RETO

CALIDAD DEL AGUA EN 5000 SITIOS

Maestría en Inteligencia Artificial Aplicada
Curso: Ciencia y Analítica de Datos
Tecnológico de Monterrey

Equipo 170
Freddy Armendariz Herrera – A01793672
Samuel Elías Flores González – A01793668

PLANTEAMIENTO DEL PROBLEMA

DESCRIPCIÓN

Se trata de un conjunto de datos con registros relacionados a la calidad del agua en distintas ubicaciones geográficas a lo largo de toda la región de México.

OBJETIVO

Se tienen dos propósitos respecto a este conjunto de datos:

- Encontrar si existe una relación entre la ubicación geográfica y la calidad del agua.
- Implementar un buen modelo clasificador con el fin de catalogar los lugares según su calidad del agua en las tres diferentes clases a partir de los contaminantes encontrados en la misma.

LIMPIEZA DE LOS DATOS

- Verificamos datos nulos

```
df.isna().sum()
```

- Eliminamos columnas con gran cantidad de valores nulos

```
df.drop(["CONTAMINANTES","SDT_mg/L"], inplace=True, axis=1)
```

- Eliminamos registros con datos nulos (cantidad despreciable con respecto al total de datos de la base de datos)

```
df.dropna(inplace = True)
```

- Nos centramos en las columnas numéricas y su salida (semáforo)

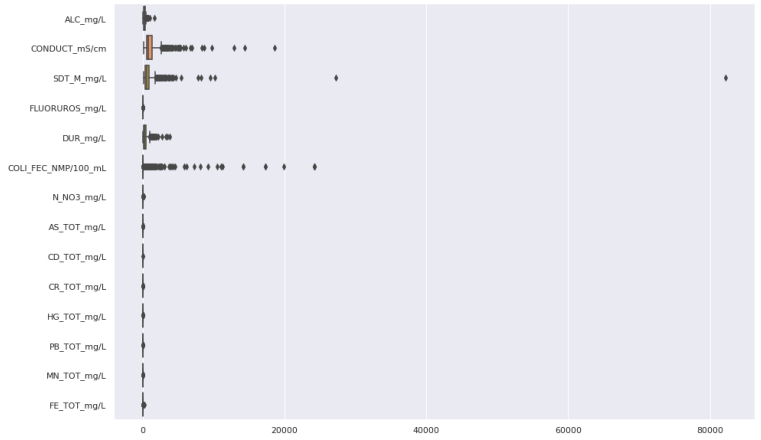
```
col_num = ['LONGITUD','LATITUD','ALC_mg/L','CONDUCT_mS/cm','SDT_M_mg/L','FLUORUROS_mg/L','DUR_mg/L','COLI_FEC_NMP/100_mL',
           'N_NO3_mg/L','AS_TOT_mg/L','CD_TOT_mg/L','CR_TOT_mg/L','HG_TOT_mg/L','PB_TOT_mg/L','MN_TOT_mg/L','FE_TOT_mg/L','SEMAFORO']
```

- Damos formato al tipo de dato y eliminamos símbolos

```
for name in col:
  df_new[name] = df_new[name].astype(str)
  df_new[name] = df_new[name].replace("<","", regex=True)
  df_new[name]= df_new[name].astype(float)
```

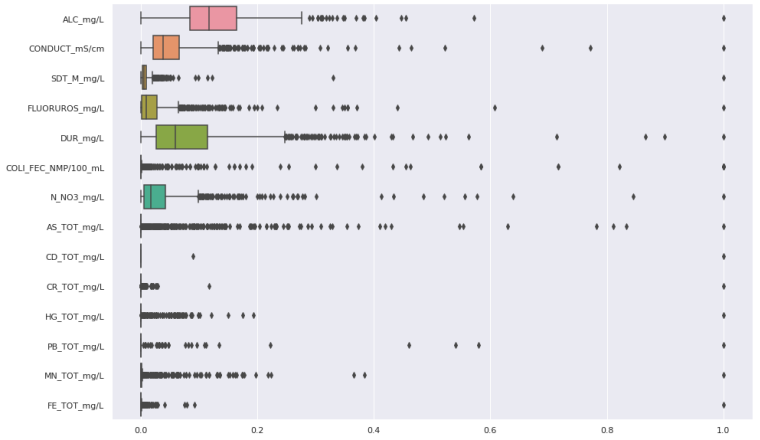
ANÁLISIS DE LOS DATOS

- Analizamos la varianza de los datos

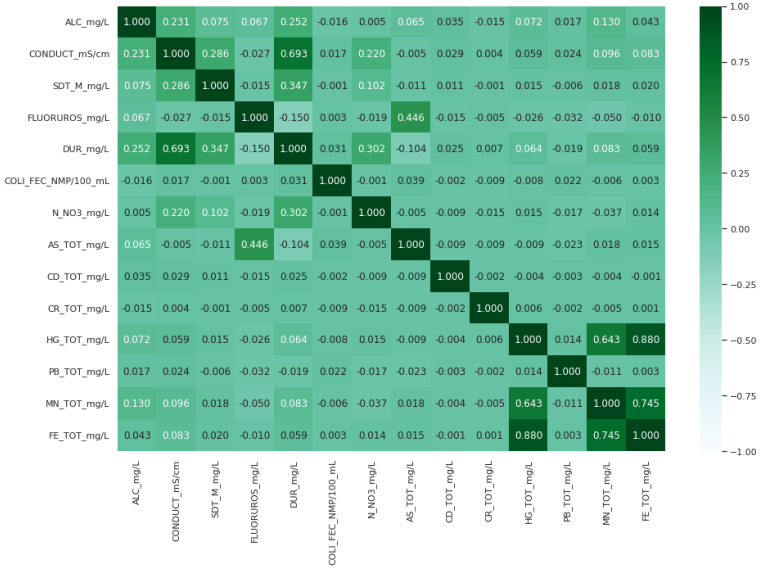


Mediante la gráfica de caja observamos el comportamiento de la varianza de los datos.

Nos damos cuenta en la gráfica que los valores de algunas de las columnas manejan diferente rango, por lo cual se tomó la decisión de escalar los valores.



Con los datos escalados podemos visualizar de una mejor manera los datos y también podemos hacer uso de ellos sin causar sesgos debido a su diferencia de magnitud.



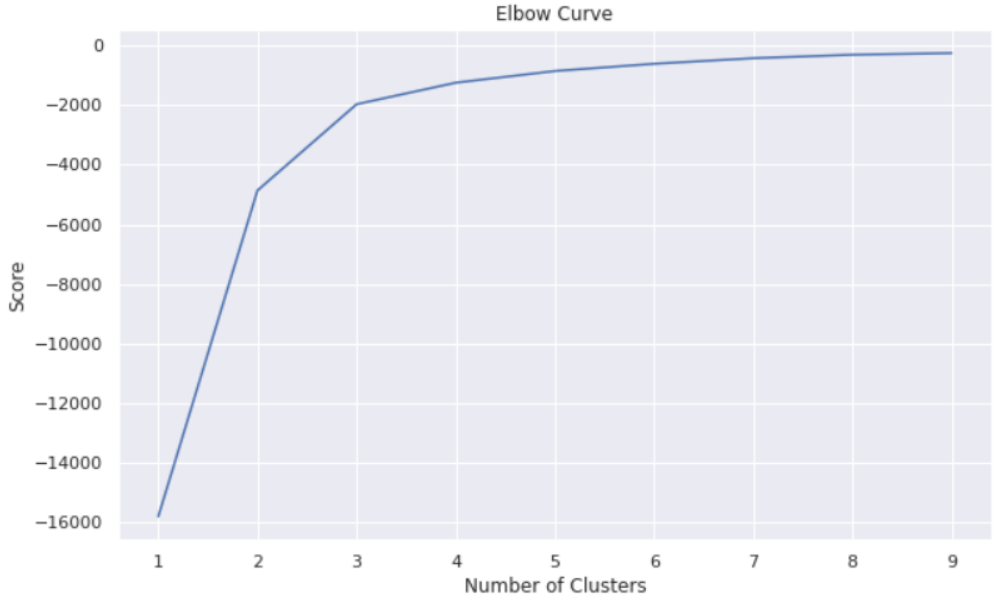
Observamos la correlación existente entre cada una de las variables del dataframe.

Notamos una correlación importante entre los contaminantes hierro y mercurio.

Además de una correlación moderada entre la dureza y la conductividad.

K-MEANS

- Generamos gráfica de codo



Determinamos que el número de clusters ideal es 3, ya que apartir del valor 4 los valores se vuelven constantes o muy similares.

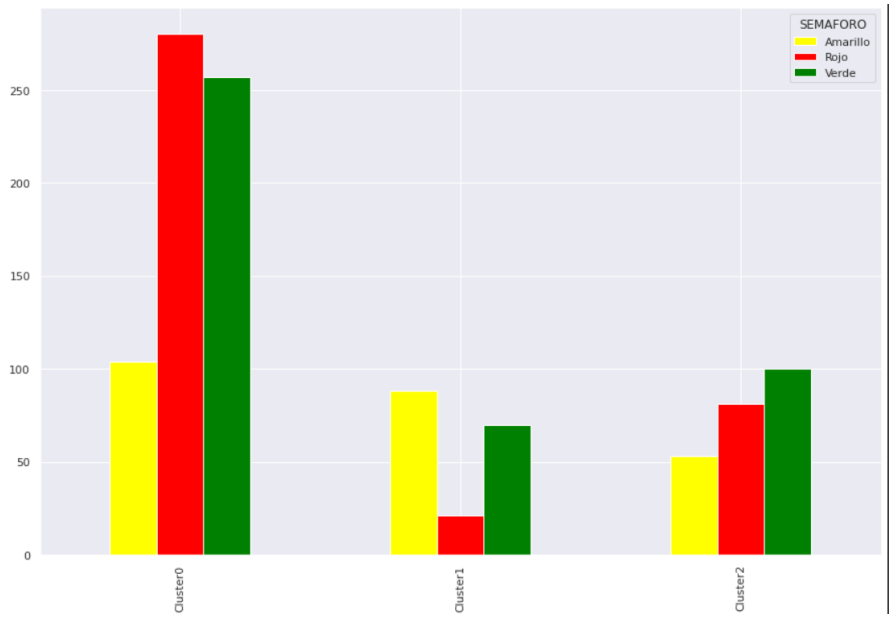
- Mediante Kmeans determinamos las coordenadas de los centroides de los 3 clusters

	0	1	Coordinates
0	-90.698434	19.475165	POINT (-90.69843 19.47516)
1	-110.740896	28.420375	POINT (-110.74090 28.42038)
2	-101.715581	22.271624	POINT (-101.71558 22.27162)

Obtenemos dataframe con los valores de latitude y longitude de cada cluster calculado con Kmeans.

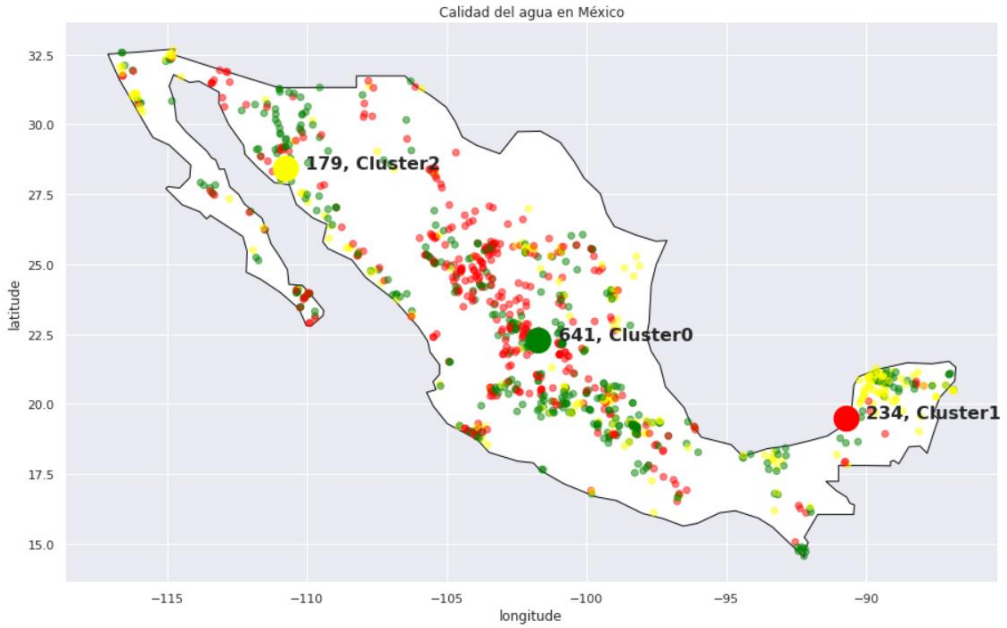
K-MEANS

- Calculamos la moda de cada cluster.



Calculamos la moda de cada uno de los clusters para determinar su color de semáforo

- Graficamos clusters y cuerpos de agua.

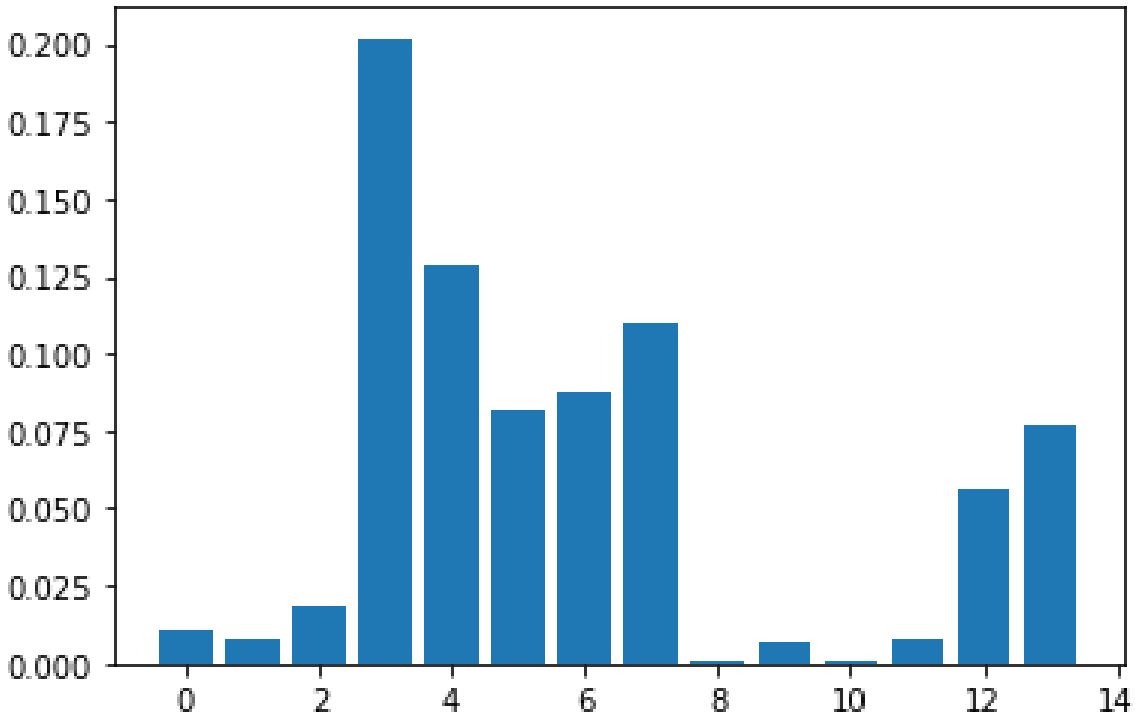


Podemos observar que la mejor calidad de agua basada en la moda del cluster es el cluster 0, el cual se encuentra en la zona central de México. Mientras que la peor calidad de agua la encontramos en el cluster 1, ubicada en la zona sur este de México

CLASIFICACION

FEATURE IMPORTANCE

ITEM	FEATURE	SCORE
0	ALC_mg/L	0.0102
1	CONDUCT_mS/cm	0.0073
2	SDT_M_mg/L	0.0183
3	FLUORUROS_mg/L	0.2017
4	DUR_mg/L	0.1287
5	COLI_FEC_NMP/100_mL	0.0816
6	N_NO3_mg/L	0.0879
7	AS_TOT_mg/L	0.1103
8	CD_TOT_mg/L	0.0011
9	CR_TOT_mg/L	0.0065
10	HG_TOT_mg/L	0.0011
11	PB_TOT_mg/L	0.0073
12	MN_TOT_mg/L	0.0561
13	FE_TOT_mg/L	0.0767



La gran parte del score se los llevan tres variables:

- 3 – FLUORUROS_mg/L
- 4 – DUR_mg/L
- 7 – AS_TOT_mg/L

CLASIFICACION

- Partición del conjunto de datos utilizada:
  - Entrenamiento y validación – 80%
  - Prueba – 20%

- Generamos modelos y entrenamos con los conjuntos de datos divididos anteriormente.

RANDOM FOREST

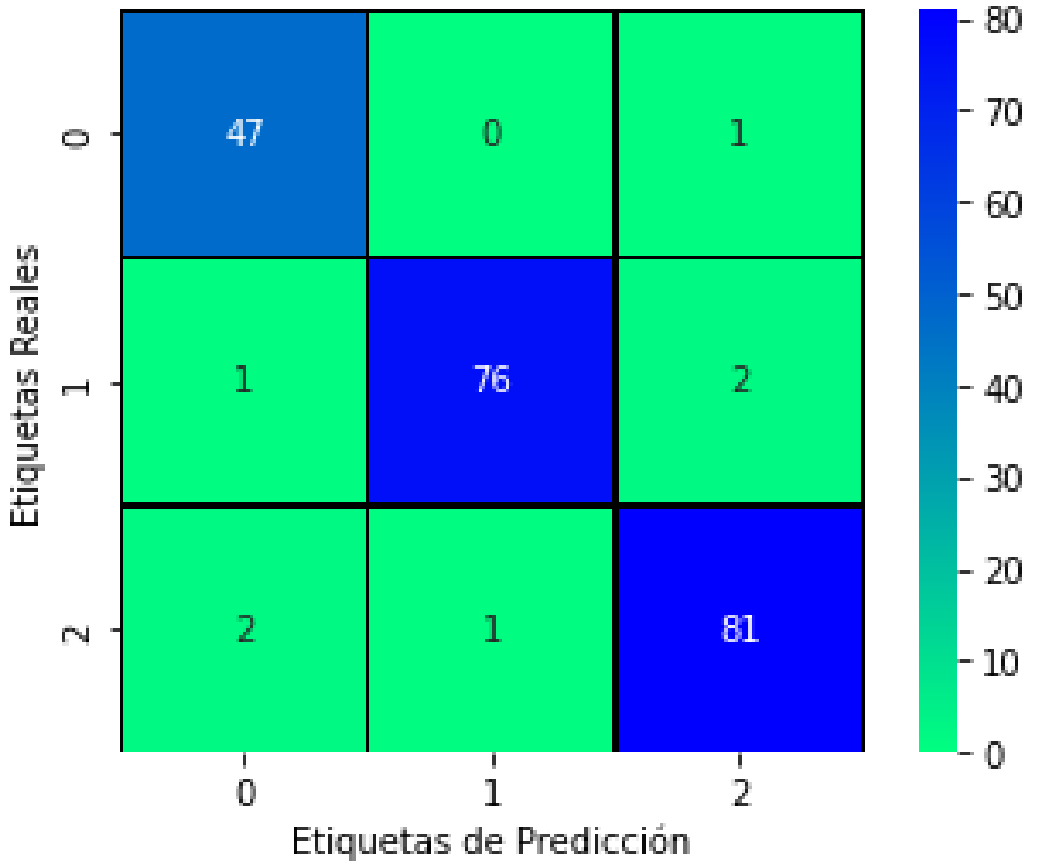
```
[ ]  mimodelo = RandomForestClassifier()

     clf = mimodelo.fit(Xtv,Ytv)

     clf.score(Xtest,Ytest)

     0.957345971563981
```

DECISION TREE

```
     mimodelo2 = DecisionTreeClassifier()

     clf2 = mimodelo.fit(Xtv,Ytv)

     clf2.score(Xtest,Ytest)

     0.966824644549763
```

RESULTADOS

Se graficó la matriz de confusion a partir de las predicciones hechas por el modelo de *Random Forest* con los datos del conjunto de prueba y se obtiene un buen desempeño ya que las falsas predicciones podrian considerarse mínimas en comparación a las verdaderas. Recordemos que el modelo en ningún momento toco los datos de prueba, es decir fueron datos completamente nuevos para el mismo.



CONCLUSIONES

- La limpieza de datos es parte importante del análisis y procesamiento de los datos, ya que un conjunto de datos limpios nos permitirá trabajar los modelos sin problemas de formato o de datos nulos.

- El análisis de los datos nos permite tener una perspectiva mas profunda sobre nuestro conjunto de datos.

- Mediante Kmeans se pudo determinar la cantidad de clusters optima, en este caso consideramos que 3 era el número adecuado.

- Los resultados de los clusters basados en la moda, nos indican que la peor calidad de agua se encuentra en el sur de México, mientras que la mejor se ubica en el centro del país.

- El preprocesamiento de los datos es de gran importancia, ya que es necesaria para que los datos puedan ser procesados por los modelos.

- Para obtener un mejor desempeño en cualquier modelo se deben encontrar y ajustar sus hiper parámetros.