**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 nos dedicamos a aplicar los contenidos brindados por la cátedra de Fundamentos de la Ciencia de Datos, durante el transcurso de este cuatrimestre. Pudimos aplicar los conocimientos aprendidos sobre el lenguaje Python. En este trabajo pudimos demostrar que sabemos manipular datos brindados desde una base de datos previamente cargada, limpiarlos, imputarlos en caso de ser necesario, hacer con ellos un análisis exploratorio. Aprendimos y desarrollamos el conocimiento de plantearnos ante los datos posibles hipótesis para su posterior demostración mediante las herramientas brindadas por la cátedra.

**ANÁLISIS DE INICIAL DE VARIABLES:**    
Al iniciar identificamos sólo dos variables cuantitativas y las demás inicialmente cualitativas.

Esto debido a que, la mayoría, poseía valores como “no se muestreó” además de los valores numéricos que sí serían de nuestro interés.

La mayoría de las variables contenían entre 168 y 169 valores no nulos, especificamente “ph”, “p\_total\_l\_mg\_l” e “ica” contenian un Nan y “calidad\_del\_agua” contenía 15 Nans.

.  
Identificamos que había posibles outliers pero decidimos no eliminarlos.

**LIMPIEZA:**

Antes de iniciar la limpieza, copiamos la base de datos inicial en otra variable (base\_limpia), para preservar los datos originales.

Una vez iniciada, lo primero que realizamos fue la eliminación de las columnas que consideramos que no brindarán información:

-”campaña”: consideramos redundante la información que brindaba en relación a la variable fecha

-”codigo”: consideramos que el identificador de cada lugar no era relevante para el análisis

-”año”: consideramos que como el año era constante (2023) no iba a ser relevante para el análisis, además de que el mismo se podía ver desde la variable fecha (redundancia).

Luego, continuamos con el reemplazo a Nan de los valores que estaban en formato de texto que no eran de nuestro interés (“en obra”, “no se midió”, “no se muestreó”, “no muestreó”, “sin equipo”, etc).

Eliminamos también el simbolo “<” de las variables que los poseían ya que no consideramos el uso de rangos. 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”.

Además, convertimos las variables cualitativas en booleanas para poder incluirlas en la exploración.

* “olores”, “color”, “espumas”, “mat\_susp” (Ausencia=0 y Presencia=1)
* “calidad\_de\_agua” (Muy deteriorada=0 y Extremadamente deteriorada=1)

Antes de iniciar con dicha exploración, decidimos aplicar imputación de las filas que poseían más del 60% de sus valores completos, debido a que si eliminabamos todas las filas con Nan perdiamos mucha información. Como todas las variables poseían distribución no normal utilizamos el algoritmo de vecinos más cercanos.

**EXPLORACIÓN:**

Realizamos los histogramas de cada variable para conocer su distribución.

Para iniciar con la exploración realizamos una copia de la copia (llamada base\_limpia\_matriz) para poder solo quedarnos con las variables cuantitativas y así realizar el cálculo de la matriz de correlaciones. Las columnas eliminadas para realizar esto fueron “sitios”, “fecha” , “olores”, “color”, “espumas”, “mat\_susp”y “calidad\_de\_agua”.

Además, nos dimos cuenta de que “hidr\_deriv\_petr\_ug\_l” al mantener un único valor (0.10) constante era reemplazada completa por Nans en la matriz. Por ende la eliminamos también.

Gracias a la matriz pudimos identificar las siguientes posibles relaciones lineales con valor absoluto mayor a 0,67 (decidimos este valor y no 0,7 porque no tenemos tantos datos):

* 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)

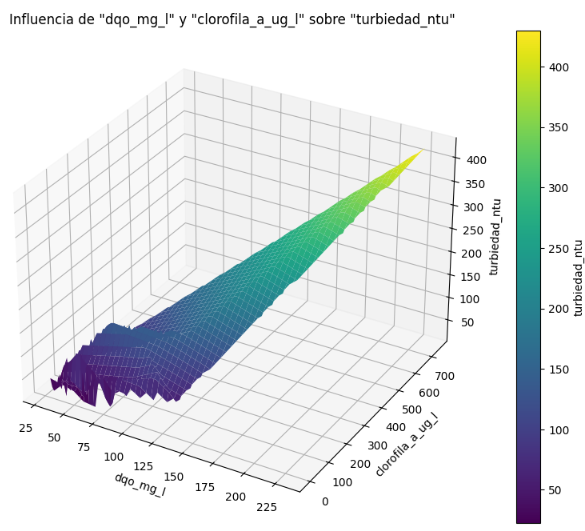
**Aclaración:** para todas nuestras hipótesis planteadas a continuación vamos a considerar un nivel de significancia de α=0,05.

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:

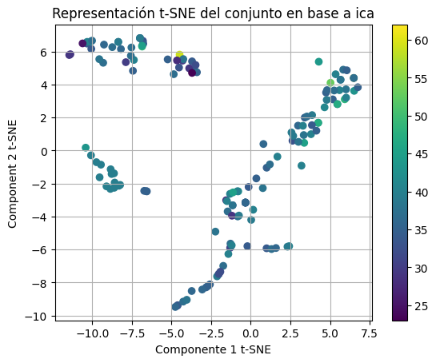
**H2: ¿El índice de calidad del agua se ve influenciado por características físico químicas y microbiológicas registradas?**

* 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 afectan al índice de la calidad del agua?

* H0 = La concentración de las bacterias fecales no presentan una diferencia de las medianas en relación a los grupos del índice de calidad del agua.
* H1 = La concentración de las bacterias fecales presentan una diferencia de las medianas en relación del índice de calidad del agua.

Comenzamos aplicando reducción de dimensionalidad mediante PCA, que nos representó un 90% de la información y nos mostró un leve comportamiento.

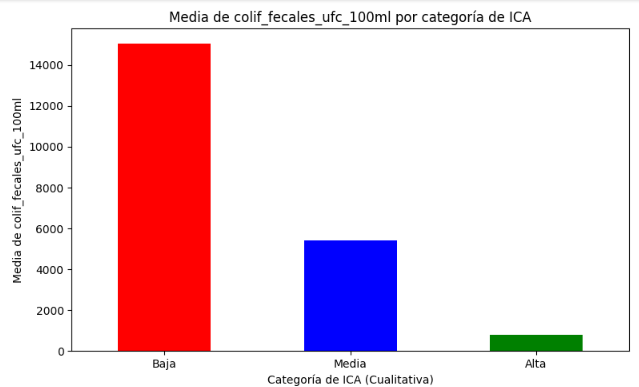
Luego decidimos convertir la variable cuantitativa “ica” a cualitativa mediante 3 rangos (alta, media, baja). Realizamos tres gráficos de barras con índice del agua en relación con cada bacteria fecal. Vimos que estos gráficos indicaban comportamientos del “ica” que parecían validar nuestra hipótesis.

Entonces continuamos con la evaluación de normalidad mediante Shapiro Wilk, y homocedasticidad mediante Levene, lo cual nos dió como resultado que los supuestos eran no normales y no homocedásticos.

Entonces frente a estos resultados y dado que la variable dependiente “ica” ahora era cualitativa con tres grupos, decidimos aplicar el test no paramétrico de Kruskal-Wallis.

Los p-valores obtenidos nos dieron menor con lo cual se rechaza la H0, por ende existe una diferencia significativa entre las medianas de las bacterias fecales para los diferentes grupos de ica cualitativa.

Finalmente la **H3 queda validada.**

****

*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.*

**H4: ¿Las bacterias fecales tienen influencia sobre la materia fecal suspendida?**

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

Comenzamos realizando reducción de la dimensionalidad, mediante PCA, el cual nos representó el 90% de la información, y se podía ver un leve comportamiento.

Realizamos gráficos de barras, uno por cada bacteria fecal en comparación con los grupos de la variable “mat\_susp”.

Continuamos entonces con el análisis de normalidad realizando histogramas, qq-plot y el test de Shapiro Wilk para cada variable fecal (ya que son cuantitativas).

Luego con el análisis de homocedasticidad mediante Fisher, ya que “mat\_susp” tiene solo dos grupos.

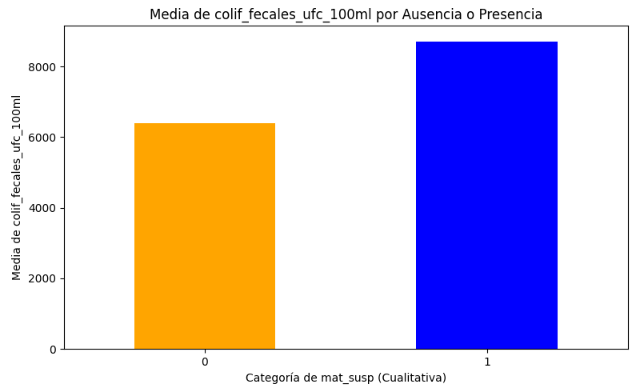
Los resultados nos demostraron que ninguna de las bacterias fecales posee una distribución normal, ya que los p-valores resultantes fueron menores a 0,05, con lo cual se rechaza la hipótesis nula de normalidad.

Por otro lado, los supuestos son homocedásticos, por tener un p-valor > 0,05, con lo cual queda rechazada la hipótesis nula de homocedasticidad.

Por ende, continuamos aplicando el test de Mann-Withney para cada una de las bacterias fecales junto con la materia fecal suspendida.

Como resultado obtuvimos que solo dos de estas bacterias fecales, colif\_fecales\_ufc\_100ml y escher\_coli\_ufc\_100ml, si tienen influencia sobre la materia fecal suspendida, porque su p-valor fue menor al 0,05, entonces rechazamos nuestra hipótesis nula (**H0**) pero que la bacteria enteroc\_ufc\_100ml no la tiene, porque su p-valor es mayor al 0,05, entonces esta bacteria no rechaza la hipótesis nula (**H0**)

Por lo tanto, la **H4 queda validada** parcialmente ya que no son todas las bacterias fecales quienes tienen influencia sobre la materia fecal suspendida.



*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 una diferencia de las medianas en relación a los grupos de las estaciones
* H1: La bacteria coliforme fecal presenta una diferencia de las medianas en relación a los grupos de las estaciones

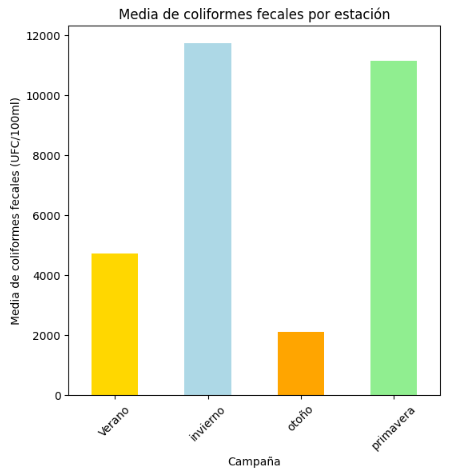
Comenzamos realizando un gráfico de barras diferenciando por las estaciones del año, mostrando la cantidad de coliformes fecales en cada una de dichas estaciones.

En el mismo pudimos observar a simple vista que la bacteria si posee un comportamiento estacional.

Continuamos con la evaluación de normalidad mediante Shapiro Wilk y homocedasticidad mediante Levene y obtuvimos que los supuestos eran no normales, porque su p-valor era menor 0.05, entonces se rechaza la hipótesis nula de normalidad.

Además, los supuestos resultaron no homocedásticos porque su p-valor fue menor al 0.05, entonces se rechaza la hipótesis nula de homocedasticidad.

Continuamos con el test no paramétrico de Kruskal Wallis ya que “campaña” posee más de dos grupos. El resultado obtenido validó lo visto en el gráfico de barras,ya que el p-valor dió menor al 0.05, con lo cual quedó rechazada nuestra hipótesis nula (**H0**), la bacteria si presenta un comportamiento estacional.

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.*

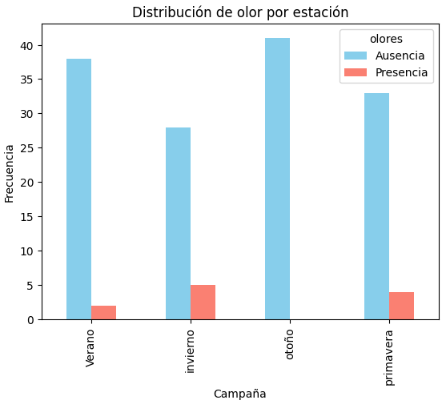
**H6: ¿El olor del agua del Río de La Plata tiene un comportamiento estacional?**

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

Comenzamos realizando un gráfico de barras diferenciando por las estaciones del año, mostrando la existencia o no de olor en cada una de dichas estaciones.

Pudimos observar con las pocas muestras que tenemos del tipo ‘presencia’ que no existe un comportamiento estacional.

Continuamos aplicando la prueba de chi-cuadrado, el cual nos arroja un resultado, con p-valor > 0,05, con lo cual se rechaza nuestra hipótesis nula (**H0**), pero no podemos afirmar que el olor tiene una diferencia significativa en cuanto a los grupos de estaciones, ya que no contamos con datos suficientes del tipo ‘Presencia’.



*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 una diferencia en relación a los grupos de estaciones.
* H1 = La materia fecal suspendida en el agua presenta una diferencia en relación a los grupos de estaciones.

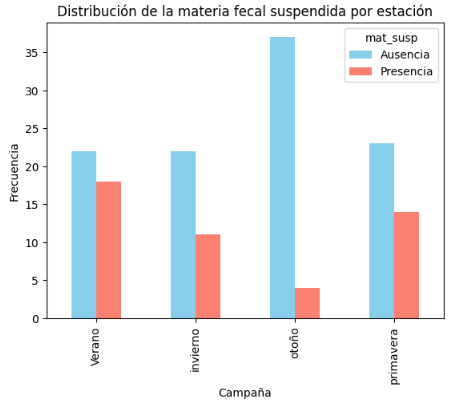
Comenzamos realizando un gráfico de barras diferenciando por las estaciones del año, mostrando la existencia o no de materia fecal suspendida en cada una de dichas estaciones.

Pudimos observar que existe una diferencia en cuanto a las estaciones en el comportamiento de materia fecal suspendida.

Continuamos aplicando la prueba de chi-cuadrado, el cual nos arroja un resultado, con p-valor < 0,05, por ende no se rechaza nuestra hipótesis nula (**H0**).

Entonces afirmamos que la materia fecal suspendida si presenta un comportamiento estacional.

La **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.*

**CONCLUSIÓN:** Los resultados obtenidos destacan algunas relaciones clave:

* **Influencia en la turbiedad**: Se observó que la turbiedad del agua está significativamente influenciada por la concentración de clorofila y la demanda química de oxígeno. Esto sugiere que la materia orgánica y la demanda química de oxígeno para oxidarla contribuyen a la turbiedad del agua del río.
* **Determinantes del índice de calidad del agua**: El índice de calidad del agua no está determinado por todos los componentes fisicoquímicos registrados, sino que se ve principalmente afectado por la presencia de bacterias fecales, específicamente escherichia coli, coliformes fecales y enterococos.
* **Relación entre materia fecal suspendida y bacterias fecales**: La cantidad de materia fecal suspendida en el agua del río parece estar influenciada principalmente por las bacterias fecales coliformes y escherichia coli, lo que refuerza la relación entre la presencia de estos microorganismos y la contaminación fecal.
* **Estacionalidad de los coliformes fecales y materia suspendida**: Los datos muestran un comportamiento estacional en los niveles de coliformes fecales y por ende en la materia suspendida. Sin embargo, no se pudo evaluar una relación estacional en el olor del agua debido a la falta de datos suficientes en esa variable.