

## **Seminario de investigación**

**Mario Andrés Hernández Moreno**

**Actividad 2: Planteamiento del problema de investigación**

**Dr. Oscar Cardona Morales**

**Universidad Autónoma de Manizales, Manizales**

**Especialización en Inteligencia Artificial**

**2024**

## **1. Contexto de la aplicación, proceso o fenómeno que se aborda**

La Enfermedad Diarreica Aguda (EDA) continúa siendo un desafío importante para la salud pública global, en especial en regiones con recursos limitados y problemas de acceso a agua potable. En el municipio de Ipiales, Nariño, una ciudad de aproximadamente 160.000 habitantes al suroccidente de Colombia, un reciente brote de EDA con 366 casos reportados en un período de seis días en septiembre de 2024, ha sido un claro ejemplo de la necesidad de implementar un sistema eficiente para la predicción y monitoreo de este tipo de enfermedades. Este brote ocurre en un contexto de crisis hídrica agravada por una sequía prolongada, que ha comprometido la calidad del agua potable, además de la contaminación del río Blanco que abastece a la ciudad, junto con la deficiencia en la infraestructura de tratamiento de agua, lo que ha incrementado el riesgo de propagación de enfermedades.

## **2. Necesidad, problema, dificultad u oportunidad identificada, que se encuentra soportada en el estado del arte.**

La necesidad principal identificada es el desarrollo de un sistema de alerta temprana, predicción y monitoreo continuo para brotes de EDA impulsado mediante inteligencia artificial. Actualmente, las autoridades dependen de algunas medidas correctivas de respuesta estandarizadas, como encuestas y análisis de muestras, lo que limita la capacidad de respuesta rápida, algo que no es suficiente para prevenir nuevos brotes a tiempo. Estudios recientes han demostrado que los sistemas de vigilancia tradicionales a menudo fallan en detectar brotes de manera temprana, especialmente en áreas con recursos limitados.

El estado del arte en vigilancia epidemiológica y salud pública digital sugiere que la implementación de sistemas basados en Inteligencia Artificial (IA) especialmente con enfoques de aprendizaje automático (ML) puede mejorar significativamente la detección temprana y la respuesta a brotes de enfermedades infecciosas. Estudios previos han demostrado que algunos modelos predictivos, como redes neuronales artificiales (ANN), bosques aleatorios (Random Forest), árboles de decisión (DT), máquinas de soporte vectorial (SVM) e incluso técnicas de interpretabilidad como XAI (eXplainable Artificial Intelligence), pueden detectar patrones decisivos en grandes volúmenes de datos para prever picos de enfermedades o tendencias de ciertas patologías. Estos modelos se eligen en función de su capacidad para manejar datos complejos y no lineales, así como su eficacia en la clasificación y predicción de la calidad del agua en tiempo real, lo que podría facilitar un monitoreo más efectivo y accesible de la calidad del agua. Esto representa una solución innovadora que podría transformar la forma en que se gestionan y monitorean los recursos hídricos, mejorando así la salud pública y la seguridad del agua.

## **3. Reto técnico**

El reto técnico principal implica el desarrollo e implementación de un sistema de IA que integre múltiples componentes, fuentes de datos y algoritmos para abordar la complejidad del problema. Basándonos en el estado del arte, los principales aspectos técnicos a considerar son:

### **3.1 Obtención de datos precisos y en tiempo real:**

Uno de los mayores retos es la recopilación y centralización de los datos relevantes para el modelo o algoritmo. Los datos epidemiológicos, microbiológicos, ambientales y socioeconómicos provienen de diversas fuentes como los hospitales, centros de salud, estaciones meteorológicas, la planta de tratamiento, usuarios y sistemas de monitoreo de la calidad del agua, cada uno con diferentes niveles de actualización y precisión. Por lo tanto, se requiere establecer grandes sistemas o mecanismos de recolección automatizada para obtener datos en tiempo real o casi real, teniendo en cuenta siempre una estandarización clara para su análisis, ya que datos inconsistentes o incompletos afectarán la precisión del modelo y la calidad del mismo. Además, la infraestructura en algunas áreas rurales por las que pasa el río y sus

vertientes pueden ser limitadas, dificultando la instalación de sensores para la obtención de y su respectivo mantenimiento.

### **3.2 Calidad y preprocesamiento de datos:**

La calidad de los datos es fundamental para el rendimiento de los modelos de aprendizaje automático. Los datos pueden contener errores, valores faltantes o estar desbalanceados entre las clases, por lo que se debe implementar técnicas de limpieza de datos, aplicar métodos de imputación para valores faltantes, utilizar técnicas de balanceo de clases como sobremuestreo o submuestreo, además de asegurar un conjunto de datos representativo y de alta calidad.

### **3.3 Selección de variables o características:**

Es necesario identificar las características más relevantes que influyen en la predicción de las patologías relacionadas con el agua, ya que un conjunto inadecuado de variables puede llevar a un modelo ineficaz, el desgaste de recursos y de la capacidad de cómputo, por lo tanto, se deben utilizar técnicas de selección de características como regresión o árboles de decisión.

### **3.4 Modelo predictivo de aprendizaje automático:**

Se requiere la selección y ajuste de algoritmos de ML que no solo ofrezcan un alto rendimiento predictivo, sino que también sean adecuados para la clasificación y regresión de los datos. Esto incluye la evaluación de diferentes modelos y la identificación del más adecuado para el contexto específico de la calidad del agua. Se propone el uso de bosques aleatorios (RF) y redes neuronales artificiales (ANN), que han demostrado ser modelos efectivos para tareas de clasificación y regresión, y pueden ser adecuadas en la detección de patrones en grandes conjuntos de datos. El desafío radica en adaptar estos modelos a los datos locales y garantizar su precisión con datos limitados, además de su validación que es crucial para asegurar la fiabilidad y precisión en diferentes condiciones y ubicaciones geográficas.

### **3.5 Integración, interpretabilidad y visualización de datos:**

Desarrollar una plataforma centralizada que integre los diferentes componentes y proporcione visualizaciones intuitivas para los tomadores de decisiones, con una arquitectura de sistemas robusta y la implementación de dashboards interactivos en tiempo real que no solo sean precisos, sino también interpretables, utilizando técnicas de XAI para proporcionar explicaciones claras sobre las predicciones y los factores que influyen en ellas.

### **3.6 Privacidad, seguridad y manejo ético de los datos:**

Implementar técnicas de encriptación para proteger la información personal de los ciudadanos mientras se mantiene la utilidad de los datos para el análisis conforme a estándares internacionales, como el Reglamento General de Protección de Datos (GDPR) o normativas nacionales de Colombia.

### **3.7 Educación y Capacitación:**

Es fundamental capacitar al personal de salud en la toma de decisiones con el uso de estas nuevas tecnologías y en la interpretación de los resultados, por lo que se debe desarrollar programas de capacitación para operadores y personal de salud, además de proporcionar recursos educativos que apoyen el proceso y permitan establecer un sistema de soporte continuo para garantizar la sostenibilidad y efectividad de la solución a largo plazo.

#### **4. Variables a tener en cuenta para la implementación de una solución**

Basándonos en el estado del arte, las variables clave a considerar para la implementación de la solución incluyen:

- Datos epidemiológicos, como el número de casos de EDA, la distribución geográfica, características demográficas de los pacientes, síntomas reportados y evolución de la enfermedad.
- Variables ambientales como la calidad del agua (niveles de contaminantes, pH, turbidez), las condiciones climáticas (temperatura, precipitación, humedad), y los niveles de contaminación del aire.
- Factores socioeconómicos como la densidad poblacional, grupos de edad, acceso a servicios de salud, condiciones de saneamiento y distribución de recursos sanitarios.
- Infraestructura tecnológica con la capacidad de procesamiento, almacenamiento y transmisión de datos en tiempo real.
- Métricas de rendimiento del sistema que incluye la precisión de las predicciones, tiempo de detección de anomalías y latencia del sistema.

#### **5. Impacto potencial de la solución**

La implementación de esta solución basada en IA tendría un impacto significativo en varios aspectos:

##### **5.1 Beneficiarios y beneficios potenciales:**

- Población general de Ipiales (aproximadamente 160,000 habitantes): Mejora en la salud pública y reducción de casos de EDA, especialmente para las personas más vulnerables.
- Autoridades sanitarias locales y departamentales: Toma de decisiones más informada y oportuna.
- Optimización de la distribución de personal médico, medicamentos, recursos y suministros.
- Reducción en el tiempo de respuesta ante brotes potenciales.

##### **5.2 Nivel de desarrollo tecnológico:**

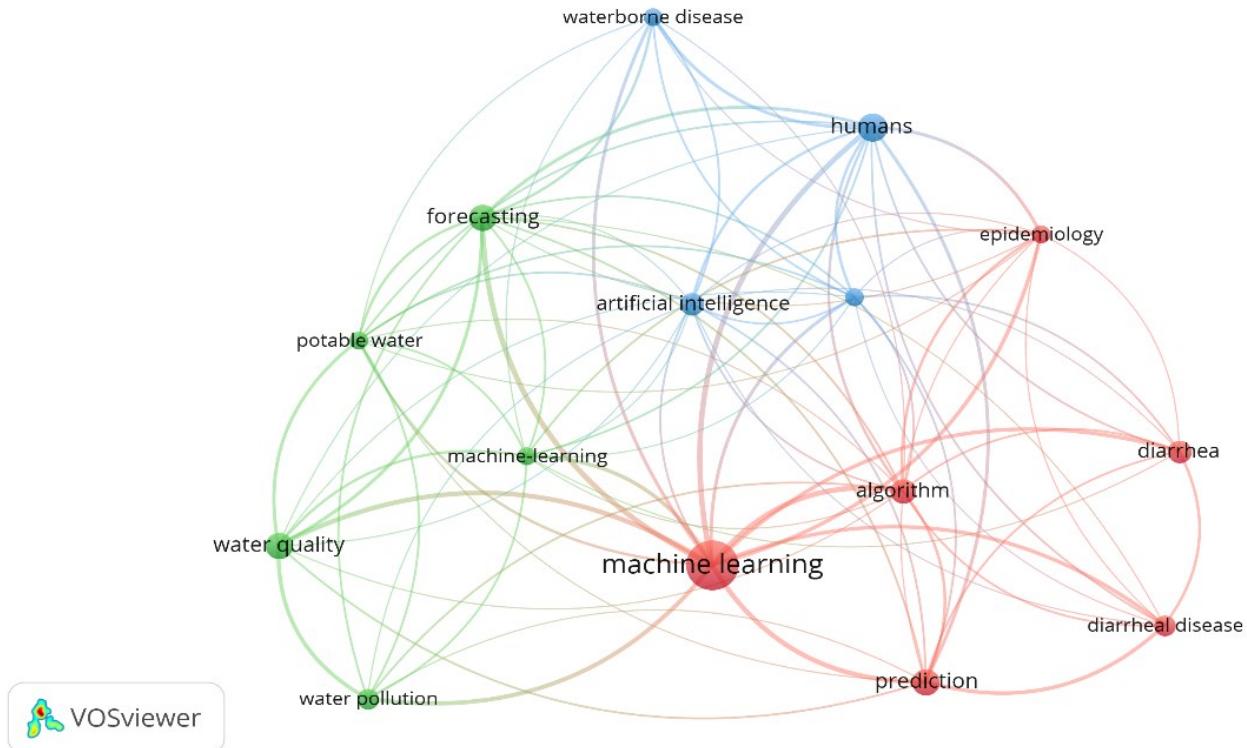
Se espera alcanzar un nivel de madurez tecnológica (TRL) 7-8, que implica la demostración del sistema en un entorno operativo real y su cualificación completa, además del potencial para escalar y adaptar la solución a otras regiones del país.

##### **5.3 Alineación con políticas y estándares:**

- Objetivos de Desarrollo Sostenible de la ONU, específicamente el ODS 3 (Salud y Bienestar) y el ODS 6 (Agua Limpia y Saneamiento).
- Plan Decenal de Salud Pública de Colombia 2022-2031.
- Estándares de la Organización Mundial de la Salud (OMS) para la vigilancia de enfermedades transmitidas por el agua.
- Cumplimiento de estándares internacionales de manejo de datos de salud, como HIPAA y GDPR.

## 6. Gráfico de VOSViewer como soporte a la búsqueda de información

Consulta principal: ( ai OR "artificial intelligence" OR "machine learning" OR "deep learning" ) AND ( "diarrheal disease" OR "waterborne disease" ) AND ( prediction OR forecasting OR "outbreak detection" )



## 7. Referencias

- World Health Organization: WHO. (2024, 7 marzo). Diarrhoeal disease. <https://www.who.int/news-room/fact-sheets/detail/diarrhoeal-disease>
- Instituto Nacional de Salud de Colombia. (2024). Boletín Epidemiológico Semanal. <https://www.ins.gov.co/busador-eventos/Paginas/Vista-Boletin-Epidemilogico.aspx>
- Ligda P, Mittas N, Kyzas GZ, Claerebout E, Sotiraki S. Machine learning and explainable artificial intelligence for the prevention of waterborne cryptosporidiosis and giardiosis. Water Res. 2024 Sep 15;262:122110. doi: 10.1016/j.watres.2024.122110. Epub 2024 Jul 22. PMID: 39042970.
- Hussain M, Cifci MA, Sehar T, Nabi S, Cheikhrouhou O, Maqsood H, Ibrahim M, Mohammad F. Machine learning based efficient prediction of positive cases of waterborne diseases. BMC Med Inform Decis Mak. 2023 Jan 18;23(1):11. doi: 10.1186/s12911-022-02092-1. PMID: 36653779; PMCID: PMC9848024.
- Pras A, Mamane H. Nowcasting of fecal coliform presence using an artificial neural network. Environ Pollut. 2023 Jun 1;326:121484. doi: 10.1016/j.envpol.2023.121484. Epub 2023 Mar 21. PMID: 36958657.
- Petrau, A., & Sagarduy, M. (2024). Computer-implemented method for analysing water quality (WO2024156901A1). European Patent Office. <https://patentscope.wipo.int>
- Jutla, A. S., Usmani, M., & Colwell, R. (2024). Predictive models for infectious diseases (US2024029894A1). United States Patent and Trademark Office. <https://worldwide.espacenet.com>