Ciencia, arte, tradición, opinión, reflexión y meditación...

Artículo: ***Análisis de calidad del agua usando Ciencia de Datos***

Autor(es):
Eduardo Orozco Pérez
Miguel Félix Mata Rivera
Issis Claudette Romero Ibarra
Cristian Barria Huidobro

Publicación: **No. 3T, vol. 2025, pp. 58 - 73**

No. 3T
Vol. 2025

Análisis de calidad del agua usando Ciencia de Datos

Eduardo Orozco Pérez *eorozco@gmail.com*
Miguel Felix Mata Rivera *mmatar@ipn.mx*
Issis Claudette Romero Ibarra *iromero@ipn.mx*
Cristian Barria Huidobro *cristian.barria@umayor.cl*

El acceso a agua limpia y segura es un derecho humano fundamental. Hoy más que nunca, conocer su calidad es esencial para proteger la salud pública y tomar decisiones informadas sobre su uso. En ese contexto, diversas organizaciones internacionales –como la Organización Mundial de la Salud han desarrollado guías para evaluar y asegurar la calidad del agua que consumimos.

En muchos países, como China o México, se han realizado estudios que analizan la calidad del agua, tanto en fuentes naturales como en sistemas urbanos. Esos trabajos han demostrado que la contaminación puede tener múltiples causas: desde fallas en el mantenimiento de tuberías hasta la cercanía de cuerpos de agua con actividades humanas o industriales. Incluso factores naturales, como el tipo de vegetación o de suelo, pueden influir en los niveles de metales pesados y sales disueltas.

También se ha observado que el agua de lluvia –especialmente la recolectada en sistemas pluviales urbanos o para riego puede presentar variaciones importantes en parámetros como el pH, la turbidez o la concentración de sodio, lo que tiene consecuencias tanto para el ambiente como para nuestra salud.

Ante este panorama, algunos estudios recientes han comenzado a aplicar herramientas de ciencia de datos e inteligencia artificial para analizar de manera más profunda los registros disponibles. Gracias a esas tecnologías, hoy es posible:

- Recolectar y procesar muestras de agua de forma sistemática;
- Almacenar y compartir los datos a través de plataformas digitales;
- Aplicar modelos analíticos que detectan patrones y permiten hacer pronósticos.

El proyecto que vamos a describir se enfoca precisamente en ese tercer punto: aprovechar los datos históricos sobre la calidad del agua de lluvia recolectada en la Ciudad de México entre 2017 y 2019. Para ello, se analizaron parámetros como el pH, la turbidez, la dureza y las concentraciones de cloruros, hierro y manganeso. Mediante técnicas de agrupamiento de datos (clustering), se identificaron patrones tanto espaciales como temporales, y se logró vincularlos con factores socioeconómicos.

Además, se diseñaron visualizaciones interactivas y herramientas en línea que permiten a cualquier persona consultar nuestros hallazgos. Así, no solo se facilita la comprensión de la problemática, sino que también se abre la puerta a políticas públicas más justas, sobre todo en zonas donde el acceso a agua de calidad sigue siendo limitado.

Problemática

Los sistemas actuales de monitoreo de calidad del agua, en su mayoría, muestran únicamente datos en tiempo real recolectados por sensores. Es decir, indican si en ese momento el agua cumple con ciertos parámetros, pero no permiten observar cómo ha cambiado esa calidad con el paso del tiempo, ni cómo influyen factores como la ubicación geográfica o el contexto social.

Además, la manera en que se almacenan los datos suele depender del tipo de sensores utilizados, lo que puede dificultar su análisis a futuro. Por ello, es importante pensar en estructuras de datos claras, bien organizadas y compatibles con herramientas digitales modernas. También resulta clave contar con tableros visuales que no solo muestren el estado actual del agua, sino que permitan comparar su evolución en distintas regiones y a lo largo del tiempo.

Motivación

En la Ciudad de México existen datos históricos sobre la calidad del agua de lluvia, pero hasta ahora no se han aprovechado a fondo para observar su comportamiento en el tiempo y el espacio. Nuestra investigación busca responder preguntas clave como:

a) ¿Qué alcaldías presentan mejor o peor calidad de agua?
b) ¿Esa calidad ha cambiado con los años?
c) ¿Qué tan potables son esas aguas según los criterios oficiales de CONAGUA?

Contar con esa información permite identificar zonas de riesgo, tomar mejores decisiones sobre dónde recolectar agua de lluvia y evaluar cuándo es necesario aplicar un tratamiento adicional. Esos datos pueden ser de gran utilidad para autoridades, ciudadanía e investigadores que trabajan en soluciones sostenibles basadas en evidencia.

Vivimos en la era de los datos, lo cual ha abierto nuevas posibilidades para abordar problemas complejos como el que nos ocupa. Herramientas como la minería de datos y el aprendizaje automático nos permiten estudiar fenómenos naturales de manera más profunda y precisa. Nuestro proyecto aplica dichas tecnologías al caso del agua de lluvia, un recurso cada vez más valioso para las ciudades. Saber qué tan limpia es, cómo varía entre regiones y cómo ha cambiado a lo largo del tiempo es fundamental para diseñar mejores políticas públicas, proteger la salud comunitaria y planificar con inteligencia la instalación de sistemas de captación. Por ello, estudiar la calidad del agua de lluvia con herramientas de ciencia de datos no solo es posible, sino urgente y necesario.

Estado del arte: conocer y estudiar la calidad del agua

La calidad del agua es un tema clave para la salud pública y la sostenibilidad ambiental. Los estudios en este campo buscan entender cómo distintos

contaminantes –como pesticidas, metales pesados o residuos industriales afectan el agua en entornos urbanos y rurales. Para ello, se aplican análisis físicos, químicos y biológicos que permiten identificar estas sustancias en fuentes naturales o sistemas de distribución.

Uno de los riesgos más importantes se presenta cuando las aguas residuales o el agua de lluvia contaminada llegan a fuentes de agua potable. Por ejemplo, en el estudio de *Ospina-Zúñiga* y *Ramírez-Arcila* (2014), se utilizaron sensores, junto con algoritmos de aprendizaje automático, para detectar esos eventos en tiempo real. Se monitorearon variables como pH, turbidez, temperatura y fluorescencia, alcanzando una precisión superior al 97% al clasificar situaciones normales, de mantenimiento o de contaminación. A diferencia de ese enfoque predictivo, el presente proyecto se enfoca en el análisis histórico de datos, utilizando inteligencia artificial para clasificar y entender el estado actual del agua, bajo parámetros establecidos por norma en cada país o región. La inteligencia artificial también ha sido aplicada en contextos agrícolas. En *Das et al.* (2021), por ejemplo, se analizaron genotipos de trigo en Australia mediante imágenes térmicas obtenidas con drones, lo que permitió predecir el rendimiento de cultivos en suelos salinos usando modelos de *machine learning*.

Otro ejemplo interesante lo encontramos en el análisis de temperatura superficial en lagos, como en el estudio de *Yousefi & Toffolon* (2022), donde se evaluaron diferentes modelos de inteligencia artificial para predecir el calentamiento del agua. Aunque los resultados fueron prometedores, los autores subrayan la importancia de comprender las dinámicas naturales, ya que los modelos no pueden explicarlo todo por sí solos. Nuestra investigación adopta un enfoque similar: aplica modelos de clasificación para etiquetar muestras de agua de lluvia y detectar patrones relacionados con el tiempo y la ubicación geográfica, considerando siempre los factores físicos y contextuales que influyen.

En cuanto al monitoreo de lluvias, el estudio de *Wei, Lin* y *Zheng* (2025) aplicó redes neuronales artificiales y algoritmos de bosque aleatorio para clasificar datos obtenidos por satélites y radares. Los modelos desarrollados lograron detectar con alta precisión las zonas con presencia de lluvia, lo que demuestra el

potencial del aprendizaje automático en el análisis climático. También existen investigaciones centradas en fuentes no convencionales de agua, como la lluvia o incluso la captación de niebla. En Sheng et al. (2018), se diseñó un sistema inalámbrico para evaluar la calidad del agua en bebederos urbanos, midiendo parámetros como el pH, la temperatura y el oxígeno disuelto, con el fin de determinar si el agua era segura para el consumo humano. De forma complementaria, en el estudio de *Ospina-Zúñiga & Ramírez-Arcila* (2014), se recolectaron muestras de agua de lluvia directamente del ambiente, evitando el contacto con superficies contaminadas, y luego se analizaron en laboratorio para evaluar su aptitud para uso doméstico.

A nivel metodológico, también se ha investigado cómo organizar y procesar grandes volúmenes de datos ambientales. El estudio de *Chapman et al.* (2000) analizan el uso del modelo CRISP-DM en ciencia de datos, una metodología que guía paso a paso la recolección, organización, análisis e interpretación de datos. Ese tipo de estructura es especialmente útil en proyectos como el nuestro, donde se manejan grandes bases de datos históricos sobre la calidad del agua, permitiendo reducir la complejidad técnica y mantener el foco en la solución del problema.

Además, existen iniciativas enfocadas en la comunicación y accesibilidad de los datos. En Liu et al. (2023) se desarrollaron herramientas avanzadas de visualización espaciotemporal de alta dimensión para analizar la calidad del aire, un enfoque moderno que podría aplicarse también al análisis del agua. En *Hognogi et al.* (2023) se desarrolló una aplicación móvil que permite a los ciudadanos reportar problemas relacionados con la calidad del agua, brindando retroalimentación directa a las autoridades. Por su parte, el estudio de Xu et al. (2024 diseñó herramientas interactivas, como mapas, gráficos 3D y nubes de puntos para visualizar cambios en la calidad del agua a lo largo del tiempo. Este tipo de soluciones facilita la comprensión de datos complejos tanto para investigadores como para tomadores de decisiones y el público en general.

Finalmente, el trabajo de *Ribeiro* y *Reynoso-Meza* (2018) demuestra cómo es posible detectar anomalías en tiempo real utilizando modelos de *machine learning* aplicados a datos provenientes de sensores. Esa tecnología podría

aplicarse en sistemas de captación de agua de lluvia, generando alertas automáticas ante posibles eventos de contaminación.

En resumen, el análisis de la calidad del agua ha evolucionado significativamente gracias a los avances en sensores, algoritmos, plataformas digitales y herramientas de visualización. Desde el monitoreo satelital hasta aplicaciones ciudadanas, pasando por técnicas de inteligencia artificial, la tendencia es clara: hoy entendemos mejor cómo se comporta el agua, y eso nos da más herramientas para cuidarla y gestionarla de forma equitativa y sustentable.

Metodología: análisis de datos sobre la calidad del agua de lluvia

Este estudio empleó herramientas de ciencia de datos para analizar la calidad del agua de lluvia en la Ciudad de México entre 2017 y 2019. El proceso se organizó en seis etapas clave que permitieron recopilar, procesar y visualizar la información de forma clara y accesible.

Diseño metodológico y recolección de datos

Se adaptó una metodología general de análisis de datos , como se ilustra en la Figura 1, que guió cada etapa del estudio, desde la obtención de datos hasta la visualización final. Las fuentes de información incluyeron instituciones confiables como SACMEX, *Data.world*, el Observatorio Hidrológico y el INEGI. Se recolectaron datos fisicoquímicos del agua de lluvia, así como datos sobre precipitación y características demográficas. Todo se integró en una base híbrida con variables como pH, turbidez, cloruros, hierro, manganeso y dureza total.

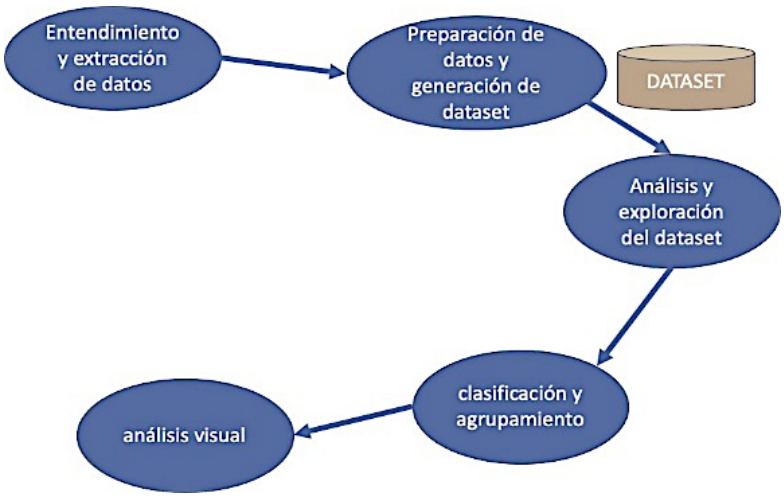

Figura 1. Metodología de datos

Preparación de los datos

Una vez recopilados, los datos fueron filtrados, organizados y estructurados por alcaldía, mes y año. Para garantizar una representación adecuada, se aplicó un muestreo aleatorio simple con detalle mensual. Los valores límite establecidos por normativas oficiales fueron reunidos en la Tabla 1, que se usó como referencia para etiquetar automáticamente los registros según su nivel de calidad.

Variable	Valor permisible
pH	6.5 a 8.5 (unidad)
Turbiedad	4.0 (UNT)
Dureza	500 (mg/l)
Cloruros	250 (mg/l)

Tabla 1. Valores límite de referencia para la clasificación de la calidad del agua acorde a la normativa

Clasificación con aprendizaje automático

Se entrenaron tres modelos de inteligencia artificial para clasificar la calidad del agua en tres categorías: excelente, aceptable y contaminada. El algoritmo de Bosque Aleatorio fue el más preciso, alcanzando una tasa de acierto del 100 %

en muestras excelentes y del 74 % en muestras contaminadas. Gracias a este modelo, se generó un conjunto de datos clasificado y listo para su análisis visual y geográfico.

Agrupamiento (clustering)

Para detectar patrones ocultos, se utilizó el algoritmo *K-Means*, que identificó cinco grupos con comportamientos similares. La cantidad óptima de grupos se determinó mediante el método del codo. Los resultados se visualizaron en un cubo tridimensional que muestra cómo varía la calidad del agua por alcaldía, año y mes (Figura 2).

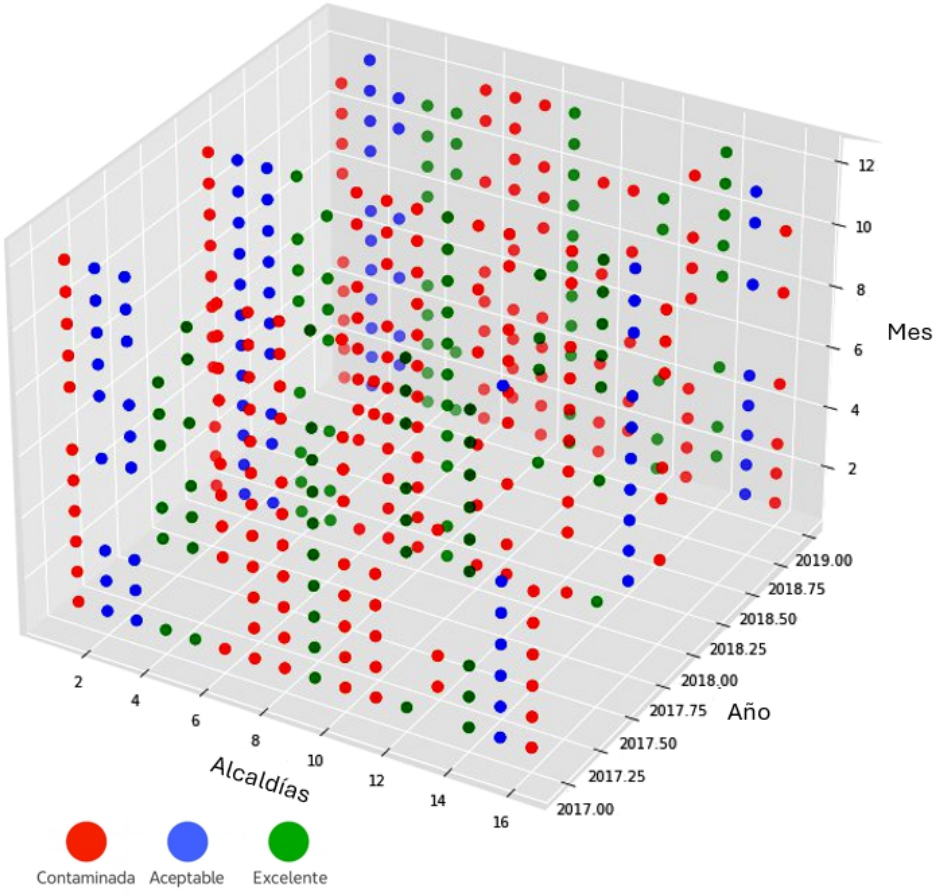

Figura 2. Cubo tridimensional de lecturas de calidad por alcaldía, año y mes

Análisis espacio-temporal

Se detectaron concentraciones específicas de agua contaminada en ciertas alcaldías y meses. Por ejemplo, se observaron cambios notables en los valores de pH y cloruros entre distintos años (Figura 3), mientras que el hierro presentó concentraciones más altas durante 2018 (Figura 4).

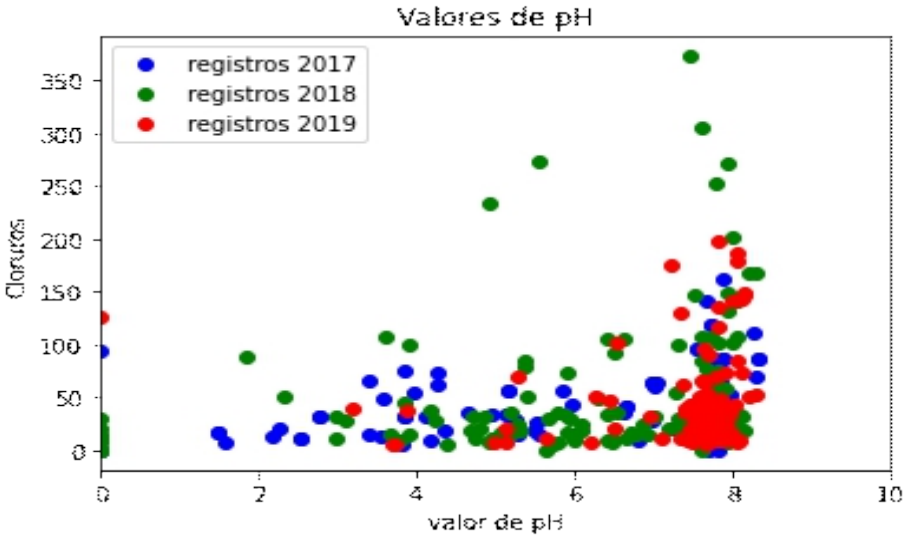

Figura 3. Exploración de datos de pH y cloruros en distintas alcaldías: distribución centralizada sin correlación aparente

Mientras que el análisis del hierro mostró concentraciones elevadas en 2018 (Figura 4).

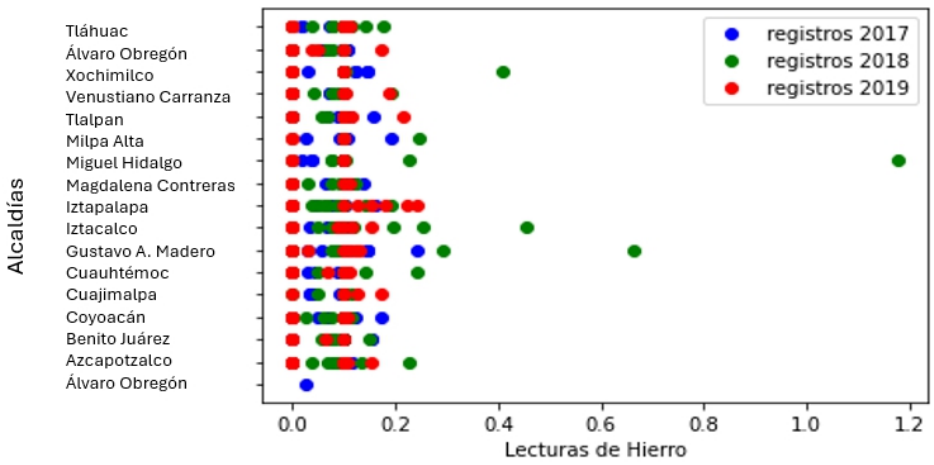

Figura 4. Evolución de los valores de hierro por año en un periodo trianual

Relación con datos socioeconómicos

Se incorporaron variables del INEGI, como la densidad de negocios y la actividad económica (número de unidades económicas registradas en el DENUE por alcaldía). El análisis reveló que, en algunas alcaldías, el crecimiento económico coincidió con una disminución en la calidad del agua. En la Figura 5 se muestran las alcaldías ordenadas de mayor a menor actividad económica: las barras indican el número de negocios por alcaldía y la línea naranja representa el porcentaje acumulado, lo que sugiere una posible relación entre el desarrollo urbano y la presión sobre los recursos hídricos.

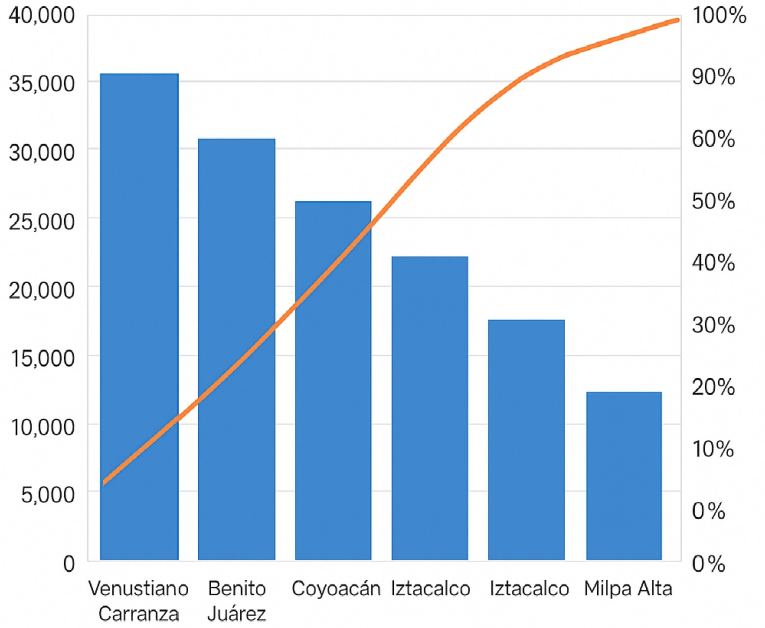

Figura 5. Aumento de la actividad económica en alcaldías durante el periodo de estudio

Resultados y visualización del análisis de calidad del agua de lluvia en CDMX

El estudio logró visualizar cómo varían los parámetros de calidad del agua de lluvia a lo largo del tiempo y entre alcaldías. Esos patrones sugieren que factores como la densidad poblacional, la ubicación geográfica y el desarrollo económico pueden influir directamente en la calidad del agua recolectada.

Almacenamiento y organización de los datos

Los datos recolectados se integraron en una base de datos local en *MySQL*, lo que facilitó su análisis y el entrenamiento de algoritmos. Esta base fue la fuente principal para los experimentos clasificatorios y visuales.

Algoritmo clasificador más eficiente

Se probaron tres algoritmos de clasificación para categorizar la calidad del agua: *Naive Bayes*, Red Neuronal y Bosque Aleatorio (*Random Forest*). Este último demostró el mejor rendimiento, clasificando correctamente la mayoría de los registros en las tres categorías: contaminada, aceptable y excelente . Con base en este algoritmo, se generó un conjunto clasificado de datos .

Análisis espacio-temporal y agrupamiento

Utilizando el cubo de datos 3D (Figura 2), se analizó la distribución mensual por alcaldía. Posteriormente, el algoritmo K-Means permitió agrupar las alcaldías según similitudes en el comportamiento de la calidad del agua (contaminada, aceptable y excelente de acuerdo con la leyenda de colores). Se identificaron cuatro grupos (identificados con una estrella) destacados con variaciones relevantes a lo largo del tiempo (Figura 6).

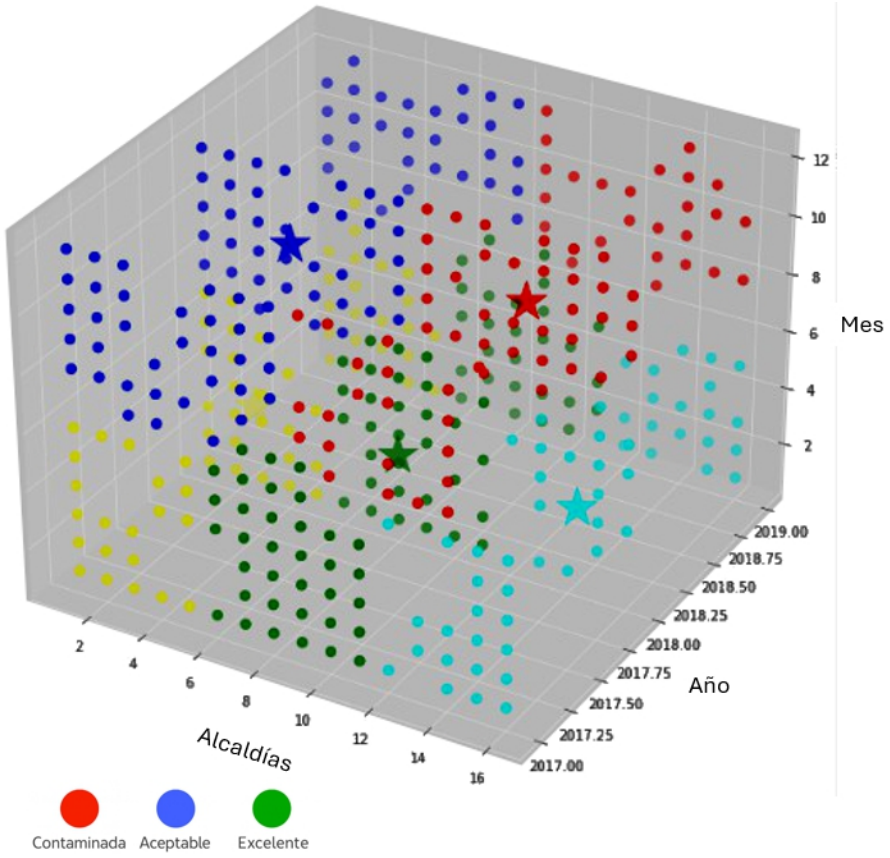

Figura 6. Visualización tridimensional de grupos y centroides en el cubo de datos

Integración con datos socioeconómicos

Al incorporar datos del DENUE (Directorio Nacional de Unidades Económicas) del INEGI, se evidenció una correlación entre la cantidad de negocios por alcaldía y los cambios en la calidad del agua, como se mostró previamente en la Figura 5.

Visualización interactiva con mapas web

Todos los resultados se integraron en un visor interactivo accesible en línea. Los usuarios pueden explorar la calidad del agua por alcaldía y consultar datos económicos relacionados. Cada punto del mapa incluye ventanas emergentes con información detallada (Figura 7), además de gráficos que explican los hallazgos para facilitar su comprensión por parte del público general.

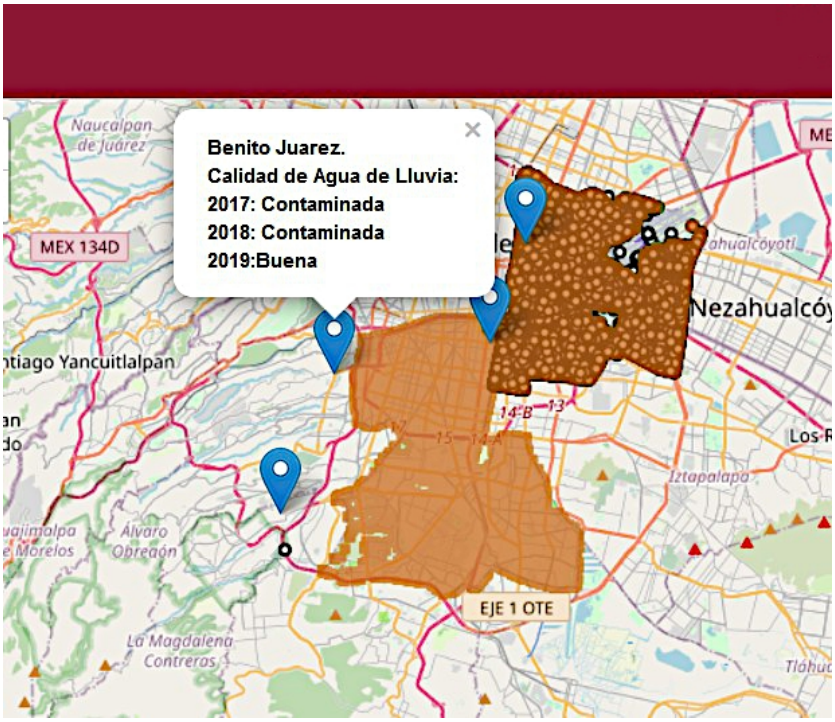

Figura 7. Ventana emergente con información del agua

Es así que los resultados obtenidos revelan que, la calidad del agua de lluvia en la Ciudad de México no es homogénea, sino que está influenciada por múltiples factores sociales, económicos y geográficos. La integración de modelos de clasificación, análisis espacio-temporal y visualización interactiva permitió transformar grandes volúmenes de datos en conocimiento accesible y útil. Este enfoque no solo facilita la comprensión de las dinámicas territoriales del agua, sino que también abre nuevas posibilidades para la toma de decisiones basada en evidencia, orientada a la equidad hídrica y la sostenibilidad urbana.

Conclusiones y trabajo futuro

Esta investigación ejemplifica el valor de aplicar herramientas de ciencia de datos e inteligencia artificial en el análisis de la calidad del agua de lluvia en contextos urbanos complejos como el de la Ciudad de México. A partir de datos abiertos recolectados entre 2017 y 2019, se identificaron patrones espacio-temporales y geosociales que revelan importantes desigualdades en el acceso a agua de calidad. Las alcaldías con mayores niveles de vulnerabilidad social

tienden a registrar una peor calidad del agua recolectada, lo cual plantea desafíos urgentes en términos de equidad ambiental y salud pública.

El uso de modelos de clasificación y agrupamiento permitió organizar y etiquetar grandes volúmenes de datos, generando un conjunto limpio y estructurado que puede ser reutilizado por otros investigadores. Además, la implementación de herramientas visuales e interactivas –como mapas web y cubos de datos– facilitó la interpretación de los resultados, tanto para la ciudadanía como para los tomadores de decisiones.

Es así como este enfoque ofrece una base para el diseño de políticas públicas orientadas a la captación de agua de lluvia más segura.

Entre las líneas de trabajo futuras derivadas de esta investigación se proponen las siguientes:

- Ampliación temporal y geográfica del estudio: Integrar datos posteriores a 2019 e incluir otras zonas metropolitanas del país para contrastar resultados y validar patrones comunes o divergentes.

- Monitoreo en tiempo real: Incorporar sensores IoT conectados a redes de baja potencia (como LoRa o Sigfox) para complementar el análisis histórico con datos en vivo, generando alertas tempranas ante eventos de contaminación.

- Modelos predictivos de calidad del agua: Desarrollar modelos de pronóstico usando aprendizaje automático supervisado, tomando en cuenta variables ambientales y socioeconómicas para anticipar riesgos por alcaldía y temporada.

- Integración con salud pública: Estudiar correlaciones entre la calidad del agua recolectada y variables epidemiológicas, con el fin de evaluar impactos directos sobre la salud comunitaria.

- Educación y participación ciudadana: Crear versiones simplificadas del visor *web* y *apps* móviles para que ciudadanos puedan consultar datos locales, reportar anomalías y participar activamente en la vigilancia de la calidad del agua.

- Sistemas de recomendación para captación: Diseñar un sistema que sugiera las mejores zonas, temporadas y métodos para recolectar agua de lluvia segura, tomando en cuenta las condiciones históricas y los modelos predictivos.

Agradecimientos

Los autores agradecen el apoyo de la Secretaría de Investigación y Posgrado (SIP) del Instituto Politécnico Nacional a través de los proyectos SIP 20253692, SIP 20250343 y SIP 20254771, los cuales hicieron posible el desarrollo de esta investigación. Asimismo, a la COFAA del IPN y la SECIHTI, por su apoyo.

Para conocer más, consulta:

1) Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T. P., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. The CRISP-DM Consortium.
2) Das, S., Christopher, J., Apan, A., Choudhury, M. R., Chapman, S., Menzies, N. W., & Dang, Y. P. (2021). *Evaluation of water status of wheat genotypes to aid prediction of yield on sodic soils using UAV-thermal imaging and machine learning*. Agricultural and Forest Meteorology, 307, 108477. https://doi.org/10.1016/j.agrformet.2021.108477
3) Hognogi, G. G., et al. (2023). *The role of citizen science mobile apps in facilitating a digital approach to water quality monitoring.* [Journal Name]. https://doi.org/10.1038/...
4) Liu, J., Wan, G., Liu, W. *et al.* High-dimensional spatiotemporal visual analysis of the air quality in China. *Sci Rep* 13, 5462 (2023). https://doi.org/10.1038/s41598-023-31645-1
5) Ospina-Zúñiga, Ó. E., & Ramírez-Arcila, H. (2014). *Evaluación de la calidad del agua de lluvia para su aprovechamiento y uso doméstico en Ibagué, Tolima, Colombia.* Ingeniería Solidaria, 10(17), 125–138. https://doi.org/10.16925/in.v9i17.812
6) Ribeiro, V. H. A., & Reynoso-Meza, G. (2018). *Online anomaly detection for drinking*

water quality using a multi-objective machine learning approach. In *Genetic and Evolutionary Computation Conference Companion (GECCO '18 Companion)*. https://doi.org/10.1145/3205651.3208202

7) Sheng, J., Weixing, W., Jieping, Y., & Zhongqiang, H. (2018). *Design of a WSN system for monitoring the safety of drinking water quality*. IFAC PapersOnLine, 17, 752–757.
8) Yousefi, A., & Toffolon, M. (2022). *Critical factors for the use of machine learning to predict lake surface water temperature*. Journal of Hydrology, 127418. https://doi.org/10.1016/j.jhydrol.2022.127418
9) Wei, N., Lin, Y., & Zheng, H. (2025). Prediction of the flood distribution caused by returning cropland to forest based on Generative Adversarial Network and multi-source remote sensing data. International Journal of Applied Earth Observation and Geoinformation, 143, 104790. https://doi.org/10.1016/j.jag.2025.104790
10) Xu, Y., Hui, M., & Qu, H. (2024). Design of a 3D Platform for the Evaluation of Water Quality in Urban Rivers Based on a Digital Twin Model. *Water*, *16*(24), 3668. https://doi.org/10.3390/w16243668