Tecnológico de Monterrey

Proyecto Aguas subterráneas.

Alumno: **Erick de Jesus Hernández Cerecedo**

Matricula: **A01066428**

Alumno: **Francisco Javier Hernandez Camarillo**

Matricula: **A00998083**

Materia: **Ciencia y analítica de datos**

Profesor: **María de la Paz Rico**

Fecha: **Viernes 18 de noviembre de 2022**

Pipeline

Los pasos para las transformaciones fueron, aplicar:

- LabelEncoder():
  - Empleado para las columnas de tipo categórico.

- Standard Scaler():
  - Empleado para las columnas de tipo numérico y evitar que alguna feature tenga más peso que las demás.

Limpieza de los datos

En la limpieza de datos se optó por remover las columnas con datos vacíos usando la función “dropna” de pandas.

La columna llamada “SDT_mg/L” se removió por completo ya que carecía de datos en todos los registros.

La columna de contaminantes igualmente se removió ya que carecía del 60% de los registros.

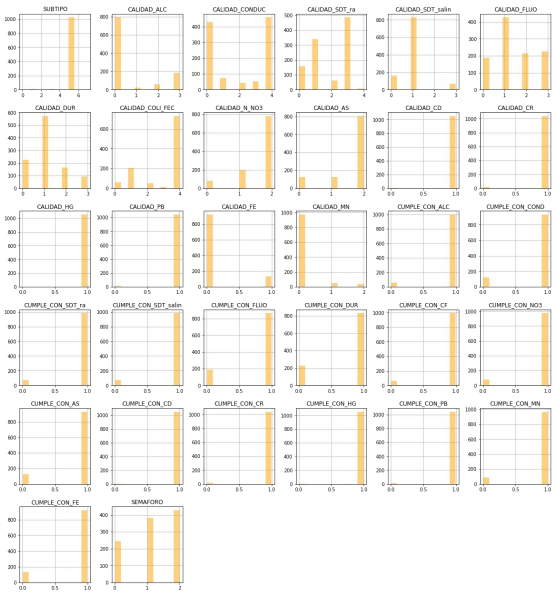
Se cambiaron el tipo de dato para las variables, numéricas float por float64 y las object se cambiaron a categoricos.

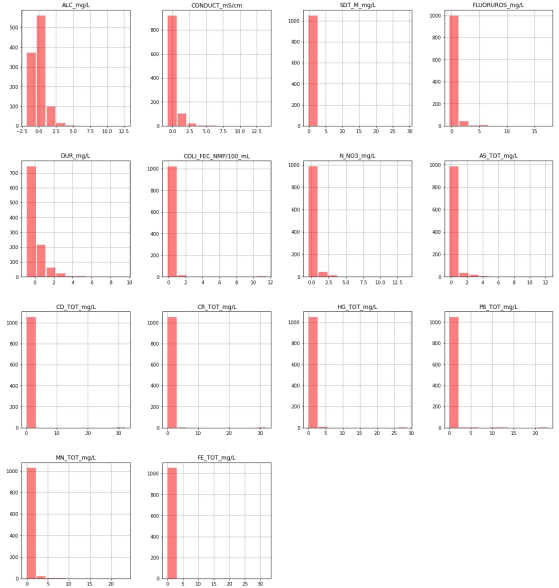
Los geopuntos se cambiaron a valores de coordenadas con la función Point de shapely.geometry para poder graficarlos en el mapa.

A las variables numéricas se le aplicó Standard scaler.

Análisis

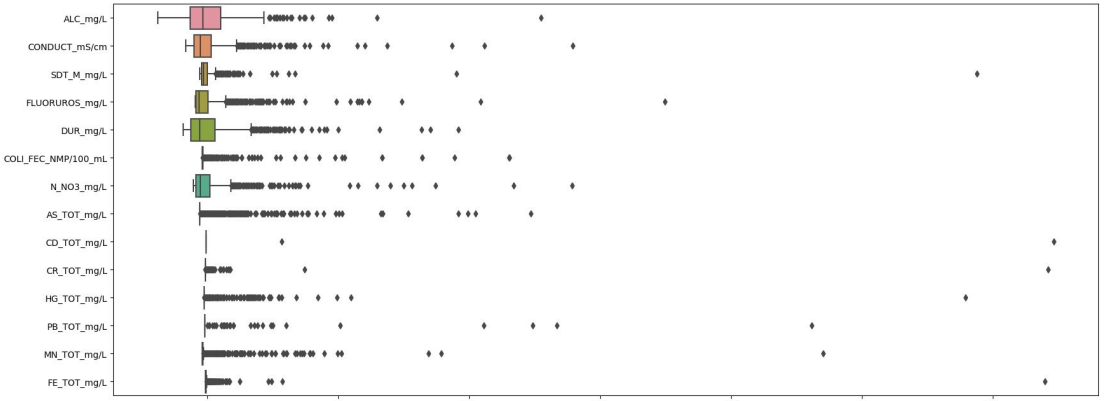
Para ambos valores, categóricos y numéricos se desplegaron histogramas para identificar visualmente las escalas obtenidas.



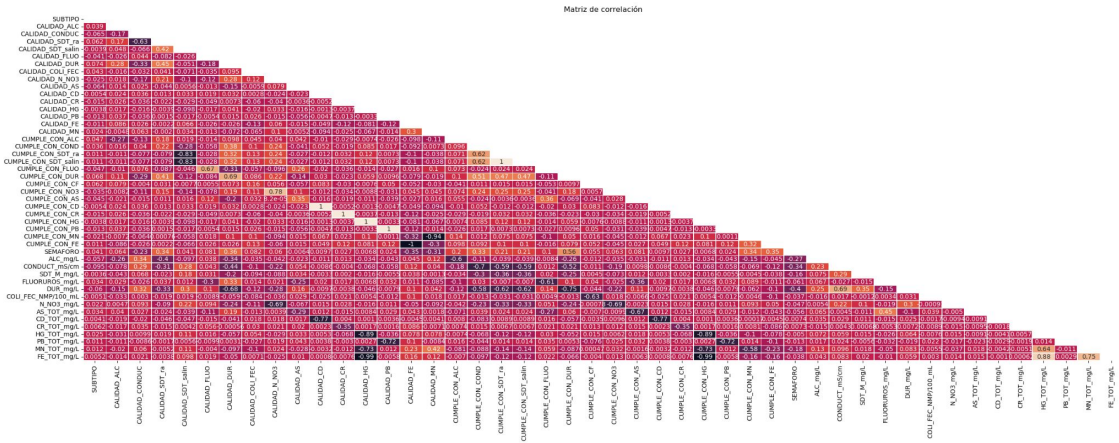


Análisis

Se crearon boxplots de las variables principales.



Obtuvimos la correlación con df.corr() de todas la variables y desplegamos la información usando un mapa de calor con la librería seaborn.



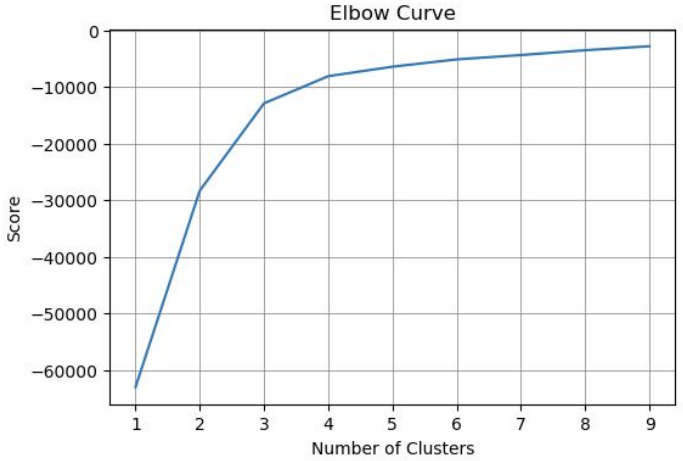
Análisis

Con df.describe() identificamos las medidas de tendencia central , Media , Desviación estándar , Man-Max y cuartiles.

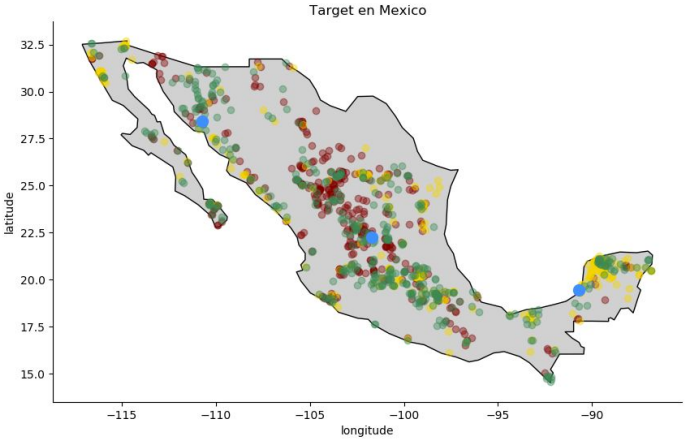
	SUBTIPO	LONGITUD	LATITUD	PERIODO	ALC_mg/L	CALIDAD_ALC	CONDUCT_mS/cm	CALIDAD_CONDUC	SDT_M_mg/L	CALIDAD_SDT_ra	CALIDAD_SDT_salin
count	1054.000000	1054.000000	1054.000000	1054.0	1.054000e+03	1054.000000	1.054000e+03	1054.000000	1.054000e+03	1054.000000	1054.000000
mean	4.945920	-101.848270	23.161796	2020.0	5.730183e-17	0.653700	-1.921297e-16	2.035104	-1.685348e-17	1.850095	0.980076
std	0.437898	6.697568	3.875005	0.0	1.000475e+00	1.176132	1.000475e+00	1.866814	1.000475e+00	1.178437	0.637973
min	0.000000	-116.664250	14.561150	2020.0	-1.872767e+00	0.000000	-8.272414e-01	0.000000	-2.878500e-01	0.000000	0.000000
25%	5.000000	-105.385170	20.224857	2020.0	-6.340312e-01	0.000000	-5.100348e-01	0.000000	-2.021728e-01	1.000000	1.000000
50%	5.000000	-102.170665	22.640705	2020.0	-1.698569e-01	0.000000	-2.585125e-01	2.000000	-1.249964e-01	2.000000	1.000000
75%	5.000000	-98.971268	25.508770	2020.0	5.241881e-01	0.750000	1.484090e-01	4.000000	6.747900e-03	3.000000	1.000000
max	7.000000	-86.864120	32.677713	2020.0	1.273958e+01	3.000000	1.396532e+01	4.000000	2.939940e+01	4.000000	3.000000

K Means, visualización

- Se eligió el número de clusters con ayuda de la gráfica “Elbow Curve”, el resultado óptimo fueron 3 clusters.



- En conclusión, el algoritmo K Means no tiene relevancia en este problema de clasificación, debido a la baja correlación que existe con la ubicación de cada depósito de agua y la calidad de esta.



Clasificación.

En la partición de los datos usamos 70% de ellos para entrenamiento y el 30% para validación.
La función con los parámetros usados quedó de la siguiente forma:

- `x_train, x_test, y_train, y_test = train_test_split(x, y, train_size=0.70, random_state=10)`

Se analizaron dos modelos diferentes: Decision tree y Random forest.

La variable de salida que tomamos fue la columna llamada “Semáforo” y sus tres diferentes clases.

```
Clase 1 = Amarillo
Clase 2 = Rojo
Clase 3 = Green
```

Siendo verde la clase con mayor ponderación.

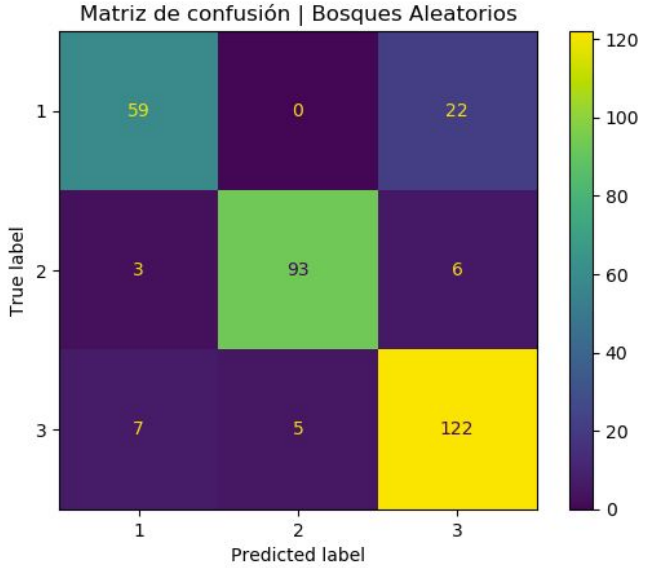
Con ayuda de classification_report() obtuvimos los valores de exactitud.

Clasificación Random forest

Los resultados para random forest fueron los siguientes:

General Accuracy: **81.38** %

	Precisión	Recall	F1-Score	Support
Class 1	0.86	0.73	0.79	81
Class 2	0.95	0.91	0.93	102
Class 3	0.81	0.91	0.86	134
Accuracy			0.86	317
Macro avg	0.87	0.85	0.86	317
Weighted avg	0.87	0.86	0.86	317

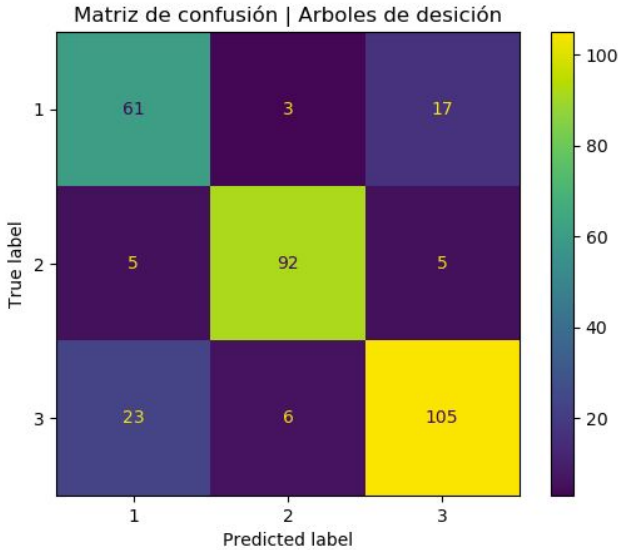


Clasificación Decision tree

Los resultados para Decision tree fueron los siguientes:

General Accuracy: **86.43** %

	Precisión	Recall	F1-Score	Support
Class 1	0.69	0.75	0.72	81
Class 2	0.91	0.90	0.91	102
Class 3	0.83	0.78	0.80	134
Accuracy			0.81	317
Macro avg	0.81	0.81	0.81	317
Weighted avg	0.82	0.81	0.82	317



Resultados y Conclusiones

- Con Decisión tree después de entrenar el modelo , fue el que mejor performance tuvo.
- Con la ayuda de la elbow curve se identificó que el punto de inflexión empezaba en 2 y se asentaba en 4 . Por lo que se decidió usar un valor de 3 para el valor de los clusters.
- Los clusters resultantes quedaron en Sonora , Guanajuato y Campeche.
- Seleccionamos 8 Features porque representaban el mayor porcentaje de pesos para el modelo, son los siguientes: `'FLUORUROS_mg/L', 'CALIDAD_FLUO', 'CUMPLE_CON_FLUO', 'DUR_mg/L', 'COLI_FEC_NMP/100_mL', 'N_NO3_mg/L', 'AS_TOT_mg/L', 'CUMPLE_CON_AS'.`