Universidad Nacional del Centro de la
Provincia de Buenos Aires

FACULTAD DE CIENCIAS EXACTAS

Ingeniería de sistemas

Trabajo Práctico especial

Fundamentos de la ciencia de datos

GRUPO 14
Valentino Courtil: vcourtil@alumnos.exa.unicen.edu.ar
Ignacio Gonzalez: igonzalez@alumnos.exa.unicen.edu.ar
Ramiro Ortiz: rortiz@alumnos.exa.unicen.edu.ar

ÍNDICE

Introduccion

En este informe se analizan mediciones del agua en diferentes sitios del Río de la Plata en el año 2022, teniendo como meta comprender la relación entre el ICA (Indice de Calidad del Agua) y variables ambientales y biológicas como la turbidez, concentración de bacterias, oxígeno disuelto en el agua, nutrientes, entre otras.

Para el desarrollo se utilizan técnicas de análisis estadístico y ciencia de datos para identificar los factores que más influyen en la calidad del agua, proporcionando así una mejor comprensión de los patrones de contaminación.

Este análisis incluyó la limpieza y preparación de los datos, planteo de hipótesis de interés y su comprobación mediante el uso de pruebas estadísticas.

Materiales

El desarrollo de este estudio se realizó a partir de un conjunto de mediciones recolectadas en diversas estaciones y lugares de muestreo del Río de La Plata durante el 2022. Este conjunto de datos incluye una serie de parámetros físicos, químicos y biológicos que señalan la calidad del agua: oxígeno disuelto, pH, turbidez, concentración de bacterias fecales, y niveles de nutrientes como fósforo total, amonio, nitratos y fosfatos. Estas variables brindan la posibilidad de evaluar la condición general del agua.

Descripción de nuestras variables

1. sitios: Localización específica donde se realizó el muestreo del agua.
2. codigo: Identificador único para cada muestra o estación de muestreo.
3. fecha: Fecha en la que se tomó la muestra de agua.
4. año: Año en que se realizó el muestreo.
5. campaña: Nombre o número de la campaña de monitoreo en la que se realizó el muestreo.
6. tem agua: Temperatura del agua en grados Celsius.
7. tem aire: Temperatura del aire en grados Celsius.
8. od: Oxígeno disuelto, medido en miligramos por litro (mg/L), esencial para la vida acuática.
9. ph: Medida de la acidez o alcalinidad del agua, en una escala de 0 a 14.
10. olores: Presencia de olores en el agua, que puede indicar contaminación.
11. color: Color del agua, que puede ser un indicador de la calidad del agua.
12. espumas: Presencia de espumas en la superficie del agua, que puede ser un signo de contaminación.
13. mat susp: Materia suspendida, que se refiere a partículas sólidas que flotan en el agua.
14. colif fecales ufc 100ml: Unidades formadoras de colonias de coliformes fecales en 100 ml de agua, un indicador de contaminación fecal.
15. escher coli ufc 100ml: Unidades formadoras de colonias de Escherichia coli en 100 ml de agua, otro indicador de contaminación fecal.
16. enteroc ufc 100ml: Unidades formadoras de colonias de enterococos en 100 ml de agua, que también indican contaminación fecal.

17. nitrato mg l: Concentración de nitratos en miligramos por litro (mg/L), que puede indicar contaminación por fertilizantes.
18. nh4 mg l: Concentración de amonio en miligramos por litro (mg/L), que puede ser un indicador de contaminación orgánica
19. p total l mg l: Fósforo total en miligramos por litro (mg/L), que incluye todas las formas de fósforo en el agua.
20. fosf ortofos mg l: Concentración de ortofosfatos en miligramos por litro (mg/L), que es un nutriente importante.
21. dbo mg l: Demanda biológica de oxígeno en miligramos por litro (mg/L), que mide la cantidad de oxígeno requerido por microorganismos para descomponer materia orgánica.
22. dqo mg l: Demanda química de oxígeno en miligramos por litro (mg/L), que mide la cantidad total de oxígeno requerido para oxidar materia orgánica e inorgánica.
23. turbiedad ntu: Turbidez del agua medida en unidades NTU (Nephelometric Turbidity Units), que indica la claridad del agua.
24. hidr deriv petr ug l: Hidrocarburos derivados del petróleo en microgramos por litro (µg/L), que indican contaminación por productos petroleros.
25. cr total mg l: Concentración total de cromo en miligramos por litro (mg/L), un metal pesado que puede ser tóxico.
26. cd total mg l: Concentración total de cadmio en miligramos por litro (mg/L), otro metal pesado que es tóxico en altas concentraciones.
27. clorofila a ug l: Concentración de clorofila a en microgramos por litro (µg/L), que indica la cantidad de fitoplancton en el agua.
28. microcistina ug l: Concentración de microcistinas en microgramos por litro (µg/L) que son toxinas producidas por ciertas algas.
29. Índice de calidad del agua, que puede ser un valor calculado para evaluar la calidad general del agua.
30. calidad de agua: Clasificación general de la calidad del agua basada en los parámetros medidos

Preparación de los datos

Durante la fase inicial del estudio, se llevaron a cabo actividades de preparación de datos para garantizar que el conjunto de datos esté preparado para el análisis estadístico. A continuación se explica el procedimiento de limpieza y transformación de los datos:

Limpieza de los datos

- **Eliminación de valores faltantes:** Valores como "no se midió" fueron reemplazados por NaN para facilitar su tratamiento.
- **Revisión de la consistencia de los datos:** Mediciones cómo “< 0.10” fueron reemplazadas por valores numéricos.
- **Eliminación de columnas:** Se eliminaron las siguientes columnas del conjunto de datos:
    - Orden: El significado de esta columna no fue proporcionado.
    - Año: Todas las mediciones fueron tomadas el mismo año.
    - Fecha: El análisis fue realizado tomando en cuenta las estaciones del año, en lugar de fechas específicas.
    - Columnas referentes al cadmio e hidrocarburos: Presentan gran cantidad de datos repetidos y faltantes.
    - Código y sitios: El análisis tiene como objetivo el estudio del Río de la Plata en su integridad, por lo que los lugares de medición y sus identificadores no resultaron relevantes.
    - Calidad del agua: Se consideró al ICA como único indicador de la calidad del agua.
- **Imputación de datos faltantes:** Los valores faltantes en el conjunto de datos fueron imputados empleando las medias de los datos de los cinco vecinos más cercanos.

Transformación de variables

Para facilitar el análisis y hacer los datos más comparables entre sí, se realizaron algunas transformaciones de las variables:

- **Conversión de tipos de datos:** Se realizó la conversión de las columnas del tipo “object” a su tipo de dato correspondiente dependiendo del contexto.
- **One-Hot Encoding:** Se aplicó la técnica one-hot encoding a la variable campaña para representar las estaciones del año como cuatro columnas binarias.

Hipótesis propuestas

El enfoque principal de este análisis se centra en la calidad del agua, medida a través del índice de calidad del agua (ICA), que se seleccionó como variable objetivo en el desarrollo de las hipótesis.

Hipótesis 1: La calidad del agua se ve afectada por la presencia de bacterias fecales:
Con esta hipótesis se busca descubrir si la presencia de bacterias fecales en el agua genera un impacto sobre la calidad de la misma. Dado que la presencia de bacterias fecales es un indicador de contaminación de origen orgánico, se espera que altos niveles de estas bacterias tengan un impacto negativo en el ICA.

Hipótesis 2: La concentración de bacterias fecales varía según la estación del año:
En esta hipótesis se tiene el objetivo de analizar si varía la concentración de bacterias entre las estaciones del año y el efecto que esto puede tener sobre la calidad del agua.

Hipótesis 3: El índice de calidad de agua varía según la turbidez:
Con esta hipótesis se busca determinar si la turbidez del agua puede ser empleada como indicador de baja calidad de la misma. Dado que se trata de un indicador visual, este facilitaría el reconocimiento de posibles problemas de contaminación.

Hipótesis 4: La concentración de fósforo total en el agua influye en el índice de calidad del agua
Con esta hipótesis se busca analizar si los niveles elevados de fósforo total pueden ser un factor determinante en la calidad del agua. Esto se basa en que el fósforo es un nutriente que, en exceso, puede desencadenar problemas en el agua.

Hipótesis 5: La cantidad de oxígeno disuelto tiene un impacto positivo en el índice de calidad del agua
Con esta hipótesis se busca explorar si los niveles de oxígeno disuelto están asociados a una mejor calidad de agua

Hipótesis 6: A partir de la concentración de coliformes fecales , fósforo total, turbiedad y oxígeno disuelto es posible predecir el ICA.
Con esta hipótesis se busca descubrir si, en conjunto, estos factores podrían permitir la estimación de la calidad del agua.

Desarrollo de hipótesis

Hipótesis 1: La calidad del agua se ve afectada por la presencia de bacterias fecales.

Para la validación de esta hipótesis, se calculó el coeficiente de correlación de Spearman entre las unidades formadoras de colonias de coliformes fecales y el ICA. Sin embargo, al obtener una resultado menor al 0,7, no fue posible obtener conclusiones significativas a partir de este análisis.

Para continuar con el análisis, se dividieron los datos en dos grupos según la mediana de los coliformes fecales, uno con los valores por debajo de la mediana y otro con los valores por encima.

Para analizar la normalidad del ICA de ambos grupos, se realizó el test de Shapiro-Wilk. Dado que ambos grupos arrojaron p-valores menores a 0.05, se obtuvo que ninguno seguía una distribución normal, lo cual llevó a optar por un test no paramétrico.

Para decidir qué test emplear, se realizó el test de Levene sobre el índice de calidad del agua en ambos grupos, con el fin de verificar la homocedasticidad de los datos. Al obtener un p-valor menor a 0.05, observamos que los datos no son homocedásticos, por lo que se procedió con la prueba de Kruskal-Wallis.

La prueba de Kruskal-Wallis es una prueba estadística no paramétrica utilizada para determinar si existen diferencias estadísticamente significativas entre dos o más grupos. Teniendo como hipótesis nula de que las medianas de todos los grupos son iguales.

Al aplicar la prueba con ambos grupos, se obtuvo un p-valor de 0, lo cual nos permitió rechazar la hipótesis nula y concluir que existen diferencias significativas entre los grupos. Esto valida nuestra hipótesis inicial: la calidad del agua se ve afectada por la presencia de bacterias fecales, ya que en grupos con diferentes concentraciones de estas bacterias se observaron diferencias en el índice de calidad del agua.

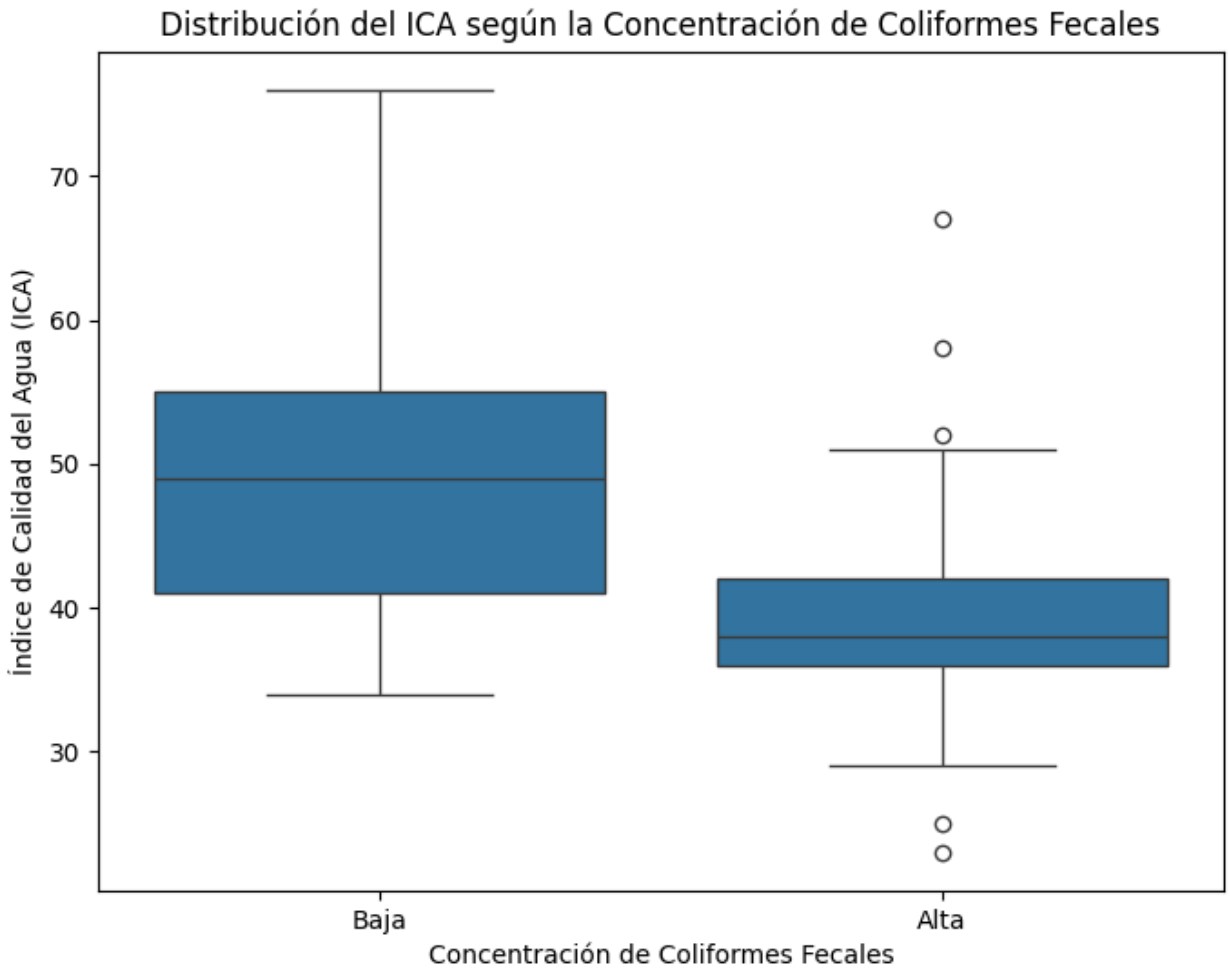

Figura 1: Boxplot del ICA según la concentración de coliformes fecales.

Como se puede observar en la figura 1, ambos grupos presentan una diferencia notoria en el índice de la calidad del agua. En los grupos con una concentración baja de unidades formadoras de colonias de coliformes fecales el índice de calidad de agua es más alto, y en grupos con alta concentración el índice es más bajo.

Hipótesis 2: La concentración de bacterias fecales varía según la estación del año

Para el planteo de esta hipótesis, se partió de analizar los resultados de algoritmos de reducción de dimensionalidad para descubrir si se presentaban grupos a en base a las estaciones del año. De estos, el que mejor formó agrupamientos fue UMAP.

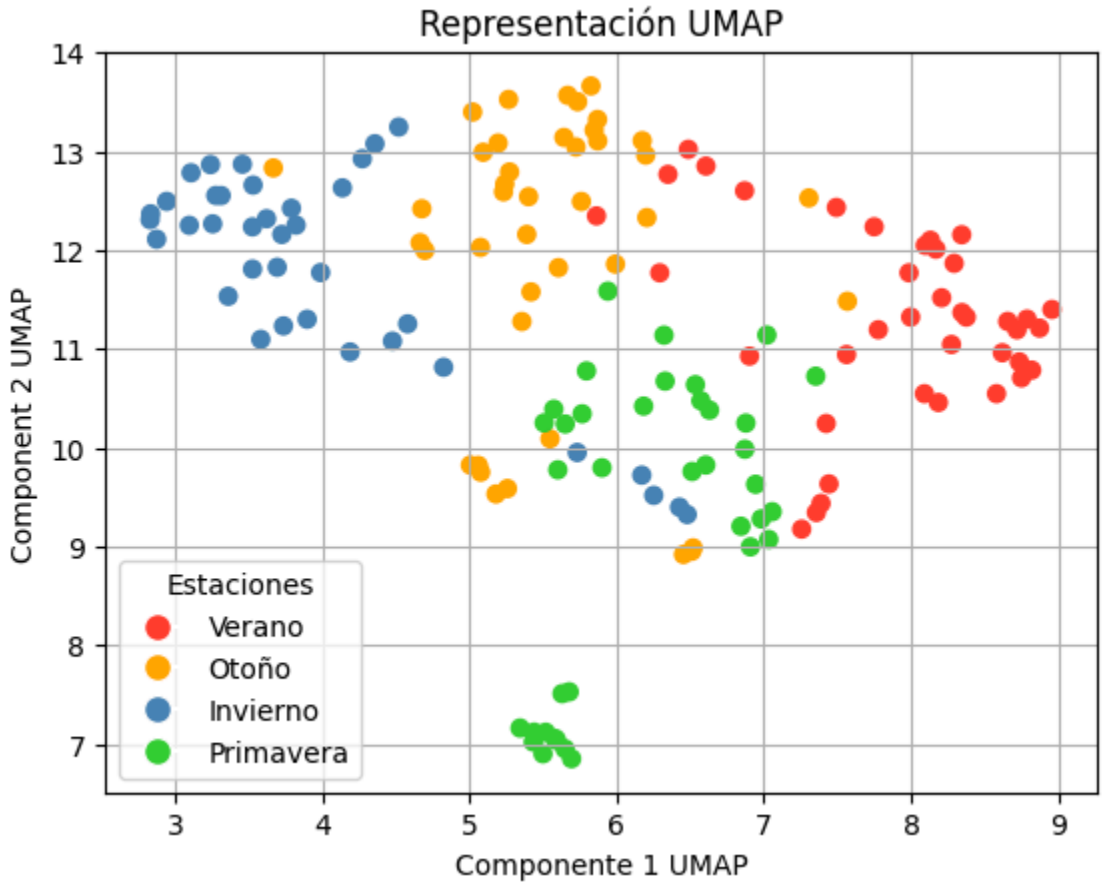

Figura 2: Representación UMAP de las diferentes estaciones

En la figura 2, se observan patrones de agrupación en función de las estaciones del año. Cada punto representa una medición y los colores indican la estación a la que pertenece. Se pueden identificar grupos específicos en diferentes áreas del gráfico, lo que sugiere que los datos presentan características distintivas según la estación.
Para la verificación de esta hipótesis, se comenzó agrupando los datos en cuatro conjuntos de acuerdo a la estación del año en que fueron obtenidos.
Posteriormente, se aplicó el test de Shapiro-Wilk a cada grupo para evaluar si la variable correspondiente a los coliformes fecales seguía una distribución normal en cada estación. Dado que todos los grupos presentaron un p-valor de 0, se concluye

que ninguno sigue una distribución normal, lo que llevó a optar por un test no paramétrico.
Para determinar el test no paramétrico adecuado, se empleó el test de Levene para verificar la homocedasticidad entre los grupos. Este análisis arrojó un p-valor de 0.009, menor a 0.05, indicando que los datos no son homocedásticos.
Finalmente, se aplicó el test de Kruskal-Wallis, el cual dio como resultado un p-valor de 0. Esto indica la existencia de diferencias significativas en las concentraciones de bacterias fecales entre las estaciones del año, validando nuestra hipótesis inicial.

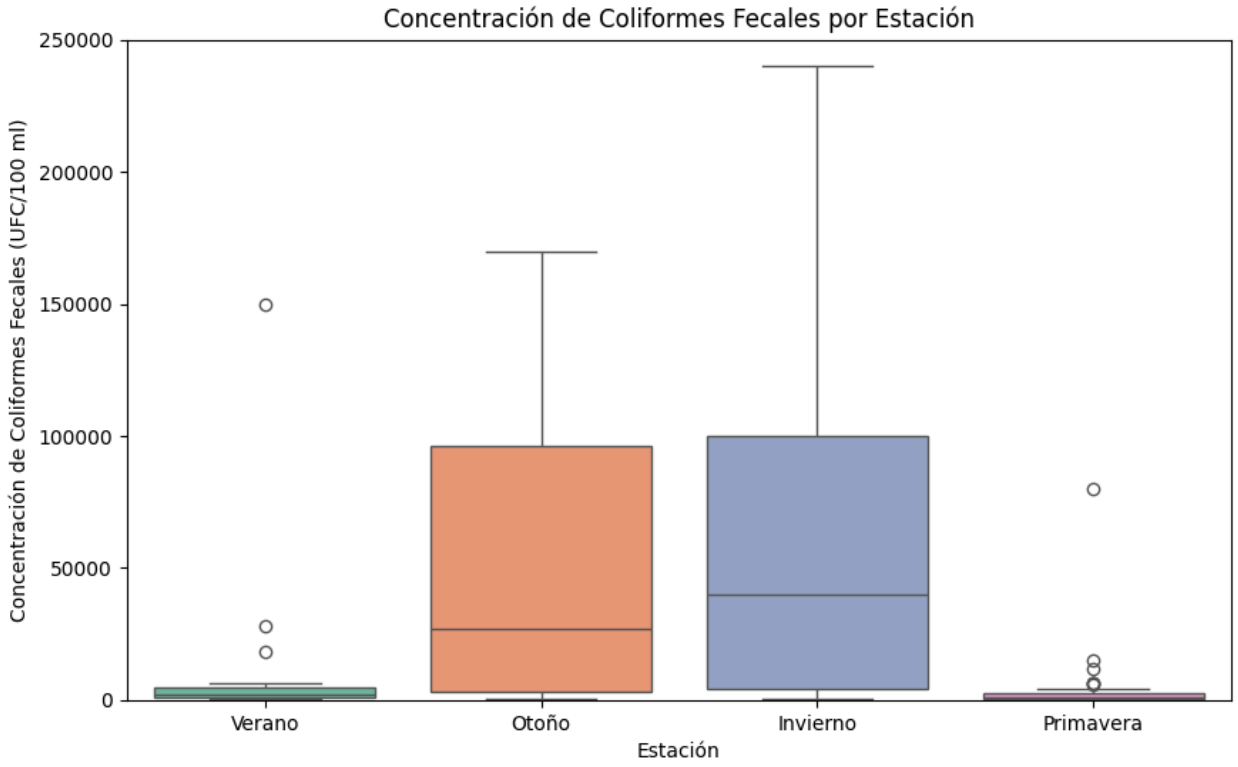

Figura 3: Boxplots de la concentración de coliformes fecales según la estación del año.

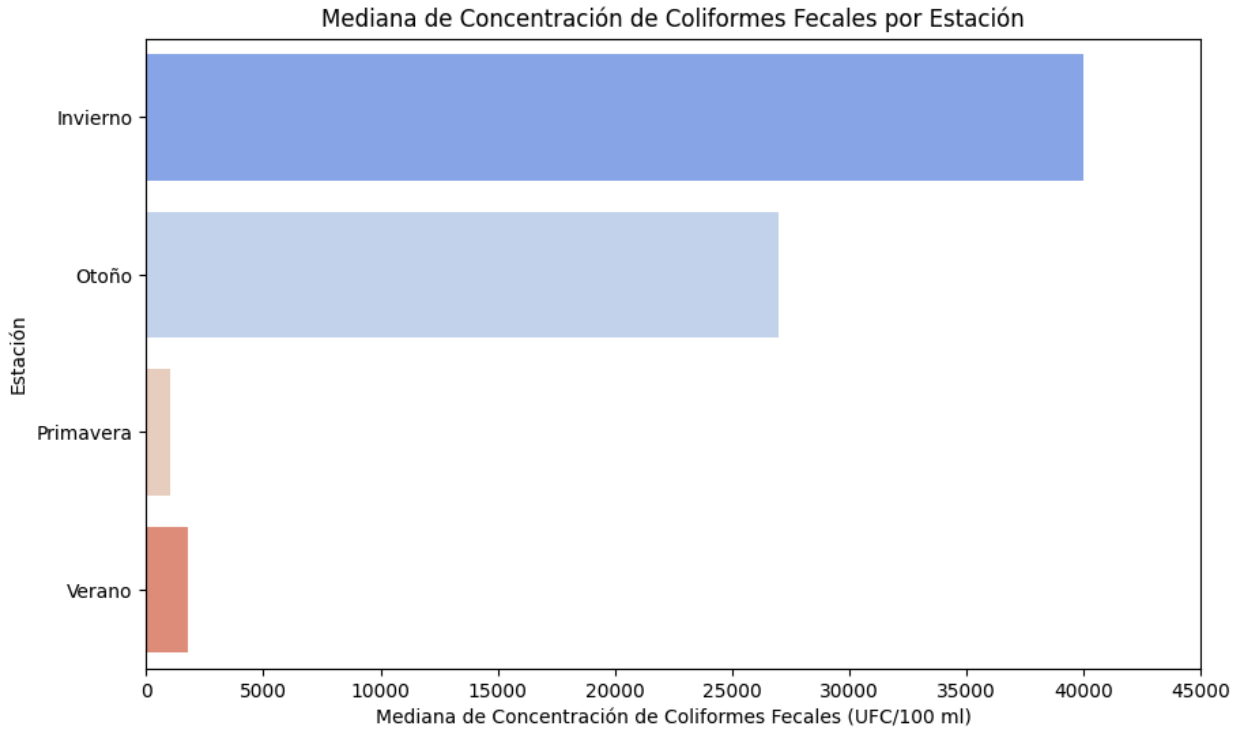

Figura 4: Gráfico de barras de la mediana de coliformes fecales por estación

Como puede apreciarse en las figuras tres y cuatro, se presentan diferencias notorias entre los grupos, por ende es posible afirmar que la concentración de bacterias fecales varía según la estación del año. Destacando que existen grandes diferencias entre todas las estaciones del año, aunque en mucho mayor medida entre el invierno y otoño en comparación al verano y primavera.

Hipótesis 3: El índice de calidad de agua varía según la turbidez.

Para la verificación de esta hipótesis se realizó la división de los datos en dos grupos según el nivel de turbidez: uno de baja turbidez y otro de alta. Esto se hizo en base a la mediana de la variable correspondiente a la turbidez.

Se aplicó el test de Shapiro-Wilk sobre el índice de calidad del agua a ambos grupos para evaluar si estos presentaban una distribución normal. Dado que en uno de los grupos se obtuvo un p-valor menor a 0.05, lo que indica que no se distribuye de forma normal, se debió aplicar un test no paramétrico.

Para la selección del test adecuado, se realizó el test de Levene para comprobar la homocedasticidad entre ambos grupos. Este análisis arrojó un p-valor de 0.068, mayor a 0.05, lo que indica que los datos son homocedásticos y es posible utilizar el test U de Mann-Whitney.

El test U de Mann-Whitney, es una prueba no paramétrica que permite comparar dos grupos independientes, el cual evalúa si las distribuciones de ambos grupos difieren significativamente. La hipótesis nula del test sostiene que no hay diferencia significativa entre los dos grupos.

Al realizar la prueba, obtuvimos un p-valor de 0.001, menor a 0.05, lo cual nos lleva a rechazar la hipótesis nula y afirmar que existen diferencias significativas en el índice de calidad del agua (ICA) entre los grupos de alta y baja turbidez.

Con este resultado se confirma la hipótesis planteada, sin embargo, se decidió representar gráficamente la diferencia entre los grupos.

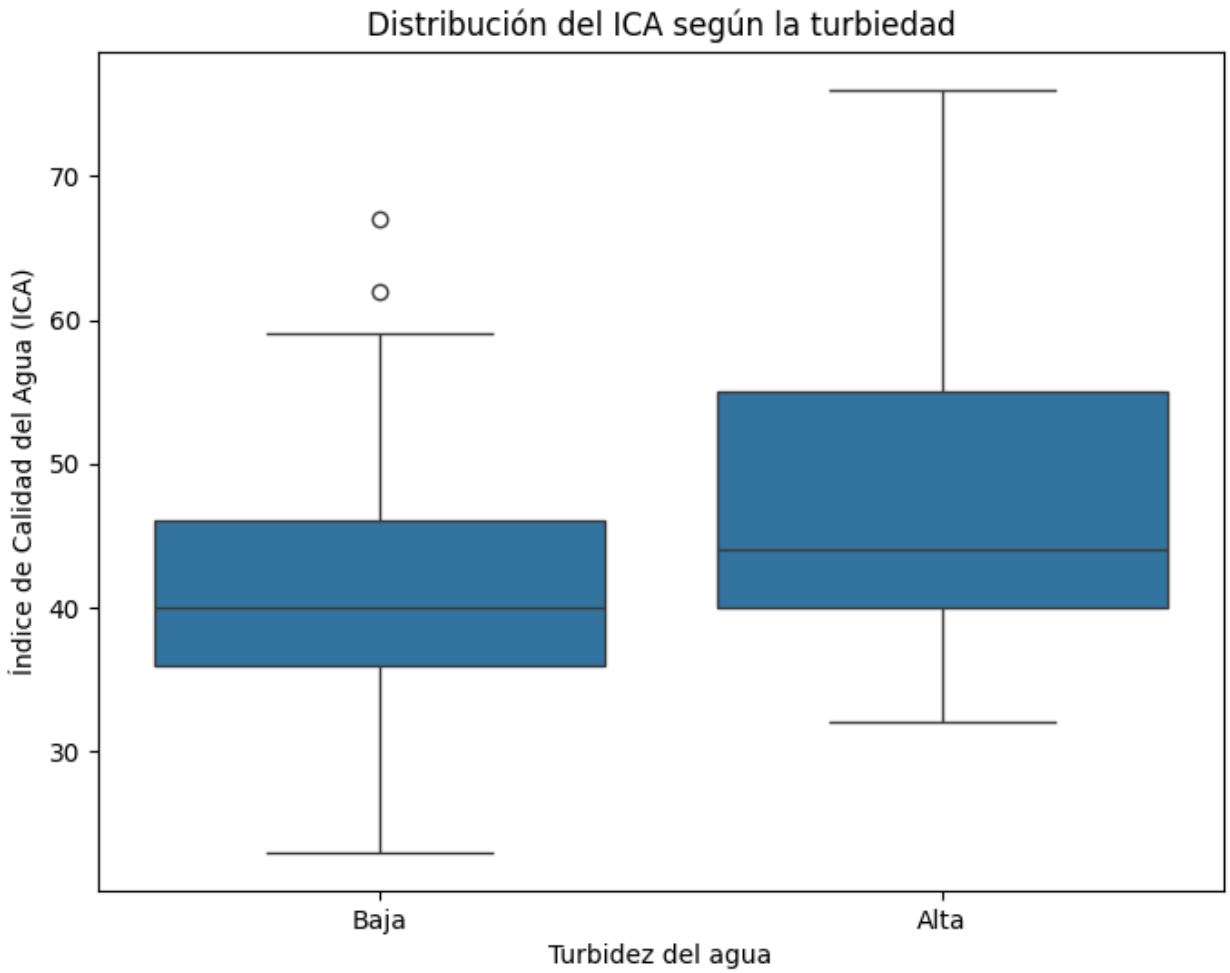

Figura 5: Boxplots del ICA en la turbidez agrupada en alta-baja

El análisis de la figura 5, revela una relación inversa a la esperada. En lugar de que un mayor nivel de turbidez se asocie con un menor ICA, los datos muestran que a mayor turbidez, el ICA es más alto. Esto sugiere que la turbidez no actúa como un indicador negativo de la calidad del agua, sino lo contrario.

Hipótesis 4: La concentración de fósforo total en el agua influye en el índice de calidad del agua

Para el desarrollo de esta hipótesis, se repitió el procedimiento de división de los datos en grupos, en este caso, según la variable correspondiente al fósforo total en el agua de acuerdo al cuartil al que pertenece la medición.
Posteriormente, se verificó la normalidad de cada grupo mediante el test de Shapiro-Wilk. Dado que algunos de los grupos no mostraban una distribución normal, se optó por un test no paramétrico.
Se aplicó el test de Levene a los cuatro grupos para evaluar la homocedasticidad. De esto, se obtuvo un p-valor de 0.010, menor a 0.05, lo cual indica que los datos no son homocedásticos. Dado esto, utilizamos el test de la mediana para determinar si existen diferencias significativas entre los grupos.
El test de la mediana es una prueba no paramétrica que compara la mediana de dos o más grupos independientes. La hipótesis nula de este test establece que las medianas de los grupos son iguales, mientras que la hipótesis alternativa sugiere que al menos una de las medianas difiere.
Al realizar el test de la mediana, obtuvimos un p-valor de 0. Esto nos permite rechazar la hipótesis nula y concluir que existen diferencias significativas entre las medianas de los grupos.

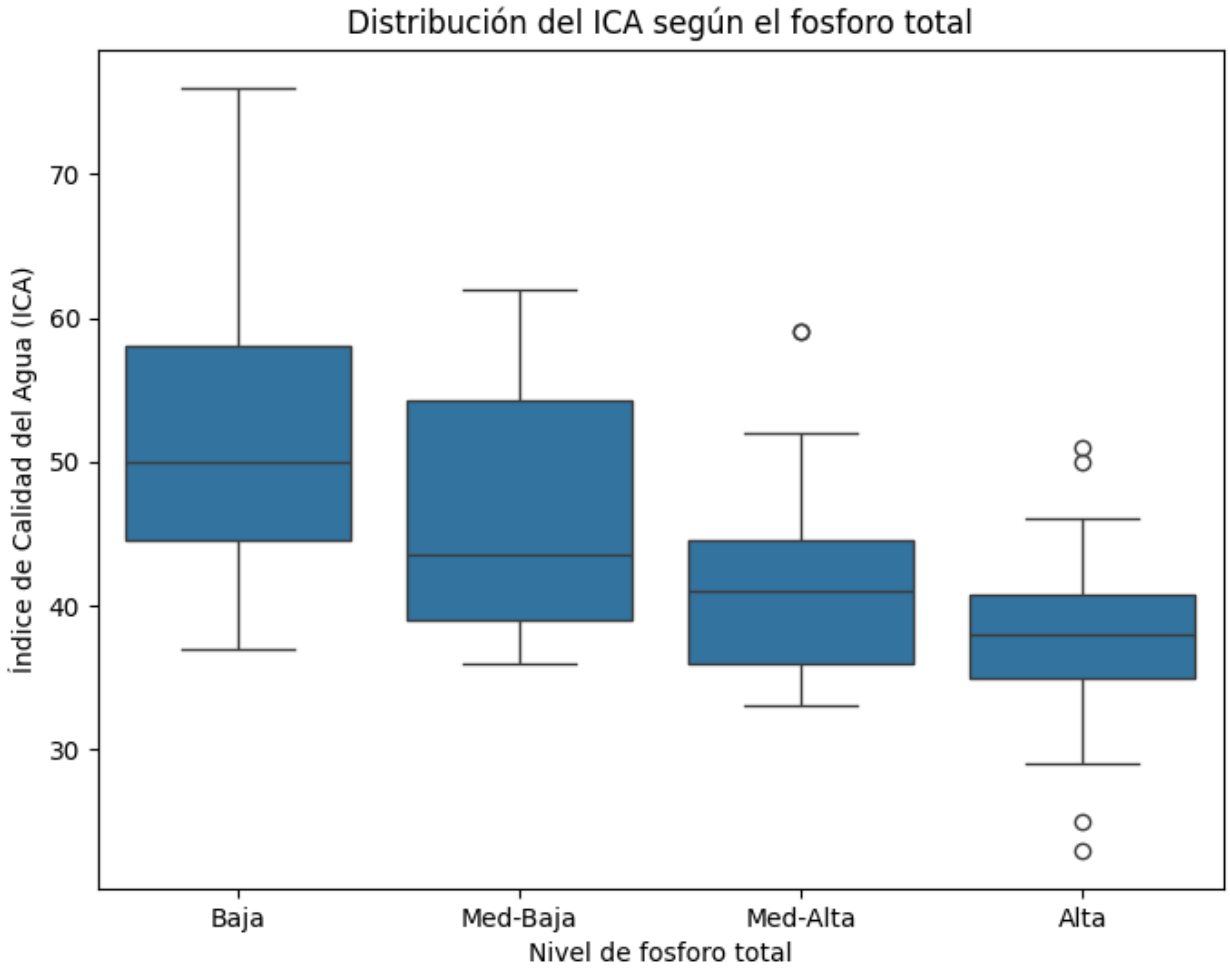

Figura 6: Boxplots del ICA según el fósforo total (distribuido en los grupos Baja/Media-Baja/Media-Alta/alta)

A partir de la figura 6 es posible observar las diferencias del ica entre los cuatro grupos y como el ICA disminuye gradualmente a medida que aumentan los niveles de fósforo total. Con esto, es posible validar la hipótesis de que la concentración de fósforo total influye en la calidad del agua.

Hipótesis 5: La cantidad de oxígeno disuelto tiene un impacto positivo en el índice de calidad del agua

Para este análisis, los datos fueron divididos en tres grupos según el oxígeno disuelto en cada uno.
Se aplicó el test de Shapiro-Wilk para verificar la normalidad del ICA en cada grupo. Los resultados mostraron p-valores menores a 0.05, indicando que los datos no siguen una distribución normal.
A su vez, se llevó a cabo el test de Levene para comprobar la homocedasticidad entre los grupos, obteniendo un p-valor de 0.106, lo cual indica que son homocedásticos.
Dado que se buscan comparar más de dos grupos independientes, se empleó el test de Kruskal-Wallis, que arrojó un p-valor de 0.006, evidenciando diferencias significativas entre los grupos.

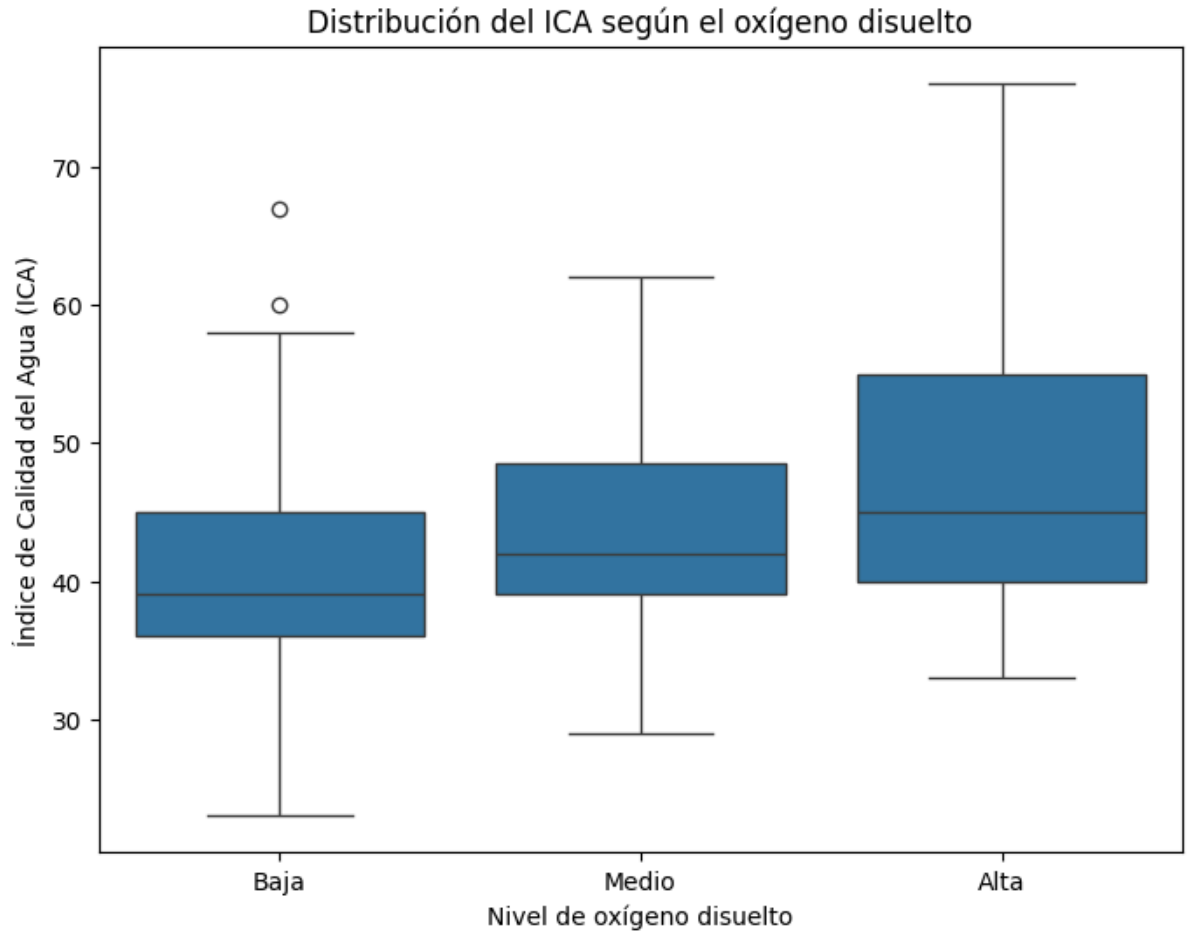

Figura 7: Boxplots del ICA en funcion del oxigeno disuelto en el agua, agrupado en tres niveles

Como puede observarse en la figura 7, a mayor concentración de oxígeno disuelto en el agua, mayor es el índice de calidad. Esto respalda la hipótesis de que la cantidad de oxígeno disuelto tiene un impacto positivo en el índice de calidad del agua.

Hipótesis 6: A partir de la concentración de coliformes fecales, fósforo total, turbiedad y oxígeno disuelto es posible predecir el ICA.

Para validar esta hipótesis, se planteó una regresión lineal múltiple, seleccionando como variables predictoras las correspondientes al fósforo total, unidades formadoras de colonias de coliformes fecales, oxígeno disuelto y turbidez del agua. La variable dependiente fue el ICA.

Se estandarizaron las variables predictoras y se realizó el análisis de la regresión lineal para evaluar si estas variables pueden predecir el ICA. Los resultados de la regresión arrojaron un R-cuadrado de 0.14, lo que indica que estas variables explican solo el 14% de la variabilidad en el índice de calidad del agua. Además, los residuos de la regresión no mostraron una distribución normal, lo cual sugiere que el modelo de regresión lineal no es adecuado para predecir el ICA con estas variables.

Si bien es posible emplear métodos alternativos de machine learning, como Lasso o Ridge Regression que podrían mejorar la predicción del ICA, se decidió no implementarlos, ya que el tamaño de la muestra puede ser una limitación. Dado que el conjunto de datos es relativamente pequeño, dichos modelos podrían no ofrecer resultados confiables. Por lo tanto, se concluye que con los parámetros planteados, no es posible predecir adecuadamente el ICA.

Conclusiones

En este análisis a partir de las mediciones en el Río de la Plata sobre la calidad del agua, exploramos seis hipótesis para entender cómo distintos factores influyen en el índice de calidad. A lo largo de las pruebas realizadas, se encontraron ciertas limitaciones impuestas por el conjunto de datos. Fue notorio que la mayoría de los datos no seguían una distribución normal ni presentaban homocedasticidad, lo que nos llevó a usar pruebas no paramétricas, las cuales tienen menos poder estadístico en comparación con los test paramétricos.

El estudio llevado a cabo ofrece una base para entender algunos de los factores que influyen en la calidad del agua, pero también evidencia que es necesario mejorar la recolección de datos en caso de requerir análisis más profundos o que se quieran construir modelos predictivos confiables. Esto, de ser llevado a cabo, podría ayudar en la toma de decisiones y gestión del Río de la Plata.

Referencias

Fuentes:

- https://www.argentina.gob.ar/salud/ambiental/agua
- https://aconsa-lab.com/parametros-calidad-agua-consumo-humano/
- https://www.unep.org/es/noticias-y-reportajes/reportajes/que-es-el-fosforo-y-por-que-aumenta-la-preocupacion-por-su-impacto