

Validación de modelos predictivos

A series of horizontal lines in teal and light blue colors, with varying lengths and offsets, creating a modern, layered effect across the middle of the slide.

¿Para qué validar el modelo?

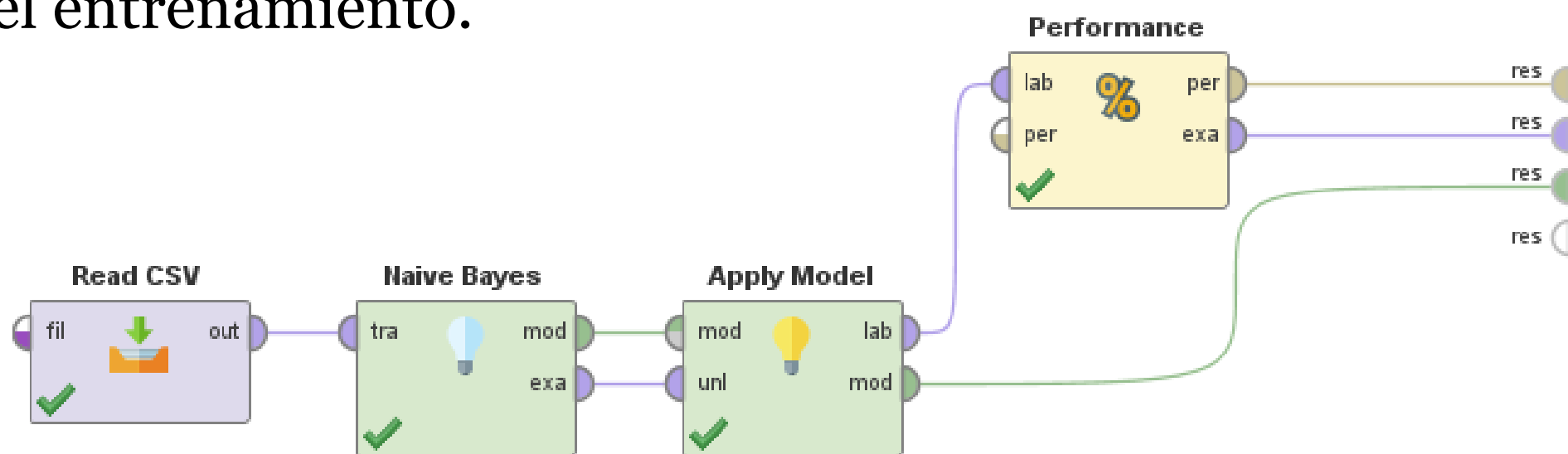
- El objetivo de la validación es medir el desempeño del modelo para predecir la respuesta esperada.
- Se busca determinar la calidad de la respuesta que brinda.
- Para que la medición sea objetiva debe hacerse sobre un conjunto de datos diferente al utilizado para generar el modelo.

Entrenamiento y testeo

- El conjunto de datos original se dividirá en dos partes
 - **Conjunto de datos de entrenamiento**
 - Se utilizarán para construir el modelo. Como el aprendizaje es supervisado el método buscará ajustar su respuesta a lo indicado en estos ejemplos.
 - **Conjunto de datos de testeo**
 - Una vez construido el modelo será utilizados para medir su calidad.
 - Se espera que la respuesta del modelo coincida lo más posible con lo indicado en estos ejemplos.

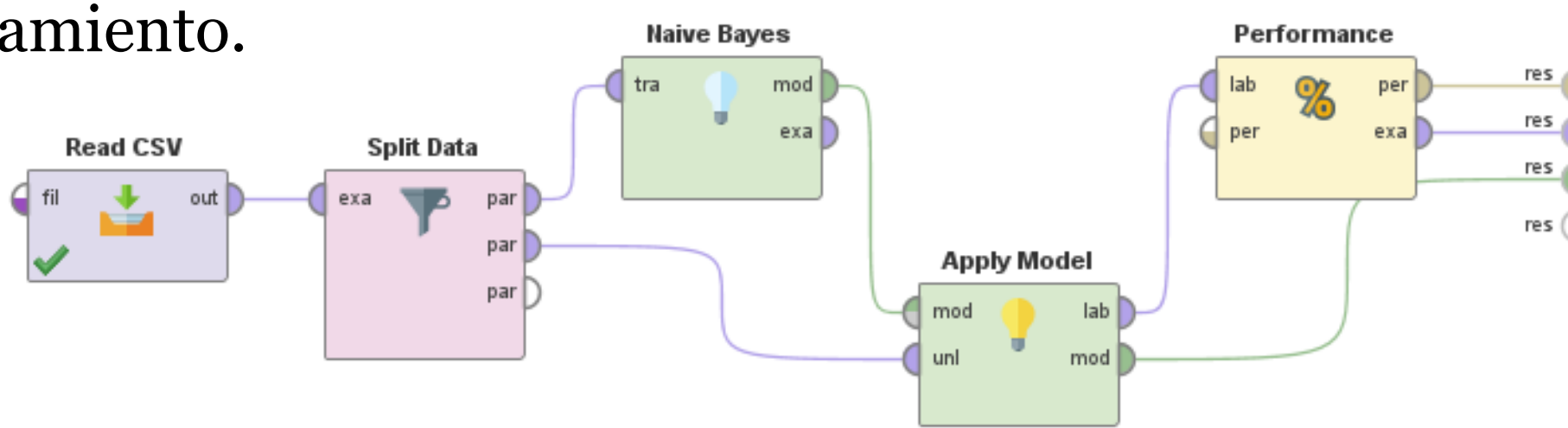
Performance sobre los datos de entrenamiento

- Se evalúa el modelo sobre los mismos datos que se usaron para construirlo.
- Mide la capacidad del modelo de ajustarse a los datos utilizados en el entrenamiento.



Performance sobre los datos de testeo

- Se evalúa el modelo sobre otros datos que NO fueron observados durante el entrenamiento.
- Mide la capacidad del modelo de generalizar a partir de los ejemplos.
- La tasa de aciertos suele ser menor que la obtenida sobre los datos de entrenamiento.



Matriz de Confusión

	True Class 1	True Class 2	Precision
Predice Clase 1	A	B	$A/(A+B)$
Predice Clase 2	C	D	$D/(C+D)$
Recall	$A/(A+C)$	$D/(B+D)$	$(A+D)/(A+B+C+D)$

accuracy

- Los **aciertos** del modelo están sobre la **diagonal** de la matriz.
- **Precision**: la proporción de **predicciones correctas** sobre **una clase**.
- **Recall**: la proporción de **ejemplos** de **una clase** que son **correctamente clasificados**.
- **Accuracy**: la performance general del modelo, sobre **todas las clases**. Es la cantidad de **aciertos** sobre el **total** de ejemplos.

Clasificación binaria

- Los resultados se etiquetan como positivos (P) o negativos (N)
- Luego, la matriz de confusión tendrá la siguiente forma:

	Clase P	Clase N
Predice P	VP	FP
Predice N	FN	VN

$P = VP + FN$ $N = FP + VN$

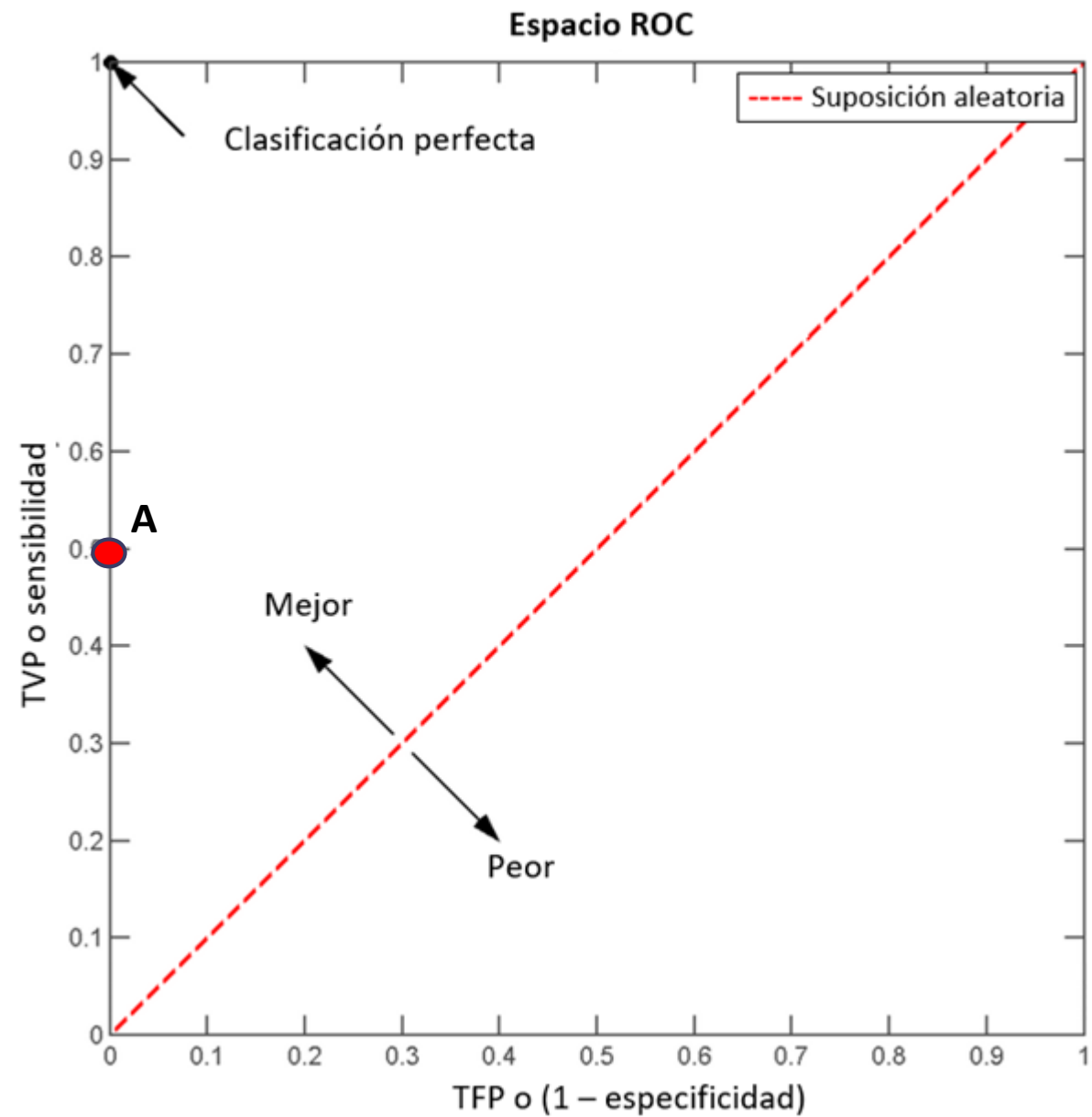
- Tasa de verdaderos positivos $\rightarrow TVP = VP / P$ (Sensibilidad)
- Tasa de verdaderos negativos $\rightarrow TVN = VN / N$ (Especificidad)
- Tasa de Falsos Positivos $\rightarrow TFP = FP / N$ (1 - Especificidad)

A

VP=6	FP=0	6
FN=6	VN=8	6
12	8	20

$$\text{TVP} = 6/12 = 0.5$$

$$\text{TFP} = 0/8 = 0$$



A

VP=6	FP=0	6
FN=6	VN=8	6
12	8	20

$$\text{TVP} = 6/12 = 0.5$$

$$\text{TFP} = 0/8 = 0$$

B

VP=9	FP=6	15
FN=3	VN=2	5
12	8	20

$$\text{TVP} = 9/12 = 0.75$$

$$\text{TFP} = 6/8 = 0.75$$

C

VP=12	FP=2	6
FN=0	VN=6	6
12	8	20

$$\text{TVP} = 12/12 = 1$$

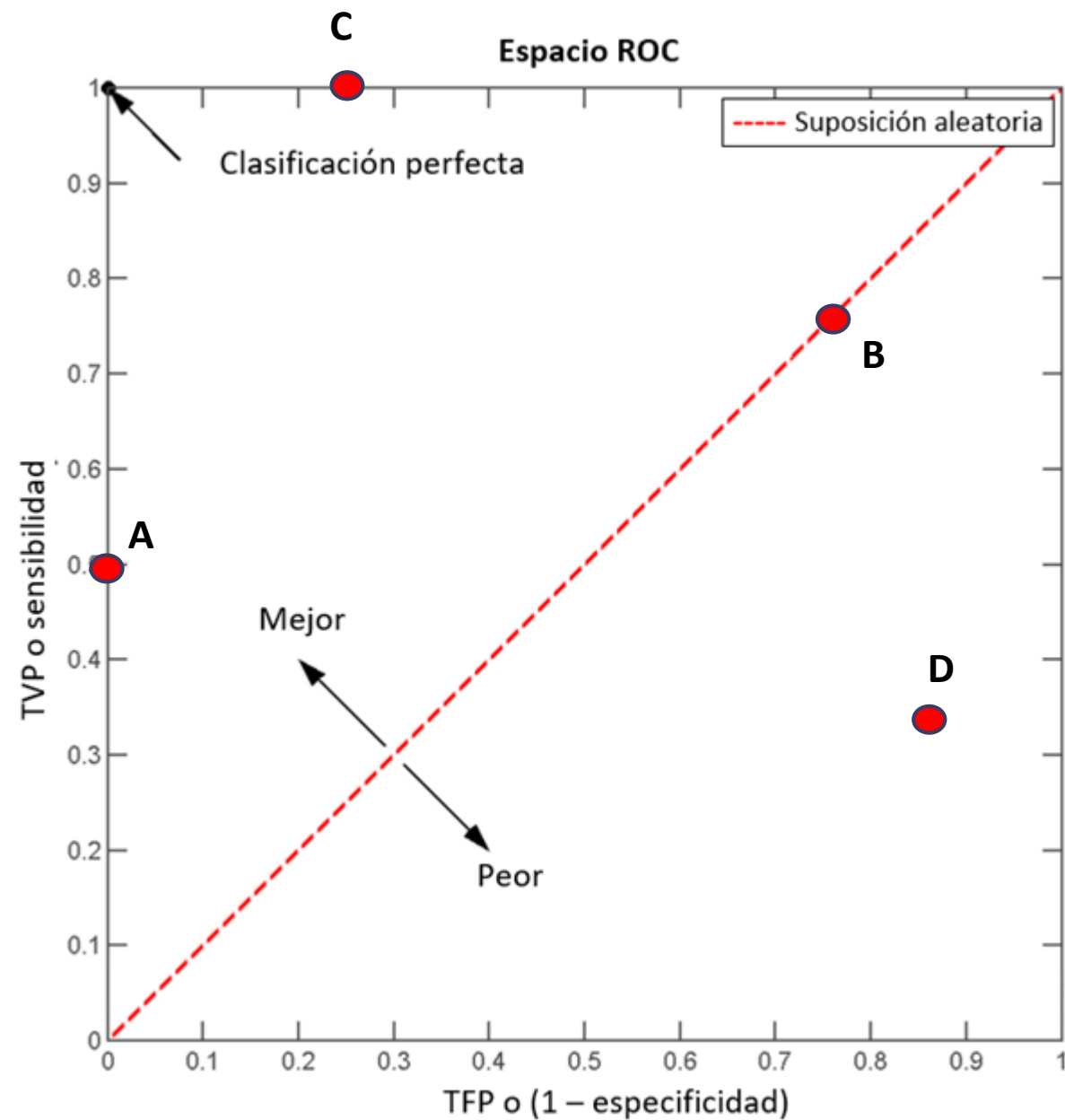
$$\text{TFP} = 2/8 = 0.25$$

D

VP=4	FP=7	11
FN=8	VN=1	9
12	8	20

$$\text{TVP} = 4/12 = 0.33$$

$$\text{TFP} = 7/8 = 0.875$$



Sonar.csv

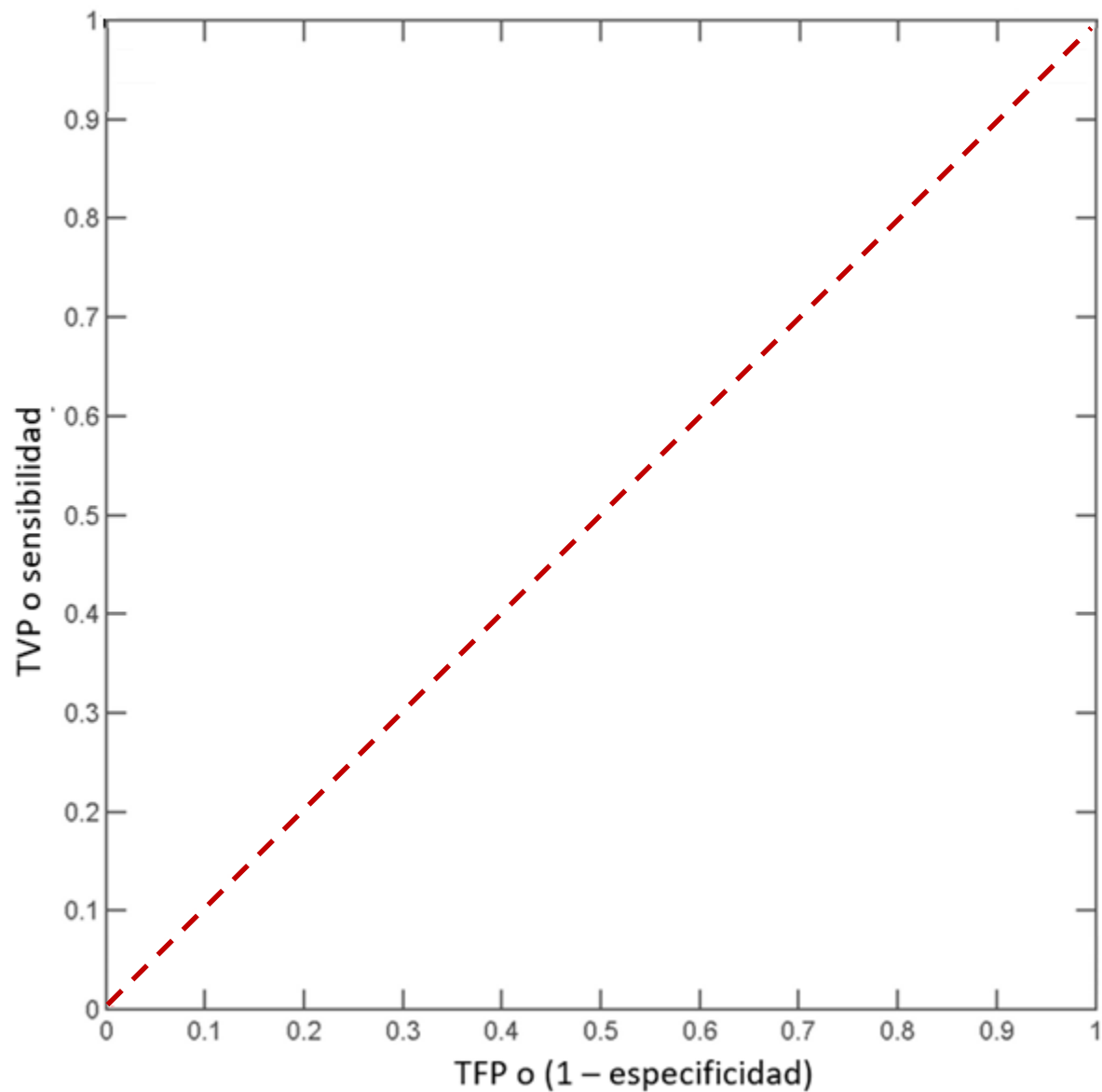
- Conjunto de datos utilizado para clasificar señales de sonar.
- La tarea es construir un modelo predictivo para discriminar entre señales de sonar rebotadas en un cilindro de metal y aquellas rebotadas en una roca más o menos cilíndrica.
 - *Cada patrón es un conjunto de 60 números en el rango de 0.0 a 1.0.*
 - *Cada número representa la energía dentro de una banda de frecuencia particular, integrada durante un cierto período de tiempo.*
 - *La etiqueta asociada a cada registro contiene "Roca" si el objeto es una roca y "Mina" si es una mina (cilindro de metal).*

<https://archive.ics.uci.edu/ml/datasets>

ID	Clase	Confianza	Predice
5	Mina	1	
7	Mina	1	
9	Mina	1	
1	Mina	0.9	
10	Mina	0.9	
20	Mina	0.9	
8	Roca	0.8	
14	Mina	0.8	
15	Mina	0.8	
18	Roca	0.8	
19	Mina	0.8	
3	Mina	0.7	
6	Mina	0.7	
12	Mina	0.65	
4	Roca	0.6	
16	Roca	0.6	
11	Roca	0.5	
2	Roca	0.4	
13	Roca	0.3	
17	Roca	0.1	

12 minas y **8 rocas**

CURVA ROC

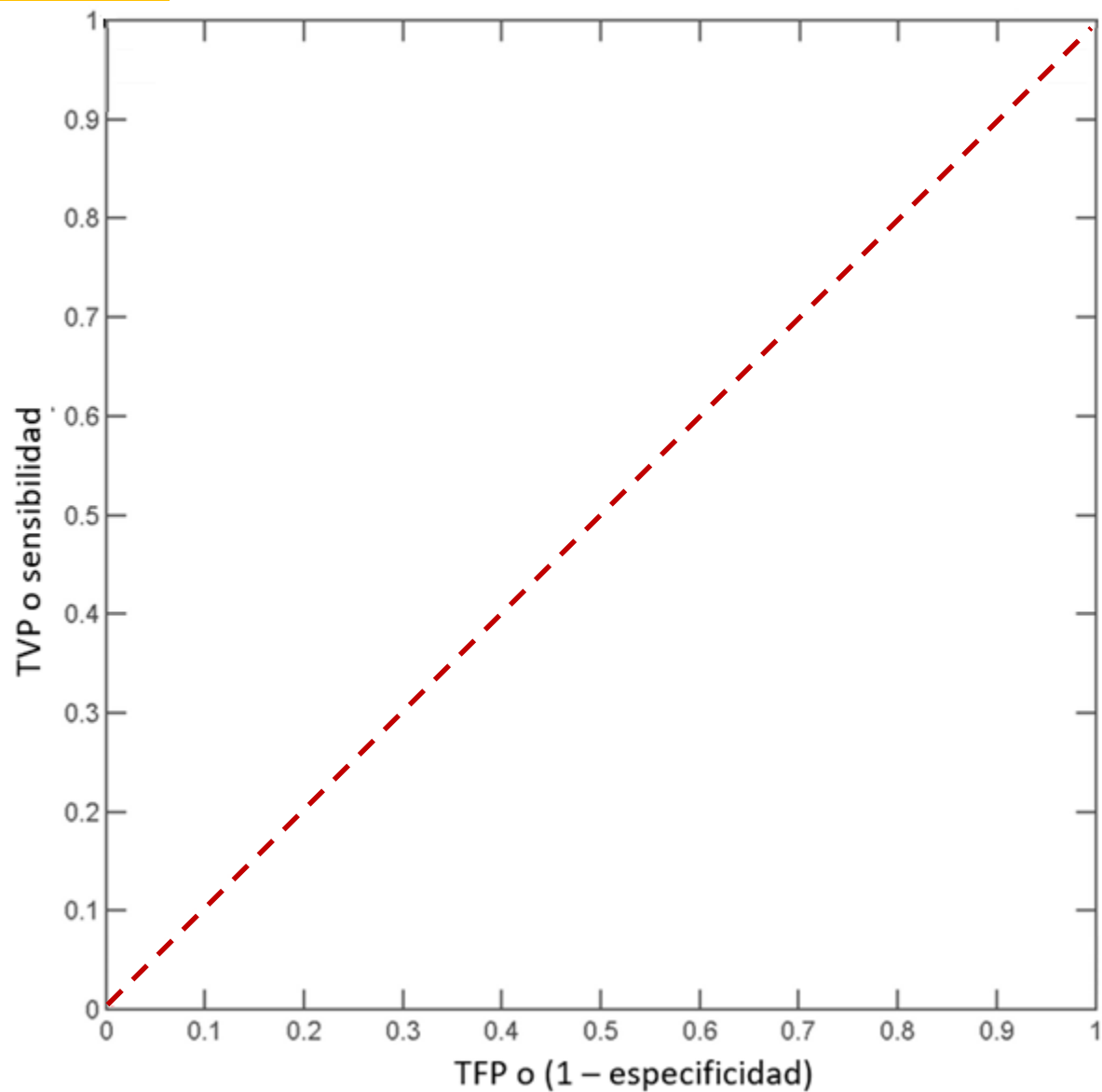


ID	Clase	Confianza	Predice
5	Mina	1	
7	Mina	1	
9	Mina	1	
1	Mina	0.9	
10	Mina	0.9	
20	Mina	0.9	
8	Roca	0.8	
14	Mina	0.8	
15	Mina	0.8	
18	Roca	0.8	
19	Mina	0.8	
3	Mina	0.7	
6	Mina	0.7	
12	Mina	0.65	
4	Roca	0.6	
16	Roca	0.6	
11	Roca	0.5	
2	Roca	0.4	
13	Roca	0.3	
17	Roca	0.1	

12 minas y **8 rocas**

Umbral = 1

CURVA ROC

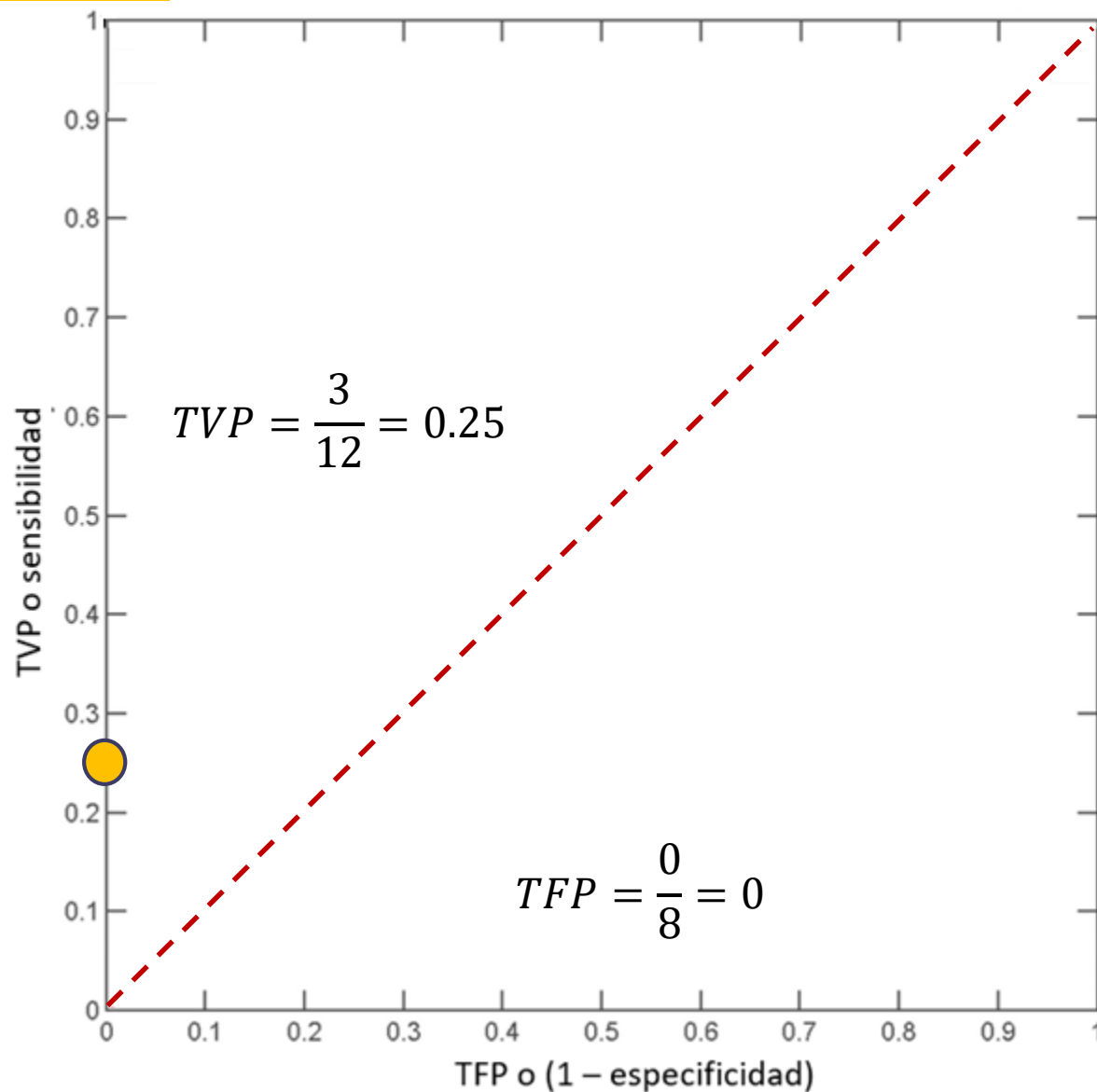


ID	Clase	Confianza	Predice
5	Mina	1	Mina
7	Mina	1	Mina
9	Mina	1	Mina
1	Mina	0.9	Roca
10	Mina	0.9	Roca
20	Mina	0.9	Roca
8	Roca	0.8	Roca
14	Mina	0.8	Roca
15	Mina	0.8	Roca
18	Roca	0.8	Roca
19	Mina	0.8	Roca
3	Mina	0.7	Roca
6	Mina	0.7	Roca
12	Mina	0.65	Roca
4	Roca	0.6	Roca
16	Roca	0.6	Roca
11	Roca	0.5	Roca
2	Roca	0.4	Roca
13	Roca	0.3	Roca
17	Roca	0.1	Roca

12 minas y **8 rocas**

Umbral = 1

CURVA ROC

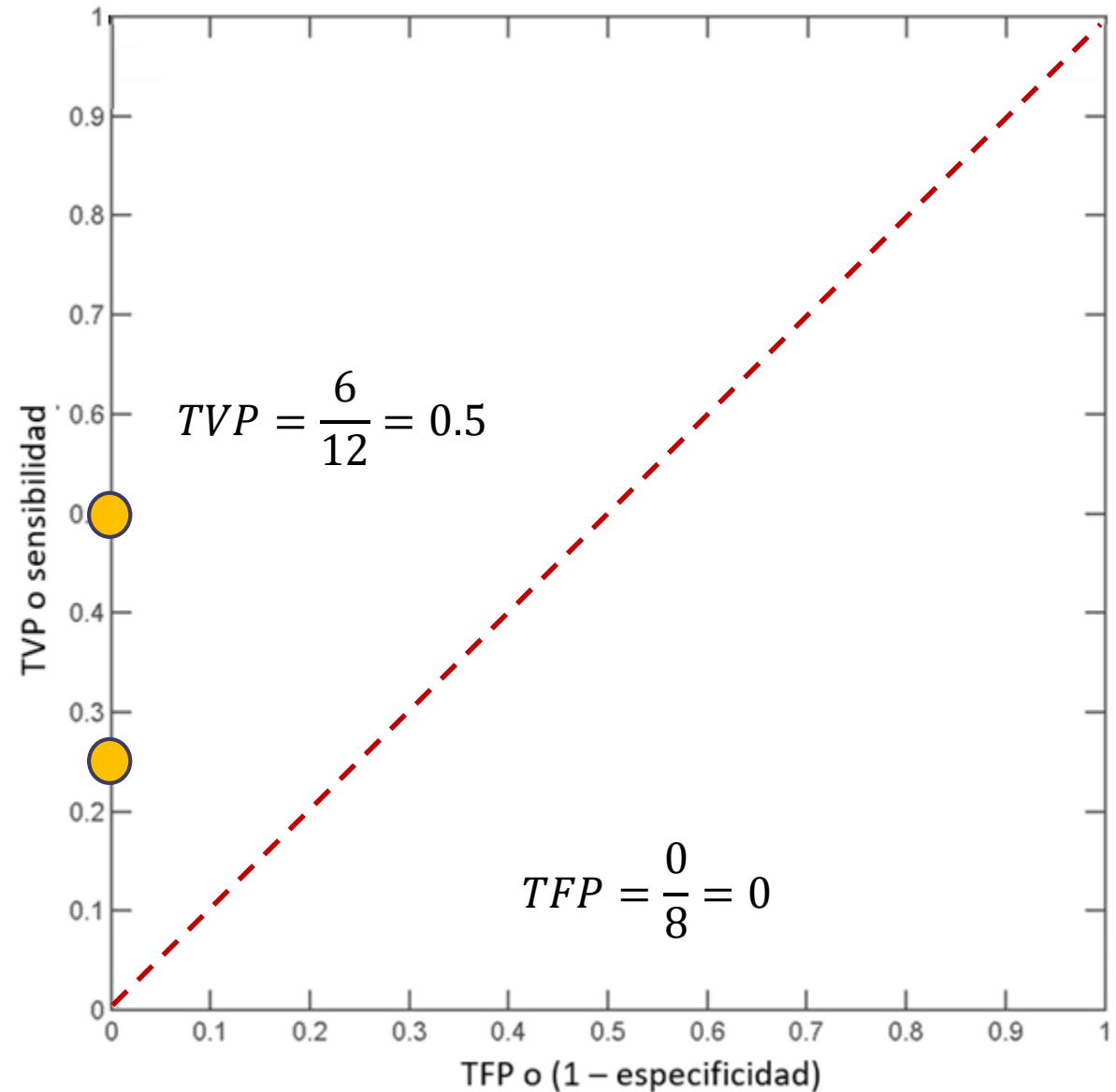


ID	Clase	Confianza	Predice
5	Mina	1	Mina
7	Mina	1	Mina
9	Mina	1	Mina
1	Mina	0.9	Mina
10	Mina	0.9	Mina
20	Mina	0.9	Mina
8	Roca	0.8	Roca
14	Mina	0.8	Roca
15	Mina	0.8	Roca
18	Roca	0.8	Roca
19	Mina	0.8	Roca
3	Mina	0.7	Roca
6	Mina	0.7	Roca
12	Mina	0.65	Roca
4	Roca	0.6	Roca
16	Roca	0.6	Roca
11	Roca	0.5	Roca
2	Roca	0.4	Roca
13	Roca	0.3	Roca
17	Roca	0.1	Roca

12 minas y 8 rocas

Umbral = 0.9

CURVA ROC

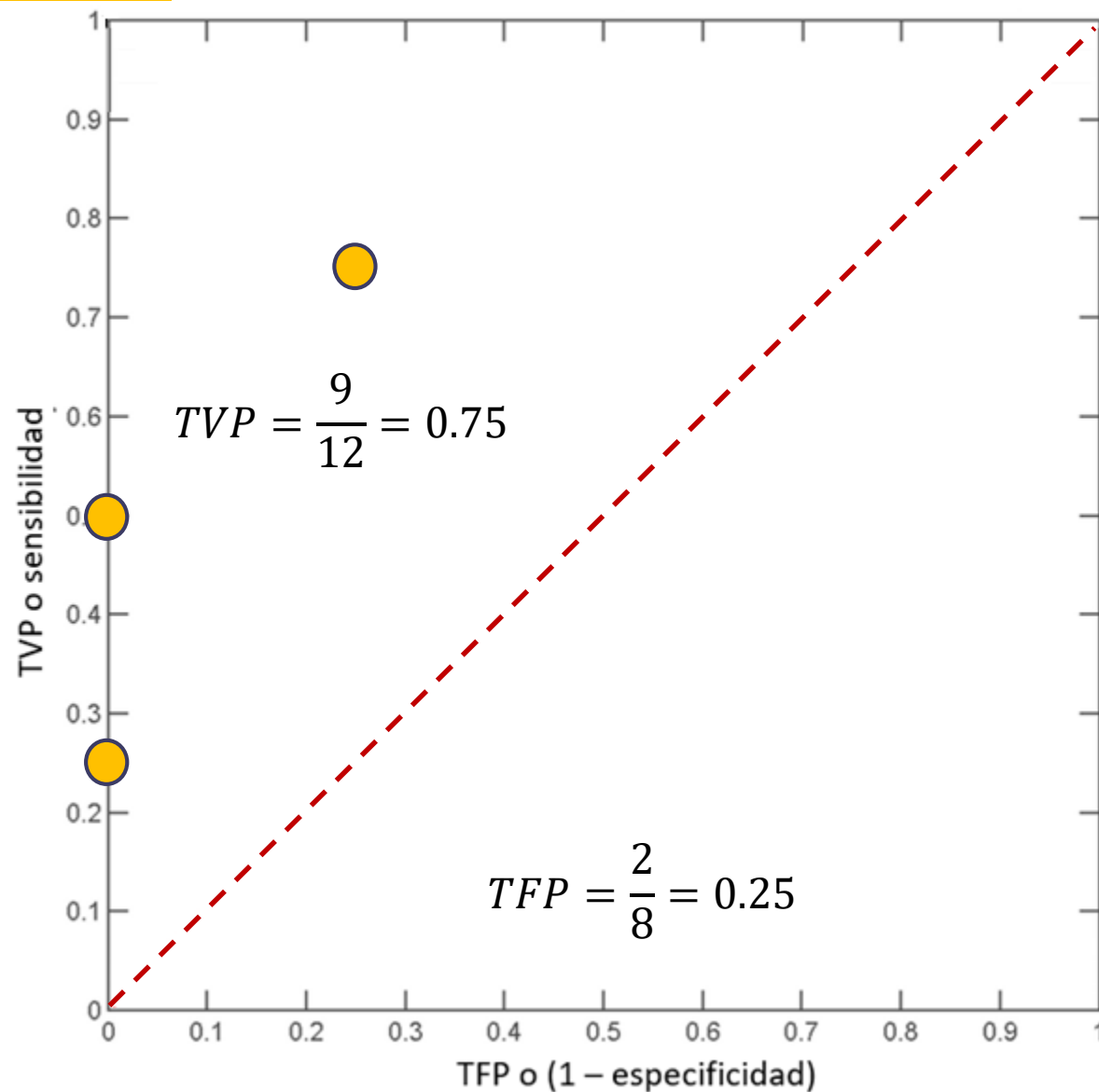


ID	Clase	Confianza	Predice
5	Mina	1	Mina
7	Mina	1	Mina
9	Mina	1	Mina
1	Mina	0.9	Mina
10	Mina	0.9	Mina
20	Mina	0.9	Mina
8	Roca	0.8	Mina
14	Mina	0.8	Mina
15	Mina	0.8	Mina
18	Roca	0.8	Mina
19	Mina	0.8	Mina
3	Mina	0.7	Roca
6	Mina	0.7	Roca
12	Mina	0.65	Roca
4	Roca	0.6	Roca
16	Roca	0.6	Roca
11	Roca	0.5	Roca
2	Roca	0.4	Roca
13	Roca	0.3	Roca
17	Roca	0.1	Roca

12 minas y 8 rocas

Umbral = 0.8

CURVA ROC

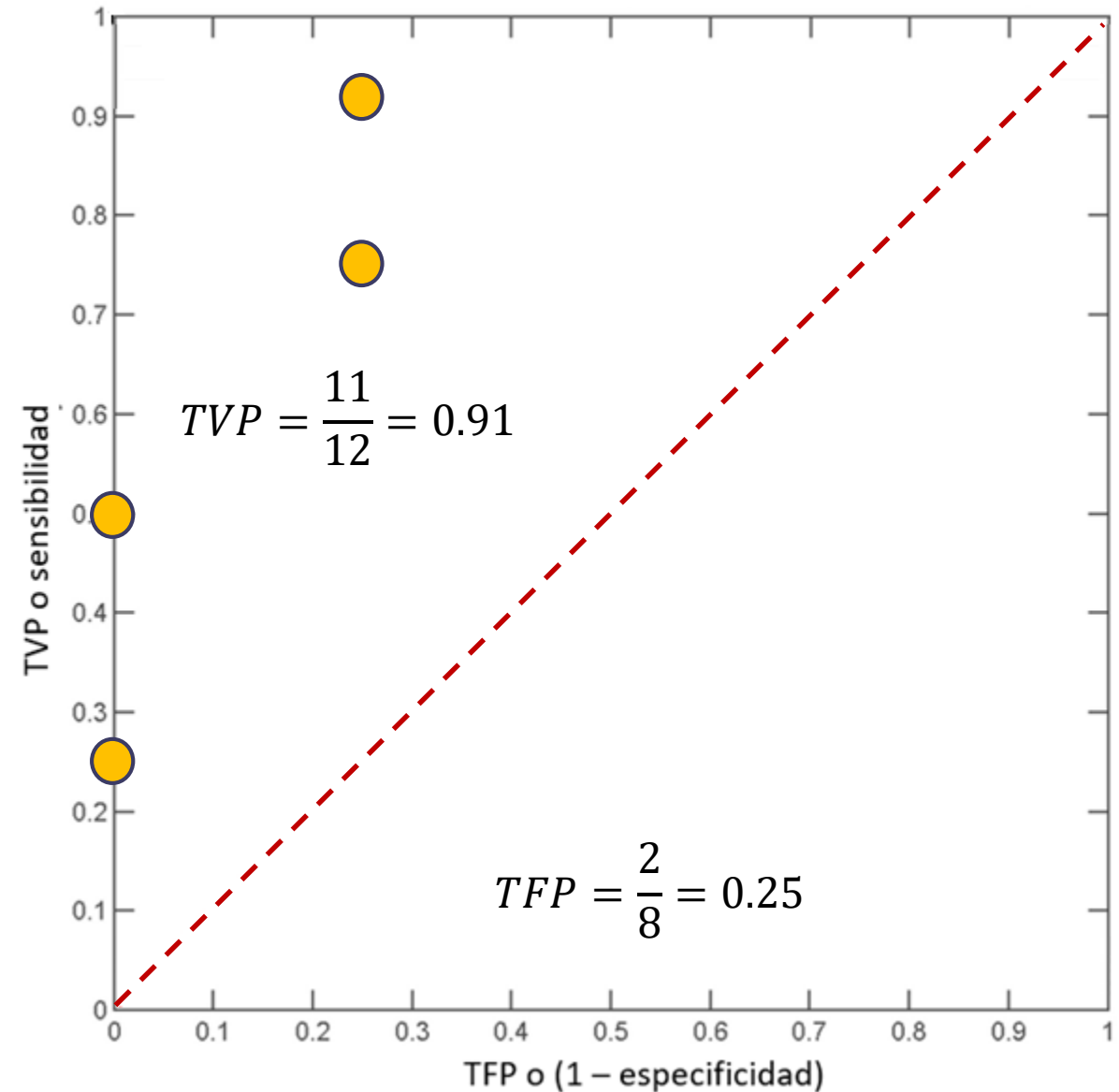


ID	Clase	Confianza	Predice
5	Mina	1	Mina
7	Mina	1	Mina
9	Mina	1	Mina
1	Mina	0.9	Mina
10	Mina	0.9	Mina
20	Mina	0.9	Mina
8	Roca	0.8	Mina
14	Mina	0.8	Mina
15	Mina	0.8	Mina
18	Roca	0.8	Mina
19	Mina	0.8	Mina
3	Mina	0.7	Mina
6	Mina	0.7	Mina
12	Mina	0.65	Roca
4	Roca	0.6	Roca
16	Roca	0.6	Roca
11	Roca	0.5	Roca
2	Roca	0.4	Roca
13	Roca	0.3	Roca
17	Roca	0.1	Roca

12 minas y 8 rocas

Umbral = 0.7

CURVA ROC

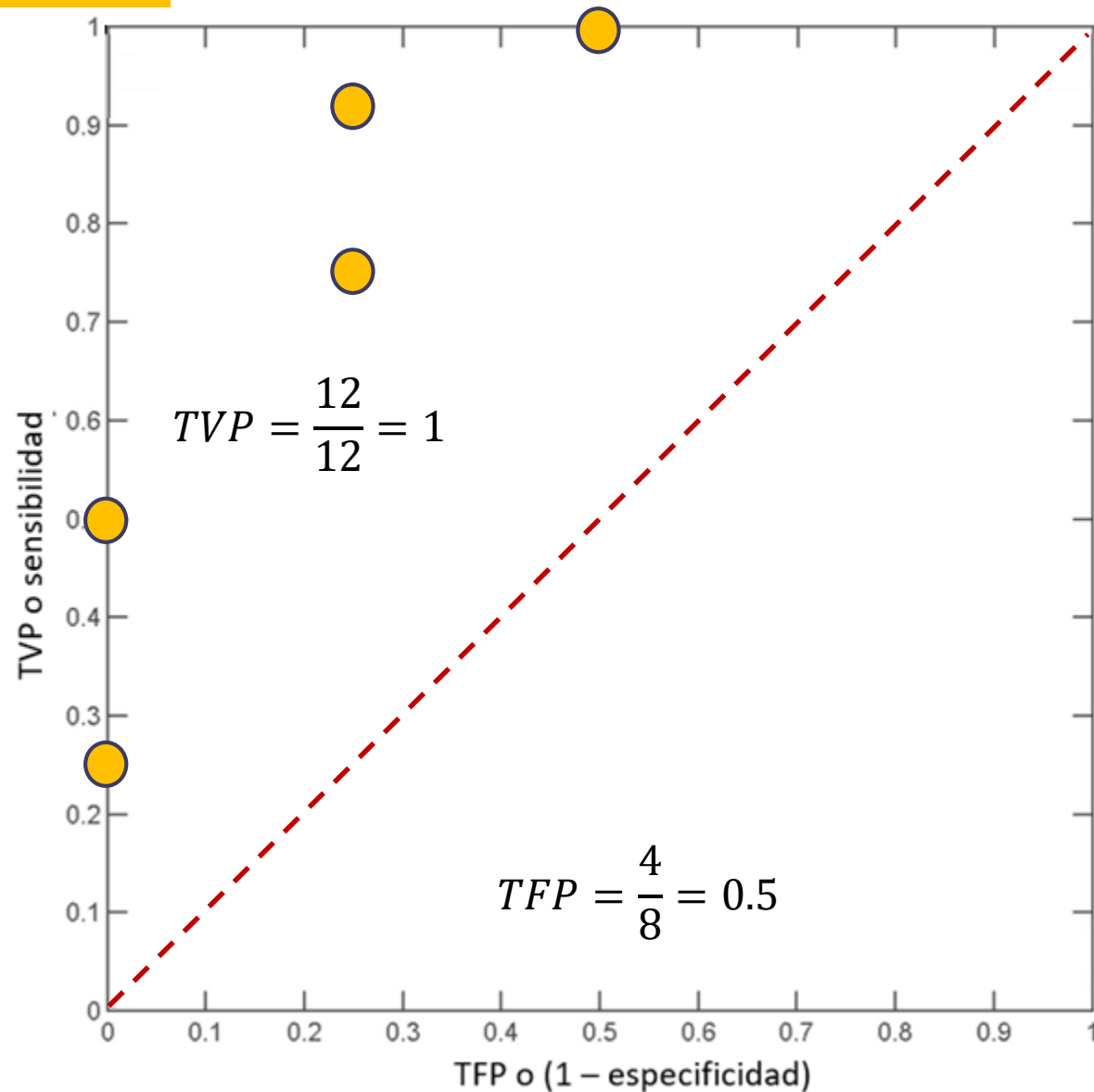


ID	Clase	Confianza	Predice
5	Mina	1	Mina
7	Mina	1	Mina
9	Mina	1	Mina
1	Mina	0.9	Mina
10	Mina	0.9	Mina
20	Mina	0.9	Mina
8	Roca	0.8	Mina
14	Mina	0.8	Mina
15	Mina	0.8	Mina
18	Roca	0.8	Mina
19	Mina	0.8	Mina
3	Mina	0.7	Mina
6	Mina	0.7	Mina
12	Mina	0.65	Mina
4	Roca	0.6	Mina
16	Roca	0.6	Mina
11	Roca	0.5	Roca
2	Roca	0.4	Roca
13	Roca	0.3	Roca
17	Roca	0.1	Roca

12 minas y 8 rocas

Umbral = 0.6

CURVA ROC

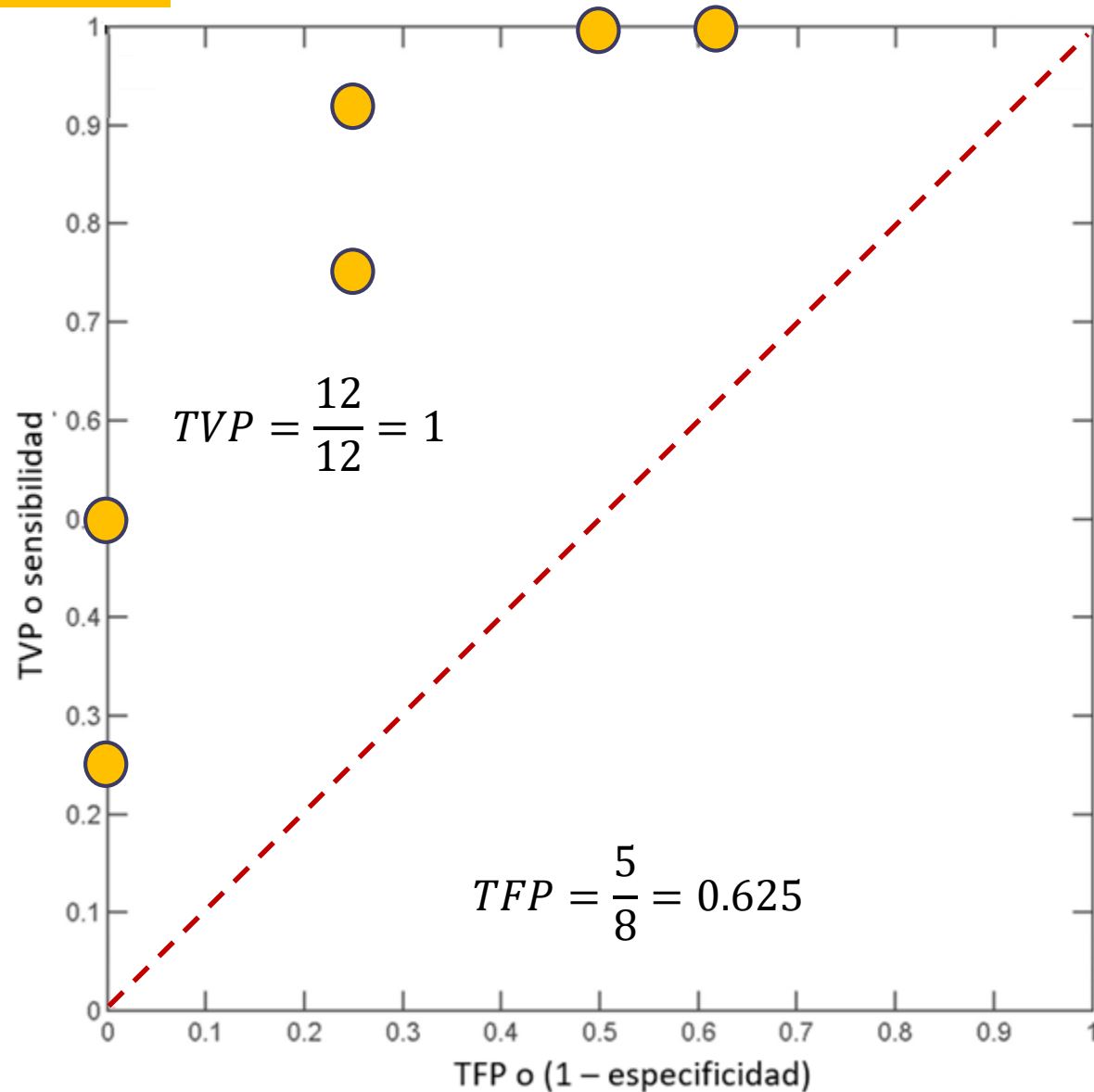


ID	Clase	Confianza	Predice
5	Mina	1	Mina
7	Mina	1	Mina
9	Mina	1	Mina
1	Mina	0.9	Mina
10	Mina	0.9	Mina
20	Mina	0.9	Mina
8	Roca	0.8	Mina
14	Mina	0.8	Mina
15	Mina	0.8	Mina
18	Roca	0.8	Mina
19	Mina	0.8	Mina
3	Mina	0.7	Mina
6	Mina	0.7	Mina
12	Mina	0.65	Mina
4	Roca	0.6	Mina
16	Roca	0.6	Mina
11	Roca	0.5	Mina
2	Roca	0.4	Roca
13	Roca	0.3	Roca
17	Roca	0.1	Roca

12 minas y 8 rocas

Umbral = 0.5

CURVA ROC

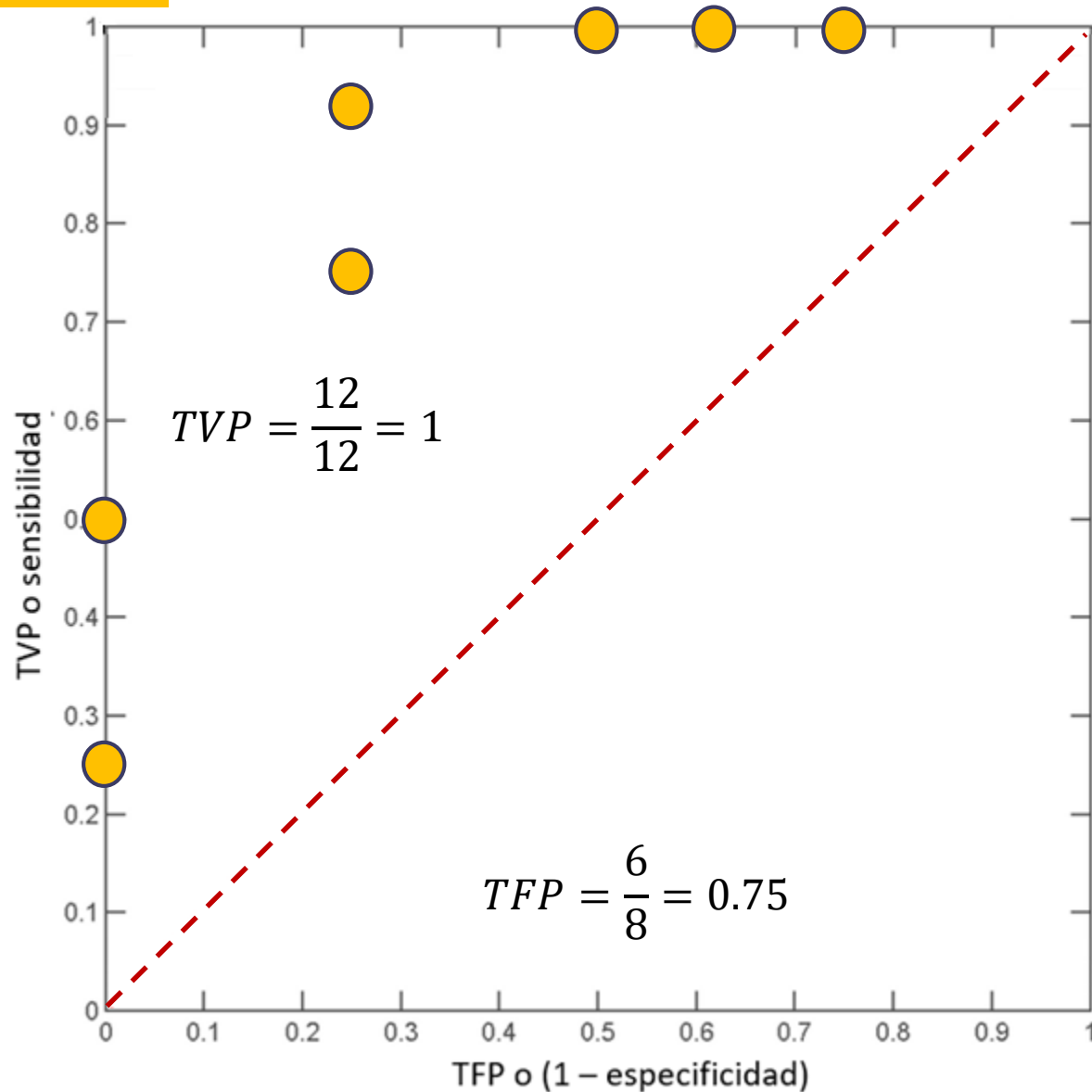


ID	Clase	Confianza	Predice
5	Mina	1	Mina
7	Mina	1	Mina
9	Mina	1	Mina
1	Mina	0.9	Mina
10	Mina	0.9	Mina
20	Mina	0.9	Mina
8	Roca	0.8	Mina
14	Mina	0.8	Mina
15	Mina	0.8	Mina
18	Roca	0.8	Mina
19	Mina	0.8	Mina
3	Mina	0.7	Mina
6	Mina	0.7	Mina
12	Mina	0.65	Mina
4	Roca	0.6	Mina
16	Roca	0.6	Mina
11	Roca	0.5	Mina
2	Roca	0.4	Mina
13	Roca	0.3	Roca
17	Roca	0.1	Roca

12 minas y 8 rocas

Umbral = 0.4

CURVA ROC

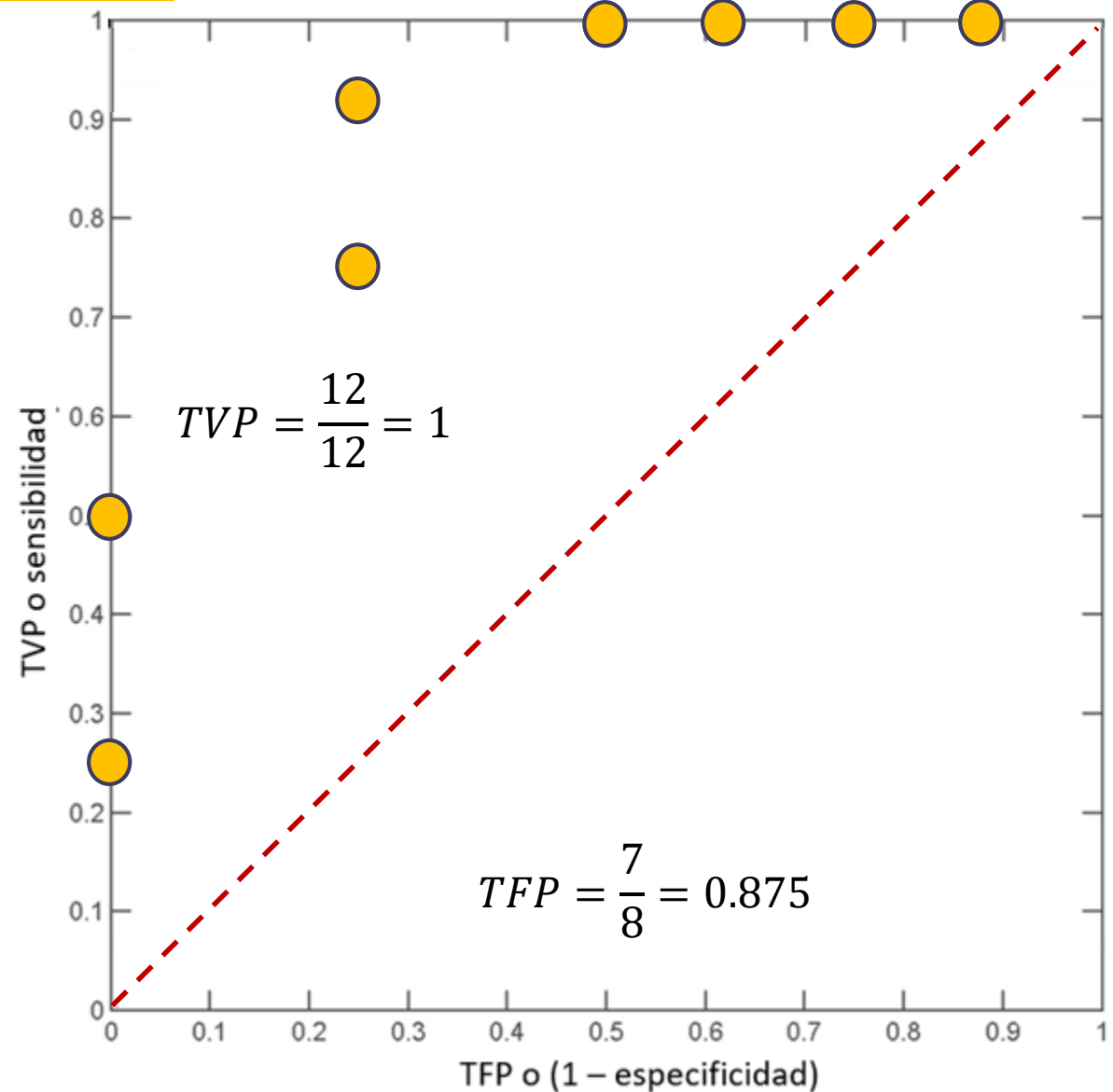


ID	Clase	Confianza	Predice
5	Mina	1	Mina
7	Mina	1	Mina
9	Mina	1	Mina
1	Mina	0.9	Mina
10	Mina	0.9	Mina
20	Mina	0.9	Mina
8	Roca	0.8	Mina
14	Mina	0.8	Mina
15	Mina	0.8	Mina
18	Roca	0.8	Mina
19	Mina	0.8	Mina
3	Mina	0.7	Mina
6	Mina	0.7	Mina
12	Mina	0.65	Mina
4	Roca	0.6	Mina
16	Roca	0.6	Mina
11	Roca	0.5	Mina
2	Roca	0.4	Mina
13	Roca	0.3	Mina
17	Roca	0.1	Roca

12 minas y 8 rocas

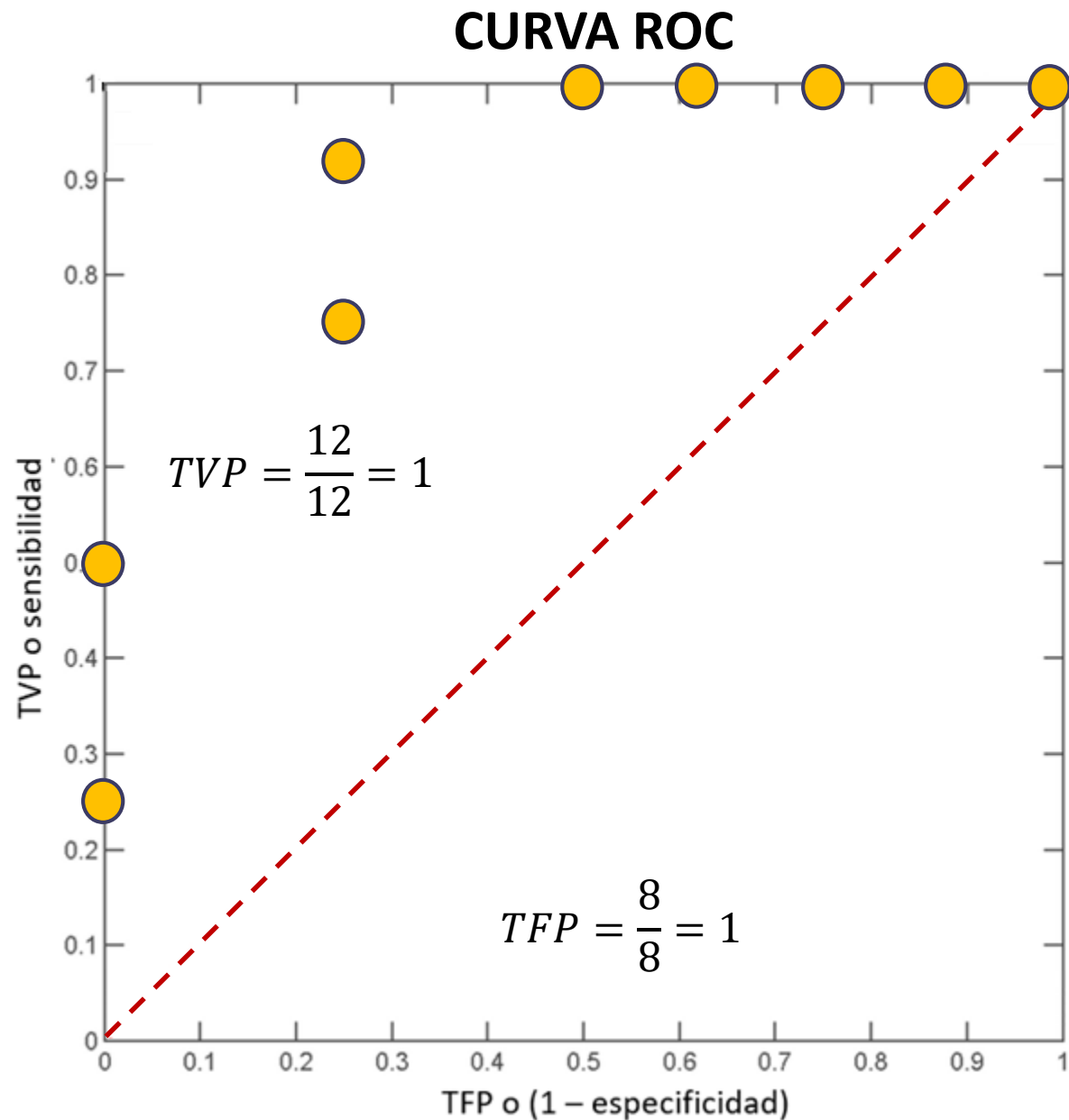
Umbral = 0.3

CURVA ROC



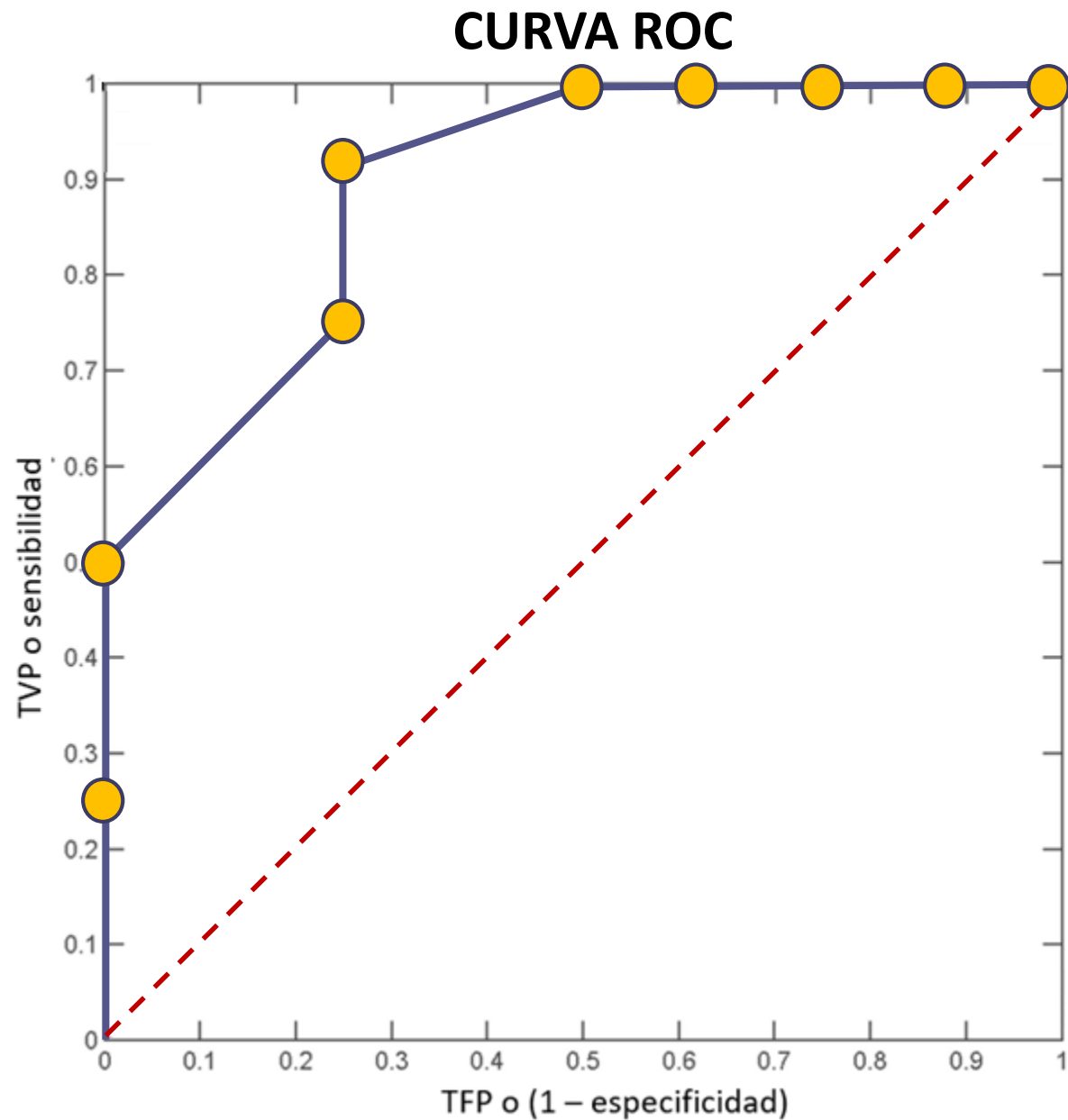
ID	Clase	Confianza	Predice
5	Mina	1	Mina
7	Mina	1	Mina
9	Mina	1	Mina
1	Mina	0.9	Mina
10	Mina	0.9	Mina
20	Mina	0.9	Mina
8	Roca	0.8	Mina
14	Mina	0.8	Mina
15	Mina	0.8	Mina
18	Roca	0.8	Mina
19	Mina	0.8	Mina
3	Mina	0.7	Mina
6	Mina	0.7	Mina
12	Mina	0.65	Mina
4	Roca	0.6	Mina
16	Roca	0.6	Mina
11	Roca	0.5	Mina
2	Roca	0.4	Mina
13	Roca	0.3	Mina
17	Roca	0.1	Mina

12 minas y 8 rocas



ID	Clase	Confianza	Predice
5	Mina	1	
7	Mina	1	
9	Mina	1	
1	Mina	0.9	
10	Mina	0.9	
20	Mina	0.9	
8	Roca	0.8	
14	Mina	0.8	
15	Mina	0.8	
18	Roca	0.8	
19	Mina	0.8	
3	Mina	0.7	
6	Mina	0.7	
12	Mina	0.65	
4	Roca	0.6	
16	Roca	0.6	
11	Roca	0.5	
2	Roca	0.4	
13	Roca	0.3	
17	Roca	0.1	

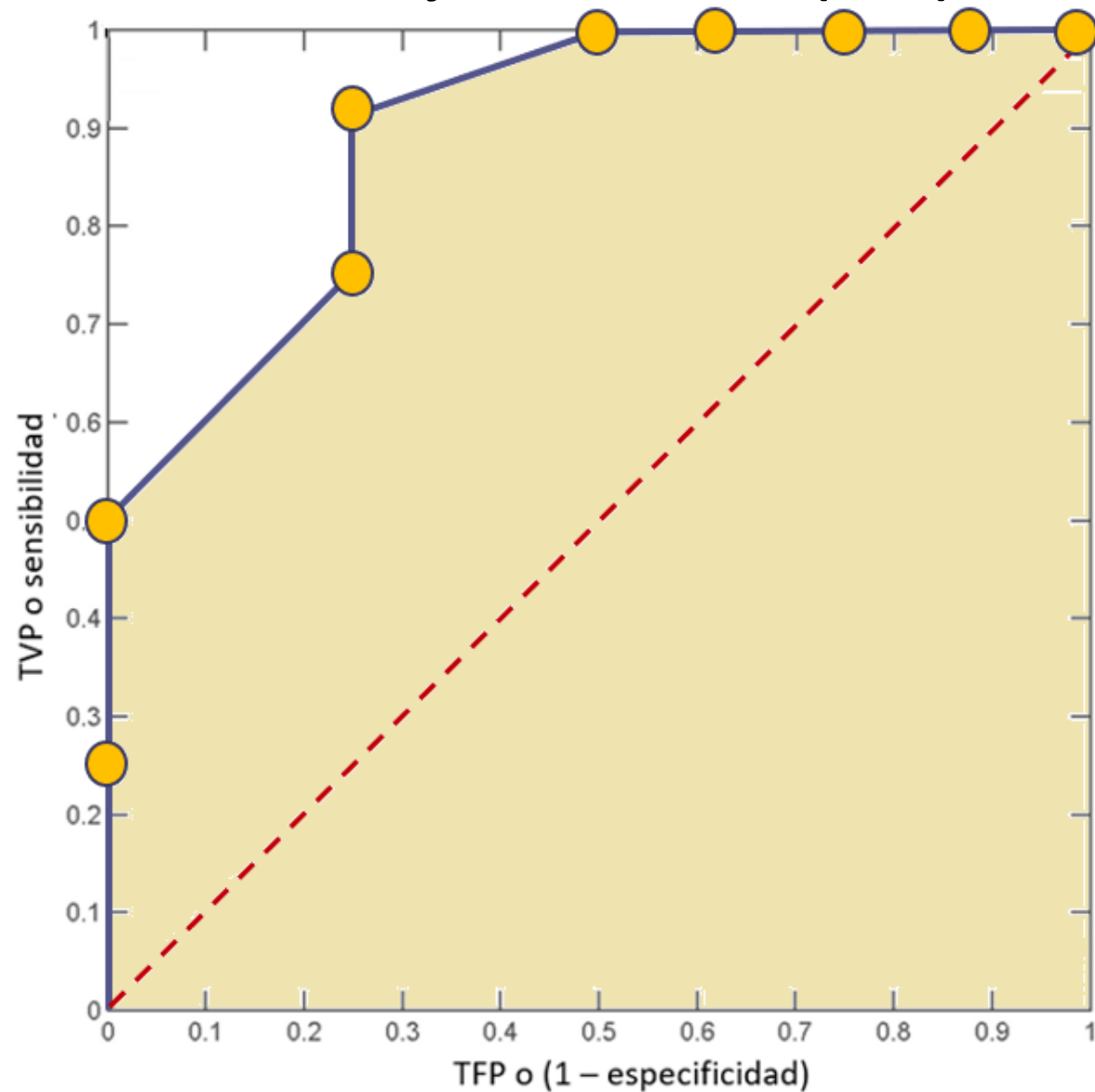
12 minas y 8 rocas



ID	Clase	Confianza	Predice
5	Mina	1	
7	Mina	1	
9	Mina	1	
1	Mina	0.9	
10	Mina	0.9	
20	Mina	0.9	
8	Roca	0.8	
14	Mina	0.8	
15	Mina	0.8	
18	Roca	0.8	
19	Mina	0.8	
3	Mina	0.7	
6	Mina	0.7	
12	Mina	0.65	
4	Roca	0.6	
16	Roca	0.6	
11	Roca	0.5	
2	Roca	0.4	
13	Roca	0.3	
17	Roca	0.1	

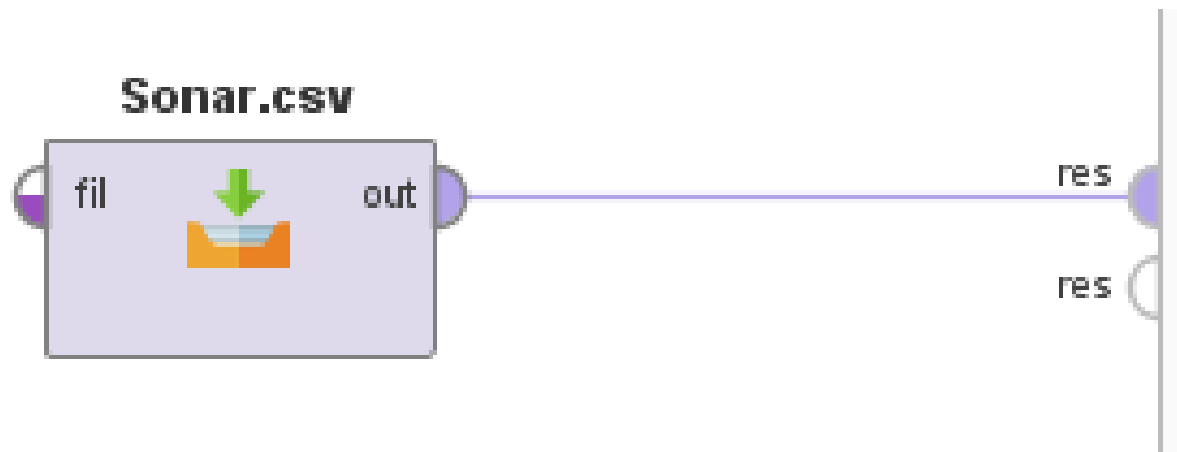
12 minas y **8 rocas**

Area bajo la curva ROC (AUC)



Sonar.csv

- Abra el archivo



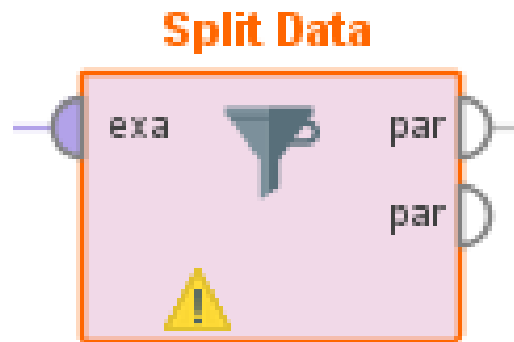
Cómo dividir los datos

- **Lineal::** separar según las proporciones indicadas tomando la ubicación actual de cada ejemplo (orden de aparición).
- **Aleatoria:** mezclar los ejemplos antes de separarlos.
- **Estratificada:** generar conjuntos con la misma proporción de ejemplos de cada clase que el conjunto original.

Cómo vamos a dividir los datos?

- Pueden generarse dos particiones utilizando un porcentaje (operador *Split data*).
- Utilizar Validación Cruzada (*cross validation*) indicando la cantidad de particiones que se desean utilizar. Se usa para evaluar un modelo garantizando que los resultados son independientes de la partición realizada para formar los conjuntos de entrenamiento y testeo.

Generando particiones



Parameters ×

Split Data

partitions Edit Enumerations...

sampling type automatic ▼ ⓘ

Edit Parameter List: partitions ×

Edit Parameter List: **partitions**
The partitions that should be created.

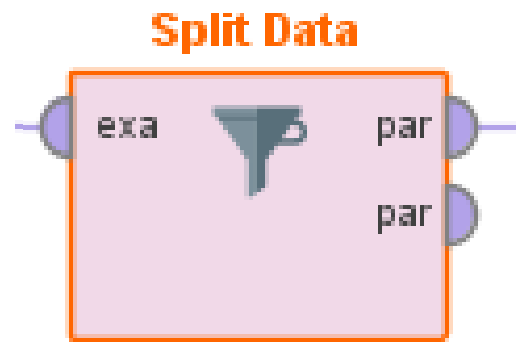
ratio

0.8


0.2


Add Entry Remove Entry OK

Generando particiones



Parameters [X]

 **Split Data**

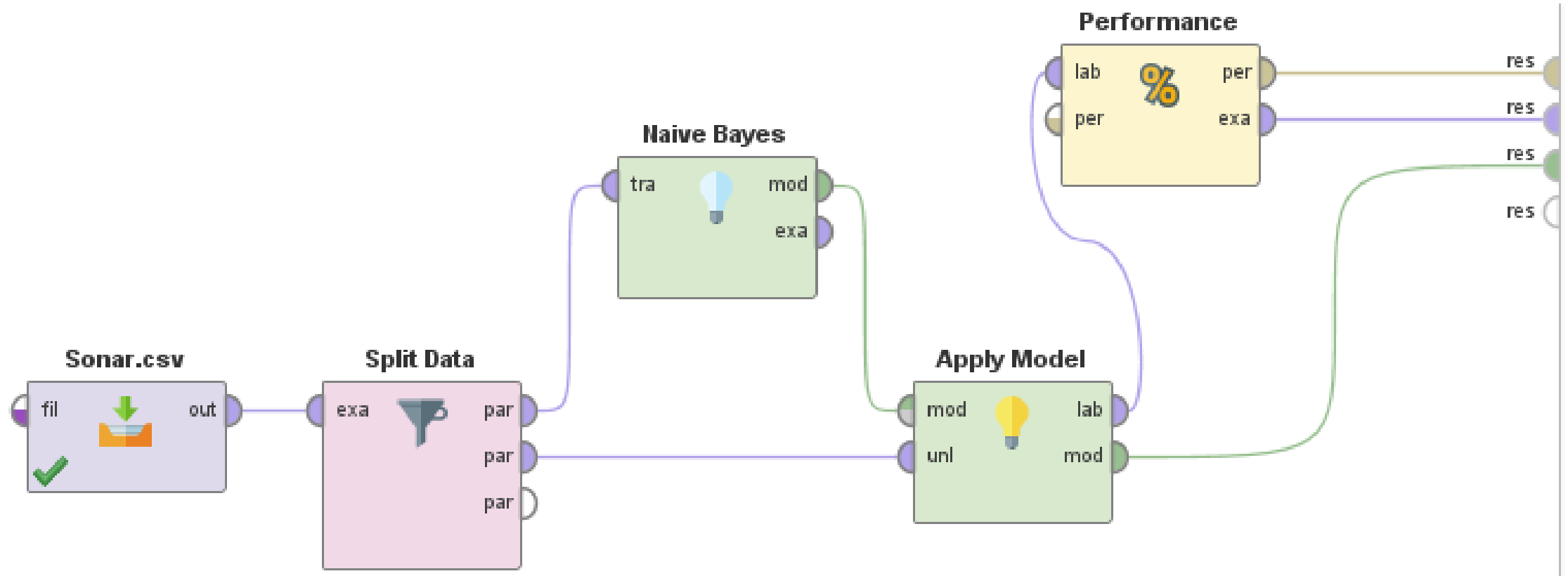
partitions  Edit Enumer...

sampling type automatic ▼

- linear sampling
- shuffled sampling
- stratified sampling
- automatic

☐ use local random

Modelo predictivo (80% train - 20% test)



Matriz de confusión

accuracy: 65.85%

Clase Positiva

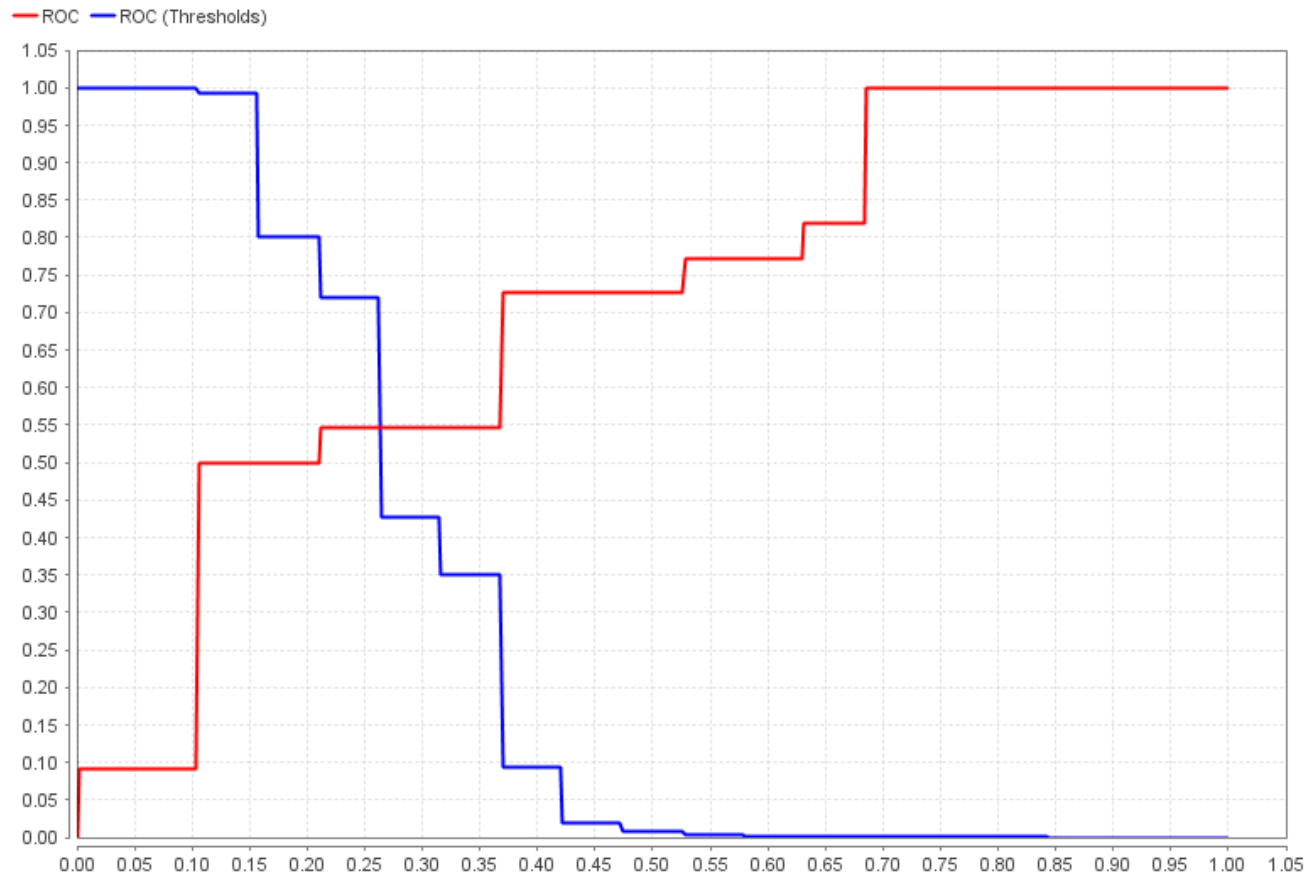
	true Rock	true Mine	class precision
pred. Rock	15	10	60.00%
pred. Mine	4	12	75.00%
class recall	78.95%	54.55%	

↑
TVN
(especificidad)

↑
TVP
(sensibilidad)

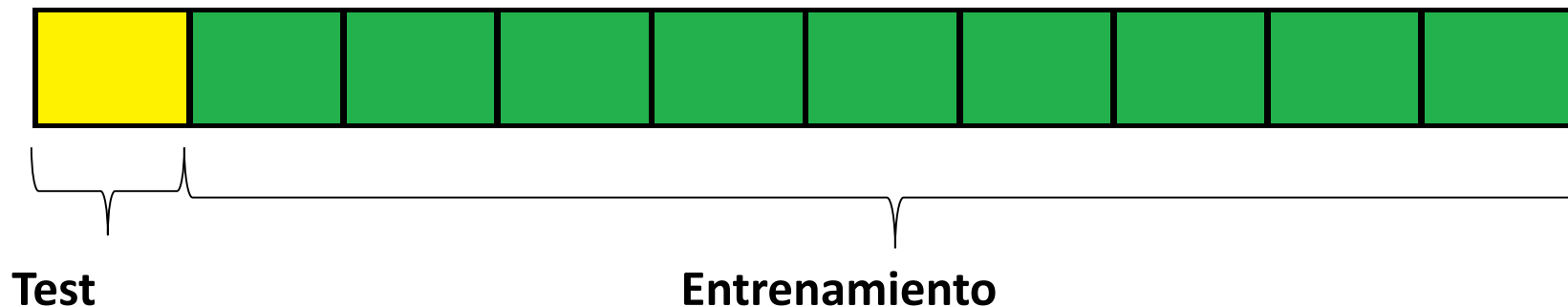
Curva ROC

AUC (optimistic): 0.703 (positive class: Mine)



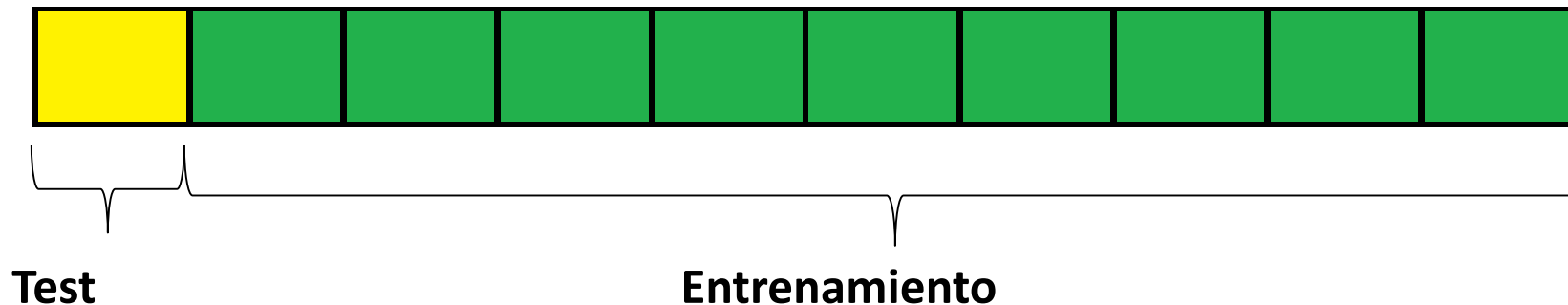
Cross Validation

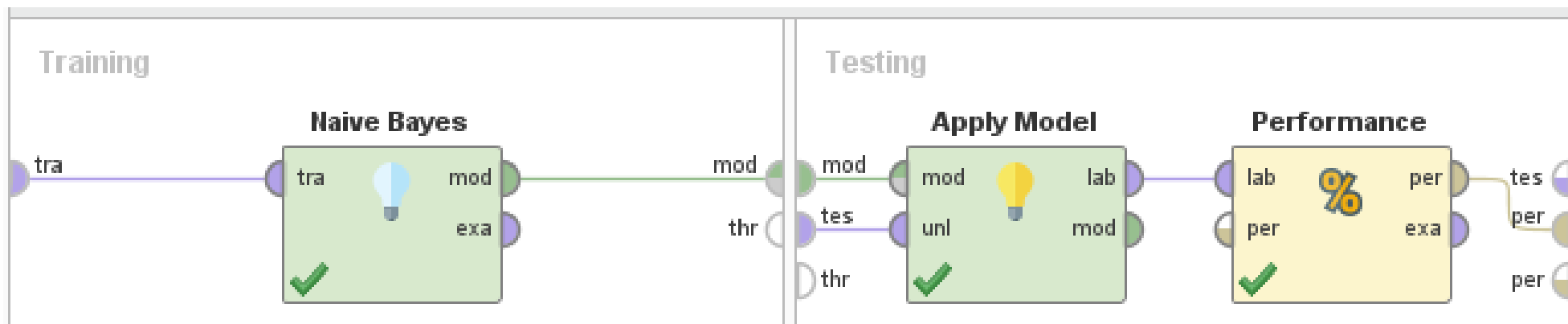
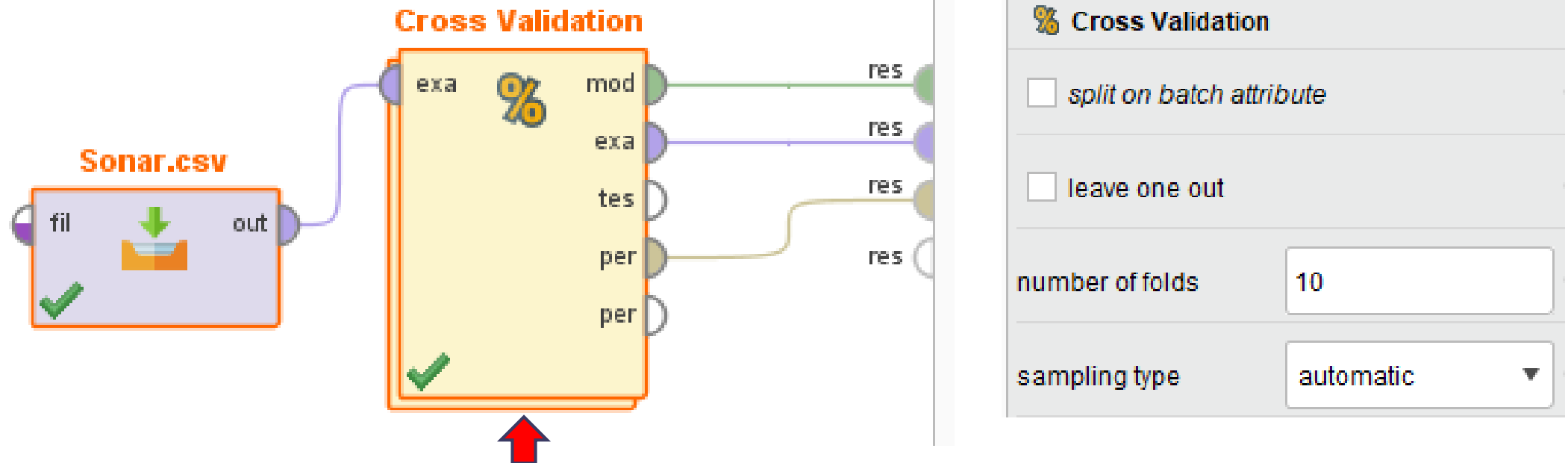
- Dividir el conjunto de ejemplos en **k** partes iguales.
- Repetir para cada una de las partes
 - Designar la parte como conjunto de **test**.
 - Construir un modelo con las partes restantes como conjunto de **entrenamiento**.
 - Evaluar este modelo sobre el conjunto de test.



Cross Validation

- Se obtienen **k** sub-modelos y **k** evaluaciones de la performance.
- El promedio de estas **k** evaluaciones es la estimación de la performance del modelo final.
- Se construye el modelo final usando todas las partes como conjunto de entrenamiento.





Operador Cross Validation

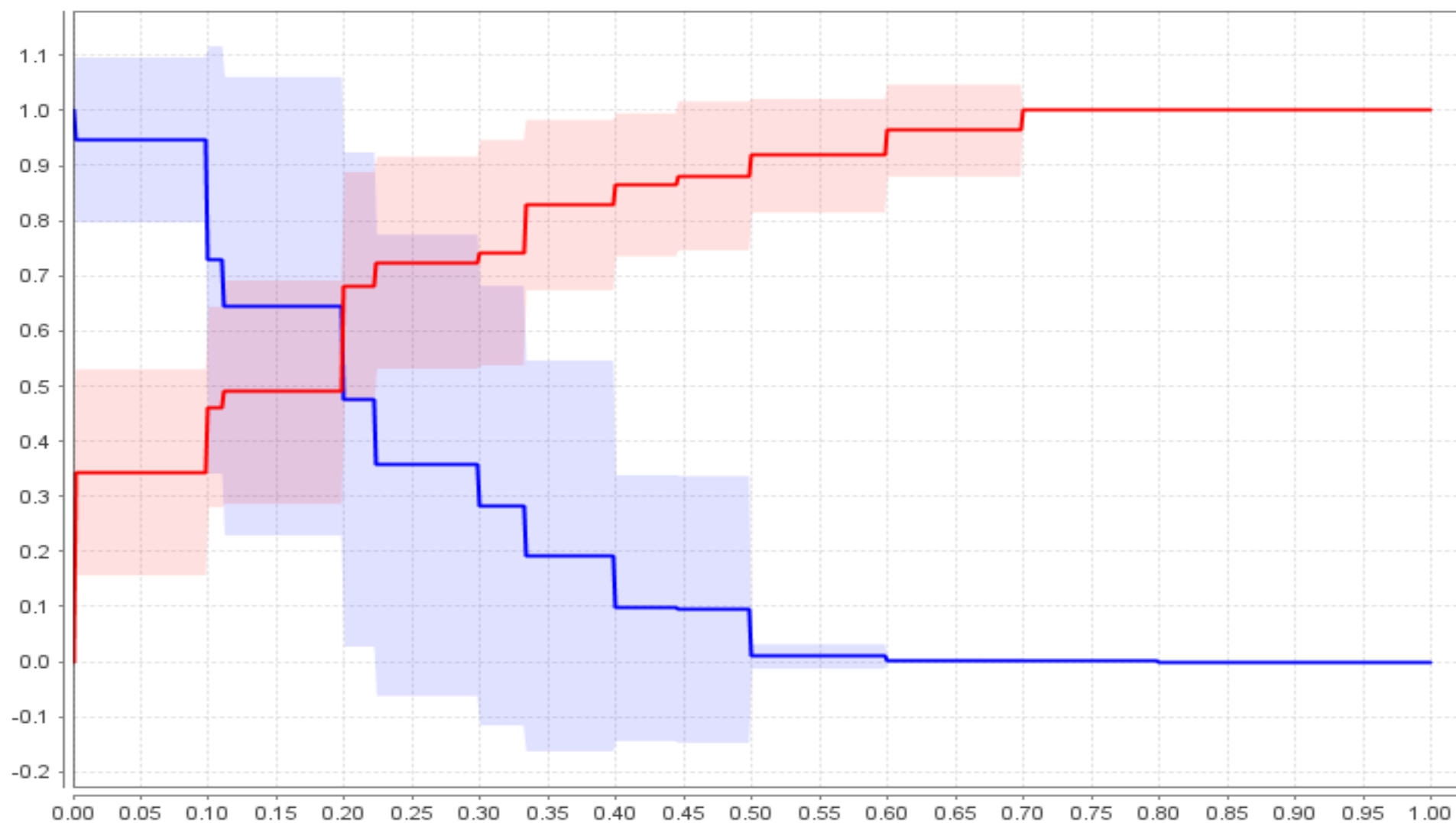
- Matriz de confusión

accuracy: 66.90% +/- 7.66% (micro average: 66.83%)

	true Rock	true Mine	class precision
pred. Rock	78	50	60.94%
pred. Mine	19	61	76.25%
class recall	80.41%	54.95%	

AUC (optimistic): 0.810 +/- 0.083 (micro average: 0.810) (positive class: Mine)

ROC ROC (Thresholds)



Clasificación de flores de Iris

- Se dispone de información de 3 tipos de flores Iris



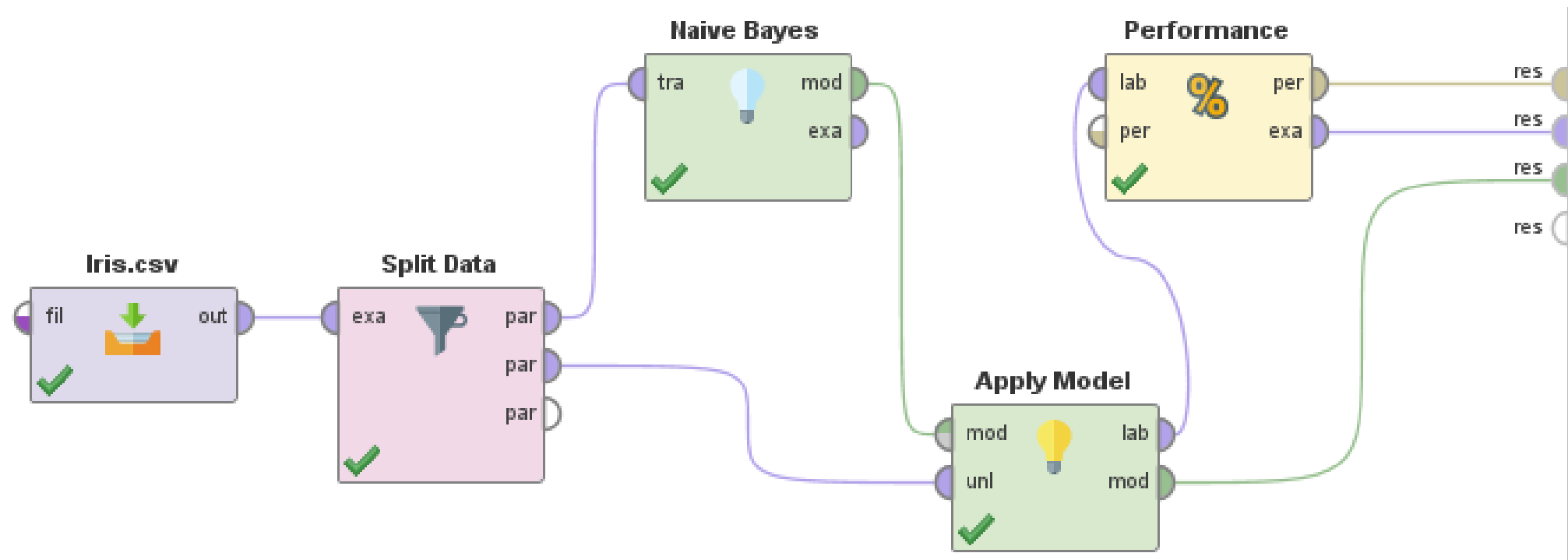
<https://archive.ics.uci.edu/ml/datasets/Iris>

Iris.csv

Id	sepalength	sepalwidth	petallength	petalwidth	class
1	5,1	3,5	1,4	0,2	Iris-setosa
2	4,9	3,0	1,4	0,2	Iris-setosa
...
95	5,6	2,7	4,2	1,3	Iris-versicolor
96	5,7	3,0	4,2	1,2	Iris-versicolor
97	5,7	2,9	4,2	1,3	Iris-versicolor
...
149	6,2	3,4	5,4	2,3	Iris-virginica
150	5,9	3,0	5,1	1,8	Iris-virginica

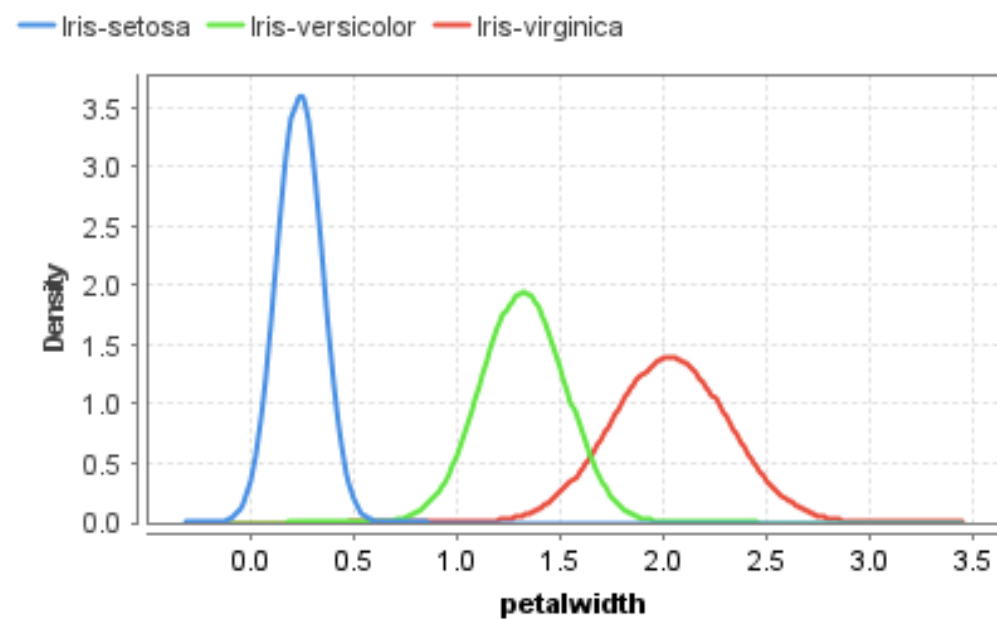
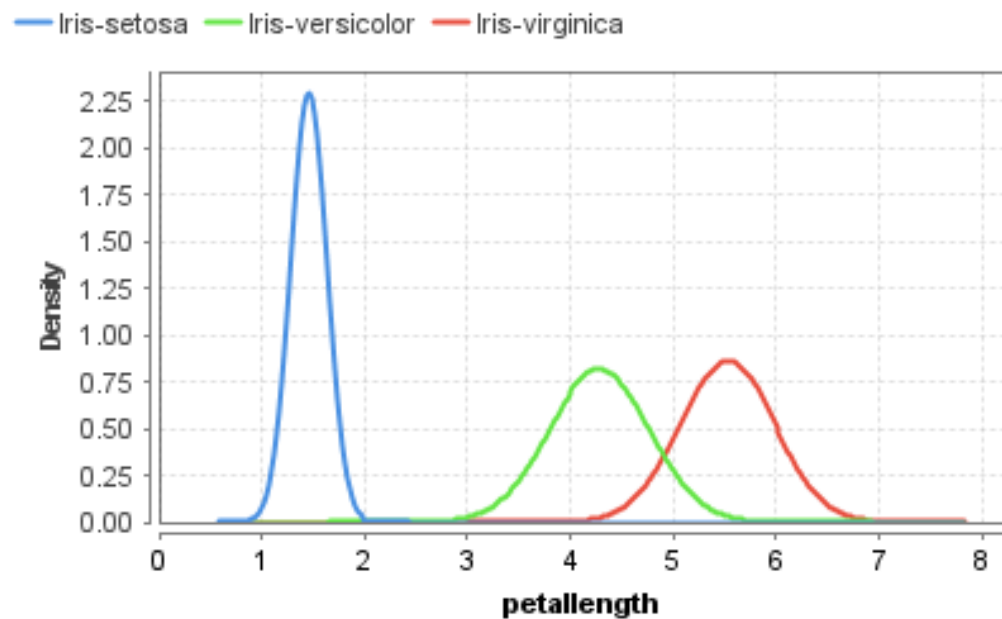
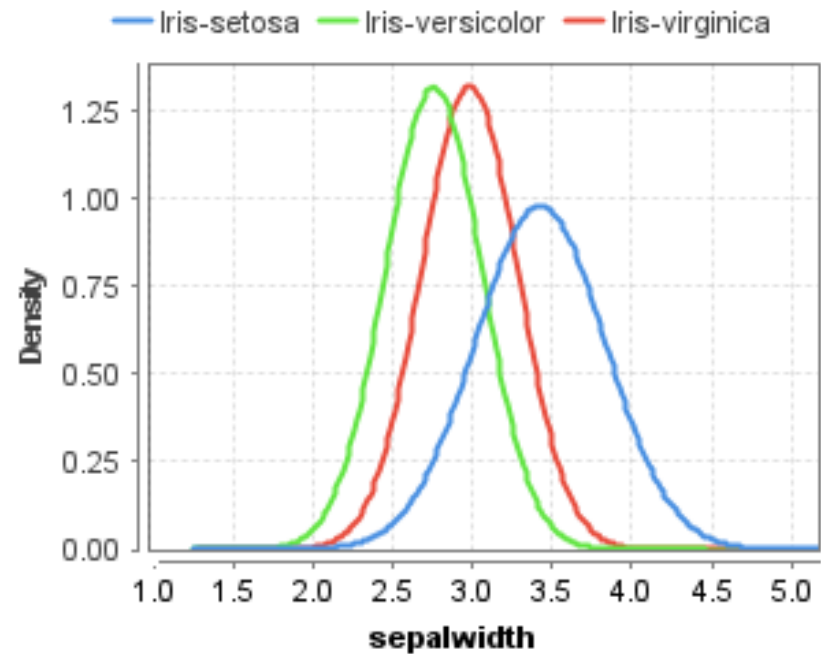
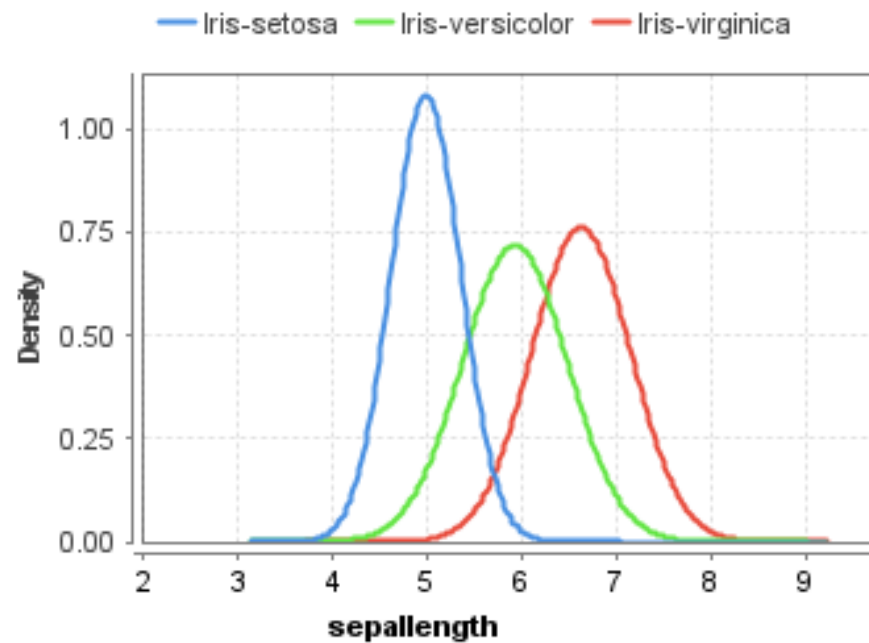
<https://archive.ics.uci.edu/ml/datasets/Iris>

Clasificación de flores de Iris



Clasificador Bayesiano

Attribute	Parameter	Iris-setosa	Iris-versicolor	Iris-virginica
sepalength	mean	4.990	5.930	6.630
sepalength	standard deviation	0.369	0.556	0.525
sepalwidth	mean	3.425	2.758	2.982
sepalwidth	standard deviation	0.407	0.303	0.302
petallength	mean	1.452	4.280	5.527
petallength	standard deviation	0.174	0.487	0.464
petalwidth	mean	0.242	1.323	2.032
petalwidth	standard deviation	0.111	0.206	0.286



Clasificación de flores de Iris

accuracy: 96.67%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	10	0	0	100.00%
pred. Iris-versicolor	0	10	1	90.91%
pred. Iris-virginica	0	0	9	100.00%
class recall	100.00%	100.00%	90.00%	

Ejercicio

- Analice la información referida a distintos tipo de hongos que se encuentra disponible en
<https://archive.ics.uci.edu/ml/datasets/Mushroom>
- Construya un clasificador bayesiano para predecir si se trata de un hongo venenoso o comestible
 - Entrene con el 80% de las muestras y testee con el 20% restante.
 - Compare la tasa de acierto con un clasificador que siempre responde por la clase mayoritaria (Operador W-ZeroR)