Reto calidad del agua.

Ciencia y Analítica de datos.

Tecnológico de Monterrey

Profesor Titular Maria de la Paz Rico.
Profesor Tutor. Mtro. Mario Alberto Solano Saldaña.

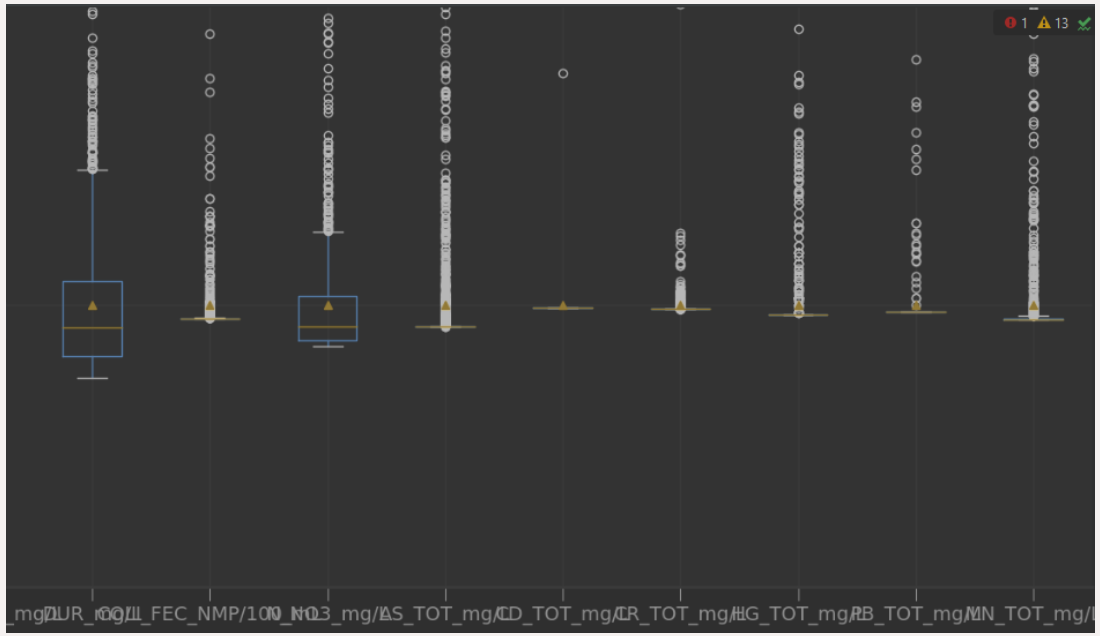
- Lázaro Lara Martínez. Matricula A01793198
- José Mtanous Treviño. Matricula A00169781

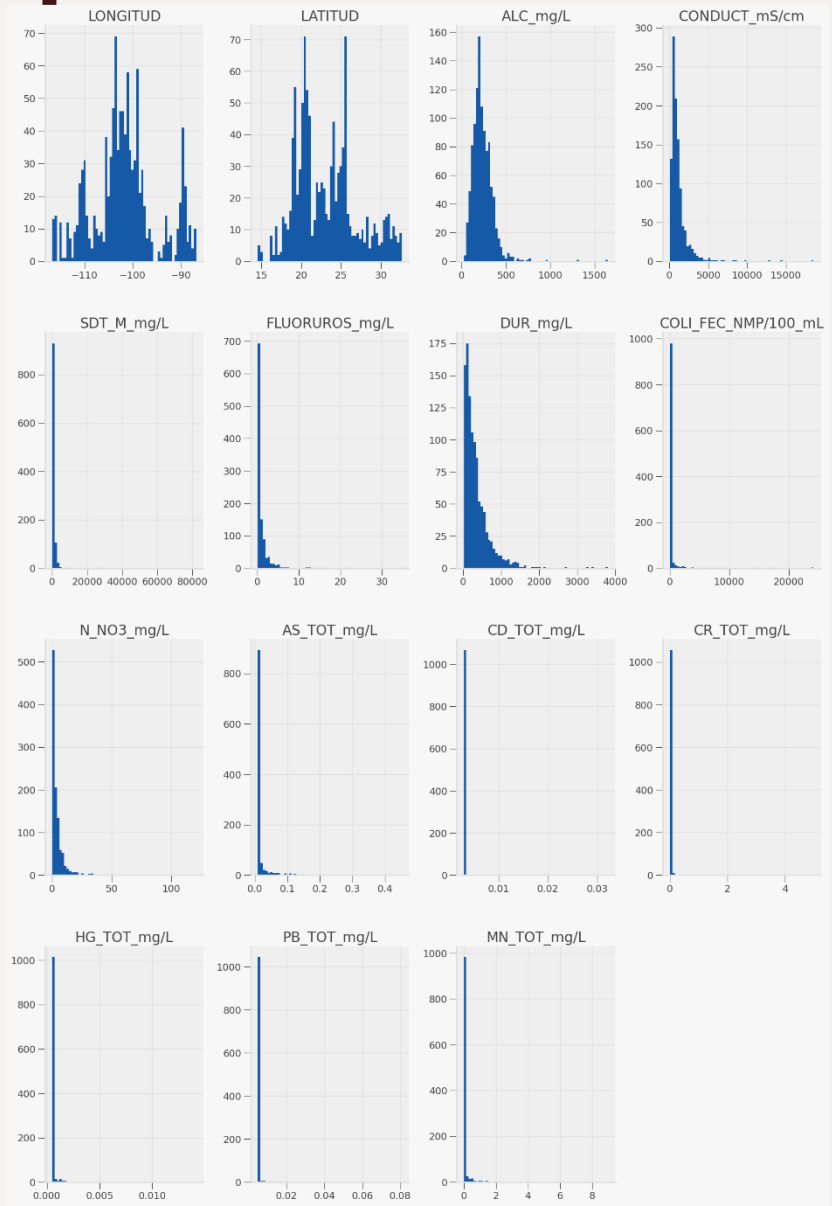
Limpieza, análisis, visualización y kmeans

- Para obtener el modelo de Kmeans, primero hicimos limpieza de los datos
- Encontramos varios problemas en los datos como:
- La columna SDT_mg/L no contenía datos asi que se eliminó.
- Campos numéricos donde algunos registros tenian el símbolo <, se eliminó y dejo solo el valor numérico, gracias a esto pudimos cambiar el tipo de dato de object a float en el dataframe.
- Columna CONTAMINANTES, Esta columna contiene una lista separada por comas de los contaminantes encontrados en las muestras, sin embargo las muestras que no tienen contaminantes tienen este registro nulo. En general un campo nulo se entiendo como una muestra no capturada o un campo donde no hay información pero en este caso no es así. Una muestra sin contaminantes indica que el agua está limpia. Además varias librerías ignoran los valores nulos al momento de procesar los datos. Para quitar esta ambigüedad, vamos a sustituir los valores nulos por el string 'NINGUNO'
- Al terminar el proceso de limpieza guardamos los datos en un nuevo archivo.

Identificando tendencias centrales promedio, media y mediana de los datos.

- Ambas gráficas Boxplot y barras muestran que las variables tienen muchos outliers, aún cuando los datos se normalizaron utilizando zscore.



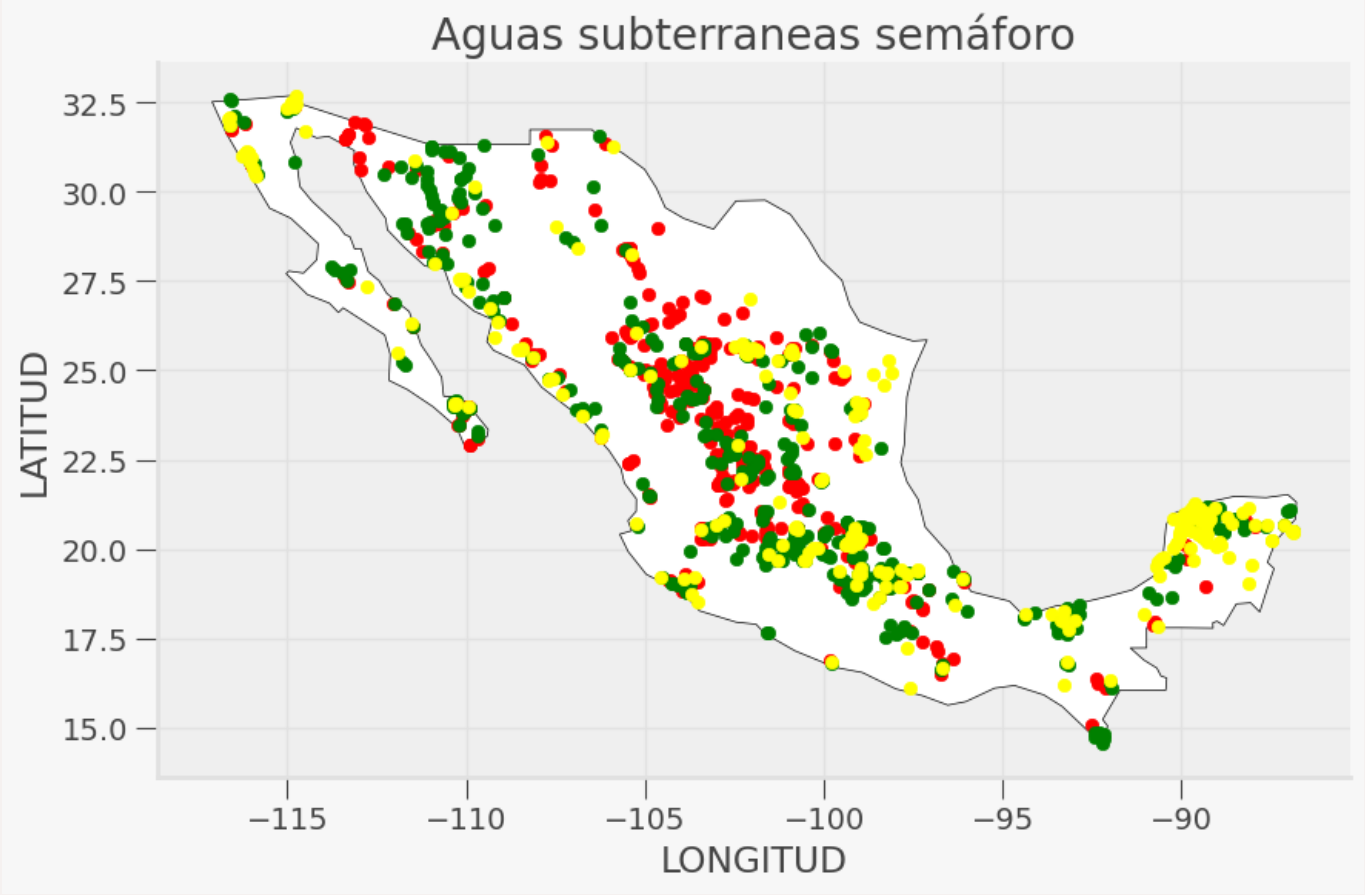


Identificar correlaciones.

- En la gráfica de correlación se puede apreciar que esta es muy baja entre las variables numéricas, a excepción de HG_TOT_mg/L vs MN_TOT_mg/L, entre las demás variable la correlación es muy baja.

Realizar análisis para encontrar si existe una relación entre la calidad del agua y su ubicación geográfica a través de K- means.

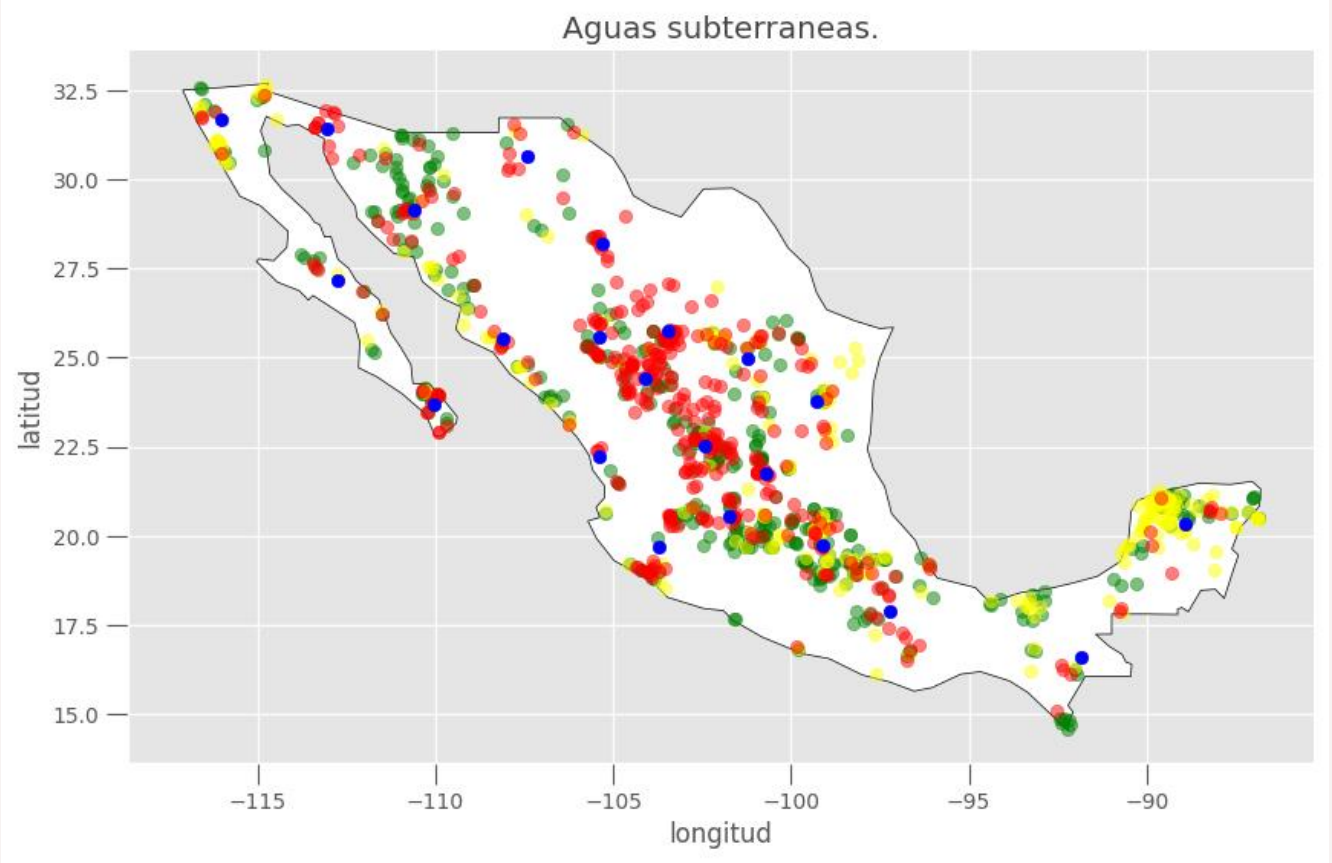
- Como primer paso ilustramos en el mapa todos los puntos en base al semáforo, para tartar de identificar a simple vista si existe una agrupación natural de la calidad del agua respect a su localización geográfica.
- Pudimos notar que no es así, a excepción del noreste del pais, todos los puntos estan mezclados con los diferentes colores.



Mostrar resultados de agrupamiento de latitudes y longitudes con K means en el mapa de México.

Buscamos el número mínimo de clusters que al menos agruparan correctamente las muestras de las Baja Californias y el número que encontramos fue de 22.

Al visualizar las muestras con diferentes colores y dibujarlas en el mapa de México, a primera vista no encontramos una relación entre la calidad del agua (semáforo) y la ubicación geográfica. Creemos que un dato relevante que se omitió en las muestras, es la profundidad de los pozos, esto es fundamental para poder entender el origen y posible relación entre mantos freáticos.



Clasificación

La meta de nuestro modelo es usar datos que sean fáciles de adquirir y de preferencia que se puedan obtener en el mismo lugar de la muestra y con esos datos predecir la calidad del agua. Este modelo tendría utilidad práctica y podría ser una primera herramienta para hacer un catastro más amplio de la calidad de agua en pozos.
Con nuestro modelo y un kit básico de medición podríamos predecir la calidad del agua de una manera rápida sin necesidad de enviar las muestras al laboratorio, con el riesgo de contaminarlas y el alto costé de usar equipo especializado.

Variables

Independientes

- Latitud
- Longitud
- Subtipo [1,2,3,4,5,6,7]
- Contaminantes

Dependientes

- Semáforo [1,2,3]

LATITUD	LONGITUD	SUBTIPO	ALC	AS	CD	CF	CONDUC	CR	DT	FE	FLUO	HG	MN	NI	NO3	PB	SDT_ra	SDT_salin
22.20887	-102.02210	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21.99958	-102.20075	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22.36685	-102.28801	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
22.18435	-102.29449	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23.45138	-110.24480	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

Clasificadores

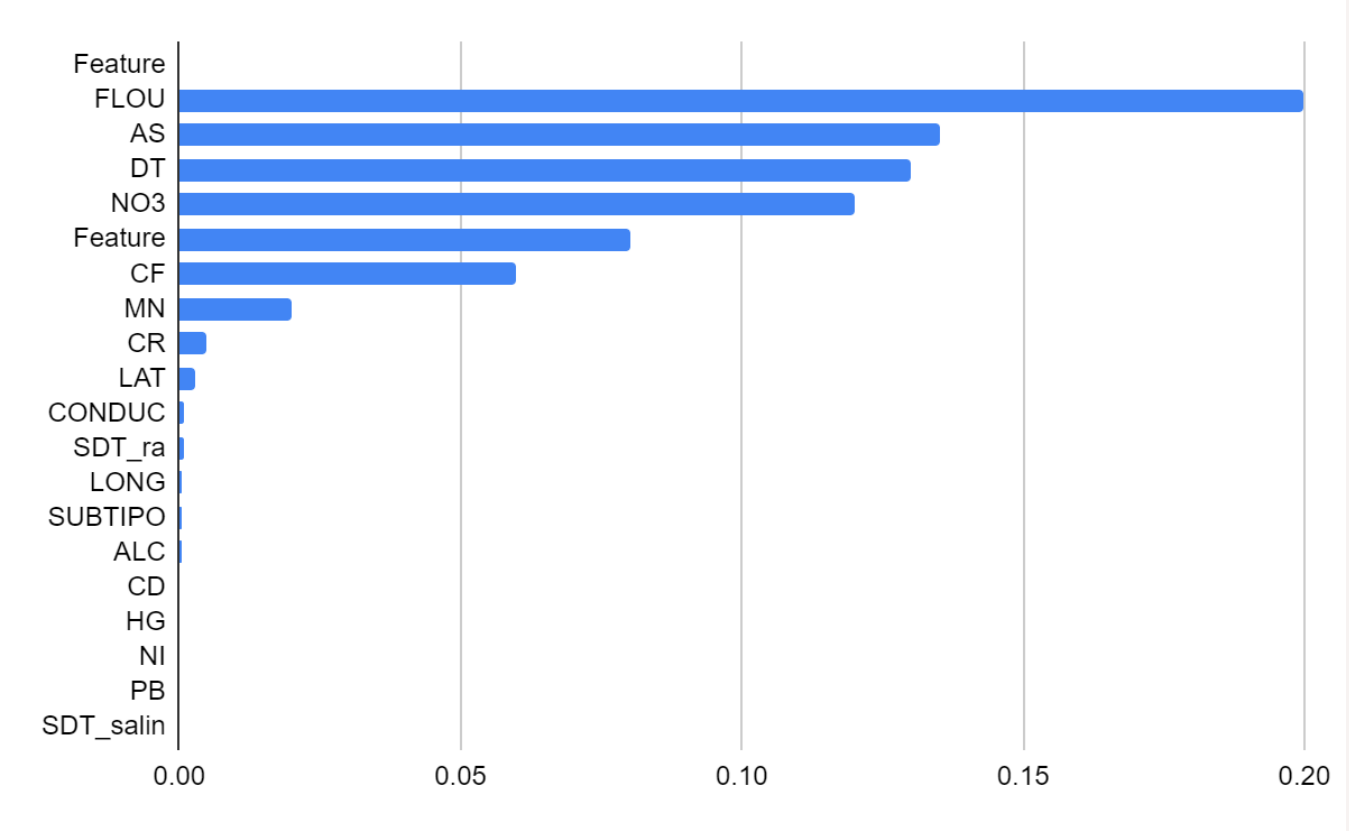
DecisionTreeClassifier

- Mejores Parámetros
- {'criterion': 'gini', 'max_depth': 5, 'min_samples_leaf': 2, 'min_samples_split': 7}
- Score: 0.88

RandomForestClassifier

- Mejores Parámetros
- {'criterion': 'entropy', 'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 7}
- Score: 0.96

Features Importance



Conclusión

La calidad del agua no tiene correlación con su ubicación geográfica

Fuerte correlación entre contaminantes y calidad del agua

Alto desempeño de los clasificadores

Reducción de costo de clasificación en campo