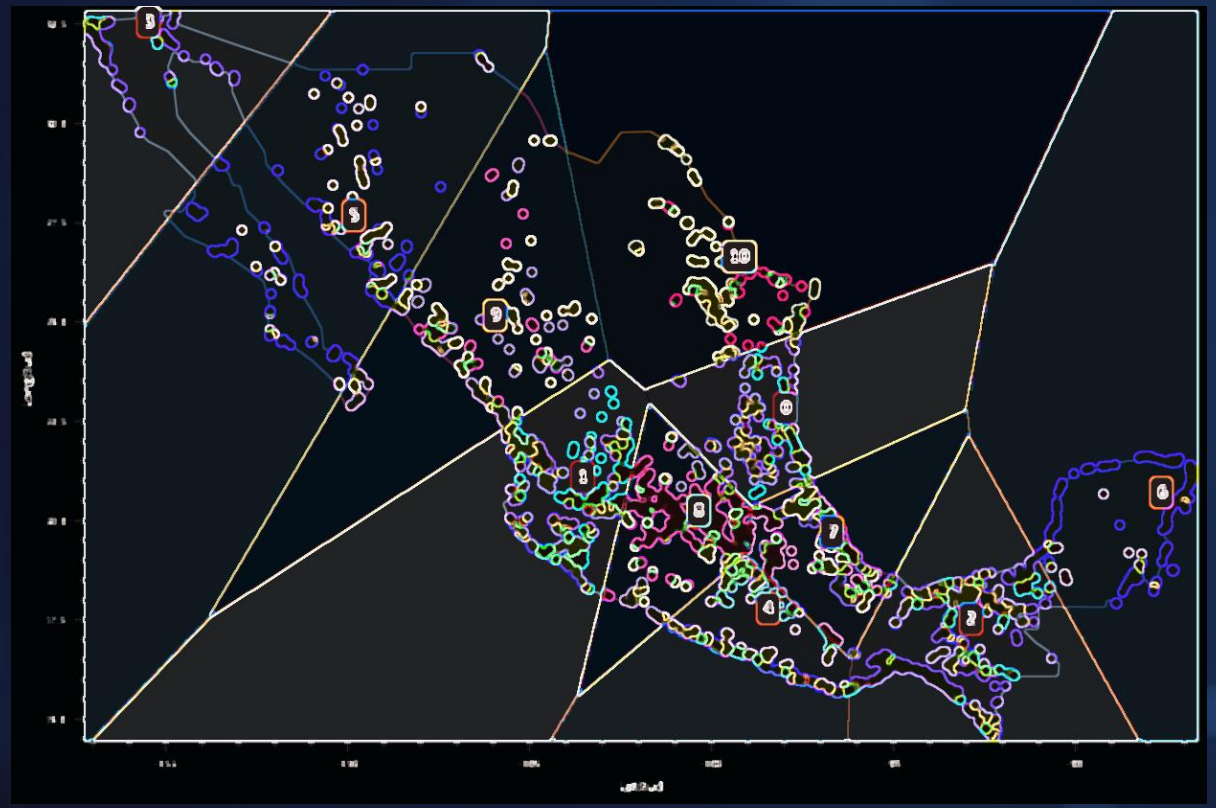
Ciencia y Analítica de Datos

Proyecto Final – Calidad del Agua



Alumnos:

- Armando Bringas Corpus (A01200230)
- Walter André Hauri Rosales (A01794237)

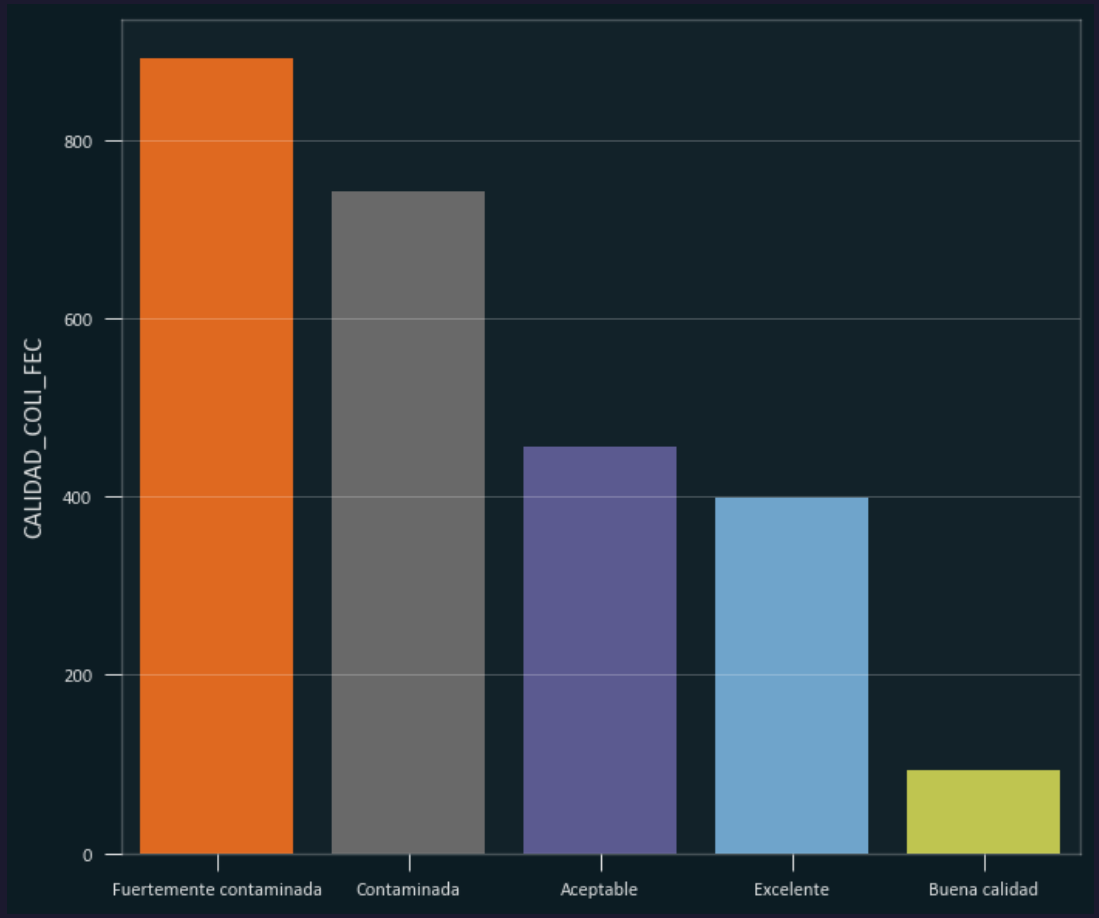
Profesores:

- Dra. María de la Paz Rico Fernández
- Mtra. Victoria Guerrero Orozco

Limpieza de Datos

Base de Datos

- Aguas Superficiales
- 227,755 registros totales
- 55 columnas (atributos)
- 648 registros con datos faltantes
- **178,173 registros finales después de limpieza**

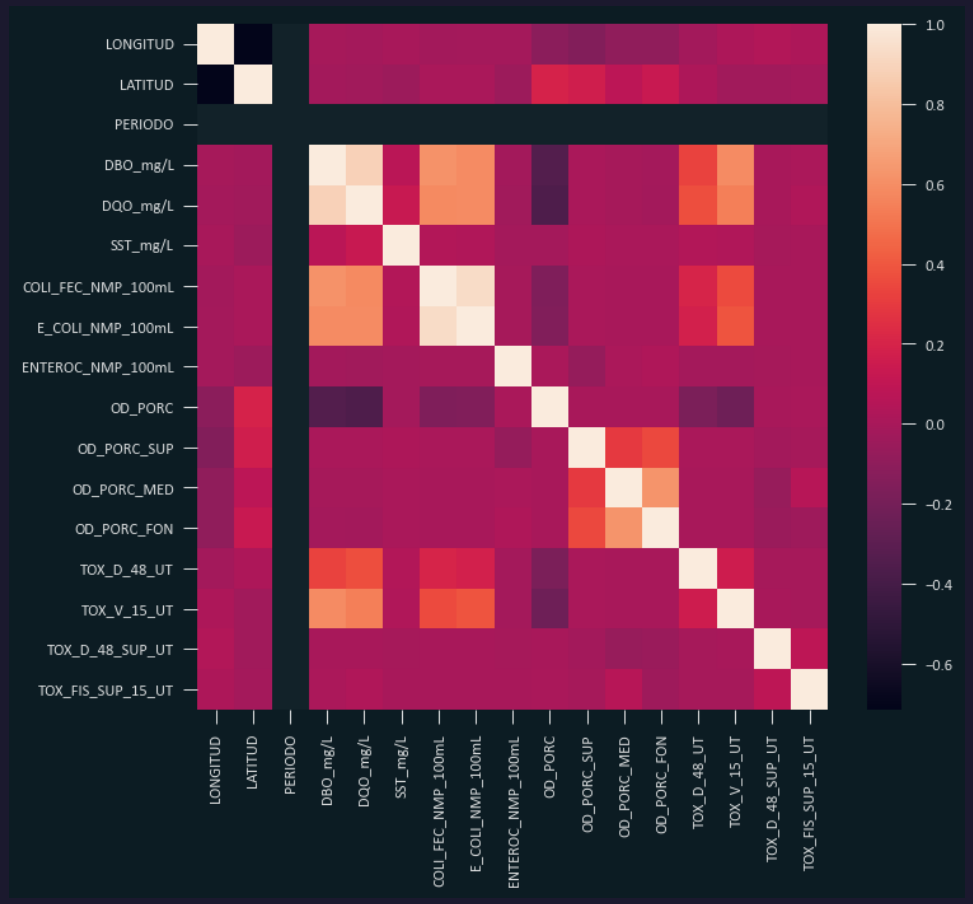


Análisis

Estadística Descriptiva

	LONGITUD	LATITUD	SEMAFORO
count	3493.000000	3493.000000	3493.00000
mean	-100.359969	21.046992	1.03779
std	6.122773	3.893696	0.82851
min	-117.124030	14.534910	0.00000
25%	-103.882310	18.396070	0.00000
50%	-99.795530	20.148980	1.00000
75%	-96.860230	22.828930	2.00000
max	-86.732150	32.706500	2.00000

Matriz de Correlación



One-hot encoding

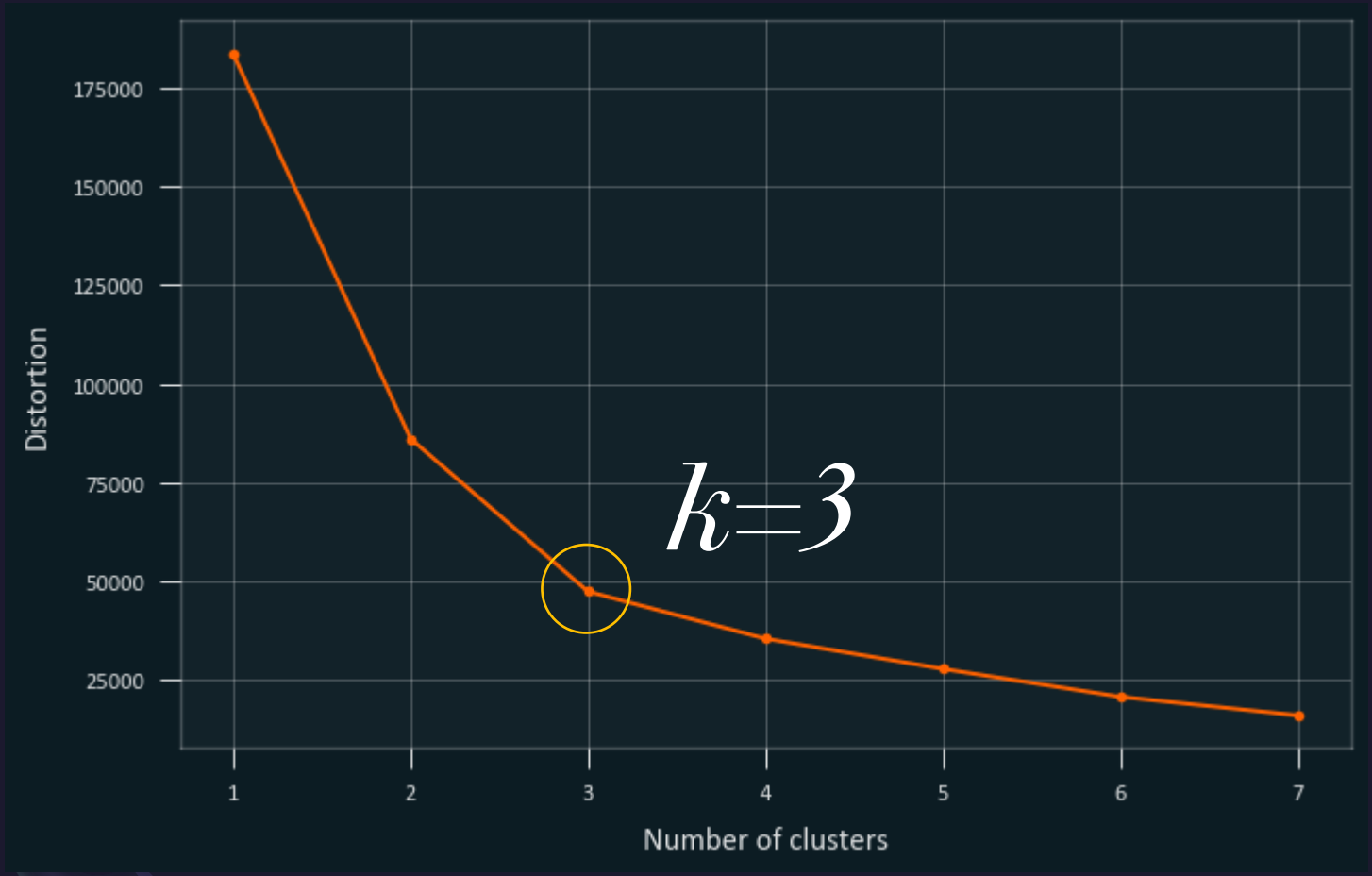
Variable Y

Semáforo

Codificación

- Amarillo: 0
- Rojo: 1
- Verde: 2

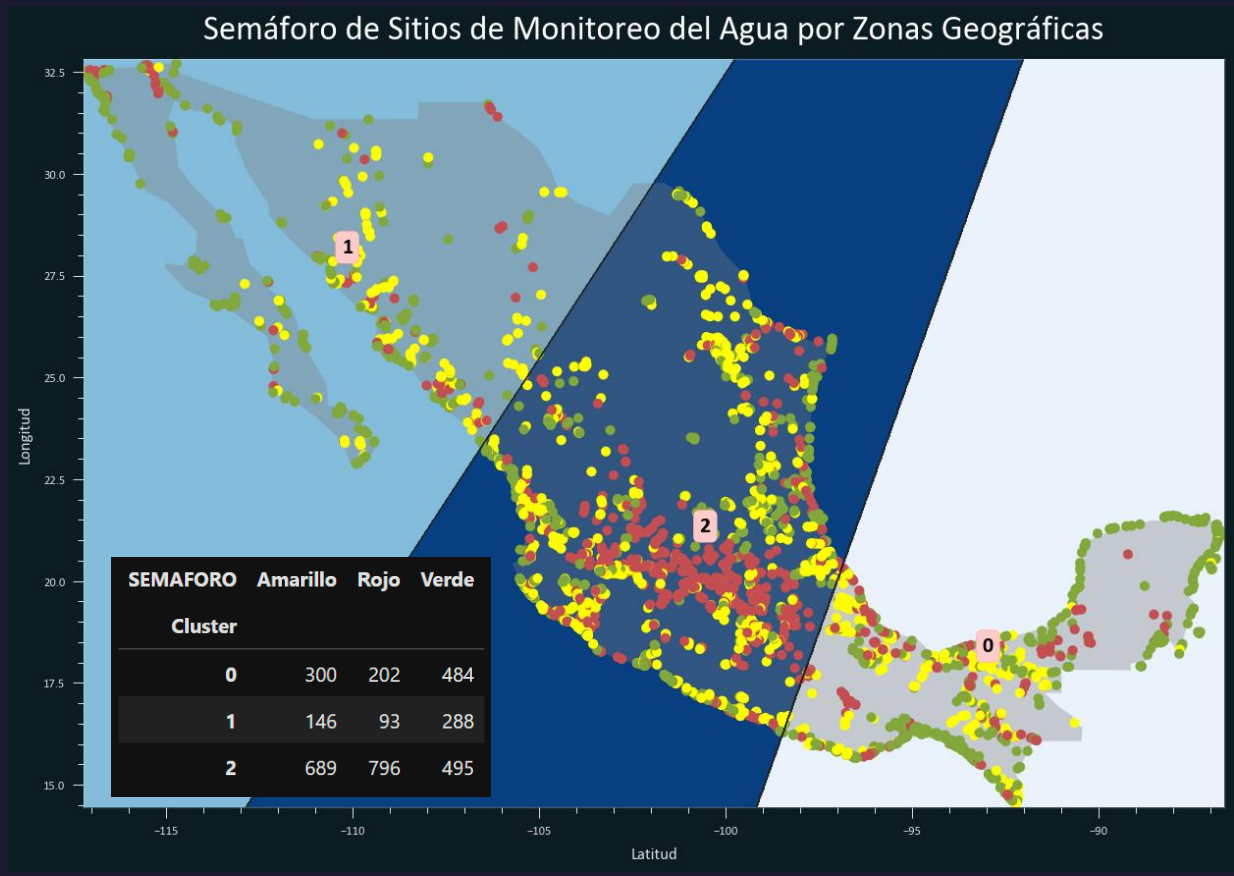
k - means



Se empleo el método de Curva del Codo ('Elbow Curve Method') para determinar el número de agrupaciones (*k*: clusters) recomendable para las longitudes y latitudes.

k=3

Las tres agrupaciones obtenidas nos permite tener una visualización y clasificación por región geográfica:
(1) – Sur, (2) – Norte, (3) – Centro.

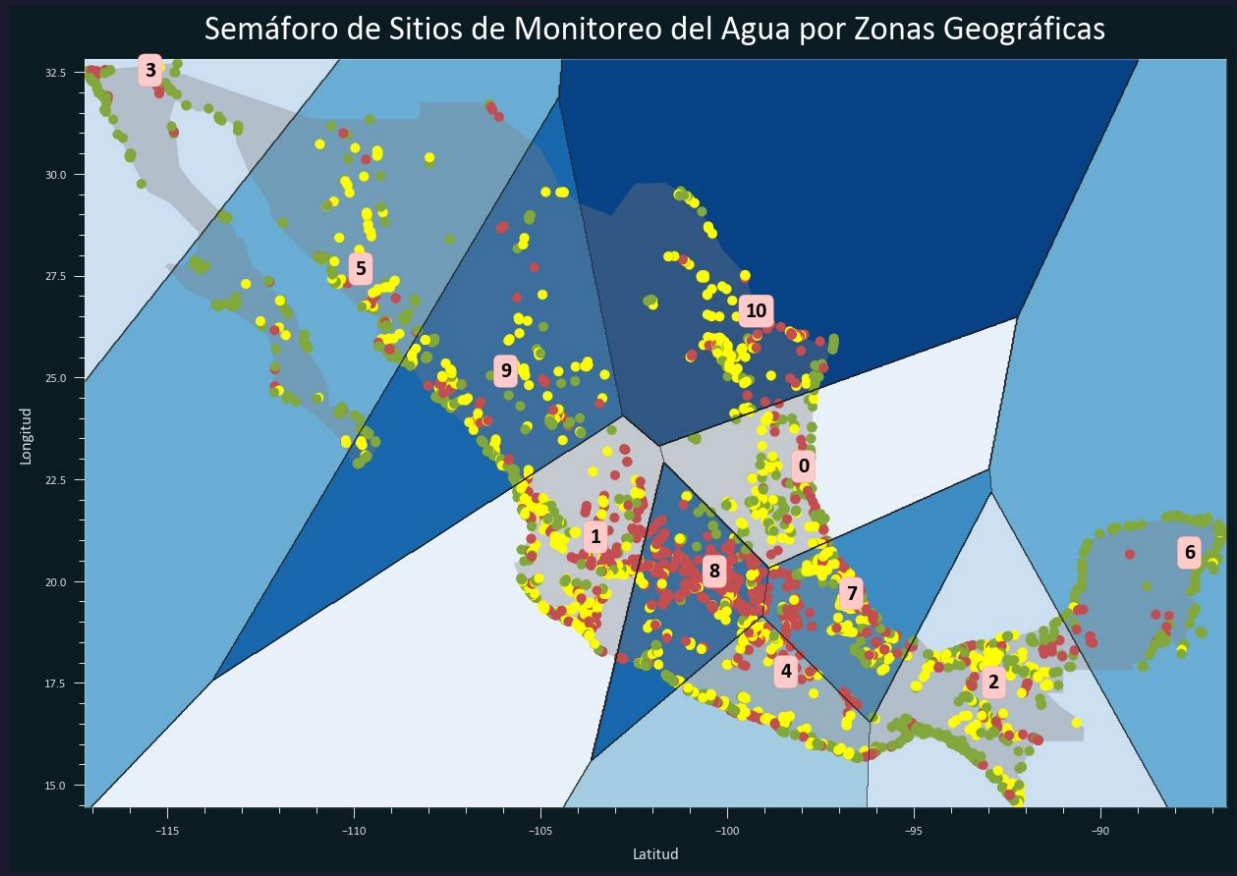


SEMAFORO	Amarillo	Rojo	Verde
Cluster			
0	300	202	484
1	146	93	288
2	689	796	495

Se puede observar que el segundo agrupamiento (Región Centro) es el que tiene mayores zonas de agua contaminada (semáforo rojo)

Aumentando el valor de *k=11*

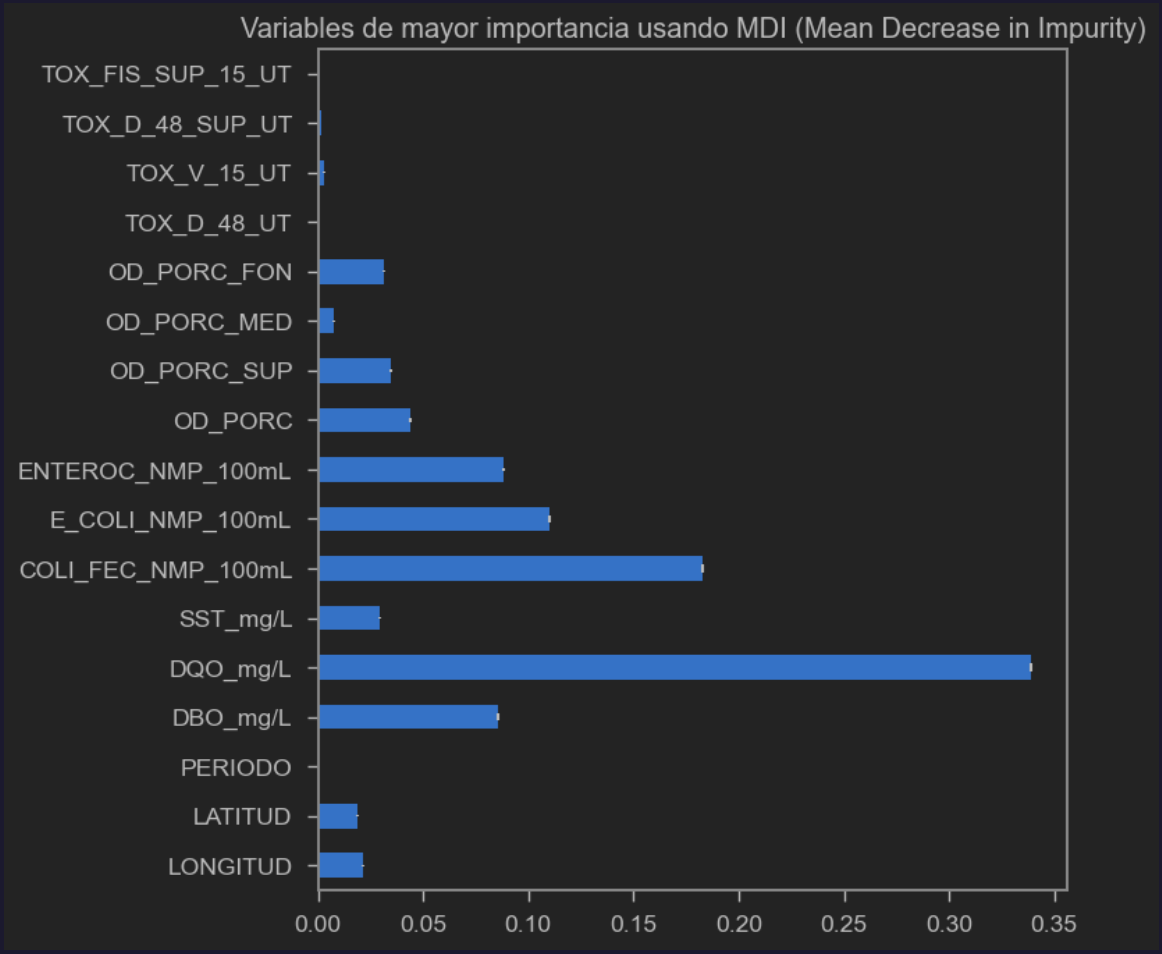
Se decidió aumentar el número de agrupaciones para obtener una mejor visualización de los semáforos.



De igual manera se puede corroborar la región centro es el que tiene mayores zonas de agua contaminada (semáforo rojo)

Clasificación

Selección de variables de mayor importancia



Se seleccionaron las variables más importantes por medio del método de *Random Forest – Mean Decrease in Impurity (MDI).*

*Variables **X** seleccionadas:*

'LONGITUD', 'LATITUD', 'DBO_mg/L', 'DQO_mg/L', 'SST_mg/L', 'COLI_FEC_NMP_100mL', 'E_COLI_NMP_100mL', 'ENTEROC_NMP_100mL', 'OD_PORC', 'OD_PORC_SUP', 'OD_PORC_FON', 'TOX_V_15_UT','TOX_D_48_SUP_UT'

Implementación de los Modelos Random Forest y Decision Tree

Random Forest

```
Mejor valor de accuracy obtenido con la mejor combinación: 0.9865192793924977
Mejor combinación de valores encontrados de los hiperparámetros: {'ccp_alpha': 0.0, 'criterion': 'gin
i', 'max_depth': 19, 'min_samples_split': 2}
Métrica utilizada: accuracy
```

	precision	recall	f1-score	support
Amarillo	0.99	0.99	0.99	227
Rojo	1.00	1.00	1.00	218
Verde	0.98	0.99	0.99	254
accuracy			0.99	699
macro avg	0.99	0.99	0.99	699
weighted avg	0.99	0.99	0.99	699

Decision Tree

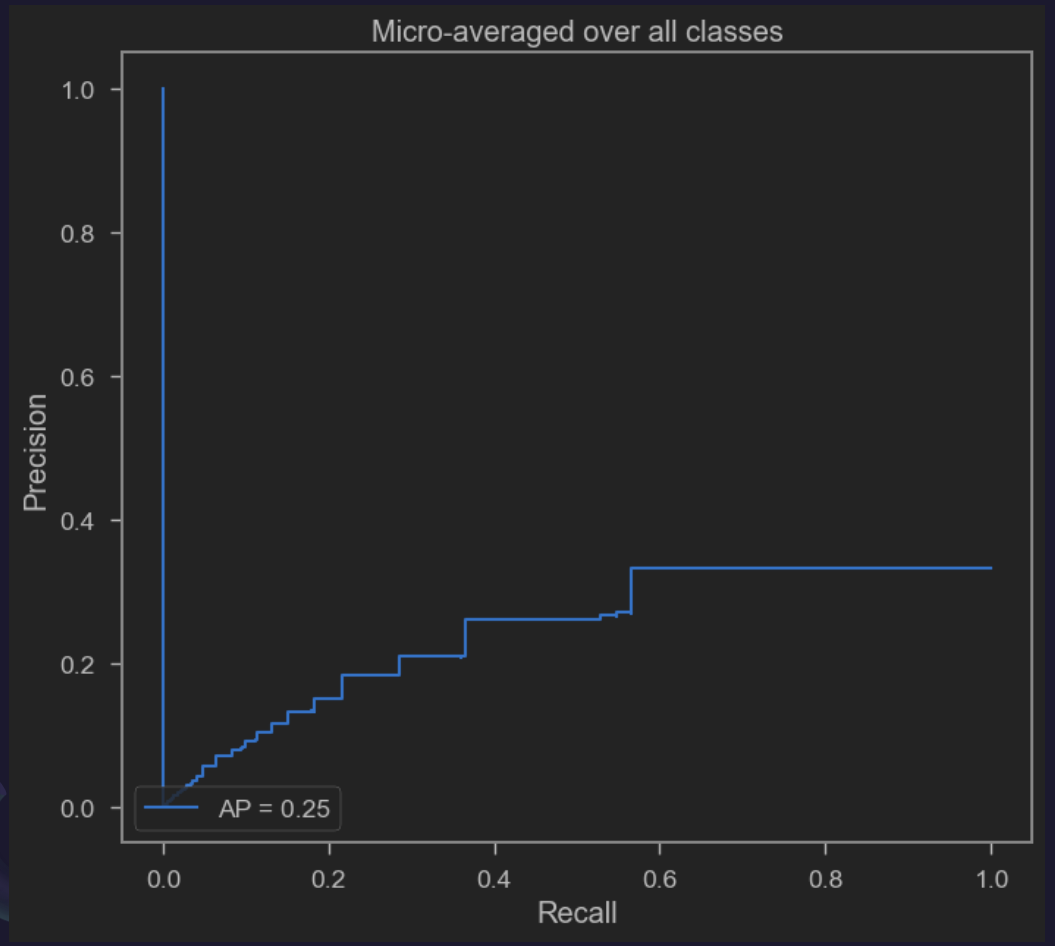
```
Mejor valor de accuracy obtenido con la mejor combinación: 0.9927231807951987
Mejor combinación de valores encontrados de los hiperparámetros: {'ccp_alpha': 0.0, 'class_weight': 'ba
lanced', 'criterion': 'entropy', 'max_depth': 13, 'min_samples_split': 2}
Métrica utilizada: accuracy
```

	precision	recall	f1-score	support
Amarillo	1.00	1.00	1.00	227
Rojo	1.00	1.00	1.00	218
Verde	1.00	1.00	1.00	254
accuracy			1.00	699
macro avg	1.00	1.00	1.00	699
weighted avg	1.00	1.00	1.00	699

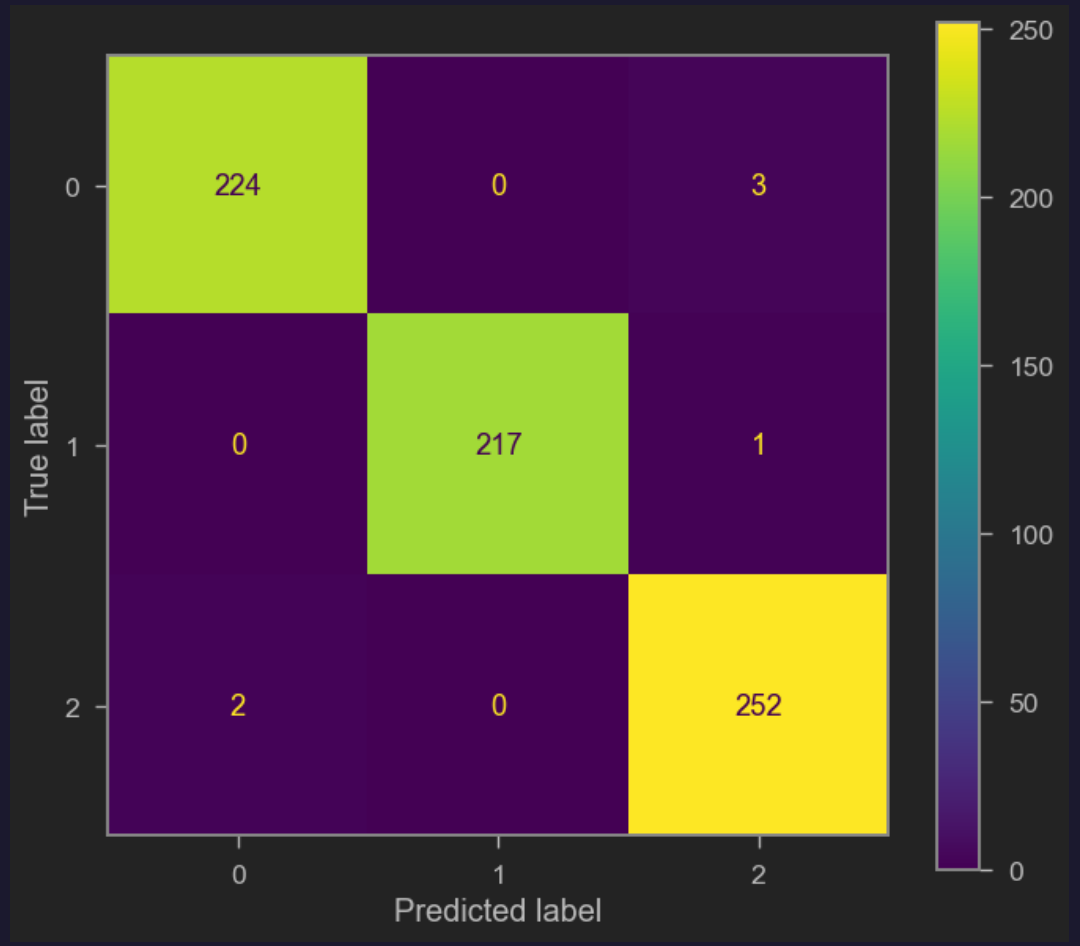
Ambos modelos tienen un valor de precisión muy aproximado

Resultados – Random Forest

Gráfica de 'Precision vs Recall'

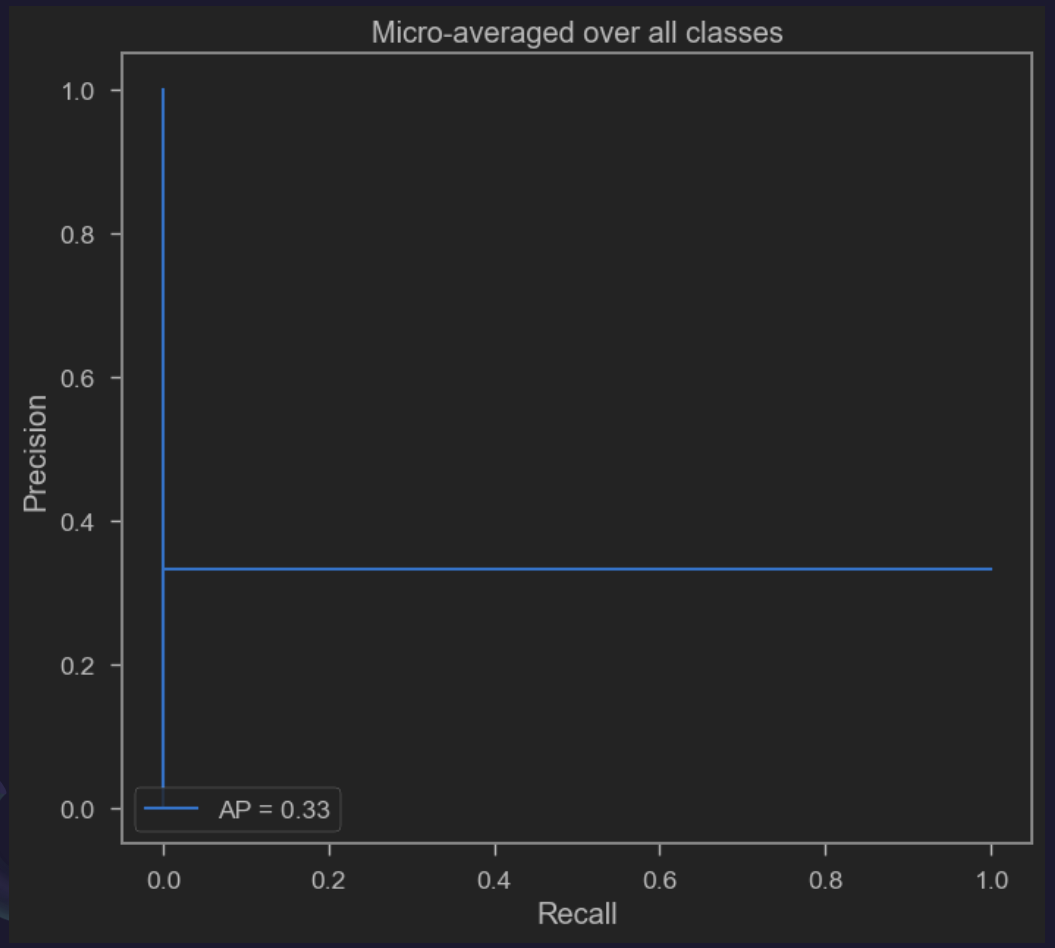


Matriz de Confusión

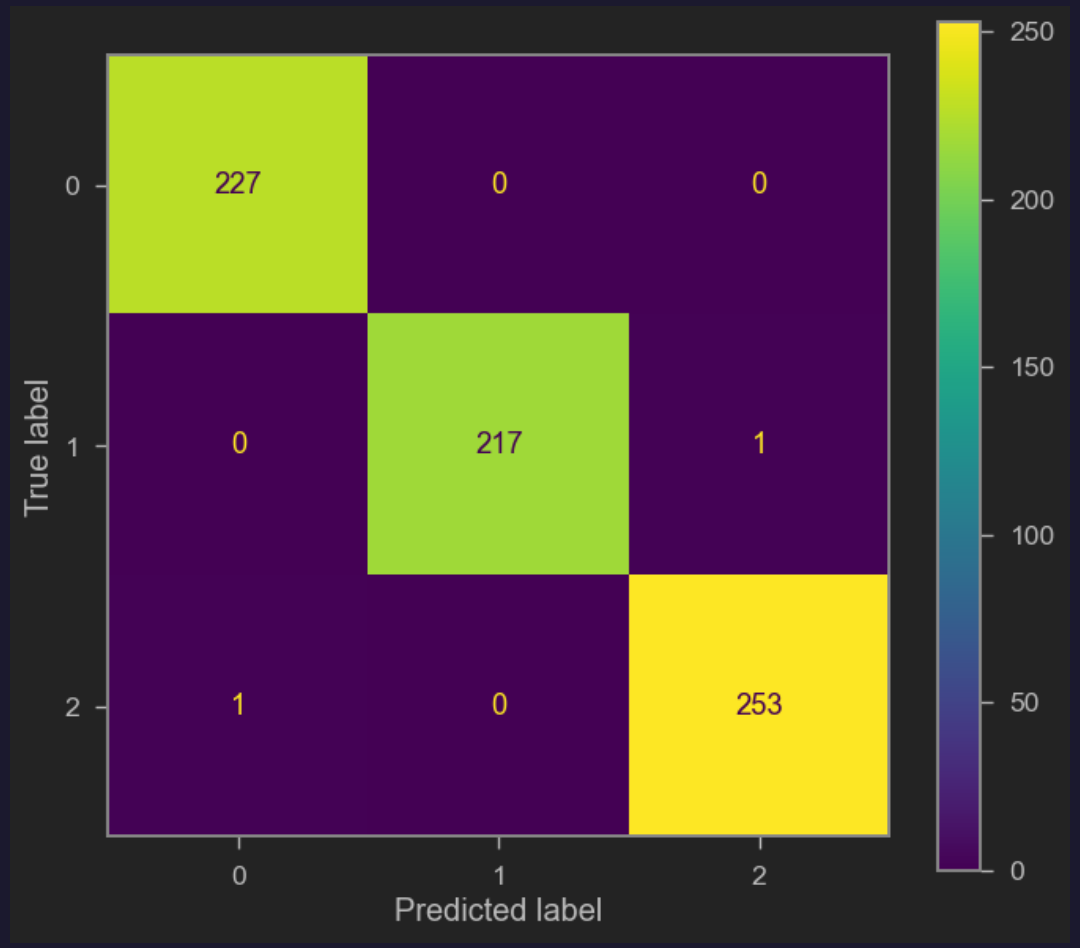


Resultados – Decision Tree

Gráfica de 'Precision vs Recall'



Matriz de Confusión



Conclusiones

De acuerdo a las predicciones que obtuvimos de los modelos de Random Forest y Decision Trees, podemos observar que la precisión de ambos algoritmos, Decision Trees (precisión 'accuracy' de %99.27) y Random Forest (precisión 'accuracy' de %98.65) es similar, los dos funcionan de forma adecuada, sin embargo, el de Random Forest consume más tiempo computacional al ejecturse para realizar el 'Grid Search' por lo que tomando en consideración este último aspecto, para este caso particular es más conveniente usar Decision Trees y tuvo una exactitud ligeramente mejor.