Reto: Base de datos calidad de agua

Ciencia y Analítica de datos

Profesor Titular: María de la Paz Rico Fernández

Maestría en Inteligencia Artificial Aplicada (MNA-V)

18/11/2022

Equipo 24

Victor Hugo Avila Felipe - A01794425

Andrés Eduardo Figueroa García - A01378536

Pipeline

- Adquisición de datos

- Preprocesamiento

- Procesamiento

- Entrenamiento

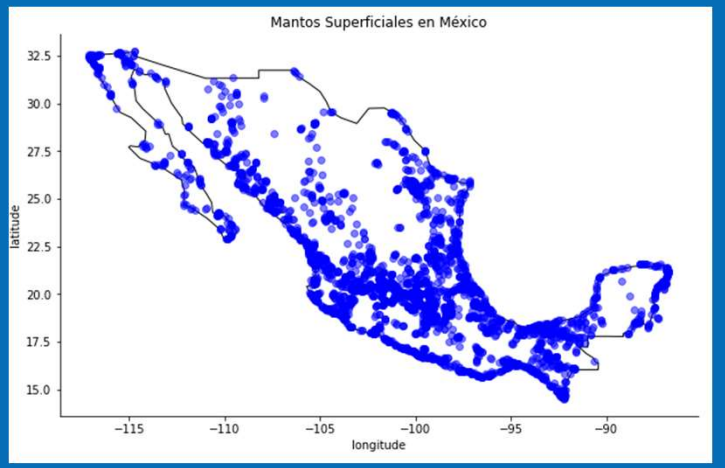
- Mantenimiento del modelo

Adquisición de Datos

- base de datos de aguas superficiales periodo 2020.csv
- Obtenidos del portal de la CONAGUA.
- Carga de datos.

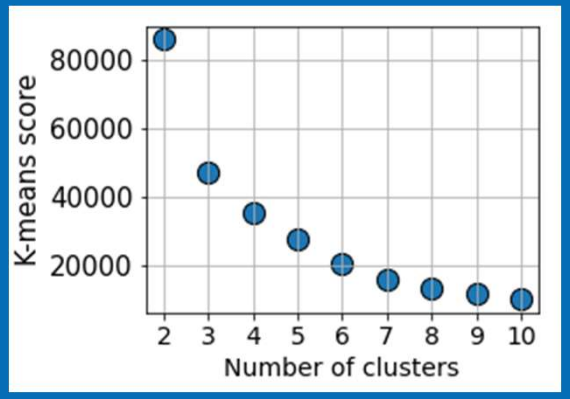
Preprocesamiento

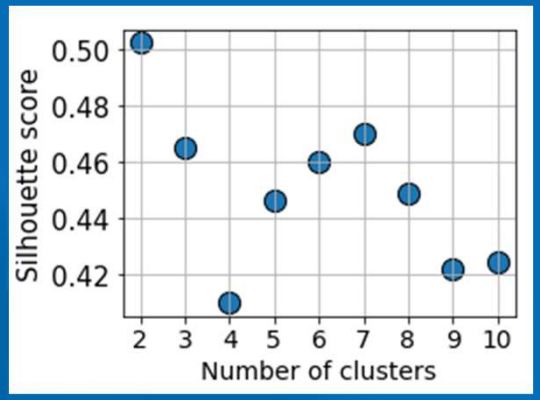
- Exploración de datos.
- Normalización de datos:
  - Eliminar N/A y convervar var. categoricas
  - Imputación de datos en registros vacíos
  - Eliminar ruido
  - Análisis de histograma de datos
  - Previsualización de coordenadas

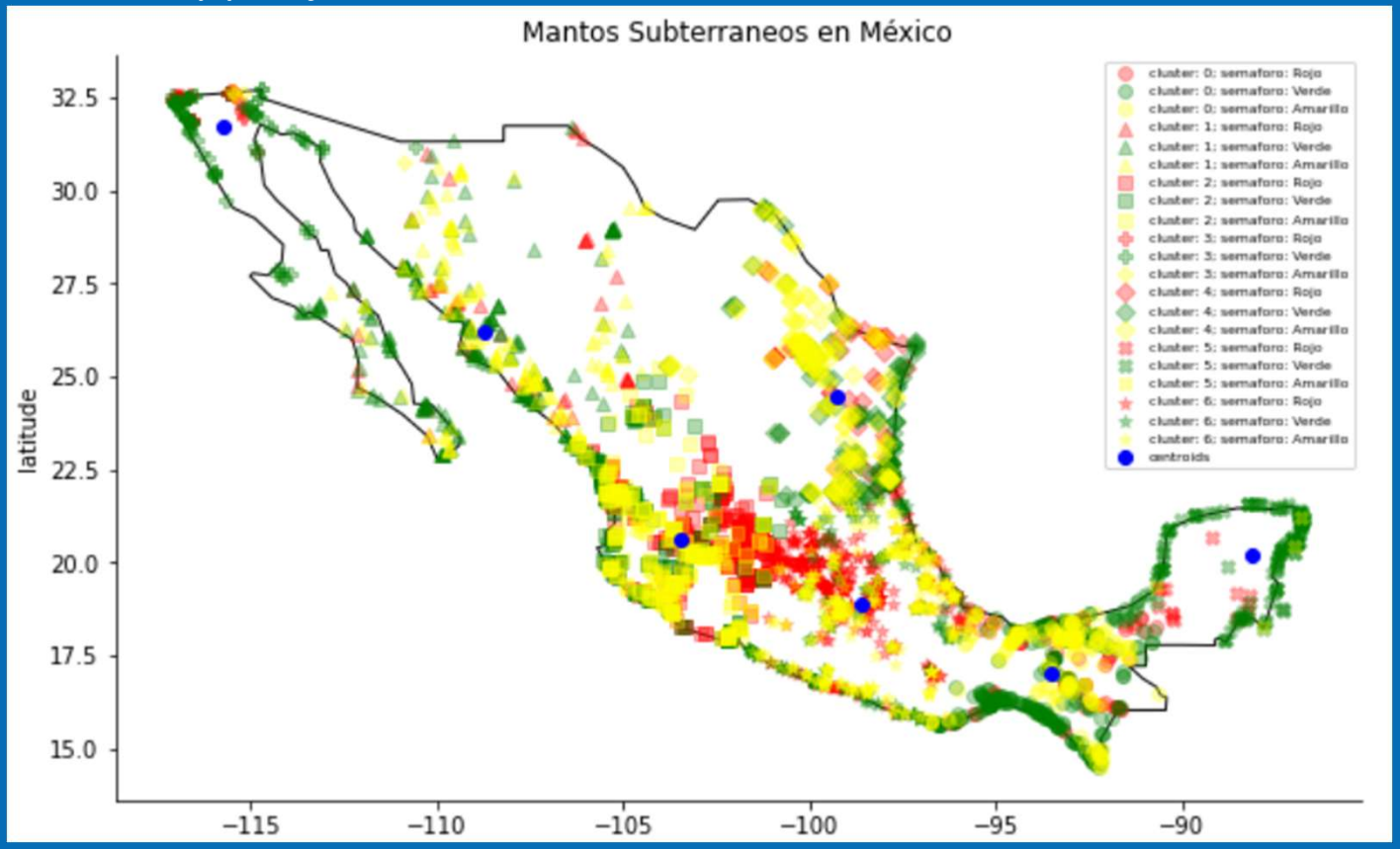


Procesamiento: K-Means

- Método de agrupamiento de variables en clusters (subconjuntos de datos).
- Algoritmo no supervisado.
- **Conclusión:** para clasificar los puntos en relación al semáforo no es suficiente la ubicación geográfica, aunque existe una tendencia hacia cierto tipo de color de semáforo, existen otros en cuyas clases estos se encuentran muy parejos.



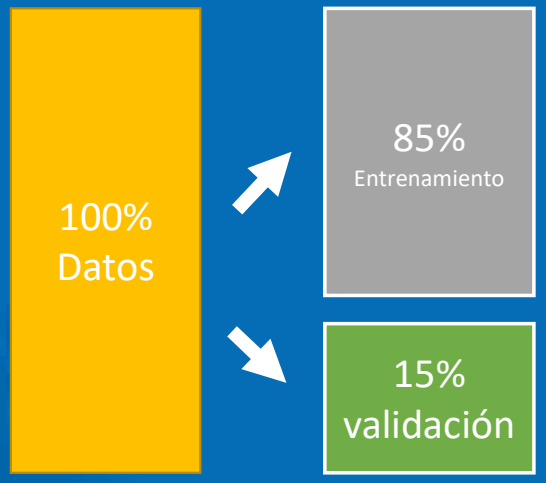


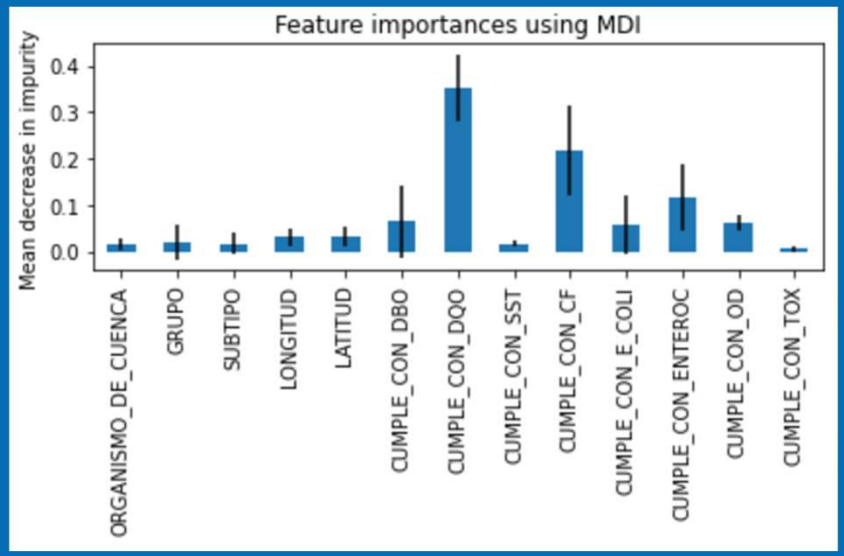


Procesamiento: Preparación datos para modelos

- Identificador/índice: CLAVE
- Variables categóricas -> LabelEncoder
- Variables dependientes X, Variable dependiente Y (Semaforo).

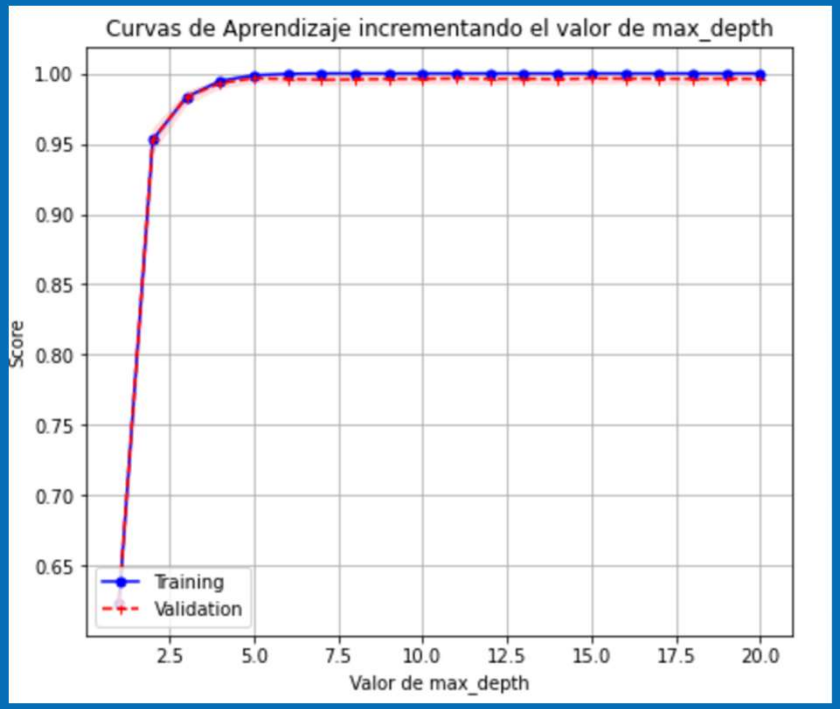
Particionamiento de datos para aprendizaje supervisado.





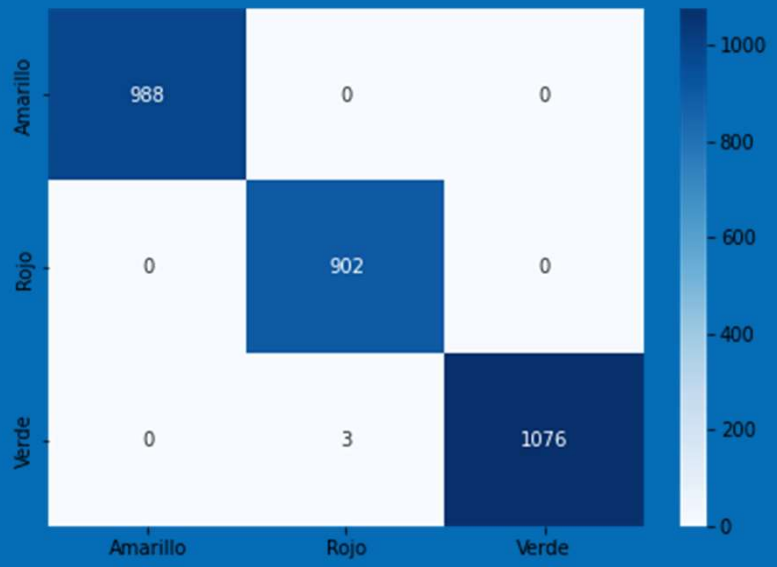
Entrenamiento - Decision trees

- Se puede observar que a partir de una produndidad de 6, el sistema tiene un score de 1 de manera constante. Se realiza una evaluación sobre los datos de entrenamiento y se muestran resultados en una matriz de confusión.

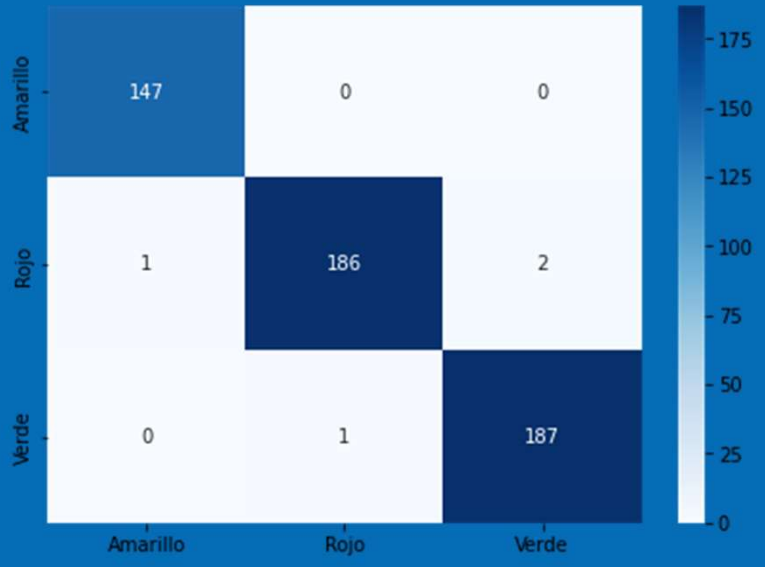


Evaluación: Matriz de Confusion

- Entrenamiento



- Validación



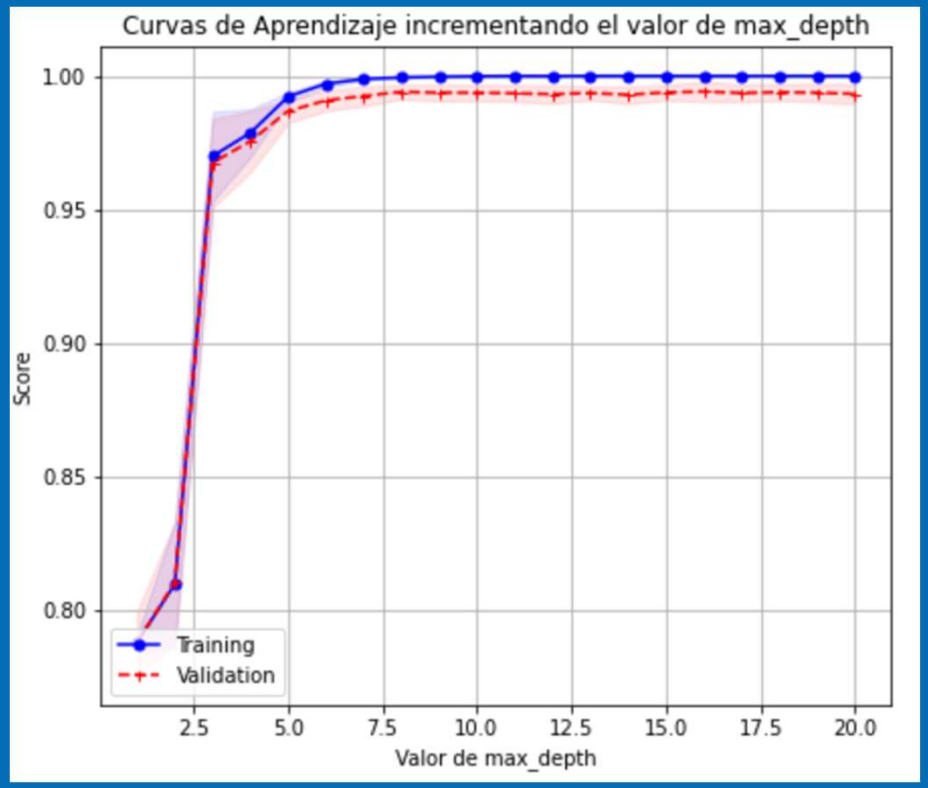
- reporte de clasificación:

```
[52] target_names = ['Amarillo', 'Rojo', 'Verde']
     print(classification_report(Y_test, Y_hat, target_names=target_names))
```

	precision	recall	f1-score	support
Amarillo	0.99	1.00	1.00	147
Rojo	0.99	0.98	0.99	189
Verde	0.99	0.99	0.99	188
accuracy			0.99	524
macro avg	0.99	0.99	0.99	524
weighted avg	0.99	0.99	0.99	524

Entrenamiento – Random Forest

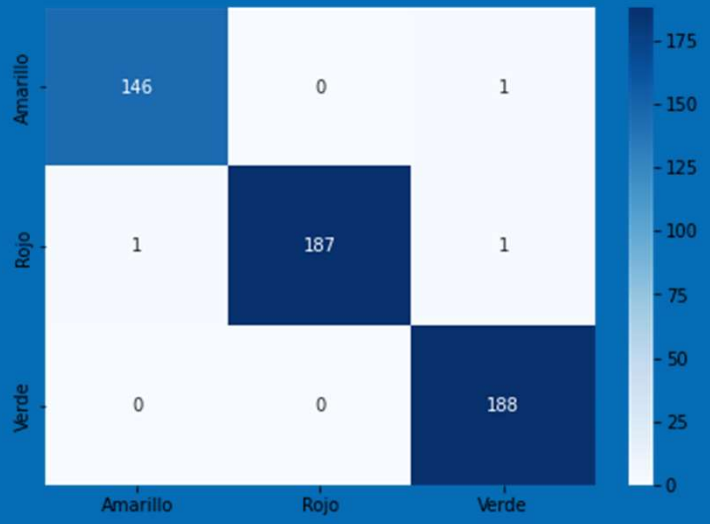
- Se puede observar que a partir de una produndidad de 6, el sistema tiene un score de 1 de manera constante. Se realiza una evaluación sobre los datos de entrenamiento y se muestran resultados en una matriz de confusión.



Evaluación: Matriz de Confusion

- Entrenamiento

- Validación



- reporte de clasificación:

```
[58] target_names = ['Amarillo', 'Rojo', 'Verde']
     print(classification_report(Y_test, Y_hat, target_names=target_names))
```

	precision	recall	f1-score	support
Amarillo	0.99	0.99	0.99	147
Rojo	1.00	0.99	0.99	189
Verde	0.99	1.00	0.99	188
accuracy			0.99	524
macro avg	0.99	0.99	0.99	524
weighted avg	0.99	0.99	0.99	524

Conclusiones

- Realizar un análisis del Feature Importance con el bosque aleatorio nos hizo ver que si bien la columna nombrada como 'CUMPLE_CON_DQO' es la que tiene más importancia en el modelo, todas tienen hasta cierto punto un impacto para este. Era de esperarse que las variables ya categorizadas a partir de los valores numéricos fueran más importantes, pero se puede ver que hay algunas variables que inician con la etiqueta 'CUMPLE' que resultaron tener menor importancia que algunas otras como el GRUPO, SUBTIPO, CUENTA e incluso las coordenadas geográficas.

- Por otro lado, se puede observar que para el Árbol de decisiones y el Bosque Aleatorio se obtiene muy buenos resultados en el conjunto de entrenamiento y que estos se manetiene en muy buenos valores para los conjuntos de prueba. Del reporte de clasificación se observa que para Precisión, Recall y F1-Score, el valor más bajo en todas las clases es de 0.98. Esto es muy bueno, porque quiere decir que no sólo clasifica de evitando falsos positivos, sino que también minimiza los falsos negativos. De igual forma la métrica de Accuracy es de 0.99 en el caso más bajo, lo cual hace que la evaluación de los modelos haya resultado de manera satisfactoria.