# CIENCIA Y ANALÍTICA DE DATOS

## RETO: CLASIFICACIÓN DE AGUAS SUBTERRÁNEAS

Profesora: Dr. María de la Paz Rico Fernández

**Equipo 37**

Karina Zafra Vallejo - A01793979

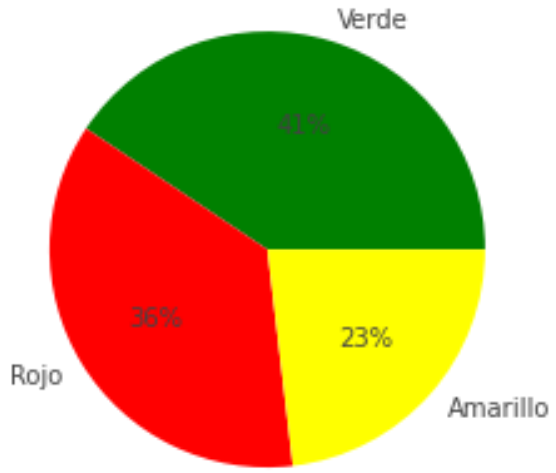
Francisco Javier Parga García - A01794380

*Viernes 18 de Noviembre de 2022*

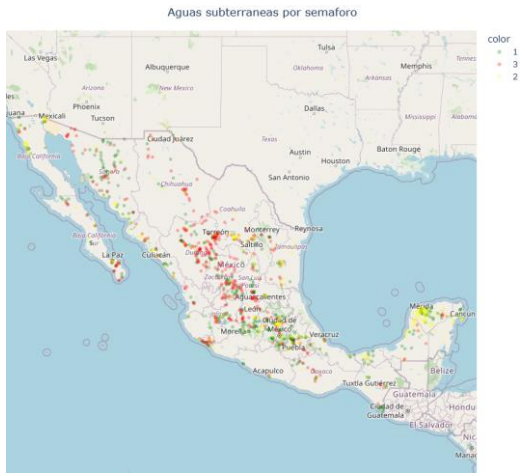
# Limpieza y acondicionamiento:

- Eliminación de caracteres “<” & “<=”
- Columnas sin información útil:
  - La columna 'SDT_mg/L' no tiene datos.
  - 'PERIODO' tiene un solo valor (2020)
- Ingeniería de datos
  - Lista de CONTAMINANTES --> Número de contaminantes
- Codificación de etiquetas:
  - SEMAFORO_cat --> 'Verde':1, 'Amarillo':2, 'Rojo':3
- Eliminación de variables
  - 'SEMAFORO', 'PERIODO', 'SDT_mg/L'

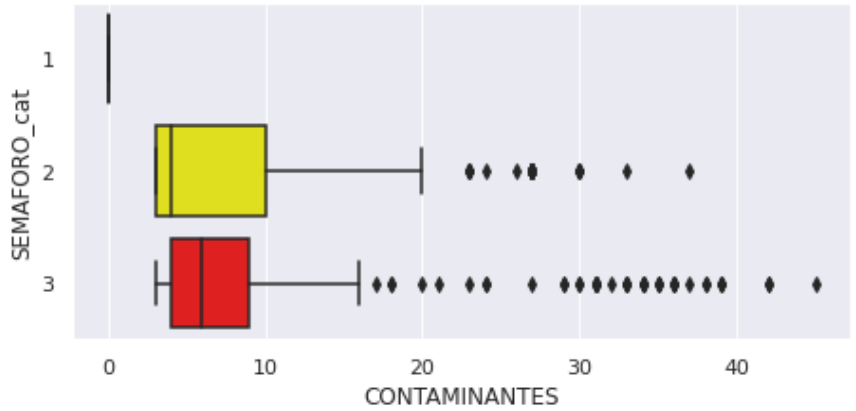
# Análisis

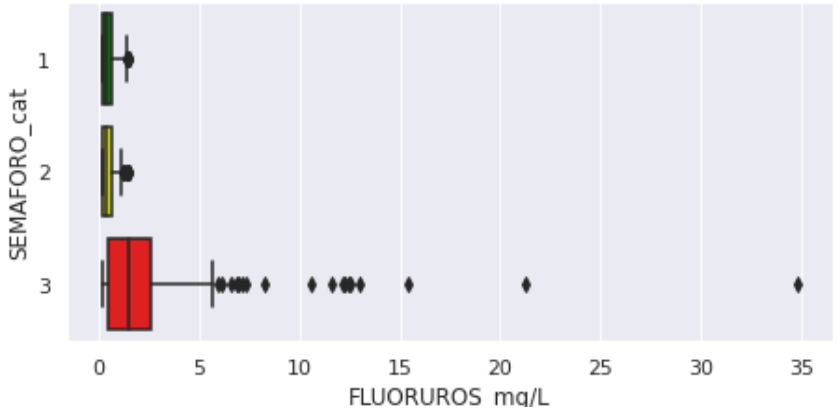



Distribución de las categorías de Semáforo




Contaminantes y Fluoruros diferenciados por semáforo



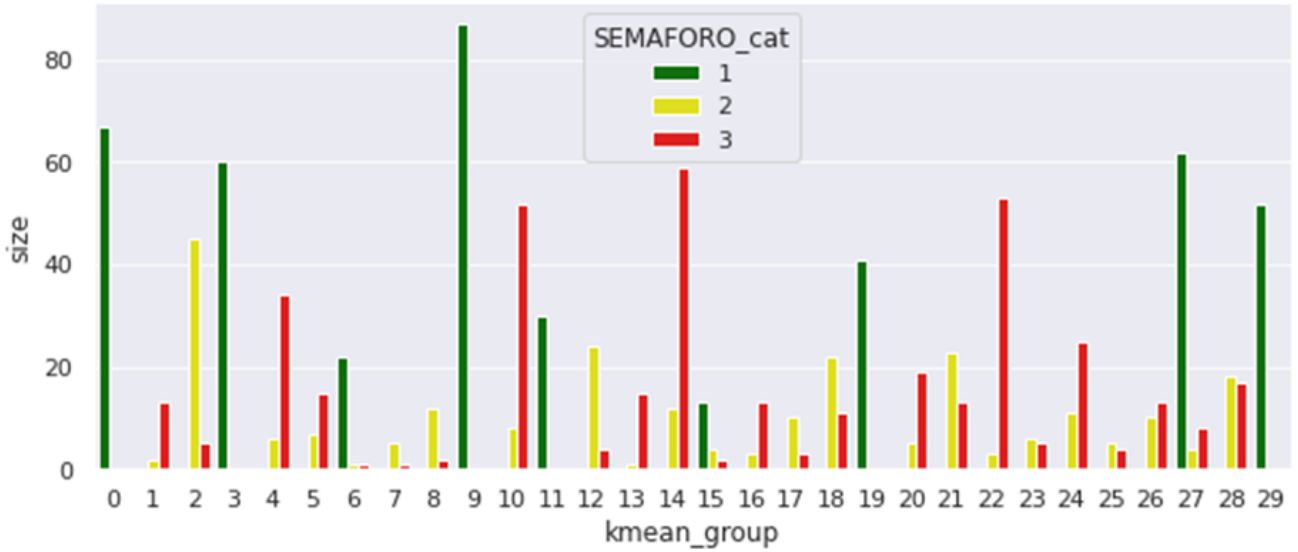


Correlación de Pearson de variables continuas con la categórica del semáforo.

# K-means:

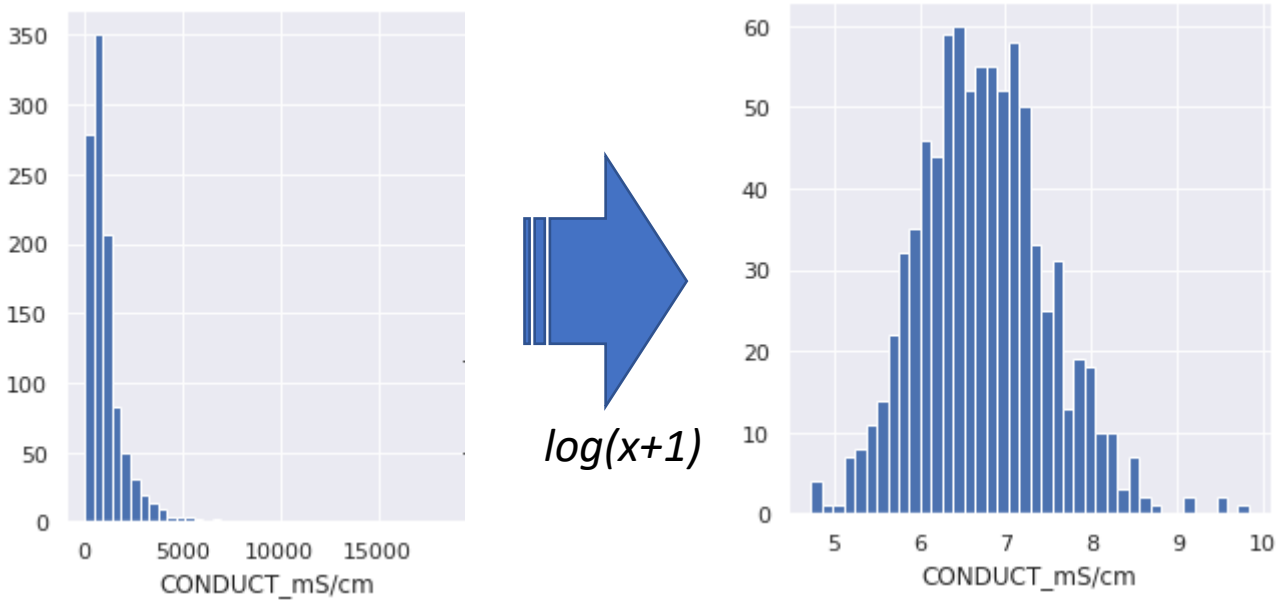
Con 30 grupos se logra separar la mayoría de las aguas con semáforo verde. Sin embargo, con esta técnica de agrupación no se logra discernir de los semáforos amarillos y rojos.

Se requiere de un modelo un poco más complejo para poder separar las clases adecuadamente, así como usar más variables como los Fluoruros, donde se observa cierta separación entre esos semáforos.



# Pipeline - Transformación de datos

- Transformación variables continuas: logaritmo



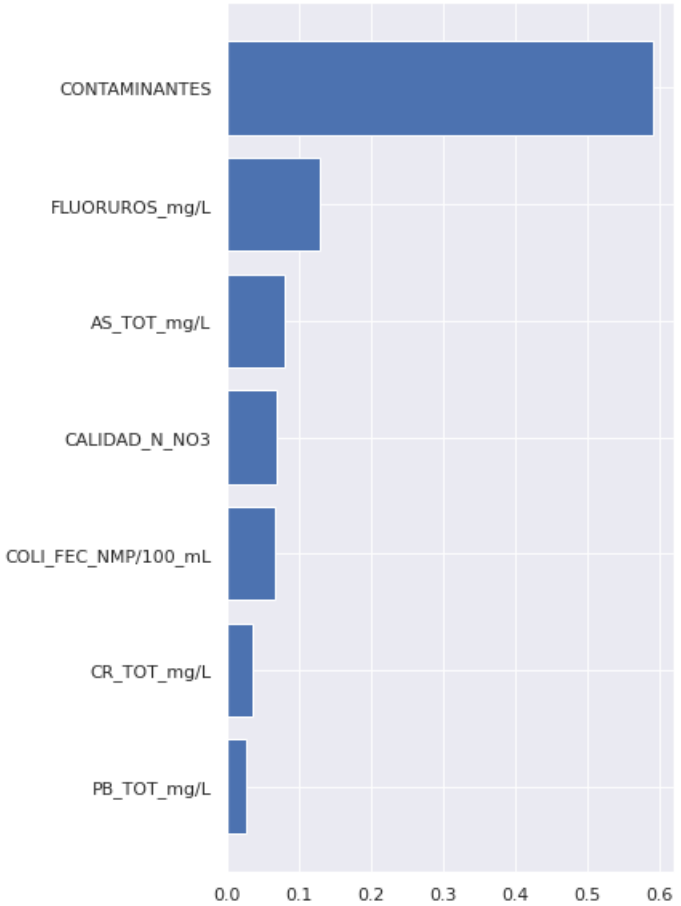

- Transformación variables categóricas: Codificación de etiquetas

| CALIDAD_COLI_FEC | CODE |
|---|---|
| 'Aceptable' | 0 |
| 'Buena calidad' | 1 |
| 'Contaminada' | 2 |
| 'Fuertemente contaminada' | 3 |
| 'Potable - Excelente' | 4 |

*Ejempo de codificación de etiquetas*

# Variables más importantes

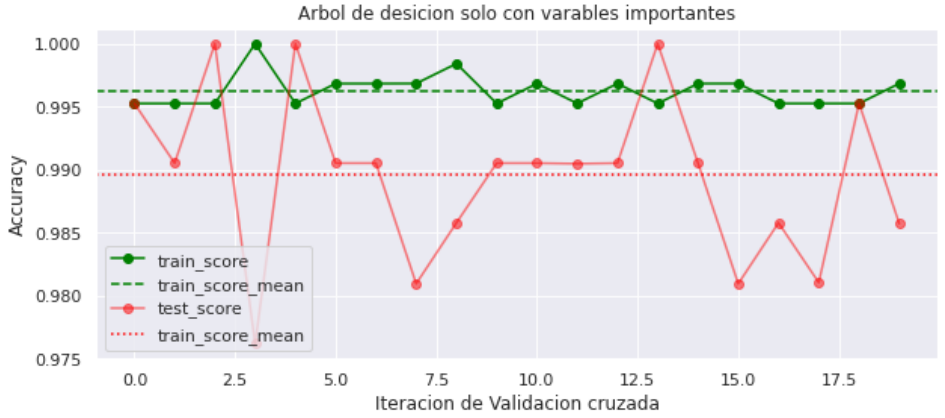
- Se logran identificar 7 variables principales



# Clasificación - Decission tree

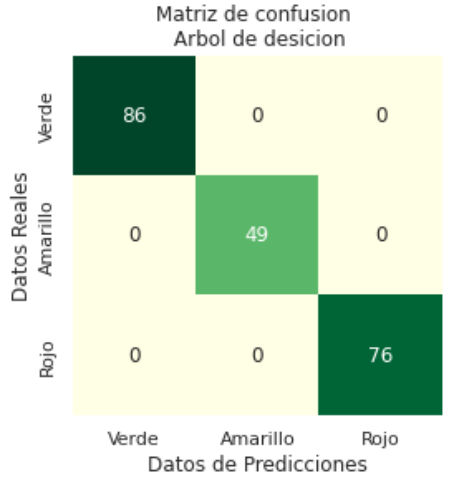
- Entrenamiento 80%,
- Prueba 20%

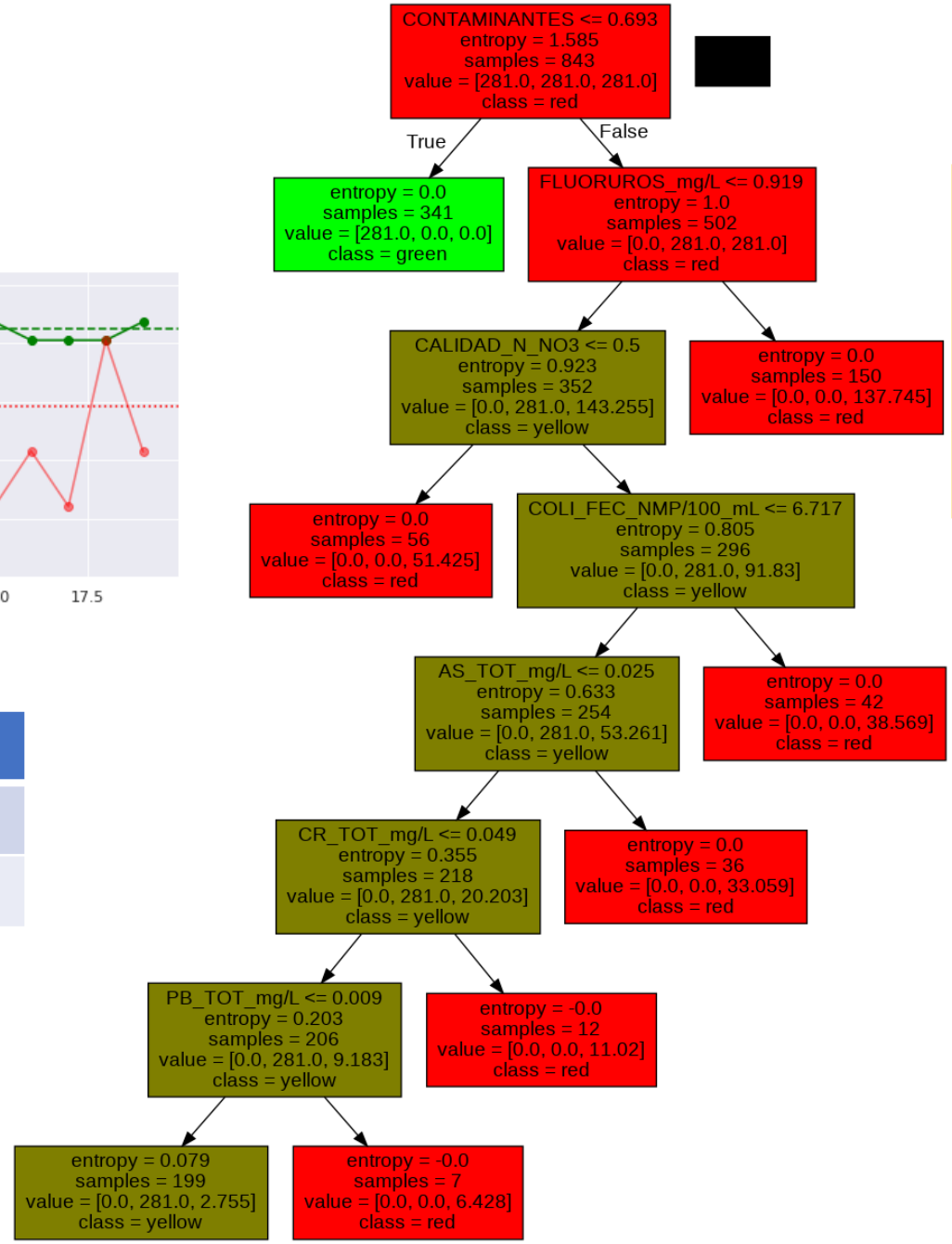
- Mejores hiperparámetros:
  - *ccp_alpha= 0,*
  - *class_weight='balanced',*
  - *criterion= 'entropy',*
  - *max_depth=7,*
  - *min_samples_split= 2,*




| Set de datos | Accuracy [%] |
|---|---|
| Entrenamiento | 99.64 |
| Prueba | 100.00 |

Datos de prueba

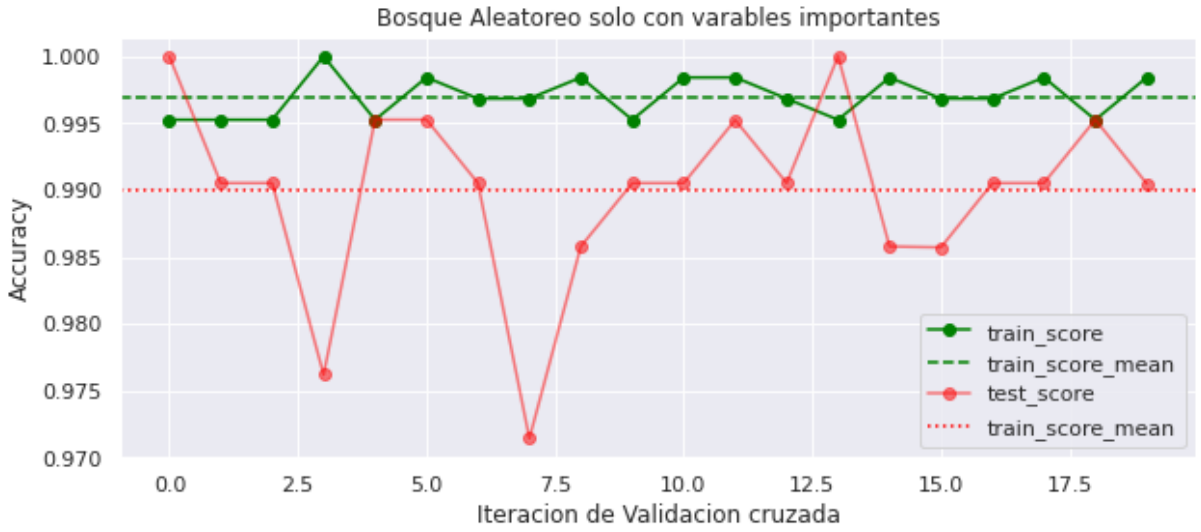


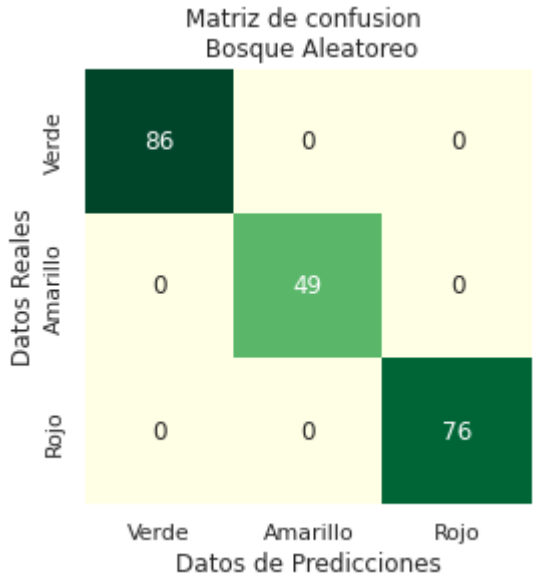
# Clasificación - Ramdom forest

- Entrenamiento 80%,
- Prueba 20%

- Mejores hiperparámetros:
  - *ccp_alpha= 0,*
  - *class_weight='balanced',*
  - *criterion= 'entropy',*
  - *max_depth=11,*
  - *min_samples_split= 3,*




Datos de prueba




| Set de datos | Accuracy [%] |
|---|---|
| Entrenamiento | 99.64 |
| Prueba | 100.00 |

# Resultados

- ‘CONTAMINANTES’ es determinante para clasificar la clase verde de ‘SEMAFORO_cat’.
- Se requieren de 6 variables adicionales para lograr minuciosamente la separación entre clase amarilla y roja de ‘SEMAFORO_cat’.
- En el árbol se observa que ‘FLUORUROS’ logra separar 150 samples de las 153 que se separan entre las otras 5 variables restantes.
- Debido a que no se observa un cambio significativo en el accuracy del Decision Tree y el Random Forest, se opta por quedarnos con Decision Tree.

# Conclusiones

- Es posible mejorar la clasificación con ingeniería de datos
  - *Contaminantes* ayuda a separar muy bien la clase *Verde*
- La validación cruzada ayudó a verificar que los modelos no están sobre entrenados y  muestran un performace muy alto
- Para el caso donde solo se usaron 7 variables, ambos modelos clasifican con un accuracy de 100% los datos de prueba