

Predicción de la calidad del agua implementado un sistema de adquisición de datos y aprendizaje automático.

Water quality prediction implemented a data acquisition and machine learning system.

Felisa Vanessa López Pineda

Instituto Tecnológico de Tuxtla Gutiérrez, Tecnológico Nacional de México. Posgrado en Ciencias de la Ingeniería, Carretera Panamericana 29050. Tuxtla Gutiérrez Chiapas, México.

d23271374@tuxtla.tecnm.mx

Héctor Ricardo Hernández de León

Instituto Tecnológico de Tuxtla Gutiérrez, Tecnológico Nacional de México. Posgrado en Ciencias de la Ingeniería. Tuxtla Gutiérrez Chiapas, México.

hector.hl@tuxtla.tecnm.mx

Elías Neftalí Escobar Gómez

Instituto Tecnológico de Tuxtla Gutiérrez, Tecnológico Nacional de México. Posgrado en Ciencias de la Ingeniería. Tuxtla Gutiérrez Chiapas, México.

elías.eg@tuxtla.tecnm.mx

José Armando Frago Mandujano

Instituto Tecnológico de Tuxtla Gutiérrez, Tecnológico Nacional de México. Posgrado en Ciencias de la Ingeniería. Tuxtla Gutiérrez Chiapas, México.

jose.fm@tuxtla.tecnm.mx

Andrés Eduardo De Paz Martínez

Instituto Tecnológico de Tuxtla Gutiérrez, Tecnológico Nacional de México. Posgrado en Ciencias de la Ingeniería Mecatrónica, Carretera Panamericana 29050. Tuxtla Gutiérrez Chiapas, México.

m17270661@tuxtla.tecnm.mx

Palabras Clave – calidad del agua; aprendizaje automático; sistema de adquisición de datos.

Resumo — El agua potable es un tema de salud pública, su disponibilidad se encuentra amenazada por factores como el crecimiento de la población, la contaminación y el cambio climático. La gestión sostenible del recurso hídrico requiere proteger los cuerpos de agua, en específico los que suministran el agua potable a la población.

Abstract — Drinking water is a public health issue; its availability is threatened by factors such as population growth, pollution, and climate change. Sustainable management of water resources requires protecting bodies of water, specifically those that supply drinking water to the population.

El monitoreo y análisis de la calidad del agua implementando un sistema de adquisición de datos para la recolección de datos que son analizados a través de aprendizaje automático permite una supervisión eficiente con respuesta a anomalías que se puedan presentar. Los datos se procesan con algoritmos de aprendizaje automático para evaluar variables y realizar toma de decisiones informadas.

The monitoring and analysis of water quality by implementing a data acquisition system, where the collected data is analyzed through machine learning, allows for efficient supervision in response to anomalies that may arise. The data is processed with machine learning algorithms to evaluate variables and make informed decisions.

En plantas de tratamiento de agua potable, los modelos de aprendizaje automático monitorean y diagnostican en tiempo real, identificando anomalías en los datos. La investigación propone un sistema de adquisición de datos con el uso de sensores que realicen mediciones de temperatura, turbidez, conductividad, sólidos disueltos totales (TDS) y pH, estos datos se analizan a través de algoritmos de aprendizaje automático como regresión lineal, random forest y k-nearest neighbors (KNN) permitiendo analizar los datos con mejor precisión y realizando toma de decisiones eficiente con respecto a la calidad del agua.

In drinking water treatment plants, machine learning models monitor and diagnose in real time, identifying anomalies in the data. The research proposes a data acquisition system using sensors that measure temperature, turbidity, conductivity, total dissolved solids, and pH. These data are analyzed through machine learning algorithms such as linear regression, random forest, and k-nearest neighbors (KNN), allowing for better precision in data analysis and efficient decision-making regarding water quality.

Keywords – water quality; machine learning; data acquisition system.

I. INTRODUCCIÓN

El agua potable es una necesidad básica y un elemento preciado para la vida cotidiana. La seguridad, sostenibilidad y accesibilidad del agua potable son preocupaciones relevantes en todo el mundo, se ha convertido en un objetivo principal en temas de investigación. La contaminación del agua (figura 1) por diferentes agentes, como residuos industriales, agroquímicos, desechos urbanos y otros contaminantes provocan graves consecuencias para la salud humana y el medio ambiente [1].

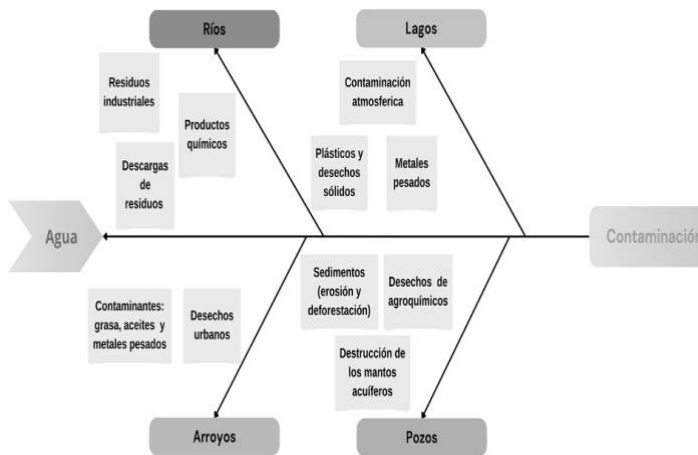


Figura 1. Tipos de fuentes de agua y contaminación

El suministro de agua para consumo humano con una correcta eficacia del agua es fundamental para prevenir la transmisión de enfermedades relacionadas con el agua, para lo cual se establece y mantienen actualizados los límites permisibles de la calidad del agua. Para establecer la calidad del agua es necesario identificar los parámetros establecidos en sus diferentes tipos como son: físicos, químicos y biológicos, a los cuales se establecen un rango de medición que brindan información del estado en que se encuentra el agua[2].

La Secretaría de Salud, propuso la emisión de la Norma Oficial Mexicana NOM-127-SSA1-2021, con el objeto de establecer un control sanitario del agua que se somete a tratamientos de potabilización con el objetivo de hacerla apta para su uso y consumo humano, de acuerdo a las necesidades de la población.

Para una gestión sostenible del recurso hídrico es fundamental implementar medidas que aseguren la protección de las fuentes de agua, así como su uso de manera responsable y eficiente. La vigilancia y el monitoreo constante de la calidad del agua son acciones imprescindibles para identificar problemas de forma preventiva y tomar medidas correctivas[3].

Para identificar la calidad del agua en las plantas potabilizadoras, se realiza una serie de procedimientos manuales, que requiere de mano de obra y tiempo, en donde se recolectan las muestras de agua cruda y durante las diferentes fases del proceso de potabilización[4], se analizan las muestras y se determina el rango en el que se encuentran los parámetros y se evalúa con respecto a la NOM-127-SSA1-2021; analizando el proceso de potabilización de forma manual, se identifica que durante el proceso de recolección de muestras y la medición efectuada de forma manual se pueden presentar deficiencias, generando errores en la medición de los datos y brindando datos incorrectos al determinar la calidad del agua.

La utilización e implementación de modelos de aprendizaje automático y tecnologías como el internet de las cosas (IoT) para el monitoreo y análisis de los datos[5] relacionados con la calidad del agua y variables representan un avance significativo. Estas herramientas permiten una supervisión eficiente y en tiempo real, lo que facilita la detección de anomalías y mejorar la capacidad de respuesta ante cambios en la calidad del agua[6].

Al realizar el proceso de producción en las plantas potabilizadoras se consideran diferentes parámetros de medición que se estudian para predecir la pureza del agua. El uso de sistemas para la recolección de datos que implementan tecnologías como IoT y aprendizaje automático, monitoreo y gestión del recurso hídrico[7], realizan análisis de los datos mediante la instrumentación de sensores que realizan la medición de parámetros. Los datos obtenidos de los sensores se registran en una base de datos para después procesar y realizar el análisis[8]. La combinación de IoT y aprendizaje automático permite la recopilación de datos continuos, diversidad de tipos de datos, conectividad de dispositivos, adaptabilidad, mejora continua en los modelos de predicción y escalabilidad[9]; ofrece una solución para recopilar y analizar datos en tiempo real, lo que puede ayudar a prevenir problemas en la calidad del agua antes que se conviertan en una amenaza para la salud pública[10].

En esta investigación se diseñó un sistema de adquisición de datos que permite el monitoreo continuo del agua a partir de sensores habilitados y calibrados, la combinación de IoT y aprendizaje automático para análisis datos, mediante un modelo de predicción implementando aprendizaje automático, que se ajuste continuamente a los nuevos datos reportados para mejorar el tiempo de procesamiento y la precisión de las mediciones, para una correcta toma de decisiones. El sistema implementa técnicas para identificar patrones y anomalías en los datos de calidad del agua que son de utilidad para mejorar la eficiencia en el monitoreo[11].

Actualmente es indispensable realizar evaluaciones continuas y en periodos cortos de tiempos para conocer la situación que guarda el agua potable para uso y consumo humano[12], [13]. La intención de este trabajo es mostrar el

trabajo de investigación realizado, en donde se implementa un sistema de adquisición de datos combinando con internet de las cosas para el almacenamiento y procesamiento de la información, y la implementación de un algoritmo de aprendizaje automático que brinde información eficiente para la correcta toma de decisiones y con ello cuidar la calidad del agua que se suministra.

II. MATERIALES Y MÉTODOS.

A. Área de estudio

El río Grijalva (figura 3) se encuentra situado en el estado de Chiapas, México, es un recurso vital del sistema hidrológico mexicano. El río se extiende desde las montañas de Chiapas hasta el Golfo de México, inmerso en diversas áreas geográficas y climáticas. Su ubicación estratégica es de importancia ya que es una fuente de agua dulce para el estado[14], siendo el principal abastecimiento de agua, se somete a un proceso de potabilización para convertir en agua potable para uso y consumo para la población de la ciudad capital Tuxtla Gutiérrez, lo cual abastece para uso en viviendas, ganadería, agricultura, industria, biodiversidad, el turismo, entre otros.

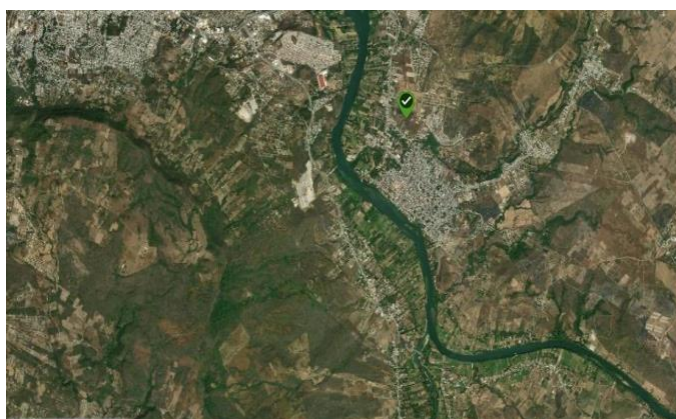


Figura 2. Río Grijalva (CONAGUA, 20023)

La relevancia del río Grijalva con referencia al agua es vital debido a los diversos desafíos que se enfrentan[15]. El caudal del río experimenta variaciones en el agua, lo que ocasiona un cambio en la calidad del agua, esto a consecuencia de factores como: temporada de lluvias, la contaminación y actividades humanas, la agricultura, turismo o la industria. Las variaciones afectan la disponibilidad del agua, la salud de los ecosistemas y la salud de los habitantes[16] de la ciudad capital, por lo que es necesario el monitoreo y análisis de los parámetros físicos químicos del agua en el río Grijalva.

Ante las variaciones que se presentan, es necesario realizar un proceso de potabilización en el agua, y obtener agua de calidad que sea apta para las diferentes actividades.

El proceso de potabilización del agua (figura 3) es una serie de pasos para convertir el agua cruda, proveniente de fuentes naturales como ríos, lagos, arroyos o pozo, en agua potable segura y apta para consumo humano. Los pasos de este proceso son captación y pretratamiento, el proceso inicia con la captación del agua cruda desde su fuente natural (ríos, lagos, arroyos o pozos). Antes de que el agua ingrese a la planta de tratamiento, puede pasar por una etapa de pretratamiento que incluye la remoción de materiales grandes como desechos (orgánicos e inorgánicos), así como la sedimentación para eliminar partículas sólidas grandes. El siguiente paso es la coagulación y floculación, en esta etapa, se añade el agua coagulante, que ayuda a unir partículas más pequeñas y sólidos en suspensión formando flóculos grandes. A continuación, es la sedimentación, el agua pasa a través de grandes tanques de sedimentación donde los flóculos más grandes se asientan en el fondo debido a su peso. Se prosigue con la filtración, el agua clarificada se somete a través de diferentes medios filtrantes, como arena y carbón activado. Estos medios ayudan a retener partículas finas y otras impurezas que puedan quedar después de la sedimentación. Lo siguiente es la desinfección, para asegurar que el agua esté libre de microorganismos, garantizando que el agua sea segura para el consumo humano. Finalmente, una vez completado el proceso de potabilización, el agua tratada se almacena en tanques para luego ser distribuida a través de una red de tuberías hacia los hogares, industrias y otros lugares donde se necesita agua potable. El proceso de potabilización del agua, es amplio y elaborado, el cuidado que se requiere en el manejo, uso de químicos y control en los parámetros requeridos para la obtención del agua potable, es un arduo trabajo, el cual se realiza en la mayoría del proceso de forma manual. Por lo que se denota la importancia de implementar aprendizaje automático en estudios de la calidad del agua.

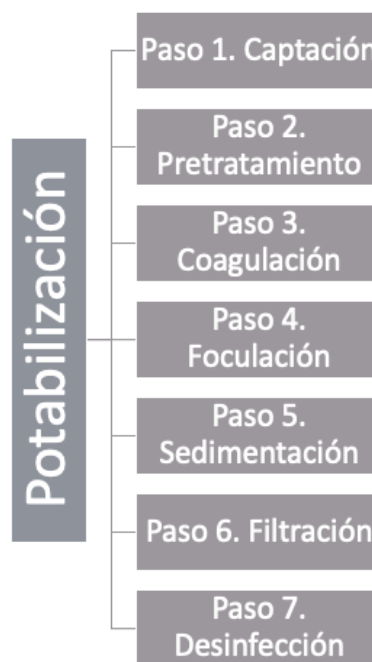


Figura 3. Proceso de potabilización del agua.

El proceso de potabilización conlleva el análisis de parámetros físicos químicos, en donde se analiza el pH, la turbidez, temperatura, conductividad y sólidos disueltos totales (TDS), entre otros parámetros establecidos en la NOM-127-SSA1-2021. Para el desarrollo de esta investigación, se identificaron estos cinco parámetros principales, que permiten determinar la calidad del agua. Sin embargo, se destaca un parámetro, la turbidez, el análisis del agua turbia en el río Grijalva es un indicador de contaminación en el agua. La turbidez se refiere a la cantidad de partículas suspendidas en el agua, como sedimentos, materia orgánica y microorganismos. Al detectarse altos niveles de turbidez afectan la transparencia del agua[17], lo que reduce la penetración de la luz solar y altera el hábitat acuático, lo que provoca problemas para el ecosistema acuático y los seres humanos que dependen del consumo de agua cruda para el proceso de potabilización del agua.

El diseño de un sistema de adquisición de datos para monitoreo y análisis predictivo para la calidad del agua es relevante debido a que permite incrementar la capacidad para abordar una serie de desafíos críticos relacionados con la disponibilidad de agua potable segura para la población. Los modelos de monitoreo y diagnóstico basado en algoritmos de aprendizaje automático[18] en una estación de tratamiento de agua potable permiten conocer el estado del comportamiento de los parámetros físicos químicos en la planta e identificar parámetros característicos que ayuden a la toma de decisiones en el proceso de potabilización.

B. Materiales

Para lograr una gestión sostenible del recurso hídrico, es fundamental implementar medidas que aseguren la protección de las fuentes de agua, así como el uso responsable y eficiente de los datos. La utilización de modelos de aprendizaje automático y tecnologías como el internet de las cosas (IoT) para el monitoreo y análisis de la calidad del agua representa un avance significativo en procesos de investigación[19].

Las plantas potabilizadoras presentan desafíos y problemas, entre ellos lo relacionado con los métodos de monitoreo tradicionales[20]. La toma de muestras es limitada y sujeta a intervalos de tiempos que dependen de la época del año y la situación ambiental, dando como consecuencia cambios significativos que no se analizan en la calidad del agua entre cada muestreo.

Por lo anterior, el trabajo de recolección de muestras del río Grijalva se enfoca en determinar la calidad del agua al ingresar a una planta potabilizadora para identificar el nivel de turbidez, temperatura, pH, conductividad y sólidos disueltos totales. A la vez, se implementa el sistema de adquisición de datos ejecutando un monitoreo continuo del agua, esto a partir de los sensores habilitados, los datos se almacenan en una plataforma[21], lo que permite generar información que se suministra a un modelo de predicción mediante el uso de aprendizaje automático.

Se realizó un estudio inicial de la calidad del agua en el río Grijalva, esto debido a que la ciudad capital Tuxtla Gutiérrez se abastece del recurso hídrico para la zona metropolitana. Se procedió a realizar mediciones manuales de los parámetros representativos y establecer un referente con respecto a la medición de los parámetros. Para el desarrollo de este primer estudio y la recolección de muestras subsecuentes se utilizó material básico de laboratorio.

La recolección de muestras se llevó a cabo en las inmediaciones del río Grijalva, principal fuente de suministro de agua de la ciudad. Para la recolección de las muestras se utilizaron los siguientes materiales: guantes de látex grado examinación; cofia de polipropileno; tubos tipo Falcon de 50 ml con tapa de polietileno, esterilizado y con parche de escritura para registrar la muestra; y contenedor de espuma aislante de 23x23x18 cm.



Figura 4. Materiales utilizados (guantes de látex, cofia, tubo tipo Falcon y contenedor).

Para realizar las mediciones de pH, temperatura, turbidez, conductividad y TDS in situ, se utilizó termómetro digital, medidor de pH (se calibro), turbidímetro, conductímetro y medidor de TDS.

De forma paralela se utilizó el sistema de adquisición de datos (figura 5) para realizar las mediciones correspondientes a los parámetros físico químicos como temperatura, turbidez, conductividad, pH y sólidos disueltos totales. Los resultados manuales de laboratorio con respecto a las muestras obtenidas en las inmediaciones del río Grijalva, se compararon con las mediciones obtenidas del sistema de adquisición de datos para verificar la eficiencia en la toma de mediciones. Los resultados obtenidos del sistema de adquisición de datos fueron son eficientes y continuos con respecto a las mediciones obtenidas de los parámetros físico químicos.



Figura 5. Sistema de adquisición de datos.

C. Método

El proceso del análisis de las muestras se llevó a cabo con un enfoque integral y cuidadoso para garantizar la representatividad de los datos recopilados. Para el desarrollo del estudio se obtuvieron datos representativos de la calidad del agua en el río Grijalva, se seleccionó un sitio de muestreo a lo largo del curso del río, este sitio es estratégico para la recolección de agua cruda que se encuentra próxima a la entrada de una planta de potabilización, elementos como diversidad geográfica, posibles fuentes de contaminación y la distribución estratégica de la zona en donde se realiza la recolección de agua, son factores importantes al realizar el muestreo.

El periodo en el que se recolectan los datos, es a lo largo de un año, para identificar las variaciones estacionales, las condiciones climáticas cambiantes, así como otros factores, entre los que se considera los desechos humanos e industriales. Se establecieron intervalos de medición, teniendo en cuenta factores como la temporalidad de las lluvias, la actividad agrícola o acuícolas de la zona. La contaminación de la población y la presencia de eventos naturales como deslizamiento de tierra en zonas cercanas.



Figura 6. Recolección de muestras río Grijalva

Las muestras se recolectaron en un punto estratégico para conocer el comportamiento de los parámetros físico químicos a lo largo del período establecido. Para la medición de los parámetros se implementó un dispositivo que integra un sistema de adquisición de datos, que se compone de sensores y una plataforma para almacenamiento y análisis de los datos recolectados[22]; los sensores implementados son: temperatura, turbidez, TDS y pH, en donde se obtiene la medición de 5 parámetros, se incluye conductividad, al aplicar un factor de conversión con el parámetro de TDS. Los datos recopilados se analizan de acuerdo con la norma NOM-127-SSA-2021, que establece los criterios de calidad del agua para uso y consumo humano en México.

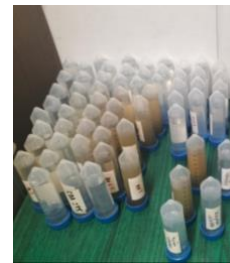


Figura 7. Muestras recolectadas en el río Grijalva

Además de los parámetros físico químicos, se registran datos contextuales como la fecha y la hora de cada medición. Estos datos adicionales son relevantes para contextualizar y analizar los resultados obtenidos, identificar posibles factores externos que podrían influir en la calidad del agua y correlacionar las observaciones con eventos ambientales específicos.

III. SISTEMA DE ADQUISICIÓN DE DATOS

El desarrollo de un sistema de monitoreo y pronóstico de la calidad del agua, requiere satisfacer las necesidades de tecnologías de la comunicación, controladores, sensores y plataformas de aplicación utilizando IoT[23].

El diseño e implementación del sistema de adquisición de datos para el monitoreo del río Grijalva representó un paso crucial en la recolección de información precisa y confiable sobre la calidad del agua. Este sistema fue concebido con el objetivo de automatizar el proceso de recolección de datos, garantizando la continuidad de las mediciones a lo largo del tiempo y permitiendo un monitoreo remoto y en tiempo real de los parámetros de interés[24].

El sistema de adquisición de datos consta de varios componentes interconectados[25][26], [27], cada uno desempeñando un papel específico en la recolección, transmisión y almacenamiento de los datos. Los componentes incluyen: dispositivo de medición automatizado, se utilizó un dispositivo específicamente diseñado para el monitoreo de la calidad del agua en entornos fluviales. Este dispositivo está equipado con una variedad de sensores de alta precisión para medir parámetros como temperatura, turbidez, TDS, conductividad y pH. La selección de sensores adecuados es fundamental para garantizar la exactitud y la fiabilidad de las mediciones.





Sensor	Modelo	Rango de medición	Unidades	Imagen
Temperatura	Omega PR-21	0-100	°C	
Turbidez	Seeed Studio 101020752	0-4000	NTU	
Sólidos disueltos totales (TDS)	LOGOELE	0-2000	PPM	
pH	SEN0161-V2	0-14	pH	

Figura 8. Sensores

La plataforma de comunicación y almacenamiento de datos en la nube[27] empleada es Amazon Web Service (AWS)[28], [29], el servicio que se utiliza para el almacenamiento de los datos es Amazon Simple Storage Service (S3) y Amazon Elastic Compute Cloud (EC2). Estos dos productos de Amazon se seleccionaron con base en características que se adecuan a las necesidades del trabajo de investigación. Los datos almacenados en Amazon S3, proporciona un servicio de almacenamiento altamente escalable y duradero. Amazon EC2 proporciona capacidad de computación escalable en la nube, permitiendo lanzar y administrar instancias de servidores virtuales.



Figura 9. AWS S3 y EC2

Los datos recopilados por el sistema de adquisición de datos mediante las mediciones se transmiten de manera inalámbrica usando un protocolo de comunicación, garantizando la integridad y la continuidad de la información. Esta transmisión se lleva a cabo utilizando un módulo de comunicación Wi-Fi integrado en el sistema de adquisición de datos. El módulo opera bajo los estándares IEEE 802.11, permitiendo la transferencia de datos y con baja latencia.

El proceso de transmisión y almacenamiento se ejecuta de la siguiente forma, la captura de datos se realiza a través de los sensores integrados en el sistema de adquisición, se miden los parámetros como temperatura, pH, turbidez, conductividad y TDS. Estos datos se digitalizan mediante un convertidor analógico digital (DAC). La transmisión inalámbrica es por el módulo Wi-Fi, operando bajo el estándar IEEE 802.11[30], se

establece una conexión con una red Wi-Fi local. Los datos digitalizados se transmiten inalámbricamente a través de la red hacia un punto de acceso. Una vez en el punto de acceso, los datos se envían a AWS utilizando un protocolo seguro[31] como HTTPS. Finalmente, los datos se almacenan en Amazon S3 y se procesan aplicando aprendizaje automático en EC2.

La infraestructura basada en la nube garantiza la disponibilidad y la integridad de los datos, así como su accesibilidad desde cualquier ubicación con conexión a internet[29]. El sistema de adquisición de datos se implementó para organizar y procesar los datos recopilados de manera eficiente. Este sistema permite la catalogación, indexación y consulta rápida de los datos, facilitando su análisis posterior y la generación de informes[32].

Además, se implementaron medidas de seguridad para proteger la confidencialidad y la integridad de los datos, garantizando su cumplimiento con las regulaciones de privacidad y seguridad de datos. El monitoreo remoto y las alertas automatizadas, ayudan para detectar y responder a cambios significativos en la calidad del agua en tiempo real. Por lo anterior, se establecieron umbrales de alerta para cada parámetro medido, y se configuraron notificaciones automáticas para informar a los usuarios sobre cualquier desviación de los valores esperados. Esto permite una respuesta rápida y eficaz ante eventos anómalo o situaciones de riesgo para la salud humana y el medio ambiente.



Figura 10. Representación del sistema de adquisición de datos, plataforma AWS y técnica de aprendizaje automático.

El sistema de adquisición de datos implementado para el monitoreo del río Grijalva representa una combinación de tecnología avanzada, diseño particular y procesos robustos, que permite la recolección de datos de forma precisa y confiable sobre la calidad del agua en tiempo real. Este sistema proporcionó una base sólida para la toma de decisiones informadas y la gestión efectiva de los recursos hídricos en la región.

A. Preprocesamiento de datos

Mediante el sistema de adquisición de datos se recopilan los datos que genera el dataset con los datos de los parámetros que se procesan para el análisis y generación de información para toma de decisiones. Antes de la implementación del

procesamiento de datos es necesario realizar un proceso a los datos adquiridos para adaptarlos y con ello obtener resultados óptimos. El dataset que se preprocesa contiene información de cinco parámetros: temperatura, turbidez, TDS, conductividad y pH, tomando como referencia la ponderación de la norma 127-SSA-2021 con respecto a la calidad del agua. La variabilidad de los parámetros está en función al rango establecido en la norma, por lo que es importante la normalización de datos.

Se realiza el preprocesamiento en el dataset obtenido del sistema de adquisición de datos, se realiza eliminación de valores atípicos y datos faltantes, esto mediante la técnica de desviación estándar para identificar la dispersión en el conjunto de datos. En donde, la fórmula para la desviación estándar de una población es:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2} \quad (1)$$

donde:

N es el número de datos en la población.

x_i son los valores individuales de los datos.

μ es la media aritmética de la población, calculada como:

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i \quad (2)$$

El siguiente paso en el preprocesamiento es el proceso de normalización de datos, en donde se ajusta los valores de las características de los datos a una escala entre 0 y 1 sin distorsionar las diferencias en los rangos de valores. Esto se realiza para facilitar la comparación de diferentes características y mejorar la convergencia y el rendimiento de los algoritmos de aprendizaje automático. Se aplica la técnica Min-Max Scaling:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (3)$$

Mediante la aplicación de la técnica Min-Max Scaling se colocan los datos en una escala común, mejorando así la eficacia y la precisión de los modelos predictivos, contribuyendo en la preparación de los datos para el procesamiento.

B. Procesamiento

El conjunto de datos recopilados por medio del sistema de adquisición de datos se encuentran contenidos en el dataset de esta investigación, los datos hacen referencia a los parámetros de temperatura, turbidez, pH, conductividad y TDS, estos relacionados con la calidad del agua. En esta investigación se realizan evaluaciones para conocer el nivel de precisión utilizado en técnicas de aprendizaje automático, como son: random forest, regresión lineal, KNN, SVM y árbol de decisión.

Para evaluar la precisión y el rendimiento de los modelos de aprendizaje automático, se utilizan métricas de evaluación de aprendizaje automático las cuales son herramientas cuantitativas que se utilizan para medir el rendimiento del modelo. Estas métricas proporcionan una manera objetiva de comparar diferentes modelos y seleccionar el mejor ajuste a los datos y al problema establecido a resolver.

Las métricas utilizadas en este trabajo de investigación es el Error Absoluto Medio (MAE), Error Cuadrático Medio (MSE) y el Coeficiente de Determinación (R2). A continuación, se explica cada una de estas métricas:

El Error Absoluto Medio (MAE) mide el promedio de los errores absolutos entre las predicciones del modelo y los valores reales. Es una medida de precisión que representa la magnitud promedio de los errores en un conjunto de predicciones. Un MAE bajo indica que las predicciones del modelo están cerca de los valores reales, lo que sugiere un mejor rendimiento del modelo. Es menos sensible a valores atípicos en comparación con el Error Cuadrático Medio (MSE).

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

Donde:

y_i es el valor real

\hat{y}_i es el valor predicho

n es el número de observaciones

El Error Cuadrático Medio (MSE) mide las diferencias al cuadrado entre los valores predichos y los valores reales. Al elevar los errores al cuadrado, se penalizan más los errores, lo que hace que el MSE sea más sensible a estos errores en comparación con el MAE.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (5)$$

Donde:

y_i es el valor real

\hat{y}_i es el valor predicho

n es el número de observaciones

Un MSE bajo indica un mejor rendimiento del modelo, ya que las predicciones están más cerca de los valores reales. Debido a que penaliza más los errores grandes, el MSE es de utilidad cuando es importante minimizar grandes discrepancias entre las predicciones y los valores reales. Sin embargo, debido a la penalización de los errores, es importante considerar tanto MAE como MSE para una evaluación completa.

El Coeficiente de Determinación (R2) es una medida estadística que indica la proporción de la variabilidad en la variable dependiente que es explicada por las variables

independientes en el modelo. Se utiliza para evaluar qué tan bien los datos se ajustan a la línea de regresión.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (6)$$

Donde:

y_i es el valor real

\hat{y}_i es el valor predicho

\bar{y} es el valor promedio de los valores reales

n es el número de observaciones

En los resultados del Coeficiente de Determinación (R^2) se obtienen los siguientes valores en donde se tiene que:

$R^2=1$; indica que el modelo explica toda la variabilidad de los datos de respuesta alrededor de su media.

$R^2=0$; indica que el modelo no explica ninguna variabilidad de los datos de respuesta alrededor de su media.

$R^2 < 0$; indica que el modelo es peor que simplemente usar la media de los datos.

Un R^2 alto indica un mejor ajuste del modelo a los datos, ya que más variabilidad en los datos de respuesta es explicada por las variables independientes. Sin embargo, un R^2 alto no siempre implica que el modelo es adecuado; también debe considerarse la significancia de los coeficientes del modelo y posibles problemas de sobreajuste.

La importancia de la aplicación de métricas de evaluación en aprendizaje automático, permite comparar el rendimiento de diferentes modelos, ayuda a identificar áreas donde el modelo requiere mejoras, informa sobre la adecuación del modelo para problemas específicos y orienta la selección y ajuste de hiperparámetros.

La aplicación de aprendizaje automático en este contexto es especialmente útil debido a la capacidad de los algoritmos para manejar datos no lineales, lo que es recurrente en los elementos naturales, en este caso el agua. A continuación, se describen los algoritmos de aprendizaje automático utilizados en esta investigación. Durante el entrenamiento se construyen varios árboles de decisión utilizando diferentes subconjuntos del conjunto de datos y diferentes subconjuntos de características.

Random Forest es un algoritmo de aprendizaje automático basado en la construcción de múltiples árboles de decisión durante el entrenamiento y la salida de la clase que es el modo de las clases o el promedio de las predicciones de los árboles individuales. Esta técnica reduce la varianza y ayuda a prevenir el sobreajuste. En el entrenamiento se construyen árboles de decisión utilizando diferentes subconjuntos del conjunto de datos y diferentes subconjuntos de características. Para el proceso de predicción cada árbol realiza una predicción y el resultado final se obtiene mediante el promedio de todas las predicciones de los árboles. Esta técnica es útil para capturar las relaciones no lineales entre los parámetros de calidad del agua y proporciona una alta precisión en las predicciones[33], [34].

La regresión lineal múltiple es una técnica estadística que modela la relación entre una variable dependiente y múltiples variables independientes utilizando una ecuación lineal. Es útil para predecir el valor de la variable dependiente basándose en los valores de las variables independientes. Esta técnica asume una relación lineal entre las variables, lo cual puede limitar su aplicabilidad en casos donde las relaciones son no lineales[35].

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad (7)$$

Donde y es la variable dependiente

x_i son las variables independientes

β_i Son los coeficientes del modelo

K-Nearest Neighbors (KNN) es un algoritmo de aprendizaje supervisado que clasifica un nuevo punto basado en los K puntos más cercanos en el conjunto de entrenamiento. Es un método no paramétrico que puede ser utilizado tanto para clasificación como para regresión. Para la predicción de la calidad del agua, KNN puede ser efectivo en identificar patrones locales en los datos, aunque su rendimiento puede verse afectado por la elección de K y la escala de datos. Este tipo de algoritmo no tiene un proceso de entrenamiento explícito; simplemente almacena los datos de entrenamiento. En la predicción se calcula la distancia entre el punto nuevo y todos los puntos en el conjunto de entrenamiento y se selecciona los k más cercanos para determinar la clase o el valor de la predicción[36][37][38].

Los modelos de predicción basados en aprendizaje automático son una herramienta efectiva para el monitoreo y predicción[39], [40] de diferentes variables como las utilizadas en la calidad del agua. Esto permite la optimización de los tiempos de procesamiento y el desarrollo de soluciones de seguridad y privacidad para el manejo de la información. Los modelos de aprendizaje automático utilizando un marco de aprendizaje adaptativo incremental que permite al modelo ajustarse continuamente a los nuevos datos recopilados, mejorando la precisión de las predicciones con el tiempo[41].

IV. RESULTADOS

En esta sección se presentan los resultados obtenidos del trabajo de investigación. En primera instancia se presenta la matriz de correlación entre los parámetros analizados, seguido de las evaluaciones de los algoritmos Random Forest, Regresión Lineal Múltiple y K-Nearest Neighbors (KNN) que fueron utilizados para predicción.

Para evaluar la precisión y el rendimiento de los modelos de aprendizaje automático, se emplearon métricas de evaluación como Error Absoluto Medio (MAE), Error Cuadrático Medio (MSE) y el Coeficiente de Determinación (R^2).

Se presenta la matriz de correlación entre los parámetros de temperatura, turbidez, pH, TDS y conductividad para evaluar la

calidad del agua. La metodología utilizada resalta la importancia de establecer niveles de correlación entre las variables obtenidas. Cada variable fue relacionada con cada una de las otras para identificar correlaciones fuertes y positivas que son relevantes para el estudio.

El análisis de correlación es esencial para entender cómo los parámetros interactúan entre sí y cómo influye conjuntamente en la calidad del agua. La matriz de correlación permite observar la magnitud y dirección de las relaciones entre los diferentes parámetros, facilitando la identificación de patrones significativos y dependencias cruciales.

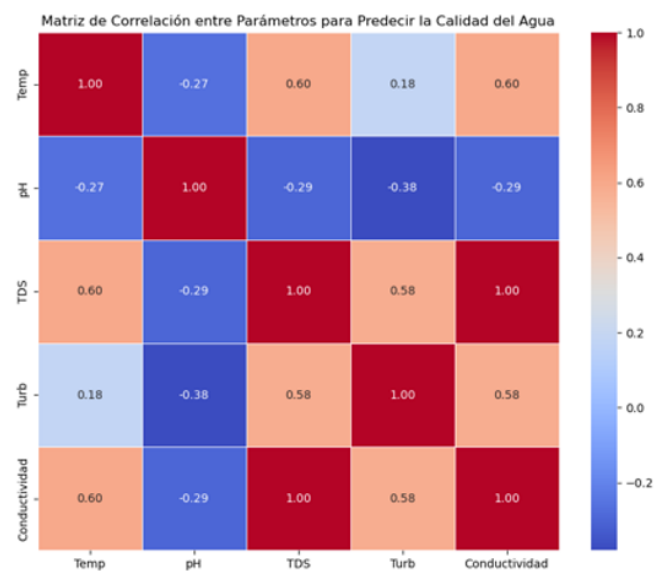


Figura 11. Matriz de correlación entre parametros físico químicos para predecir la calidad del agua.

Las correlaciones fuertes y positivas indican que cuando uno de los parámetros aumenta, es probable que el otro también lo haga, proporcionando información valiosa sobre la dinámica interna de la calidad del agua.

En la figura 11 se muestra visualmente la fuerza y dirección de las relaciones entre los parámetros. Esta matriz proporciona una base sólida para la toma de decisiones en la gestión de la calidad del agua, ayudando a identificar que parámetros deben ser monitoreados y controlados con mayor interés para asegurar una buena calidad del agua. Cabe hacer mención que este análisis es crucial para el desarrollo de modelos predictivos precisos y para la implementación de estrategias efectivas de manejo y conservación del recurso hídrico.

A continuación, se presentan las evaluaciones de los algoritmos que fueron utilizados para predecir la calidad del agua. El conjunto de datos resultantes está conformado por la ejecución y evaluación de los tres modelos de aprendizaje automático explicados. La evaluación se llevó a cabo utilizando el dataset obtenido del sistema de adquisición de datos en un período de un año.

Para medir la precisión y el desempeño de los modelos de aprendizaje automático, se implementaron métricas de evaluación como el Error Absoluto Medio (MAE), Error Cuadrático Medio (MSE) y el Coeficiente de Determinación (R2). Estas métricas proporcionan una visión cuantitativa de la capacidad predictiva de cada modelo, permitiendo identificar cuál de ellos ofrece una mejor precisión y menor error en la predicción de la calidad del agua.

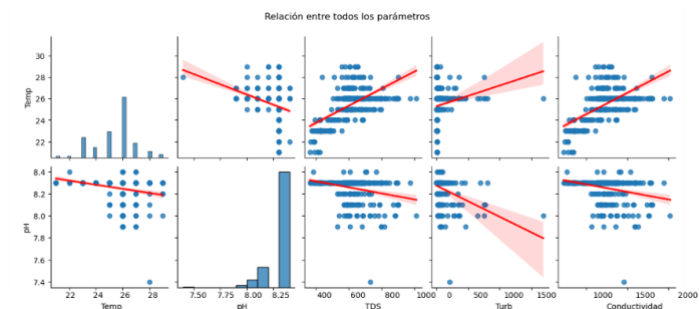


Figura 12. Análisis de relación del parametro de temperatura y pH con los 5 parámetros estudiados.

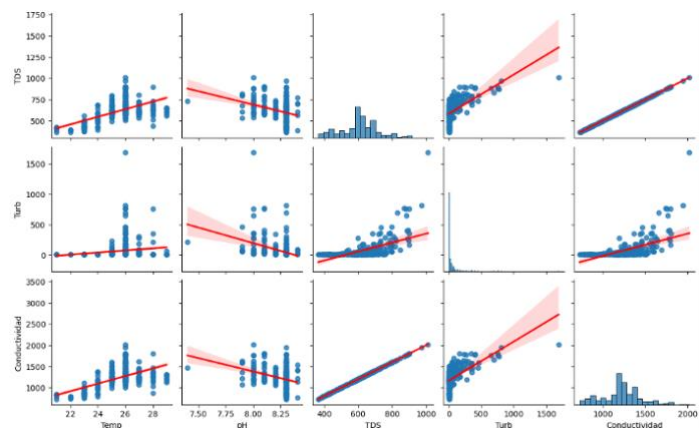


Figura 13. Análisis de relación del parametro de TDS, turbidez y conductividad con los 5 parámetros estudiados.

Se realizo el análisis predictivo de los siguientes modelos, con sus respectivas evaluaciones métricas, para identificar al algoritmo que presenta mejor precisión y desempeño del modelo.

El modelo con el mejor desempeño fue la implementación y evaluación del modelo Random Forest, en donde se observa que se tiene una continuidad en los datos generados.

En la evaluación de métricas se obtiene un MSE de 0.008, un MAE de 0.02 y R2 de 0.97, se identifica con una precisión y desempeño alto con respecto a los modelos de Regresión Lineal Múltiple y KNN.

Random Forest.

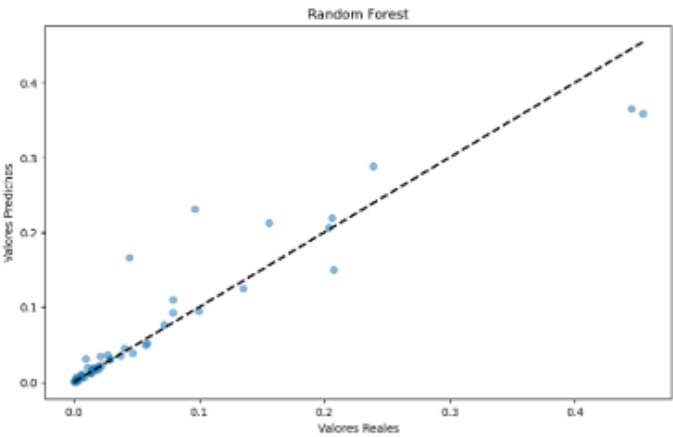


Figura 14. Análisis de predictivo utilizando el algoritmo Random Forest.

Métrica	
MSE	0.008010
MAE	0.021206
R2	0.971548

KNN.

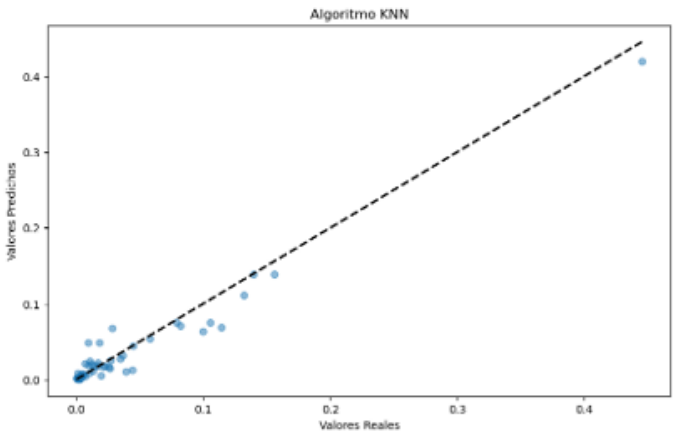


Figura 16. Análisis de predictivo utilizando el algoritmo KNN.

Métrica	
MSE	0.006729
MAE	0.023393
R2	0.956633

Regresión lineal múltiple.

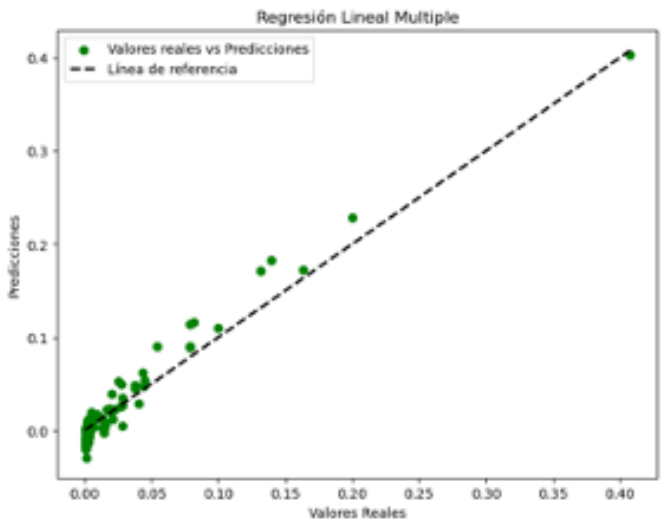


Figura 15. Análisis de predictivo utilizando el algoritmo de regresión lineal.

Métrica	
MSE	0.000164
MAE	0.009234
R2	0.934937

V. CONCLUSIONES

En esta investigación, se realizó un análisis predictivo de la calidad del agua utilizando algunos modelos de aprendizaje automático e implementando un sistema de adquisición de datos. Los modelos evaluados son Random Forest, Regresión Lineal Múltiple y K-Nearest Neighbors (KNN). La intención fue identificar el algoritmo que ofreciera la mejor precisión y rendimiento en la predicción de parámetros de calidad del agua, como temperatura, turbidez, TDS, pH y conductividad.

El análisis mostró que el modelo de Random Forest proporcionó el mejor desempeño entre los modelos evaluados. Los resultados métricos indicaron un MSE de 0.008, un MAE de 0.02 y R2 de 0.97 para el modelo Random Forest. Estos valores denotan una alta precisión y un rendimiento predictivo, mostrando la capacidad del modelo para generar predicciones coherentes y continuas.

La implementación de un sistema de adquisición de datos con el uso de técnicas avanzadas de aprendizaje automático, en específico Random Forest, ha demostrado ser una metodología robusta y precisa para el monitoreo y predicción de la calidad del agua. Este enfoque facilita una gestión eficiente de los recursos hídricos, permitiendo la detección temprana de problemas y la toma de decisiones informadas para la protección del medio ambiente y la salud pública.

AGRADECIMIENTOS

Expresamos nuestro agradecimiento al Tecnológico Nacional de México campus Tuxtla Gutiérrez por su apoyo institucional y por proporcionar los recursos necesarios para la realización de esta investigación. Su compromiso con la educación superior y la investigación científica es crucial para el desarrollo de este trabajo.

Agradezco profundamente a los doctores, Dr. Héctor Ricardo Hernández de León y Dr. Elías Neftalí Escobar Gómez por su invaluable guía y mentoría a lo largo de este proyecto. Su experiencia y consejos son fundamentales para superar los desafíos técnicos y metodológicos que se han encontrado.

También extendiendo el reconocimiento al Dr. José Armando Fragoso Mandujano por su constante apoyo y su valioso aporte científicos. Su conocimiento y orientación han sido esencial para la dirección y fortalecimiento de la investigación.

Finalmente, expreso un agradecimiento al Ing. Andrés Eduardo De Paz Martínez por su dedicación y colaboración en la recopilación, análisis y desarrollo de la investigación. Su arduo trabajo ha sido vital para el éxito de esta investigación.

A todos los colaboradores, mi más profundo agradecimiento por su contribución y apoyo a este proyecto de investigación.

REFERÊNCIAS BIBLIOGRÁFICA

- [1] G. Guan et al., "Water-Quality Assessment and Pollution-Risk Early-Warning System Based on Web Crawler Technology and LSTM," *Int J Environ Res Public Health*, vol. 19, no. 18, 2022, doi: 10.3390/ijerph191811818.
- [2] P. Chen, B. Wang, Y. Wu, Q. Wang, Z. Huang, and C. Wang, "Urban river water quality monitoring based on self-optimizing machine learning method using multi-source remote sensing data," *Ecol Indic*, vol. 146, 2023, doi: 10.1016/j.ecolind.2022.109750.
- [3] B. K. Jha, "Cloud-Based Smart Water Quality Monitoring System using IoT Sensors and Machine Learning," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 3, pp. 3403–3409, Jun. 2020, doi: 10.30534/ijatcse/2020/141932020.
- [4] I. Kurniawan, G. Hayder, and H. M. Mustafa, "Predicting Water Quality Parameters in a Complex River System," *Journal of Ecological Engineering*, vol. 22, no. 1, 2020, doi: 10.12911/22998993/129579.
- [5] C. Z. Zulkifli et al., "IoT-Based Water Monitoring Systems: A Systematic Review," *Water (Switzerland)*, vol. 14, no. 22, 2022, doi: 10.3390/w14223621.
- [6] N. Geetha, "IoT based smart water quality monitoring system," *International Journal of Nonlinear Analysis and Applications*, vol. 12, no. Special Issue, 2021, doi: 10.22075/IJNAA.2021.5853.
- [7] H. Aftab, K. Gilani, J. E. Lee, L. Nkenyereye, S. M. Jeong, and J. S. Song, "Analysis of identifiers in IoT platforms," *Digital Communications and Networks*, vol. 6, no. 3, 2020, doi: 10.1016/j.dcan.2019.05.003.
- [8] V. Radhakrishnan and W. Wu, "IoT Technology for Smart Water System," in *Proceedings - 20th International Conference on High Performance Computing and Communications, 16th International Conference on Smart City and 4th International Conference on Data Science and Systems, HPCC/SmartCity/DSS*, 2018, 2019, doi: 10.1109/HPCC/SmartCity/DSS.2018.00246.
- [9] V. Lakshmikantha, A. Hiriyanagowda, A. Manjunath, A. Patted, J. Basavaiah, and A. A. Anthony, "IoT based smart water quality monitoring system," *Global Transitions Proceedings*, vol. 2, no. 2, 2021, doi: 10.1016/j.gltp.2021.08.062.
- [10] P. Boccadoro, V. Daniele, P. Di Gennaro, D. Lofù, and P. Tedeschi, "Water quality prediction on a Sigfox-compliant IoT device: The road ahead of WaterS," *Ad Hoc Networks*, vol. 126, 2022, doi: 10.1016/j.adhoc.2021.102749.
- [11] A. C. Aguilar Aguilar and F. F. Obando - Díaz, "Aprendizaje automático para la predicción de calidad de agua potable," *Ingeniare*, no. 28, 2020, doi: 10.18041/1909-2458/ingeniare.28.6215.
- [12] G. Hayder, I. Kurniawan, and H. M. Mustafa, "Implementation of machine learning methods for monitoring and predicting water quality parameters," *Biointerface Res Appl Chem*, vol. 11, no. 2, 2021, doi: 10.33263/BRIAC112.92859295.
- [13] R. Huang, C. Ma, J. Ma, X. Huangfu, and Q. He, "Machine learning in natural and engineered water systems," *Water Research*, vol. 205, 2021, doi: 10.1016/j.watres.2021.117666.
- [14] V. A. Gallardo Zavaleta, "Legado histórico de la Comisión del Río Grijalva en Chiapas y Tabasco," *Relaciones Estudios de Historia y Sociedad*, vol. 44, no. 174, 2023, doi: 10.24901/rehs.v44i174.910.
- [15] A. A. Pease et al., "Rivers of Mexico," in *Rivers of North America*, Second Edition, 2023, doi: 10.1016/B978-0-12-818847-7.00004-5.
- [16] C. Rivera Farfán, "Desplazamiento ambiental forzado. La pertinencia de una reflexión conceptual," *Maya America: Journal of Essays, Commentary, and Analysis*, vol. 5, no. 1, 2023, doi: 10.32727/26.2023.13.
- [17] R. Hu, W. Xu, W. Yan, T. Wu, X. He, and N. Cheng, "Comparison between Machine-Learning-Based Turbidity Models Developed for Different Lake Zones in a Large Shallow Lake," *Water (Switzerland)*, vol. 15, no. 3, 2023, doi: 10.3390/w15030387.
- [18] M. Lowe, R. Qin, and X. Mao, "A Review on Machine Learning, Artificial Intelligence, and Smart Technology in Water Treatment and Monitoring," *Water (Switzerland)*, vol. 14, no. 9, 2022, doi: 10.3390/w14091384.
- [19] A. Bhardwaj et al., "Smart IoT and Machine Learning-based Framework for Water Quality Assessment and Device Component Monitoring," *Environmental Science and Pollution Research*, vol. 29, no. 30, 2022, doi: 10.1007/s11356-022-19014-3.
- [20] F. Jan, N. Min-Allah, and D. Düşteğör, "IoT based smart water quality monitoring: Recent techniques, trends and challenges for domestic applications," *Water (Switzerland)*, vol. 13, no. 13, 2021, doi: 10.3390/w13131729.
- [21] A. H. Miry and G. A. Aramice, "Water monitoring and analytic based ThingSpeak," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 4, 2020, doi: 10.11591/ijece.v10i4.pp3588-3595.
- [22] J. N. Kabi, C. wa Maina, E. T. Mharakurwa, and S. W. Mathenge, "Low cost, LoRa based river water level data acquisition system," *HardwareX*, vol. 14, 2023, doi: 10.1016/j.ohx.2023.e00414.
- [23] E. Kaur, "IoT Regulated Water Quality Prediction Through Machine Learning for Smart Environments," in *Intelligent Systems Reference Library*, vol. 121, 2022, doi: 10.1007/978-3-030-97516-6_3.
- [24] F. Sanchez-Sutil and A. Cano-Ortega, "Smart plug for monitoring and controlling electrical devices with a wireless communication system integrated in a LoRaWAN," *Expert Syst Appl*, vol. 213, 2023, doi: 10.1016/j.eswa.2022.118976.
- [25] E. E. D. Hemdan, Y. M. Essa, M. Shouman, A. El-Sayed, and A. N. Moustafa, "An efficient IoT based smart water quality

- monitoring system,” *Multimed Tools Appl*, 2023, doi: 10.1007/s11042-023-14504-z.
- [26] S. Srivastava, “Study of IoT Based Smart Water Quality Monitoring System,” *Int J Res Appl Sci Eng Technol*, vol. 9, no. VIII, 2021, doi: 10.22214/ijraset.2021.37483.
- [27] S. Iranpak, A. Shabbahrami, and H. Shakeri, “Remote patient monitoring and classifying using the internet of things platform combined with cloud computing,” *J Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00507-w.
- [28] Abhishek Saini, Chaman Sharma, Nadeem Khan, Rohit Chauchan, and Gurjeet Singh, “A REVIEW PAPER ON AWS,” *EPRA International Journal of Multidisciplinary Research (IJMR)*, 2024, doi: 10.36713/epra15444.
- [29] S. Chakraborty and P. S. Aithal, “Let Us Create An IoT Inside the AWS Cloud,” *International Journal of Case Studies in Business, IT, and Education*, 2023, doi: 10.47992/ijcsbe.2581.6942.0253.
- [30] S. I. Pella and H. F. Lami, “Integrating iee802.11 and lorawan for wireless sensor network data transaction in non-infrastructure area,” *Jurnal Media Elektro*, 2023, doi: 10.35508/jme.v0i01.10105.
- [31] Ita Ita, Muhlis Tahir, and Edem Vincentius, “Analisis Komparatif Layanan Cloud: Microsoft Azure, Aws, Dan Google Cloud Platform (GCP),” *Jurnal Informasi, Sains dan Teknologi*, vol. 6, no. 02, 2023, doi: 10.55606/isaintek.v6i02.127.
- [32] A. Simkin, A. Kopp, and O. Olkhovyi, “Research on the optimization model for building an efficient IT infrastructure using the AWS platform,” *InterConf*, no. 38(175), 2023, doi: 10.51582/interconf.19-20.10.2023.027.
- [33] V. Khandelwal and S. Khandelwal, “Ground Water Quality Index Prediction Using Random Forest Model,” in *Lecture Notes in Networks and Systems*, 2023, doi: 10.1007/978-981-19-8825-7_40.
- [34] P. K. Jena, S. M. Rahaman, P. K. Das Mohapatra, D. P. Barik, and D. S. Patra, “Surface water quality assessment by Random Forest,” *Water Pract Technol*, vol. 18, no. 1, 2023, doi: 10.2166/wpt.2022.156.
- [35] H. Ghosh, M. A. Tusher, I. S. Rahat, S. Khasim, and S. N. Mohanty, “Water Quality Assessment Through Predictive Machine Learning,” in *Lecture Notes in Networks and Systems*, 2023, doi: 10.1007/978-981-99-3177-4_6.
- [36] S. Y. Abuzir and Y. S. Abuzir, “Machine learning for water quality classification,” *Water Quality Research Journal*, vol. 57, no. 3, 2022, doi: 10.2166/wqrj.2022.004.
- [37] A. Najah Ahmed et al., “Machine learning methods for better water quality prediction,” *J Hydrol (Amst)*, vol. 578, 2019, doi: 10.1016/j.jhydrol.2019.124084.
- [38] S. Singha, S. Pasupuleti, S. S. Singha, R. Singh, and S. Kumar, “Prediction of groundwater quality using efficient machine learning technique,” *Chemosphere*, vol. 276, 2021, doi: 10.1016/j.chemosphere.2021.130265.
- [39] O. K. Pal, “The Quality of Drinkable Water using Machine Learning Techniques,” *International Journal of Advanced Engineering Research and Science*, vol. 9, no. 6, 2022, doi: 10.22161/ijaers.96.2.
- [40] M. Zhu et al., “A review of the application of machine learning in water quality evaluation,” *Eco-Environment & Health*, vol. 1, no. 2, pp. 107–116, Jun. 2022, doi: 10.1016/j.eehl.2022.06.001.
- [41] L. Li, S. Rong, R. Wang, and S. Yu, “Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review,” *Chemical Engineering Journal*, vol. 405, 2021, doi: 10.1016/j.cej.2020.126673.