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.

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

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)

2. Análisis exploratorio de datos

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

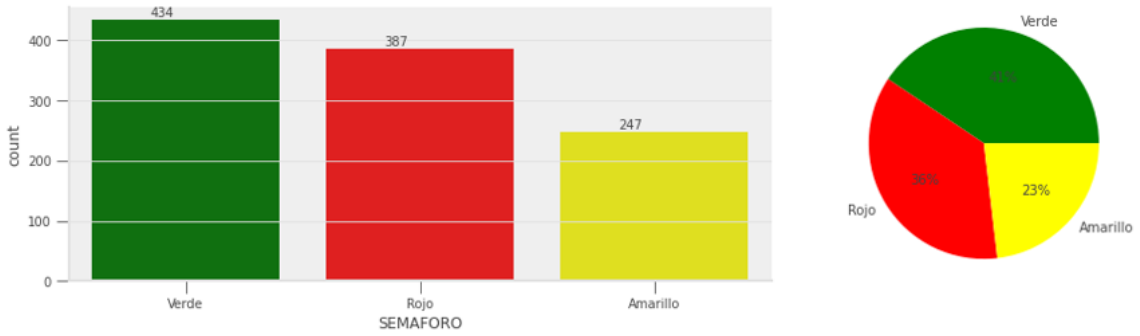


Figura 2-1 Distribución de la variable SEMÁFORO (variable objetivo para clasificación)

2.1. 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:

CALIDAD_CR	Potable - Excelente
HG_TOT_mg/L	< 0.0005

→

CALIDAD_CR	Potable - Excelente
HG_TOT_mg/L	0.0005

Figura 2-2 Transformación de texto a valores numéricos

2.2. 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.

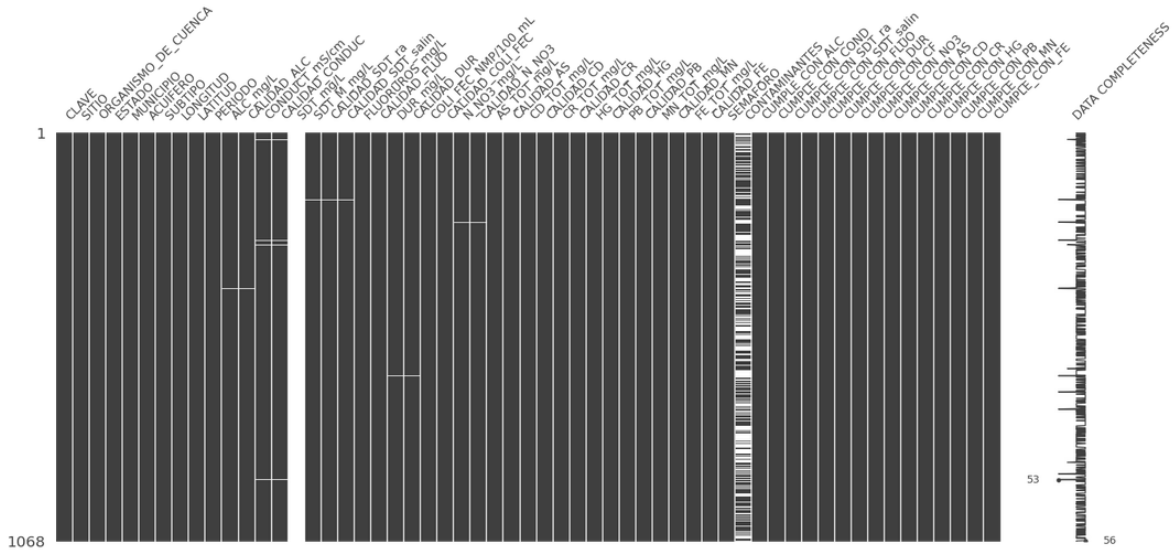


Figura 2-3 Representación visual de datos faltantes (color blanco representa un dato faltante en la variable)

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.

2.3. Distribución de los datos continuos

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

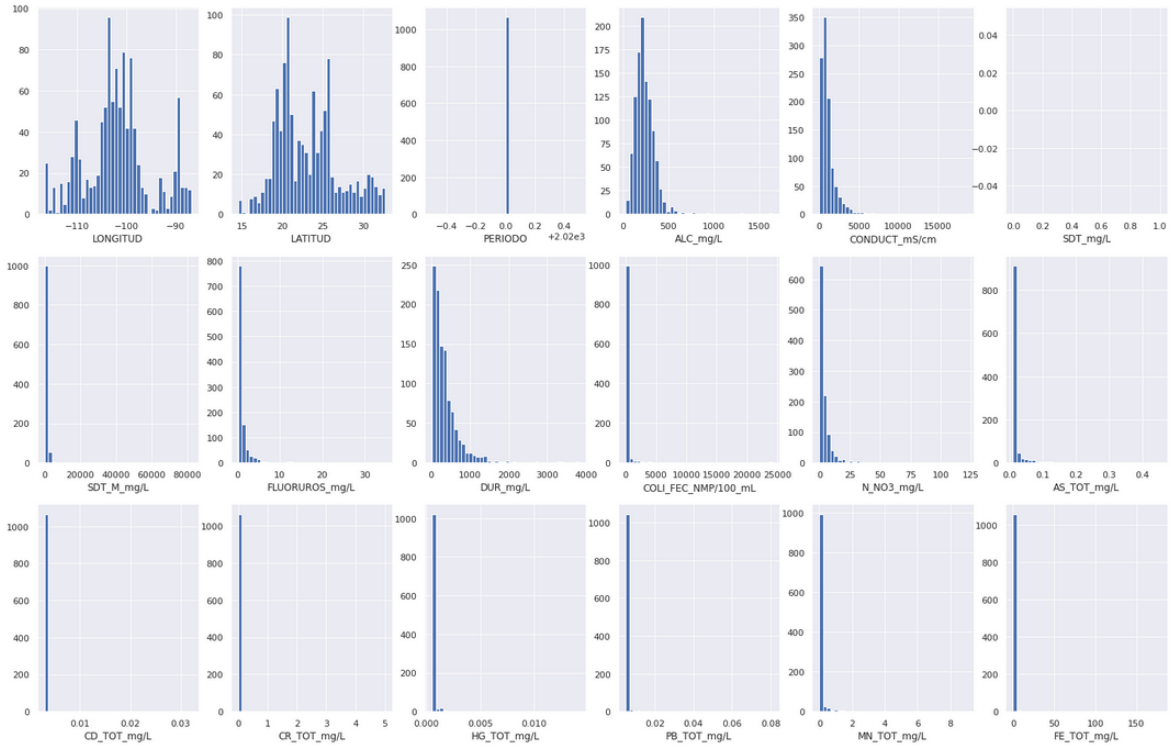


Figura 2-4 Histogramas de las variables continuas.

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

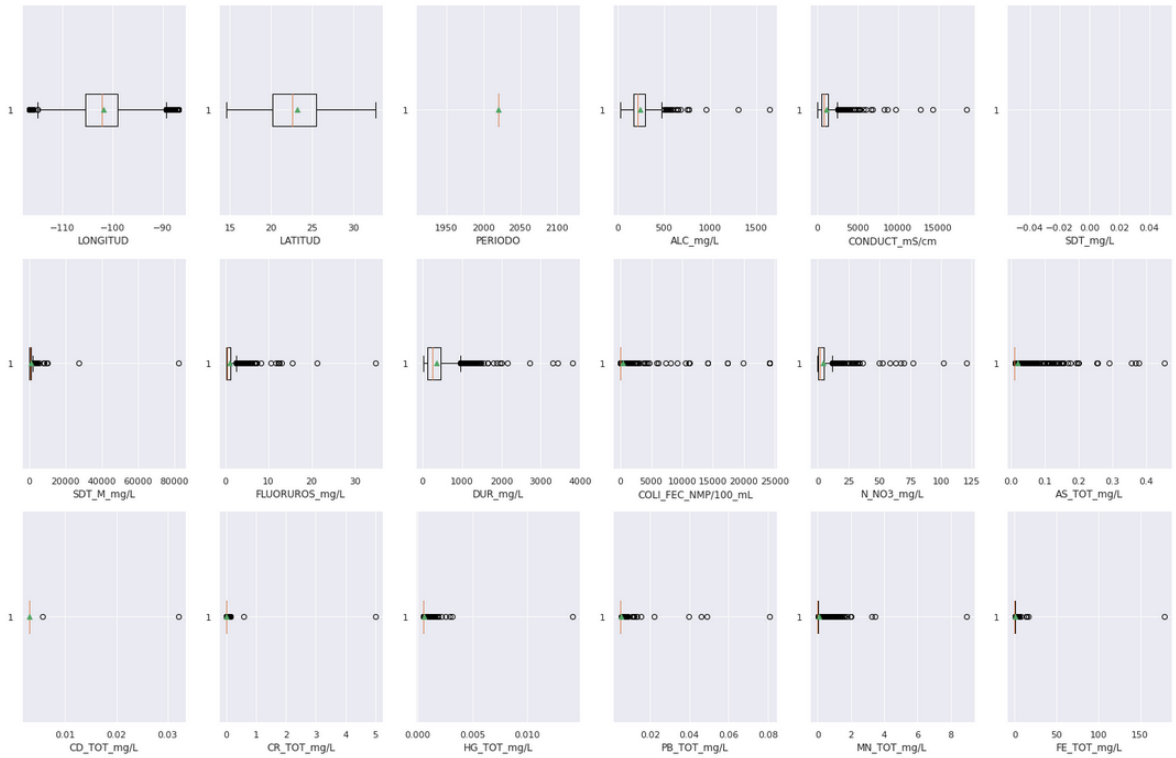


Figura 2-5 Gráfica de caja de bigotes para las variables continuas. Se observan varios datos extremos y mayormente un sesgo a la derecha.

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

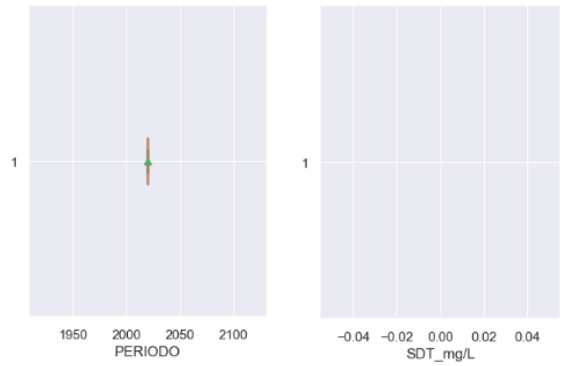


Figura 2-6 Variables sin aporte de información

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

Figura 2-7 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.

2.4.Análisis de la variable CONTAMINANTES:

Esta variable contiene una lista de tipos contaminantes por cada toma

SEMAFORO	Verde	Verde	Rojo	Verde	Rojo	Rojo	Verde	Rojo
CONTAMINANTES	NaN	NaN	FLUO,AS,	NaN	NO3,	CF,	NaN	CONDUC,NO3,

Figura 2-8 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.

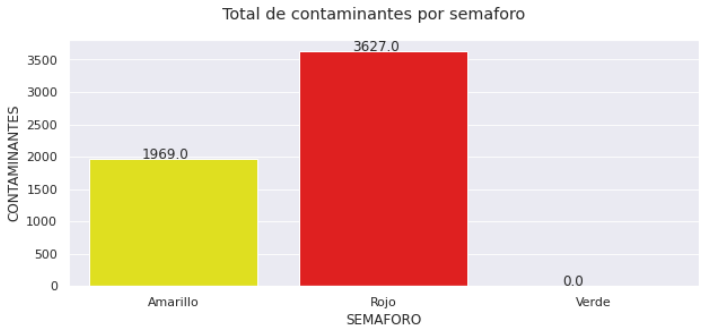


Figura 2-9 Conteo de contaminantes por semáforo

2.5.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.

2.6.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.

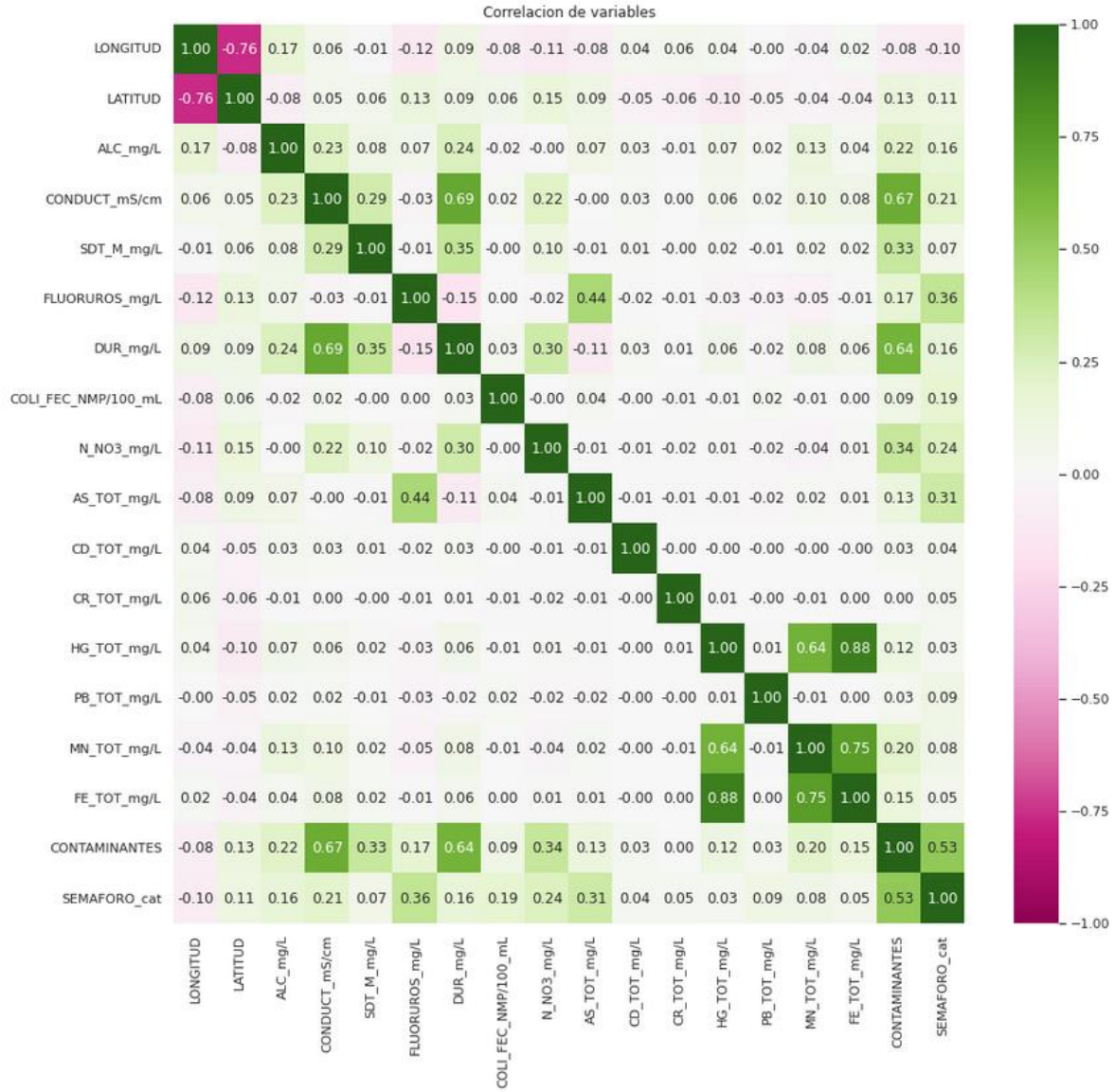


Figura 2-10 Matriz de correlación

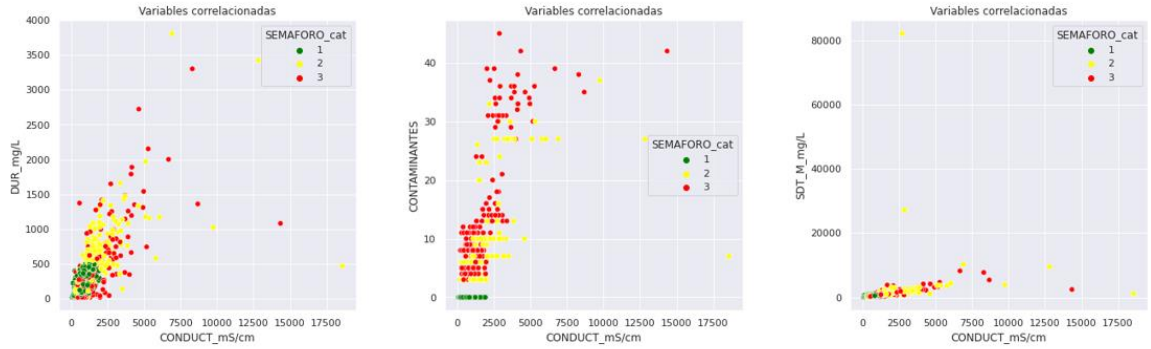


Figura 2-11 Correlación de la Dureza, Contaminantes y Sólidos en fusión de la conductividad.

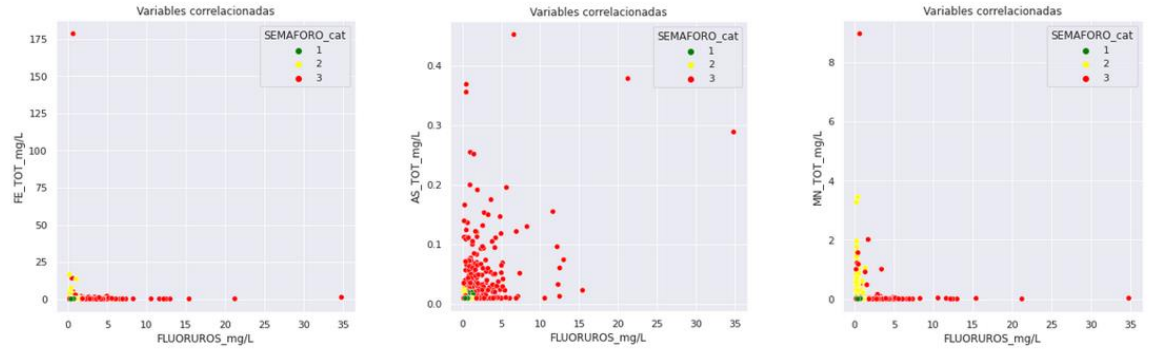


Figura 2-12 FE, AS y MN en función de los Fluoruros. Se observa que el semáforo rojo se mantiene en valores altos de Fluoruros.

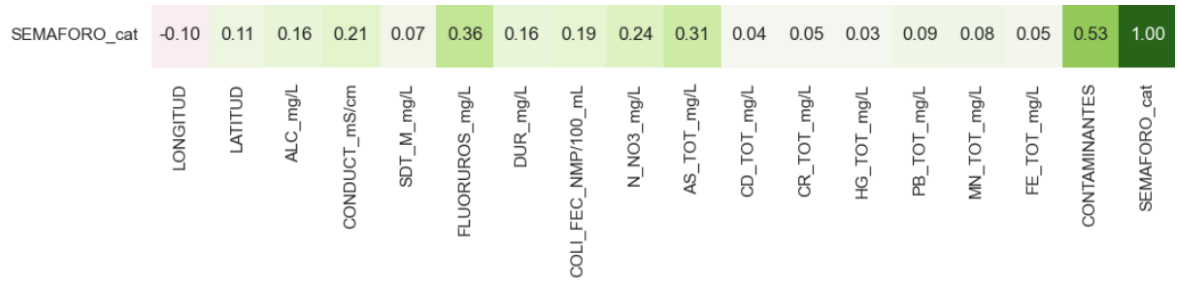


Figura 2-13 Correlación del semáforo en función de las variables numéricas independientes.

Importante podemos destacar que la variable que más relacionada está con 'SEMAFORO_cat' es 'CONTAMINANTES'.

3. Geolocalización de las aguas subterráneas

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

Figura 3-1 Ubicación geográfica de las aguas subterráneas, codificadas por el semáforo y transparencia.

4. 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.

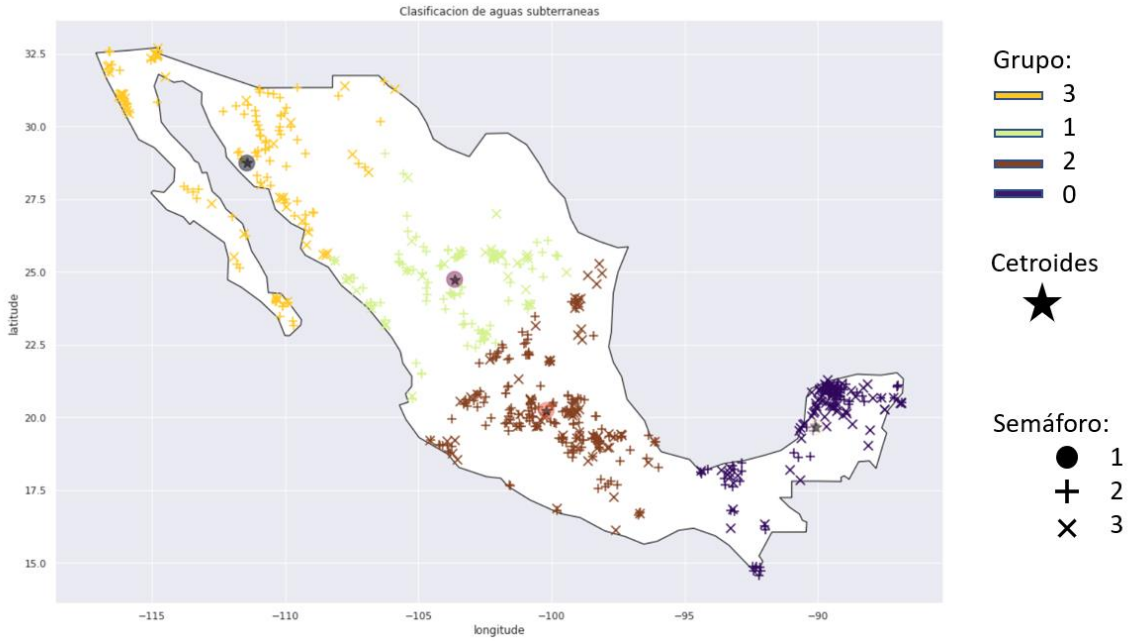


Figura 4-1 Agrupación por geolocalización con 4 clústeres.

Figura 4-2 Clústeres geolocalizados

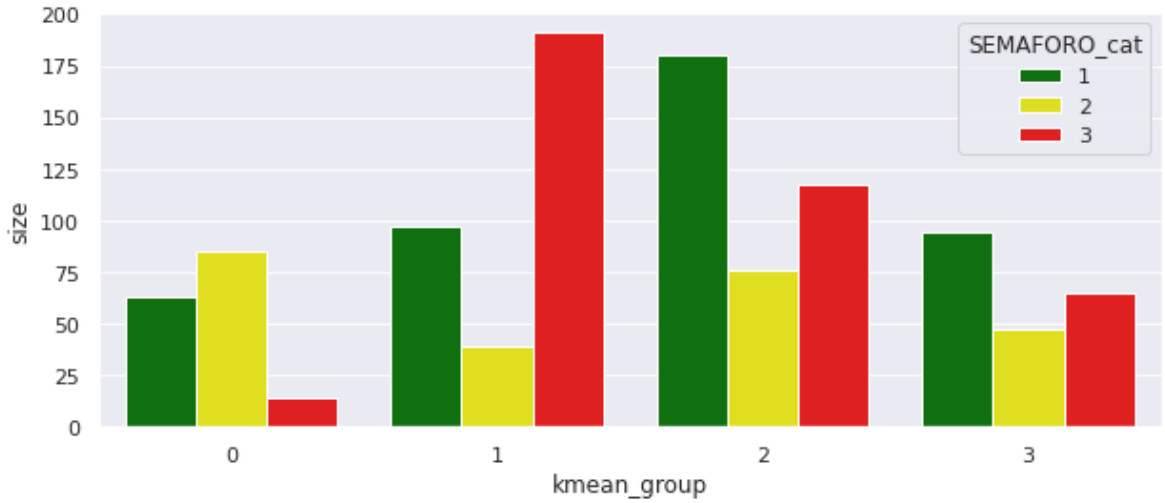


Figura 4-3 Distribución de semáforo en los grupos de 4 clústeres

Tabla 1 Contenido de clases en cada clúster

kmean_group	SEMAFORO_cat	size
0	1	63
0	2	**85**
0	3	14
1	1	97
1	2	39
1	3	**191**
2	1	**180**
2	2	76
2	3	117
3	1	**94**
3	2	47
3	3	65

4.1.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

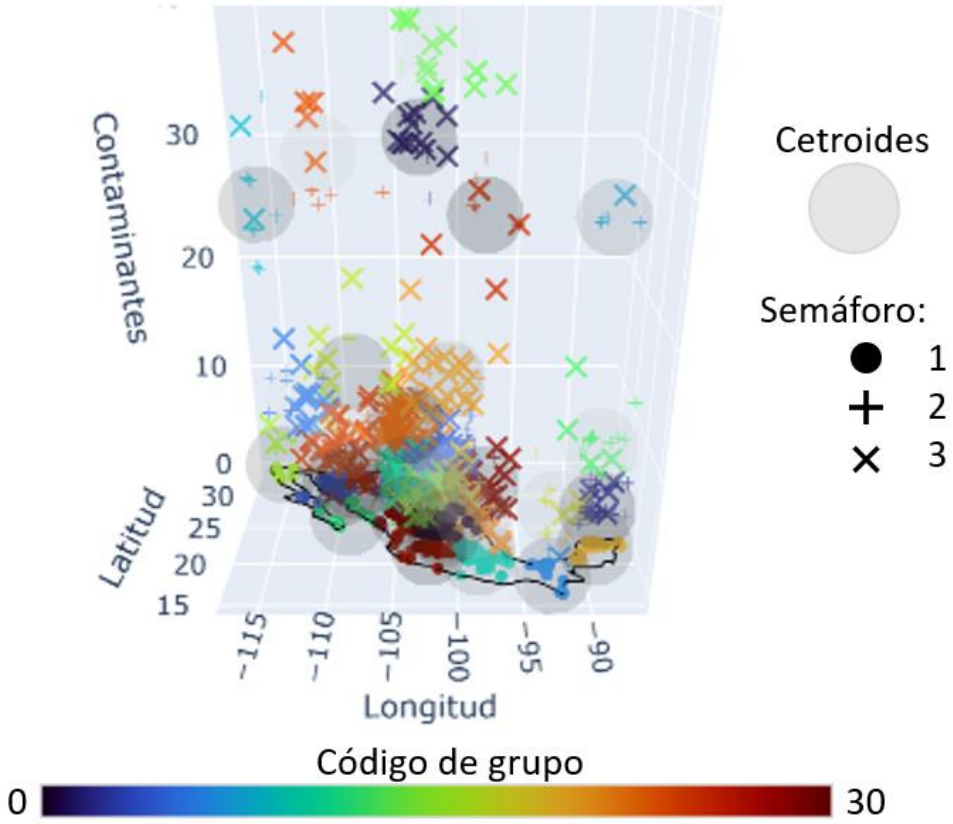


Figura 4-4 Vista 3D de clústeres agrupados por localización espacial y cantidad de contaminantes.

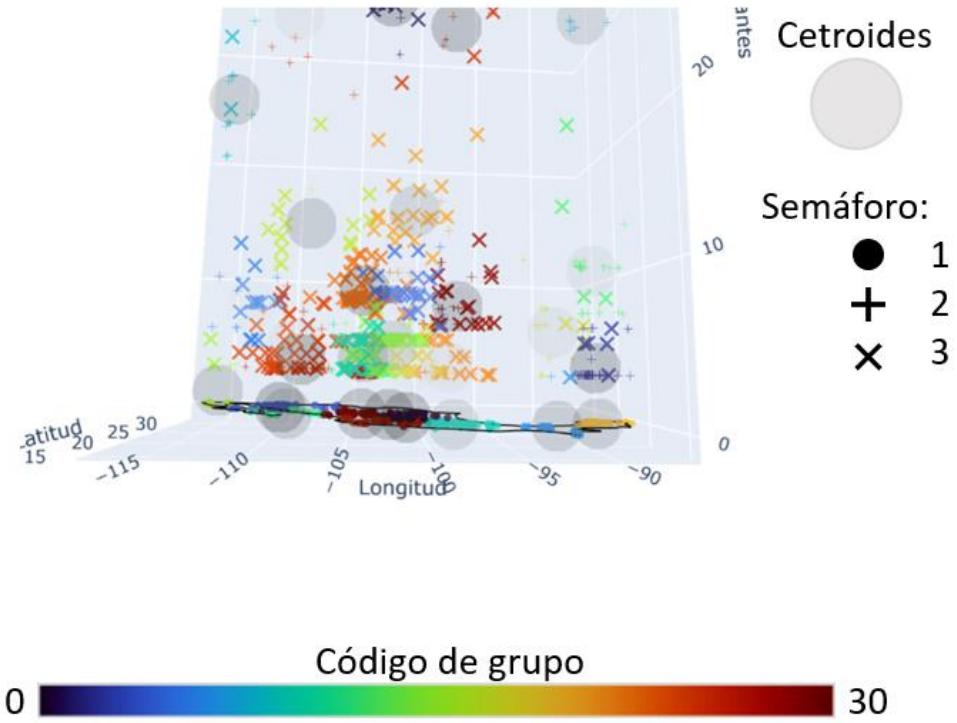


Figura 4-5 Vista 3D de clústeres agrupados por localización espacial y cantidad de contaminantes: Desde perspectiva vertical. Se observa la separación de semáforo verde(1) respecto a los amarillos y rojos (2&3)

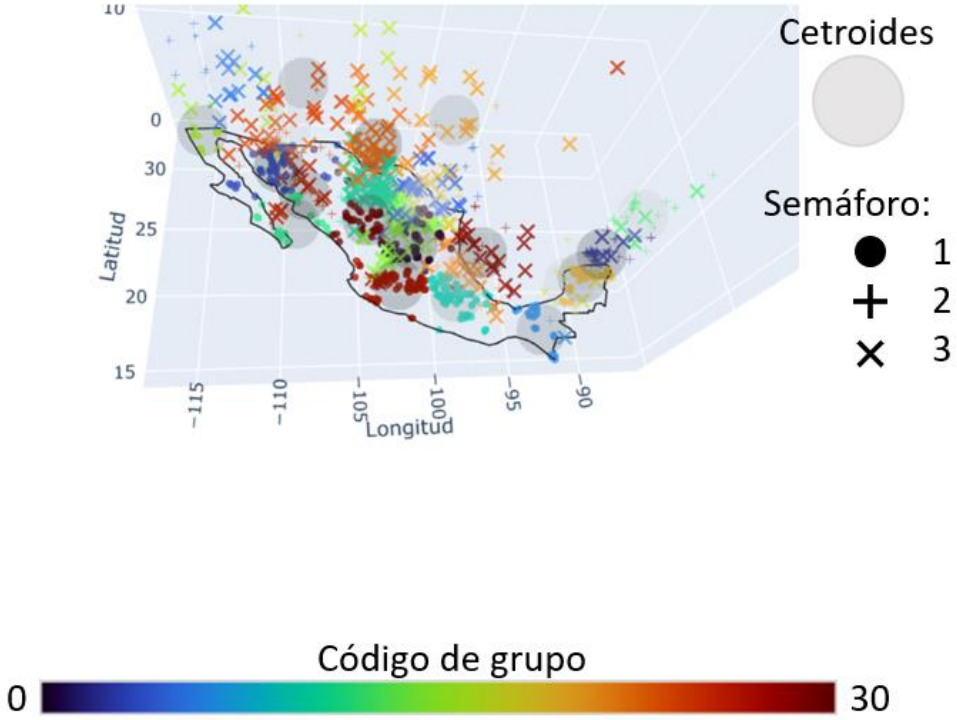


Figura 4-6 Vista 3D de clústeres agrupados por localización espacial y cantidad de contaminantes.

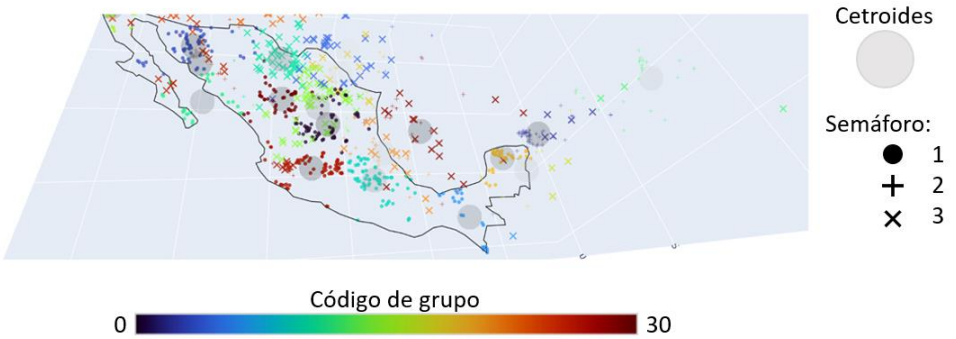


Figura 4-7 Vista 3D de clústeres agrupados por localización espacial y cantidad de contaminantes.

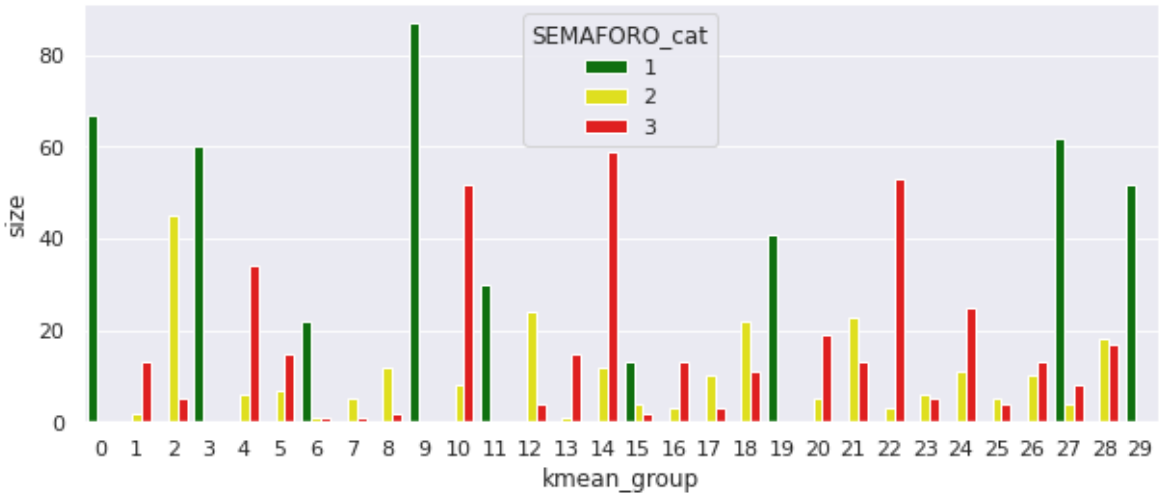


Figura 4-8 Distribución de semáforo en los grupos de 30 clústeres

5. 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 2 Descripción de variables numéricas

Variable	count	mean	std	min	25%	50%	75%	max
LONGITUD	1068	-101.891007	6.703263	-116.66425	-105.388865	-102.17418	-98.974716	-86.86412
LATITUD	1068	23.163618	3.88767	14.56115	20.212055	22.61719	25.510285	32.677713
PERIODO	1068	2020	0	2020	2020	2020	2020	2020
ALC_mg/L	1064	235.633759	116.874291	26.64	164	215.5275	292.71	1650
CONDUCT_ mS/cm	1062	1138.953013	1245.563674	50.4	501.75	815	1322.75	18577
SDT_mg/L	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
SDT_M_mg/	1066	896.101567	2751.53059	25	337.5	550.4	916.1	82170
FLUORUROS _mg/L	1068	1.0756	1.924278	0.2	0.267175	0.5035	1.13985	34.8033
DUR_mg/L	1067	347.938073	359.669452	20	121.1948	245.3358	453.93	3810.6922
COLI_FEC_N MP/100_mL	1068	355.490356	2052.457014	1.1	1.1	1.1	13.25	24196
N_NO3_mg/	1067	4.319759	8.345134	0.02	0.650294	2.080932	5.201698	121.007813
AS_TOT_mg/	1068	0.019618	0.035209	0.01	0.01	0.01	0.01	0.4522
CD_TOT_mg/	1068	0.00303	0.000894	0.003	0.003	0.003	0.003	0.03211
CR_TOT_mg/	1068	0.013276	0.154391	0.005	0.005	0.005	0.005	5.0032
HG_TOT_mg	1068	0.000557	0.000467	0.0005	0.0005	0.0005	0.0005	0.01415
PB_TOT_mg/	1068	0.005282	0.003254	0.005	0.005	0.005	0.005	0.0809
MN_TOT_mg	1068	0.072478	0.376512	0.0015	0.0015	0.0015	0.009947	8.982
FE_TOT_mg/	1068	0.410387	5.537974	0.025	0.025	0.04696	0.17338	178.615
CONTAMINA NTES	1068	5.2397	7.86702	0	0	3	7	45
SEMAFORO_	1068	1.955993	0.876076	1	1	2	3	3

Tabla 3 Descripción de variables categóricas

Variable	count	unique	top	freq
CLAVE	1068	1068	DLAGU6	1
SITIO	1068	1066	EL FUERTE	2
ORGANISMO_DE_CUENCA	1068	13	CUENCAS CENTRALES DEL NORTE	232
ESTADO	1068	32	DURANGO	121
MUNICIPIO	1068	452	LA PAZ	27
ACUIFERO	1068	273	PENINSULA DE YUCATAN	119
SUBTIPO	1068	8	POZO	1039
CALIDAD_ALC	1064	4	Alta	794
CALIDAD_CONDUC	1062	5	Permisible para riego	460
CALIDAD_SDT_ra	1066	5	Excelente para riego	491
CALIDAD_SDT_salin	1066	4	Potable - Dulce	834
CALIDAD_FLUO	1068	4	Baja	434
CALIDAD_DUR	1067	4	Potable - Dura	577
CALIDAD_COLI_FEC	1068	5	Potable - Excelente	739
CALIDAD_N_NO3	1067	3	Potable - Excelente	788
CALIDAD_AS	1068	3	Potable - Excelente	816
CALIDAD_CD	1068	2	Potable - Excelente	1066
CALIDAD_CR	1068	2	Potable - Excelente	1053
CALIDAD_HG	1068	2	Potable - Excelente	1067
CALIDAD_PB	1068	2	Potable - Excelente	1056
CALIDAD_MN	1068	3	Potable - Excelente	982
CALIDAD_FE	1068	2	Potable - Excelente	932
SEMAFORO	1068	3	Verde	434
CUMPLE_CON_ALC	1068	3	SI	1005
CUMPLE_CON_COND	1068	3	SI	939
CUMPLE_CON_SDT_ra	1068	3	SI	995
CUMPLE_CON_SDT_salin	1068	3	SI	995
CUMPLE_CON_FLUO	1068	2	SI	876
CUMPLE_CON_DUR	1068	3	SI	841
CUMPLE_CON_CF	1068	2	SI	1007
CUMPLE_CON_NO3	1068	3	SI	985
CUMPLE_CON_AS	1068	2	SI	941
CUMPLE_CON_CD	1068	2	SI	1066
CUMPLE_CON_CR	1068	2	SI	1053
CUMPLE_CON_HG	1068	2	SI	1067
CUMPLE_CON_PB	1068	2	SI	1056
CUMPLE_CON_MN	1068	2	SI	982
CUMPLE_CON_FE	1068	2	SI	932

6. 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`.

6.1. 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.

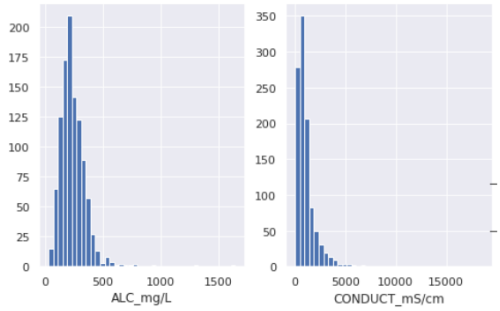


Ilustración 1-Histograma con los datos originales

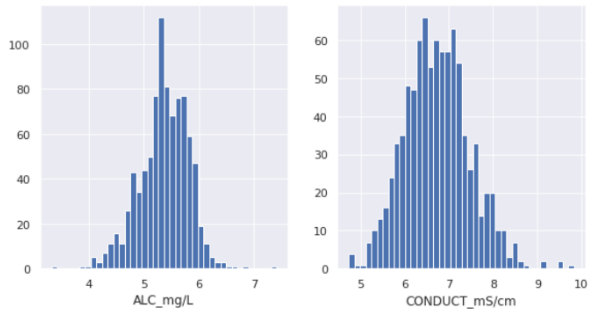


Ilustración 2- Histograma con la transformación logarítmica

6.2. 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.

6.3. 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.

6.4. 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 3-Gráfica de accuracy por cada iteración para decisión tree

`Aplicamos un GridSearchCV` para encontrar los mejores parámetros de este modelo y encontramos que p ara 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

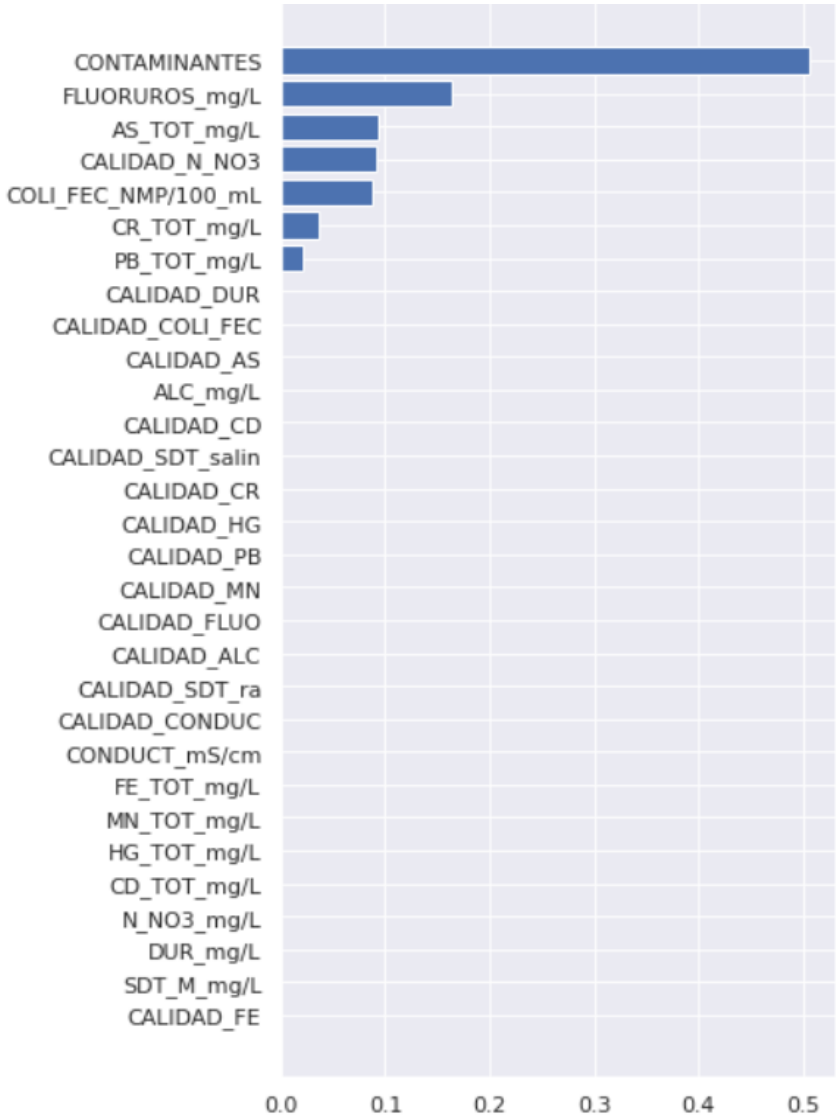


Ilustración 4- Variables más importantes para modelo DT

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
```

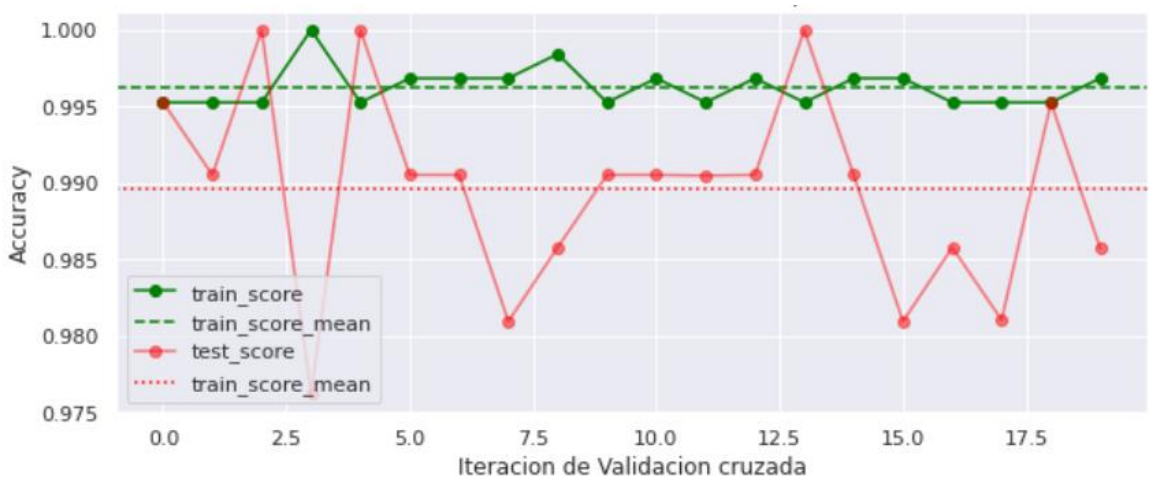


Ilustración 5-Gráfica de accuracy por cada iteración para Decision Tree

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

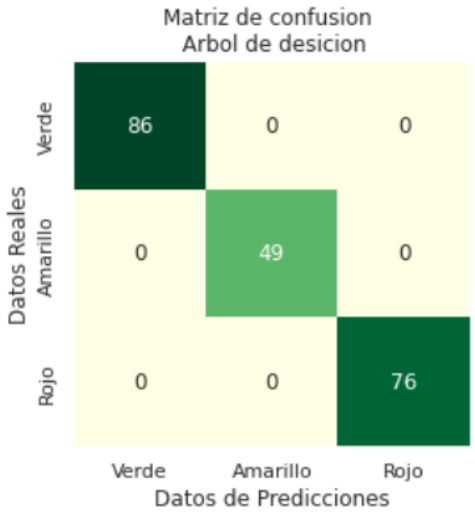


Ilustración 6- Matriz de confusión para modelo DT con best parameters del GridSearchCV y solo con las variables más importantes

Random Forest

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

`train: 0.9969944207810908`

`test: 0.9900282103362674`

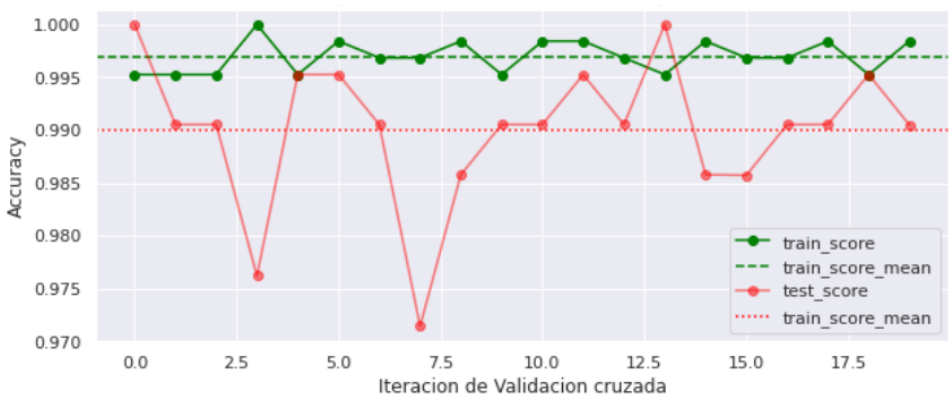


Ilustración 7-Gráfica de accuracy por cada iteración para Random Forest

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

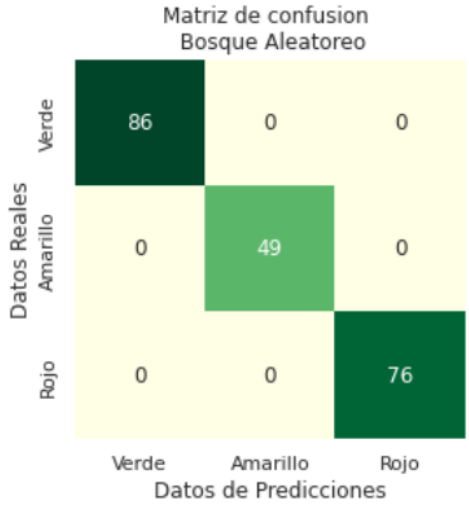


Ilustración 8-Matriz de confusión para modelo RF con best parameters del GridSearchCV y solo con las variables más importantes

6.5. Reporte de clasificación

Decisión Tree

```
Accuracy train final:  0.99644128113879
Accuracy test final:  1.0
El reporte de Clasificación :
              precision    recall  f1-score   support

           1       1.00      1.00      1.00        86
           2       1.00      1.00      1.00        49
           3       1.00      1.00      1.00        76

    accuracy                           1.00       211
   macro avg       1.00      1.00      1.00       211
weighted avg       1.00      1.00      1.00       211
```

Random Forest

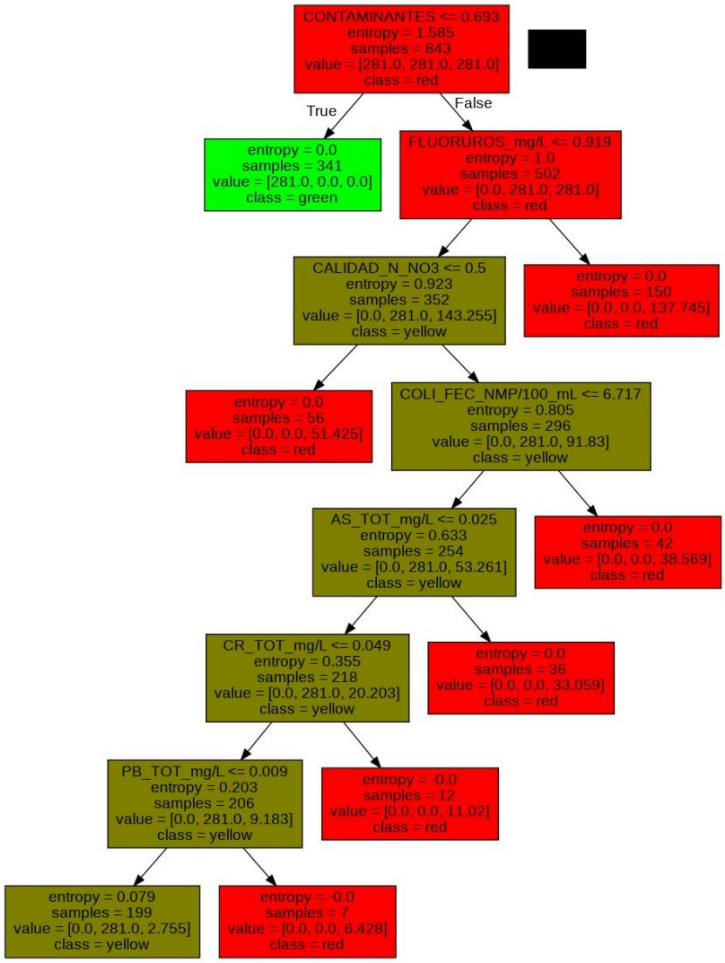
```
Accuracy train final:  0.99644128113879
Accuracy test final:  1.0
El reporte de Clasificación :
              precision    recall  f1-score   support

           1       1.00      1.00      1.00        86
           2       1.00      1.00      1.00        49
           3       1.00      1.00      1.00        76

    accuracy                           1.00       211
   macro avg       1.00      1.00      1.00       211
weighted avg       1.00      1.00      1.00       211
```

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.