Reporte: Exploración en aguas superficiales

EQUIPO 104

Integrantes:
Eddie Guadalupe Elorza Ruiz | **A01793547**
Yves Turley Macias Vargas. | **A00813752**

Materia: Ciencia y analítica de datos (Gpo 10)

Profesor Titular: PhD. María de la Paz Rico Fdz

Profesor Tutor: Victoria Guerrero Orozco

18 de noviembre 2022

Metodología

01

Limpieza

Tecnicas para limpiar la base de datos

02

Análisis

Estrategía para analizar la información

03

Clasificación

Metodo y argumentación para ello

04

Resultados

Visualización de los resultados

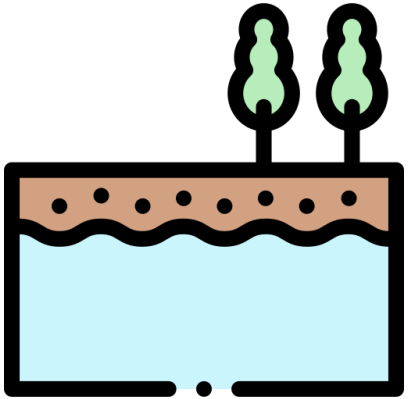
05

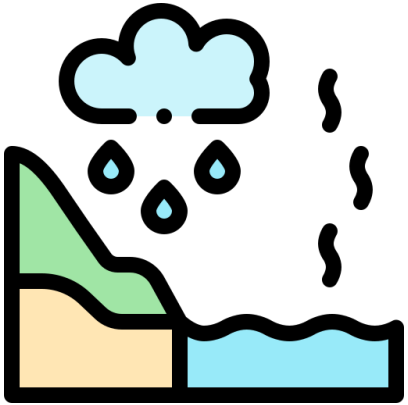
Conclusiones

Hallazgos tras ver los resultados

Selección de Dataset

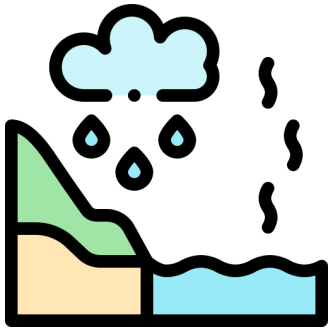
La primera tarea asignada fue la de seleccion de un dataset. Se nos dieron dos opciones. La de cuerpos de agua superficiales y cuerpos de agua subterraneos en México.

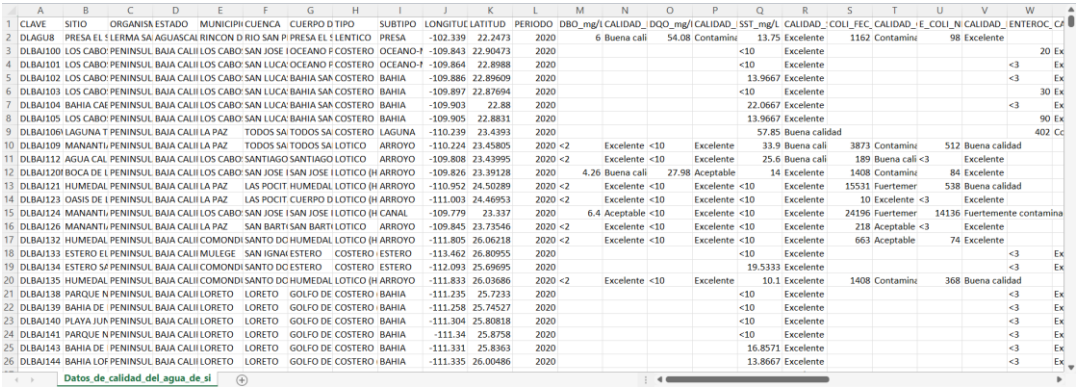




Justificación de selección de “Dataset”

Decidimos optar por el de cuerpos de agua superficiales en México.
El dataset es un campo de datos pero con mucha oportunidad para aplicar diferentes técnicas de limpieza y poner en practica métodos aprendidos a lo largo del curso. Se puede visualizar espacios en blanco en la imagen.





	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	CLAVE	SITIO	ORGANISM	ESTADO	MUNICIPI	CUENCA	CUERPO D	TIPO	SUBTIPO	LONGITUD	LATITUD	PERIODO	DBO_mg/L	CALIDAD_	DQO_mg/L	CALIDAD_	SST_mg/L	CALIDAD_	COLI_FEC	CALIDAD_	E_COLI_N	CALIDAD_	ENTEROC_CA
2	DLAGU8	PRESA EL S	LERMA SA	AGUASCAL	RINCON D	RIO SAN P	PRESA EL S	LENTICO	PRESA	-102.339	22.2473	2020	6	Buena cali	54.08	Contamina	13.75	Excelente	1162	Contamina	98	Excelente	
3	DLBAJ100	LOS CABO	PENINSUL	BAJA CALI	LOS CABO	SAN JOSE I	OCEANO P	COSTERO	OCEANO-I	-109.843	22.90473	2020					<10	Excelente					20
4	DLBAJ101	LOS CABO	PENINSUL	BAJA CALI	LOS CABO	SAN LUCA	OCEANO P	COSTERO	OCEANO-I	-109.864	22.8988	2020					<10	Excelente					<3
5	DLBAJ102	LOS CABO	PENINSUL	BAJA CALI	LOS CABO	SAN LUCA	BAHIA SAN	COSTERO	BAHIA	-109.886	22.89609	2020					13.9667	Excelente					<3
6	DLBAJ103	LOS CABO	PENINSUL	BAJA CALI	LOS CABO	SAN LUCA	BAHIA SAN	COSTERO	BAHIA	-109.897	22.87694	2020					<10	Excelente					30
7	DLBAJ104	BAHIA CAE	PENINSUL	BAJA CALI	LOS CABO	SAN LUCA	BAHIA SAN	COSTERO	BAHIA	-109.903	22.88	2020					22.0667	Excelente					<3
8	DLBAJ105	LOS CABO	PENINSUL	BAJA CALI	LOS CABO	SAN LUCA	BAHIA SAN	COSTERO	BAHIA	-109.905	22.8831	2020					13.9667	Excelente					90
9	DLBAJ106	LAGUNA T	PENINSUL	BAJA CALI	LA PAZ	TODOS SA	TODOS SA	COSTERO	LAGUNA	-110.239	23.4393	2020					57.85	Buena calidad					402
10	DLBAJ109	MANANTI	PENINSUL	BAJA CALI	LA PAZ	TODOS SA	TODOS SA	LOTICO	ARROYO	-110.224	23.45805	2020	<2	Excelente	<10	Excelente	33.9	Buena cali	3873	Contamina	512	Buena calidad	
11	DLBAJ112	AGUA CAL	PENINSUL	BAJA CALI	LOS CABO	SANTIAGO	SANTIAGO	LOTICO	ARROYO	-109.808	23.43995	2020	<2	Excelente	<10	Excelente	25.6	Buena cali	189	Buena cali	<3	Excelente	
12	DLBAJ120	BOCA DE L	PENINSUL	BAJA CALI	LOS CABO	SAN JOSE I	SAN JOSE I	LOTICO	ARROYO	-109.826	23.39128	2020	4.26	Buena cali	27.98	Aceptable	14	Excelente	1408	Contamina	84	Excelente	
13	DLBAJ121	HUMEDAL	PENINSUL	BAJA CALI	LA PAZ	LAS POCIT	HUMEDAL	LOTICO	ARROYO	-110.952	24.50289	2020	<2	Excelente	<10	Excelente	<10	Excelente	15531	Fuertemer	538	Buena calidad	
14	DLBAJ123	OASIS DE	PENINSUL	BAJA CALI	LA PAZ	LAS POCIT	CUERPO D	LOTICO	ARROYO	-111.003	24.46953	2020	<2	Excelente	<10	Excelente	<10	Excelente	10	Excelente	<3	Excelente	
15	DLBAJ124	MANANTI	PENINSUL	BAJA CALI	LOS CABO	SAN JOSE I	SAN JOSE I	LOTICO	CANAL	-109.779	23.337	2020	6.4	Aceptable	<10	Excelente	<10	Excelente	24196	Fuertemer	14136	Fuertemente contamina	
16	DLBAJ126	MANANTI	PENINSUL	BAJA CALI	LA PAZ	SAN BARTI	SAN BARTI	LOTICO	ARROYO	-109.845	23.73546	2020	<2	Excelente	<10	Excelente	<10	Excelente	218	Aceptable	<3	Excelente	
17	DLBAJ132	HUMEDAL	PENINSUL	BAJA CALI	COMONDU	SANTO DO	HUMEDAL	LOTICO	ARROYO	-111.805	26.06218	2020	<2	Excelente	<10	Excelente	<10	Excelente	663	Aceptable	74	Excelente	
18	DLBAJ133	ESTERO EL	PENINSUL	BAJA CALI	MULEGE	SAN IGNA	ESTERO	COSTERO	ESTERO	-113.462	26.80955	2020					<10	Excelente					<3
19	DLBAJ134	ESTERO SA	PENINSUL	BAJA CALI	COMONDU	SANTO DO	ESTERO	COSTERO	ESTERO	-112.093	25.69695	2020					19.5333	Excelente					<3
20	DLBAJ135	HUMEDAL	PENINSUL	BAJA CALI	COMONDU	SANTO DO	HUMEDAL	LOTICO	ARROYO	-111.833	26.03686	2020	<2	Excelente	<10	Excelente	10.1	Excelente	1408	Contamina	368	Buena calidad	
21	DLBAJ138	PARQUE N	PENINSUL	BAJA CALI	LORETO	LORETO	GOLFO DE	COSTERO	BAHIA	-111.235	25.7233	2020					<10	Excelente					<3
22	DLBAJ139	BAHIA DE	PENINSUL	BAJA CALI	LORETO	LORETO	GOLFO DE	COSTERO	BAHIA	-111.258	25.74527	2020					<10	Excelente					<3
23	DLBAJ140	PLAYA JUN	PENINSUL	BAJA CALI	LORETO	LORETO	GOLFO DE	COSTERO	BAHIA	-111.304	25.80818	2020					<10	Excelente					<3
24	DLBAJ141	PARQUE N	PENINSUL	BAJA CALI	LORETO	LORETO	GOLFO DE	COSTERO	BAHIA	-111.34	25.8758	2020					<10	Excelente					<3
25	DLBAJ143	BAHIA DE	PENINSUL	BAJA CALI	LORETO	LORETO	GOLFO DE	COSTERO	BAHIA	-111.331	25.8363	2020					16.8571	Excelente					<3
26	DLBAJ144	BAHIA LOF	PENINSUL	BAJA CALI	LORETO	LORETO	GOLFO DE	COSTERO	BAHIA	-111.335	26.00486	2020					13.8667	Excelente					<3

Datos_de_calidad_del_agua_de_si

Imagen 1. Dataset elegido, referencia visual

Limpieza

Se corrieron métodos para identificar espacios vacíos dentro de la base de datos. Así como para identificar el tipo de elemento para trabajar. La cantidad de elementos faltantes dentro de los atributos.
Técnicas que se implementaron:
-Se eliminaron columnas y filas que no agregaban valor al análisis
-Se cambiaron formatos para poder hacer las interpretaciones adecuadas

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3493 entries, 0 to 3492
Data columns (total 55 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   CLAVE                 3493 non-null   object
 1   SITIO                 3493 non-null   object
 2   ORGANISMO_DE_CUENCA   3493 non-null   object
 3   ESTADO                3493 non-null   object
 4   MUNICIPIO             3493 non-null   object
 5   CUENCA                3492 non-null   object
 6   CUERPO DE AGUA        3479 non-null   object
 7   TIPO                  3493 non-null   object
 8   SUBTIPO               3479 non-null   object
 9   LONGITUD              3493 non-null   float64
 10  LATITUD               3493 non-null   float64
 11  PERIODO               3493 non-null   int64
```

Imagen 2. Formato de los datos.

```
[ ]  #Revisamso si tenemos valores faltantes
     df.isnull().values.any()

     True
```

```
     #El resultado anterior nos marco en efect
     #En nuestra primer etapa de validación de
     #decidimos primero ubicar las columnas co
     df.isna().any()

     CLAVE                 False
     SITIO                 False
     ORGANISMO_DE_CUENCA   False
     ESTADO                False
     MUNICIPIO             False
     CUENCA                 True
     CUERPO DE AGUA         True
     TIPO                  False
     SUBTIPO                True
     LONGITUD              False
```

Imagen 3. Presencia de elementos faltantes

Análisis

Definimos los atributos con los que vamos a trabajar.
1.- Nos limitamos a trabajar con variables numéricas y graficamos su comportamiento.

Imagen 4. Elementos nicho y su comportamiento

2 Revisamos que los atributos nichos con los que trabajamos tengan una correlación directa o una correlación indirecta alta. Nos percatamos de una correlación directa alta y nos permite seguir con el análisis.

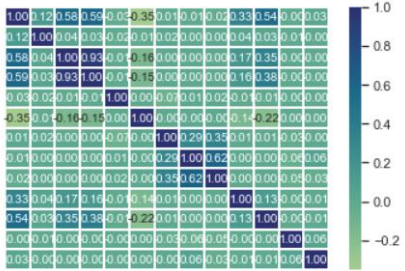


Imagen 5. Confirmación de correlacion

3.Todo el análisis parece bien realizado, es necesario identificar si tenemos “outliers” en una presencia significativa como para afectar nuestro análisis. En la siguiente grafica vemos que si tenemos pero son minoría como para poder ser despreciados en las interpretaciones, sin la necedad de una normalización extra.

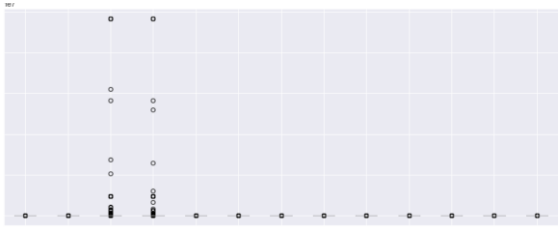


Imagen 6. Mapa de comportamiento.

Clasificación

La clasificación y ubicación de los centroides fue revisada con un método de matriz de confusión y se determino la precisión de este. Una precisión superior al 99%. Por ende podemos decir que los datos son muy conclusivos en la siguiente imagen.
El método utilizada para mostrar lo siguiente es denominado “K-means”.

El método agrupa las formas de agua superficiales en una categoría de semáforos. La verde siento la mejor condición, amarilla poco recomendada y la roja no recomendada.

A su vez podemos ubicar los centroides marcados por puntos azules en la grafica.

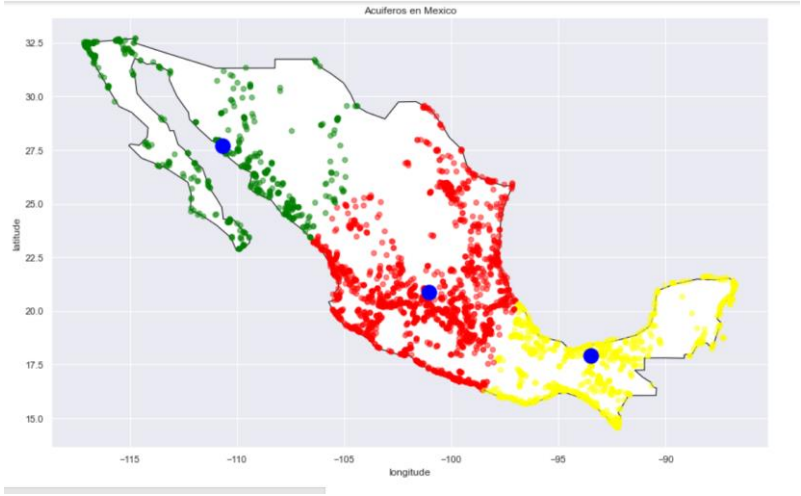


Imagen 7. Mapa de México usando el método “K-means”

Resultados

Es necesario analizar las siguientes graficas para tener una mejor perspectiva. Los mantos aquiferos superficiales fueron analizados de manera individual y podemos ver que tras tomar en cuenta presencias de metales pesados, cargaso bacterianas, así como otros aspectos nocivos para la salud humana podemos identificar en las costas de nuestro país una carga notoria de agua “buena” para nuestra humanidad pero si utilizamos un método de ML para hacer este análisis y otro método para corraborrar su efectividad podemos ver que hay una fuerte tendencia en la zona nor-oeste del país a una calidad de agua superior, denotando que la zona centro y zona nor-este un “no-recomendado” y finalmente en la zona sur y la península una presencia de agua “poco recomendada”.

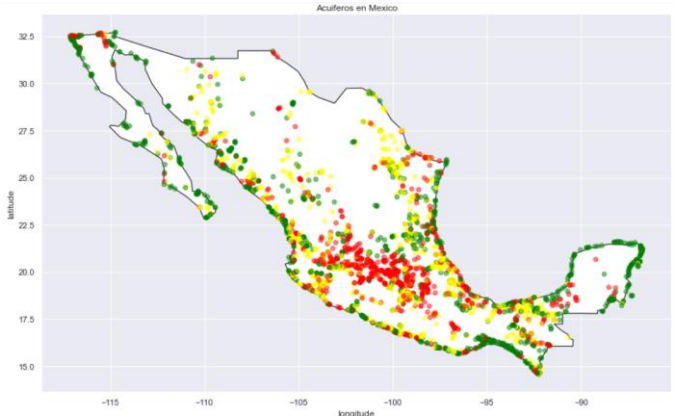


Imagen 8. Mapa de México de la calidad de los mantos acuíferos

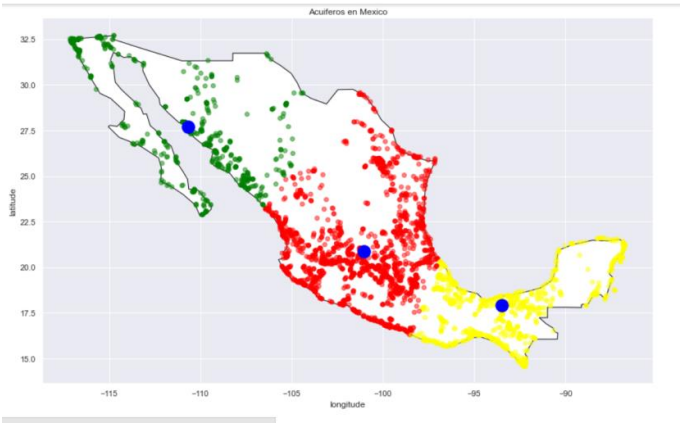
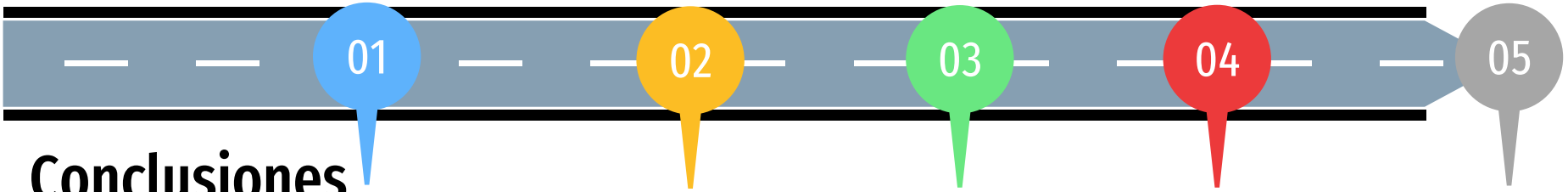


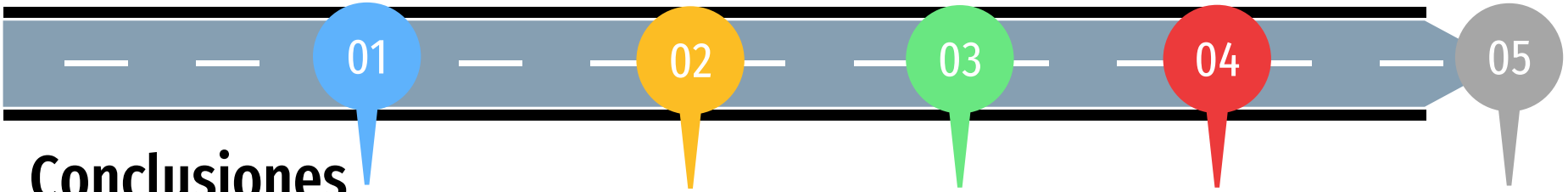
Imagen 9. Mapa de México usando el método “K-means”



Conclusiones

Las conclusiones de nuestro muestreo son las siguientes:

1.- La zona norte-oeste tiene mayor calidad de agua, centro y nor-este un “no recomendada” y la zona sur y península “poco recomendada”.
2.- Este ejercicio no permite acciones conclusivas, un análisis de ciencia de datos debe estar basado en problemáticas del negocio, la única indicación adecuada que se puede entregar es que puede servir para las estrategias de organismos gubernamentales en sus planes de acción.
3.- El ejercicio y los métodos utilizados son de gran poder y permiten una perspectiva global de la geografía del país en temas de mantos acuíferos superficiales, la comparación de estos hallazgos con las cuencas de agua entregaría mayor valor.



Conclusiones

Las conclusiones de nuestro muestreo son las siguientes:

4.- El análisis implementado requirió de la necesidad de borrar una gran cantidad de atributos debido al manejo del banco de datos, una propuesta directa del equipo de ciencia de datos es que el equipo de recolección de información no tiene un “modern data warehouse”. Estos análisis pueden entregar mayor valor al tener la infraestructura técnica adecuada.
5.- Es necesario hacer estos análisis comparativos a lo largo de 20 o 30 años para ver una evolución en los tratamientos y acciones de la sociedad junto con gobierno, o en su defecto ver los efectos nocivos.

Proyecto

Actividad de cierre de la materia de Ciencia de datos
Maestría en Inteligencia artificial aplicada
ITESM