

Clasificación: Aspectos Prácticos

Medidas para evaluar y comparar el
rendimiento de los clasificadores

Inteligencia de Negocio
Curso 2024-2025

E.T.S. de Ingenierías Informática y de Telecomunicación
Universidad de Granada

Caso Práctico

- Conjunto de datos sobre Cáncer de Mama: *Breast Cancer Dataset*
(<https://archive.ics.uci.edu/ml/datasets/breast+cancer>)
 - Atributos: 9 atributos que caracterizan a 286 mujeres que han sobrevivido a un cáncer de mama
 - Clase: Si hay recurrencia del cáncer de mama tras 5 años
 - 201 mujeres no sufrieron recurrencia del cáncer, 85 sí
- Una vez que construyamos un modelo de predicción, tendremos que responder la pregunta: ¿Cómo de bueno es este modelo?

Precisión y Matriz de Confusión

- Precisión o *Accuracy*: dividir las predicciones correctas entre el número total de predicciones.
 - Ejemplo: $Acc = 0.803$ significa que nuestro modelo acierta en el 80.3% de los casos

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

- Matriz de Confusión

		PREDICCIÓN	
		POSITIVO	NEGATIVO
CLASE REAL	POSITIVO	True Positives (TP)	False Negatives (FN)
	NEGATIVO	False Positives (FP)	True Negatives (TN)

Paradoja del *Accuracy*

- Suponemos que evaluamos tres modelos para nuestro problema de cáncer de mama:
 - Modelo 1: Predice siempre 'No' $\rightarrow Acc = 201/286 = 0.7028$
 - Modelo 2: Predice siempre 'Sí' $\rightarrow Acc = 85/286 = 0.2972$
 - Modelo 3: Árbol de decisión $\rightarrow Acc = 0.6923$

¿Cuál es el mejor modelo?

Paradoja del *Accuracy*

Predice siempre 'No'

Clase / pred	NO	SI
NO	201	0
SI	85	0
Total	286	0

Predice siempre 'Sí'

Clase / pred	NO	SI
NO	0	201
SI	0	85
Total	0	286

Árbol de decisión

Clase / pred	NO	SI
NO	188	13
SI	75	10
Total	263	23

Paradoja del *Accuracy*

- ¿Es siempre *accuracy* una buena medida del *poder predictivo*?
- ¿Por qué en este problema el *accuracy* no es efectivo?

EL PROBLEMA DE LAS CLASES NO BALANCEADAS

Cuando las clases no están completamente balanceadas conviene emplear otras medidas de evaluación que tienen en cuenta la distribución del error entre clases

Sensitivity y Specificity

- True Positive Rate (TPR), *Recall* o *Sensitivity*: predicciones positivas correctas entre el nº total de positivos

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

- True Negative Rate (TNR) o *Specificity*: predicciones negativas correctas entre el nº total de negativos

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP}$$

False Positive Rate y False Negative Rate

- False Positive Rate (FPR): predicciones positivas incorrectas entre nº total de negativos

$$\mathbf{FPR = \frac{FP}{N} = \frac{FP}{TN + FP}}$$

- False Negative Rate (FNR): predicciones negativas incorrectas entre nº total de positivos

$$\mathbf{FNR = \frac{FN}{P} = \frac{FN}{TP + FN}}$$

Positive Predictive Value

- Positive Predictive Value (PPV) o *Precision*: predicciones positivas correctas entre el nº total de predicciones positivas

$$PPV = \frac{TP}{TP + FP}$$

Resultados Caso Práctico

	TPR	TNR	FPR	FNR
Siempre NO	0	1	0	1
Siempre SI	1	0	1	0
Árbol de decisión	0,12	0,94	0,065	0,85

Ninguna de estas métricas por sí misma, individualmente, es representativa del poder predictivo del clasificador. Es necesario combinarlas

Receiver Operating Characteristic Curve (Curva ROC)

- En problemas complejos un clasificador aumentará el número de *True Positives* (TP) a costa de incrementar también el de *False Positives* (FP) –no al mismo ritmo
- Se busca un clasificador que sea capaz de incrementar TP a un ritmo (mucho) mayor que FP

Receiver Operating Characteristic Curve (Curva ROC)

- Los gráficos ROC son gráficos bidimensionales en los cuales se representa FPR (*False Positive Rate*) en el eje X y TPR (*True Positive Rate*) en el eje Y
- Un gráfico ROC muestra el compromiso entre beneficio (*True Positives*) y coste (*False Positives*)

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861-874.

Receiver Operating Characteristic Curve (Curva ROC)

Dos tipos de clasificadores:

- A. Discretos: solo devuelven una etiqueta de clase para cada ejemplo de test
- B. No discretos: devuelven probabilidades o *scores*

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861-874.

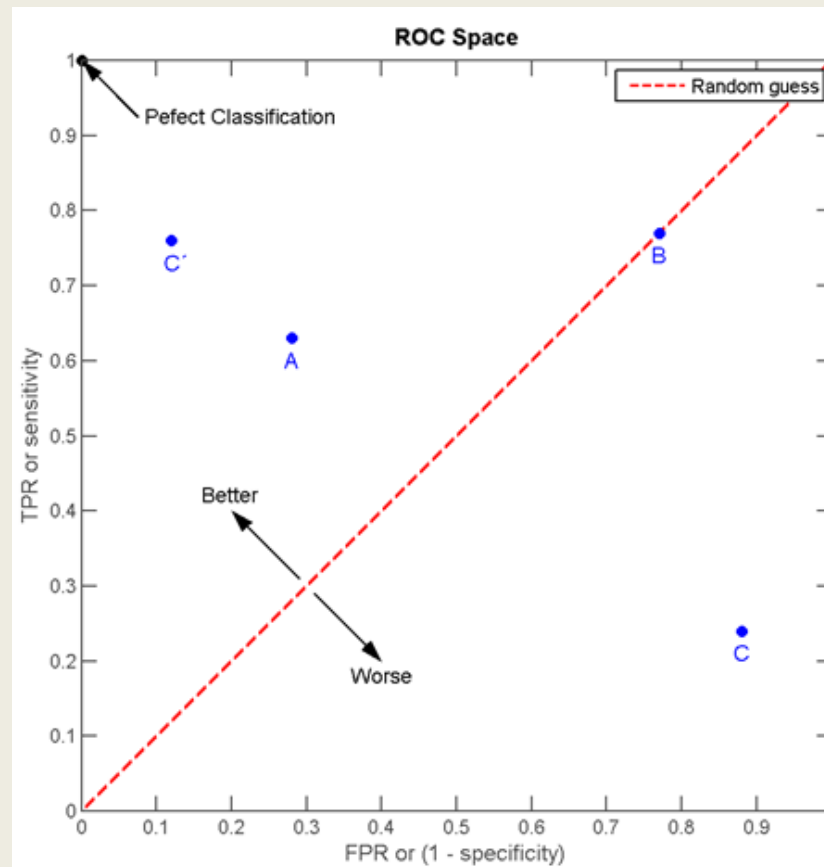
Receiver Operating Characteristic Curve (Curva ROC)

Dos tipos de clasificadores:

- A. Discretos: solo devuelven una etiqueta de clase para cada ejemplo de test
 - Tienen valores concretos de TPR y FPR que se corresponden con un único punto en el espacio ROC (*punto ROC*)
 - Un punto es mejor que otro si está al noroeste de él

Curva ROC

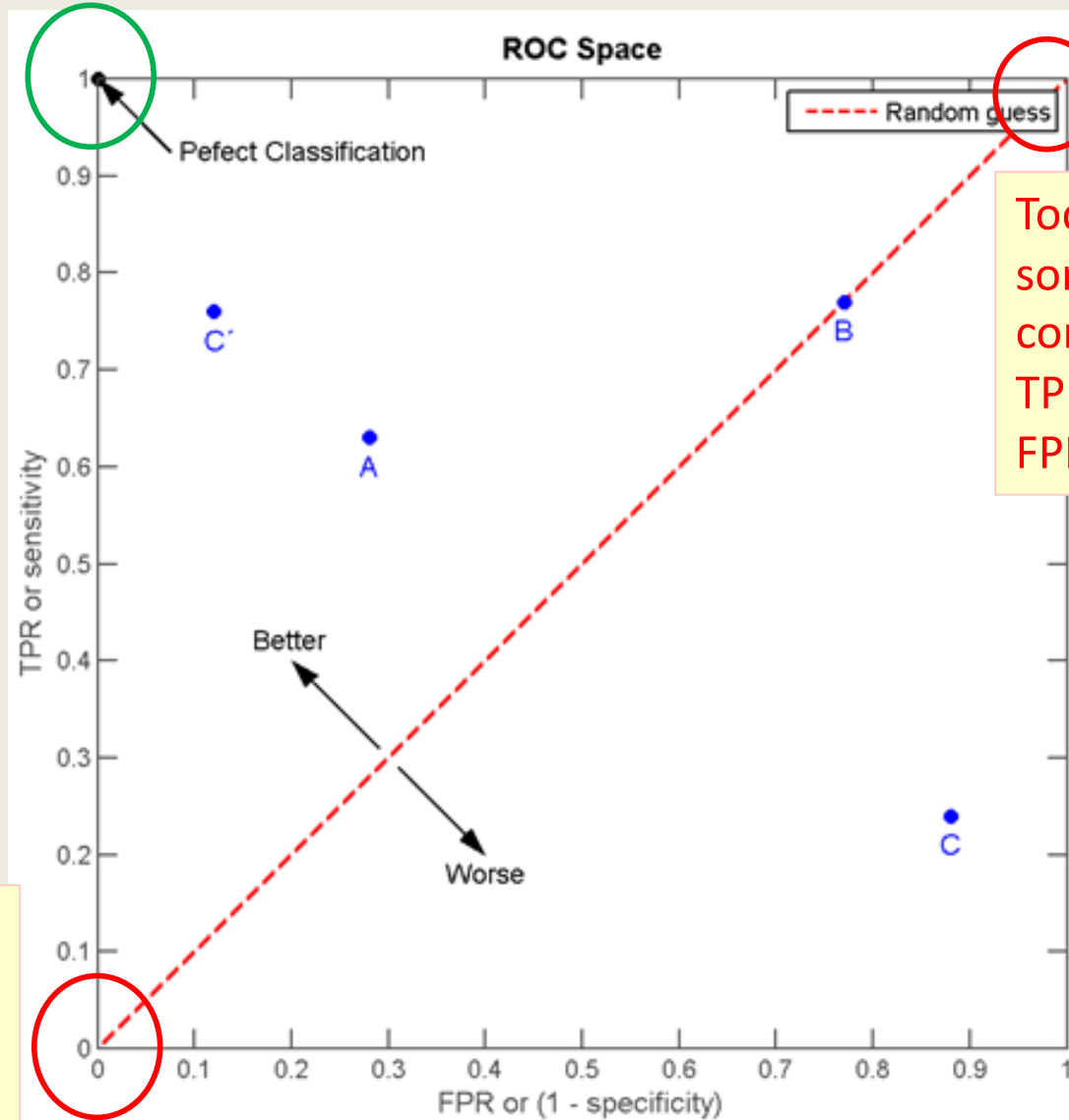
- Clasificadores discretos: a cada clasificador le corresponde un punto en el espacio ROC



By ROC_space.png: Indonderivative work: Kai walz (talk) - ROC_space.png, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=8326140>

Curva ROC

Clasificación
Perfecta



Todos los ejemplos
son clasificados
como positivos
 $TPR = 1$
 $FPR = 1$

Ningún ejemplo es
clasificado como
positivo
 $TPR = 0$
 $FPR = 0$

Curva ROC

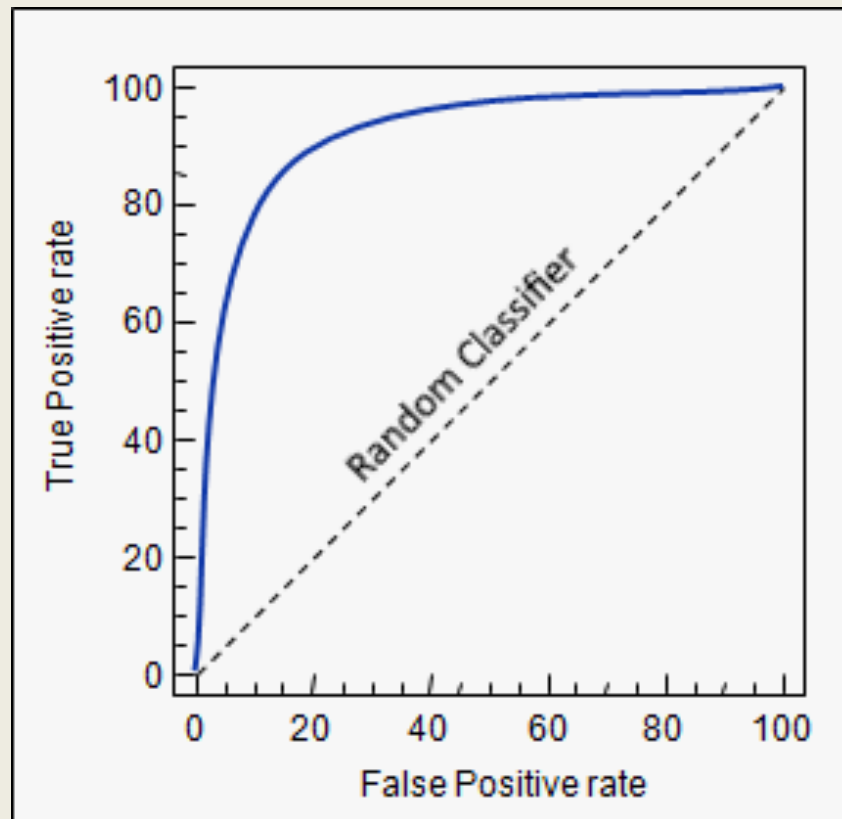
Dos tipos de clasificadores:

B. Clasificadores no discretos: devuelven probabilidades o scores

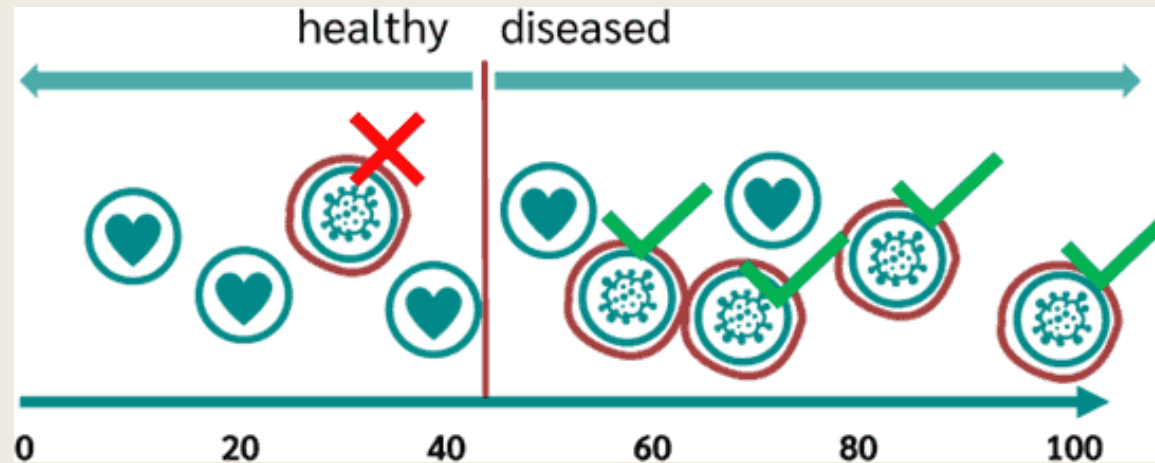
- Se pueden discretizar fijando un umbral
- Conceptualmente, si variamos dicho umbral desde $-\infty$ hasta $+\infty$ podemos dibujar los puntos de la curva ROC
- Para cada valor del umbral tengo una discretización distinta (como un clasificador discreto distinto)

Curva ROC

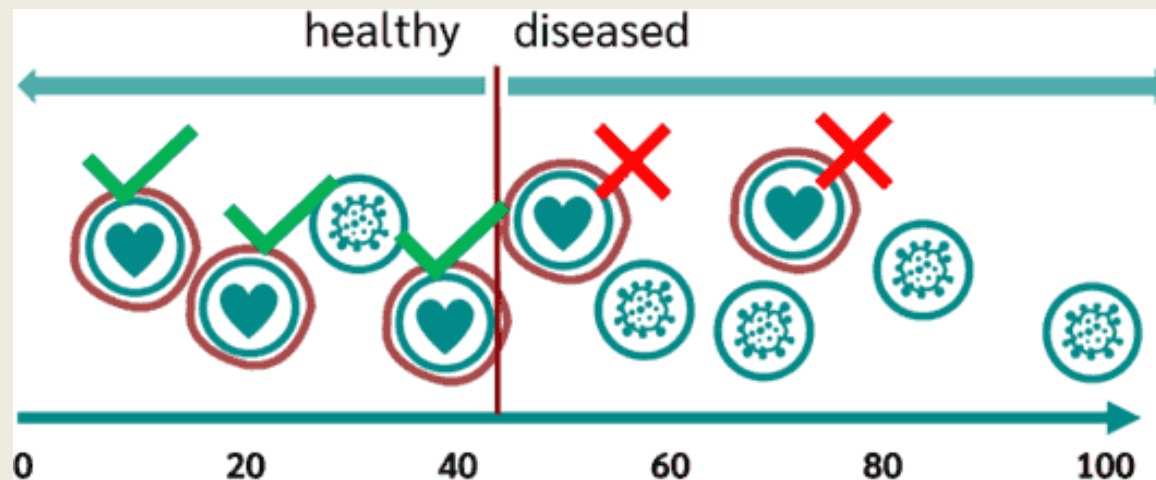
- Clasificadores no discretos: para cada posible valor del umbral tengo un TPR y un FPR



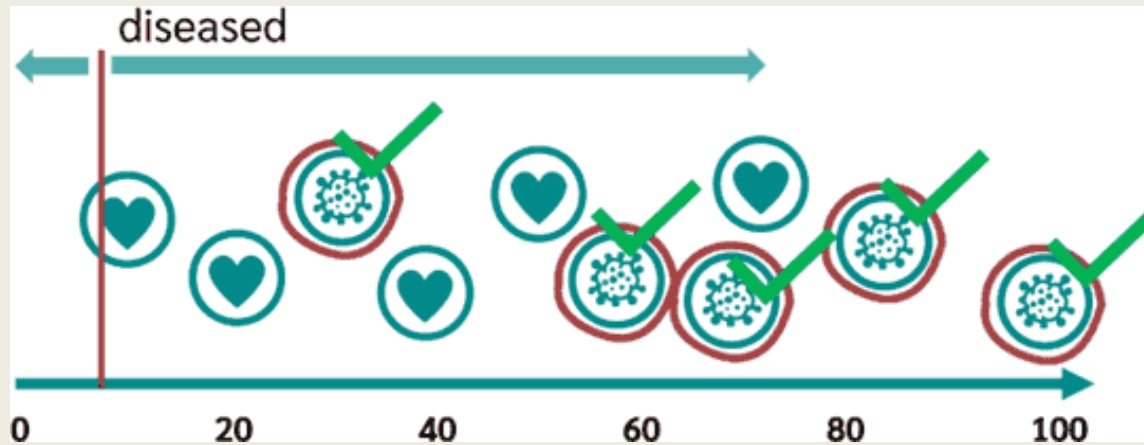
Curva ROC



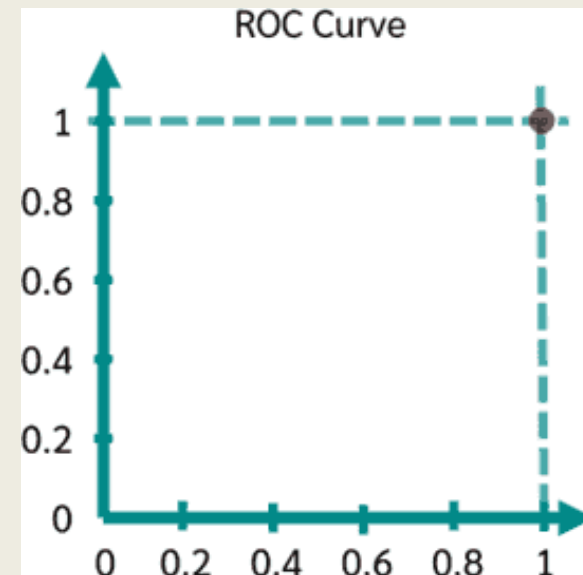
Umbral 45%



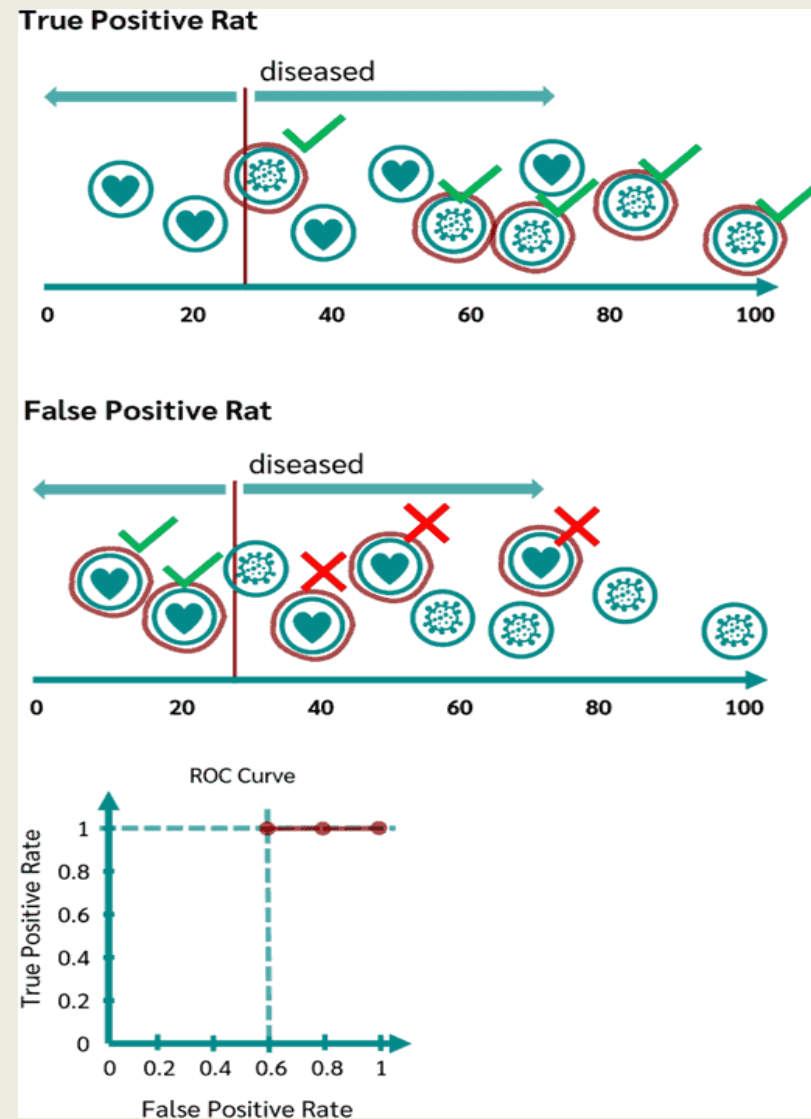
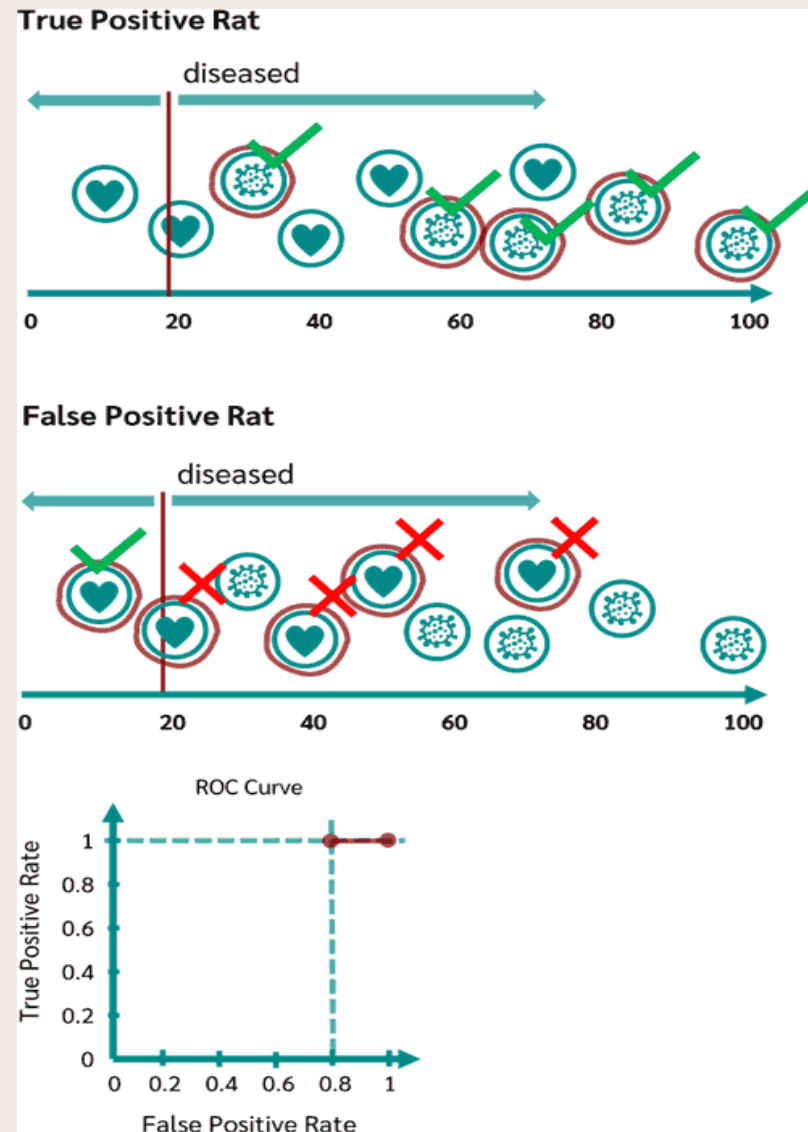
Curva ROC



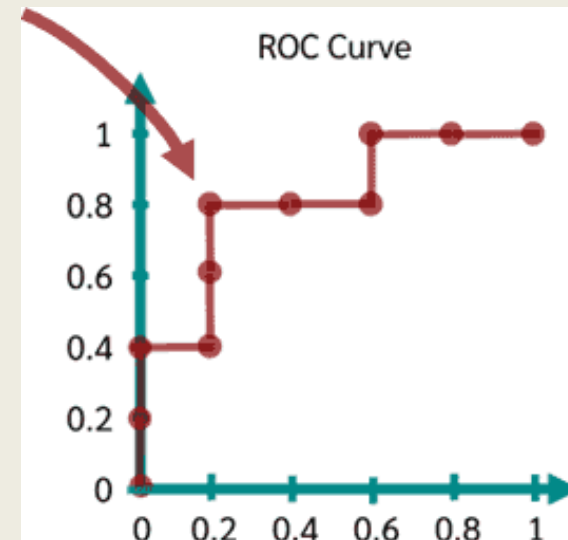
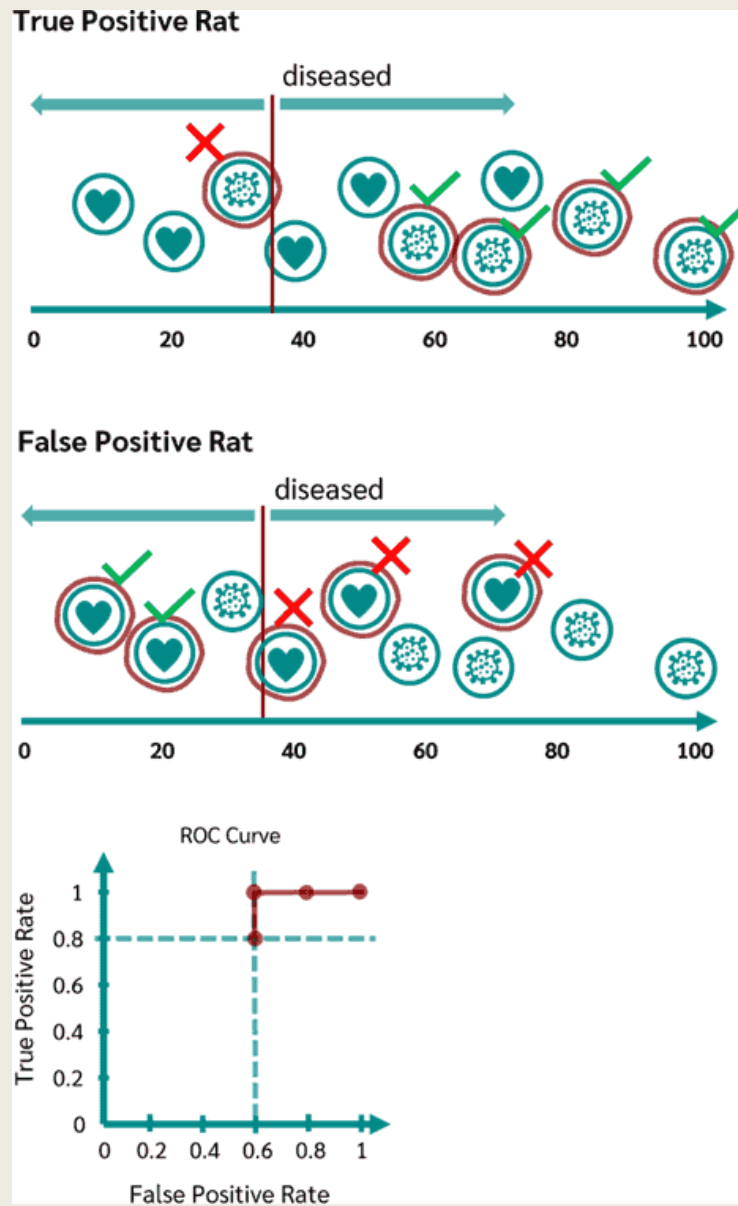
Umbral Muy Bajo



Curva ROC



Curva ROC



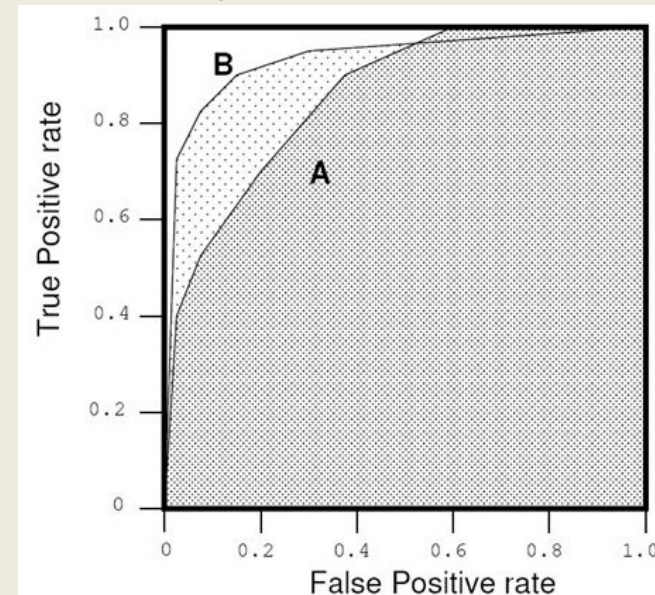
Area Under Curve (AUC)

- Una curva ROC es una representación bidimensional del rendimiento de un clasificador
- El área bajo la curva ROC (AUC) permite representar en un único valor el rendimiento del clasificador. Esto puede resultar útil para realizar comparativas entre clasificadores

$$AUC = \frac{1 + TPR - FPR}{2}$$

Máximo AUC posible = 1

AUC del clasificador aleatorio = 0.5



Resultados Caso Práctico

$$AUC = \frac{1 + TPR - FPR}{2}$$

	TPR	TNR	FPR	FNR	AUC
Siempre NO	0	1	0	1	0,5
Siempre SÍ	1	0	1	0	0,5
Árbol de decisión	0,12	0,94	0,065	0,85	0,53

AUC trata de maximizar el acierto en ambas clases con un buen balance entre ambas

G-mean

- *G-mean*: media geométrica de TPR y TNR

$$Gmean = \sqrt{\frac{TP}{TP + FN} \cdot \frac{TN}{TN + FP}} = \sqrt{TPR \cdot TNR}$$

G-mean trata de maximizar el acierto en ambas clases con un buen balance entre ambas

Resultados Caso Práctico

$$Gmean = \sqrt{\frac{TP}{TP + FN} \cdot \frac{TN}{TN + FP}} = \sqrt{TPR \cdot TNR}$$

	TPR	TNR	FPR	FNR	AUC	G-mean
Siempre NO	0	1	0	1	0,5	0
Siempre Sí	1	0	1	0	0,5	0
Árbol de decisión	0,12	0,94	0,065	0,85	0,53	0,33

G-mean, al igual que AUC, trata de maximizar el acierto en ambas clases (Siempre NO y Siempre Sí coinciden en ambas medidas)

Sin embargo, en ciertos problemas puede que consideremos más relevantes los errores en una clase que en otra

F1-score

- *F1-score*: media armónica de PPV y TPR (también llamado *F-score* o *F-measure*)

$$F1 = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR}$$

$$PPV = \frac{TP}{TP + FP}$$

$$TPR = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

Resultados Caso Práctico

$$F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

	TPR	TNR	FPR	FNR	AUC	G-mean	F1
Siempre NO	0	1	0	1	0,5	0	0
Siempre Sí	1	0	1	0	0,5	0	0,46
Árbol de decisión	0,12	0,94	0,065	0,85	0,53	0,33	0,19

F1-score penaliza más los errores al clasificar ejemplos positivos (FN) que los errores al clasificar ejemplos negativos (FP) [Siempre Sí es el que consigue mejor F1]

G-measure

- *G-measure*: media geométrica de PPV y TPR

$$Gmeasure = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}} = \sqrt{PPV \cdot TPR}$$

Resultados Caso Práctico

$$Gmeasure = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}} = \sqrt{PPV \cdot TPR}$$

	TPR	TNR	FPR	FNR	AUC	G-mean	F1	G-measure
Siempre NO	0	1	0	1	0,5	0	0	0
Siempre Sí	1	0	1	0	0,5	0	0,46	0,297
Árbol de decisión	0,12	0,94	0,065	0,85	0,53	0,33	0,19	0,052

G-measure, de forma similar a F1-score, penaliza más los errores al clasificar ejemplos positivos (FN) que los errores al clasificar ejemplos negativos (FP) [Siempre Sí es el que consigue un G-measure más alto]

Bibliografía

- <https://sci2s.ugr.es/imbalanced>
- <https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/>
- [https://en.wikipedia.org/wiki/Sensitivity and specificity](https://en.wikipedia.org/wiki/Sensitivity_and_specificity)
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861-874
- [https://en.wikipedia.org/wiki/Receiver operating characteristic#Area under the curve](https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_the_curve)
- [https://en.wikipedia.org/wiki/F1 score](https://en.wikipedia.org/wiki/F1_score)
- [https://en.wikipedia.org/wiki/Fowlkes%E2%80%93Mallows index](https://en.wikipedia.org/wiki/Fowlkes%E2%80%93Mallows_index)