|  |  |
| --- | --- |
| Ciencia y analítica de datos  RETO: Clasificación de aguas Subterráneas. | Abstract  La Comisión Nacional del Agua (CONAGUA) lleva a cabo, a través de la Red Nacional de Medición de Calidad del Agua, el monitoreo de los principales cuerpos de agua del país. En este trabajo se analiza la calidad de agua del año 2020 para crear un clasificador con ayuda de un modelo de inteligencia artificial que ayude a determinar si los cuerpos analizados son seguros para el consumo humano o no.  Karina Zafra Vallejo A01793979;Francisco Javier Parga Garcia A01794380  Profesor: Dr. María de la Paz Rico Fernandez  **Equipo 37**  Karina Zafra Vallejo *A01793979*  Francisco Javier Parga Garcia *A01794380*  TC4029 |

Contenido

[1. Base de datos 2](#_Toc119713609)

[2. Análisis exploratorio de datos 2](#_Toc119713610)

[2.1. Acondicionamiento inicial de los datos 2](#_Toc119713611)

[2.2. Datos faltantes 3](#_Toc119713612)

[2.3. Distribución de los datos continuos 3](#_Toc119713613)

[2.4. Análisis de la variable CONTAMINANTES: 6](#_Toc119713614)

[2.5. Acondicionamiento adicional de datos 6](#_Toc119713615)

[2.6. Correlación entre variables numéricas 7](#_Toc119713616)

[3. Geolocalización de las aguas subterráneas 8](#_Toc119713617)

[4. Agrupación por Kmeans 9](#_Toc119713618)

[4.1. Geolocalización y cantidad de contaminantes 12](#_Toc119713619)

[5. Conclusiones 17](#_Toc119713620)

[6. RETO PARTE 2 21](#_Toc119713621)

[6.1. Transformación de variables 21](#_Toc119713622)

[6.2. Partición de datos 21](#_Toc119713623)

[6.3. Pipeline 21](#_Toc119713624)

[6.4. Modelos clasificadores 22](#_Toc119713625)

[**Decision Tree** 22](#_Toc119713626)

[**Random Forest** 24](#_Toc119713627)

[6.5. Reporte de clasificación 26](#_Toc119713628)

[Decisión Tree 26](#_Toc119713629)

[Random Forest 26](#_Toc119713630)

[CONCLUSIONES: 27](#_Toc119713631)

# Base de datos

Para el siguiente análisis, utilizaremos la base de datos:

Datos\_de\_calidad\_del\_agua\_de\_sitios\_de\_monitoreo\_de\_aguas\_subterraneas\_2020.csv

Tenemos un total de 57 columnas y 1068 datos, de las cuales 39 son categóricas y 18 son numéricas.

Adicionalmente, también sabemos que la variable *PERIODO* solo es 2020. Así que más adelante se podría descartar ya que no aporta.

Fuente: CONAGUA (<https://files.conagua.gob.mx/aguasnacionales/Calidad%20del%20Agua%20Subterranea%20p.xlsx>)

# Análisis exploratorio de datos

Procedemos primero a observar la variable semáforo que sería nuestra variable de salida. Tenemos la siguiente distribución:

Chart, bar chart

Description automatically generated

Figura 2‑1 Distribución de la variable SEMÁFORO (variable objetivo para clasificación)

## Acondicionamiento inicial de los datos

Se identificó que las columnas que miden contaminantes en mg/L tienen valores asignados con texto como “<0.005” por ello se decidió transformarlo a numéricos removiendo los caracteres “<” & “<=”(ver Figura 2‑2)

Ejemplo: 

Figura ‑ Transformación de texto a valores numéricos

## Datos faltantes

Comenzamos con la identificación de los datos faltantes, con el objetivo de determinar qué tratamiento se les dará a estos datos: si se eliminarán o se realizará algún método de imputación.

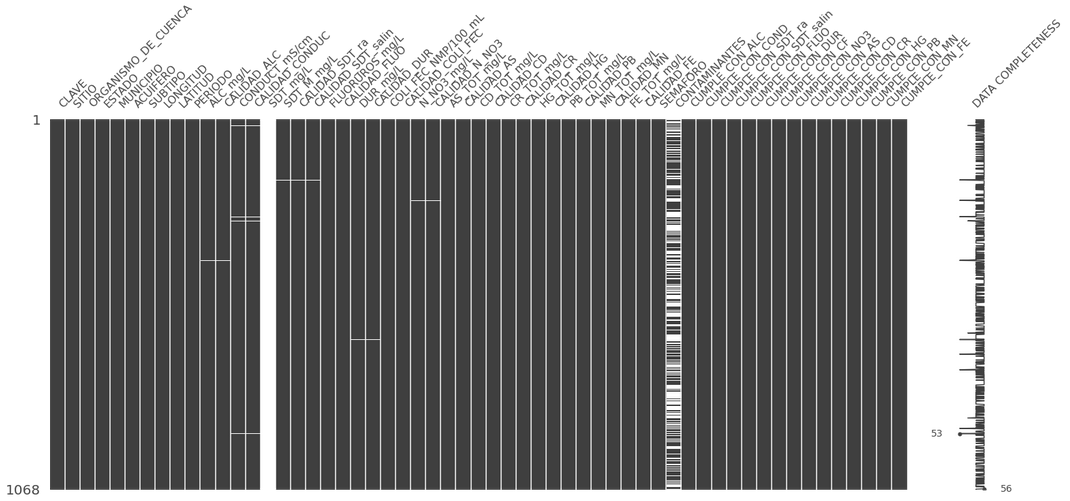


Figura ‑ Representación visual de datos faltantes (color blanco representa un dato faltante en la variable)

De esta gráfica podemos concluir que:

* La columna ‘SDT\_mg/L’ no tiene datos.
* Al parecer la columna ‘CONTAMINANTES’ tiene muchos datos faltantes, sin embargo vamos a analizar a más profundidad esta variable para identificar que sucede.

## Distribución de los datos continuos

Con las variables continuas, generamos los histogramas y boxplot para ver la distribución de los datos continuos.

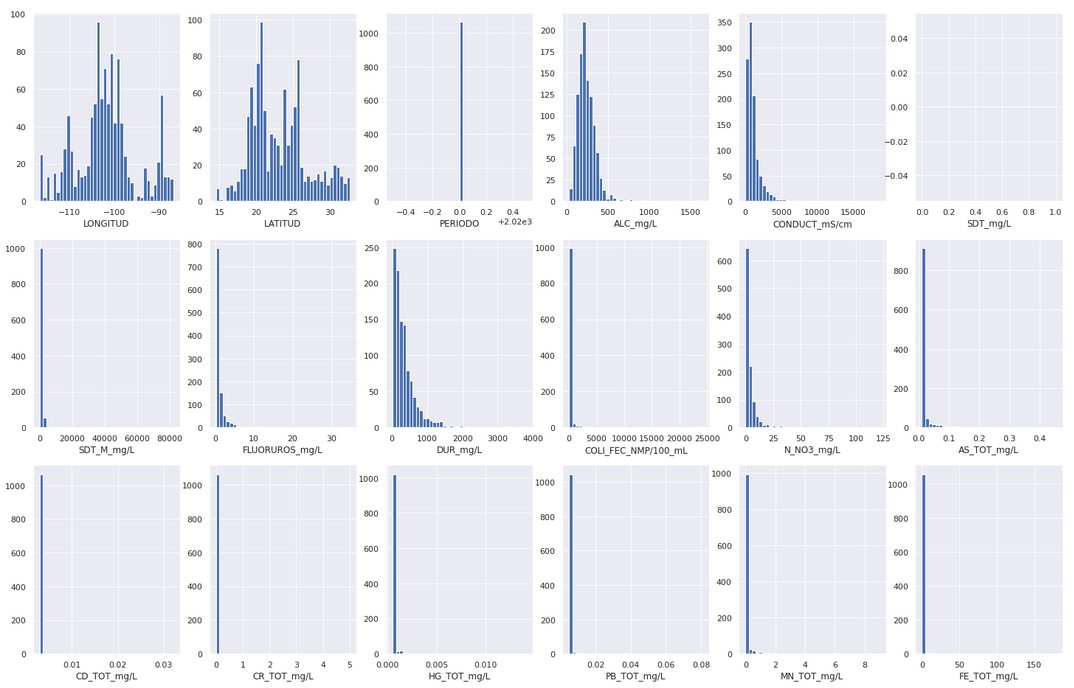


Figura ‑ Histogramas de las variables continuas.

Podemos observar que los datos se encuentran en diferentes magnitudes y escalas.

Calendar

Description automatically generated

Figura ‑ Gráfica de caja de bigotes para las variables continuas. Se observan varios datos extremos y mayormente un sesgo a la derecha.

Adicionalmente, como se observó anteriormente, ‘PERIODO’ tiene un solo valor (2020) y ‘SDT\_mg/L’ no tiene valores

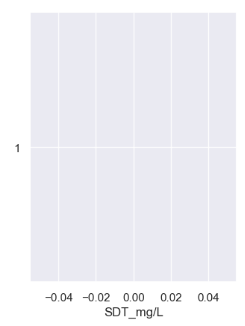
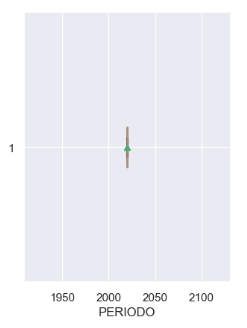


Figura ‑ Variables sin aporte de información

Se realiza una primera matriz de correlación entre las variables numéricas, se observa nuevamente que ‘PERIODO’ y ‘SDT\_mg/L’ no aportan información útil.

Chart

Description automatically generated with medium confidence

Figura ‑ Cajas de bigotes en función de la clasificación de semáforo.

Para más información sobre los promedios, dispersión, máximos y mínimos de las variables, ver las Tabla 2 y Tabla 3.

## Análisis de la variable CONTAMINANTES:

Esta variable contiene una lista de tipos contaminantes por cada toma

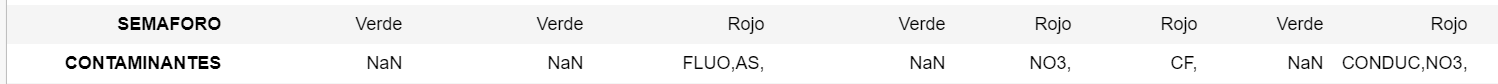


Figura ‑ Ejemplo de tipos de contaminantes por dato. El semáforo es altamente dependiente de los contaminantes.

Vamos a convertir esta lista de contaminantes por un número que representará la cantidad de contaminantes de cada muestra.

Al hacer el conteo de la lista, los datos *NaN* es porque no tienen contaminantes, por lo tanto, se convierte a 0.

Y graficamos para ver cuantos contaminantes en total tienen todas las muestras.

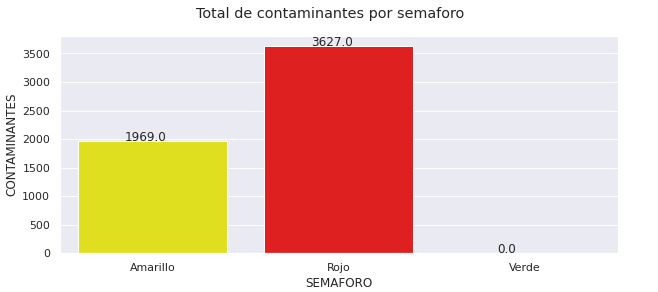


Figura ‑ Conteo de contaminantes por semáforo

## Acondicionamiento adicional de datos

Se crea una columna SEMAFORO\_cat que convierte el color en un número de la siguiente manera:

'Verde':1,

'Amarillo':2,

'Rojo':3

Adicionalmente, se eliminan las variables 'SEMAFORO', 'PERIODO', 'SDT\_mg/L' y guardamos el dataframe con un nuevo nombre.

## Correlación entre variables numéricas

Analizando las distribuciones y realizando gráficas cruzadas entre las variables, se observan algunas correlaciones lineales:

* Dureza, Contamiantes y sólidos se correlacionan con la conductividad
* El semáforo rojo se agrupa en función de FE, AS y MN en conjunto con los Fluoruros: La correlación no es lineal.

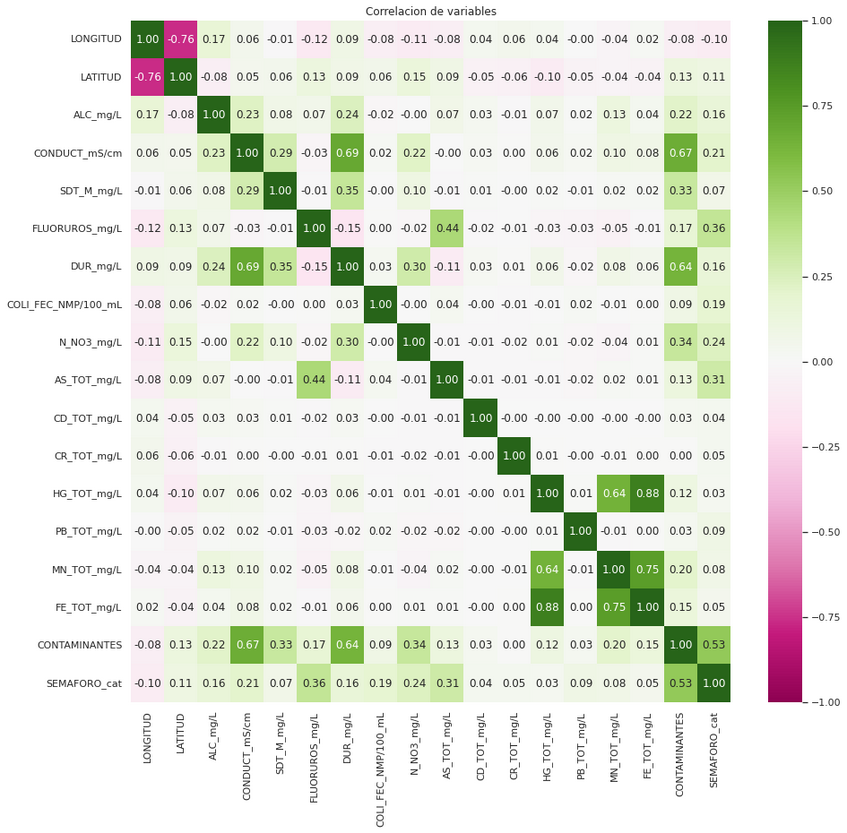


Figura ‑ Matriz de correlación

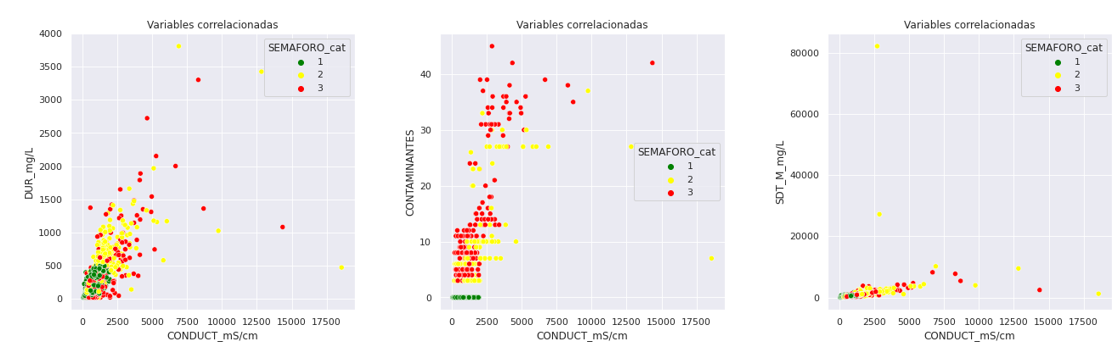


Figura ‑ Correlación de la Dureza, Contaminantes y Sólidos en fusión de la conductividad.

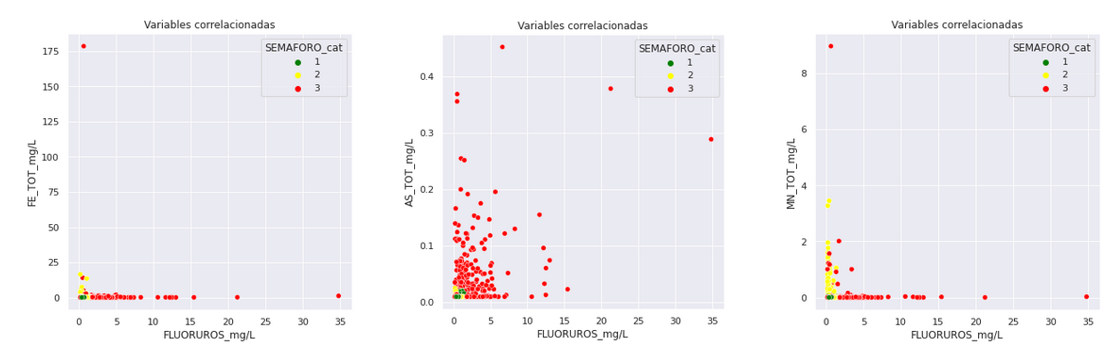


Figura ‑ FE, AS y MN en función de los Fluoruros. Se observa que el semáforo rojo se mantiene en valores altos de Fluoruros.



Figura ‑ Correlación del semáforo en función de las variables numéricas independientes.

Importante podemos destacar que la variable que más relacionada está con ‘SEMAFORO\_cat’ es ‘CONTAMINANTES’.

# Geolocalización de las aguas subterráneas

Utilizando la librería de geopandas y mapbox se puede localizar espacialmente los acuíferos subterráneos (ver Figura 3‑1)

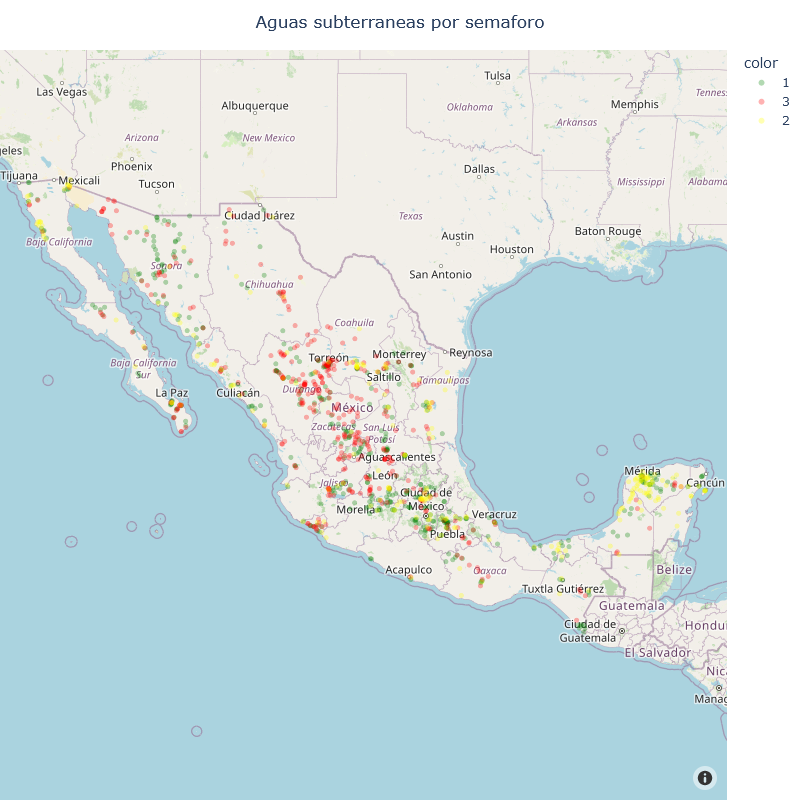


Figura ‑ Ubicación geográfica de las aguas subterráneas, codificadas por el semáforo y transparencia.

# Agrupación por Kmeans

Considerando los datos de Longitud, Latitud y la cantidad de contaminantes, se realizó un técnica de agrupación usando la librería de Kmeans.

Sin embargo, no se observa una dependencia espacial en latitud y longitud absoluta de la calidad del agua: ver Figura 4‑1, Figura 4‑2 y Tabla 1.

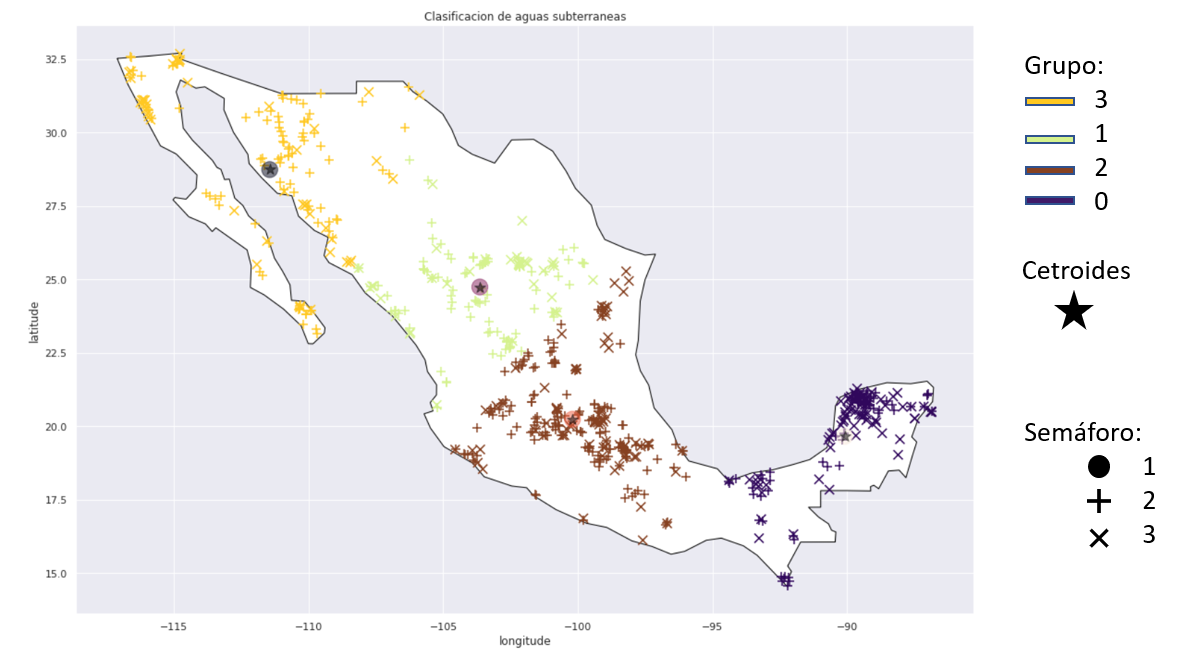


Figura ‑ Agrupación por geolocalización con 4 clústeres.

Map

Description automatically generated

Figura ‑ Clústeres geolocalizados

Chart, bar chart

Description automatically generated

Figura ‑ Distribución de semáforo en los grupos de 4 clústeres

Tabla Contenido de clases en cada clúster



## Geolocalización y cantidad de contaminantes

Se observó que los datos con semáforo verde no contienen contaminantes y al usar 30 clusters para agrupación se pueden separar completamente aquellas aguas subterráneas con semáforo verde de aquellas amarrillas y rojas, ver

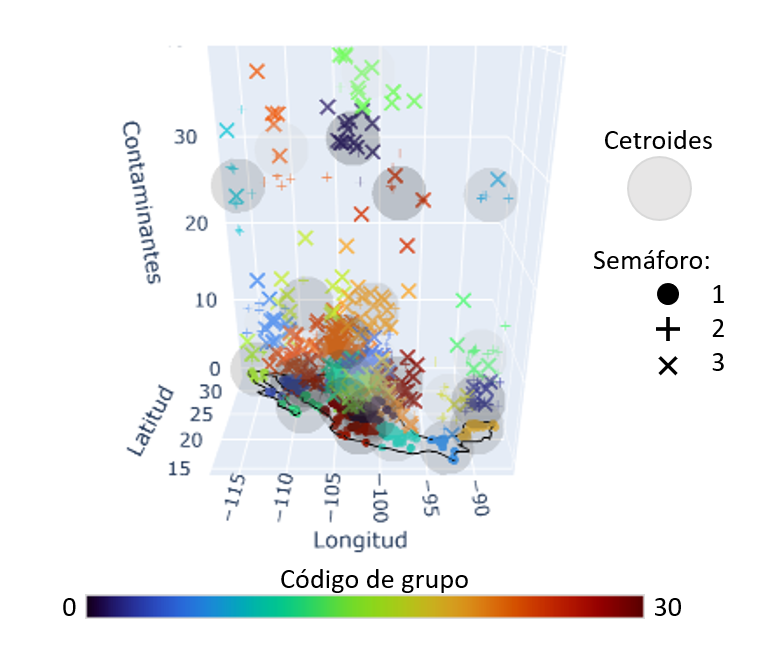


Figura ‑ Vista 3D de clústeres agrupados por localización espacial y cantidad de contaminantes.

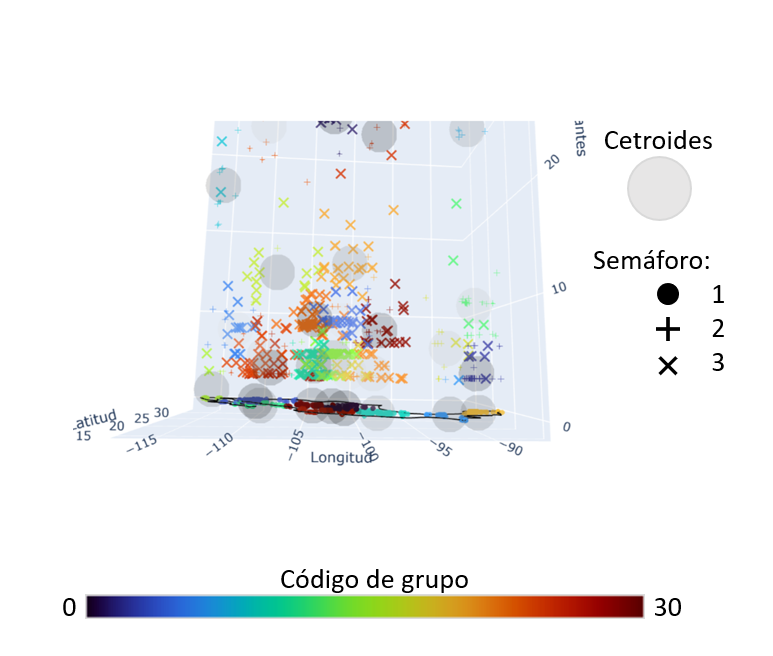


Figura ‑ Vista 3D de clústeres agrupados por localización espacial y cantidad de contaminantes: Desde perspectiva vertical. Se observa la separación de semáforo verde(1) respecto a los amarillos y rojos (2&3)

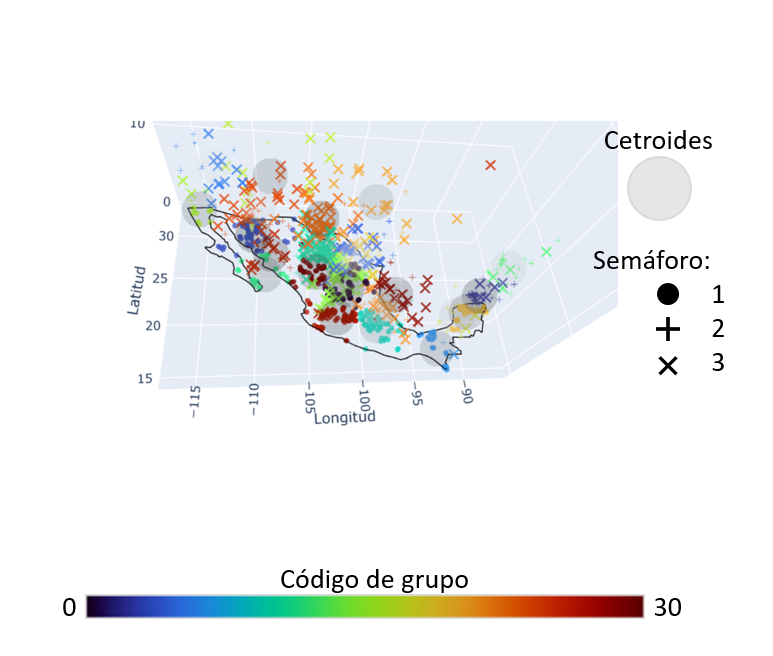


Figura ‑ Vista 3D de clústeres agrupados por localización espacial y cantidad de contaminantes.

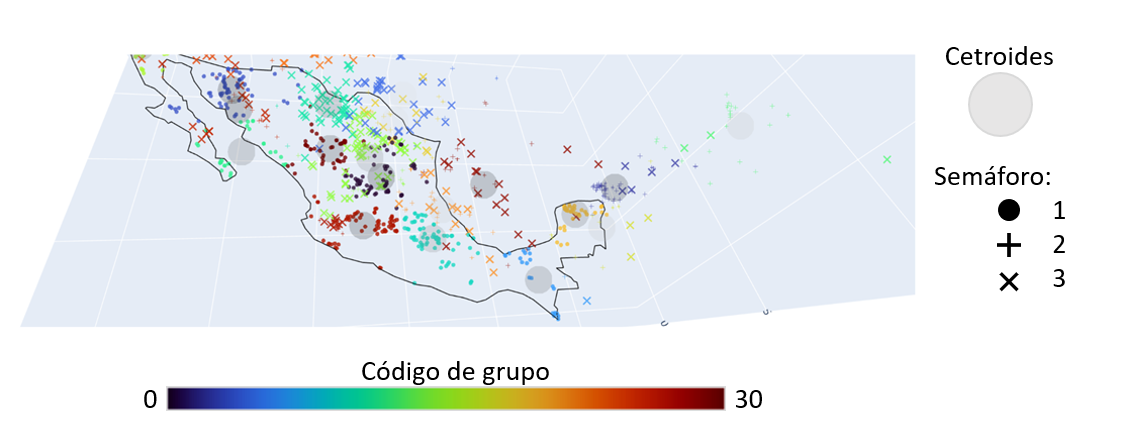


Figura ‑ Vista 3D de clústeres agrupados por localización espacial y cantidad de contaminantes.

Chart

Description automatically generated

Figura ‑ Distribución de semáforo en los grupos de 30 clústeres

# Conclusiones

* Se analizaron los datos de semáforo de calidad de agua y se observa que con 4 clústeres se identifica una agrupación coincidente con semáforo amarillo y verde hacia la península de Yucatán. El clúster #1 contiene una mayor cantidad de aguas con semáforo rojo. El clúster 2 contiene mayor concentración de aguas con semáforo verde, mientras que el clúster 3 tiene una distribución similar de las 3 clases (Tabla 1)
* Con 30 grupos se logra separar la mayoría de las aguas con semáforo verde. Sin embargo, con esta técnica de agrupación no se logra discernir de los semáforos amarillos y rojos.
* Se requiere de un modelo un poco más complejo para poder separar las clases adecuadamente, así como usar más variables como los Fluoruros, donde se observa cierta separación entre esos semáforos.

Anexos

Tabla Descripción de variables numéricas



Tabla Descripción de variables categóricas



# RETO PARTE 2

Para esta sección seleccionamos primero las variables feature X:

'ALC\_mg/L', 'CALIDAD\_ALC', 'CONDUCT\_mS/cm', 'CALIDAD\_CONDUC', 'SDT\_M\_mg/L', 'CALIDAD\_SDT\_ra', 'CALIDAD\_SDT\_salin', 'FLUORUROS\_mg/L', 'CALIDAD\_FLUO', 'DUR\_mg/L', 'CALIDAD\_DUR', 'COLI\_FEC\_NMP/100\_mL', 'CALIDAD\_COLI\_FEC', 'N\_NO3\_mg/L', 'CALIDAD\_N\_NO3', 'AS\_TOT\_mg/L', 'CALIDAD\_AS', 'CD\_TOT\_mg/L', 'CALIDAD\_CD', 'CR\_TOT\_mg/L', 'CALIDAD\_CR', 'HG\_TOT\_mg/L', 'CALIDAD\_HG', 'PB\_TOT\_mg/L', 'CALIDAD\_PB', 'MN\_TOT\_mg/L', 'CALIDAD\_MN', 'FE\_TOT\_mg/L', 'CALIDAD\_FE', 'CONTAMINANTES'

y la variable target Y que será 'SEMAFORO\_cat'

Adicionalmente se realiza una limpieza de datos NaN, pasamos de tener 1068 entradas a 1054.

## Transformación de variables

Para las variables categóricas se realiza el Label Encoder

Para las variables numéricas por el momento no se le realiza ninguna transformación, pero visualmente aplicamos la transformación logaritmo para ver el comportamiento de las variables ya que se observa que todas tienen sesgo para la derecha.

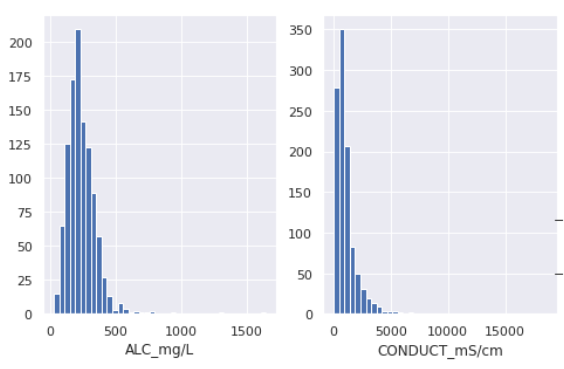
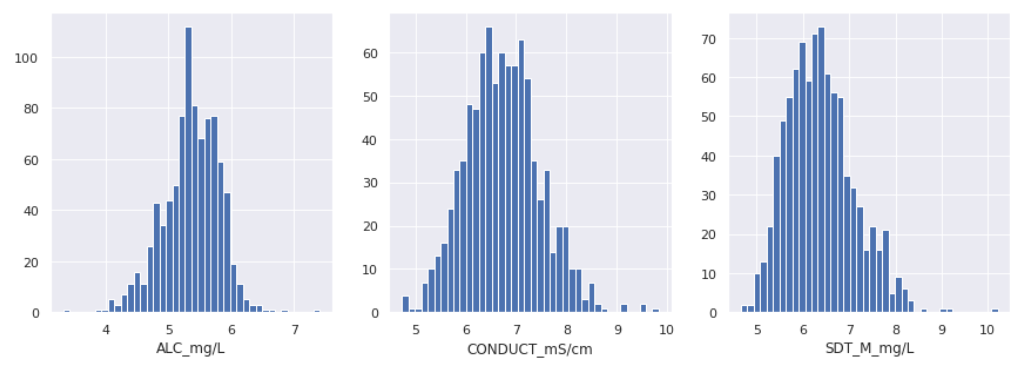
 

Ilustración -Histograma con los datos originales Ilustración - Histograma con la transformación logarítmica

## Partición de datos

Se genera la partición de datos en set de training (80%) y set de test (20%) y manteniendo la estratificación de la variable SEMAFORO, con el objetivo de mantener la distribución de las categorias durante el entrenamiento.

## Pipeline

Para las variables categóricas se establece imputación por la moda y para las numéricas por la mediana.

Adicionalmente para las numéricas se aplica transformación logarítmica.

## Modelos clasificadores

Realizamos primero un GridSearchCV para obtener los mejores parámetros, luego corremos un nuevo modelo usando los mejores parámetros obtenido anteriormente y por último generamos un modelo final con las variables más importantes obtenidas con el Feature Importance.

### **Decision Tree**

En el primer modelo tenemos:

train: 0.9964409982602437

test: 0.9881313473256601

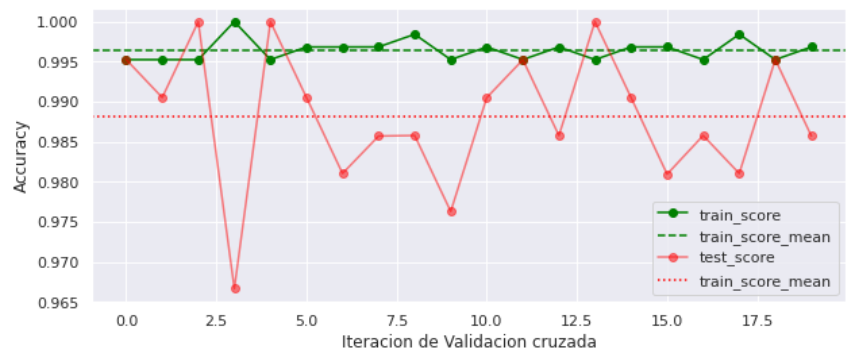


Ilustración -Gráfica de accuracy por cada iteración para decisión tree

Aplicamos un GridSearchCV para encontrar los mejores parámetros de este modelo y encontramos que para 20 particiones: n\_splits=4, n\_repeats=5 con la mejor combinación de parámetros, obtenemos:

Accuracy train final: 0.99644128113879

Accuracy test final: 1.0

Ahora procedemos a encontrar las feature importance

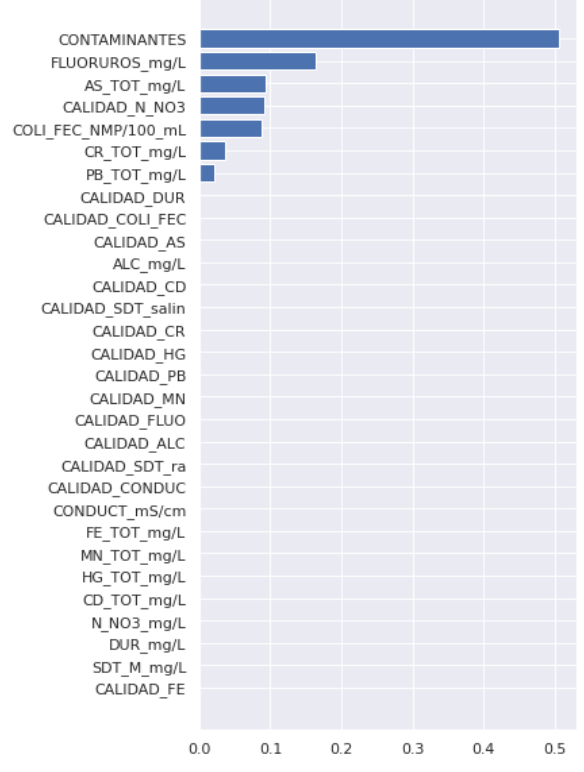


Ilustración - Variables más importantes para modelo DT

Encontramos que las variables seleccionadas, realmente solo 7 son las más importantes y dentro de estas podemos resaltar a ‘CONTAMINANTES’ y ‘FLUORUROS\_mg/L’ como las más relevantes.

Ahora se genera el modelo con los mejores hiperparámetros y solo con las variables más importantes, obteniendo:

train: 0.9962827704121425

test: 0.9895531482735276

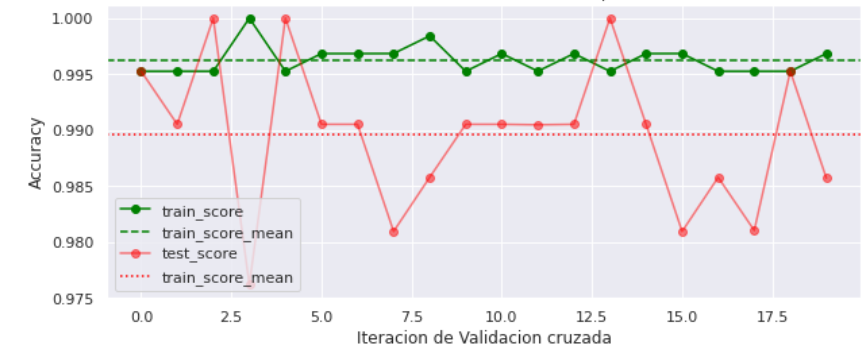


Ilustración -Gráfica de accuracy por cada iteración para Decision Tree

Para finalizar, evaluamos la matriz de confusión con los datos de prueba. Observamos que este modelo logra clasificar correctamente las clases.

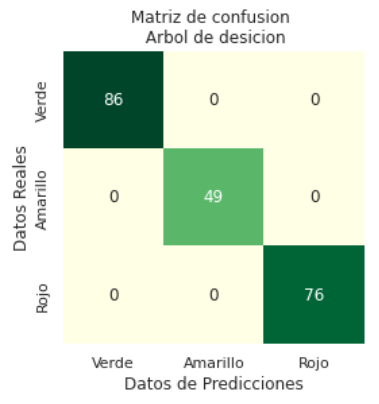


Ilustración - Matriz de confusión para modelo DT con best parameters del GridSearchCV y solo con las variables más importantes

### **Random Forest**

Se realiza el mismo procedimiento que para el Decision Tree obteniendo resultados muy similares:

train: 0.9969944207810908

test: 0.9900282103362674

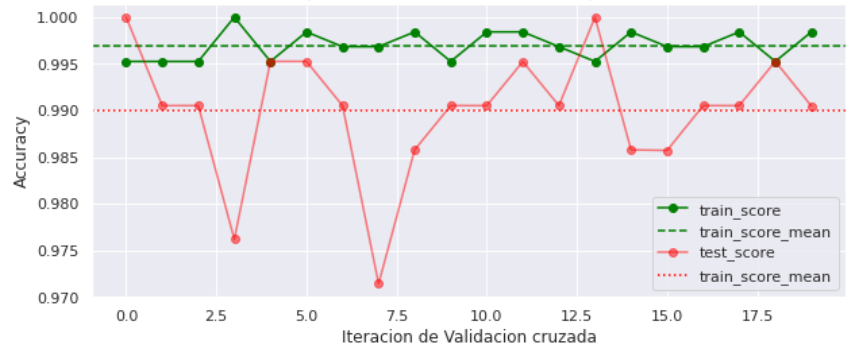


Ilustración -Gráfica de accuracy por cada iteración para Random Forest

Para finalizar, evaluamos la matriz de confusión con los datos de prueba. Observamos que este modelo logra clasificar correctamente las clases.

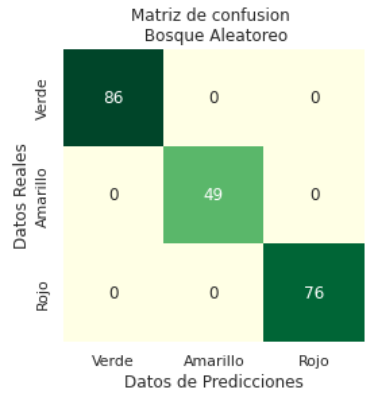
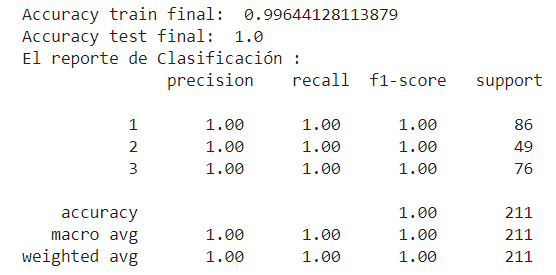


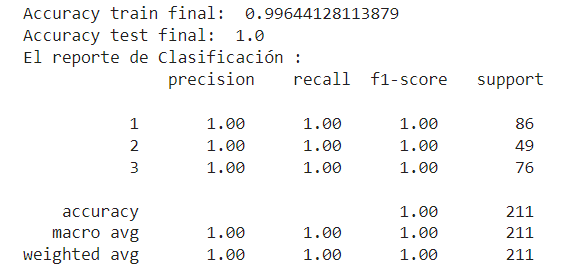
Ilustración -Matriz de confusión para modelo RF con best parameters del GridSearchCV y solo con las variables más importantes

## Reporte de clasificación

### Decisión Tree

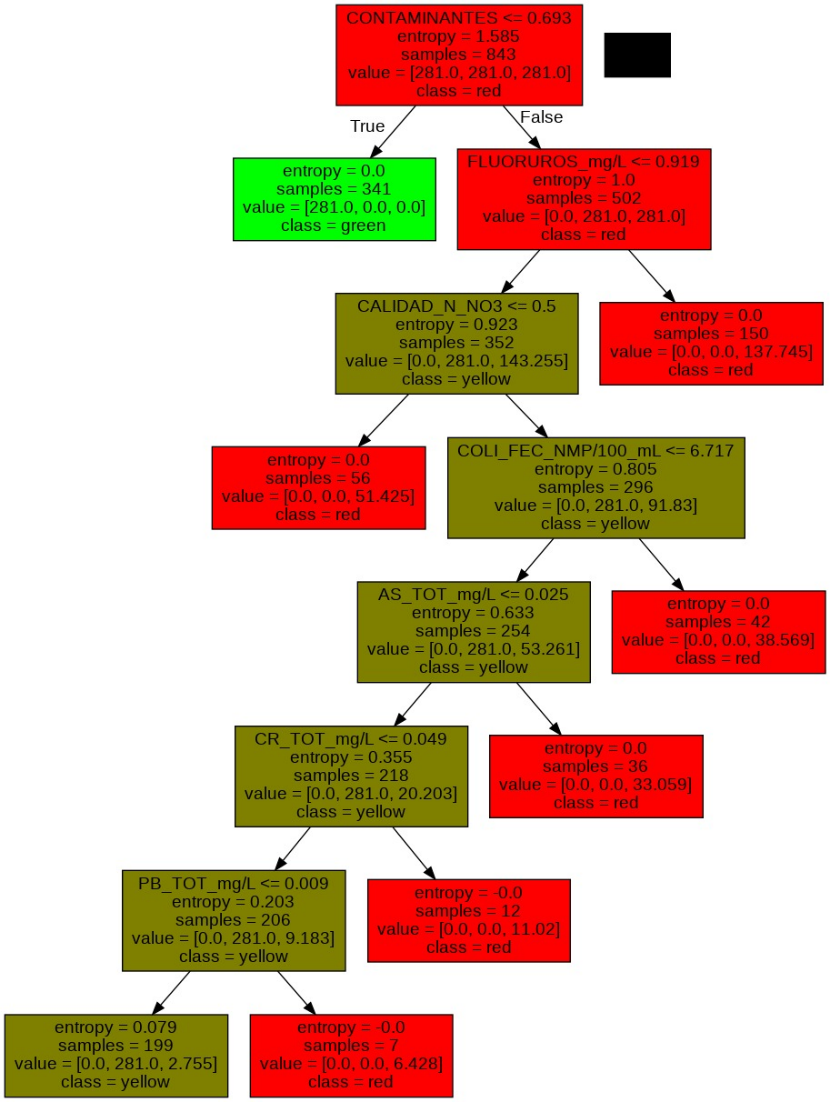


### Random Forest



Al final escogemos el Decision Tree ya que es más sencillo y arroja los mismos resultados que el Random Forest.

CONCLUSIONES:



De nuestra variable de salida ‘SEMAFORO\_cat’. La clase verde fue bastante sencilla de clasificar usando la variable ‘CONTAMINANTES’. Sin embargo, para el amarillo y el rojo, no dependía de la cantidad como lo habíamos determinado en los primeros análisis sino de los tipos de contaminantes, así que se requieren de más variables para poder clasificar correctamente entre estas dos clases. En el árbol se observa que se requieren de 6 variables adicionales para lograr minuciosamente esta separación, iniciando con ‘FLUORUROS’ que como vimos en las feature importance, tenía segundo grado de relevancia y en nuestro algoritmo logra separar 150 samples de las 153 que se separan entre las otras 5 variables restantes.