Reto Ciencia de Datos

Equipo 55

Edgar López Valdes A01339939

Alejandro Juarez Corona A01168444

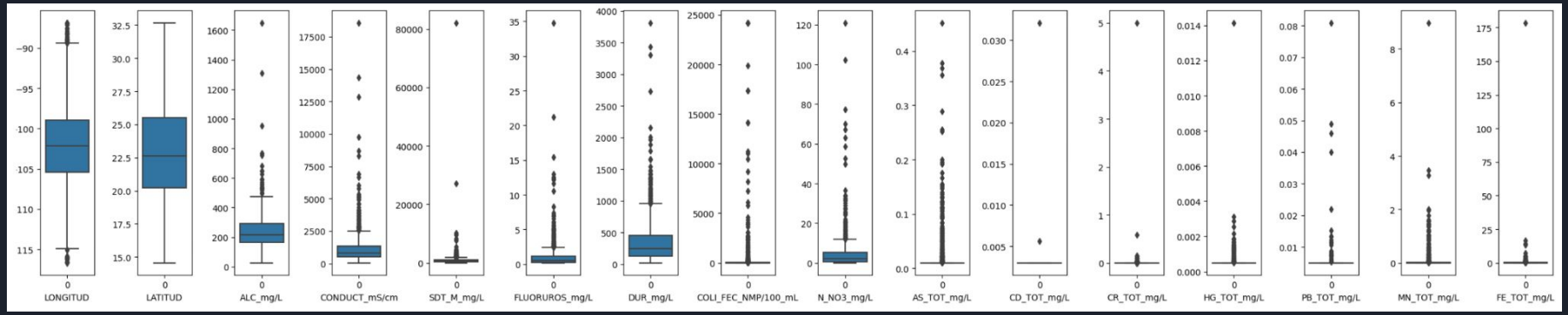
Datos

Aguas subterráneas, consideraciones sobre los datos:

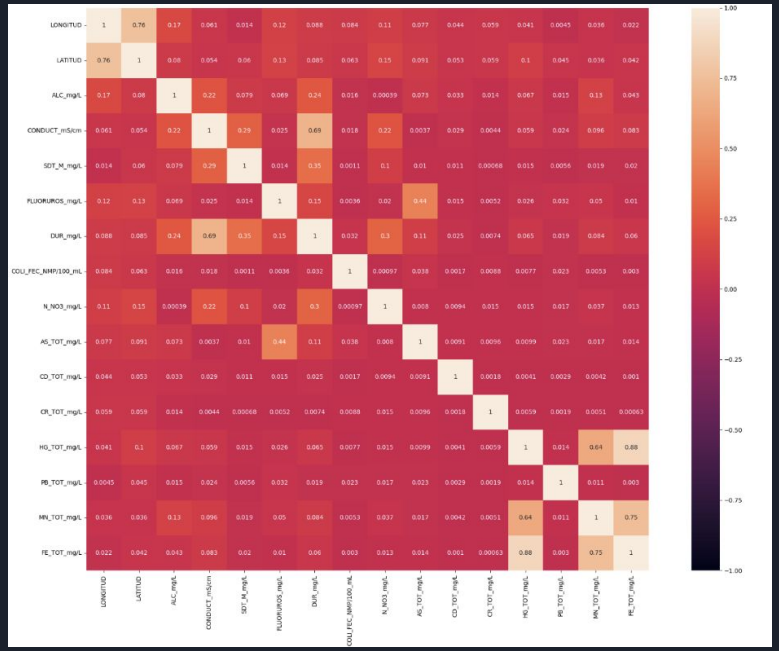
- 1068 registros
- 57 columnas
- Algunas dan información respecto a la ubicación, mientras que otras indican concentraciones.
- Datos numéricos y categóricos
- Variable de salida semaforo (verde, amarillo, rojo)

Limpieza

Se encontró que dos columnas son descartables y otras fueron convertidas de *object* a *float.* Se optó por reemplazar los valores numéricos por la media de la columna para los registros con nulos.



Análisis de variables (correlación) etc



A pesar de que algunos valores mostraban una alta correlación como la presencia de hierro y la presencia de mercurio, se determinó no reducir el número de componentes ya que en su mayoría las columnas no presentaban correlaciones significativas.

K means

- Prueba del codo - 4 clusters
- Centros en Campeche, Durango, Sonora y Querétaro

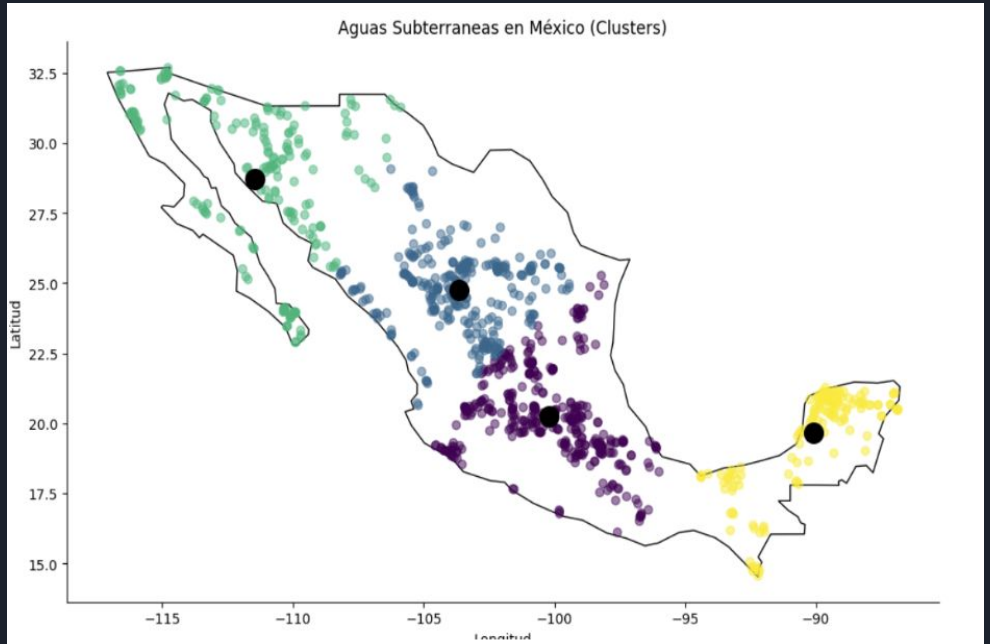
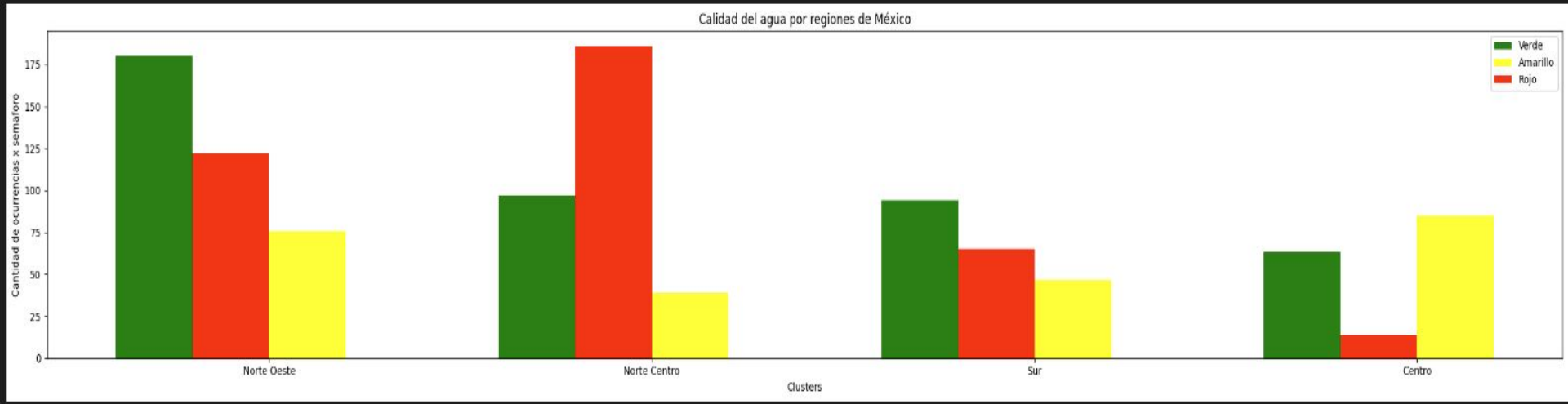


Fig 4. Clusters sobre el mapa de México

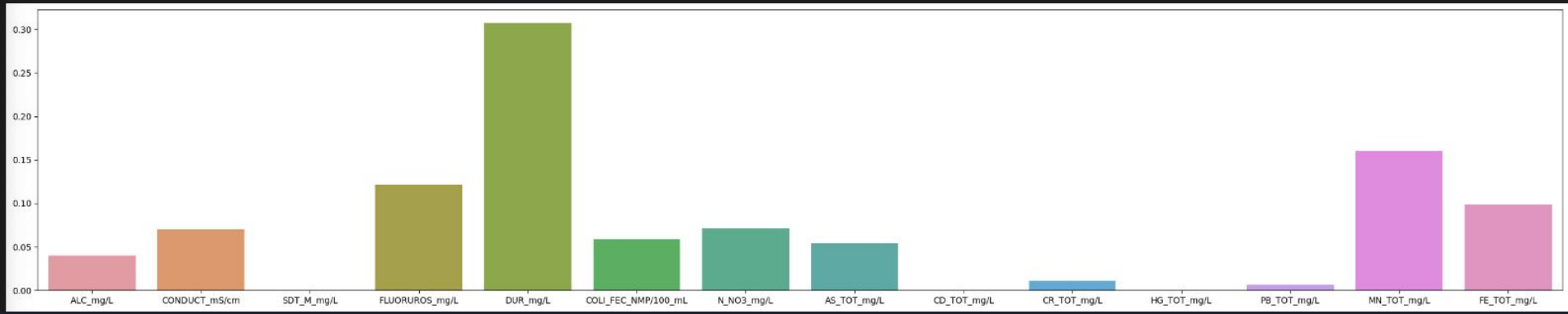


Determinamos que sí existe una relación entre la ubicación y la calidad del agua, podemos observar que en los clusters más al norte existen comportamientos similares, aunque en la zona centro norte la calidad del agua es muy mala en algunos puntos, podemos pensar que esto está relacionado a la grandes ciudades industriales, mismo caso con la zona centro. La zona sur (cabe mencionar que es la que presenta mayor vegetación del país) tiene los valores más altos de calidad buena e intermedia, mientras que presenta pocos puntos con mala calidad.

Árbol de decisión

```
X_train,X_test,y_train,y_test = train_test_split(dfX,
dfOutEncoded, test_size=0.30, random_state=7)
arbol = DecisionTreeRegressor(max_depth=100, max_leaf_nodes=20
,random_state=7)
```

Se realiza grid search para determinar los mejores hiperparametros. Finalmente se obtiene la gráfica de *feature importance*.



Random Forest

Mismo caso, se utilizan tanto los valores de entrenamiento como de validación para el modelo Random Forest tras haber corrido el grid search nuevamente. Se generan las matrices de confusión para ambos modelos.

RF

```
[[ 73   0   1]
 [  3 104   1]
 [  3   1 135]]
```

DT

```
[[ 72   2   0]
 [ 11  97   0]
 [  3   0 136]]
```

Comparativa RF vs DT

Resultados Random Forest

	precision	recall	f1-score	support
0	0.92	0.99	0.95	74
1	0.99	0.96	0.98	108
2	0.99	0.97	0.98	139
accuracy			0.97	321
macro avg	0.97	0.97	0.97	321
weighted avg	0.97	0.97	0.97	321

Resultados Arbol de Decision

	precision	recall	f1-score	support
0	0.84	0.97	0.90	74
1	0.98	0.90	0.94	108
2	1.00	0.98	0.99	139
accuracy			0.95	321
macro avg	0.94	0.95	0.94	321
weighted avg	0.96	0.95	0.95	321

Conclusiones

- Comparación de modelos.
- Importancia de la visualización
- Trabajo futuro