Tecnológico de Monterrey

CIENCIA Y ANALÍTICA DE DATOS

Reto

Profesor: María de la Paz Rico Fernández

Alumno: Juan Sebastián Ortega Briones A01794327

Equipo 13

18 de Noviembre del 2022

AGENDA

Datos

Limpieza

Análisis

Kmeans

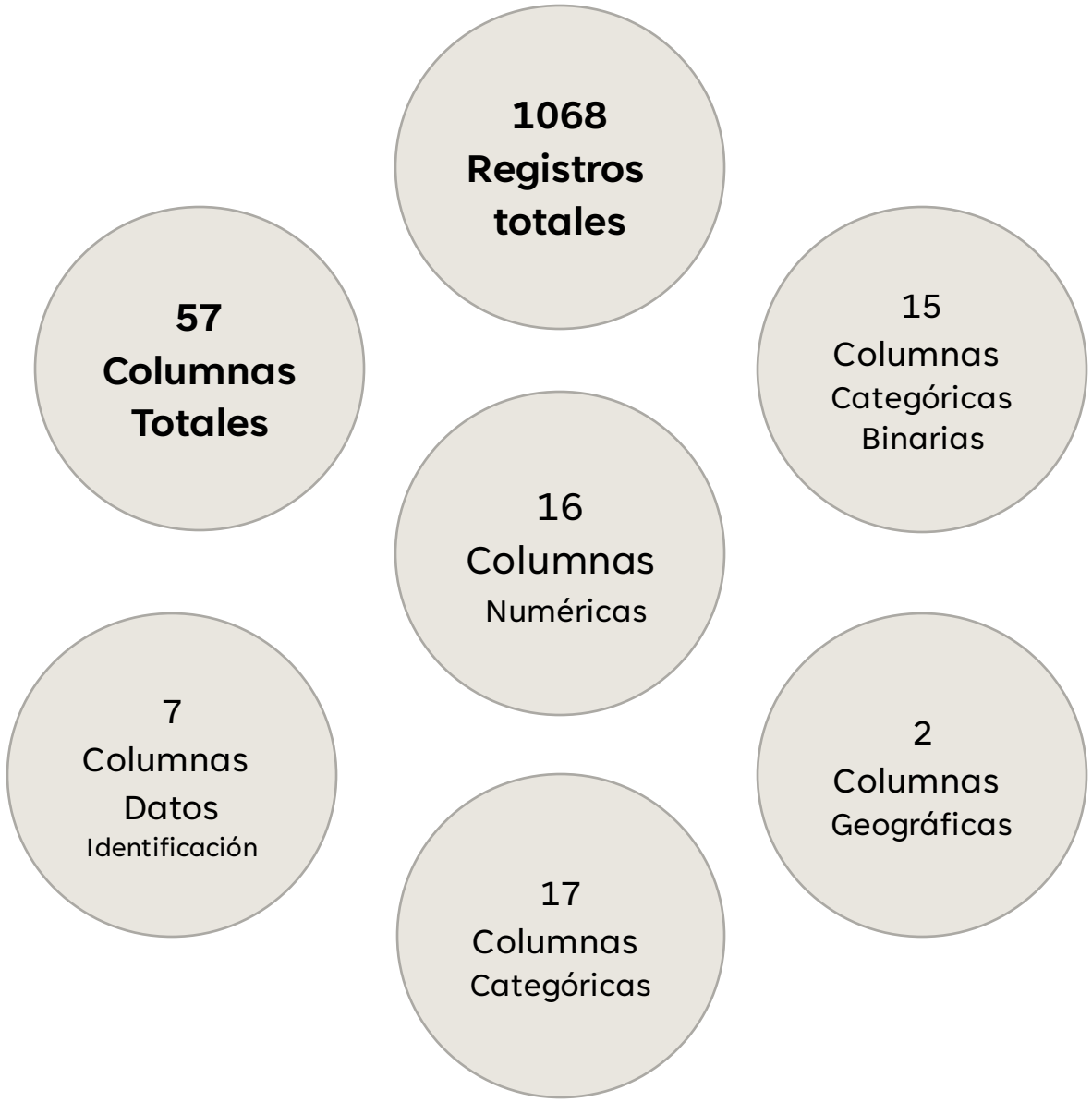
Clasificación

Resultados

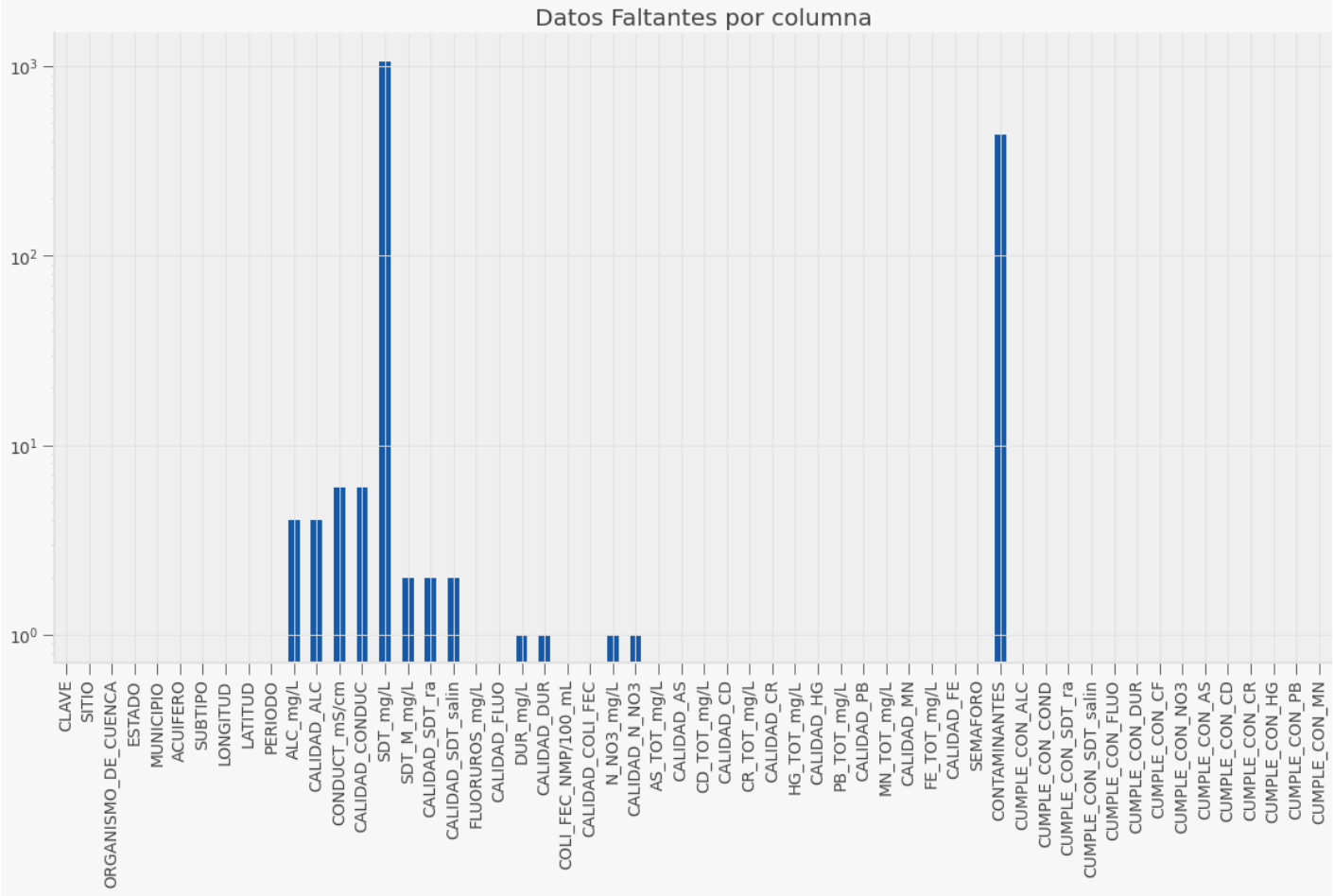
Conclusiones

Los datos:

Datos de calidad del agua subterránea de 5000 estaciones a nivel nacional del año 2020



LIMPIEZA



Muchos datos númericos estaban dados en formato '< x'

Se sustituyeron por 0 asumiendo que dejaron de ser significativos

14 filas contenian datos vacios (NaN)

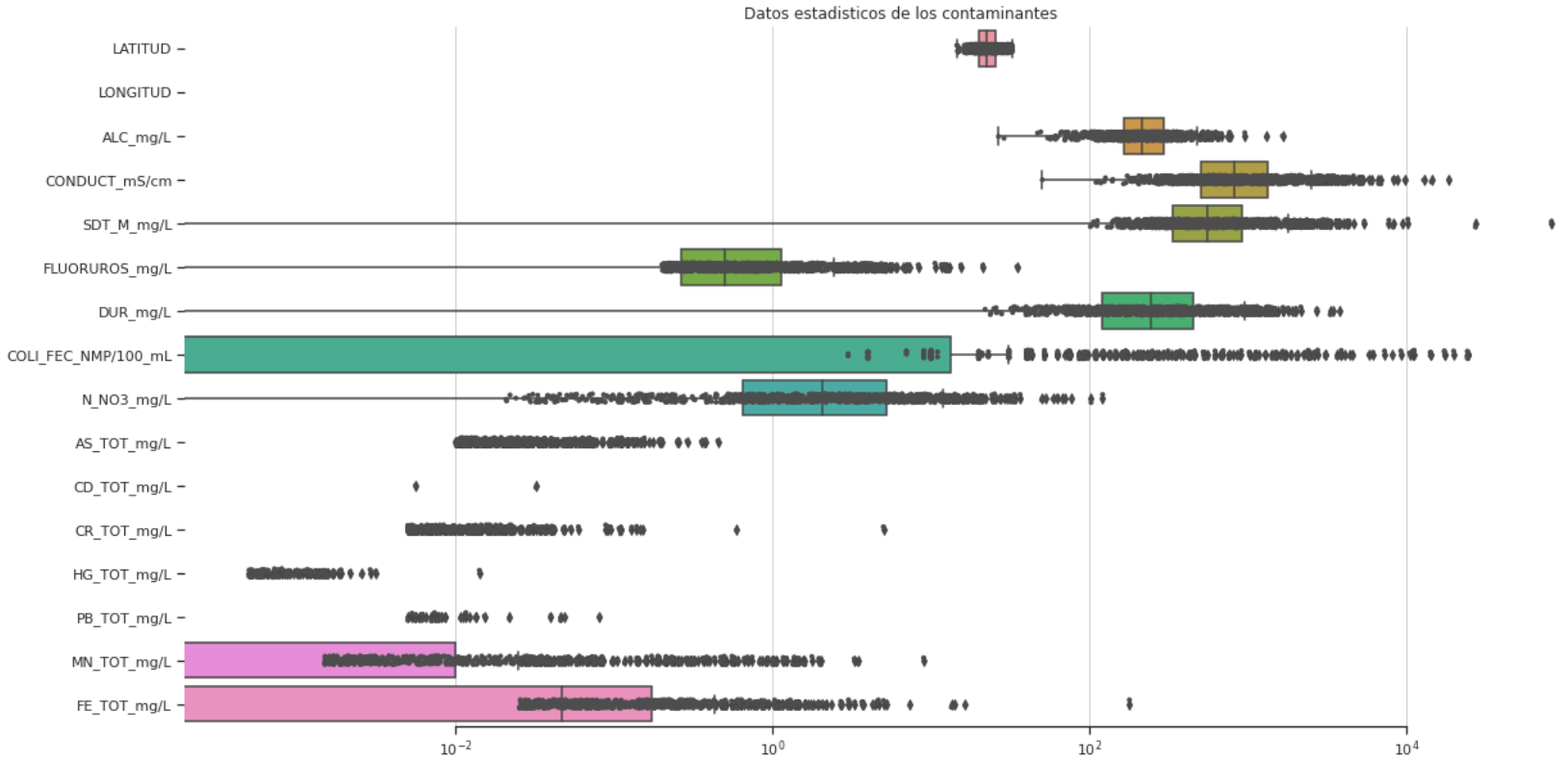
Se eliminaron

La columna 'SDT_mg/L' columna no contenia ningun dato

Se elimino la columna

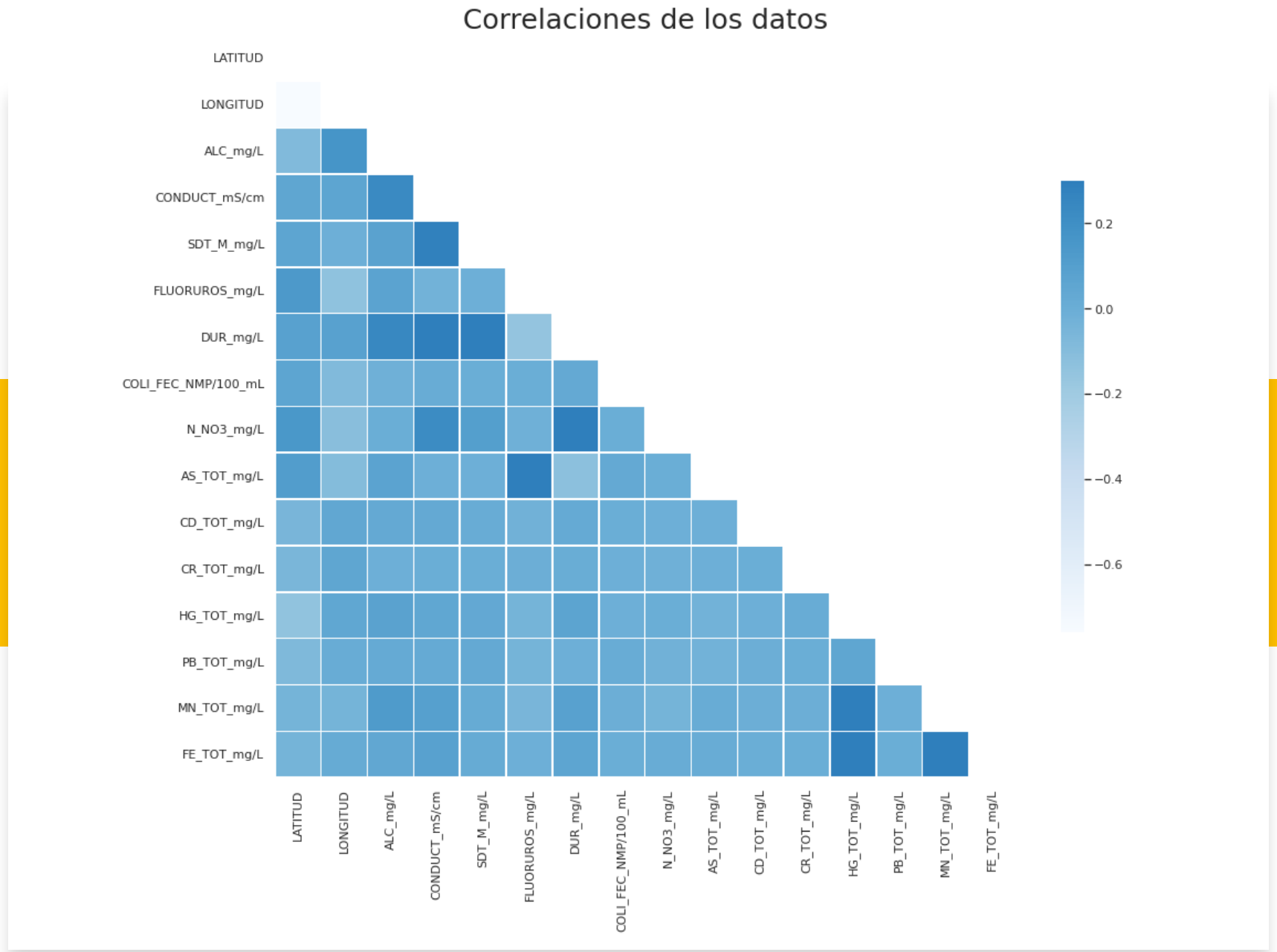
La columna 'Contaminantes' es redundante con las categoricas

Se elimino



ANALISIS

Calidad del
agua subterránea 2022

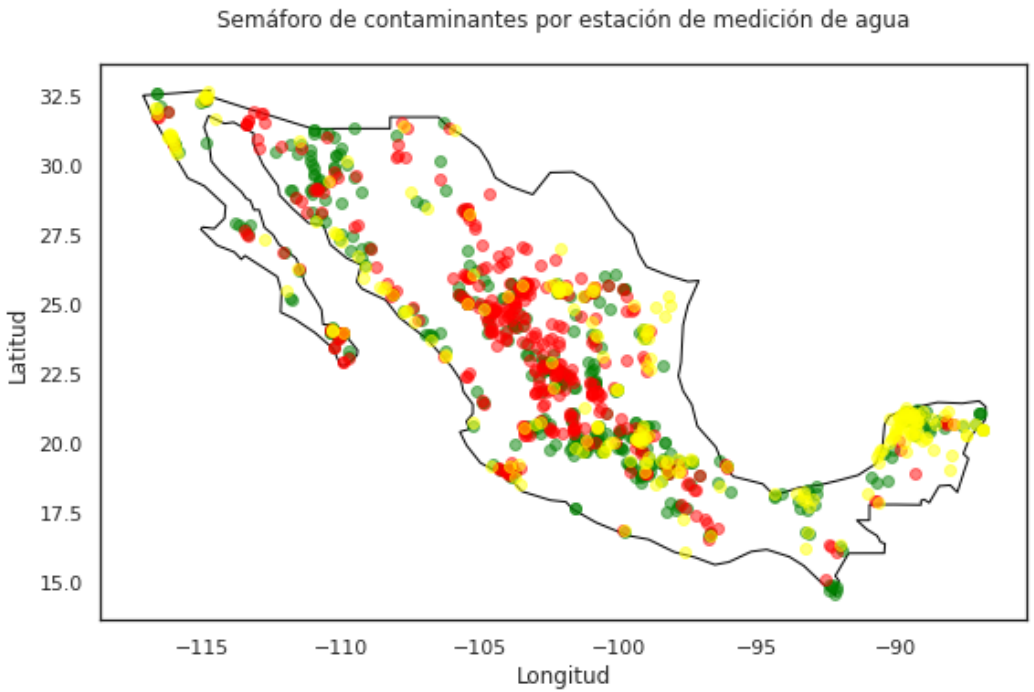


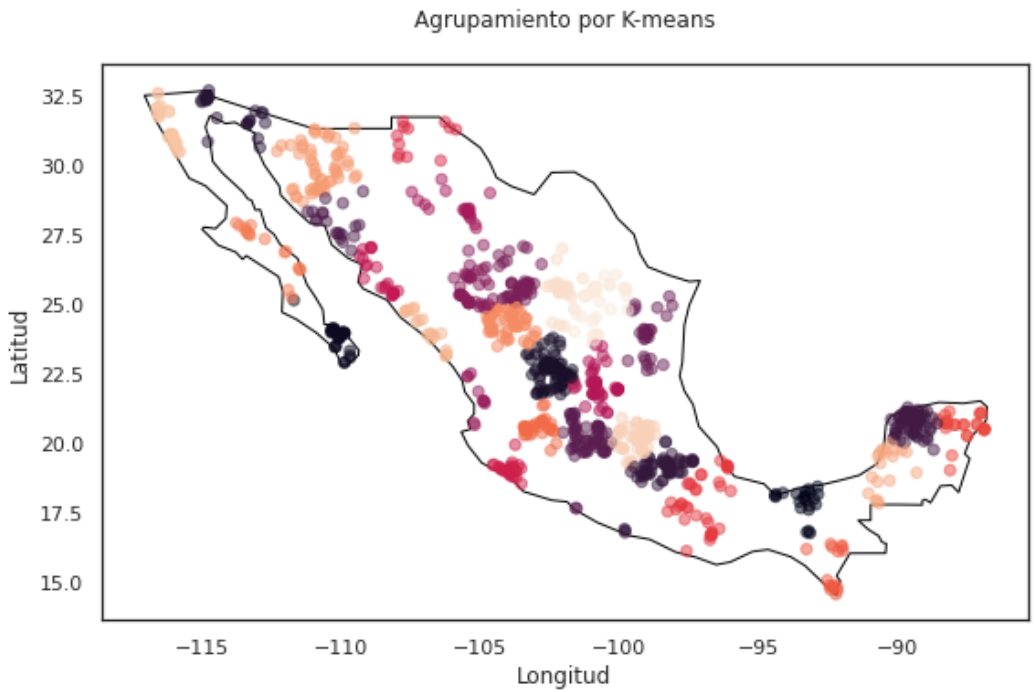
ANALISIS

Calidad del
agua subterránea 2022

KMEANS

Comparativo de semáforo contra predicción de kmeans

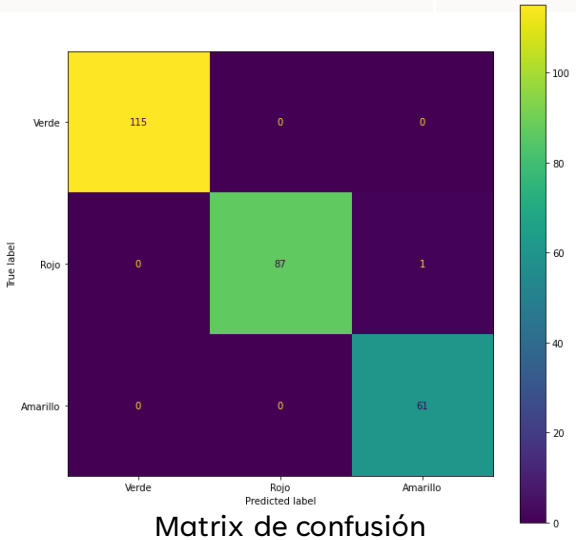




CLASIFICACIÓN

Después del análisis, se puede ver que el método más eficiente es SVC

Metodo	**Eficiencia**
LogisticRegression	0.977
RandomForestClassifier	0.992
SVC	0.996
VotingClassifier	0.996



Matrix de confusión

790 (75%)
Datos de entrenamiento

264 (25%)
Datos de prueba

Variables más importantes

Porcentaje de importancia	**Variable**
43.68 %	CUMPLE_CON_DUR
18.49 %	CUMPLE_CON_MN
17.13 %	CUMPLE_CON_FLUO
13.07 %	CUMPLE_CON_FE
4.40 %	CUMPLE_CON_NO3
2.43 %	CUMPLE_CON_CF
0.76 %	CUMPLE_CON_AS

RESULTADOS

Datos y análisis

La base de datos es muy buena, tiene muy pocos datos vacíos

La correlación más alta es entre dureza y fluoruros.

Es interesante la correlación entre el plomo y la latitud que es la segunda más alta.

kmeans

Usando kmeans con latitudes y longitudes me parece que no se observa una relación clara entre custers y calidad del agua, aunque si es clara en clusters y cercanía de estaciones de medición.

Clasificación

Inicialmente use los datos numéricos de la BD y la eficiencia mayor fue con random forest con un 97% y la más baja con SVC con 60%, al cambiar los datos numéricos por los datos categóricos binarios la eficiencia del modelo subió casi a 100% siendo SVC el más eficiente con 99.60% y random fores el menos eficiente con 99.20% al pasar la votación llego al 99.60%, viendo la matriz de confusión solo fallo un solo dato.

CONCLUCIONES

Usando el método de ensamble con votación suave es posible determinar con gran eficiencia (>99%), la potabilidad del agua subterránea usando las 7 variables principales.

Para determinar clusters geográficos que determinen la calidad del agua es necesario usar estas variables en kmeans: DUR_mg/L, MN_TOT_mg/L, FLUORUROS_mg/L, FE_TOT_mg/L