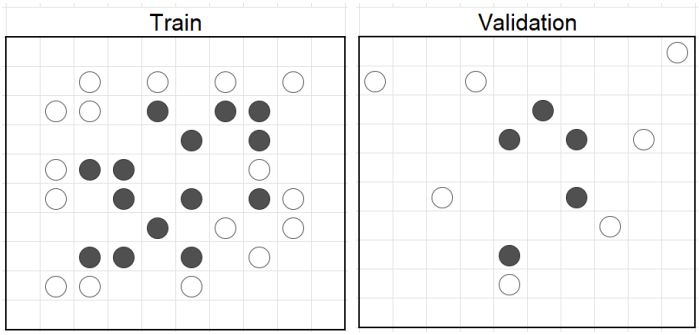
Guía de ejercicios ML-IA 2023

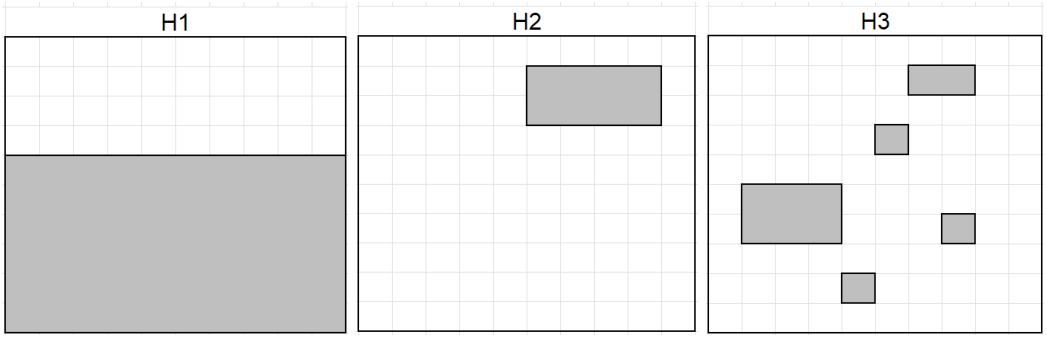
Ejercicio 1

Considere el conjunto de datos de la siguiente figura. En cada casillero hay a lo sumo un punto.



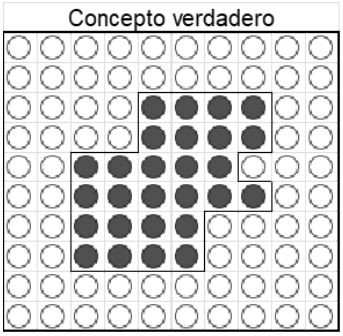
1. Describa el problema de aprendizaje representado.
2. Explique el objetivo de la partición realizada.
3. A continuación se describen tres sesgos inductivos posibles para el problema de la figura anterior:
    - Cada hipótesis de H1 corresponde a la elección de una línea horizontal y el color negro para la parte inferior que la misma define.
    - Cada hipótesis de H2 corresponde a la elección de un rectángulo de lados paralelos a los ejes y el color negro para su interior.
    - Cada hipótesis de H3 corresponde a la elección de una cantidad arbitraria de rectángulos de lados paralelos a los ejes y el color negro para sus interiores.

    Ejemplos:



A. Para cada sesgo inductivo descrito, elija una hipótesis que minimice el error en train y calcule su error en validación.
B. ¿Qué sesgo inductivo elegiría y por qué?

C. Suponga que la distribución de los datos es uniforme en la grilla y que el concepto verdadero es como se muestra en la siguiente figura:



Calcular el error verdadero de las hipótesis obtenidas en A.

Ejercicio 2

Se desea identificar una especie de flor, representada por la variable Y (cuyo valor es 1 si la flor es de la especie buscada y 0 si no), en base al largo de su pétalo (en mm) representado por la variable X. Para ello se realiza una regresión logística $\Psi(f(X))$ con $f(X) = \beta_0 + \beta_1 X$ y Ψ la función sigmoidea, utilizando un gran dataset de entrenamiento.

En la siguiente tabla se muestra el resultado en un pequeño dataset de validación:

Id	X: largo del pétalo	Y: especie	Ψ(f(X))
1	11	0	0.40
2	13	0	0.45
3	20	1	0.62
4	15	0	0.50
5	19	1	0.60
6	22	1	0.67
7	17	0	0.55
8	13	1	0.45
9	12	0	0.43
10	16	1	0.52

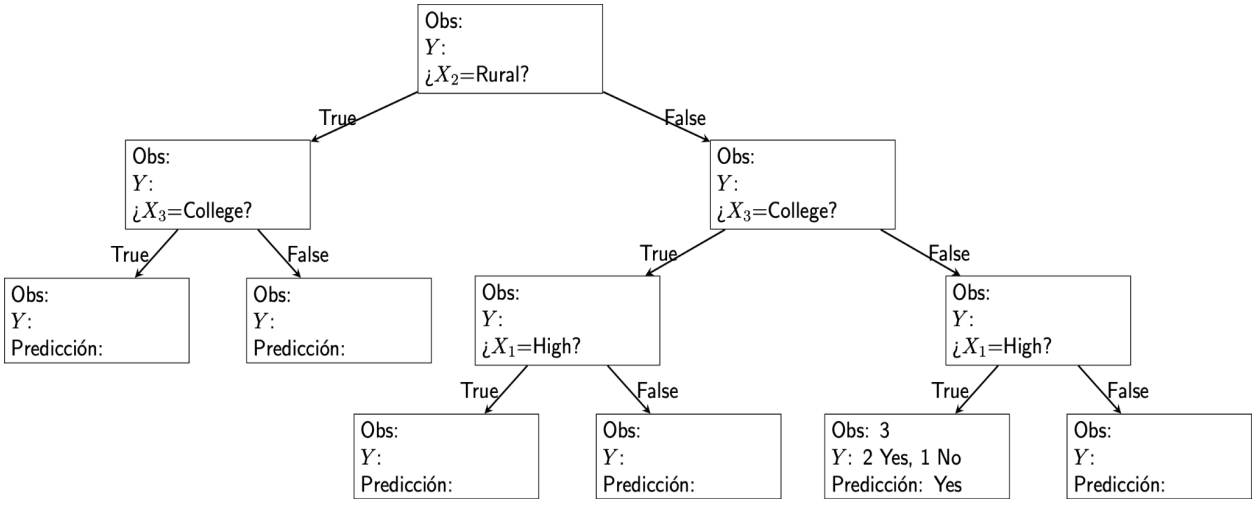
1. ¿Qué estima el valor de la columna Ψ(f(X)) correspondiente al valor calculado por la hipótesis obtenida?
2. Grafique la columna Ψ(f(X)) en función de X.
3. Para dicho dataset, calcule la exactitud (accuracy) para el siguiente conjunto de umbrales (1 si >= umbral): 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65 y 0.70.
4. ¿Qué valor de umbral elegiría? Justifique.
5. ¿Existe algún umbral que tiene exactitud igual a 1? Justifique.

Ejercicio 3

La siguiente tabla representa un conjunto de datos S para entrenar un algoritmo de aprendizaje con el objetivo de predecir la variable Y. Para ello se decide utilizar un árbol de decisión.

ID	X_1: Income	X_2: Location	X_3: Education	Y: TV at home
1	High	Urban	College	Yes
2	High	Urban	High School	Yes
3	Low	Urban	High School	No
4	High	Rural	College	No
5	Low	Rural	High School	Yes
6	Low	Urban	College	No
7	High	Urban	High School	Yes
8	Low	Rural	College	No
9	High	Urban	High School	No
10	High	Urban	College	Yes

1. Considere el árbol de la siguiente figura:



   a. Completar la información faltante en el árbol, relativa a las observaciones, la variable Y y la predicción, de forma análoga a como esta es presentada en la hoja de ejemplo.
   b. Calcule el error de dicho árbol en el dataset *S*.
2. Determine si dicho árbol corresponde a la salida del algoritmo visto en clase. Justifique su respuesta.

Fórmulas útiles:

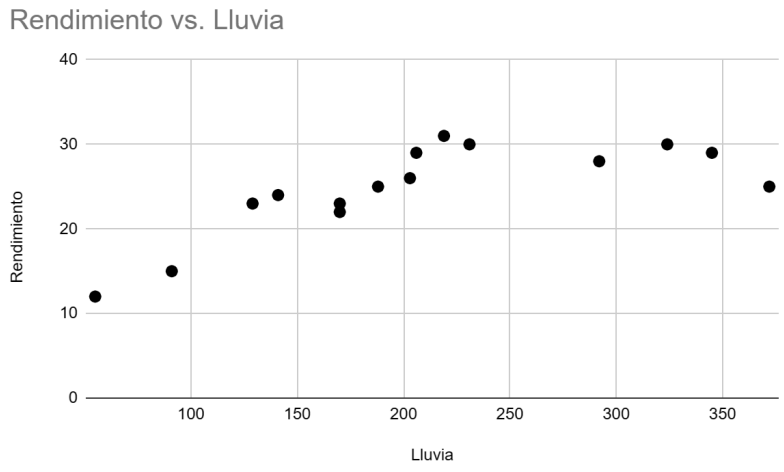
- Información: $h(x) = -\log_2(\mathbf{P}(X = x))$.
- Entropía: $H(S) = -\sum_c \mathbf{P}(Y = c \mid S) \log_2(\mathbf{P}(Y = c \mid S))$.

- Entropía esperada: $H_P(S) = \sum_a \mathbf{P}(\text{respuesta a } P = a \mid S)\, H(S^a)$.

Ejercicio 4

El cuadro a continuación muestra datos de rendimiento de cultivos de papa (en ton/ha) en función de la lluvia acumulada recibida (en mm):

Lluvia	Rendimiento
55	12
91	15
129	23
141	24
170	23
170	22
188	25
203	26
206	29
219	31
231	30
292	28
324	30
345	29
372	25



1. Se desea predecir el Rendimiento en función de la Lluvia
    a. Indique de qué tipo de problema de machine learning se trata.
    b. ¿Qué función de pérdida (loss) utilizaría?
2. ¿Cómo se descompone el error? ¿Hay error irreducible en el dataset?
3. Considere un ensemble de 3 hipótesis H1, H2 y H3 para las cuales se obtienen individualmente las siguientes predicciones:

Lluvia	Rendimiento	H1	H2	H3
55	12	10	14	21
170	22	25	20	30
231	30	35	34	51

    a. Calcule la predicción del ensemble asumiendo pesos iguales
    b. Calcule el error de cada hipótesis y la del ensemble

c. ¿Puede obtener un ensemble mejor? En caso afirmativo argumente por qué, diga cuál es y calcule su error. En caso negativo justifique.

Ejercicio 5

1. Usted y su colega desean utilizar el algoritmo de Random Forest para un problema de clasificación binaria.
    a. Su colega desea que en la construcción de cada árbol se utilicen todas las variables predictoras disponibles. Explique si está de acuerdo con su colega, para esto exponga ventajas o desventajas de dicha decisión.
    b. En caso de utilizar un número limitado de variables, explique qué procedimiento seguir para elegir la cantidad más conveniente.

2. El cuadro a continuación representa el resultado intermedio de un ensemble de 5 modelos de clasificación binaria en donde la variable a predecir es Y con valores 0 o 1.

    a. Calcular la exactitud (accuracy) de cada uno de los clasificadores y su promedio.
    b. ¿Cúal es la exactitud del ensemble si se aplica voto mayoritario?
    c. ¿El promedio calculado en a. es mayor que la exactitud del ensemble? En caso afirmativo, argumente por qué.
    d. ¿Existe un ensemble con mejor accuracy? Justifique

id	h1	h2	h3	h4	h5	Y
1	0	0	0	0	0	**0**
2	0	1	1	1	0	**1**
3	1	0	1	1	1	**1**
4	1	1	1	1	0	**0**
5	0	0	0	0	0	**0**
6	0	0	0	0	0	**1**
7	1	1	1	1	1	**1**
8	1	0	0	1	1	**0**
9	1	0	1	0	0	**1**
10	0	0	1	0	0	**0**

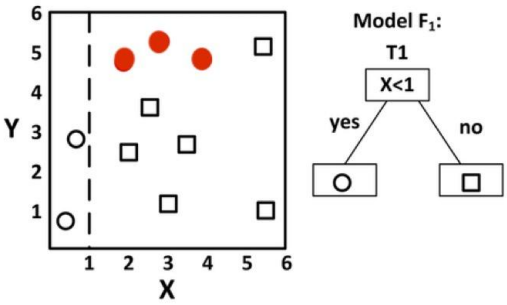
Ejercicio 6

Se dispone de un dataset cuyas variables predictoras son X e Y, siendo la variable a predecir Z cuyos valores son “círculo” y “cuadrado”. Considere las siguientes iteraciones de un algoritmo que construye un ensemble de hipótesis.

Se pide:

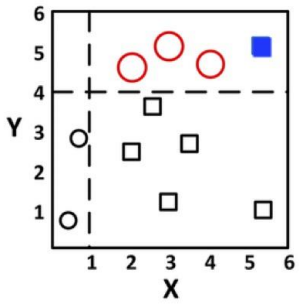
1. Describir y justificar de qué algoritmo se trata.
2. Completar las iteraciones 2 y 3.
3. Calcular el error del ensemble obtenido en cada iteración.

Iteration 1



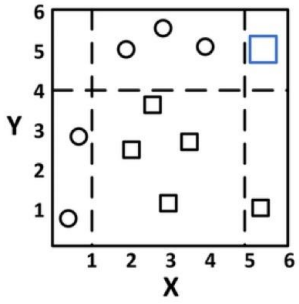
Model F_1:

Iteration 2



Model F_2:

Iteration 3



Model F_3:

Ejercicio 7

La tabla de la derecha muestra las predicciones de 4 hipótesis, h1, h2, h3 y h4, para un dataset de test de 15 elementos, numerados de 0 a 14. La última columna de la tabla es el valor verdadero de la variable a predecir Y.

Se pide:

1. Calcular la *accuracy* de cada una de las hipótesis
2. Calcular la *accuracy* del ensemble (voto mayoritario) compuesto por las 4 hipótesis. En caso de empate se predice el valor 0.
3. ¿El promedio de las *accuracies* de las hipótesis es mejor que la del ensemble? En caso negativo, argumente por qué.
4. ¿Existe un ensemble con mejor *accuracy*? Justifique su respuesta.

	h1	h2	h3	h4	Y
0	0	1	0	0	0
1	0	1	0	0	0
2	0	1	0	0	0
3	1	1	1	1	0
4	0	0	1	1	1
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	1	1	1	1	1
10	0	0	1	0	0
11	1	1	1	1	0
12	1	1	0	1	0
13	1	0	1	0	1
14	0	0	1	0	0

Ejercicio 8

Se desea construir un árbol de decisión para predecir si un préstamo concedido a un individuo terminará o no en una cancelación. Para esto se dispone de un dataset de 30 individuos: 16 pertenecen a la clase de préstamos cancelados y los otros 14 a la clase de préstamos no cancelados. Tenemos dos variables predictoras (o atributos), a saber, "Saldo", que puede tomar dos valores: "<50.000" o ">=50.000", y "Residencia", que puede tomar tres valores: "PROPIA" o "ALQUILADA".

Las siguientes tablas resumen la cantidad de individuos según el valor de cada atributo:

Saldo	Individuos	Cancelado	No cancelado
<50.000	13	12	1
>=50.000	17	4	13

Residencia	Individuos	Cancelado	No cancelado
PROPIA	15	9	6
ALQUILADA	15	7	8

Se pide:

1. Calcular la entropía de la variable a predecir (o entropía del dataset):

2. Calcular la entropía esperada de particionar el dataset para cada par (atributo, valor):

3. ¿Cuál es el par (atributo, valor) utilizado para particionar la raíz del árbol de decisión? Justifique.

Las siguientes fórmulas pueden serle de utilidad:

$$H(S, X) = \Sigma_x P(X = x \mid S).H(S_x)$$
$$H(S) = -\Sigma_{c \in C} P(Y = c \mid S).log_2(P(Y = c \mid S))$$

donde S_x es el resultado de particionar *S* por el valor *x* del atributo *X*.

Ejercicio 9

1. ¿Cómo se descompone el error? Grafique las curvas típicas.
2. ¿Qué es el sobreajuste? Identifique la zona de sobreajuste en el gráfico.
3. ¿Qué se entiende por regularización?
4. ¿Cuál es el objetivo de la regularización en términos del sobreajuste?
5. ¿Eliminar variables predictoras es una técnica de regularización? Justifique.
6. ¿Qué técnica de regularización tiene como efecto eliminar variables? Explique.
7. ¿Qué algoritmo de machine learning se apoya en la idea de no usar todas las variables?
8. Considere la siguiente tabla.

Ejercicio 10

Elemento	C1	C2	C3	C4	C5
A	*		*	*	*
B	*		*	*	*
C	*	*		*	*
D	*	*			
E		*	*		

La tabla muestra los resultados obtenidos por 5 clasificadores C1, C2, C3, C4 y C5, sobre un conjunto de datos compuesto por 5 elementos, identificados como A, B, C, D y E. Una * indica que el elemento fue clasificado correctamente.

a. Calcule el error cometido por el ensemble compuesto por los 5 clasificadores, asumiendo peso uniforme y la esperanza del error de los clasificadores.
b. ¿Se verifica la propiedad fundamental del error del ensemble con respecto a los errores de los clasificadores? Justifique.
c. En caso que la respuesta sea negativa, construya un ensemble a partir de los clasificadores dados que cumpla la propiedad. Justifique.

Ejercicio 11

Dentro de las técnicas de ensamble existen diferentes mecanismos para combinar o agregar los resultados de varias hipótesis. Consideremos un problema de clasificación binaria, donde las hipótesis que se componen son regresiones logísticas. Dos mecanismos posibles para combinarlas con el objetivo de formar el ensamble son:

- El voto mayoritario. Es decir, se elige la clase que tiene la mayor cantidad de votos (es decir que es predicha más veces por los clasificadores individuales).
- Realizar el promedio del output de las hipótesis intermedias y emitir el veredicto en función de este promedio.

Discuta:
¿Son estos criterios de combinación de resultados equivalentes? En caso afirmativo, justifique, en caso negativo, brinde un contraejemplo.

Ejercicio 12

1. ¿En qué casos considera usted que la accuracy es una métrica de poca utilidad? Justifique con un ejemplo.
2. Dada la siguiente tabla de Precision vs Recall:

		Precision	**Recall**
Thresholds	**0.0**	0.38	1.0
	0.26	0.43	1.0
	0.51	0.6	1.0
	0.73	0.5	0.67
	1.0	0.5	0.34
	1.2	1.0	0.34

   A. Defina Precision y Recall (fórmulas).
   B. Grafique la curva Precision vs Recall.
   C. El F1-score se define como la media armónica entre precision y recall. ¿Cuál es el umbral que maximiza este valor?
   D. Un compañero de equipo le presenta la tabla anterior diciendo que realizó los cálculos de la parte B para seleccionar el mejor umbral con datos de test, ¿Qué comentario le haría?

Ejercicio 13

1. ¿Qué es un árbol de decisión y cómo se puede utilizar en problemas de clasificación y regresión? Describa los conceptos clave, como nodos, ramas, criterios de división y predicciones.

2. Explique en detalle el algoritmo de aprendizaje de los árboles de decisión utilizando el criterio de ganancia de información. Proporcione un pseudocódigo completo que ilustre el proceso de construcción del árbol.

3. Discuta en profundidad la relación entre la entropía y la ganancia de información en los árboles de decisión. Explique cómo se calcula la entropía y cómo se utiliza para medir la impureza de un conjunto de datos. Describa cómo se aplica la ganancia de información para seleccionar las mejores características de división.

Ejercicio 14

1. Describa en qué consiste una técnica de ensamble, como el bagging o el boosting. Explique cómo se combinan múltiples modelos individuales para mejorar la precisión de las predicciones. Proporcione ejemplos concretos de algoritmos de ensamble, como Random Forest y Gradient Boosting.
2. Analice los problemas que se pueden resolver con técnicas de ensamble. Explique cómo estas técnicas pueden abordar desafíos como el sobreajuste, la falta de generalización y la mejora del rendimiento en conjuntos de datos desbalanceados. Ilustre sus respuestas con ejemplos de casos de uso relevantes.
3. Compare la varianza de un Random Forest compuesto por 100 árboles con la de un Random Forest compuesto por 1000 árboles. Explique cómo la cantidad de árboles en un ensamble afecta la varianza del modelo y su capacidad para reducir el sobreajuste.
4. Detalle las diferencias fundamentales entre el bagging y el boosting como técnicas de ensamble. Compare cómo se construyen y combinan los modelos individuales en cada técnica y cómo se realiza la ponderación de las predicciones. Analice los efectos en el rendimiento y la capacidad de generalización de los modelos.

Ejercicio 15

1. Identifique los elementos que determinan el error total y explique cada uno, así como la relación que existe entre ellos (se recomienda utilizar la gráfica vista en clase como apoyo a la explicación).
2. Hay muchas causas que afectan la precisión y el rendimiento de un modelo de Machine Learning, una de ellas es el overfitting o sobreajuste.
    a. Describa brevemente a qué se refiere dicho concepto y qué consecuencias negativas puede tener.
    b. Mencione y describa dos técnicas que se pueden utilizar para evitar el sobreajuste.

Ejercicio 16

Para cada uno de los escenarios propuestos a continuación, determine si es un problema de clasificación o regresión y qué métrica de selección elegiría. Justifique brevemente su respuesta.

1. Detección de cáncer.
2. Predecir el límite de crédito de un préstamo.
3. Detectar si un correo electrónico es spam.
4. Determinar el consumo de datos de un cliente en un período de tiempo.

Ejercicio 17

Recordando la descomposición de sesgo-varianza:

1. Elabore un bosquejo de una única gráfica típica que muestre las curvas de sesgo (al cuadrado), varianza, error de train, error de test y error irreducible, considerando en el eje de las abscisas diferentes complejidades de hipótesis (de menor complejidad a mayor complejidad). Esto es, el “eje de las x” debería representar la flexibilidad del método de aprendizaje y el “eje de las y” los valores de cada curva. Por favor, agregue una etiqueta a cada curva para identificarlas.
2. Explique por qué cada una de las cinco curvas, tiene la forma expuesta en (a).

Ejercicio 18

a. Explique en qué consiste la técnica de bagging.
b. Suponga que se generan 10 muestras “bootstrapped” a partir de un conjunto de datos que contiene las clases “rojo” (R) y “verde” (V). Sobre cada una de estas muestras, se aplica un árbol de decisión para clasificar y, para cada valor de X, se producen 10 predicciones de $P(Clase = R \mid X)$ según el siguiente detalle:

0.1	0.15	0.2	0.2	0.55	0.6	0.6	0.65	0.7	0.75

Considere los siguientes enfoques para combinar estos resultados en una única predicción:

a. Clasificación basada en el “voto mayoritario”.
b. Clasificación basada en la probabilidad promedio.

Para los resultados anteriores, indique la clasificación final basada en las opciones (1) y (2).

c. Repita la parte (b) considerando ahora los siguientes datos:

0.1	0.15	0.2	0.25	0.55	0.65	0.7	0.8	0.85	0.9

Compare los resultados obtenidos y explique.

Ejercicio 19

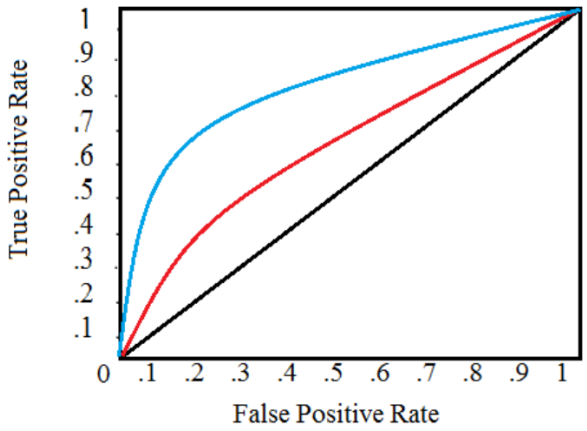
1. ¿Cuál es el objetivo de particionar los datos que se tienen para desarrollar un sistema inteligente en train, test y validation? Desarrolle brevemente, y mencione algunos desafíos que podrían surgir al realizar esta partición.

2. Al seleccionar conjuntos de datos para el desarrollo de modelos de Machine Learning, ¿por qué los datasets de dev (train + validation) y test deben venir de la misma distribución? Explique la importancia de la estacionariedad de la distribución y mencione posibles soluciones cuando los datos disponibles son limitados.

3. Explique cómo funciona y para qué sirve la técnica k-folds cross-validation.

Ejercicio 20

Dada la siguiente imagen de curvas ROC:



1. ¿Qué clasificador tiene mejor relación entre el recall de la clase positiva y el recall de la clase negativa?
2. ¿Por qué?
3. Marque en la imagen el umbral qué logra la mejor relación e indique su FPR y TPR aproximado.

Las siguientes fórmulas pueden serle de utilidad:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{TNR}$$

Ejercicio 21

Dada la siguiente tabla (asumiendo la variable a predecir, y que la frecuencia relativa de los valores de cada variable es representativa de la probabilidad de ocurrencia):

X1	X2	X3	Y
A	Z	F	No
B	Z	D	Si
A	C	D	Si
B	C	D	Si

Responda:

1. ¿Qué valor (y de qué atributo) maximiza la información?
2. ¿Qué atributo (o atributos) tiene mayor entropía?
3. ¿Cúal es la entropía del dataset?

Las siguientes fórmulas pueden serle de utilidad:

$$i(x) = -\log_2(P(x))$$
$$I(X) = -\Sigma_x P(x) \cdot \log_2(P(x))$$
$$I(S) = -\Sigma_{c \in C} P(Y = c \mid S) \cdot \log_2(P(Y = c \mid S))$$