**REPORTE SOBRE CALIDAD DE  
AGUAS SUBTERRÁNEAS EN SITIOS DE MONITOREO**

**RETO 1:**

1) Respecto a la limpieza de base de datos, se identificó que había datos nulos, y se eliminaron las dos columnas con el mayor número de datos nulos (SDT\_mg/L y CONTAMINANTES), una de las columnas tenía 100% de datos nulos (SDT\_mg/L).

Las variables cuantitativas se completaron con la media, y las variables cualitativas o categóricas se completaron con la moda.

2) Se exploraron  los datos (empleando diversos comandos como describe(), mean(), plot, boxplot de pandas, entre otros). Se identificaron las tendencias centrales como promedio, media y mediana de los datos.También medidas de dispersión, así como el máximo y mínimo, y se identificaron las medidas de posición no centrales, los cuartiles y outliers. Par ello fueron útiles los diagramas de caja y bigote.

También se identificaron correlaciones, para lo cual se ejecutaron matrices de correlación con las diversas variables cuantitativas, y se prepararon los datos para ejecutar modelos de K-means.

3) Respecto al análisis para encontrar si existe una relación entre la calidad del agua y su ubicación geográfica a través de K- means, al mostrar los resultados de agrupamiento de latitudes y longitudes con K means en el mapa de México, se identifican los siguientes aspectos:

- Hay pozos de calidad baja, media y alta, a lo largo del territorio nacional.

- Se identifica una mayor concentración de pozos con calidad alta en el centro de la República Mexicana.

- Se identifica una mayor concentración de pozos con calidad alta en el centro de la República Mexicana

- Se identifica una mayor concentración de pozos con calidad baja en el Sureste de la República Mexicana, particularmente en la Península de Yucatán, en contraste, se identifica una menor concentración de pozos con calidad alta en esta misma región.

En general, no se identifica que exista una relación significativa entre la calidad del agua y su ubicación geográfica, una vez emplados los modelos de K- means. Solamente se observan mayores concentraciones en ciertas regiones, pero en todas las regiones hay pozos de calidad baja, media y alta.

4) Con relación a mostrar resultados de agrupamiento de latitudes y longitudes con K means en el mapa de México, cabe señalar que el método del codo (Elbow) permite identificar el número de clústers con sus respectivos centroides, y utiliza los valores de la inercia obtenidos tras aplicar el K-means a diferente número de Clusters (desde 1 a N Clusters), siendo la inercia la suma de las distancias al cuadrado de cada objeto del Cluster a su centroide.

En este caso cuando se aplica a todos los datos, se identifica un número óptimo de 9 clústers con sus respectivos centroides.

Cuando se aplica el método Elbow con los datos de:

a) En los pozos de calidad alta, se identifican 4 clústers con sus respectivos centroides. Están ubicados en los siguientes lugares:

- Grupo 0 Opodepe, Sonora, México

- Grupo 1 El Capulin, Dolores Hidalgo Cuna de la Independencia Nacional, Guanajuato, 37849, México

- Grupo 2 Palizada, Campeche, México

- Grupo 3 Ramal a Coneto de Comonfort, Arroyo del Caballo, San Francisco Javier de Lajas, Coneto de Comonfort, Durango, México

Sus respectivas coordenadas geográficas son:

- 0: 29.905152 -110.920269

- 1: 21.126431 -101.066704

- 2: 18.488777 -92.037791

- 3: 24.940960 -104.759225

b) En los pozos de calidad baja, se identifican 4 clústers con sus respectivos centroides.Están ubicados en los siguientes lugares:

- Grupo 0 Jiquipilco, Estado de México, México

- Grupo 1 None

- Grupo 2 Rancho Peeta Elias, Hopelchén, Campeche, México

- Grupo 3 Parras, Coahuila, México

Sus respectivas coordenadas geográficas son:

- 0 19.492886 -99.649272

- 1 28.489008 -112.105470

- 2 19.826435 -89.929530

- 3 25.123905 -102.226471

c) Los pozos de calidad media, se identifican 3 clústers con sus respectivos centroides. Están ubicados en los siguientes lugares:

- Grupo 0 Cerrito De Los Hernandez, San Felipe, Guanajuato, 37606, México

- Grupo 1 El Coyote, Cajeme, Sonora, México

- Grupo 2 Hopelchén, Campeche, México

Sus respectivas coordenadas geográficas son:

- 0:  21.323652 -101.151358

- 1:  27.783949 -110.096769

- 2:  19.763112 -90.016450

Se considera que los pozos de buena calidad podrían estar relacionados con la presencia de cordilleras que permiten un flujo continuo de agua hacia los mantos acuíferos subterráneos. Es probable que por ello exista mayor presencia de pozos de agua con calidad buena en locaciones cercanas a la Sierra Madre Oriental y Occidental.

Asimismo, la presencia de pozos de mala calidad de agua podrían estar relacionados con la presencia de ciudades con alta densidad demográfica, lo cual se puede observar por las ubicaciones de las ciudades de las principales entidades federativas del centro del país, tales como: Ciudad de México, Querétaro, Hidalgo, Puebla, etc.

**RETO 2:**

Con relación al Reporte de Clasificación, tenemos lo siguiente:

- La exactitud alcanzó 0.96, lo cual se considera un buen desempeño. Es decir, cuántos predijo bien del total.

- La clase 2 es la que clasifica mejor, conforme a las métricas obtenidas.

- Con relación a la métrica precision, la clase que mejor predijo el modelo es la 1.

- Con relación a la métrica recall, la clase que mejor predijo el modelo es la 2.

- Con relación a la métrica f1-score, la clase que mejor predijo el modelo es la 2.

Con relación a la Curva de Precision Recall, se tiene lo siguiente:

Para la clase 0, parece que la clasifica muy bien.

Para la clase 1, parece que la clasifica muy bien precision hasta tener alrededor de .95 de recall, ya que pasando ese valor desciende el recall.

Para la clase 2 se tiene un comportamiento con mayor variabilidad entre precision y recall. A partir de .85 de recall desciende la precision.

Con relación a la matriz de confusión, tenemos lo siguiente:

Hubo cinco que etiquetó con etiqueta cero, pero que realmente eran etiqueta 1 (2.8%).

Hubo cuatro que etiquetó como etiqueta 2, pero realmente eran etiqueta cero (1.9%).

Por lo anterior, se considera que el algoritmo si es bueno para predecir la calidad del agua.

De la clase 0, tenemos 45 de 49 datos que predijo bien para la clase cero, y que representan el 21% del total de datos.

De la clase 1, tenemos 72 de 78 datos que predijo bien para la clase 1, y que representan el 33.6% del total de datos.

De la clase 2, tenemos 87 de 87 datos que predijo bien para la clase 2, y que representan el 40.7% del total de datos.

Referencias:

Can-Chulim et al. (2011). ORIGEN Y CALIDAD DEL AGUA SUBTERRÁNEA EN LA CUENCA ORIENTAL DE MÉXICO.  Recuperado de: https://www.scielo.org.mx/scielo.php?script=sci\_arttext&pid=S0187-57792011000200189

INEGI (2022). Agua Subterránea. Recuperado de: https://cuentame.inegi.org.mx/territorio/agua/cuerpos.aspx?tema=T

Macías y del Arenal (2007). La distribución del agua subterránea en México. Recuperado de: https://www.revistaciencia.amc.edu.mx/images/revista/58\_3/PDF/08-547.pdf