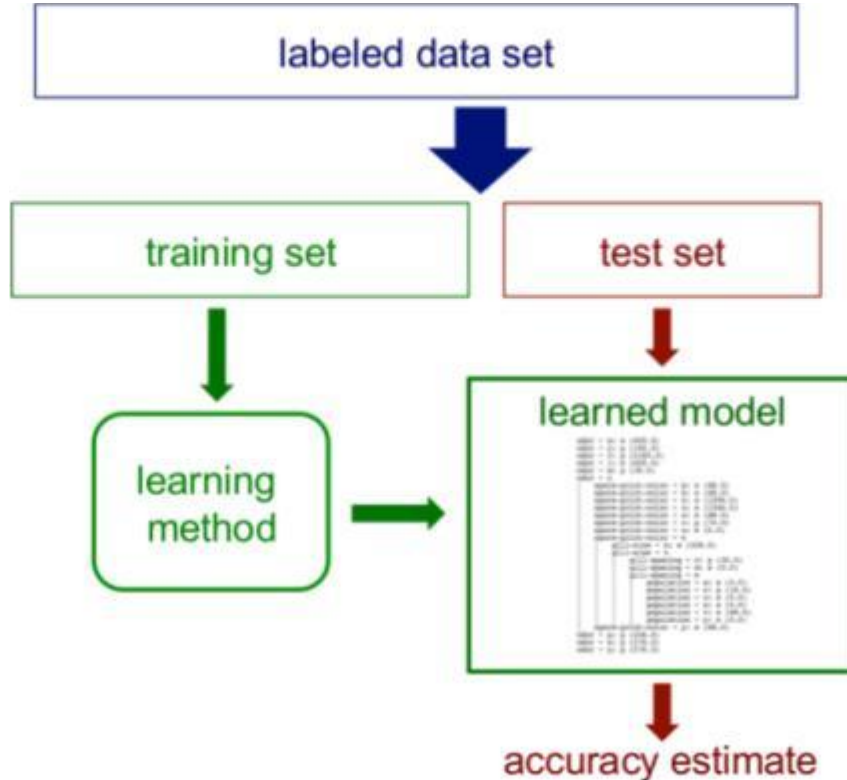


Técnicas de Validación. Métricas y medidas de performance



Conjunto de test

¿ Cómo obtener una estimación insesgada de la performance del modelo?



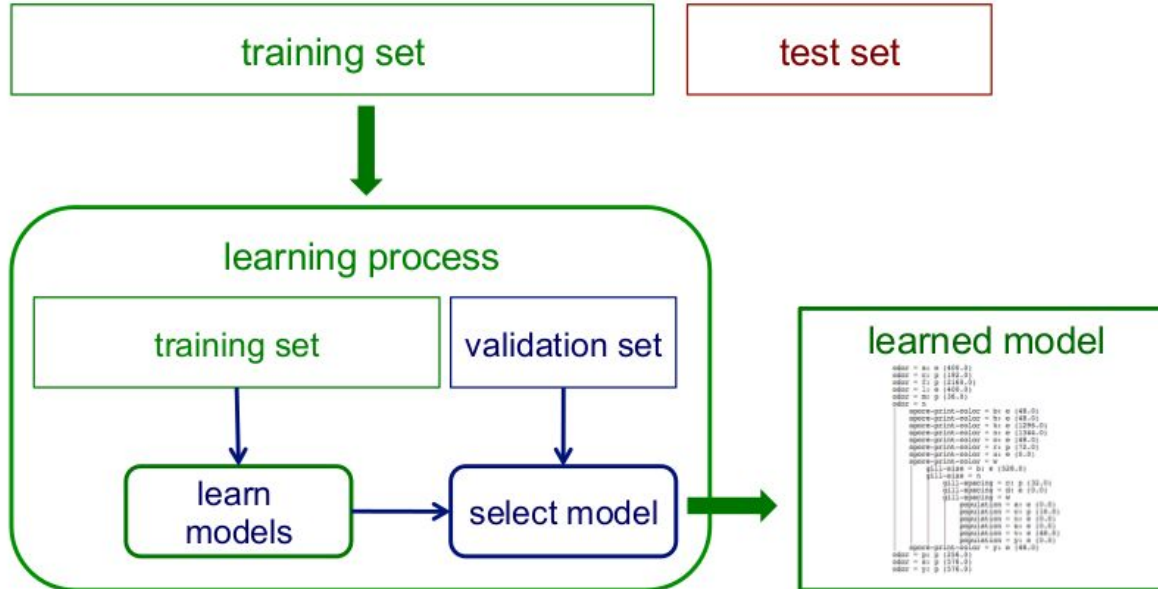
Conjunto de test

¿ Cómo obtener una estimación insesgada de la performance del modelo?

- Durante el aprendizaje el modelo no debe acceder bajo ningún motivo a datos del conjunto de test.
- Si las anotaciones en el conjunto de test influyen de **cualquier manera** el aprendizaje, las estimaciones de performance estarán sesgadas.

Conjunto de validación

¿ Cómo obtener una estimación insesgada de la performance del modelo durante el entrenamiento? (ajuste de hiperparámetros)

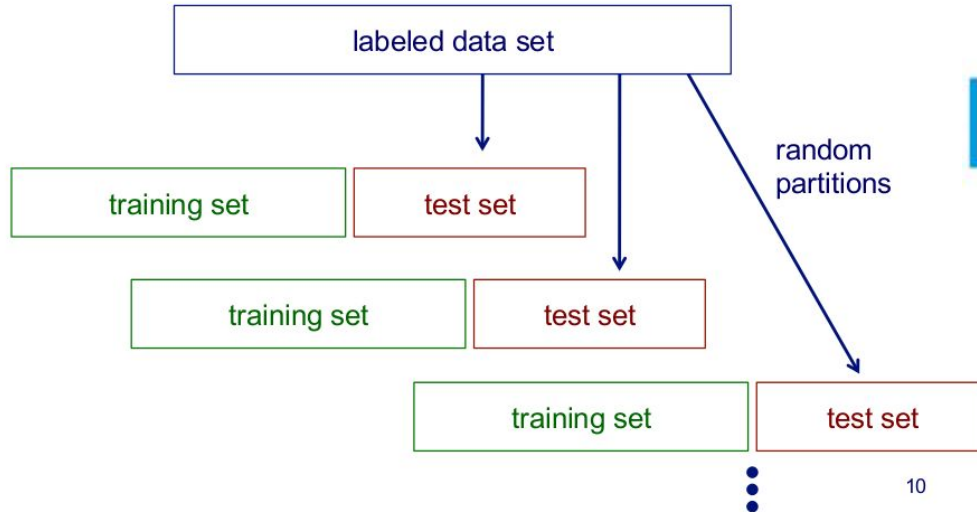


Limitación de usar sólo un conjunto de train/test

- Los datos pueden ser insuficientes para crear conjuntos de entrenamiento y test lo suficientemente grandes.
 - Un conjunto de test grande nos da una mejor medida de la performance del modelo y menos variable del rendimiento del modelo.
 - Un conjunto de entrenamiento grande es más representativo del universo de entradas posibles y permite aprender y generalizar a datos nuevos.
- Un solo conjunto de entrenamiento no nos da información sobre la sensibilidad del modelo ante cambios en los conjuntos de entrada.

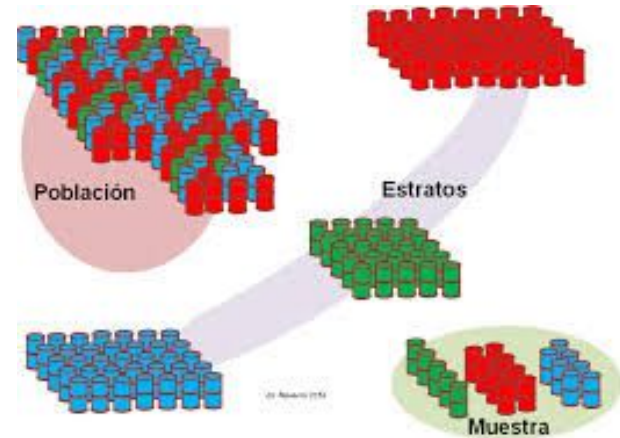
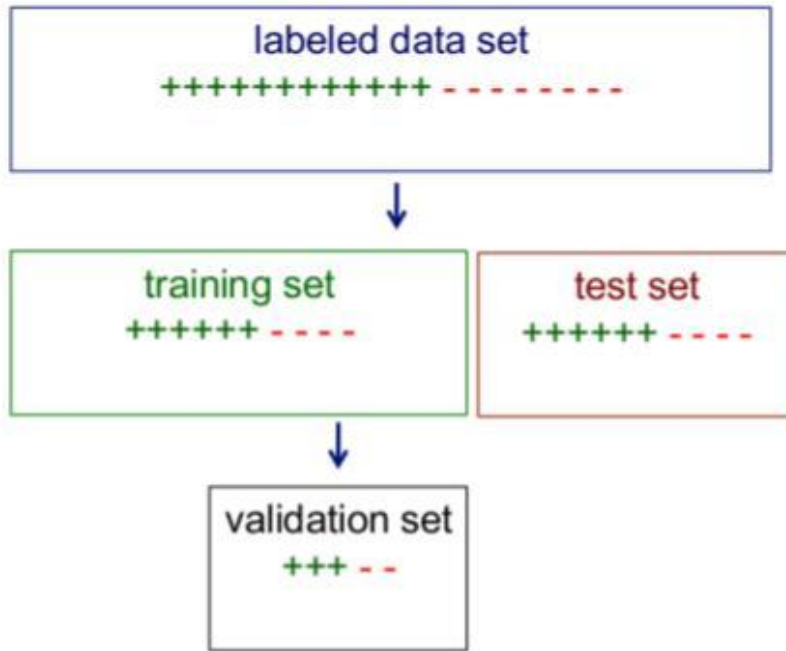
Remuestreo aleatorio

- podemos abordar el segundo punto mediante remuestreo



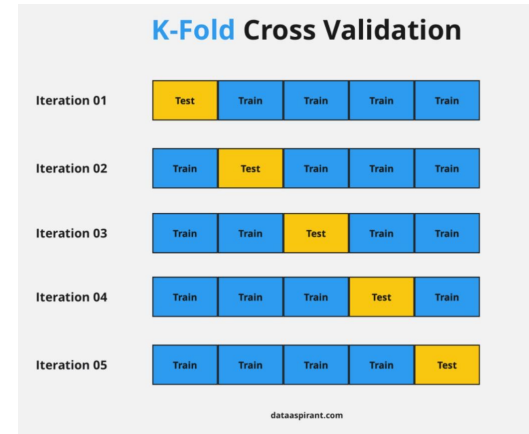
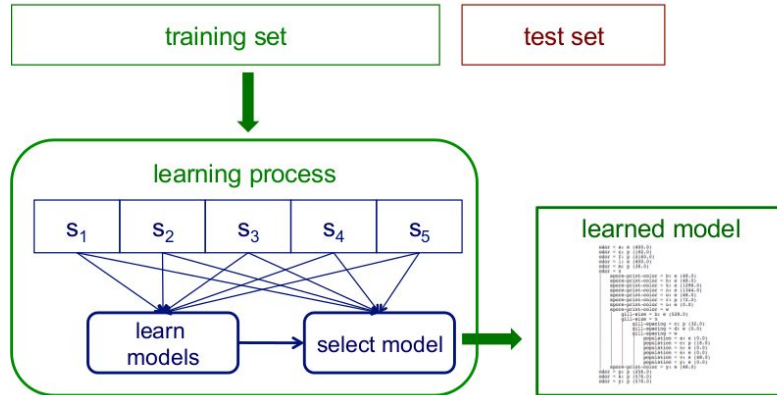
Muestreo estratificado

- Podemos requerir que las proporciones de clases se mantengan en cada subconjunto



Validación cruzada (cross validation)

- podemos considerar conjuntos de validación independientes y obtener una estimación respecto de la sensibilidad



Demo con notebook

06 Selección de Modelos.ipynb

Métricas (clasificación)

Función de costo vs métricas

- La función de costo es solo un proxy al problema en el mundo real
- Las métricas ayudan a capturar objetivos reales en forma cuantitativa
(no todos los errores son iguales)
- Ayudan a la organización del trabajo de los equipos en función de los requerimientos del problema
- Permiten cuantificar diferencias en:
 - performance deseada vs modelo base
 - performance deseada vs actual
 - evolución del tiempo
- Deberían ser el objetivo del entrenamiento, pero a veces es difícil

Clasificación binaria

- Entrada: x , salida y (valores 0/1 0 -1/+1)
- Predicción del modelo: $\hat{y} = h(x)$
- Dos tipos de modelos:
 - Modelos que predicen directamente una variable categórica (kNN, árboles de decisión)
 - Modelos que predicen un puntaje (score) (SVM, regresión logística)
 - Se necesita elegir un umbral (func. de decisión)
 - Nos enfocaremos en esta última clase de modelo. Los anteriores se pueden ver como un caso especial

Modelos basados en score

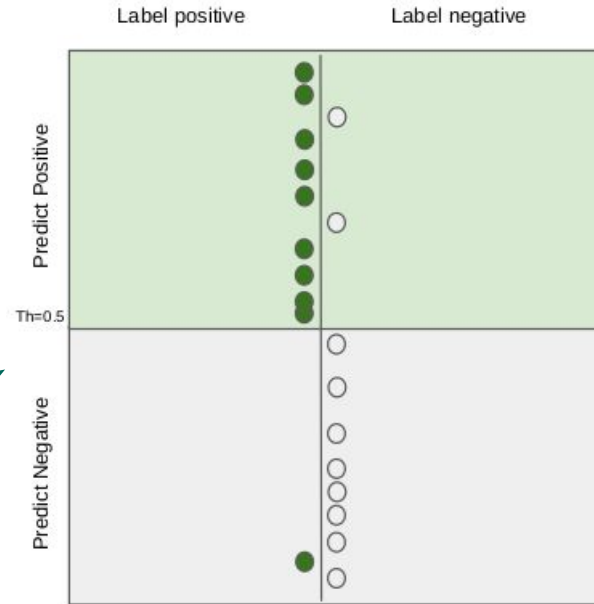
Score = 1



Score = 0

●	Positive example
○	Negative example

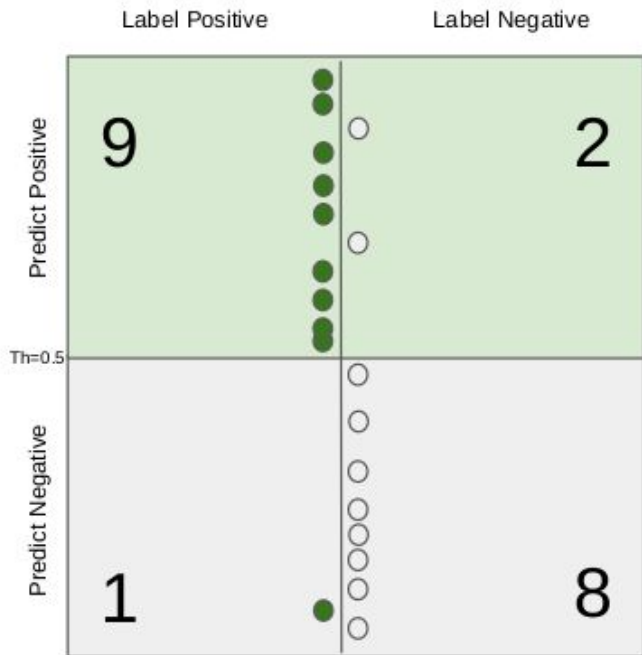
Matriz de confusión: es una tabla de contingencia de los valores observados (clases observadas) versus las clases predichas por el modelo



Th

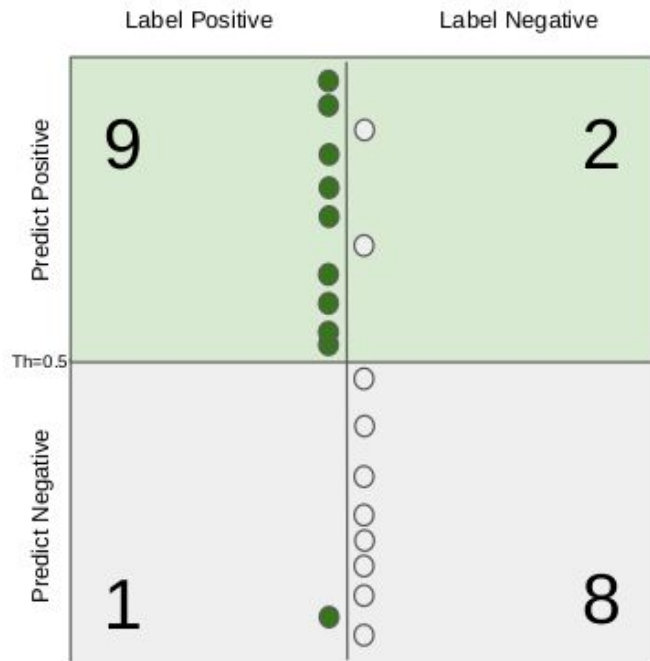
0.5

Matriz de confusión



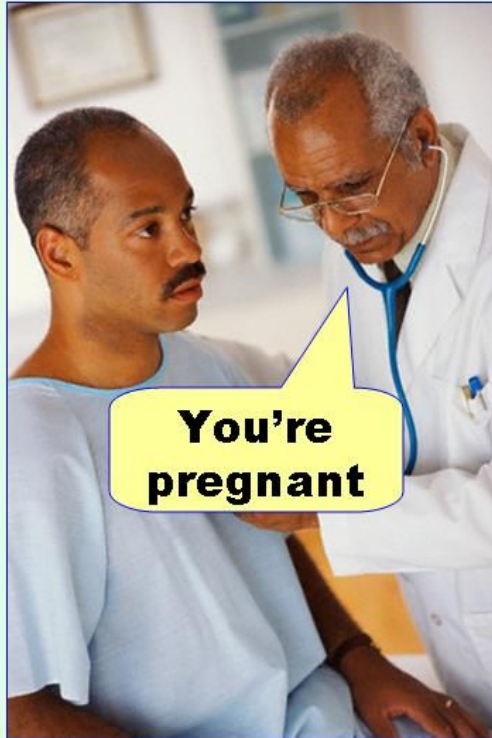
- la suma total de todos los valores es fija (total de la muestra)
- la suma por columnas es fija (muestras por clase)
- la calidad del modelo y el valor del umbral de deciden el agrupamiento de filas
- queremos que los elementos diagonales tengan valores grandes y lo nos diagonales valores chicos

Matriz de confusión



- true positives (TP)=9
- true negatives (TN)=8
- false positives (FP)=2
- false negatives (FN)=1

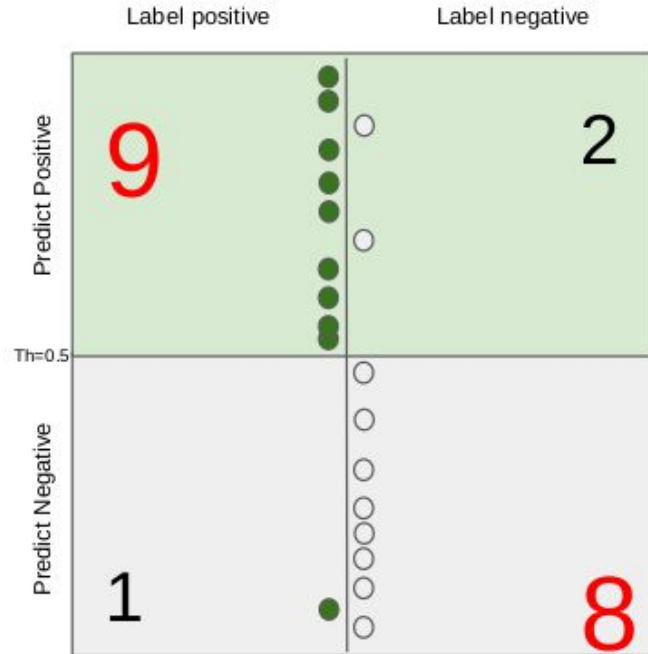
Type I error
(false positive)



Type II error
(false negative)



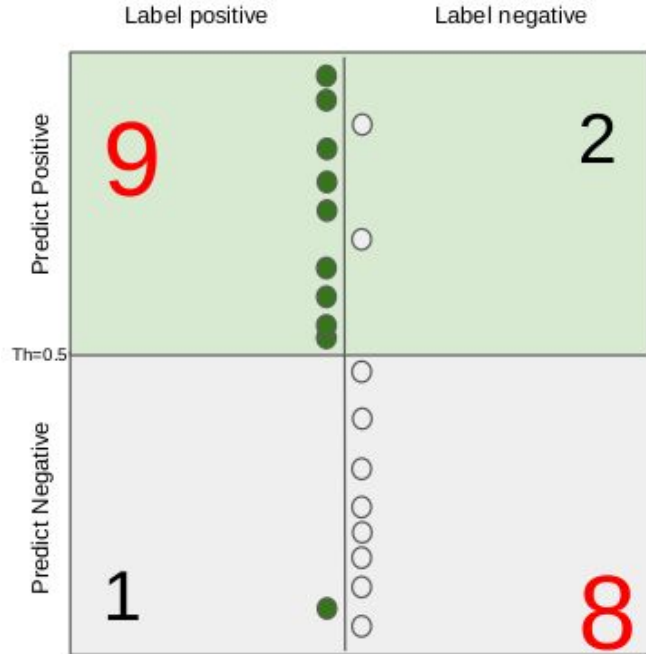
Métricas puntuales: exactitud (accuracy)



Th	TP	TN	FP	FN	Acc
0.5	9	8	2	1	0.85

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{N}$$

Métricas puntuales: precisión

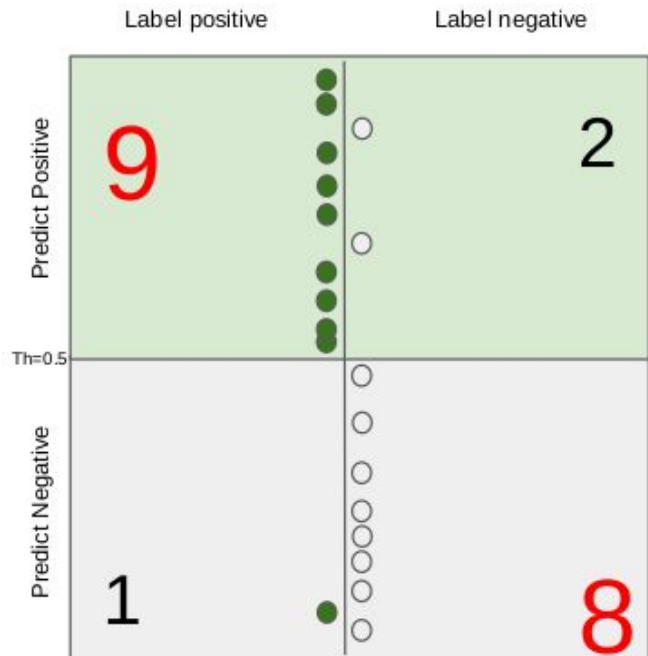


Th	TP	TN	FP	FN	Acc	Pr
0.5	9	8	2	1	0.85	0.81

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- De todas las instancias predichas como positivas por el modelo, ¿cuántas son realmente positivas?
- Prec 100%= todos bajo el umbral salvo el de score más alto (siempre que sea correcto)

Métricas puntuales: sensibilidad (recall)

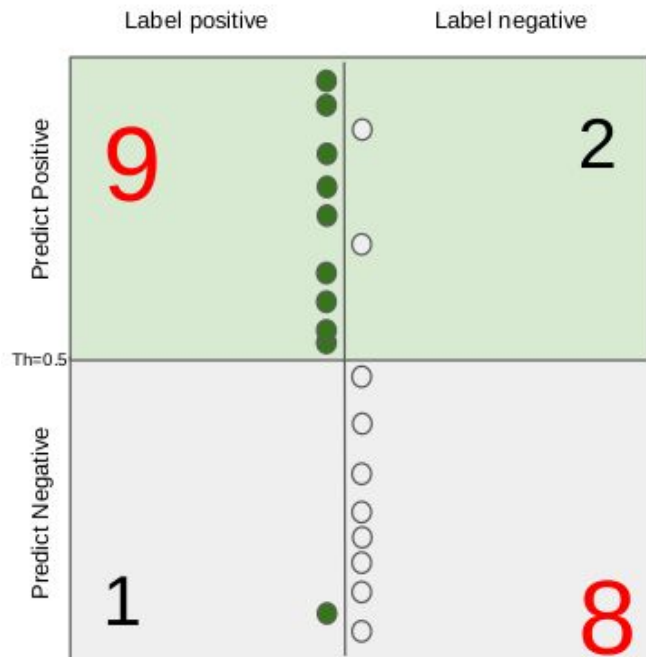


Th	TP	TN	FP	FN	Acc	Pr	Recal
0.5	9	8	2	1	0.85	0.81	0.9

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- De todas las instancias que son realmente positivas, ¿cuántas fueron correctamente identificadas por el modelo?
- Recall 100%= todos los puntos por encima del umbral

Métricas puntuales: F1-score



Th	TP	TN	FP	FN	Acc	Pr	Recal	F1
0.5	9	8	2	1	0.85	0.81	0.9	0.857

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

- **Valor:** El F1-Score varía entre 0 y 1, donde 1 es el mejor valor posible, indicando una precisión y sensibilidad perfectas, y 0 es el peor valor posible.
- **Equilibrio:** Penaliza más fuertemente valores extremos de una métrica en detrimento de la otra.
- **Uso:** donde se tiene un desequilibrio de clases

Métricas puntuales: cambio de umbral



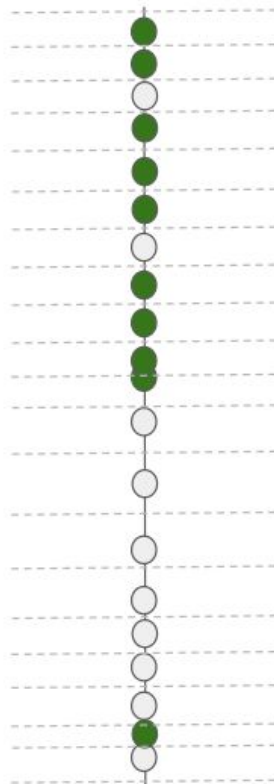
Th	TP	TN	FP	FN	Acc	Pr	Recal	F1
0.6	7	8	2	3	0.75	0.77	0.7	0.733

umbrales efectivos= # ejemplos +1

Threshold Scanning

Score = 1

Threshold = 1.00



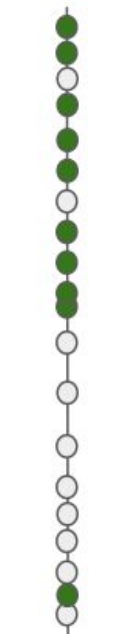
Threshold = 0.00

Score = 0

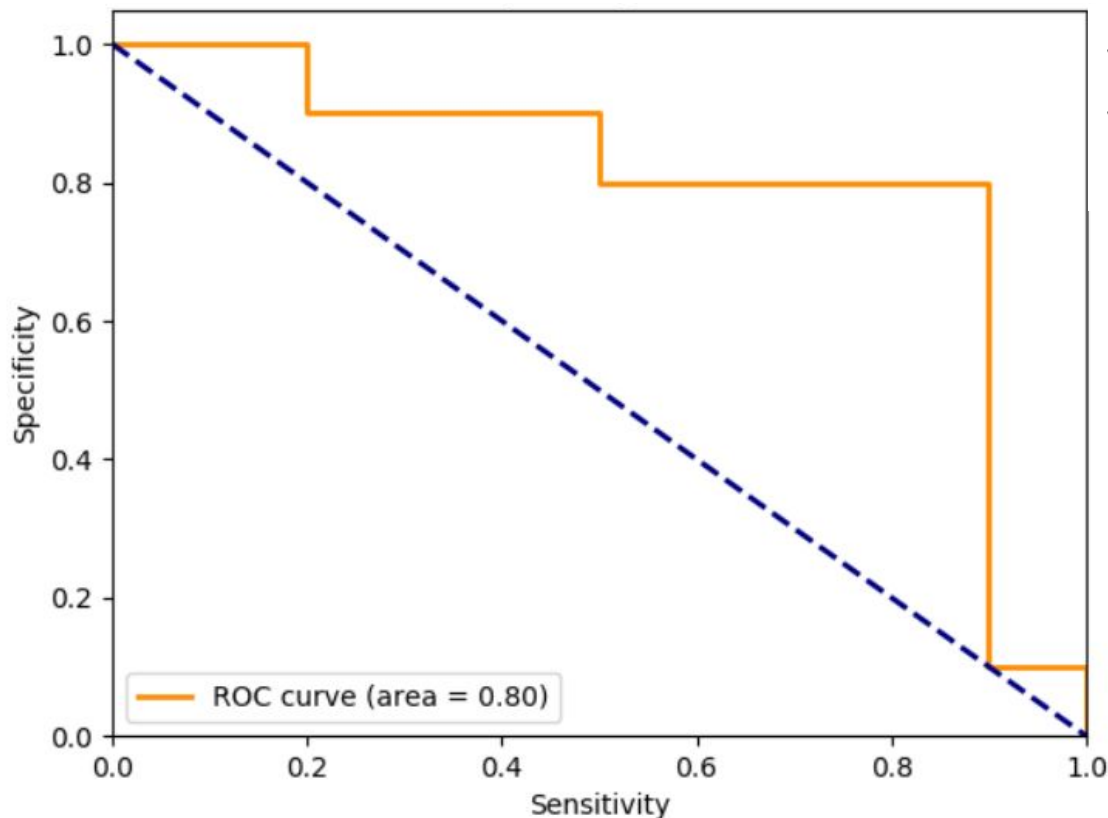
Threshold	TP	TN	FP	FN	Accuracy	Precision	Recall	Specificity	F1
1.00	0	10	0	10	0.50	1	0	1	0
0.95	1	10	0	9	0.55	1	0.1	1	0.182
0.90	2	10	0	8	0.60	1	0.2	1	0.333
0.85	2	9	1	8	0.55	0.667	0.2	0.9	0.308
0.80	3	9	1	7	0.60	0.750	0.3	0.9	0.429
0.75	4	9	1	6	0.65	0.800	0.4	0.9	0.533
0.70	5	9	1	5	0.70	0.833	0.5	0.9	0.625
0.65	5	8	2	5	0.65	0.714	0.5	0.8	0.588
0.60	6	8	2	4	0.70	0.750	0.6	0.8	0.667
0.55	7	8	2	3	0.75	0.778	0.7	0.8	0.737
0.50	8	8	2	2	0.80	0.800	0.8	0.8	0.800
0.45	9	8	2	1	0.85	0.818	0.9	0.8	0.857
0.40	9	7	3	1	0.80	0.750	0.9	0.7	0.818
0.35	9	6	4	1	0.75	0.692	0.9	0.6	0.783
0.30	9	5	5	1	0.70	0.643	0.9	0.5	0.750
0.25	9	4	6	1	0.65	0.600	0.9	0.4	0.720
0.20	9	3	7	1	0.60	0.562	0.9	0.3	0.692
0.15	9	2	8	1	0.55	0.529	0.9	0.2	0.667
0.10	9	1	9	1	0.50	0.500	0.9	0.1	0.643
0.05	10	1	9	0	0.55	0.526	1	0.1	0.690
0.00	10	0	10	0	0.50	0.500	1	0	0.667

Métricas resumen : curvas ROC (rotada)

Score = 1



Score = 0



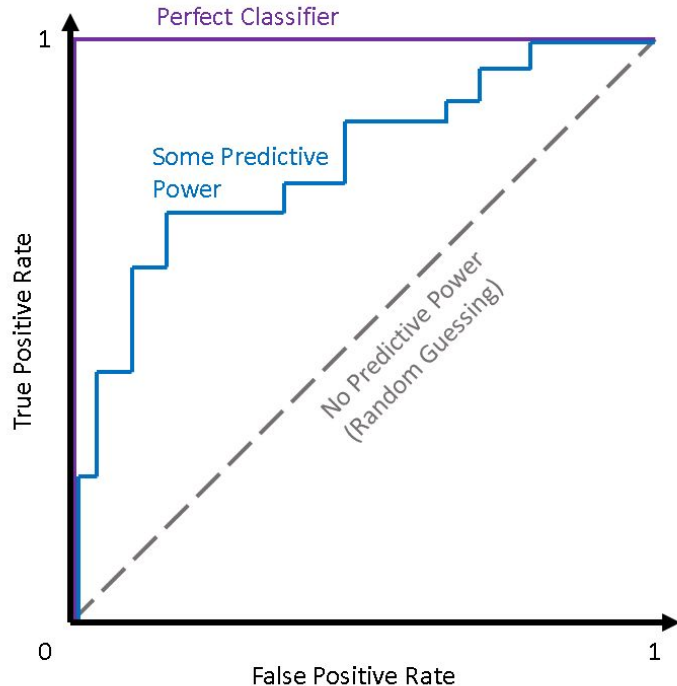
$specificity = tnr = \frac{TN}{Neg} = \frac{TN}{(TN+FP)}$

$sensitivity = tpr = \frac{TP}{Pos} = \frac{TP}{(TP+FN)}$

métrica AUC = área bajo la curva ROC

- **AUC = 1:** Representa un modelo perfecto.
- **AUC = 0.5:** Indica un modelo que no tiene capacidad de discriminación, equivalente a una clasificación aleatoria.
- **AUC < 0.5:** Indica un rendimiento peor que el azar

Métricas resumen : curvas ROC (rotada)



Uso Práctico de la Curva ROC

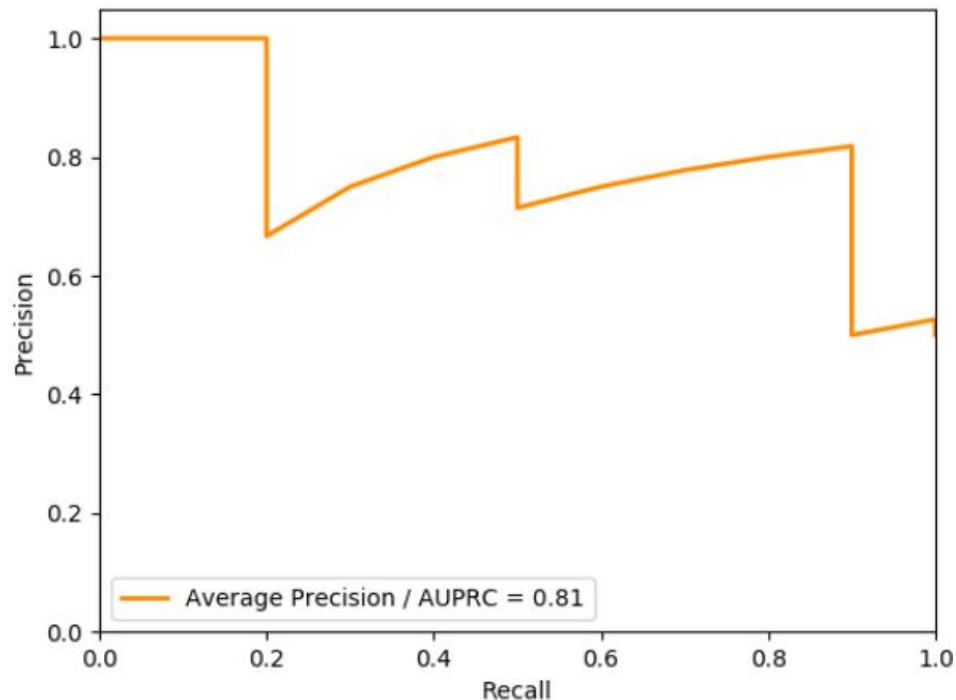
1. **Comparación de Modelos:** Las curvas ROC permiten comparar diferentes modelos de clasificación. Un modelo con una curva ROC más cerca del punto (0,1) y un AUC mayor es generalmente mejor.
2. **Selección del Umbral de Decisión:** La curva ROC puede ayudar a seleccionar un umbral de decisión que balancee la tasa de verdaderos positivos y la tasa de falsos positivos de acuerdo a las necesidades específicas del problema.
3. **Evaluación General del Modelo:** Proporciona una evaluación completa del rendimiento del modelo a lo largo

Métricas de resumen: curvas PR

Score = 1



Score = 0



precisión = $TP / (\text{pos predichos})$ (=VPP)

recall = $TP / (\text{pos verdaderos})$
(=SENSIBILIDAD)

AUPR = área bajo la curva PR

La curva de
Precisión-Recall es
especialmente útil
cuando nos interesa más
minimizar los falsos
positivos que los falsos
negativos

Curvas ROC y PR en validación cruzada

Opción 1:

- Asumir que magnitudes de los scores son comparables entre corridas
- Acumular predicciones de todas las corridas
- Trazar la curva usando predicciones acumuladas

Opción 2:

- Trazar las curvas individuales para cada partición
- Considerar la “curva promedio”

Resumen curvas ROC y PR

- Permiten evaluaciones cuantitativas a distintos niveles de “confianza”
- Asumen problemas binarios
- Se pueden resumir en medidas del tipo “área bajo la curva”
- Las curvas ROC son insensibles a cambios en la distribución de clases en el conjunto de test
- Las curvas PR muestran la fracción de las predicciones que son FP
- Las curvas PR son útiles en problemas con una proporción de muestras negativas muy alta
- Permiten determinar umbrales óptimos para distintos puntos de operación

Problemas multiclases

- Problemas con N clases \implies matriz de confusión $N \times N$
- La mayoría de las métricas se analizan como N problemas binarios (OVA)
 - El desbalance crece con el número de clases

		Estimate		
		$c_0 \dots c_{k-1}$	c_k	$c_{k+1} \dots c_n$
annotated ground truth	$c_0 \dots c_{k-1}$	TN	FP	TN
	c_k	FN	TP	FN
	$c_{k+1} \dots c_n$	TN	FP	TN

TN

 true negative

TP

 true positive

FN

 false negative

FP




 false positive

Problemas multiclase: macro avg

- Variantes multiclase de métricas AUC:
 - micros vs macro average
 - macro: calcula la métrica de rendimiento de cada clase y luego toma la media aritmética de todas las clases.
 - Útil cuando te interesa evaluar el rendimiento del modelo en cada clase por igual, sin importar el desbalance de clases.

$$\text{Precision}_{\text{Class A}} = \frac{TP_{\text{Class A}}}{TP_{\text{Class A}} + FP_{\text{Class A}}}$$




$$\text{Recall}_{\text{Class A}} = \frac{TP_{\text{Class A}}}{TP_{\text{Class A}} + FN_{\text{Class A}}}$$

Label	True Positive (TP)	False Positive (FP)	False Negative (FN)	Precision	Recall	F1 Score	Macro-Averaged F1 Score
 Airplane	2	1	1	0.67	0.67	$2 * (0.67 * 0.67) / (0.67 + 0.67) = \mathbf{0.67}$	$\frac{0.67 + 0.40 + 0.67}{3} = \mathbf{0.58}$
 Boat	1	3	0	0.25	1.00	$2 * (0.25 * 1.00) / (0.25 + 1.00) = \mathbf{0.40}$	
 Car	3	0	3	1.00	0.50	$2 * (1.00 * 0.50) / (1.00 + 0.50) = \mathbf{0.67}$	

Problemas multiclase: micro avg

- Útil cuando te interesa el rendimiento global del modelo, considerando todas las instancias por igual.
- Es especialmente relevante en situaciones donde las clases están desbalanceadas y quieres que el modelo tenga un buen rendimiento general

$$(F_1)_{micro} = 2 \cdot \frac{\text{precision}_{micro} \text{recall}_{micro}}{\text{precision}_{micro} + \text{recall}_{micro}} = \frac{\sum TP}{\sum (TP+FP)}$$

Label	True Positive (TP)	False Positive (FP)	False Negative (FN)	Micro-Averaged F1 Score
 Airplane	2	1	1	$\frac{TP}{TP + \frac{1}{2}(FP+FN)} = \frac{6}{6 + \frac{1}{2}(4+4)}$ = 0.60
 Boat	1	3	0	
 Car	3	0	3	
TOTAL	6	4	4	

Demo con notebook

07 Metricas.ipynb

—