



Bryan Acevedo - Daniel Toribio

# Aprendizaje supervisado y procesamiento de datos



# Aprendizaje Supervisado



# Definición

Algoritmo que se entrena para **predecir o clasificar** datos basándose en ejemplos etiquetados.

Necesita un conjunto de **datos de entrenamiento** que consta de entradas (**características**) y las salidas deseadas (**etiquetas**)

Aprende a hacer predicciones o tomar decisiones basadas en estas etiquetas.



# Set De Entrenamiento

id	Attribute 1	Attribute 2	Attribute 3	Attribute 4
1	1	2	2	2
2	2	3	1	3
3	3	4	3	1
4	4	5	2	3

Etiqueta
1
2
3
4



# Set De Testeo

id	Attribute 1	Attribute 2	Attribute 3	Attribute 4
1	1	2	2	2
2	2	3	1	3
3	3	4	3	1
4	4	5	2	3



# Set De Testeo

El modelo toma este set de testing y hace una **predicción**

id	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Predicción
1	1	2	2	2	1
2	2	3	1	3	2
3	3	4	3	1	3
4	4	5	2	3	4

# Matriz de Confusión



		Actual Values	
		Yes	No
Predicted Values	Yes	True Positive TP	False Positive FP
	No	False Negative FN	True Negative TN



# Accuracy

Proporción de **predicciones correctas** que hace un modelo  
(en relación al número total de predicciones hechas)

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$





# Precisión

Mide la capacidad de un modelo para hacer predicciones **positivas correctas** en relación con **todas las predicciones positivas** realizadas.

Es útil cuando se desean minimizar los falsos positivos (o maximizar los VP), ya que se enfoca en la calidad de las predicciones positivas.

$$\text{Precision} = \text{TP} / ( \text{TP} + \text{FP} )$$



# Recall o Sensibilidad

Mide la capacidad de un modelo para **identificar todos los positivos** en un conjunto de datos

→ De todos los valores positivos originales ¿cuántos fueron correctamente clasificados?

Es útil cuando se trata de problemas en los que los falsos negativos (omisiones) son costosos o críticos (como en casos médicos).

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FN})$$

# Cross-Validation



En la **validación cruzada** se debe dividir el conjunto de datos en varias particiones y realizar múltiples iteraciones de entrenamiento y evaluación del modelo. La validación cruzada se lleva a cabo de la siguiente manera:

1. El conjunto de datos se divide en  $k$  particiones de aproximadamente el mismo tamaño.
2. El modelo se **entrena** en  $k-1$  particiones y se **evalúa** en la partición restante. Esto se repite  $k$  veces, de manera que cada una de las  $k$  particiones se utilice como **conjunto de prueba** exactamente una vez. Se calcula una métrica de rendimiento (ej *accuracy*) para cada evaluación.
3. Finalmente, se calcula el **promedio** de las métricas de rendimiento obtenidas en cada partición. Esto proporciona una **estimación del rendimiento general del modelo** en el conjunto de datos.

# Cross-Validation



## 4-fold validation (k=4)





# Procesamiento de datos

**Limpieza de datos:** Eliminar datos ruidosos (*outliers*), faltantes o inconsistentes. Es importante **tratar los valores nulos**.

**Normalización y estandarización:** Ajustar las escalas y unidades de las características para que sean comparables.

**Codificación de variables categóricas:** Si los datos contienen variables categóricas, es necesario convertirlas en representaciones numéricas (por ejemplo si hay una columna con “Verdadero” y “Falso”, convertirlo a 1 y 0)

**Selección de características:** Identificar y seleccionar las características más relevantes para el problema.



# Procesamiento de datos

**Reducción de la dimensionalidad:** En casos en los que el número de características sea alto, se puede realizar reducción de dimensionalidad.

**Manejo de desequilibrio de clases:** Ajustar si las clases objetivas están desequilibradas.

**Gestión de valores atípicos:** Identificar y gestionar valores atípicos.