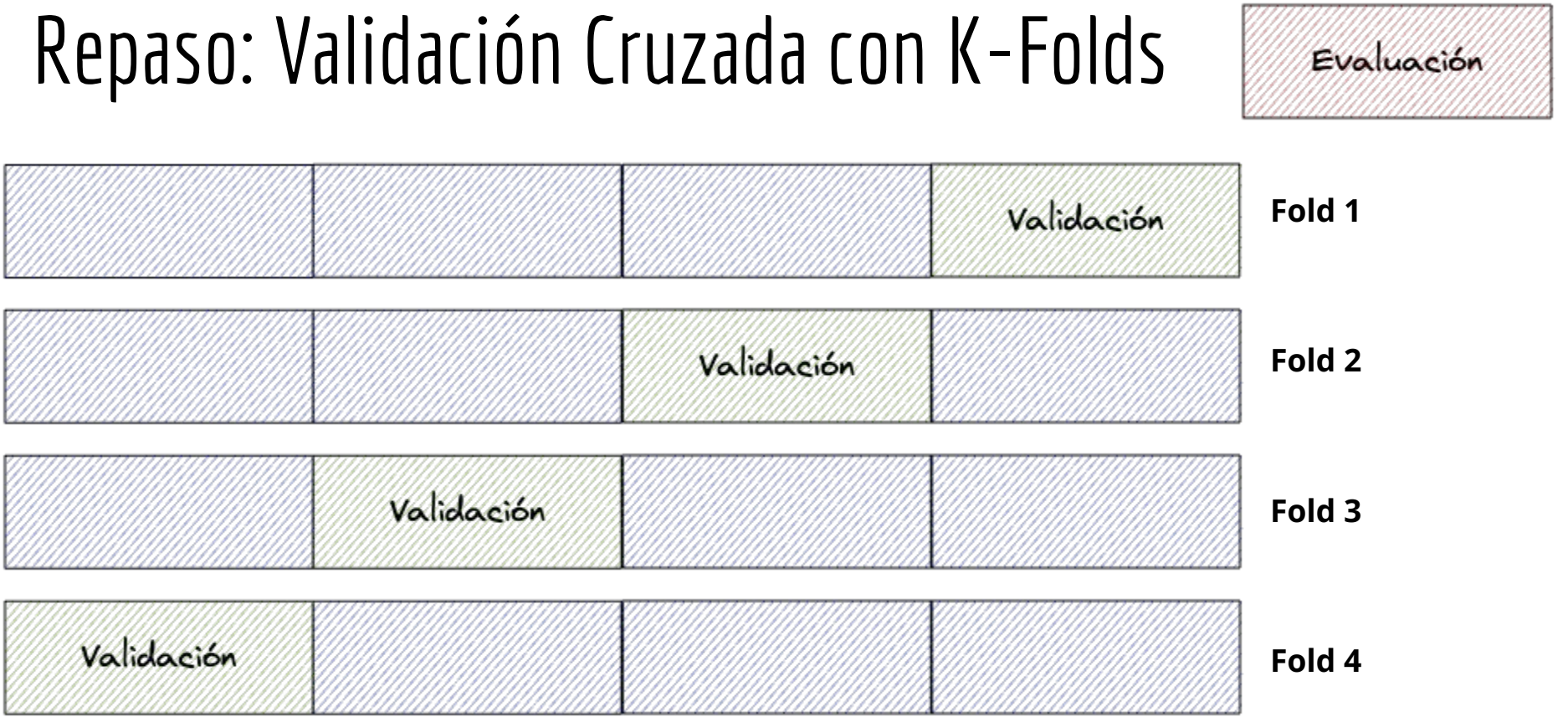
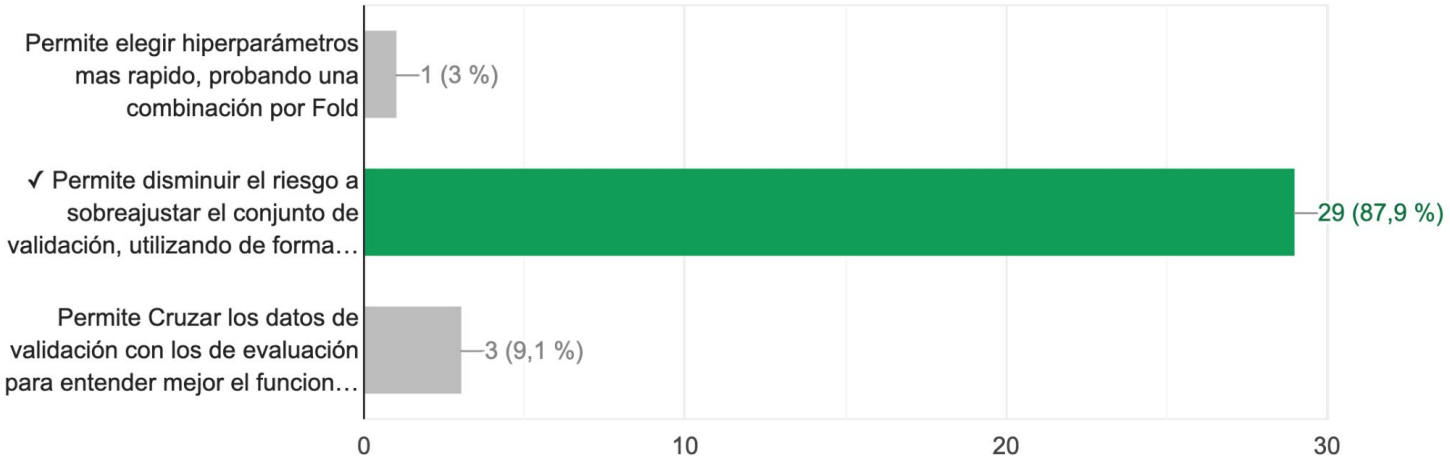
Repaso: Validación Cruzada con K-Folds



Form: Validación Cruzada

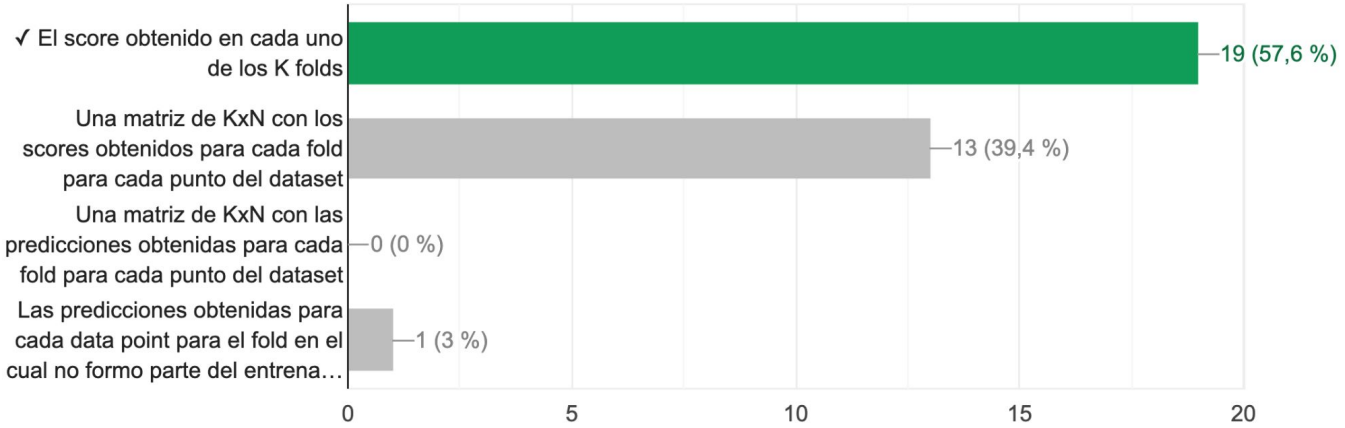
¿Que permite hacer la Validación Cruzada?

29 de 33 respuestas correctas



Que devuelve la función cross_val_score con K-Folds sobre un dataset de N samples?

19 de 33 respuestas correctas



IAA-2023
Clase 9: Clasificación Avanzada

UNSAM
UNIVERSIDAD
NACIONAL DE
SAN MARTÍN

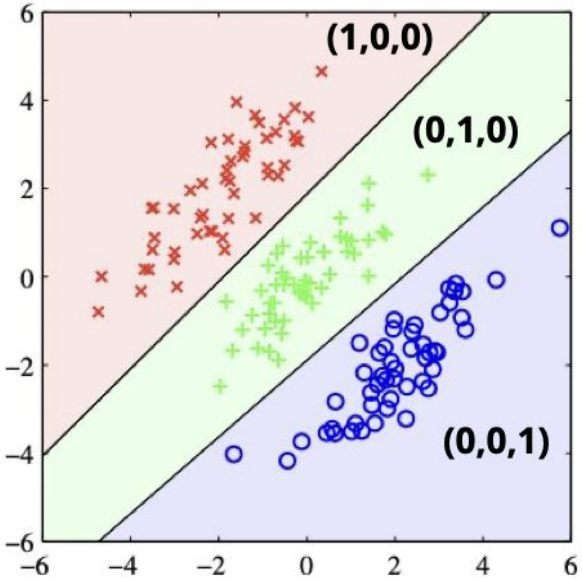
Repaso: Clasificación

Un modelo de clasificación se caracteriza porque el objetivo (*target*) de una muestra x es un mapeo a una clase C_k (k=1, ... K).

Para esto, lo primero a definir es cómo codificamos esta clase de destino en un número: Para un punto del dataset (x, C_k)

- Label Encoding:
  Asignamos un número entero a cada clase de destino (las numeramos)
  $$t = k$$

- One-Hot Encoding:
  Asignamos un vector de componentes nulas, salvo por la k-ésima que vale 1
  $$t = (0, \ldots, 0, 1, 0, \ldots, 0)$$



Repaso: Clasificación Binaria

Clasificación binaria es el caso particular en que hay solo 2 clases (K=2).
En este caso se suele usar label encoding ya que es equivalente al one-hot.

	Caso General	Caso Binario (K=2)
Label Encoding	$t \in [0, 1, \ldots, k-1]$	$t = 0$ ó 1
One-Hot Encoding	$t = (0,..,0,1,0,...0)$	$t = (1, 0)$ ó $(0,1)$

Métricas de Clasificación Binaria

La métrica que hemos estado utilizando para clasificación binaria es la *exactitud (accuracy),* que mide **el porcentaje de aciertos** de nuestro modelo.

Esta métrica tiene un sesgo hacia la clase mayoritaria, ¿qué significa esto?

Métricas de Clasificación Binaria

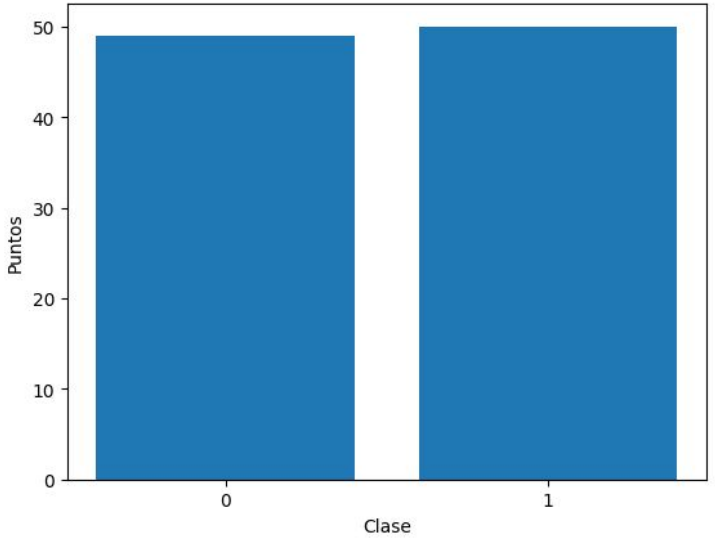
La métrica que hemos estado utilizando para clasificación binaria es la *exactitud (accuracy),* que mide **el porcentaje de aciertos** de nuestro modelo.

Esta métrica tiene un sesgo hacia la clase mayoritaria, ¿qué significa esto?

Imaginemos un escenario *balanceado*:

Tenemos la misma cantidad de puntos en cada clase.

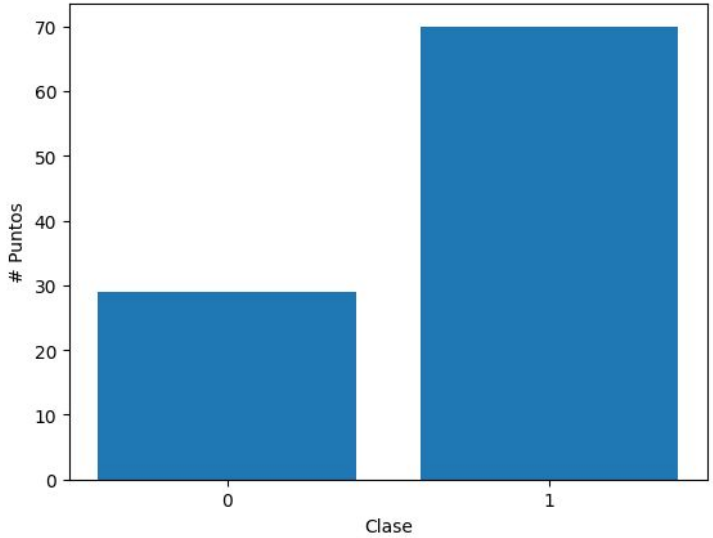
¿Cuál es el valor esperado para la exactitud de una clasificación *al azar*?



Métricas de Clasificación Binaria

Ahora imaginemos un caso *desbalanceado*: Hay muchos mas puntos en una de las dos clases.

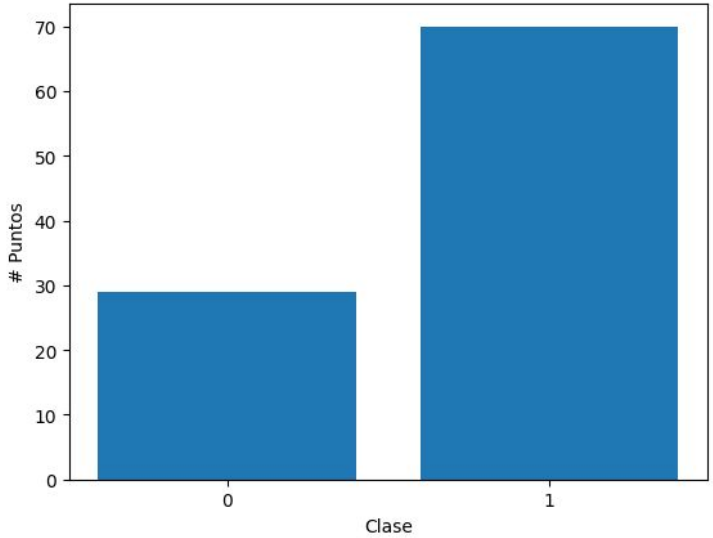
¿Cuál es el valor esperado para la exactitud de una clasificación *al azar*?



Métricas de Clasificación Binaria

Ahora imaginemos un caso *desbalanceado*: Hay muchos mas puntos en una de las dos clases.

¿Cuál es el valor esperado para la exactitud de una clasificación *al azar*?



¿Y si en vez del azar clasificamos todos como de la clase mayoritaria?

Métricas de Clasificación Binaria Desbalanceada

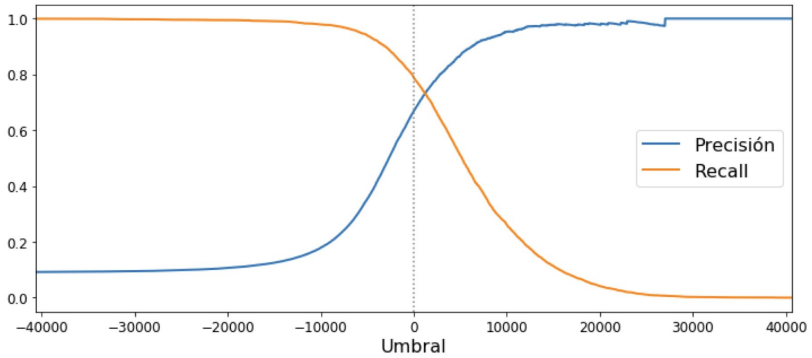
Para estos casos es que consideramos métricas especiales:

- Precisión: $\frac{\#(\text{Puntos correctamente clasificados como clase 1})}{\#(\text{Puntos clasificados como clase 1})}$

- Exhaustividad (*recall)*: $\frac{\#(\text{Puntos correctamente clasificados como clase 1})}{\#(\text{Puntos de la clase 1})}$

Métricas de Clasificación Binaria Desbalanceada

Sesgando mi modelo hacia una clase o la otra, puedo favorecer una u otra métrica.



Pero solo entrenando *mejor* puedo mejorar ambas

- F1-score: $\frac{1}{F_1} = \frac{1}{2}\left(\frac{1}{precision} + \frac{1}{exhaustividad}\right)$

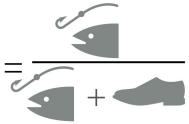
$$F_1 = \frac{2\, precision \cdot exhaustividad}{precision + exhaustividad}$$

Métricas de Clasificación Binaria Desbalanceada

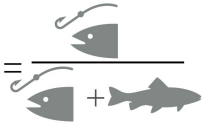
La información completa está dada por la ***matriz de confusión***

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

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



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



Predicciones

Target Real	0	1
0	TN	FP
1	FN	TP

Métricas de Clasificación Binaria Desbalanceada

Estrategias para lidiar con un dataset desbalanceado:

- Usar métricas adecuadas (nada de exactitud)

- Compensar el sesgo natural a la clase mayoritaria:
  - Dar mayor peso en la loss function a la clase minoritaria (class_weights)
  - Sobre-samplear la clase minoritaria (sklearn.utils.resample)

- Ajustar el umbral para maximizar la métrica de interés
  (más sobre esto la clase próxima)

Clasificación MultiClase

¿Cómo lidiar con la clasificación multiclase?
La forma estándar es entrenar muchos clasificadores binarios, usando diferentes estrategias:

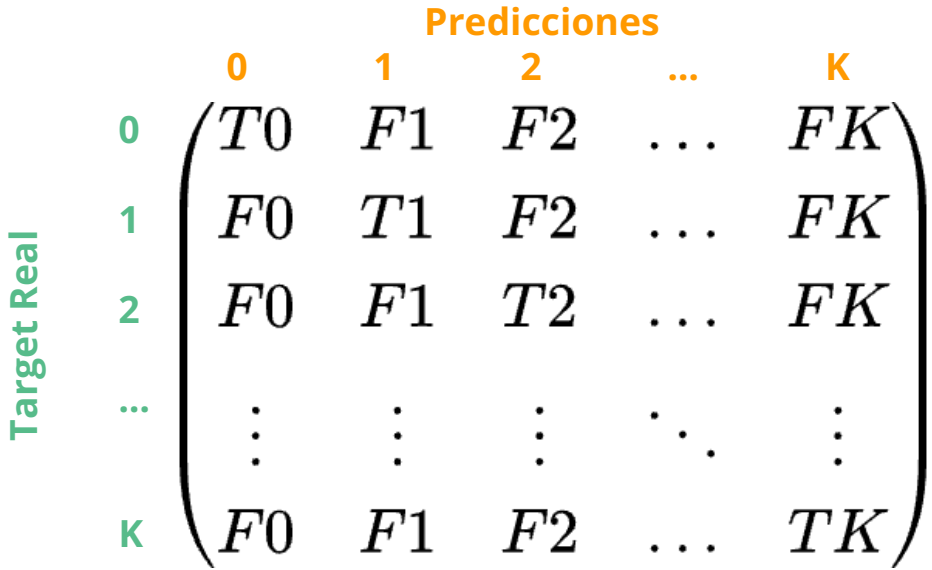
- **Uno-vs-el resto**
  Si tengo K clases, entreno K clasificadores: Cada uno clasifica si un sample es de la clase K o no. Luego clasifico con la clase de mayor probabilidad
  (y puedo normalizar las probabilidades para obtener una probabilidad multiclase)

- **Uno-vs-uno**
  Por cada par de clases, entreno un clasificador binario, es decir tengo K(K-1)/2 clasificadores. Luego clasifico a la clase que recibe mas votos.
  Contra: Escala como K^2.
  Pros: Cada clasificador sólo entrena en un subset (quizas balanceado) del dataset.

Clasificación MultiClase

¿Y qué pasa con las métricas?

- La exactitud sigue teniendo sentido, es el porcentaje de aciertos
- Tanto la exhaustividad como la precisión se definen una para cada clase.
- La matriz de confusión sigue teniendo toda la información relevante:



TP Final

Consideraciones

- La presentación del proyecto se hará en forma oral con uso de diapositivas, en las últimas dos semanas de cursada.

- Se suplementará la presentación con el envío por email del notebook / código fuente utilizado para obtener los resultados, junto a la presentación en PDF. De contar con varios archivos de código, estos deberan ser comprimidos en formato .zip.

- El envío por email es a la casilla ifabre@unsam.edu.ar o a cualquier otro docente. Deberá ser enviado antes del dia de la presentación.

Evaluación

La evaluación se hará en base a **la presentación oral.** Los siguientes puntos servirán como guía:

- Respetar las etapas del flujo de trabajo:
  - Presentación del problema y cómo se abordará usando ML
  - Presentación del dataset (de donde fue obtenido, créditos, etc.)
  - Exploración del dataset
  - Preparación de los datos
  - Elección de métricas y modelos
  - Ajuste de hiper-parámetros con técnicas de validación cruzada
  - Evaluación del modelo resultante, y análisis de sus resultados
  - Conclusiones

- La correcta implementación de las técnicas vistas en clase
- El entendimiento y análisis de los resultados obtenidos en cada paso
- La clara exposición de estos pasos, y presentación general.

Recuperatorio Parcial

- Mismo formato, Jueves 15 de Junio
- La nota reemplaza a la del parcial, tanto para mejor como para peor
- Todavía pueden promocionar
- El resto del alumnado podrá presentarse a hacer consultas respectivas al trabajo final.