Edgar López Valdes

A01339939

Alejandro Juarez Corona

A01168444

Ciencia de Datos

Reporte Reto

Equipo 55

18 de noviembre de 2022

Datos

Para este reto se seleccionó la base de datos de aguas profundas, esta contiene información relevante a los diferentes cuerpos de agua en la república mexicana que cumplen con las características para formar parte de este grupo. Para el trabajo se utilizó un archivo en formato csv que contiene 1068 registros y un total de 57 columnas con información relevante para determinar si cierto cuerpo de agua puede clasificarse en calidad alta, media o baja, esto a partir del uso de la variable semáforo. Entre las columnas más importantes están las que describen las concentraciones de químicos en el agua (numéricas) en su mayoría presentan concentraciones de tipo mg/L y las categorías que indican la ubicación, el tipo de acuífero y finalmente el semáforo, que para fines del análisis es la variable de clasificación y a predecir.

Análisis y limpieza

Para el análisis de variables primero se buscaron las columnas que no tenían valor, el boxplot nos ayuda a determinar que por ejemplo el período y el SDT_mg/L no tenían información relevante, por lo que se llegó a la conclusión de que sería mejor eliminarlas del análisis.

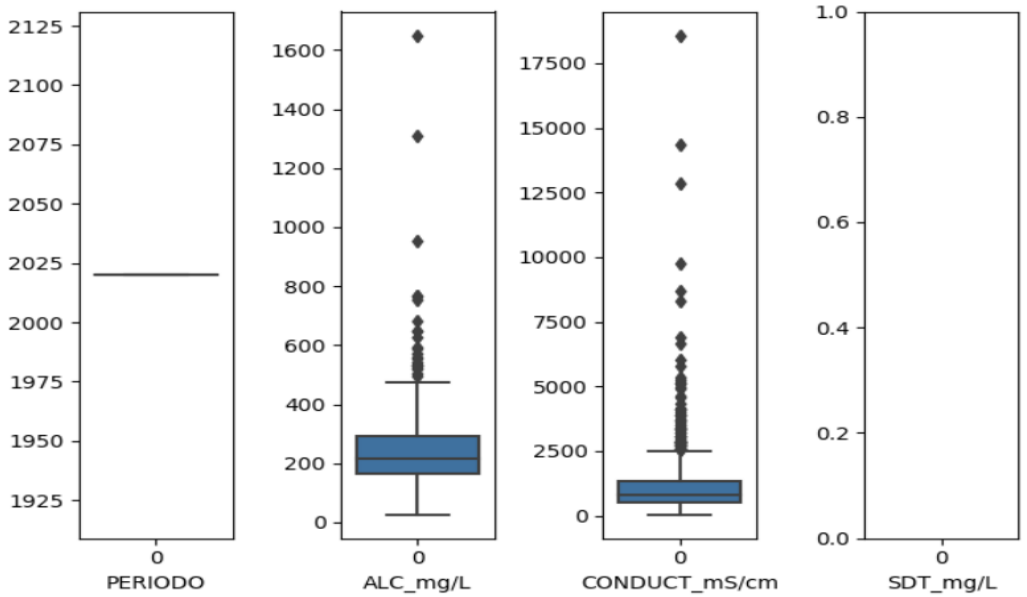

Fig 1. Boxplot variables numéricas, se eliminan las que no tienen valor

Se obtuvó la lista de variables numéricas finales previo al análisis por región -

```
New Numeric Columns: ['LONGITUD' 'LATITUD' 'ALC_mg/L' 'CONDUCT_mS/cm' 'SDT_M_mg/L'
 'FLUORUROS_mg/L' 'DUR_mg/L' 'COLI_FEC_NMP/100_mL' 'N_NO3_mg/L'
 'AS_TOT_mg/L' 'CD_TOT_mg/L' 'CR_TOT_mg/L' 'HG_TOT_mg/L' 'PB_TOT_mg/L'
 'MN_TOT_mg/L' 'FE_TOT_mg/L']
```

Cabe mencionar que se utilizaron las librerías pandas, numpy, sklearn y matplotlib principalmente. Pandas se utilizó sobre todo para el manejo del conjunto de datos, así que con el nuevo data frame se analizó si la mejor opción sería borrar todos los registros con valores nulos, esta metodología probó no ser la más apropiada ya que se perdería una gran cantidad de datos. Se optó finalmente por reemplazar los valores numéricos por la media de la columna, esto funciona ya que no había outliers tan marcados y en general era el mejor de los posibles valores. Como nota general, se obtuvieron todas las medidas de dispersión, entre ellas la moda, mediana, media, std, min y max para cada columna. Se obtuvieron los siguientes resultados para el caso particular de la media.

LONGITUD	-101.891007
LATITUD	23.163618
ALC_mg/L	235.633759
CONDUCT_mS/cm	1138.953013
SDT_M_mg/L	896.101567
FLUORUROS_mg/L	1.075600
DUR_mg/L	347.938073
COLI_FEC_NMP/100_mL	355.490356
N_NO3_mg/L	4.319759
AS_TOT_mg/L	0.019618
CD_TOT_mg/L	0.003030
CR_TOT_mg/L	0.013276
HG_TOT_mg/L	0.000557
PB_TOT_mg/L	0.005282
MN_TOT_mg/L	0.072478
FE_TOT_mg/L	0.410387

Finalmente, con el conjunto de datos limpio y sin nulos, se graficó una matriz de correlación, simplemente para determinar si sería necesario aplicar alguna técnica de reducción de dimensiones, a pesar de que algunos valores mostraban una alta correlación como la presencia de hierro y la presencia de mercurio, se determinó no reducir el número de componentes ya que en su mayoría las columnas no

presentaban correlaciones significativas, además de que se quería entender verdaderamente cuáles de estas columnas eran de mayor importancia en los siguientes pasos.

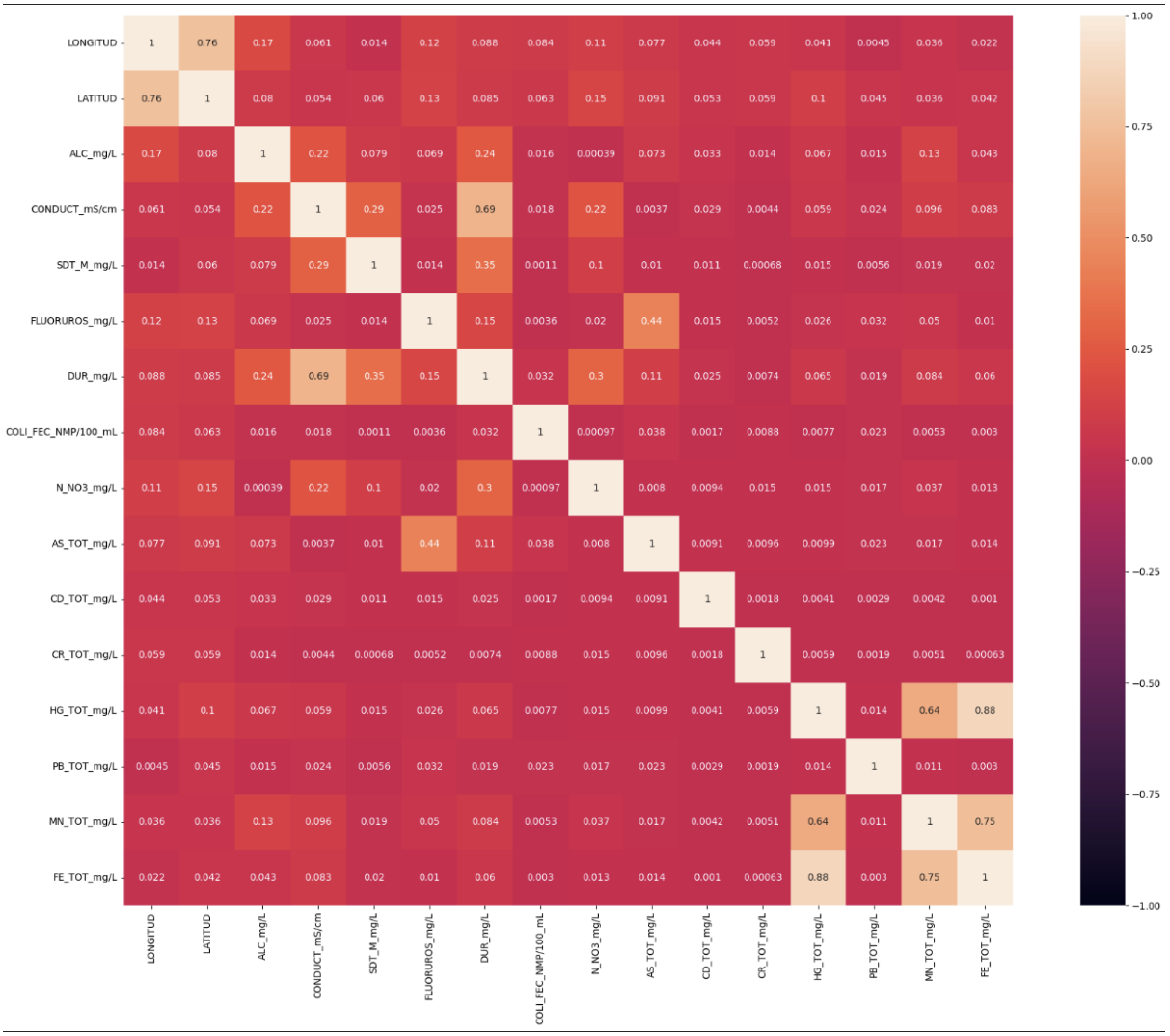

Fig 2. Matriz de correlación

K means

El siguiente paso lógico sería realizar un análisis de vecinos cercanos para determinar si había alguna relación entre la ubicación de los acuíferos y la calidad del agua. Lo primero sería determinar el número óptimo de clusters para la agrupación de los datos. Se realizaron diversas pruebas con valores diversos en cuanto al número de clusters, pero al final se determinó que el número apropiado

sería 4, como podemos observar en la siguiente figura (prueba del codo), es ahí donde se aprecia el punto clave de inflexión.

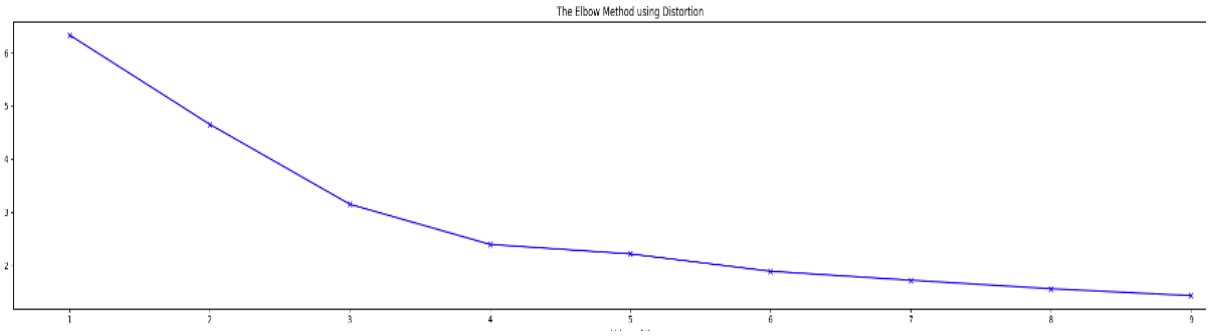

Fig 3. Prueba del codo

Con los datos del conjunto se utilizó la función Kmeans para 4 clusters y fue posible encontrar las coordenadas (latitud y longitud) de los centroides y los puntos de cada región, todo esto se condensó en un mapa, apoyados en la librería matplotlib.

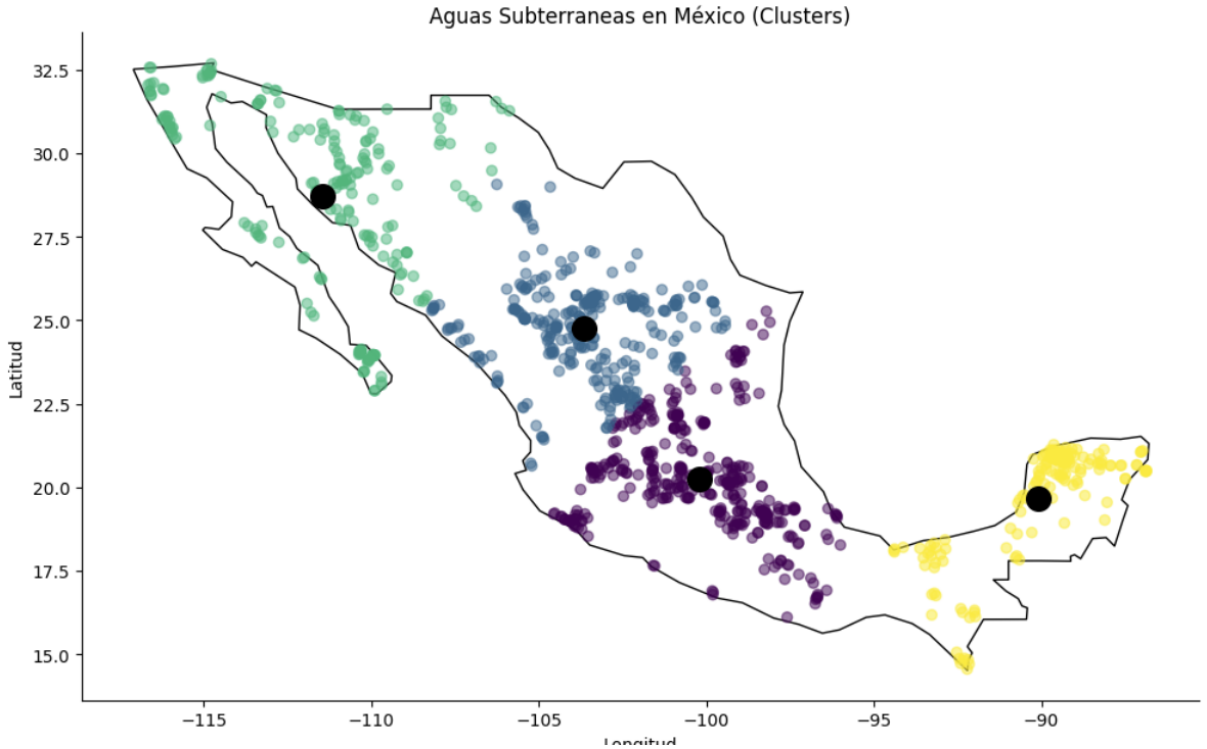

Fig 4. Clusters sobre el mapa de México

Los resultados fueron prometedores, se logra apreciar la distribución de cuerpos de agua equitativa para cada uno de los clusters, con los que se definieron las zonas como - Norte Oeste, Norte Centro, Centro y Sur. Python es una herramienta poderosa, por lo que a partir de las coordenadas de los centroides y la librería geopy, fue posible extrapolar la ubicación exacta de los centroides.

```
array([[-100.2286467 ,   20.26114492],
       [-103.66584078,   24.77478631],
       [-111.44537124,   28.7340166 ],
       [ -90.09271578,   19.6502625 ]])

Municipio de Campeche, Campeche, México
Cuencamé, Durango, México
Granja el Charro, Hermosillo, Sonora, México
Amealco de Bonfil, Querétaro, México
```

Estas ubicaciones corresponden a los puntos negros sobre el mapa de México presentado anteriormente. Tras encontrar estos clusters, se dividió el conjunto en cuatro dataframes, cada uno representando una región con la intención de comparar la calidad de agua en cada lugar. Al mostrar la moda de semáforos para cada cluster se obtuvieron los siguientes resultados.

```
0    Verde
Name: SEMAFORO, dtype: object
0    Rojo
Name: SEMAFORO, dtype: object
0    Verde
Name: SEMAFORO, dtype: object
0    Amarillo
Name: SEMAFORO, dtype: object
```

Vemos que la región de baja california en el primer cluster con centro en sonora presentan la mejor calidad de agua,cluster sur tiene calidad intermedia, mientras la zona centro presenta calidad verde en la mayor parte de los sitios, finalmente la zona norte presenta la peor calidad del agua con moda de Semáforo en rojo, hagamos una gráfica de barras para visualizar esto. En la gráfica se realizó la cuenta de apariciones de cada valor del semáforo para cada una de las regiones definidas previamente a partir de los clusters.

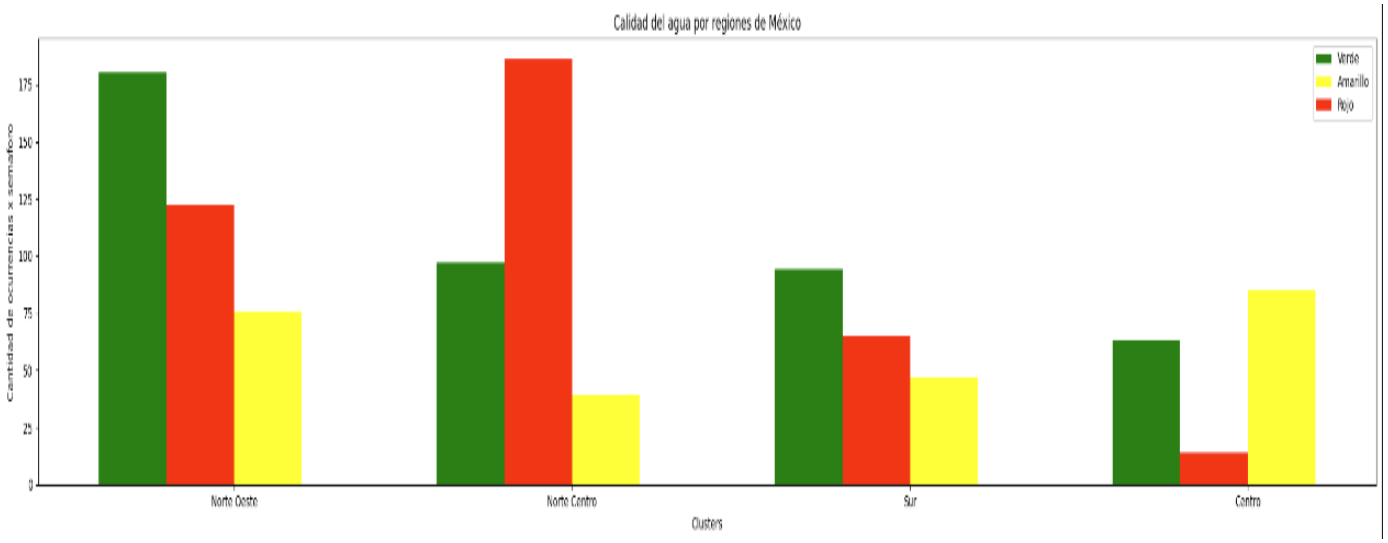

Fig 5. Calidad del agua por región

Determinamos que sí existe una relación entre la ubicación y la calidad del agua, podemos observar que en los clusters más al norte existen comportamientos similares, aunque en la zona centro norte la calidad del agua es muy mala en algunos puntos, podemos pensar que esto está relacionado a la grandes ciudades industriales, mismo caso con la zona centro, si bien las regiones rurales presentan calidad intermedia o buena del agua, podemos asumir que las zonas con más densidad poblacional representan el factor rojo. Mientras que la zona sur (cabe mencionar que es la que presenta mayor vegetación del país) tiene los valores más altos de calidad buena e intermedia, mientras que presenta pocos puntos con mala calidad.

Random Forest y Árbol de decisión

Una vez obtenida la información, el siguiente paso era entrenar un modelo que fuera capaz de predecir la variable de salida (semáforo) a partir de las entradas. Lo primero que se hizo fue cambiar los valores de salida “Verde, amarillo y rojo” por 0,1 y 2, por lo que se aplicó un label encoder.

Tras lo anterior, resulta interesante encontrar las columnas de mayor importancia, por lo que se usaría el modelo de árbol de decisión para graficar las variables por importancia. Como nota adicional, antes de aplicar cualquier modelo de clasificación, se partió el conjunto de datos en conjuntos de entrenamiento y validación con proporciones correspondientes 70%-30%. Así mismo, antes de

aplicar el modelo se realizó un GridSearch con variación en los hiper parámetros tanto para Random Forest como para Árbol de decisión, para poder asegurar el mejor desempeño. Se obtuvo que la profundidad apropiada era 100 niveles y 20 nodos.

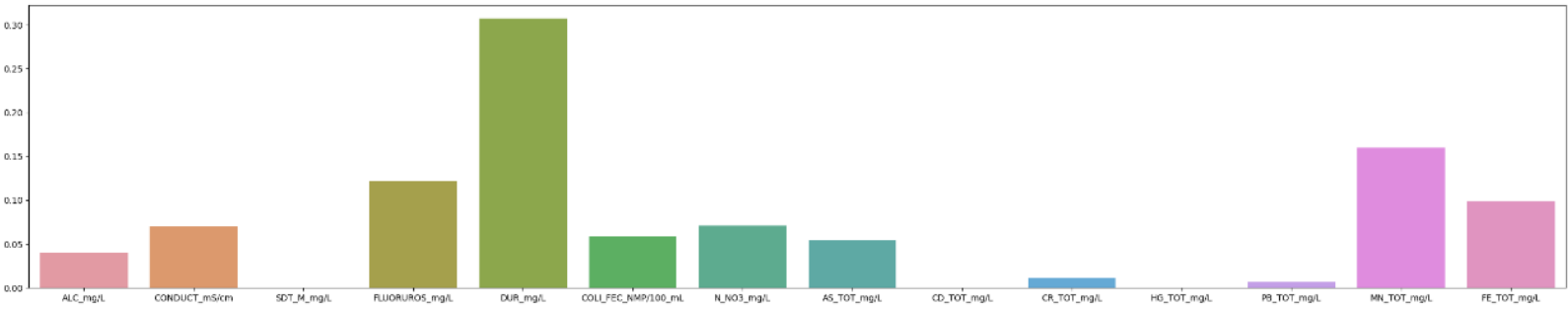

Fig 6. Feature importance

A partir de la gráfica de barras de feature importance se concluye que las sustancias de mayor relevancia son DUR_mg/L, Fluoruro_mg/L y Mn_Tot_mg/L. Entendemos que su presencia es un factor determinante en la variable de salida.

Se entrenaron y probaron ambos modelos con sus respectivos parámetros, contrastando las predicciones con los valores separados durante el split inicial. Se obtuvieron los siguientes resultados para las métricas de desempeño.

Resultados Random Forest

	precision	recall	f1-score	support
0	0.92	0.99	0.95	74
1	0.99	0.96	0.98	108
2	0.99	0.97	0.98	139
accuracy			0.97	321
macro avg	0.97	0.97	0.97	321
weighted avg	0.97	0.97	0.97	321

Resultados Arbol de Decision

	precision	recall	f1-score	support
0	0.84	0.97	0.90	74
1	0.98	0.90	0.94	108
2	1.00	0.98	0.99	139
accuracy			0.95	321
macro avg	0.94	0.95	0.94	321
weighted avg	0.96	0.95	0.95	321

Además de las matrices de confusión.

DT

```
[[ 72   2   0]
 [ 11  97   0]
 [  3   0 136]]
```

RF

```
[[ 73   0   1]
 [  3 104   1]
 [  3   1 135]]
```

Los resultados son muy prometedores, hubo una gran cantidad de aciertos, minimizando tanto los falsos positivos como los negativos. En realidad podríamos llegar a pensar que hay un sesgo. La manera en que nosotros podríamos determinar si hay un sobreentrenamiento es probando el modelo con otro tipo de características, tal vez mostrarle valores de cuerpos de agua en otros países para determinar si las clasificaciones en realidad son adecuadas, o si los modelos están bajo el sesgo del territorio mexicano para su entrenamiento. Aun así podemos concluir que el desempeño es ligeramente mejor con Random Forest al comparar las métricas de precisión, recall y f1.

La actividad fue muy interesante ya que nos permitió tener un entendimiento integral de los modelos de clasificación y la importancia de K Means. Sería interesante analizar otro tipo de entradas para trabajo en el futuro, además de que nos permitiría confirmar los valores importantes para la clasificación. Entender qué factores son los que afectan la calidad del agua es el primer paso para crear estrategias de mejora.