Resumen Ejecutivo
Ciencia y Analítica de datos

Profesora: María de la Paz Rico Fernandez

Equipo 71:
- Ignacio Antonio Quintero Chávez | A01794419
- Francisco Xavier Bastidas Moreno | A01794188

Fecha: 18 de Noviembre del 2022.

Contenido

Describe los principales hallazgos del análisis y clasificación de los datos por medio de los modelos de Decision Trees y Random Forest. La presentación deberá incluir todos los pasos del pipeline seguidos, limpieza, análisis, kmeans, clasificación, resultados y conclusiones.

01

Dataset

Introducción al data set "Datos de calidad del agua de sitios de monitoreo de aguas subterráneas"

02

Limpieza y analisis

Selecciona de variables independientes y dependiente (semáforo).

03

Modelo y metricas

Balanceo de clases y creación del clasificador

04

Técnicas utilizadas

Desarrollo de Decision Trees y Random Forest

05

Resultados

Presentación de las métricas seleccionadas y grafica de confusión del modelo.

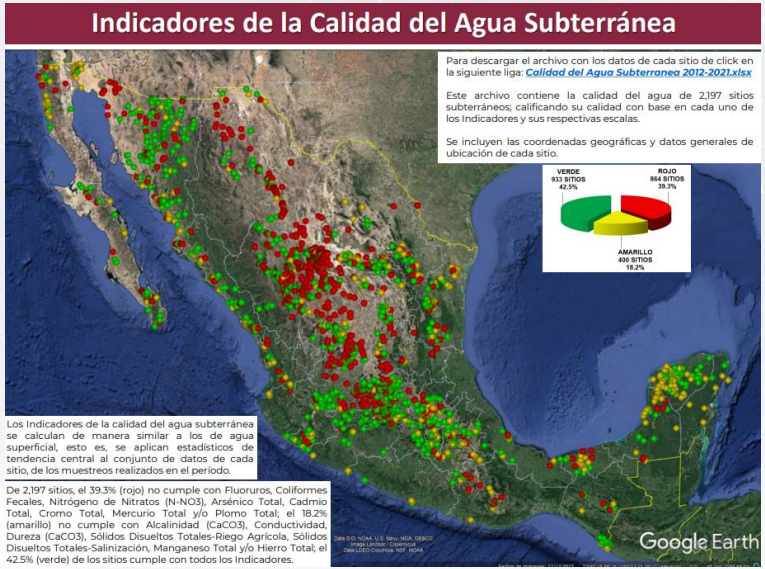
06

Conclusiones

Principales hallazgos y análisis final

1. Dataset

- **Nombre del dataset:** Datos_de_calidad_del_agua_de_sitios_de_monitoreo_de_aguas_subterraneas_2020.
- **Columnas:** 57
- **Filas:** 1068
- Algunas de las columnas representan información general, mientras que otras presentan información numérica que es relevante para el análisis.



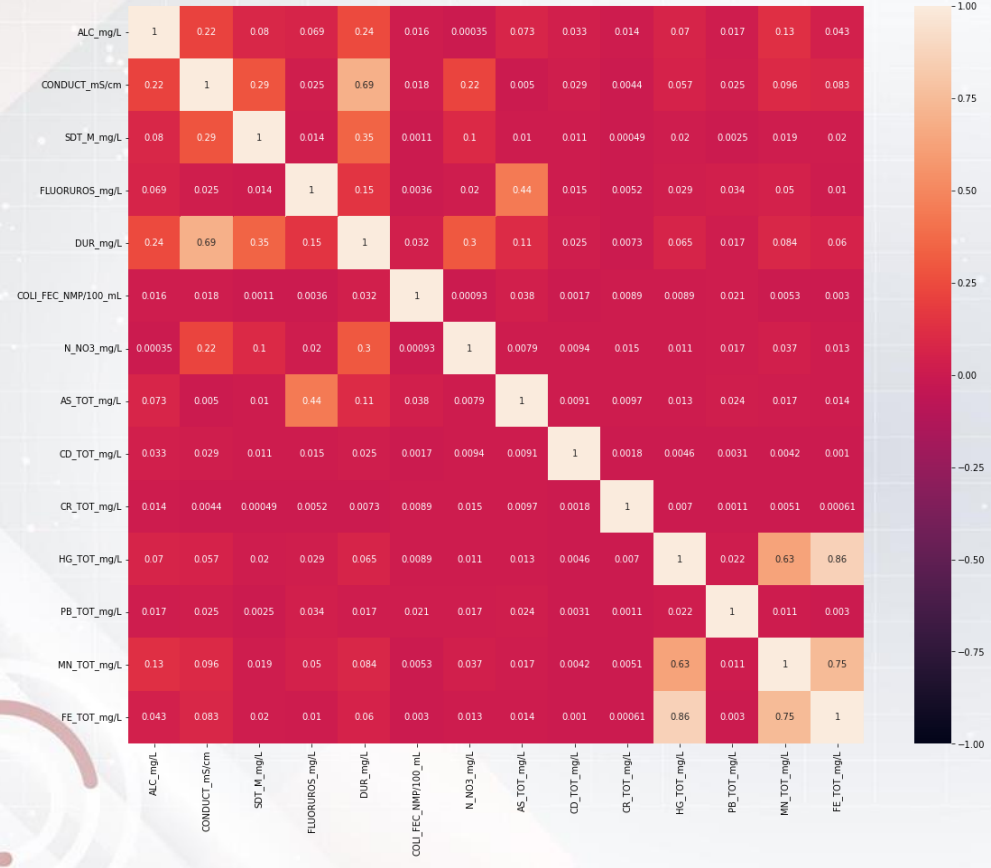
2. Limpieza y Análisis

1. Se exploraron los datos y que tipo son.
2. Se obtuvo la suma de valores perdidos por cada columna.
3. Determinamos las columnas que son numéricas.
4. Obtenemos los valores únicos de cada columna.
5. Se convirtieron los datos de tipo objeto a tipo flotante.
6. Se utilizó la mediana para imputar los datos.
7. Identificamos correlaciones y utilizamos un diagrama de caja y bigote para identificar outliers.
8. Se utilizó un mapa de calor para identificar la correlación de nuestras variables.

	ALC_mg/L	CONDUCT_mS/cm	SDT_mg/L	SDT_M_mg/L	FLUORUROS_mg/L	DUR_mg/L	COLI_FEC_NMP/100_mL	N_NO3_mg/L	AS_TOT_mg/L	CD_TOT_mg/L	CR_TOT_mg/L	HG_TOT_mg/L	PB_TOT_mg/L	MN_TOT_mg/L	FE_TOT_mg/L
0	229.990	940.0	NaN	603.6	0.9766	213.732	<1.1	4.184656	0.0161	<0.003	<0.005	<0.0005	<0.005	<0.0015	0.0891
1	231.990	608.0	NaN	445.4	0.9298	185.0514	<1.1	5.75011	0.0134	<0.003	<0.005	<0.0005	<0.005	<0.0015	<0.025
2	204.920	532.0	NaN	342	1.8045	120.719	<1.1	1.449803	0.037	<0.003	<0.005	<0.0005	<0.005	<0.0015	<0.025
3	327.000	686.0	NaN	478.6	1.1229	199.879	<1.1	1.258597	0.0154	<0.003	0.005	<0.0005	<0.005	<0.0015	<0.025
4	309.885	1841.0	NaN	1179	0.2343	476.9872	291	15.672251	<0.01	<0.003	<0.005	<0.0005	<0.005	<0.0015	<0.025
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1063	231.045	2350.0	NaN	1545.8	<0.2	752.096	<1.1	14.615488	<0.01	<0.003	<0.005	<0.0005	<0.005	<0.0015	<0.025
1064	256.000	529.0	NaN	297	<0.2	273	<1.1	77.392	<0.01	<0.003	<0.005	<0.0005	<0.005	0.00709	0.07578
1065	330.690	2600.0	NaN	1873	0.7574	660.2126	620	36.477104	<0.01	<0.003	<0.005	<0.0005	<0.005	0.0242	0.2129
1066	193.140	873.0	NaN	690.6667	0.7108	406.368	<1.1	<0.02	<0.01	<0.003	<0.005	<0.0005	<0.005	0.012	0.1786
1067	263.070	817.0	NaN	495	0.4002	362.544	<1.1	0.811876	<0.01	<0.003	<0.005	<0.0005	<0.005	<0.0015	<0.025

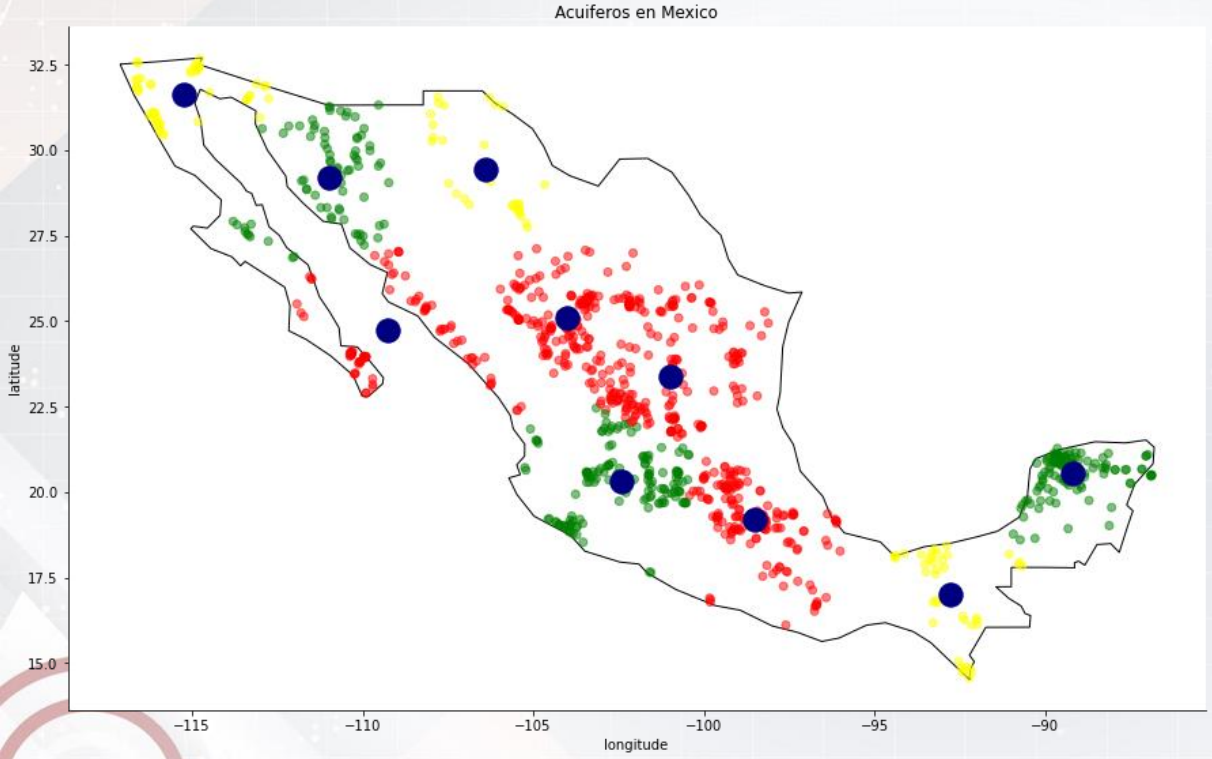
1068 rows × 15 columns

3. Modelo y métricas



Mediante un mapa de calor podemos notar que existe una fuerte correlacion positiva entre las columnas MN_TOT_mg/L (valor de Manganeso Total, en miligramos por litro), FE_TOT_mg/L (Valor de Hierro Total, en miligramos por litro), y HG_TOT_mg/L (Valor de Mercurio Total, en miligramos por litro).

Resultados de agrupamiento de latitudes y longitudes con K means en el mapa de la República Mexicana



Potable - Excelente: 739 acuíferos.

Buena Calidad: 208 acuíferos.

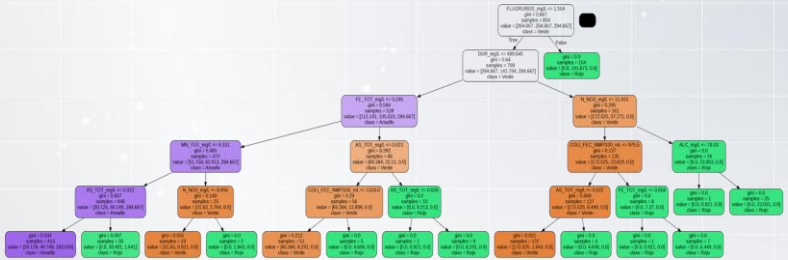
Aceptable: 60 acuíferos.

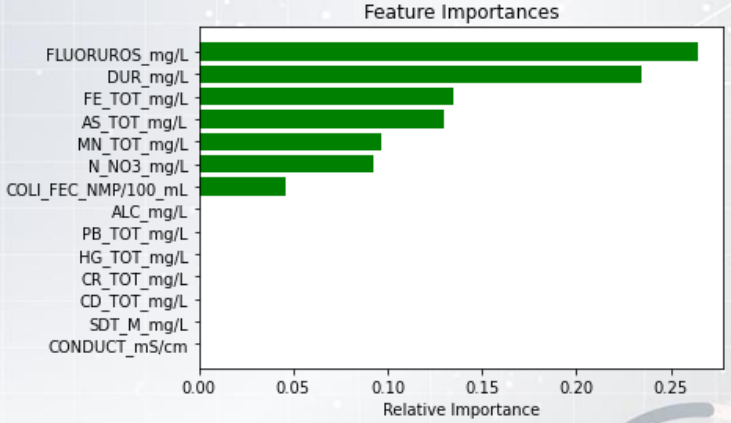
Contaminada: 49 acuíferos.

Fuertemente contaminada: 12 acuíferos.

4. Técnicas utilizadas

Decision Tree

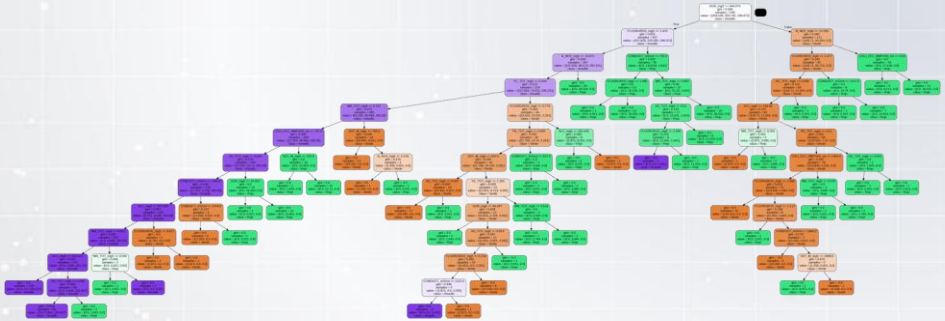


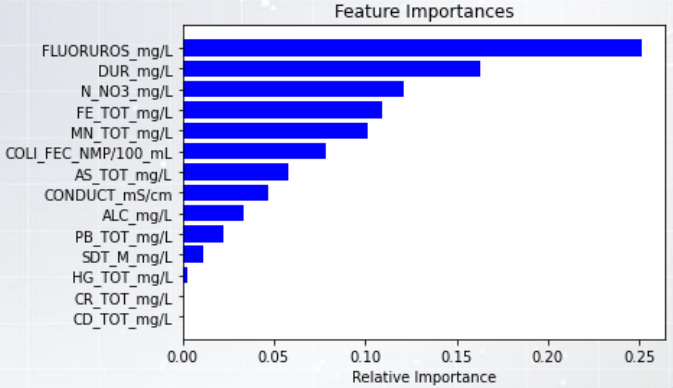


Exactitud: 0.8971962615156478

4. Técnicas utilizadas

Random Forest

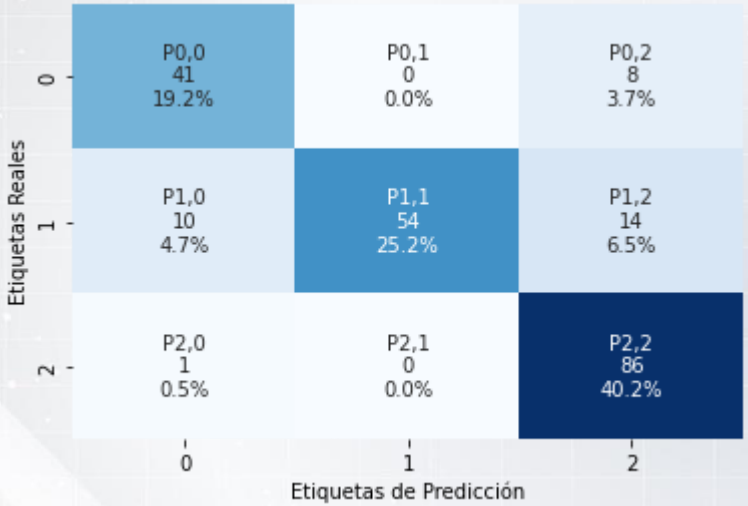




Exactitud: 0.9408099686726584

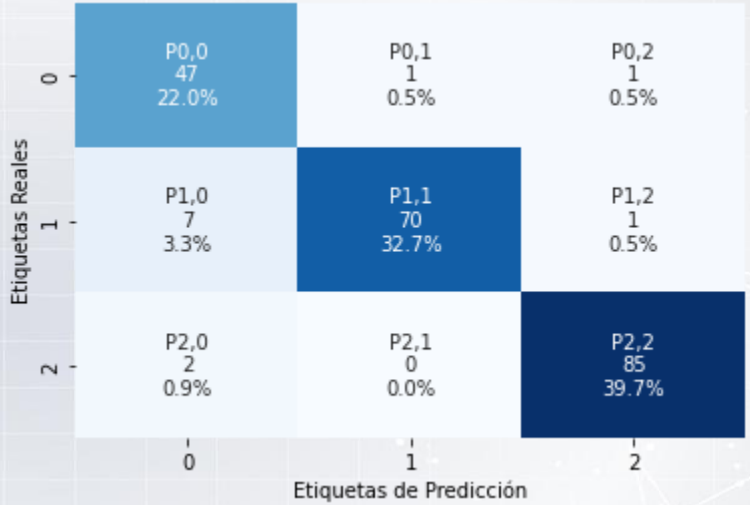
5. Resultados

Decision Tree



Exactitud: 0.8971962615156478

Random Forest



Exactitud: 0.9408099686726584

6. Conclusiones

- Una de las primeras conclusiones es con referencia a la limpieza de datos. Encontrar las columnas que son relevantes para el análisis fue un proceso directo debido a que para este estudio se debían contar con las numéricas. Se utilizó la mediana para imputar los datos faltantes.

- Utilizando K-Means, pudimos encontrar una relación entre la calidad del agua y su ubicación dentro de la República Mexicana. La zona Norte y zona Sur presentan una calidad de agua excelente y aceptable. Mientras que en ciertas ubicaciones de la zona Centro, la calidad del agua no es tan buena.

- Para los dos técnicas DT y RF se encontró que las variables con más importancia eran "Fluoruros" y "DUR" (mg/l) siendo las que mayor impacto generaban el modelo, debido a la complejidad de RF esta empezó a encontrar mayores relaciones y fue esta la que mostró mayor exactitud para predecir con un porcentaje de 94% siendo esta mayor por 5% sobre DT