**TRABAJO PRÁCTICO ESPECIAL**

**FUNDAMENTOS DE LA CIENCIA DE DATOS**

**Grupo número:** 29

**Integrantes:**

Acevedo Maria Belen

Artaza Sheila Nerea

Meresman Eden Denise

Base: calidad\_del\_agua\_2023

**INTRODUCCIÓN:**

En este proyecto, aplicamos los conocimientos adquiridos en la cátedra de Fundamentos de la Ciencia de Datos a lo largo de este cuatrimestre. A través de este trabajo, demostramos nuestra habilidad para manipular datos provenientes de una base de datos cargada previamente, realizar su limpieza, imputación cuando fue necesario, y llevar a cabo un análisis exploratorio exhaustivo. Además, aprendimos a plantear hipótesis basadas en los datos y a validarlas utilizando las herramientas y metodologías proporcionadas en el curso.

**ANÁLISIS INICIAL DE VARIABLES:**

Al comenzar, identificamos que el conjunto de datos contenía principalmente variables cualitativas, excepto dos variables cuantitativas. Esto se debió a que muchas de las variables contenían valores como "no se muestreó", que eran irrelevantes para el análisis. En cuanto a los datos numéricos, la mayoría de las variables tenían entre 168 y 169 valores no nulos, excepto "ph", "p\_total\_l\_mg\_l" e "ica", que contenían valores NaN, y "calidad\_del\_agua", que tenía 15 valores faltantes.

En esta etapa, también identificamos algunos posibles outliers en las variables, pero decidimos no eliminarlos, ya que consideramos que podrían contener información relevante para el análisis.

**LIMPIEZA DE DATOS:**

Antes de comenzar con la limpieza, creamos una copia del conjunto de datos original (denominada *base\_limpia*) para preservar los datos iniciales.

A continuación, realizamos las siguientes acciones durante el proceso de limpieza:

* **Eliminación de columnas irrelevantes**: Decidimos eliminar las columnas que no aportaban valor al análisis:
  + "campaña", ya que la información era redundante respecto a la variable "fecha".
  + "codigo", dado que el identificador de cada lugar no era relevante para los análisis.
  + "año", ya que el valor era constante (2023) y podía deducirse de la variable "fecha".
* **Reemplazo de valores no relevantes**: Reemplazamos los valores de texto irrelevantes, como "en obra", "no se midió", "no se muestreó", "sin equipo", etc., por valores NaN, para que no afectaran los análisis.
* **Eliminación de rangos en variables**: Eliminamos el símbolo "<" en las variables que contenían rangos, ya que no consideramos apropiado su uso en este análisis. Estas variables son: "nitrato\_mg\_l", "nh4\_mg\_l", "fosf\_ortofos\_mg\_l", "dbo\_mg\_l", "dqo\_mg\_l", "hidr\_deriv\_petr\_ug\_l", "cr\_total\_mg\_l", "cd\_total\_mg\_l", "clorofila\_a\_ug\_l" y "microcistina\_ug\_l".
* **Conversión de variables cualitativas a booleanas**: Para facilitar el análisis, convertimos las variables cualitativas en variables booleanas. Las variables convertidas fueron:
  + "olores", "color", "espumas", "mat\_susp" (con valores 0 para ausencia y 1 para presencia).
  + "calidad\_de\_agua" (con valores 0 para "Muy deteriorada" y 1 para "Extremadamente deteriorada").
* **Imputación de valores faltantes**: Decidimos imputar las filas con más del 60% de sus valores incompletos, dado que eliminar todas las filas con valores faltantes habría provocado una pérdida significativa de datos. Como las variables no seguían una distribución normal, utilizamos el algoritmo de *K-nearest neighbors* (KNN) para la imputación de los valores NaN.

**EXPLORACIÓN DE DATOS:**

Para comenzar con la exploración, realizamos histogramas para entender la distribución de cada variable. Luego, trabajamos con una copia de los datos limpios (*base\_limpia\_matriz*) y nos enfocamos únicamente en las variables cuantitativas para calcular la matriz de correlaciones. Durante este proceso, eliminamos las columnas "sitios", "fecha", "olores", "color", "espumas", "mat\_susp" y "calidad\_de\_agua" para simplificar el análisis.

También notamos que la variable "hidr\_deriv\_petr\_ug\_l" mantenía un único valor constante (0.10), lo que provocaba que toda la columna se reemplazara por valores NaN. Por lo tanto, decidimos eliminar esta columna.

A partir de la matriz de correlaciones, identificamos las siguientes relaciones lineales con un valor absoluto mayor a 0.67, lo que indica una posible relación significativa:

* "tem\_agua" y "tem\_aire" (0.84)
* "clorofila\_a\_ug\_l" y "turbiedad\_ntu" (0.83)
* "turbiedad\_ntu" y "dqo\_mg\_l" (0.75)
* "fosf\_ortofos" y "nh4\_mg\_l" (0.71)
* "calidad\_de\_agua" y "ica" (-0.69)
* "enteroc\_ufc\_100ml" y "colif\_fecales\_ufc\_100ml" (0.68)
* "clorofila\_a\_ug\_l" y "dqo\_mg\_l" (0.68)

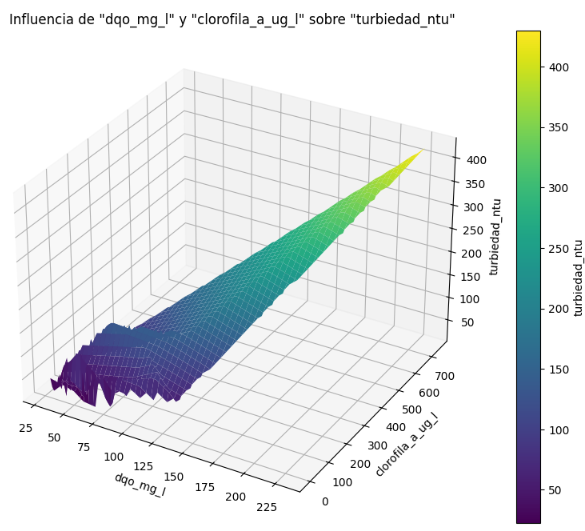
**Hipótesis planteadas:** Para el análisis posterior, decidimos trabajar con un nivel de significancia de α=0.05 para validar nuestras hipótesis. La primera de estas hipótesis se detalla a continuación:

Surge entonces nuestra primer hipótesis a validar:

**H1: ¿La turbiedad del agua está influenciada directamente por la clorofila y la demanda química de oxígeno?**

* **H0** = La clorofila y la demanda química de oxígeno no tienen una influencia directa en la turbidez del agua.
* **H1** = La clorofila y la demanda química de oxígeno tienen una influencia directa en la turbidez del agua.

Realizamos un gráfico de dispersión de las variables “clorofila\_a\_ug\_l” y “dqo\_mg\_l” pintadas por “turbiedad\_ntu” que no mostró un gran comportamiento, realizamos además un gráfico en 3D, que si mostraba más información.



*Gráfico 3D que muestra que a menor medida de clorofila y dqo, menor valor de turbiedad.*

La siguiente que surgió fue:

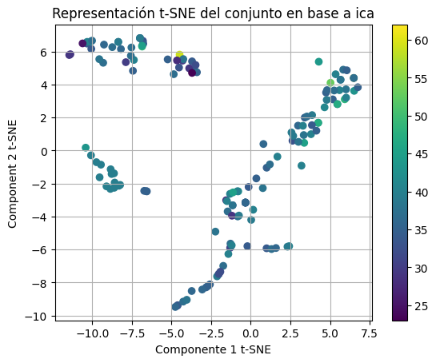
**H2: ¿El índice de calidad del agua se ve influenciado por características físico químicas y microbiológicas registradas?**

* H0 = Las características físico químicas y microbiológicas no tienen influencia en el índice de la calidad del agua.
* H1 = Las características físico químicas y microbiológicas tienen influencia en el índice de la calidad del agua.

Comenzamos normalizando para poder aplicar reducción de dimensionalidad mediante PCA. Lo que buscábamos era ver si las variables independientes afectan el comportamiento del índice de calidad del agua.

El resultado de esta reducción sólo representaba un 35% con lo cual decidimos aplicar t-sne para ver si esta reducción, nos mostraría un cierto comportamiento.

El resultado de esta aplicación no nos mostró ningún comportamiento fuerte, con lo cual la **H2** **queda refutada.**

****

*Gráfico T-sne que si bien parece mostrar agrupamientos, el comportamiento no parece variar entre los grupos.*

**H3: ¿La concentración de las bacterias fecales afecta al índice de calidad del agua?**

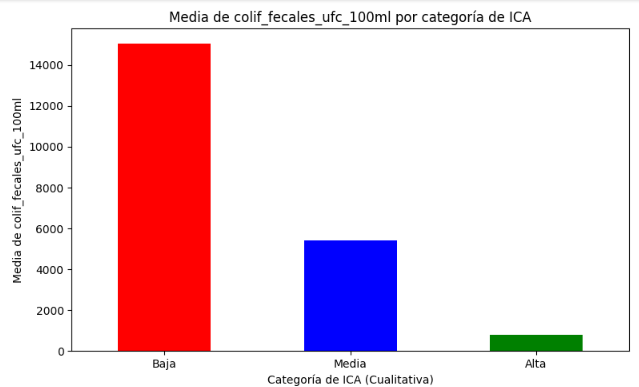
**H0**: La concentración de las bacterias fecales no presenta diferencias en las medianas en relación a los grupos del índice de calidad del agua (ICA).  
**H1**: La concentración de las bacterias fecales presenta diferencias en las medianas en relación al índice de calidad del agua.

Para abordar esta hipótesis, comenzamos aplicando una técnica de reducción de dimensionalidad mediante el análisis de componentes principales (PCA), lo que nos permitió representar el 90% de la información de las variables bacterianas. Aunque la proyección obtenida mostró un comportamiento leve, nos ayudó a orientar el análisis.

A continuación, transformamos la variable cuantitativa **ICA** en una variable cualitativa, dividiéndola en tres rangos (alta, media, baja). Realizamos tres gráficos de barras que mostraban la relación entre las concentraciones de las bacterias fecales y los diferentes niveles de **ICA**. Estos gráficos sugieren una tendencia que parecía respaldar nuestra hipótesis de que la concentración bacteriana influye en la calidad del agua.

Luego, procedimos a evaluar la normalidad de los datos utilizando el test de **Shapiro-Wilk**, y la homocedasticidad mediante el test de **Levene**, obteniendo resultados que indicaron que los datos no seguían una distribución normal y no presentaban homocedasticidad.

Con estos resultados y considerando que la variable dependiente **ICA** ahora era cualitativa con tres grupos, decidimos aplicar el **test no paramétrico de Kruskal-Wallis**. Los valores p obtenidos fueron significativamente bajos, lo que nos llevó a rechazar la hipótesis nula (H0) y concluir que existe una diferencia significativa en las medianas de las concentraciones de las bacterias fecales entre los diferentes grupos de **ICA**. Esto valida nuestra hipótesis de que las bacterias fecales influyen en el índice de calidad del agua.

****

*Uno de los gráficos de barra realizados, en este caso entre coliformes fecales e ica(cualitativa). Muestra que a mayor cantidad de coliformes fecales, menor índice de la calidad del agua.*

Podemos concluir que la hipótesis **(H3)** planteada es válida, ya que las colonias fecales ejercen una influencia significativa en la calidad del agua. El análisis de las bacterias presentes en el agua, como Escherichia coli, Enterococcus y las coliformes fecales, ha mostrado una correlación clara con la variación en los niveles del índice de calidad del agua (ICA).

A raíz de este análisis, surge una nueva interrogante que debe ser abordada para comprender más a fondo los factores que afectan la calidad del agua. Para ello, decidimos expandir el análisis e incorporar las variables de NH₄ y fosfato, con el fin de evaluar si estas también ejercen influencia en los niveles del índice de calidad del agua. Al sumar estos nuevos componentes, buscamos obtener una visión más completa y precisa de los factores que contribuyen a la variabilidad en la calidad del agua del río.

Los resultados obtenidos del test de Kruskal-Wallis para las nuevas variables revelan que existen diferencias significativas en las concentraciones de coliformes fecales, Escherichia coli, Enterococcus, fosfato y amonio NH₄ entre las distintas categorías del índice de calidad del agua (ICA).. Al comparar estos hallazgos con los resultados previos para las variables bacterianas, podemos concluir que las cinco variables estudiadas varían en función de la categoría del ICA, lo que sugiere que estos factores, tanto microbiológicos como químicos, están estrechamente relacionados con las variaciones en la calidad del agua.

**H4: ¿Las bacterias fecales influyen sobre la presencia de materia fecal suspendida?**

**H0**: Las bacterias fecales no presentan diferencias en las medianas en relación a los grupos de materia fecal suspendida.  
**H1**: Las bacterias fecales presentan diferencias en las medianas en relación a los grupos de materia fecal suspendida.

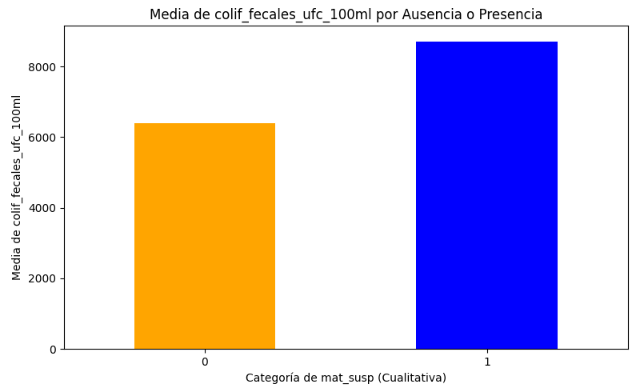
Para abordar esta hipótesis, comenzamos aplicando la técnica de reducción de dimensionalidad mediante PCA, lo que permitió representar el 90% de la información de las variables bacterianas. Aunque la proyección mostró un comportamiento leve, nos proporcionó una visión general del patrón de las bacterias.

A continuación, elaboramos gráficos de barras para cada una de las bacterias fecales, comparándolas con los grupos de la variable **"mat\_susp"** (materia fecal suspendida). Estos gráficos indicaron posibles diferencias entre las concentraciones bacterianas y la presencia de materia fecal suspendida.

Como en hipotesis anteriores corroboramos la no distribucion normal de las variables bacterianas, directamente pasamos a evaluar la homocedasticidad utilizando el test de **Fisher**, ya que la variable **"mat\_susp"** solo tiene dos grupos. Los resultados fueron satisfactorios, ya que el p-valor fue mayor a 0,05, indicando que no hay evidencia suficiente para rechazar la hipótesis nula de homocedasticidad.

Con estos resultados, procedimos a aplicar el test no paramétrico de **Mann-Whitney** para cada bacteria fecal en relación con los dos grupos de materia fecal suspendida (Ausencia y Presencia).. Los resultados mostraron que **colif\_fecales\_ufc\_100ml** y **escher\_coli\_ufc\_100ml** tienen una influencia significativa sobre la materia fecal suspendida, ya que los p-valores fueron menores a 0,05, lo que llevó al rechazo de la hipótesis nula en esos casos. Sin embargo, para **enteroc\_ufc\_100ml**, el p-valor fue mayor a 0,05, lo que significa que no se rechaza la hipótesis nula y no hay evidencia de que esta bacteria influya sobre la materia fecal suspendida.

En conclusión, la **H4** queda validada parcialmente, ya que solo dos de las bacterias fecales (**coliformes fecales** y **Escherichia coli**) muestran influencia sobre la materia fecal suspendida, mientras que **Enterococcus** no presenta esta relación.



*Gráfico de barras diferenciando por ausencia y presencia de materia suspendida mostrando la cantidad de coliformes fecales. Muestra la variación de coliformes fecales en relación a la ausencia o presencia de materia fecal suspendida.*

**H5: ¿La concentración de la bacteria coliforme fecal tiene un comportamiento estacional?**

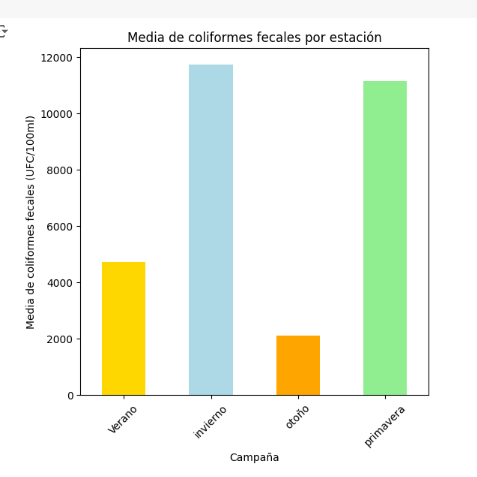
H0: La bacteria coliforme fecal no presenta diferencias en las medianas en relación a los grupos de las estaciones.  
H1: La bacteria coliforme fecal presenta diferencias en las medianas en relación a los grupos de las estaciones.

Para abordar esta hipótesis, comenzamos generando un gráfico de barras que comparaba la concentración de coliformes fecales entre las distintas estaciones del año. A simple vista, el gráfico mostró una variabilidad estacional en los niveles de la bacteria, lo que sugirió un posible comportamiento estacional.

A continuación, realizamos un análisis de normalidad utilizando el test de Shapiro-Wilk y un análisis de homocedasticidad con el test de Levene. Los resultados indicaron que los datos no seguían una distribución normal, dado que el p-valor fue menor a 0.05, por lo que rechazamos la hipótesis nula de normalidad. Además, los datos no eran homocedásticos, ya que el p-valor también fue menor a 0.05, lo que llevó a rechazar la hipótesis nula de homocedasticidad.

Dado que la variable "campaña" tiene más de dos grupos, decidimos aplicar el test no paramétrico de Kruskal-Wallis. Los resultados de este test confirmaron lo observado en el gráfico de barras, ya que el p-valor obtenido fue menor a 0.05, lo que permitió rechazar la hipótesis nula (H0) y concluir que la bacteria coliforme fecal presenta un comportamiento estacional.

**Por lo tanto, la H5 queda validada.**

****

*Gráfico de barras diferenciando por estación del año mostrando la cantidad de coliformes fecales. Muestra la variación estacional de las coliformes.*

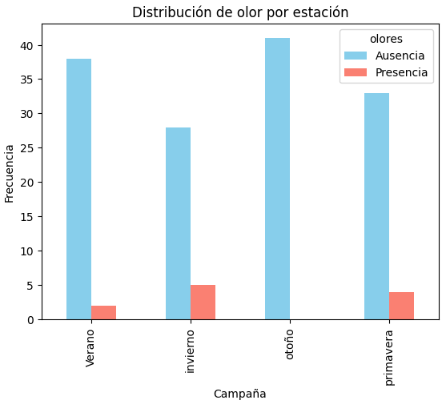
**H6: ¿El olor del agua del Río de La Plata tiene un comportamiento estacional?**

**H0**: El olor del agua del Río de La Plata no presenta diferencias en relación a los grupos de estaciones.  
**H1**: El olor del agua del Río de La Plata presenta diferencias en relación a los grupos de estaciones.

Para abordar esta hipótesis, comenzamos generando un gráfico de barras que comparaba la presencia o ausencia de olor en el agua durante las distintas estaciones del año. A partir de los datos disponibles, observamos que las muestras correspondientes a la categoría "presencia" eran limitadas, lo que dificulta identificar un patrón claro de comportamiento estacional.

Luego, realizamos una prueba de **Chi-cuadrado** para evaluar la relación entre la presencia de olor y las estaciones. Los resultados arrojaron un **p-valor mayor a 0.05**, lo que nos llevó a no rechazar la hipótesis nula (H0). Sin embargo, dado el tamaño limitado de la muestra en la categoría "presencia", no podemos concluir de manera firme que el olor del agua varíe significativamente entre las estaciones.

En resumen, aunque los resultados sugieren que no hay una diferencia significativa en el olor del agua entre las estaciones, la falta de suficientes datos en la categoría "presencia" limita la validez de esta conclusión.



*Gráfico de barras agrupadas por cada estación mostrando la ausencia y presencia de olor en cada una. No muestra un comportamiento estacional pero las muestras del tipo “Presencia” no eran suficientes.*

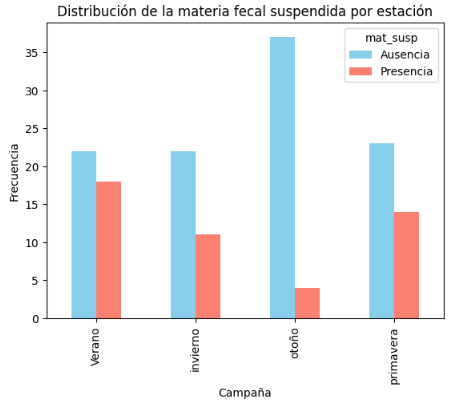
**H7: ¿La materia fecal suspendida en el agua tiene un comportamiento estacional?**

**H0**: La materia fecal suspendida en el agua no presenta diferencias en relación a los grupos de estaciones.  
**H1**: La materia fecal suspendida en el agua presenta diferencias en relación a los grupos de estaciones.

Iniciamos el análisis con la creación de un gráfico de barras, diferenciando las estaciones del año y mostrando la presencia o ausencia de materia fecal suspendida en cada una de ellas. A partir de los resultados visuales, observamos una variabilidad notable en el comportamiento de la materia fecal suspendida a lo largo de las estaciones.

Luego, aplicamos la prueba de **Chi-cuadrado**, que nos dio un **p-valor menor a 0.05**, lo que nos llevó a rechazar la hipótesis nula (H0) y concluir que **existen diferencias significativas** en la presencia de materia fecal suspendida entre las estaciones del año.

En consecuencia, podemos afirmar que la materia fecal suspendida en el agua sí presenta un comportamiento estacional. Así, la hipótesis H7 queda validada.



*Gráfico de barras agrupadas por cada estación mostrando la ausencia y presencia de materia fecal suspendida en cada una. Muestra un comportamiento estacional de la materia fecal suspendida.*

**CONCLUSIONES:**

Los resultados obtenidos revelan relaciones significativas entre diversos factores que afectan la calidad del agua del río:

1. Influencia en la turbiedad: Se constató que la turbiedad del agua está fuertemente influenciada por la concentración de clorofila y la demanda química de oxígeno (DQO). Esto sugiere que tanto la presencia de materia orgánica como la demanda de oxígeno para su descomposición contribuyen a la turbiedad del agua, destacando su relevancia como indicador de contaminación en cuerpos de agua.
2. Determinantes del índice de calidad del agua: El análisis muestra que el índice de calidad del agua (ICA) no está determinado por todos los componentes fisicoquímicos medidos, sino que se ve principalmente afectado por la presencia de bacterias fecales, en particular *Escherichia coli*, coliformes fecales y enterococos. Además, al incorporar el análisis de amonio (NH₄) y fosfato (FOSF), se observó que estos dos compuestos también influyen en la calidad del agua, lo que resalta la importancia de monitorear tanto los indicadores microbiológicos como los nutrientes en el agua.
3. Relación entre materia fecal suspendida y bacterias fecales: Los hallazgos sugieren que la cantidad de materia fecal suspendida en el agua está estrechamente vinculada con la presencia de bacterias fecales, especialmente coliformes fecales y *Escherichia coli*. Este vínculo refuerza la hipótesis de que la contaminación fecal contribuye a la presencia de materia suspendida, un factor clave en la evaluación de la calidad del agua.
4. Estacionalidad de los coliformes fecales y materia suspendida: Se observó un comportamiento estacional en los niveles de coliformes fecales, lo que a su vez influye en la variabilidad de la materia fecal suspendida a lo largo del año. Sin embargo, debido a la falta de datos suficientes, no fue posible evaluar una relación estacional en el olor del agua, lo que sugiere la necesidad de una mayor recopilación de datos para futuras investigaciones en esta área.