir la calidad del agua y además se incluyan etiquetas como "potable.º "no potable", entonces puede aplicarse un algoritmo de ML capaz de aprender la relación entre las características (los datos de entrada) y las etiquetas de salida (tipo de agua).

Entre las principales ventajas del uso de modelos de ML en este contexto, está la posibilidad de gererar predicciones rápidas y escalables, aún en escenarios con grandes volúmenes de datos recolectados por sensores en tiempo real "debido a su capacidad para modelar interacciones complejas y no lineales entre variables ambientales [9]". No obstante, una de las preocupaciones principales es la dependencia de la calidad de los datos.

Los modelos de ML tienen muchas aplicaciones, en el contexto de la clasificación de la calidad del agua ya existen investigaciones o trabajos que han demostrado un gran potencial utilizando las técnicas de aprendizaje automático en el campo del monitoreo y la gestión de la calidad del agua. Por ejemplo, un estudio se obtuvieron los siguientes resultados: "El enfoque alcanza una alta Precisión de Entrenamiento del 98 % y una Precisión de Prueba del 94 %, lo que indica su potencial para el monitoreo y la gestión de la calidad del agua en tiempo real. El modelo desarrollado en este estudio puede predecir la calidad del agua como Excelente, Buena, Mala y Muy Mala, lo que permite diversas aplicaciones como el tratamiento de aguas, el monitoreo ambiental y la gestión de la vida acuática" [11] (Pág. 7). Lo cual demuestra que es factible desarrollar sistemas de monitoreo de calidad del agua basados en ML.

Otras investigaciones han profundizado en el uso de modelos de clasificación y regresión basados en el aprendizaje automático para determinar la calidad del agua, esto en escenarios binarios (como agua potable o no potable) o en distintas categorías. Estos modelos utilizan datos que son recolectados en tiempo real a través de sensores, a los cuales se les asignan etiquetas con base a sus características que posteriormente son clasificadas según la relevancia que tengan. A diferencia de los sistemas tradicionales basados en lógica difusa, el desarrollo de la IA permite un análisis cuantitativo más preciso. La comprobación de estos modelos se ha realizado mediante métricas como la exactitud, la precisión, la recuperación y la puntuación confirmando se eficacia para clasificar fuentes de agua, predecir su potabilidad e identificar contaminantes en distintas muestras [10].

5.2. Conjunto de datos (Dataset)

El conjunto de datos constituye la base fundamental para el desarrollo de cualquier modelo de aprendizaje automático, ya que de su calidad, volumen y representatividad dependerá en gran medida el rendimiento del sistema. En el contexto del análisis de la calidad del agua, contar con un dataset confiable permite entrenar, validar y evaluar algoritmos de clasificación capaces de identificar patrones entre los parámetros y relacionarlos con categorías de potabilidad o niveles de calidad. Ya que "los buenos datos son vitales para el éxito de los sistemas de ML"[10], por lo que la eficacia de un modelo de ML no depende únicamente del algoritmo utilizado, sino también de la calidad de los datos, su preparación y limpieza.

5.2.1. Obtención de datos del campo

La recolección de datos para el entrenamiento del modelo es de suma y vital importancia, ya que si le brindamos datos erróneos o incompletos al modelo, no se obtendrán resultados deseados, se obtendrían datos erróneos que podrían poner en riesgo la salud de quienes consuman agua que no fue analizada correctamente. Para poder obtener estos datos de forma segura y confiable es necesario primero definir los parámetros clave de medición, como por ejemplo, los fisicoquímicos como pH, temperatura, oxígeno disuelto, turbidez, temperatura, después están los químicos como nitratos, nitritos, fosfatos y dureza, y, con los biológicos serían coliformes totales y fecales. Además, se necesita identificar las áreas de muestreo como ríos, lagos o cualquier cuerpo de agua que este destinado para purificación y consumo humano. La recolección requiere de equipo físico como botellas estériles, sondas, hieleras y material de conservación. El protocolo de muestreo se llevaría durante meses o incluso años para obtener un buen volumen de datos. Después del protocolo de muestreo se tendría que llevar a analizar a laboratorios, ya que, parámetros como Nitratos y Coliformes no pueden medirse directamente en el campo con dispositivos electrónicos y, es necesario esperar para la obtención de dichos resultados. El siguiente paso sería digitalizarlos, ya sea en un archivo de Excel o CSV, lo que permitirá obtener un dataset con resultados propios y confiables.

5.2.2. Obtención de datos digitales

Existen muchos dataset en internet, de diferentes desarrolladores o científicos, los cuales se dedican a la obtención de datos históricos para poder hacer uso de ellos. Para el contexto del proyecto, se necesita un dataset que contenga los parámetros biológicos, fisicoquímicos y químicos, para poder medir la calidad del agua, como lo podría ser el dataset de "Water Quality" de Kaggle [8] o también datos proporcionados por organismos del Gobierno de México como CONAGUA [3]. Para un proyecto ambicioso lo ideal siempre es obtener la recolección de datos de su lugar de origen, pero el hacer eso restaría tiempo e implicaría más esfuerzos, es por eso que, el utilizar un dataset público permite enfocarse en el objetivo, que es realizar un sistema que integre un modelo de

ML entrenado por medio de algoritmos de IA para clasificar la calidad del agua, ya que en sí, el proyecto es demostrar que con el uso de la IA es posible ayudar a controlar la contaminación del agua por medio de análisis de clasificación y así obtener resultados certeros y rápidos.

5.3. Preprocesamiento de datos

En todo proyecto que se requiera entrenamiento de algún modelo de ML, es importante llevar acabo un preprocesamiento de los datos, dado que no todos están completos. En el contexto del proyecto, puede que algunos datos de los parámetros necesarios para llevar acabo un análisis de calidad del agua no estén completos, como se muestra en la figura 2 del dataset "Water Quality" de kaggle. Esto puede ocasionar problemas al momento de entrenar el modelo con datos vacíos. Algunos pasos que se pueden emplear en el preprocesamiento del dataset es la limpieza de datos haciendo manejo de valores nulos, identificar y eliminar las filas o columnas con valores faltantes o imputarlos usando la media, mediana o moda; otra técnica sería el manejo de valores atípicos, siendo aquellos valores que se desvían significativamente del resto de los datos, también está la codificación de variables categóricas para convertir aquellas variables de texto en formato numérico que los algoritmos puedan procesar y por ultimo se puede emplear una reducción de datos, eliminando aquellas variables innecesarias para el entrenamiento y así evitar problemas al momento de obtener resultados. Existen diversas técnicas o practicas para llevar acabo un preprocesamiento de datos, necesario para eliminar el ruido e inconsistencia y de esta manera poder entrenar correctamente cualquier modelo.

ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
	204.8904554713363	20791.31898074702	7.300211873184757	368.516441349803	564.3086541722439	10.37978308	86.99097046	2.963135380631640	0
3.716080075	129.4229205149442	18630.05785797034	6.635245884		592.8853591348523	15.18001311635725	56.32907628451764	4.500656274942408	0
8.099124189298397	224.2362593935577	19909.54173229239	9.275883602694089		418.6062130644815	16.86863692955097	66.42009251176368	3.055933749664168	0
8.316765884214679	214.3733940856225	22018.41744077529	8.059332377	356.8861356430566	363.2665161642437	18.43652449549330	100.3416743650800	4.628770536837084	0
9.092223456290965	181.1015092361252	17978.98633892625	6.546599974207941	310.1357375242044	398.4108133818446	11.55827944344639	31.99799272742473	4.075075425430034	0
5.584086638456089	188.3133237696164	28748.68773904612	7.544868789	326.6783629116736	280.4679159334877	6.399734640152758	54.91786184199446	2.559708227556521	1
10.22386216452877	248.0717352701399	28749.71654352823	7.513408465831302	393.6633955150964	283.6516335078445	13.78969531751988	84.60355617402357	2.672988736934779	0
8.635848718500734	203.3615225845705	13672.09176390163	4.563008685599703	303.3097711592812	474.6076449424485	12.36381669870525	62.79830896292515	4.401424715445482	0
	118.9885790902518	14285.58385422451	7.804173553073094	268.6469407	389.3755658712614	12.70604896865791	53.92884576751223	3.595017180957615	0
11.18028447072159	227.2314692379745	25484.50849098786	9.077200016914393	404.0416346840899	563.8854814810949	17.92780641128502	71.97660103221915	4.370561936655497	0
7.360640105838258	165.5207972595286	32452.61440914388	7.550700906704114	326.6243534556016	425.3834194953873	15.58681043803312	78.74001566430479	3.662291782852457	1
7.974521648923869	218.6933004886664	18767.65668181348	8.110384501123875		364.0982304620486	14.52574569759320	76.48591117965157	4.011718108339787	0
7.119824384264552	156.7049933403921	18730.81365334271	3.606036090505720	282.3440504739606	347.7150272619437	15.92953590882569	79.50077834	3.445756223321899	1
	150.1749233951362	27331.36196192775	6.838223470687509	299.4157813468584	379.7618348257724	19.37080718123212	76.50999553	4.413974182974902	0
7.496232208	205.3449821581851	28388.00488673697	5.072557773840631		444.6453523327066	13.22831109922452	70.30021264692436	4.777382337225378	0
6.347271760539316	186.7328806605761	41065.23476453935	9.629596276480584	364.4876872467604	516.7432819	11.53978119153941	75.07161728663777	4.376348290691898	0
7.051785800016845	211.0494060605457	30980.60078678886	10.09479601166142	6	315.1412672443021	20.39702184072246	56.65160378979331	4.268428857506186	0
9.181560007151536	273.8138066598009	24041.32628006128	6.904989726470096	398.3505168222779	477.9746418621779	13.38734078022554	71.45736221	4.503660796179122	1
8.975464347533963	279.3571667700923	19460.39813123211	6.204320858892474		431.4439899903489	12.88875905430399	63.82123709666397	2.436085590305273	0
7.371050302429531	214.4966104571565	25630.32003699972	4.432669290372123	335.7544385960652	469.9145514792358	12.50916394049869	62.79727715266126	2.560299147614914	0

Figura 2: Dataset de Water Quality

5.4. Algoritmos de clasificación

5.4.1. Random Forest

El algoritmo de clasificación Random Forest (Bosques Aleatorios) es una "técnica que utiliza el concepto de árboles de decisión, construyendo un conjunto de árboles de decisión y agregando sus resultados para generar la predicción definitiva. Cada árbol de decisión dentro de un bosque aleatorio se construye utilizando subconjuntos aleatorios de datos, y cada árbol individual se entrena con una parte del conjunto de datos. Posteriormente, los resultados de todos los árboles de decisión se combinan para obtener el pronóstico definitivo." [5] (Pág. 1) En esencia el algoritmo se comporta como se muestra en la figura 3 tomando los datos del dataset y generando múltiples arboles de decisión que tendrán un resultado, los cuales se juntarán para obtener un promedio de resultados generados por cada árbol individual, conformando un bosque.

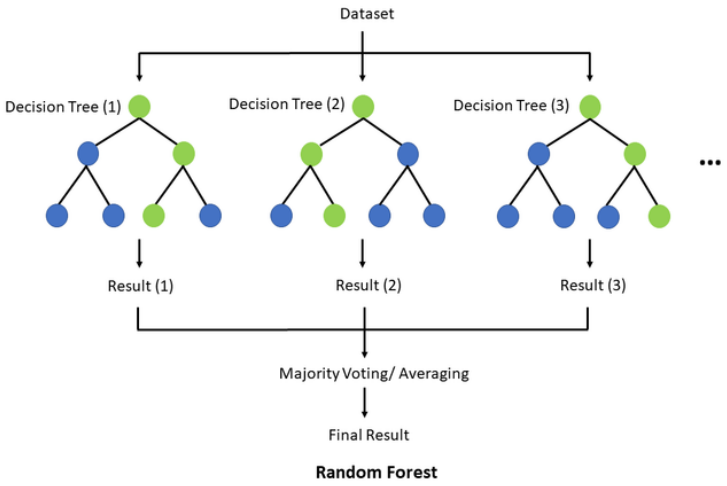

Figura 3: Comportamiento del algoritmo Random Forest

Este algoritmo es adecuado por su alta precisión, ya que ofrece buenos resultados con poco ajuste de hiperparámetros. También es resistente al overfitting gracias a su naturaleza de ensamble, además tiene un buen manejo de datos numéricos y categóricos, adaptándose bien al tipo de datos que manejan los datasets de calidad del agua, estos suelen incluir los parámetros ya antes mencionados (químicos, fisicoquímicos, biológicos) que son datos numéricos y en algunos casos cuentan, con etiquetas categóricas. Para el entrenamiento del algoritmo Random Forest, se recomienda utilizar un 80 % del dataset para entrenamiento y validaciones, dejando un 20 % para las pruebas, con esta división permite que el algoritmo construya un gran número de árboles con suficientes datos.

5.4.2. XGBoost

El algoritmo de clasificación XGBoost (eXtreme Gradient Boosting) "es un algoritmo ML de conjunto que depende de árboles de decisión y emplea un método de refuerzo de gradiente" [15] (Pág. 6). Es necesario comprender que esta implementación avanzada del framework Gradient Boosting, esta diseñada para ser computacionalmente eficiente y ofrecer un rendimiento superior. A diferencia del algoritmo Random Forest, el cual se encarga de construir árboles de forma independiente y los promedia, como se aprecia en la figura 4 XGBoost funciona con un proceso de boosting secuencial, en el que cada árbol de decisión se entrena para corregir los errores residuales cometidos por el conjunto de árboles anteriores.

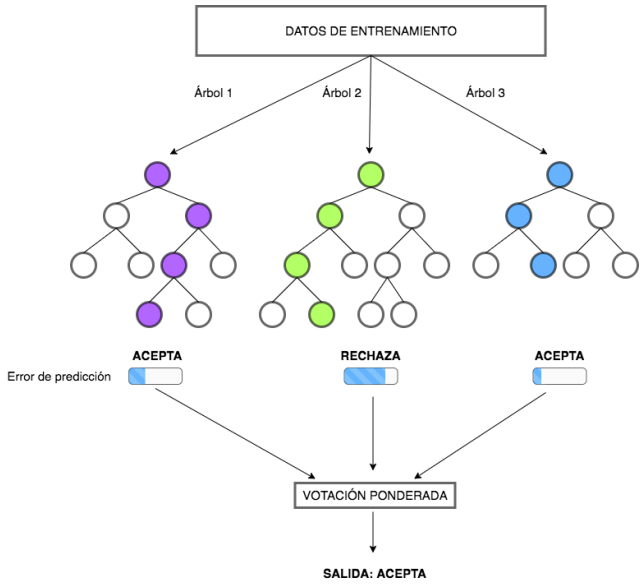

Figura 4: Comportamiento del algoritmo XGBoost

Este algoritmo es ideal para el proyecto gracias a su poder predictivo, el cual es altamente utilizado en competencias de ML, la eficiencia y velocidad de entrenamiento lo hacen manejable incluso en equipos que no requieren un gran poder de cómputo, además que cuenta con la capacidad de manejar valores faltantes y realizar selección de características, lo cual añade robustez al pipeline de preprocesamiento. En el contexto del proyecto, el algoritmo es adecuado para datasets tabulares como los de calidad del agua donde puede capturar relaciones complejas e interacciones entre múltiples parámetros fisicoquímico. Para el entrenamiento del algoritmo se recomienda una división de 80 % de entrenamiento y un 20 % reservado para la evaluación, así garantizando un desempeño superior que se mida de forma justa sobre datos no vistos.

5.4.3. KNN

El algoritmo de clasificación K-Nearest Neighbord (K-Vecinos más cercanos) "es un clasificador de aprendizaje supervisado no paramétrico, que emplea la proximidad para realizar clasificaciones o predicciones sobre la agrupación de un punto de datos individual" [7]. Si bien este algoritmo se puede usar tanto para problemas de regresión o clasificación, generalmente se usa como un algoritmo de clasificación partiendo de que se pueden encontrar puntos similares cerca uno del otro como se muestra en la figura 5. En sí, este algoritmo para clasificar una nueva muestra, calcula la distancia entre ésta y todas las muestras del conjunto de entrenamiento. La clase asignada será la clase mayoritaria entre sus K vecinos más cercanos.

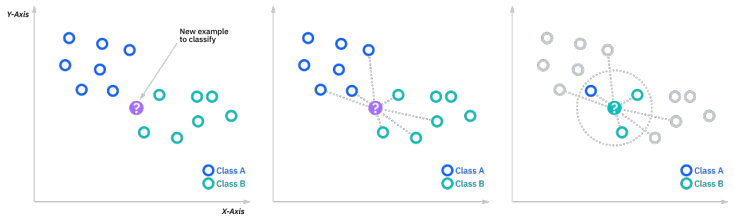

Figura 5: Comportamiento del algoritmo KNN

KNN es una buena opción de algoritmo de clasificación en el contexto del proyecto, ya que gracias a su simplicidad y falta de supuestos sobre la distribución de los datos lo hace muy adaptable al momento de explorar patrones complejos y no lineales en los parámetros de calidad del agua. El funcionamiento intuitivo, por ejemplo: "si se parece a muestras potables, entonces es potable" ayuda a la interpretación de los resultados y la validación de la importancia de las características seleccionadas. Sin embargo, su rendimiento óptimo necesita que todas las variables estén escaladas de forma correcta, es decir que cada variable (columna) del set de datos tenga aproximadamente el mismo rango de valores, pero esto se puede arreglar aplicando un buen preprocesamiento de los datos. Los porcentajes de entrenamiento en el caso de KNN es una división 70-80 % para entrenamiento, debido a que el algoritmo "memoriza" el conjunto de entrenamiento completo para realizar las predicciones, un conjunto de entrenamiento más grande mejora la densidad de la representación de las clases en el espacio de características, aumentando así, la precisión al encontrar vecinos más representativos. Y el 20-30 % se reserva para las pruebas.

5.5. Entrenamiento y evaluación

En esta fase se tiene como objetivo entrenar a los algoritmos de machine learning: Random Forest, KNN y XGBoost con el conjunto de datos ya preprocesado, también optimizar sus hiperparámetros y así evaluar de manera efectiva el rendimiento de los modelos y poder escoger el más optimo. Hoy en día, el lenguaje de preferencia para trabajos de inteligencia artificial es Python, utilizando las bibliotecas de código abierto scikit-learn, pandas, numpy, entre otros, los cuales sirven para la manipulación de datos y implementación de los algoritmos.

5.5.1. División de los datos

El conjunto de datos preprocesados se divide en dos subconjuntos independientes utilizando la función train-test-split de scikit-learn. En el conjunto de entrenamiento un 80 % será utilizado para el ajuste de los parámetros de los modelos y la validación cruzada. Para el conjunto de pruebas con un 20 % reservado exclusivamente para la evaluación final e imparcial del modelo, simulando así, datos nunca vistos. Los tres modelos serán entrenados con los mismos porcentajes de entrenamiento y prueba para ser sometidos a una evaluación justa.

5.5.2. Evaluación del desempeño

El desempeño de los modelos se evalúa con estándares en problemas de clasificación binaria, las cuales son calculadas a partir de los resultados obtenidos en el conjunto de pruebas: Accuracy es la proporción de predicciones correctas sobre el total. Útil para una visión general, pero puede ser engañosa si las clases están desbalanceadas. Precisión es la capacidad del modelo de minimizar falsos positivos, como por ejemplo, evitar etiquetar agua contaminada como potable. Recall es la capacidad del modelo de encontrar todos los positivos reales, por ejemplo, detectar toda el agua que está realmente contaminada. F1-Score es la media armónica entre precisión y sensibilidad, es la métrica principal para evaluar el balance del modelo cuando las clases están desbalanceadas. Y por ultimo la matriz de confusión en la cual se pueden visualizar los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos, permitiendo un análisis detallado de los tipos de error. Después del proceso de evaluación del desempeño de los modelos, se tiene que hacer una comparación de las métricas obtenidas por cada algoritmo en el conjunto de prueba. No se debe de elegir únicamente aquel modelo con una mayor precisión, si no más bien por el que tenga mejor equilibrio según su puntaje F1, porque este puede penalizar aquellos modelos que tienen un desempeño pobre en algunas de las clases, lo que es crucial para un sistema de seguridad como la calidad del agua.

5.6. Diseño de la solución propuesta

La implementación del sistema para detectar la calidad de agua es prácticamente una etapa futura de este proyecto, ya que en esta primera etapa se dió a la tarea de demostrar que con el uso de IA se puede resolver un problema como lo es la calidad del agua, utilizando modelos de ML entrenados con algoritmos de clasificación. Encontrando que existen distintos algoritmos, unos mejores que otros, pero con los cuales es factible desarrollar dicho modelo. Viendo que es posible desarrollar la primera etapa, en la segunda sería desarrollar el sistema o software en sí para la detección de la calidad del agua en tiempo real, utilizando claramente el mejor modelo de ML después de obtener los resultados del entrenamiento y pruebas. El sistema se puede emplear mediante una arquitectura de tres capas (presentación, negocio, datos) siendo esto una breve explicación de que se puede emplear en cada una: Para la obtención de los datos en tiempo real se pueden emplear sensores IoT para la medición de los parámetros como pH, turbidez, conductividad, incluso se pueden obtener muestras un poco más complejas como lo son metales pesados y coliformes. El almacenamiento siempre es la clave de todo sistema, es por eso que una buena opción sería una BD en la nube, lo que permitiría el manejo de grandes volúmenes de información y la realización de consultas eficientes. La cuestión del preprocesamiento se puede implementar mediante scripts automatizados (con Pandas) para la limpieza, imputación de valores faltantes y estandarización de los datos. El modelo escogido en la primera será serializado y empaquetado para su despliegue en un entorno de producción. Para la visualización, existen varias opciones como: una aplicación web con PowerBi, una aplicación móvil, entre otras. La más viable es una aplicación web para obtener los resultados de predicción en tiempo casi-real, tendencias históricas de la calidad del agua, etc. Con PowerBi poder generar dashboard para mostrar de mejor forma los resultados o datos. También mediante un sistema de alertas como notificaciones push, SMS o email para alertar cuando el modelo prediga un estado "No Potable" o cuando los parámetros sean peligrosos para el consumo. A grandes rasgos podemos decir que el flujo de operación del sistema sería:

- Ingesta: Los datos obtenidos de los sensores IoT y laboratorios se ingieren y almacenan en la BD.
- Preprocesamiento: Con los scripts de preprocesamiento se ejecutarán de forma automática para dejar los datos listos para el modelo.
- Predicción: El servicio API recibe las mediciones preprocesadas, las pasa al modelo entrenado y retorna la clasificación.
- Visualización: Los resultados se almacenarán y mostrarán en un dashboard. En caso de un parámetro peligroso se emitirá una alerta.

6. Agenda

El proyecto de investigación consto de 18 días para su elaboración, como se puede apreciar en la figura 6 la cual representa como es que dividieron los días para poder realizar dicho proyecto, siendo los cuadros de color verde los días esperados para su elaboración y rojos los días en que se realizó.

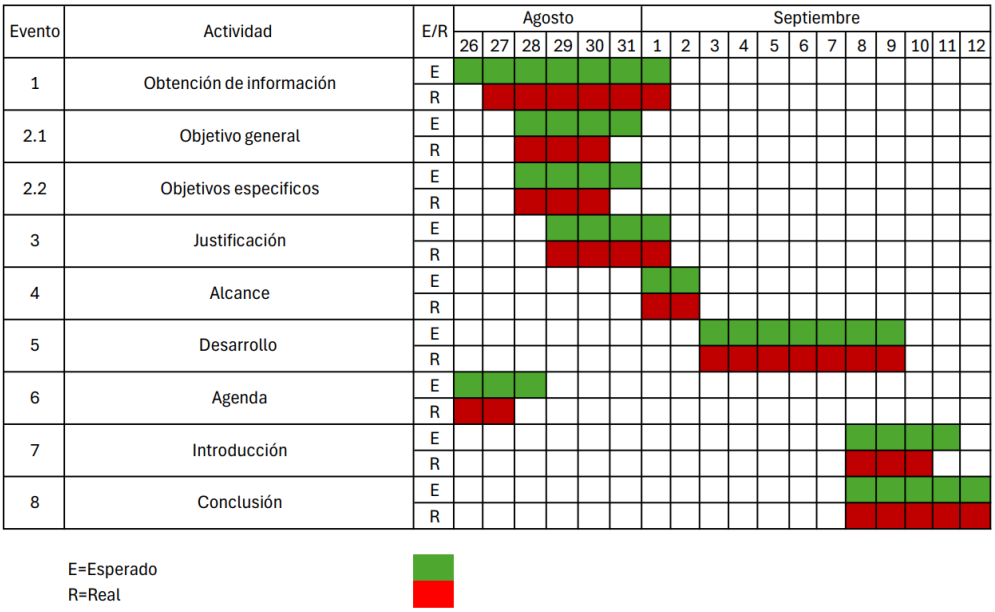

Figura 6: Cronograma de actividades

7. Conclusión

El uso de la inteligencia artificial es algo que se ha vuelto cotidiano en la vida de todos los seres humanos, ya que estas nos ayudan a realizar tareas en las cuales podríamos demorar horas en completarlas, además la IA nos ayuda ha automatizar varios procesos y hacernos más productivos y eficientes. Es por eso por lo que utilizar estas herramientas en favor del medio ambiente no es algo descabellado o imposible de hacer, ya que se pueden emplear en varias áreas que se necesita de atención, como bien lo es la calidad del agua, ya que año tras año hay incremento de población y eso ocasiona que haya más desperdicio de agua que ocasionan aguas residuales, las cuales demandan soluciones eficientes.

Se demuestra que con la ayuda de modelos de machine learning, específicamente los algoritmos de clasificación se pueden realizar análisis y predicción de la calidad del agua con una muy buena precisión. Demostrando así que implementar un sistema capas de analizar la calidad del agua, ayudaría a cientos de personas, comunidades, ciudades que necesitan mejorar la obtención de resultados para obtener de forma más rápida los procesos de purificación de agua, ofreciendo resultados rápidos, pero a la vez confiables, evitando cada vez más el uso de métodos tradicionales que requieren tiempo y recursos.

La contaminación del agua y en general, es una problemática que no debemos de minimizar ya que esta provoca cambios desfavorables al planeta con el paso del tiempo, además que es muy fácil subestimarla argumentando como es que el impacto individual es irrelevante, pero que pasa si millones de personas piensan lo mismo.

Referencias

[1] Azmir Alam. «What is Machine Learning?» En: *University of Dhaka* (2023), págs. 1-7. DOI: 10.5281/zenodo.8231580.

[2] Tasnim H.K. Albaldawi Amer F.A.H. Alnuaimi. «An overview of machine learning classification techniques». En: *Fifth International Scientific Conference of Alkafeel University* (2024), págs. 1-24. DOI: 10.1051/bioconf/20249700133.

[3] CONAGUA. *Indicadores de Calidad del Agua*. https://www.gob.mx/conagua/articulos/indicadores-de-calidad-del-agua. Último acceso: 04 de septiembre de 2025. 2025.

[4] Maribel Jeménez Gloria Castañeda. «EVALUACIÓN DE RÍOS DEL VALLE DE CULIACÁN, MÉXICO, COMO RESERVORIOS DE SEROTIPOS DE Salmonella RESISTENTES A ANTIBIÓTICOS». En: *Rev. Int. Contam. Ambie* (2018), págs. 1-11. DOI: 10.20937/RICA.2018.34.02.01.

[5] Ali Kalakech Hasan Ahmed Salman. «Random Forest Algorithm Overview». En: *Babylonian Journal of Machine Learning* (2024), págs. 1-11. DOI: 10.58496/BJML/2024/007.

[6] Akter L. Hassan M.M. «Efficient Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms». En: (2021), págs. 1-12. DOI: 10.2991/hcis.k.211203.001.

[7] IBM. *¿Qué es el algoritmo de k vecinos más cercanos (KNN)?* https://goo.su/k3WW7. Último acceso: 06 de septiembre de 2025. 2025.

[8] Aditya Kadiwal. *Water Quality, Drinking water potability*. https://www.kaggle.com/datasets/adityakadiwal/water-potability. Último acceso: 04 de septiembre de 2025. 2021.

[9] WZW Aziz NAA Lokman A. Ismail. «A Review of Water Quality Forecasting and Classification Using Machine Learning Models and Statistical Analysis». En: *Water 2025* (2025). DOI: https://doi.org/10.3390/w17152243.

[10] Gangadevi E. Shri M.L. Nallakaruppan M.K. «Reliable water quality prediction and parametric analysis using explainable AI models.» En: *Enfoque UTE* (2024). DOI: https://doi.org/10.1038/s41598-024-56775-y.

[11] Mr. R. Ambikapathy S Nandakumari. «WATER QUALITY CLASSIFICATION USING MACHINE LEARNING». En: *International Research Journal of Modernization in Engineering Technology and Science* (2024), págs. 1-8. DOI: 10.56726/IRJMETS60790.

[12] Yousef S. Abuzir Saleh Y. Abuzir. «Machine learning for water quality classification». En: *Water Quality Research Journal* (2022). DOI: 10.2166/wqrj.2022.004.

[13] SEMARNAT. *AGUA*. https://apps1.semarnat.gob.mx:8443/dgeia/informe15/tema/cap6.html. Último acceso: 31 de agosto de 2025. 2014.

[14] Henry J. Vaux William A. Jury. «The Emerging Global Water Crisis: Managing Scarcity and Conflict Between Water Users». En: (2007), págs. 1-76. DOI: https://doi.org/10.1016/S0065-2113(07)95001-4.

[15] Ziyad H. Abduljabbar Zeravan Arif Ali. «Exploring the Power of eXtreme Gradient Boosting Algorithm in Machine Learning: a Review». En: *Academic Journal of Nawroz University* (2023), págs. 1-15. DOI: 10.25007/ajnu.v12n2a1612.