Tecnológico de Monterrey

Actividad Semanal 8 - Entrega 2

Tema: Limpieza, análisis, visualización y kmeans

Materia:

TC4029.10 - Ciencia y Análitica de Datos

Profesor Titular: Dra. María de la Paz Rico Fernández

Alumnos:

- A01793625 - Luis Angel Hermenegildo Dominguez
- A01332665 - Hector Montañez Alvarez

18 de noviembre de 2022

Limpieza de datos

Comenzamos eligiendo la base de datos de: ¨Calidad de aguas subterráneas¨ ya que es el dataset más completo en comparación con ¨Calidad de aguas superficiales¨ Pasos a seguir :

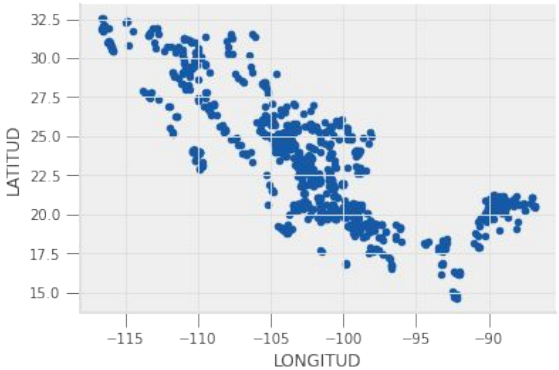
1) Checar tamaño del dataset e index.
2) Verificamos que no existan valores faltantes o inválidos (Sustituimos con valores promedio. [Imputaciones]
3) Verificamos que el tipo de dato corresponda correctamente.
4) Graficamos los datos para ver la distribución de los mismos.
5) Boxpolot para identificar los outliers, mismos que se pueden tratar con minmax
6) Verificamos el balance de la variable de salida.

K-Means

Para esta parte del ejercicio seleccionamos la variable de ´Calidad_MN´ para rankear cada registro de agua subterránea. Las dividimos en 3 familias.

Para cada valor
1) Potable Excelente
2) Sin efectos en la salud
3) Puede afectar a la salud

En la siguiente gráfica se puede observar la posición geográfica del primer grupo.

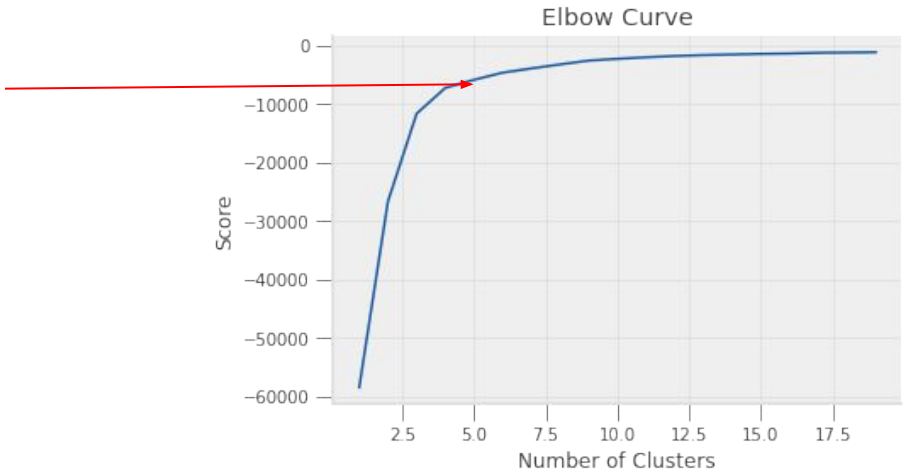


Ahora la tarea recae en encontrar los puntos más céntricos al total de estos registros

K-Means

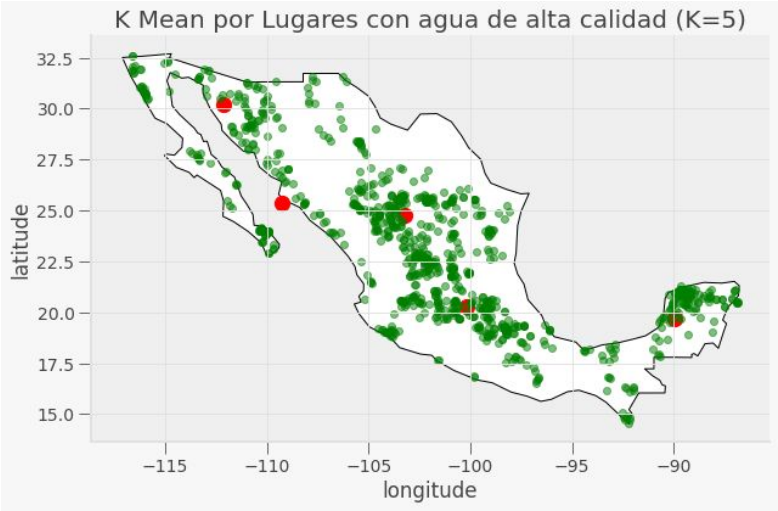
Para este paso utilizamos Elbow Curvo que nos gráfica el número de clusters VS el Score obtenido.

La decisión es responsabilidad del científico de datos, nosotros elegimos [5]



K-Means

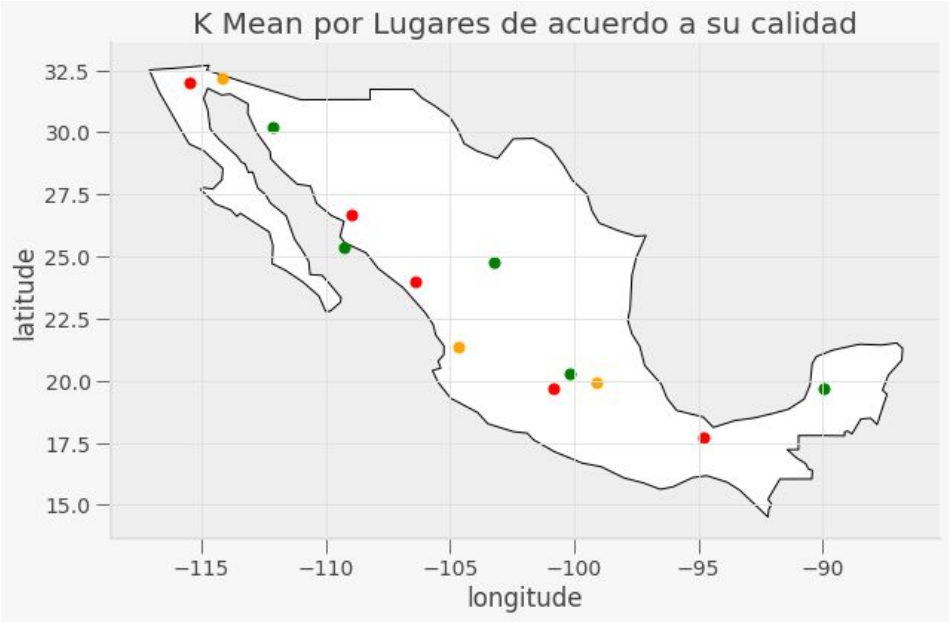
Ya teniendo el número de clusters, se obtienen las coordenadas de los mismos y se grafican.



Los puntos rojos son los 5 centroides que se encuentran más cercanos al total de Registros subterráneos con calidad: Potable - Excelente

K-Means

Lo mismo se repitió para las otras dos familias y por último graficamos únicamente los centroides.

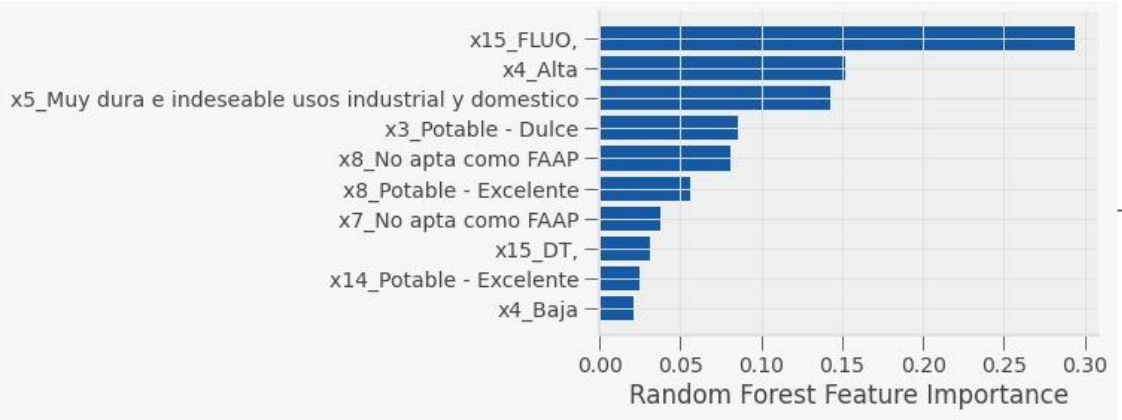


Verde: Alta calidad
Amarillo: Media calidad
Negro: Baja calidad

Con esta información se pueden tomar decisiones de en dónde colocar embotelladoras de agua en donde matemáticamente tenga los pozos de agua equidistantes.

Random Forest

Como parte del ejercicio, renombramos los valores de la variable de salida [Y] para mantenerlos de forma numérica. Posteriormente realizamos la partición: 80% de entrenamiento y 20% de validación . Mediante el modelo de Random forest obtuvimos los valores que más influencia tienen en la variable de salida.

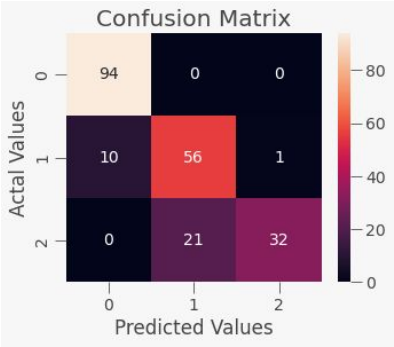


1) CONTAMINANTES
2) CALIDAD_FLUO
3) CALIDAD_DUR
4) CALIDAD_SDT_ra
5) CALIDAD_AS

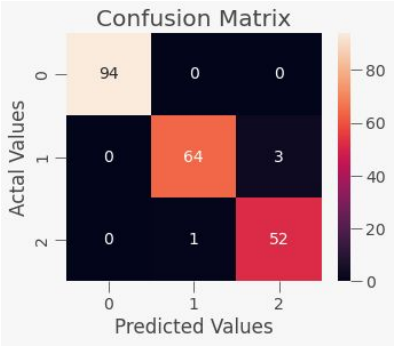
Entrenamiento

Obtuvimos las matrices de confusión de:

1) Modelo (Random Forest y Decision Tree) entrenado con todas las categorías



2) Modelo (Random Forest y Decision Tree) entrenado con las categorías obtenidas de feature importance



Pocos valores falsos y Scores altos como podemos observar.

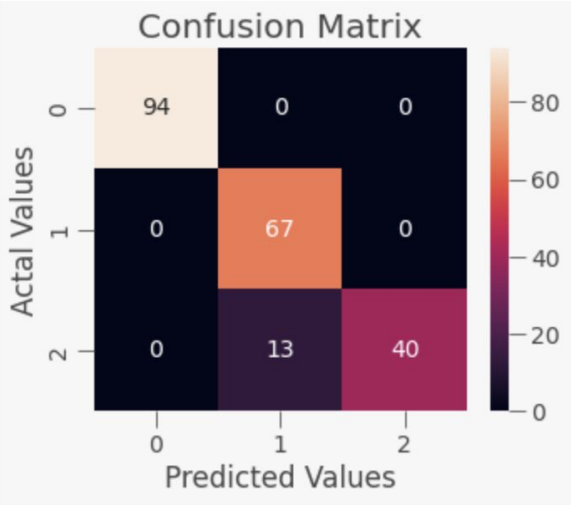
Modelo RF

	precision	recall	f1-score	support
VERDE	1.00	1.00	1.00	94
ROJO	0.98	0.96	0.97	67
AMARILLO	0.95	0.98	0.96	53
accuracy			0.98	214
macro avg	0.98	0.98	0.98	214
weighted avg	0.98	0.98	0.98	214

Resultados

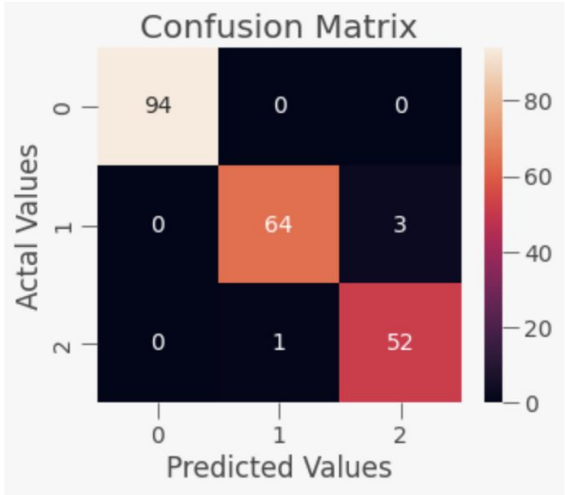
Modelo DT

	precision	recall	f1-score	support
VERDE	1.00	1.00	1.00	94
ROJO	0.84	1.00	0.91	67
AMARILLO	1.00	0.75	0.86	53
accuracy			0.94	214
macro avg	0.95	0.92	0.92	214
weighted avg	0.95	0.94	0.94	214



Modelo RF

	precision	recall	f1-score	support
VERDE	1.00	1.00	1.00	94
ROJO	0.98	0.96	0.97	67
AMARILLO	0.95	0.98	0.96	53
accuracy			0.98	214
macro avg	0.98	0.98	0.98	214
weighted avg	0.98	0.98	0.98	214



Conclusiones

Los resultados de ambos modelos [Random Forest y Decision Tree] son bastante buenos pero si analizamos por F1-Score Random forest gana.

Con este ejercicio entrenamos un modelo que es capaz de predecir con precisión [95% - 98%] el resultado de semáforo de acuerdo a 5 parámetros.

Partimos de un dataset con 56 categorías que posteriormente mediante la metodología aprendida en el curso limpiamos para así llegar con precisión a utilizar únicamente 5 y mantener predicciones altamente confiables.