



Curso Métodos y Modelos

Profesora:
Karen Ballesteros-González PhD.

Modelos Estadísticos: Modelos de Aprendizaje Automático

c). Modelos de Aprendizaje Automático

Teoría:

- Diferencia entre aprendizaje supervisado y no supervisado.
- Algoritmos: K-Means, DBSCAN, Clustering Jerárquico.

Aplicación práctica:

- Segmentación de clientes o zonas de monitoreo ambiental.

Modelos Estadísticos: Modelos de Aprendizaje Automático

Tipos de Aprendizaje Automático

El aprendizaje automático (machine learning) es una rama de la inteligencia artificial que permite a los sistemas aprender de los datos y hacer predicciones o descubrir patrones sin estar explícitamente programados para ello.

Aprendizaje Supervisado

- Regresión
- Clasificación – Árboles de Decisión

Entrenamiento con etiquetas conocidas: se le “enseña” al modelo qué resultado debe aprender.

Aplicaciones:

- Predecir la concentración de PM2.5 en una estación (regresión).
- Clasificar si un día será “bueno”, “regular” o “malo” según la calidad del aire (clasificación).

Ejemplo ambiental:

Entrenar un modelo con datos históricos de temperatura, humedad, velocidad del viento y PM10, para predecir los niveles de PM2.5.

Aprendizaje No Supervisado

- Clustering - K-means

No hay etiquetas en los datos: el modelo explora y encuentra patrones o agrupaciones ocultas.

Aplicaciones:

- Identificar zonas de la ciudad con patrones similares de contaminación.
- Agrupar usuarios según comportamiento energético.

Ejemplo ambiental:

Usar clustering para segmentar estaciones de monitoreo según similitud en variables atmosféricas, sin saber de antemano a qué categoría pertenecen.

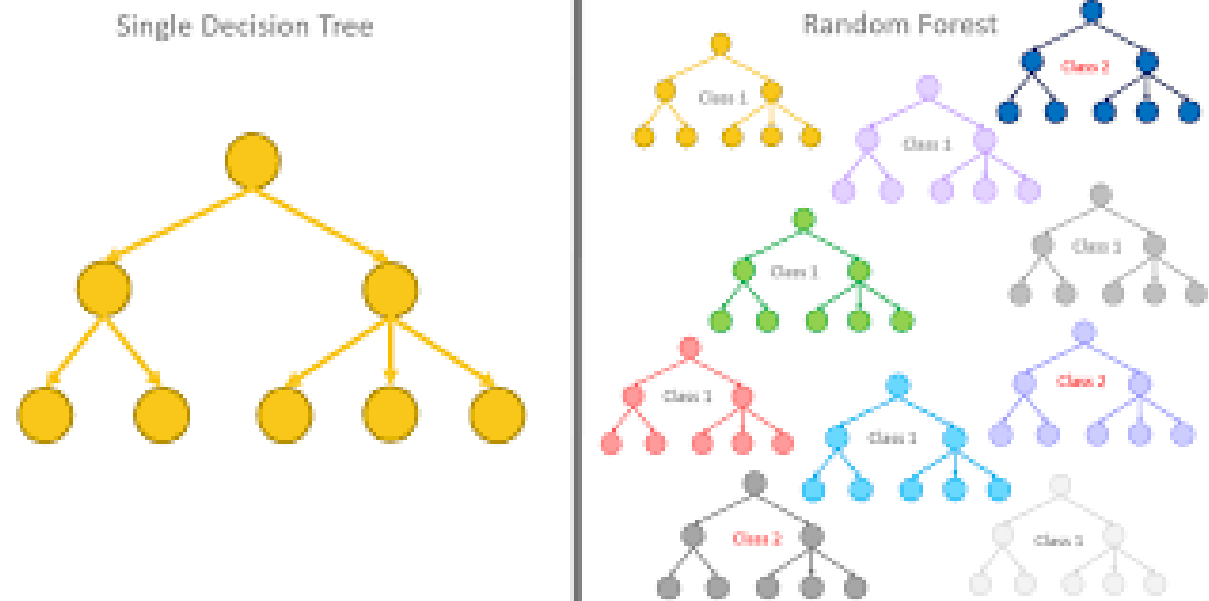
Modelos Estadísticos: Modelos de Aprendizaje Supervisado

Aprendizaje Supervisado - Clasificación

¿Qué es la clasificación supervisada?

Es un método de aprendizaje automático supervisado que consiste en entrenar un modelo para que aprenda a asociar patrones en los datos con una etiqueta de clase. El objetivo es que el modelo pueda, luego, predecir la clase correcta para nuevos datos.

Arboles de Decisión



Modelos Estadísticos: Modelos de Aprendizaje Supervisado

¿Qué es un Árbol de Decisión?

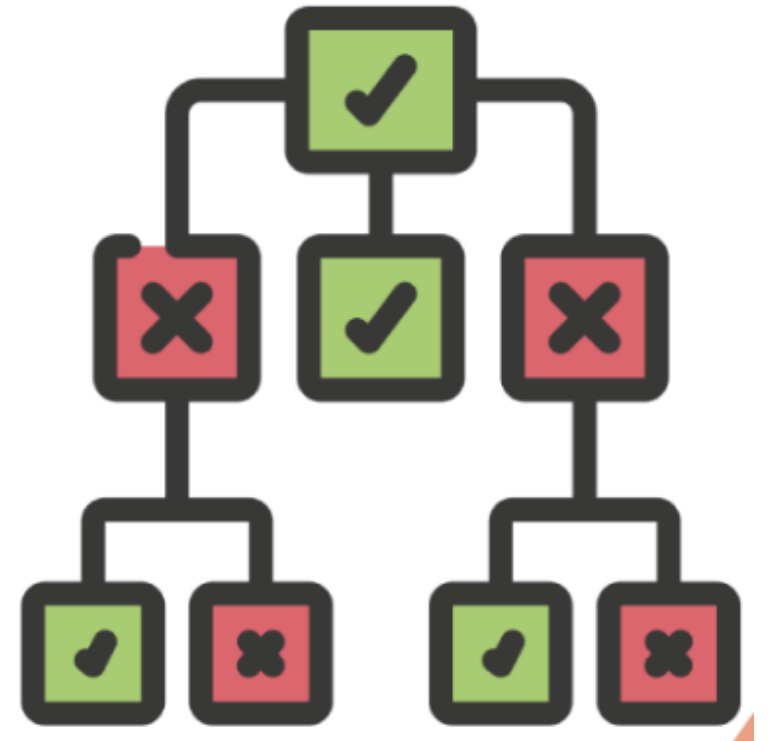
Un **árbol de decisión** es un modelo predictivo que representa una serie de decisiones organizadas en forma de árbol. Se utiliza tanto para problemas de **clasificación** como de **regresión**.

¿Cómo funciona?

Un árbol toma decisiones **dividiendo los datos en pasos lógicos** (ramas) basados en las **características** de las observaciones, hasta llegar a una **predicción final** (hoja).

Estructura básica:

- **Nodo raíz:** es el punto inicial del árbol (donde comienza la primera división).
- **Nodos internos:** son los puntos donde se hacen divisiones según condiciones (por ejemplo: “¿pH > 6.5?”).
- **Hojas:** son los resultados o predicciones (por ejemplo: “Calidad del agua: Buena”).



Modelos Estadísticos: Modelos de Aprendizaje Supervisado

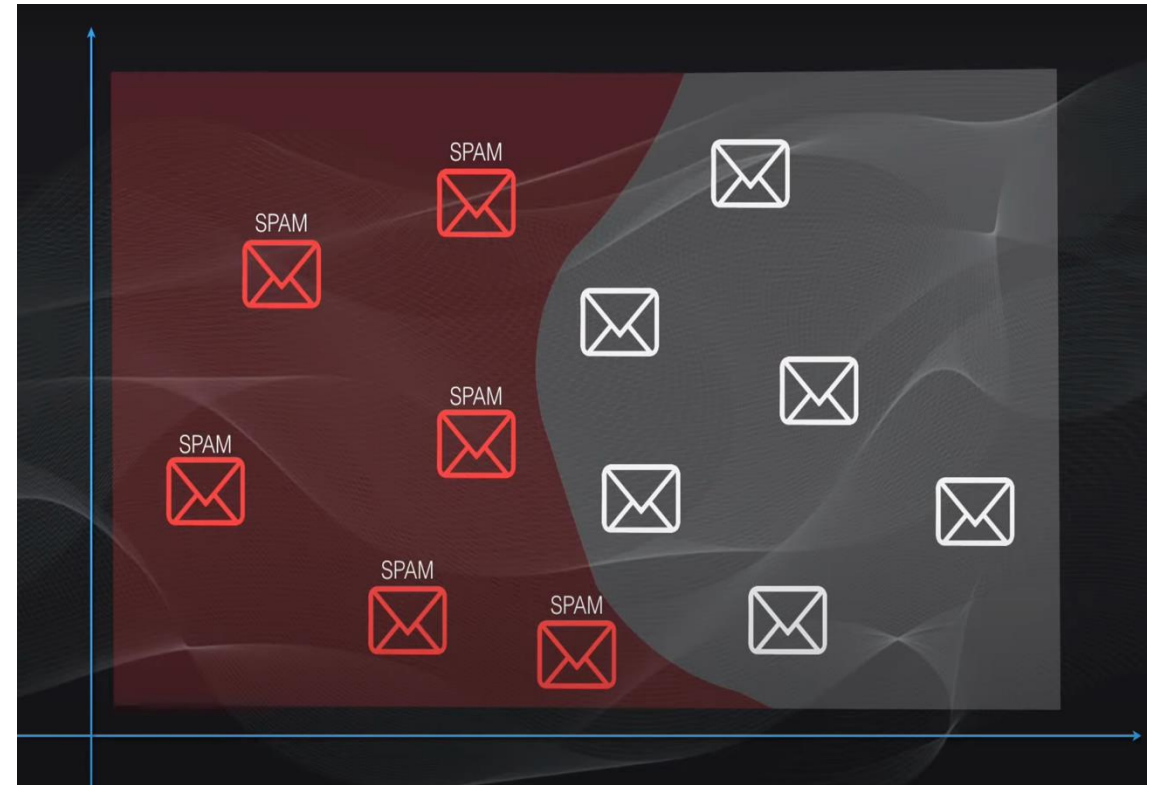
¿Cómo funciona el Árbol de Decisión?

Un árbol toma decisiones **dividiendo los datos en pasos lógicos** (ramas) basados en las **características** de las observaciones, hasta llegar a una **predicción final** (hoja).

Estructura básica:

- **Nodo raíz:** es el punto inicial del árbol (donde comienza la primera división).
- **Nodos internos:** son los puntos donde se hacen divisiones según condiciones (por ejemplo: “¿pH > 6.5?”).
- **Hojas:** son los resultados o predicciones (por ejemplo: “Calidad del agua: Buena”).

CART: Árboles de Clasificación y Regresión

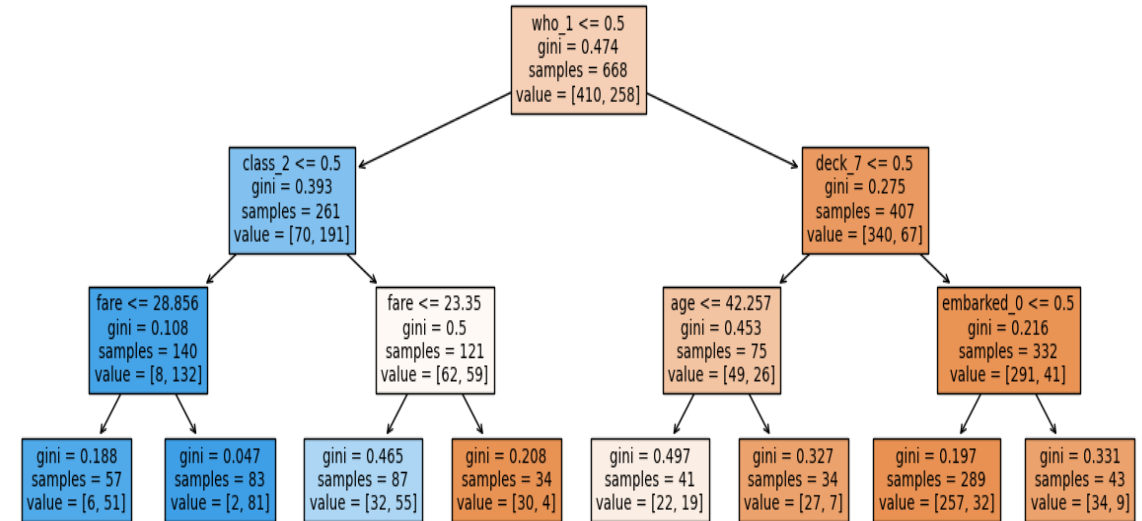


Modelos Estadísticos: Modelos de Aprendizaje Supervisado

CART: Árboles de Clasificación y Regresión: Es uno de los algoritmos más populares para construir **árboles de decisión**.

Es un algoritmo desarrollado por Breiman et al. (1986) que permite construir árboles de decisión para:

- **Clasificación:** si la variable objetivo es **categorica** (ej. calidad del agua: *Buena*, *Aceptable*, *Contaminada*).
- **Regresión:** si la variable objetivo es **numérica** (ej. concentración de DBO).



¿Cómo funciona CART?

División binaria recursiva

CART siempre divide los datos en **dos ramas** (no más), usando condiciones del tipo:

- ¿La variable X es menor o mayor que un cierto umbral?
- ¿Oxígeno disuelto > 6.5 mg/L?
- Sí o No

Modelos Estadísticos: Modelos de Aprendizaje Supervisado

CART: Árboles de Clasificación y Regresión

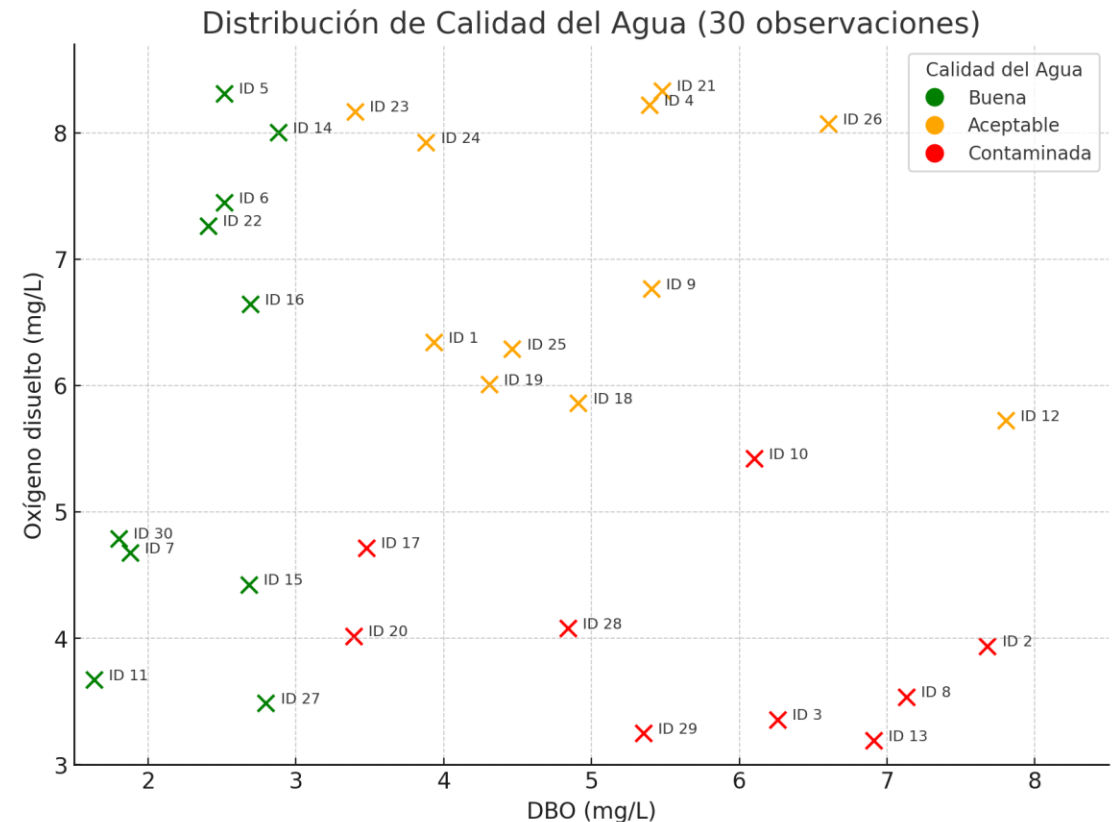
Ejercicio:

Clasificar la calidad del agua (Buena, Aceptable, Contaminada) usando las variables:

- DBO (Demanda Biológica de Oxígeno)
- Oxígeno disuelto

Result

	ID	DBO	Oxigeno_disuelto	Calidad_Agua
0	1	3.934511	6.341497	Aceptable
1	2	7.679643	3.937883	Contaminada
2	3	6.257961	3.357784	Contaminada
3	4	5.391280	8.218870	Aceptable
4	5	2.514121	8.310976	Buena



Modelos Estadísticos: Modelos de Aprendizaje Supervisado

CART: Árboles de Clasificación y Regresión





Índice Gini

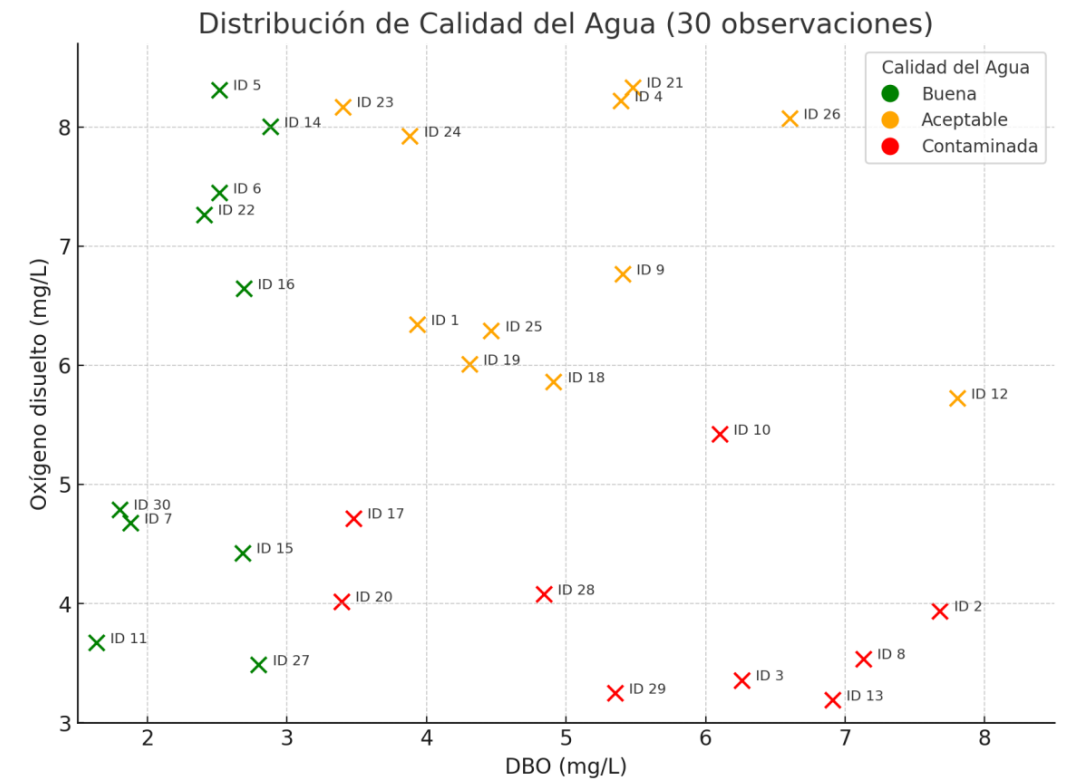
$$Gini_{raíz} = 1 - \sum p_i^2$$

Donde:

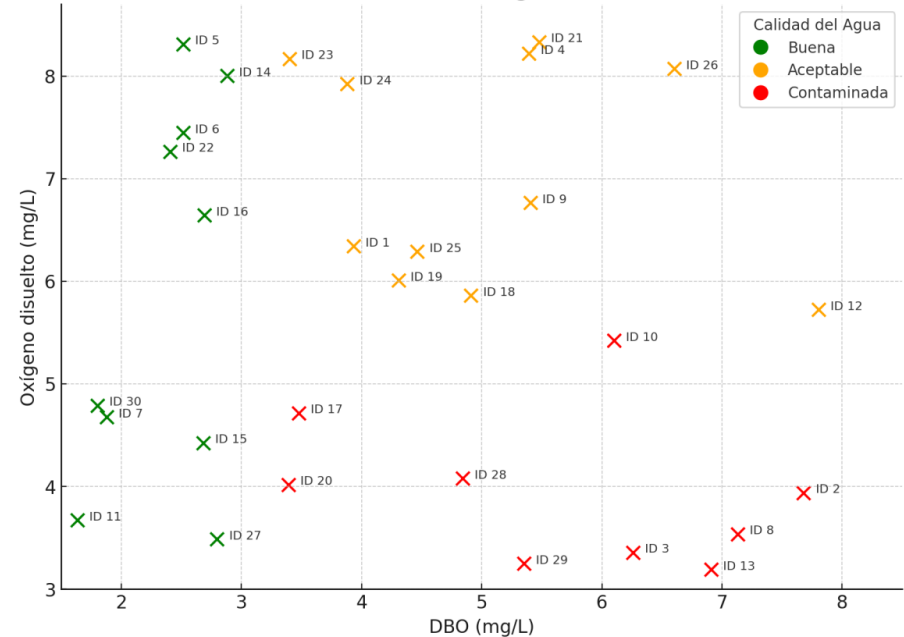
p_i es la proporción de cada clase (Buena, Aceptable, Contaminada) en el conjunto total.

Result				
	ID	DBO	Oxigeno_disuelto	Calidad_Agua
0	1	3.934511	6.341497	Aceptable
1	2	7.679643	3.937883	Contaminada
2	3	6.257961	3.357784	Contaminada
3	4	5.391280	8.218870	Aceptable
4	5	2.514121	8.310976	Buena

Nodos Puros		$1 - (10/10)^2 - (0/10)^2 = 1 - 1 - 0 = 0$
		$1 - (0/10)^2 - (10/10)^2 = 1 - 0 - 1 = 0$
Nodos In-Puros		$1 - (7/10)^2 - (3/10)^2 = 1 - 0.49 - 0.09 = 0.42$
		$1 - (5/10)^2 - (5/10)^2 = 1 - 0.25 - 0.25 = 0.5$



Distribución de Calidad del Agua (30 observaciones)

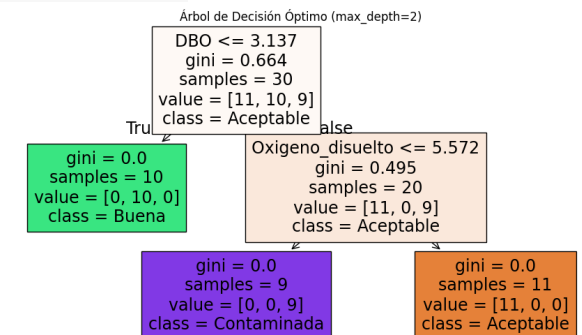


Modelos Estadísticos: Modelos de Aprendizaje Supervisado

CART: Árboles de Clasificación y Regresión

✓ Cálculo del Arbol de Decisión con Python

```
1 from sklearn.tree import plot_tree, DecisionTreeClassifier
2
3 # Variables predictoras y objetivo
4 X = df_agua[['DBO', 'Oxigeno_disuelto']]
5 y = df_agua['Calidad_Agua']
6
7 # Entrenar el mejor árbol de decisión con toda la data
8 mejor_arbol = DecisionTreeClassifier(criterion='gini', max_depth=2, random_state=42)
9 mejor_arbol.fit(X, y)
10
11 # Visualizar el árbol
12 plt.figure(figsize=(12, 6))
13 plot_tree(mejor_arbol, feature_names=X.columns, class_names=mejor_arbol.classes_, filled=True)
14 plt.title("Árbol de Decisión Óptimo (max_depth=2)")
15 plt.show()
```

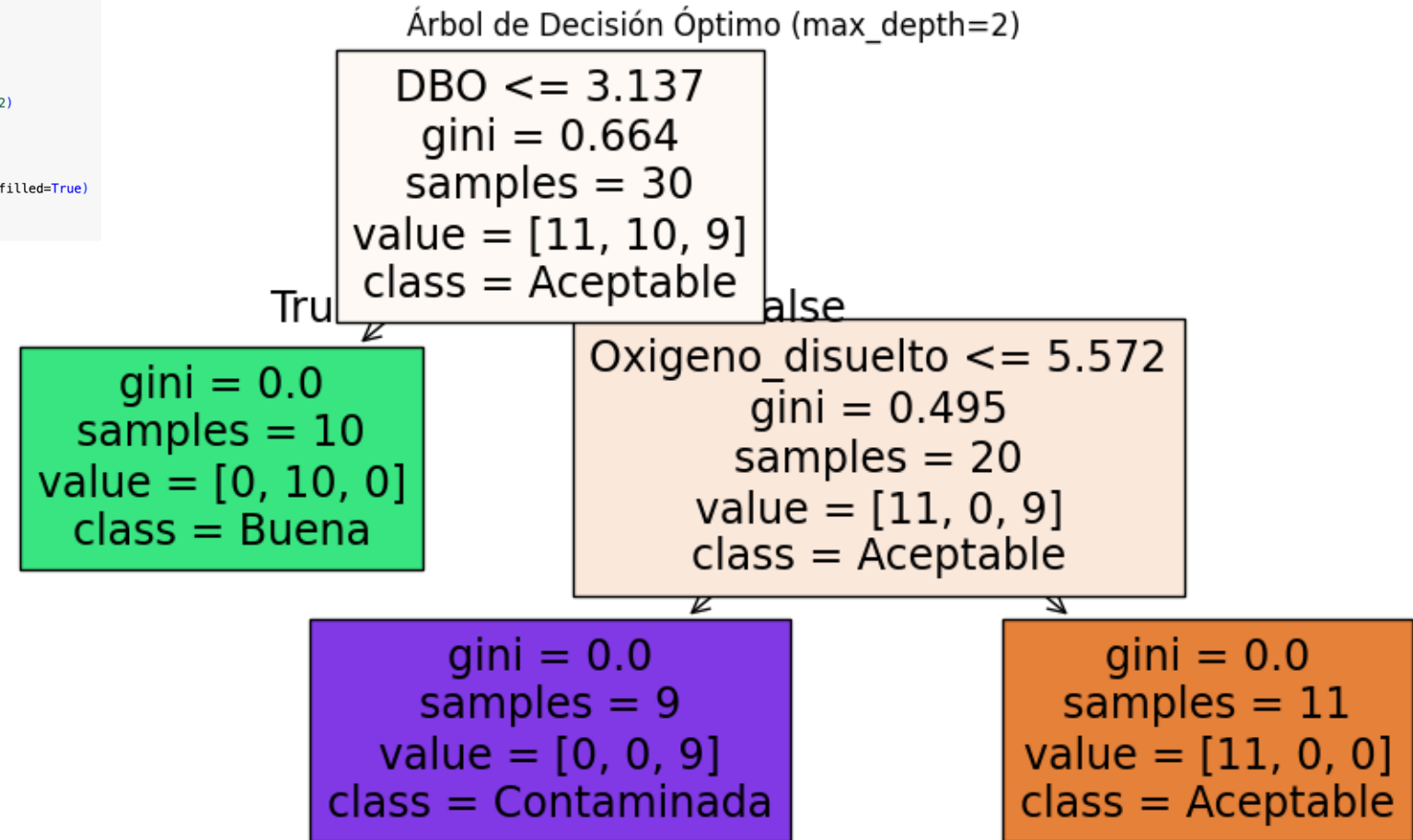


Modelos Estadísticos: Modelos de Aprendizaje Supervisado

CART: Árboles de Clasificación y Regresión

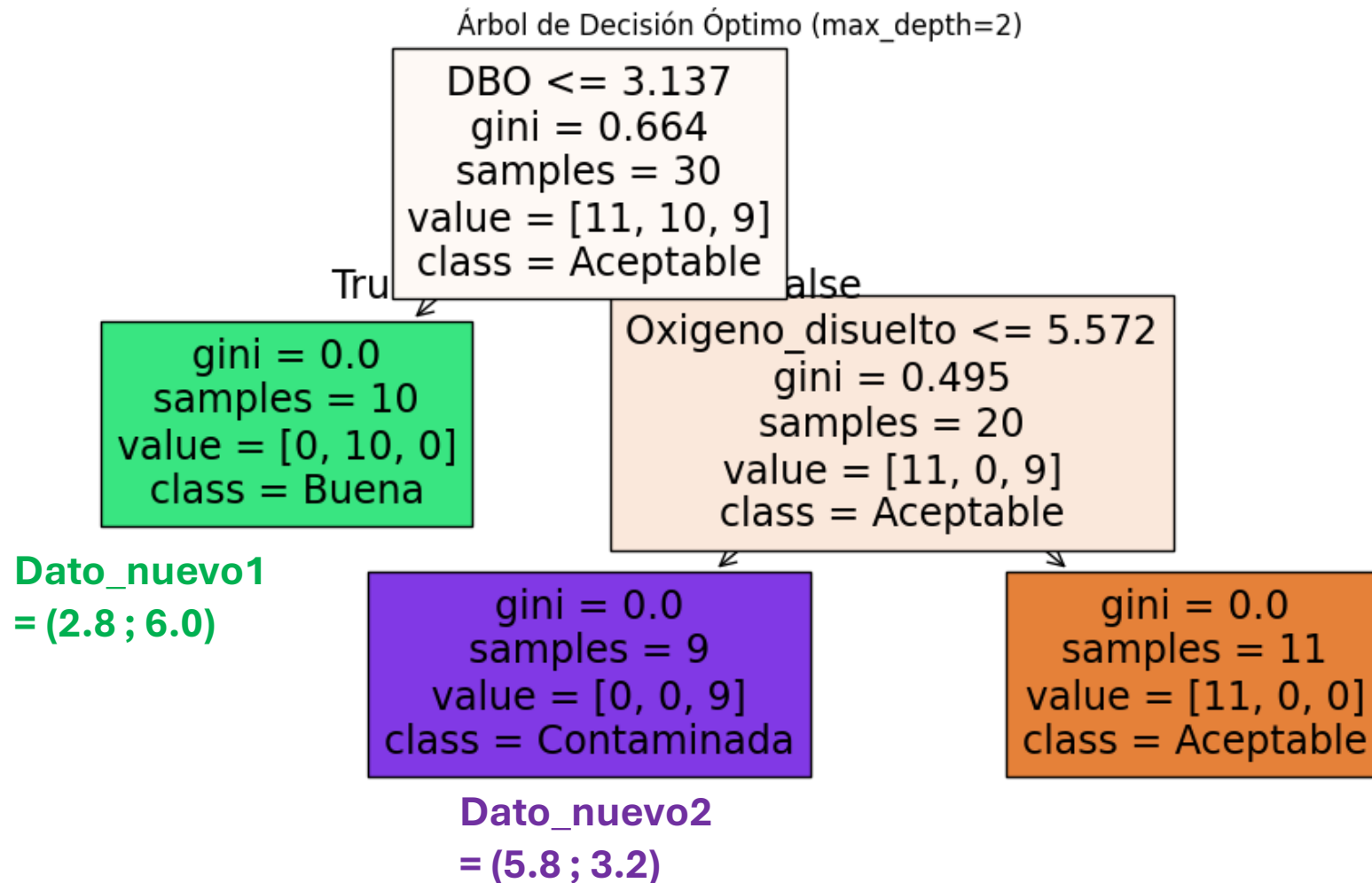
▼ Cálculo del Arbol de Decisión con Python

```
1 from sklearn.tree import plot_tree, DecisionTreeClassifier
2
3 # Variables predictoras y objetivo
4 X = df_agua[['DBO', 'Oxigeno_disuelto']]
5 y = df_agua['Calidad_Agua']
6
7 # Entrenar el mejor árbol de decisión con toda la data
8 mejor_arbol = DecisionTreeClassifier(criterion='gini', max_depth=2, random_state=42)
9 mejor_arbol.fit(X, y)
10
11 # Visualizar el árbol
12 plt.figure(figsize=(12, 6))
13 plot_tree(mejor_arbol, feature_names=X.columns, class_names=mejor_arbol.classes_, filled=True)
14 plt.title("Árbol de Decisión Óptimo (max_depth=2)")
15 plt.show()
```



Modelos Estadísticos: Modelos de Aprendizaje Supervisado

Predecir: Con modelo ya ENTRENADO



Nuevo Dato:

- DBO = 2.8 mg/L
- Oxígeno disuelto = 6.0 mg/L

Dato_nuevo1 = (2.8 ; 6.0)

Dato_nuevo2 = (5.8 ; 3.2)

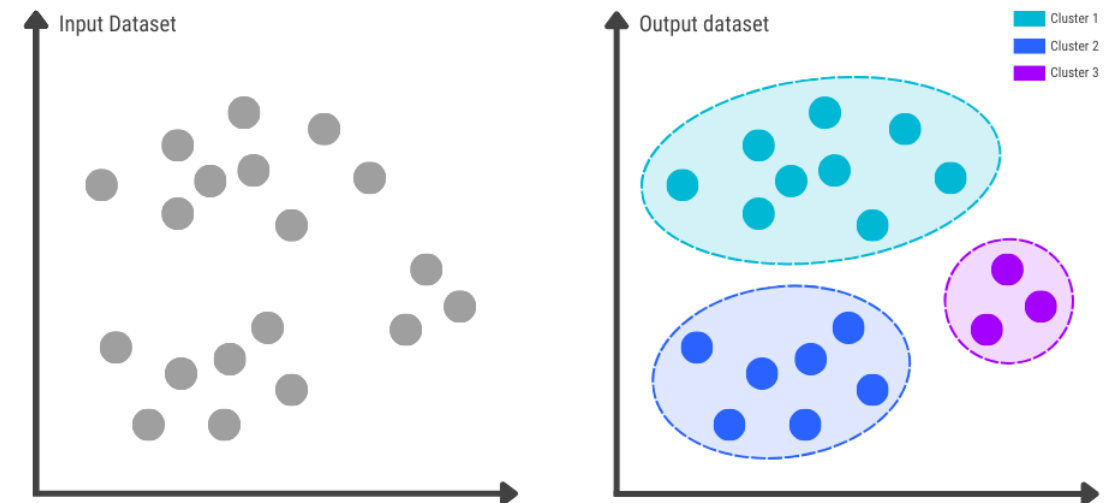
Modelos Estadísticos: Modelos de Aprendizaje **NO** Supervisado

Clustering / K-means

¿Qué es el Clustering?

El **clustering** o **agrupamiento** es una técnica de aprendizaje no supervisado que permite **identificar patrones o grupos naturales en los datos**, sin necesidad de etiquetas previas.

- Identificar perfiles de usuarios.
- Clasificar regiones con características similares (como estaciones de monitoreo).
- Detectar comportamientos anómalos (outliers).
- Agrupar series temporales, resultados de sensores, etc.



Modelos Estadísticos: Modelos de Aprendizaje **NO** Supervisado

Clustering / K-means

Clasificación del Clustering

Existen varios tipos de algoritmos de clustering, los más comunes son:

Tipo de algoritmo	Ejemplo	Características
Clustering Particional	K-Means	Divide en K grupos predefinidos. Simple y eficiente.
Clustering Jerárquico	Agglomerative Clustering	Construye un árbol de agrupaciones (dendrograma). No requiere definir K.
Clustering Basado en Densidad	DBSCAN	Agrupar puntos densamente conectados. Detecta outliers.
Modelos de Mezcla	Gaussian Mixture Models (GMM)	Asume que los datos provienen de varias distribuciones probabilísticas.

Modelos Estadísticos: Modelos de Aprendizaje **NO** Supervisado

Clustering / K-means

¿Qué hace K-Means?

K-Means busca **particionar un conjunto de datos en K grupos (clusters)**, de manera que los puntos dentro de cada grupo sean lo más similares entre sí y lo más diferentes posible a los de otros grupos. La similitud se mide normalmente con la distancia euclidiana.

Paso 1: Normalización de datos (preprocesamiento)

Paso 2: Inicialización de centroides (Clusters = k)

Paso 2.1: ¿Cómo elegir K ?

Paso 3: Asignación de puntos al cluster más cercano

Paso 4: Recalcular centroides

Paso 5: Repetir pasos 3 y 4 hasta convergencia



Modelos Estadísticos: Modelos de Aprendizaje **NO** Supervisado

Clustering / K-means

Paso 1: Normalización de datos (preprocesamiento)

Antes de aplicar K-Means, es fundamental que todas las variables estén en la **misma escala**. Por ejemplo, si temperatura está en °C y precipitación en mm, una dominará a la otra.

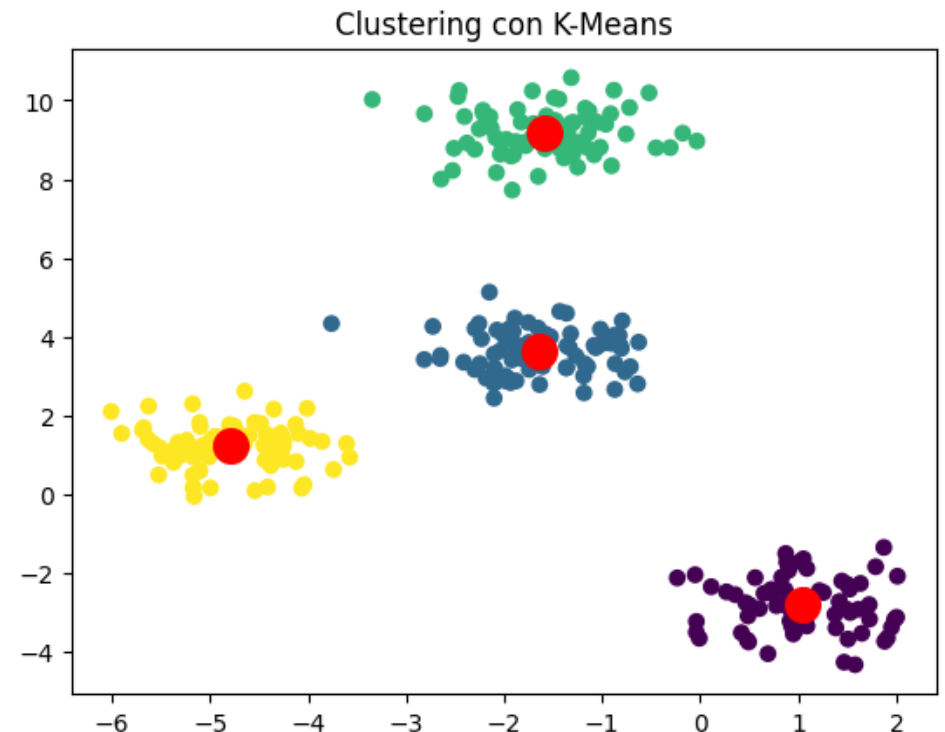
Estandarización (Z-score):

$$Z_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$$

X_{ij} : valor de la estación i en la variable j

μ_j : media de la variable j

σ_j : desviación estándar de la variable j



Modelos Estadísticos: Modelos de Aprendizaje **NO** Supervisado

Clustering / K-means

Paso 2: Inicialización de centroides

Se seleccionan aleatoriamente K puntos como centroides iniciales. Cada uno representará el "centro" de un cluster.

¿Cómo elegir K?

El algoritmo K-Means necesita que tú elijas cuántos grupos (K) deseas formar. Si eliges muy pocos, los grupos serán muy amplios y poco informativos; si eliges demasiados, habrá sobreajuste y perderás interpretabilidad.

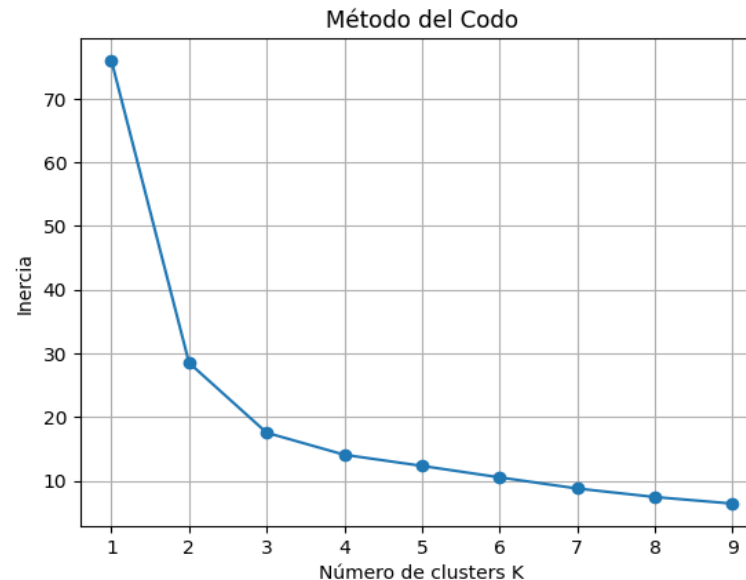
- **Método del codo (elbow method)**

¿Qué mide el método del codo?

Se basa en la Suma de Errores al Cuadrado (SSE):

$$SEE = \sum_{i=1}^n \|x_i - \mu_c\|^2$$

- x_i : es un punto de datos.
- μ_c : es el centroide de su cluster.
- $\| \|^2$: es la distancia euclidiana al cuadrado.



¿Dónde está el "codo"?

- A medida que se aumenta K, la SSE disminuye (porque más grupos = menos error).
- Pero hay un punto donde la mejora ya no es significativa.
- Ese punto de cambio en la pendiente es lo que se llama "el codo".

Modelos Estadísticos: Modelos de Aprendizaje **NO** Supervisado

Clustering / K-means

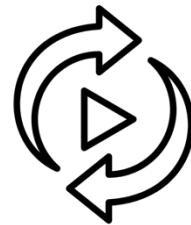
Paso 3: Asignación de puntos al cluster más cercano

- Para cada punto de datos, se calcula su **distancia euclidiana** a cada centroide.
- Y se asigna al cluster con el centroide más cercano.

Paso 4: Recalcular centroides

Una vez que cada punto pertenece a un cluster, se recalcula el centroide de cada grupo como la media de los puntos asignados a un cluster específico:

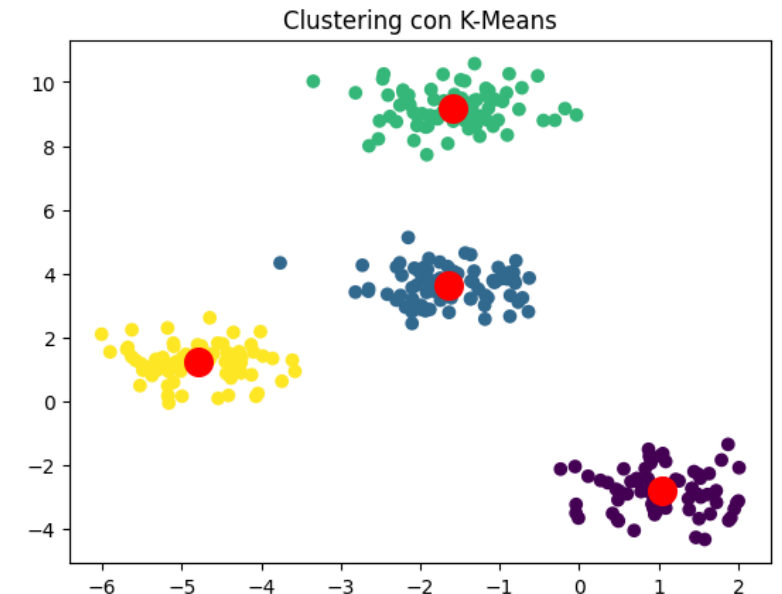
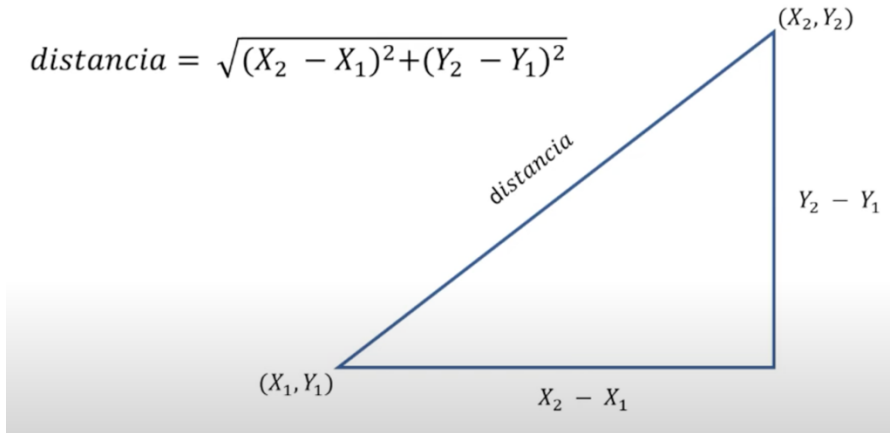
$$\mu_K = \frac{1}{n} \sum x$$



Paso 5: Repetir pasos 3 y 4 hasta convergencia:

- Los centroides ya no cambian significativamente.
- O se alcanza un número máximo de iteraciones.

Distancia euclidiana



Modelos Estadísticos: Modelos de Aprendizaje **NO** Supervisado

Clustering / K-means

Ejercicio:

Este ejercicio utiliza registros georreferenciados de atención médica por enfermedades respiratorias en Bogotá durante los años 2021 y 2022, con énfasis en casos de *rinofaringitis aguda (resfriado común)*. A través del algoritmo de clustering K-Means, se identifican zonas de la ciudad donde se concentran casos similares, lo que permite reconocer posibles patrones espaciales de ocurrencia y aportar evidencia para la priorización de intervenciones en salud pública.

Preguntas a responder:

- ¿Cuáles son las zonas de Bogotá con mayor concentración de casos similares?
- ¿Cuál es el promedio de edad por cluster?
- ¿Cuál es la distribución de Género por cluster?

