TÍTULO: "HALLAZGO DE PATRONES DE DATOS EN MUESTRAS DE AGUA DEL RÍO DE LA RECONQUISTA"

Ferrán, Natanael Emir [a]; Pastorini, Martín Ezequiel [a]; López, Carlos Gustavo [a]; Fernández, Enrique [a]; Esperanza, Marcelo Fabián [a]
a Instituto de Ingenierías y Nuevas Tecnologías - Universidad Nacional del Oeste
nferran@uno.edu.ar

Resumen

Este artículo se centra en el análisis de los datos obtenidos en el marco del trabajo "Análisis Fisicoquímico del Río de la Reconquista", presentado por este mismo equipo de investigación, con el objetivo de identificar patrones que permitan comprender el comportamiento y la situación actual del Río de la Reconquista. Explica el tratamiento del dataset utilizado y el algoritmo de inteligencia artificial empleado. Explica los resultados obtenidos y las hipótesis generadas. En última instancia, concluye con una reflexión sobre las implicancias de los hallazgos producidos y se sugieren áreas para futuras investigaciones.

Abstract

This article focuses on the analysis of data obtained as part of the work "Physicochemical Analysis of the Reconquista River", presented by the same research team, with the aim of identifying patterns that allow us to understand the behaviour and current situation of the Reconquista River. It explains the treatment of the dataset used and the artificial intelligence algorithm applied. It explains the results obtained and the hypothesis generated. Finally, it concludes with a reflection on the implications of the findings and suggests areas for future research.

***Palabras clave**:* Agua, Fisicoquímica, Inteligencia Artificial, Clustering.

INTRODUCCIÓN

La cuenca del Río de la Reconquista abarca un total de 82 km, desde su naciente en la confluencia de los arroyos La Choza y El Durazno, hasta su desembocadura en el Río Luján. En la figura 1 podemos visualizar su alcance.

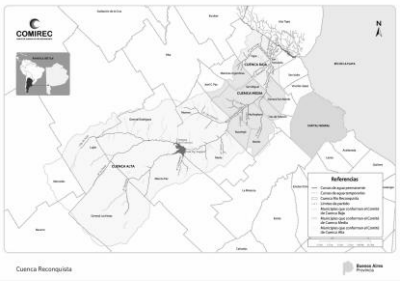

Figura 1. Mapa provisto por la COMIREC con las delimitaciones de la cuenca del Río de la Reconquista.

Existen numerosas industrias cuyos residuos son descargados en el cauce del río o de sus afluentes. Hoy en día es la segunda cuenca más contaminada de Buenos Aires (solo superada por la cuenca Matanza-Riachuelo) [1], [2]. En la figura 2 se muestra una imagen que intenta sintetizar el estado del Río de la Reconquista.

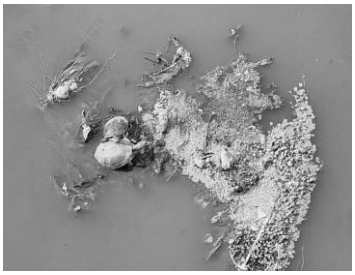

Figura 2. Tortugas sobre bolsas de consorcio y otros residuos.

A fin de tomar dimensiones de la problemática en cuestión, se menciona que según el censo nacional llevado a cabo en el año 2022, existe una población total de más de 1.000.000 de personas que viven entre Merlo y Moreno [3]. Estas personas se exponen día a día, directa o indirectamente, a condiciones preocupantes de insalubridad.

Este trabajo forma parte del proyecto de investigación y desarrollo denominado *"Elaboración de un modelo para la determinación del Índice de Calidad del Agua (ICA) en aguas superficiales de la cuenca alta del río Reconquista en el partido de Merlo con desarrollo e implementación de una herramienta de Inteligencia Artificial"* que lleva adelante el Instituto de Ingenierías y Nuevas Tecnologías de la Universidad Nacional del Oeste.

OBJETIVOS

Analizar el comportamiento y la situación actual de un tramo de la cuenca alta del río de la Reconquista, así como de los arroyos que desembocan en este, utilizando herramientas estadísticas y de inteligencia artificial para buscar patrones o relaciones entre los datos de variables fisicoquímicas de muestras de agua.

METODOLOGÍA

Desde fines de mayo del año 2023, este mismo equipo de investigación ha estado recolectando muestras, analizando de cada una de ellas un total de 10 parámetros fisicoquímicos con métodos validados, obteniendo un total de 220 datos al día de la fecha. Esta labor ha sido presentada bajo el nombre *"Análisis del Río de la Reconquista"* en la revista Ingeniería Sanitaria y Ambiental [4].

Este artículo, se centra en la utilización de técnicas de aprendizaje no supervisado para obtener información relevante basada en el análisis de dichos datos.

El aprendizaje no supervisado nos permite descubrir estructuras ocultas en los datos. A diferencia del aprendizaje supervisado, en este caso no trabajamos con datos etiquetados para entrenar los modelos, ya que el objetivo no es predecir la clase de salida [5]. De esta manera, al no contar con datos etiquetados, sólo podemos descubrir los patrones que se producen de forma natural en el conjunto de datos. Una de las principales técnicas del aprendizaje no supervisado es el *clustering*. El objetivo de este método consiste en encontrar grupos de instancias (llamados *clusters*) que están relacionados entre sí. Esta técnica tiene innumerables aplicaciones, como detección de *outliers*, segmentación de clientes o sistemas de recomendación. Hay diferentes tipos de algoritmos de clustering y cada uno de ellos identifica los clusters de manera distinta. En este artículo se utiliza el popular algoritmo *k-means*. Sus pasos son los siguientes:

1. Inicialización: se eligen *k* centroides iniciales de los datos, ya sea seleccionándolos aleatoriamente o usando técnicas específicas para elegirlos.
2. Asignación de datos a centroides: cada dato se asocia con el centroide más cercano, lo que implica agruparlo con la opción cuya distancia (usualmente euclidiana) sea la mínima.
3. Actualización de centroides: luego de que los datos se han distribuido entre los diferentes clusters, los centroides se ajustan recalculando la media de todos los datos en cada grupo. Esto posiciona los centroides en nuevas ubicaciones basadas en los datos agrupados.
4. Iteración: Los pasos 2 y 3 se repiten hasta que la distribución de los datos ya no cambia significativamente o se alcanza un límite de iteraciones.

El objetivo de k-means es minimizar la variabilidad intra-cluster y maximizar la variabilidad inter-cluster, logrando así un agrupamiento significativo de los datos. Sin embargo, puede ser sensible a la elección de los centroides iniciales y a veces puede quedar atrapado en óptimos locales [6].

El escalamiento es un paso crucial previo a la aplicación del algoritmo k-means [7]. Esta técnica consiste en normalizar las características de los datos para que tengan una escala uniforme y comparable. Al aplicar K-means a datos no escalados, las características con rangos más amplios pueden dominar el proceso de agrupamiento, lo que podría llevar a resultados sesgados o inexactos. Escalar los datos garantiza que todas las características tengan un peso equitativo en el análisis, permitiendo que el algoritmo identifique patrones y estructuras intrínsecas de manera más precisa. Esto mejora la eficiencia y eficacia del algoritmo, contribuyendo a la obtención de clusters más representativos.

La técnica utilizada para escalar los datos, es la denominada normalización *z-score*, cuya aplicación consiste en la ecuación 1:

$$z = \frac{x-\mu}{\sigma} \tag{1}$$

Donde:

- z: valor normalizado (z-score)
- x: valor sin normalizar
- μ: valor de la media
- σ: desviación estándar

En cuanto a la implementación del algoritmo k-means,
los siguientes parámetros fueron seleccionados:

- Un valor *k* igual a 3.
- El método de inicialización *k-means++*, un método de selección inteligente de centroides iniciales [8].

RESULTADOS Y DISCUSIÓN

Luego de correr el algoritmo, se obtuvo lo que se muestra en la figura 3.

Lo que se observa son tres clústers de datos. A continuación se hace una discusión de cada uno de ellos:

- El de color rojo, contiene puntos en un rango de pH más bajo que el resto y en un rango de oxígeno disuelto (OD) también bajo. Es por ello que se decidió denominarlo como "muestras ácidas con pobre nivel de O2".
- El de color verde, contiene puntos en un rango de pH más alto que el resto y en un rango de de OD también alto. Se decidió denominarlo "Muestras alcalinas con buen nivel de O2".
- El de color amarillo, se lo considera un intermedio entre los dos anteriores. Contiene puntos en un rango de pH intermedio, y lo mismo ocurre con el rango de OD. Se decidió denominarlo "Muestras neutras con pobre nivel de O2".

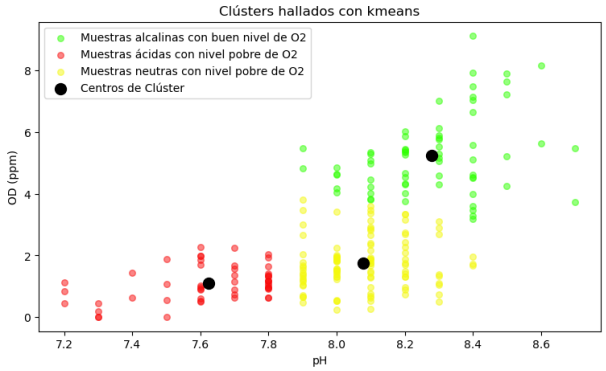

Figura 3. Resultados del algoritmo de clustering aplicado al diagrama de dispersión OD vs. pH.

La bibliografía indica que la concentración mínima de oxígeno disuelto para que la vida acuática pueda desarrollarse en un cuerpo de agua, es de 4 partes por millón, mientras que valores menores a este número indican un problema grave [9]. Notar que la mayoría de los puntos contenidos en el clúster de color verde tienen una concentración de OD mayor al número mencionado.

A continuación, se analizó el mismo diagrama de dispersión discriminando los valores por punto de muestreo (figura 4) y por mes en que se realizó el muestreo (figura 5).

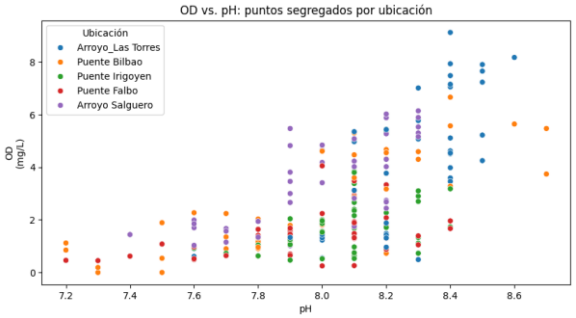

Figura 4. Diagrama de dispersión de OD vs. pH con valores segregados por punto de muestreo.

En la figura 4 se observa una densidad de puntos celestes en la esquina superior derecha del gráfico, es decir, en rangos altos tanto de pH como de OD. Estos puntos pertenecen al arroyo Las Torres, lo cual podría tener relación con algún tipo de tratamiento aguas arriba. Por otro lado, se observa una densidad de puntos rojos en rangos bajos de OD, los cuales pertenecen al Puente Falbo. Esto último podría tener relación con el efecto de diferentes residuos que pudieran ser arrojados aguas arriba.

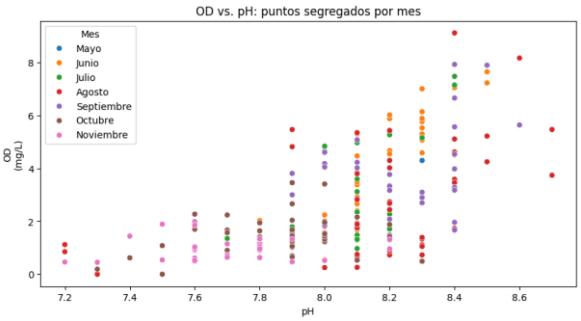

Figura 5. Diagrama de dispersión de OD vs. pH con valores segregados por el mes del muestreo.

En cuanto a la figura 5, se observa una densidad de puntos marrones y rosas en un bajo rango de OD, correspondientes a los meses de octubre y noviembre, respectivamente. Se trata de los meses más calurosos en que se han realizado muestreos. Por otra parte, se observa una densidad de puntos naranjas en un rango alto de OD y pH, los cuales corresponden al mes de Junio, uno de los meses más fríos en que se han realizado muestreos. Este último análisis podría sugerir una correlación entre la calidad del agua y la estación del año.

CONCLUSIONES

A pesar de que el proyecto se encuentra aún en un estadío inicial y la masa de datos es aún pequeña, los resultados presentados demuestran que es posible generar hipótesis y sacar ciertas conclusiones acerca del comportamiento y la situación actual del Río de la Reconquista, así como también de los arroyos que desembocan en este, cumpliendo el objetivo propuesto inicialmente. Estas hipótesis y conclusiones ganarán robustez estadística con mayor cantidad de datos, ya sea ampliando la cantidad de parámetros medidos, y/o aumentando la frecuencia de muestreos. Se busca que ambas cuestiones sean abordadas en el futuro por el equipo de investigación.

REFERENCIAS

[1] Nader, G.M. (2015). *Evaluación de la calidad del agua en un río urbano. Universidad Nacional de San Martín. Instituto de Investigaciones e Ingeniería Ambiental.*

[2] Castilla V., Canevaro, S., López, B. (2021). *Migración, degradación ambiental y percepciones del riesgo en la cuenca del río (Buenos Aires, Argentina), Revista de Estudios Sociales,* (76).

[3] *Indicadores demográficos, por sexo y edad* (2023). *Instituto Nacional de Estadística y Censos.*

[4] Ferrán, N.E., Pastorini, M.E., López, C.G., Fernández, E., Esperanza, M.F. (2023). *Análisis del Río de la Reconquista. Ingeniería Sanitaria y Ambiental*, (149), 36-38.

[5] Mohammed J. Zaki, Wagner Meira, Jr., (2020). *Data Mining and Machine Learning: Fundamental Concepts and Algorithms, Cambridge University Press*, (2).

[6] Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal, (2016) *Data Mining: Practical Machine Learning Tools and Techniques, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann.*

[7] Raschka S. (2014). *About Feature Scaling and Normalization and the effect of standardization for machine learning algorithms, Sebastian Raschka.*

[8] Vassilvitskii, S., Arthur, D. (2007). *k-means++: the advantages of careful seeding, Stanford University.*
[9] Chang J., Pulla E. (2007). *Calidad de agua: trabajo de investigación, oxígeno disuelto (OD).*