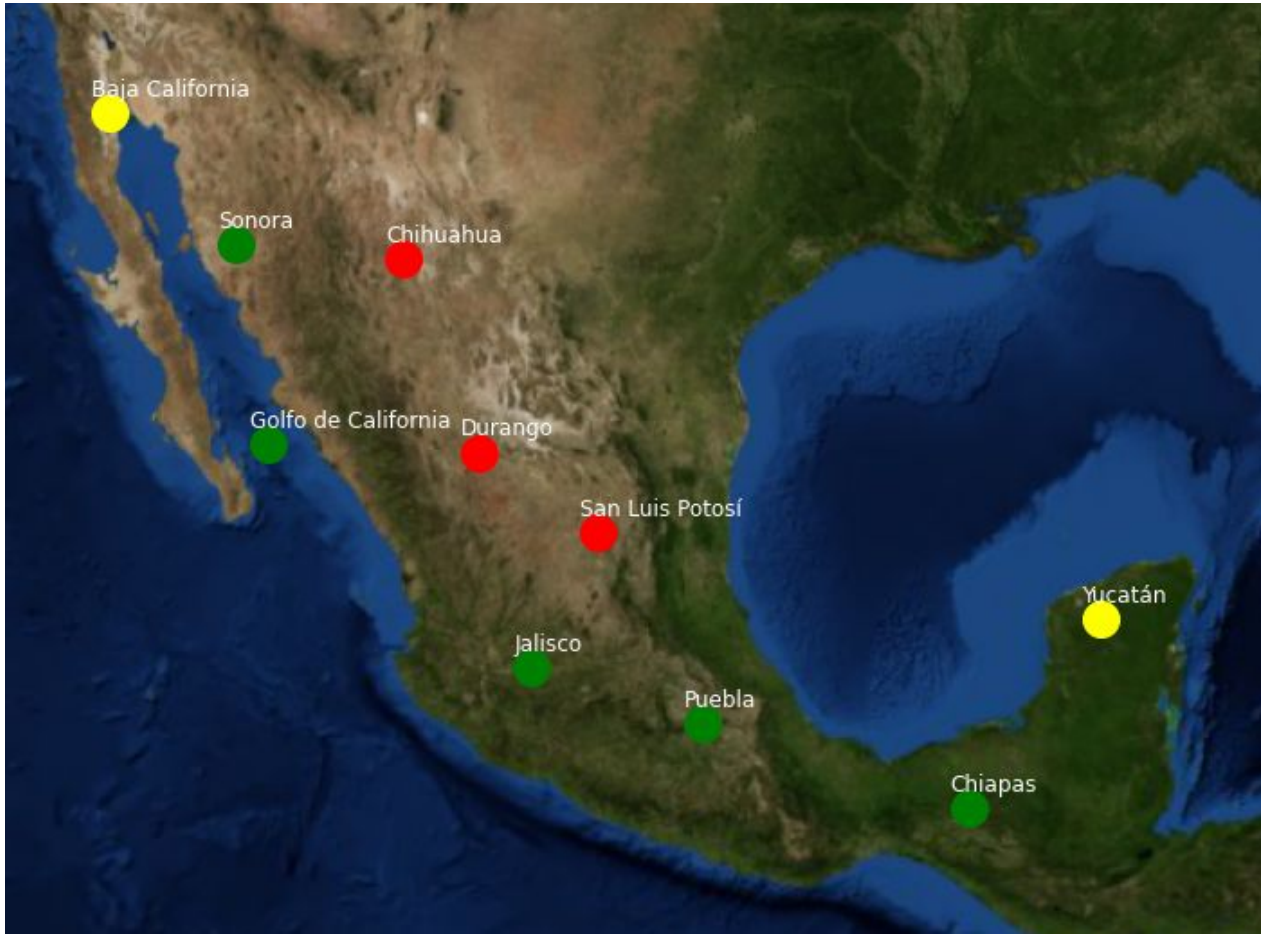
Tecnológico de Monterrey



Clasificador de Calidad en Aguas Subterráneas

Ernesto Enríquez Rubio: A01228409

Jonathan Garza Bennet: A01793038

Ciencia y Analítica de Datos

Pipeline

- Se encontraron variables tanto numéricas como categóricas.
- Aunque los puntos de muestra se encuentran en diferentes ubicaciones, varios pertenecen al mismo acuífero.
- Para sustituir los valores faltantes se utilizó la media del acuífero correspondiente en los valores numéricos y la moda para los categóricos.
- Se eliminó la columna SDT_mg/L que no contenía ningún valor.
- En los parámetros numéricos, se eliminó el carácter (<) en las que contenían un rango para poder realizar operaciones.
- En la variable categórica de interés, se aplicó un codificador de etiquetas; transformando la categoría en un entero.

Ejemplo:

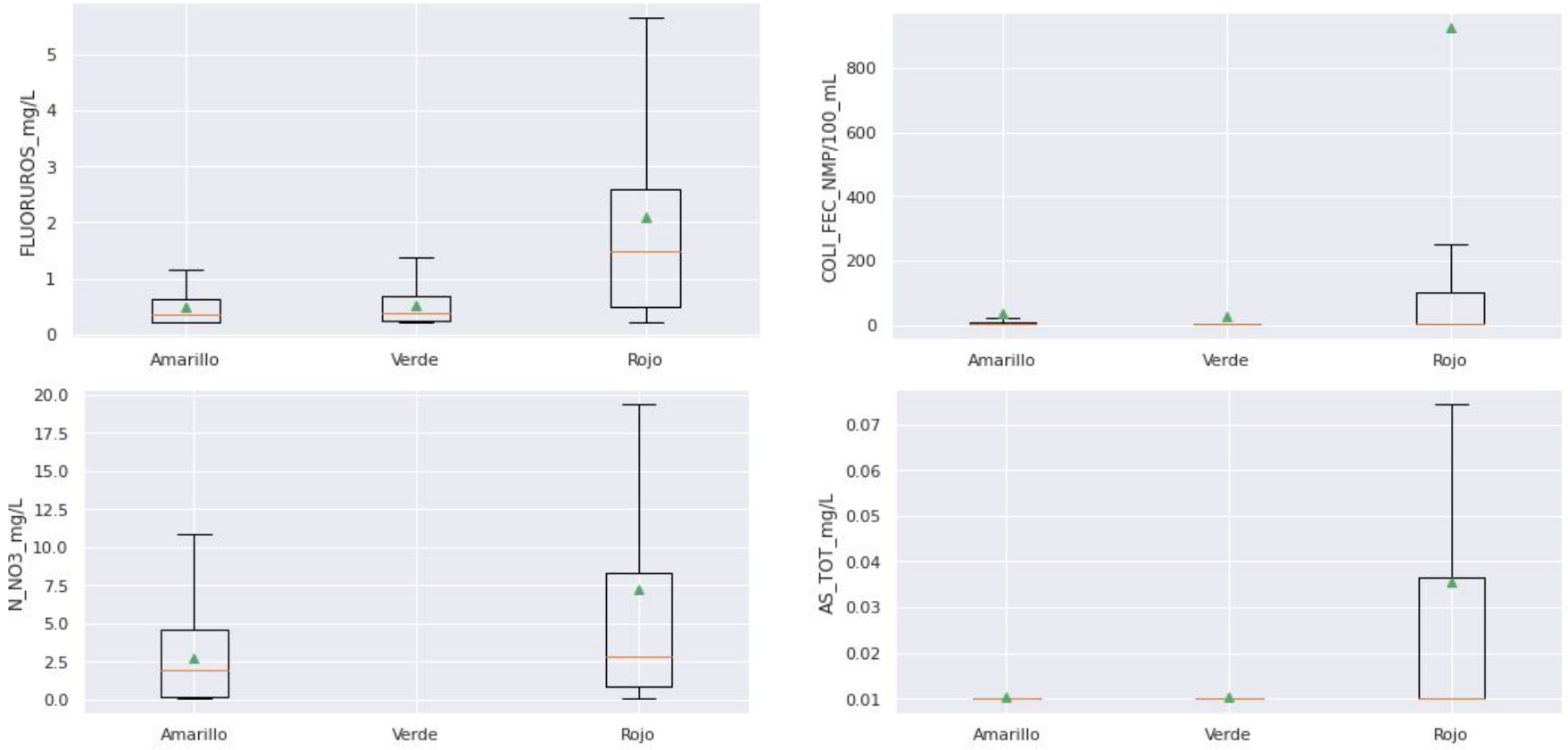
```
df_clean = fill_missing_values_with_median(df_clean, 'ACUIFERO', 'N_NO3_mg/L')

Existen valores faltantes en:

234    VALLE DE CANATLAN
Name: ACUIFERO, dtype: object

La mediana de VALLE DE CANATLAN es: 0.729592

ACUIFERO: VALLE DE CANATLAN,  Índice: 234

Valor anterior: nan

Nuevo valor: 0.729592

La cantidad de datos faltantes en la columna N_NO3_mg/L es: 0
```

Variable	Datos Faltantes	Tipo
ALC_mg/L	4	Numérico
CALIDAD_ALC	4	Categórico
CONDUCT_mS/cm	6	Numérico
CALIDAD_CONDUC	6	Categórico
SDT_mg/L	1068	?
SDT_M_mg/L	2	Numérico
CALIDAD_SDT_ra	2	Categórico
CALIDAD_SDT_salin	2	Categórico
DUR_mg/L	1	Numérico
CALIDAD_DUR	1	Categórico
N_NO3_mg/L	1	Numérico
CALIDAD_N_NO3	1	Categórico

En total se sustituyeron valores 30 faltantes, por lo que no se perdió ningún renglón del conjunto de datos.

Análisis Visual de Datos

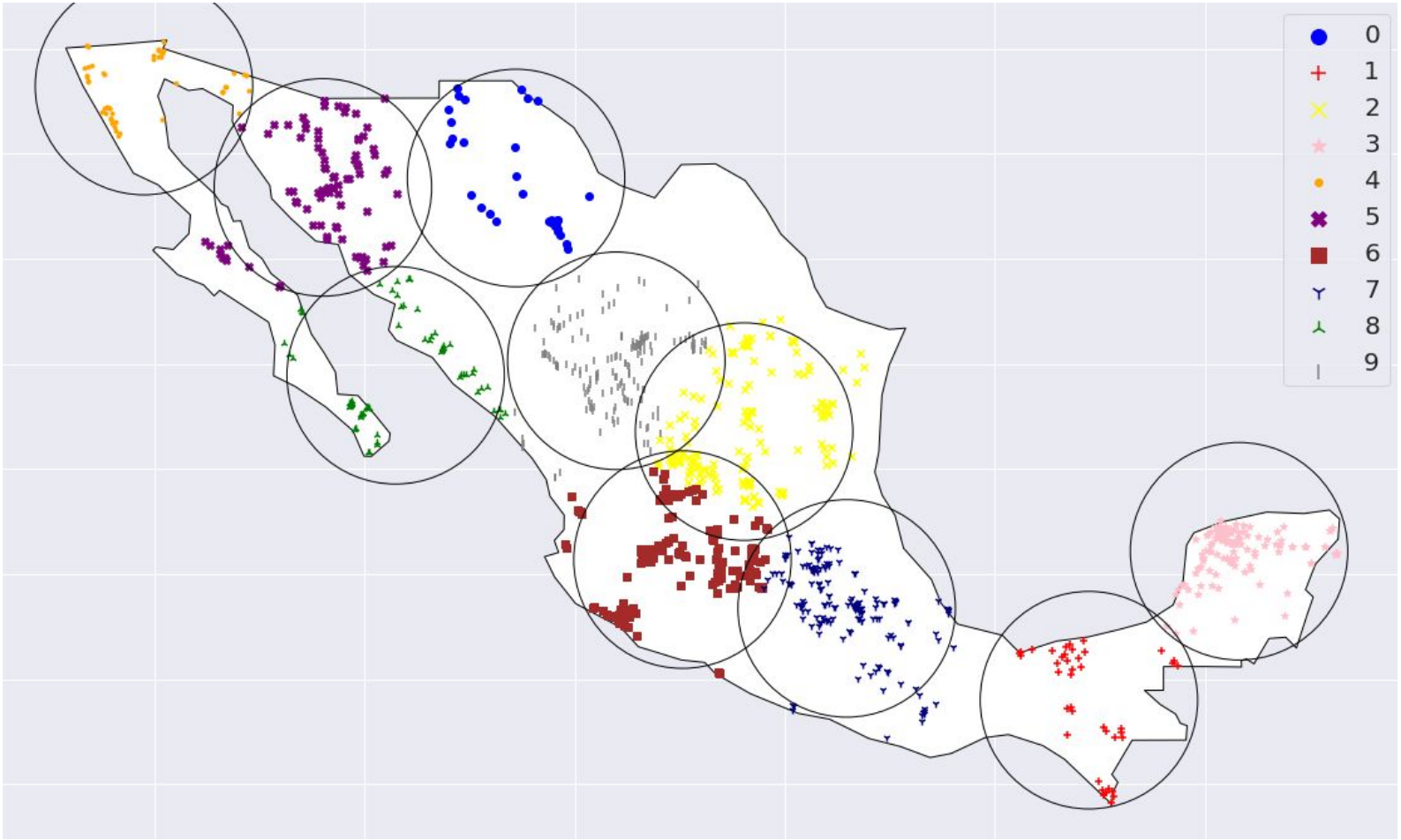
Se realizaron gráficos de caja y bigotes, se encontraron 4 variables con tendencia claramente ascendente con relación al su respectivo color de semáforo:

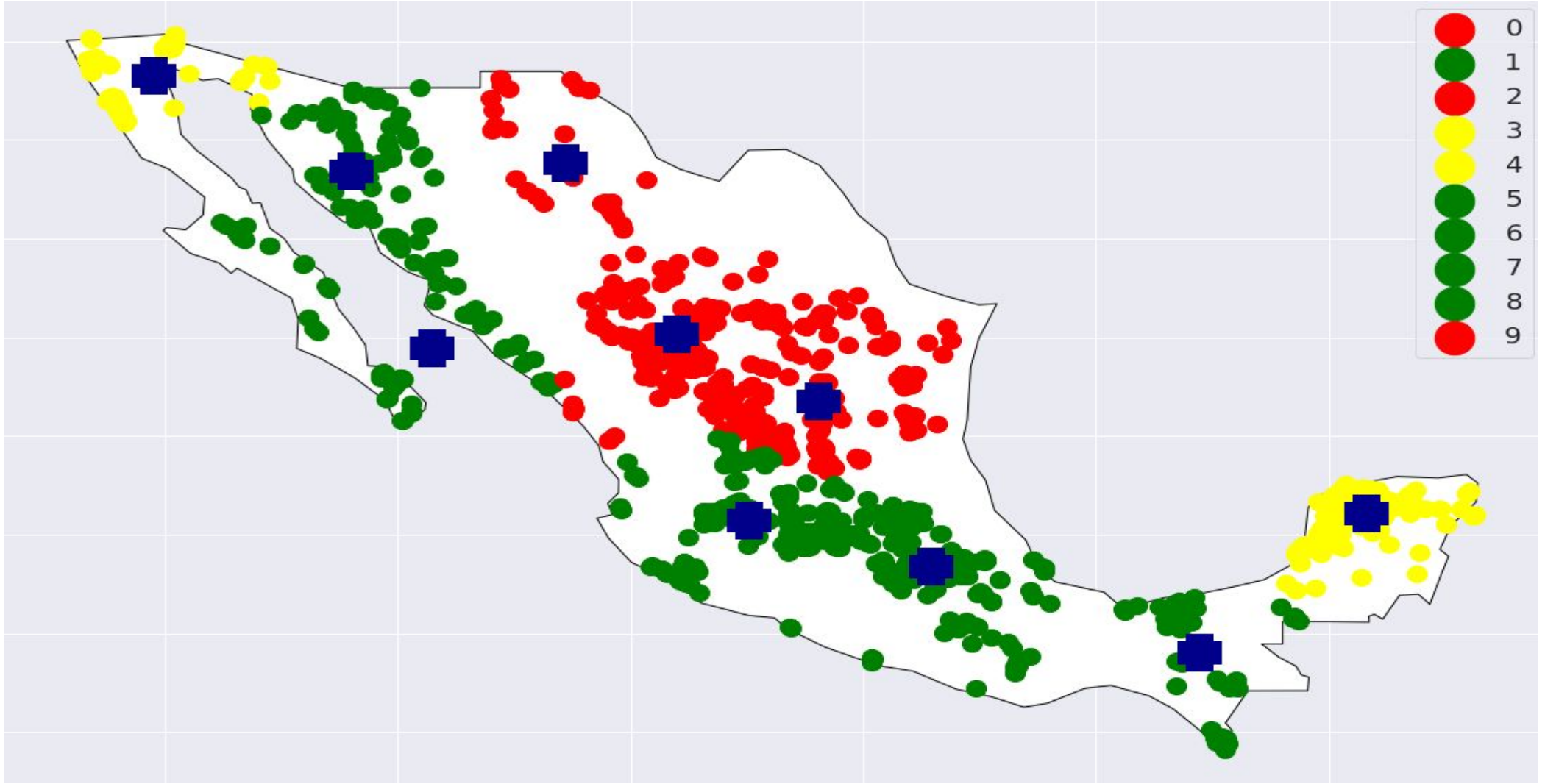


El resultado del análisis visual nos muestra una tendencia inicial de variables de importancia sin embargo, se realizará un análisis importancia de cada variable de forma individual

Análisis de Agrupación Geográfica (Kmeans)

En base a los valores de la inercia y el puntaje de Silhoette, se determinó que el número de agrupamientos para cubrir la mayor parte de las zonas con aguas subterráneas y poder llegar al objetivo de encontrar una relación entre la zona geográfica y la calidad es de 10. A continuación se muestran las diferentes regiones en la que se analizaron la calidad del agua.





Analizando la dominancia de la calidad del agua en cada región se generó un mapa para visualizar las zonas geográficas con mayor tendencia a una calidad mala, buena o regular.

Modelos de Clasificación

- Las variables contienen tanto datos numéricos como categóricos para determinar la calidad de las aguas subterráneas por medio de un semáforo (verde, amarillo y rojo), clasificando el agua como buena, regular y mala.
- Los datos categóricos para cada parámetro toman su valor de las variables numéricas.
- Se estudia el desempeño de dos clasificadores el primero utilizando las variables categóricas resultantes de los parámetros numéricos y el segundo utilizando directamente los valores numéricos de los parámetros.

Variables Categóricas

- **CUMPLE_CON_ALC**
- **CUMPLE_CON_COND**
- **CUMPLE_CON_SDT_ra**
- **CUMPLE_CON_SDT_salin**
- **CUMPLE_CON_FLUO**
- **CUMPLE_CON_DUR**
- **CUMPLE_CON_CF**
- **CUMPLE_CON_NO3**
- **CUMPLE_CON_AS**
- **CUMPLE_CON_CD**
- **CUMPLE_CON_CR**
- **CUMPLE_CON_HG**
- **CUMPLE_CON_PB**
- **CUMPLE_CON_MN**
- **CUMPLE_CON_FE**
- **SEMAFORO**

Variables Numéricas

- **ALC_mg/L**
- **CONDUCT_mS/cm**
- **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**

Árbol de decisiones VS Bosques Aleatorios

- Los híper parámetros empleados fueron los mismos entre ambos métodos: max_Depth = 15 y random_state=42
- Para ambos modelos (categórico y numérico), se realizó el entrenamiento y validación cruzada por ambos métodos, obteniendo los siguientes resultados.

Modelo Categórico

Prueba

DT:
mean Precision: 0.989 (0.0062)
mean Recall: 0.989 (0.0062)

Prueba

RF:
mean Precision: 0.988 (0.0058)
mean Recall: 0.988 (0.0058)

Entrenamiento

DT:
mean Precision: 1.000 (0.0000)
mean Recall: 1.000 (0.0000)

Entrenamiento

RF:
mean Precision: 1.000 (0.0000)
mean Recall: 1.000 (0.0000)

Modelo Numérico

Prueba

DT:
mean Precision: 0.966 (0.0131)
mean Recall: 0.966 (0.0131)

Prueba

RF:
mean Precision: 0.958 (0.0153)
mean Recall: 0.958 (0.0153)

Entrenamiento

DT:
mean Precision: 1.000 (0.0000)
mean Recall: 1.000 (0.0000)

Entrenamiento

RF:
mean Precision: 1.000 (0.0000)
mean Recall: 1.000 (0.0000)

Los resultados de ambos clasificadores muestran valores muy cercanos tanto en el conjunto de entrenamiento como en el de prueba en ambas métricas de desempeño. El mejor resultado lo obtiene el modelo de árbol de decisiones ligeramente.

Feature Importance

- Se obtuvo el nivel de importancia de las variables de ambos modelos para conocer qué tanto influyen al resultado final.

Modelo Categórico

- 21.1701 % - CUMPLE_CON_ALC
- 17.0557 % - CUMPLE_CON_COND
- 14.4559 % - CUMPLE_CON_SDT_ra
- 10.9388 % - CUMPLE_CON_SDT_salin
- 9.6629 % - CUMPLE_CON_FLUO
- 8.8955 % - CUMPLE_CON_DUR
- 5.1742 % - CUMPLE_CON_CF
- 3.4293 % - CUMPLE_CON_NO3
- 3.4194 % - CUMPLE_CON_AS
- 2.6194 % - CUMPLE_CON_CD
- 2.5203 % - CUMPLE_CON_CR
- 0.3319 % - CUMPLE_CON_HG
- 0.3267 % - CUMPLE_CON_PB
- 0.0 % - CUMPLE_CON_MN
- 0.0 % - CUMPLE_CON_FE

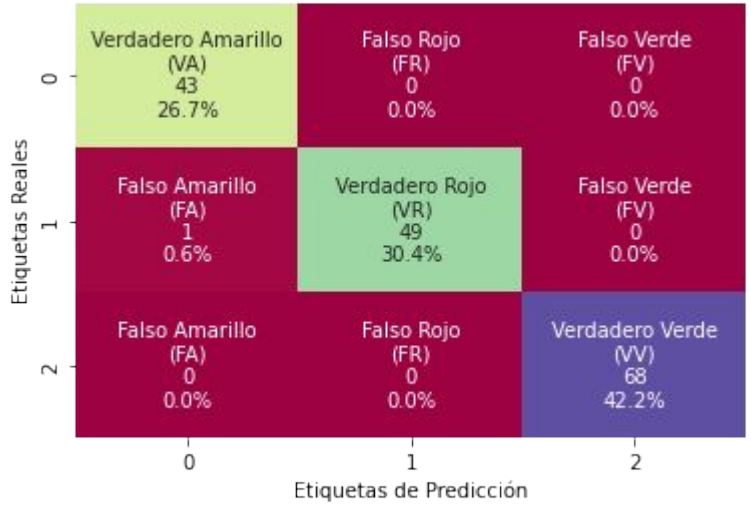
Modelo Numérico

21.5668 % - ALC_mg/L
16.7261 % - CONDUCT_mS/cm
14.1462 % - SDT_M_mg/L
11.4685 % - FLUORUROS_mg/L
9.5596 % - DUR_mg/L
8.8955 % - COLI_FEC_NMP/100_mL
5.1742 % - N_NO3_mg/L
3.1387% - AS_TOT_mg/L
2.7898 % - CD_TOT_mg/L
2.5354 % - CR_TOT_mg/L
2.2473 % - HG_TOT_mg/L
1.0981 % - PB_TOT_mg/L
0.3318 % - MN_TOT_mg/L
0.3221 % - FE_TOT_mg/L

En el modelo categórico, dos variables no son muy relevantes para el modelo de clasificación, mientras que para el numérico, todas aportan de cierta forma aunque los valores de las últimas dos también son muy pequeños. Se corrobora que los parámetros numéricos identificados de forma visual, forman parte de los de mayor importancia.

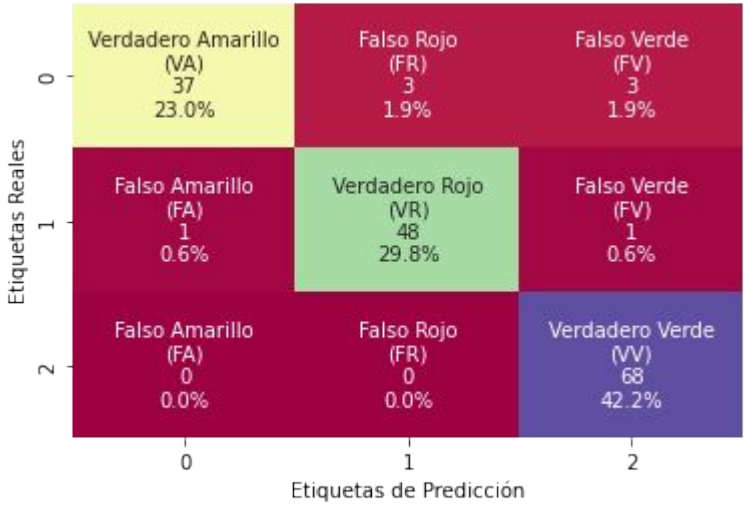
Matriz de Confusión y Reporte de Clasificación

Modelo Categórico



	precision	recall	f1-score	support
Amarillo 0	0.97	0.86	0.91	43
Rojo 1	0.94	0.96	0.95	50
Verde 2	0.94	1.00	0.97	68
accuracy			0.95	161
macro avg	0.95	0.94	0.95	161
weighted avg	0.95	0.95	0.95	161

Modelo Numérico



	precision	recall	f1-score	support
Amarillo 0	0.98	1.00	0.99	43
Rojo 1	1.00	0.98	0.99	50
Verde 2	1.00	1.00	1.00	68
accuracy			0.99	161
macro avg	0.99	0.99	0.99	161
weighted avg	0.99	0.99	0.99	161

Ambos modelos resultan en métricas de desempeño muy similares, alrededor o por encima del 90%

Conclusiones

Los resultados obtenidos con ambos clasificadores tanto numérico como categórico, mostraron resultados ligeramente superiores para el modelo de árboles de decisión, por lo que se determinó como mejor modelo comparado con el de bosques aleatorios.

Posteriormente, se estudiaron los resultados de los modelos de clasificación con variables predictoras categóricas y numéricas. Los resultados de precisión y recall para ambos fueron bastante similares, mostrando un desempeño de clasificación bastante elevado, por encima de 95% de recall y 100% de precisión en ambos. Los resultados se ven reflejados en las matrices de confusión correspondientes, los resultados fuera de la diagonal principal son prácticamente nulos.

Ambas opciones estudiadas en este ejercicio resultan en clasificadores robustos para predecir la calidad del agua de forma bastante precisa.