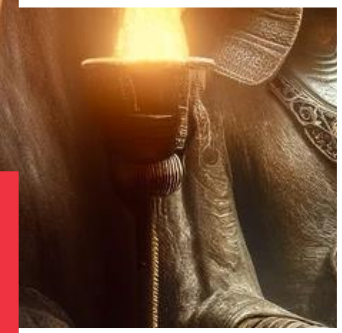
THE BRIDGE

Problemas de Clasificacion

Concepto y Métricas (I)

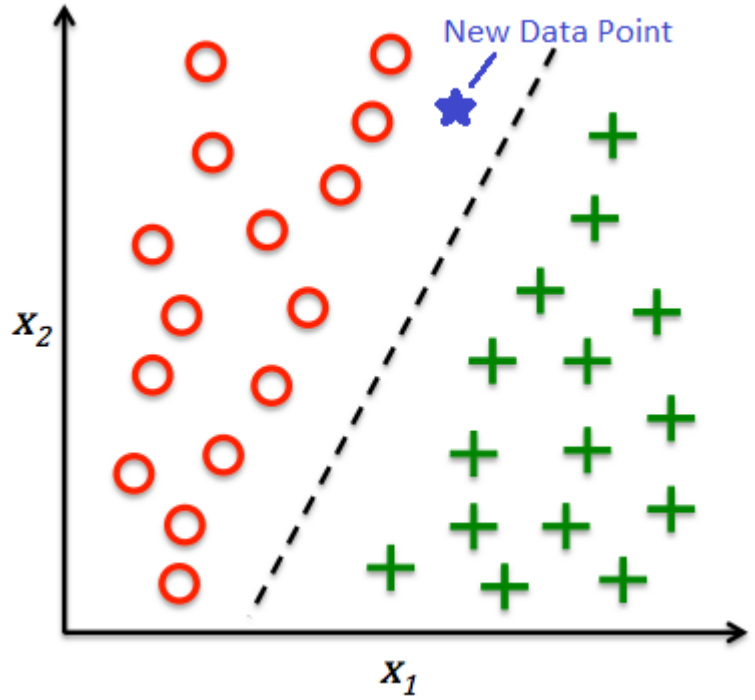


Algoritmos de clasificación

Los algoritmos de clasificación son algoritmos de aprendizaje supervisado cuyo objetivo es predecir etiquetas de clase categóricas de las nuevas instancias.

Dos tipos principales:

- *Clasificación binaria*: solo hay dos clases posibles. Ejemplo: correo spam o no spam (1 o 0)
- *Clasificación multi-clase*: más de dos clases. Ejemplo: identificación de dígitos (0 a 9)



Algoritmos de clasificación más comunes

Regresión logística
Árbol de decisión
KNN
Naive Bayes
SVC
Random Forest
Deep Learning

Métricas: Accuracy

Métricas: Accuracy

Simplemente cantidad de aciertos vs fallos.

Accuracy = nº aciertos en predicción/total muestras predicción

$$\text{Accuracy} = \frac{correct\ predictions}{all\ predictions}$$

¿Cómo se qué clasificador es el mejor? El que tenga un accuracy mas alto… Veamos si es así

Métricas: Accuracy

Imagina que tienes pacientes en una consulta y el objetivo es clasificar si tienen diabetes o no. Creamos un modelo que dadas las características e historial de los pacientes clasifica los pacientes en diabéticos y no diabéticos

Métricas: Accuracy

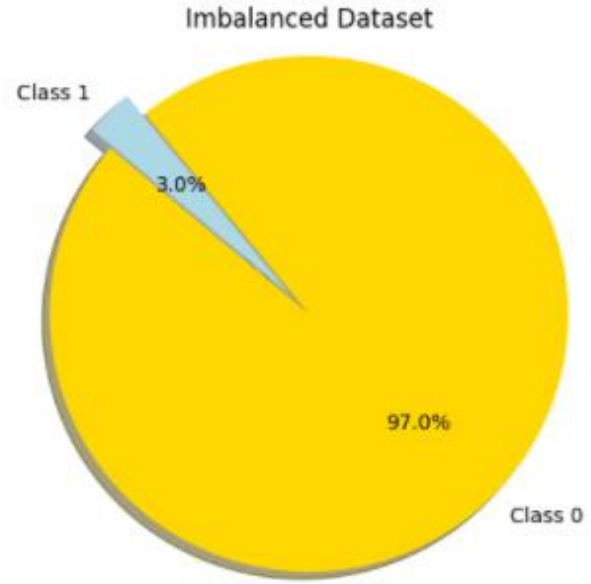
Imagina que tienes pacientes en una consulta y el objetivo es clasificar si tienen diabetes o no. Creamos un modelo que dadas las características e historial de los pacientes clasifica los pacientes en diabéticos y no diabéticos

Calculamos el accuracy: 97% de precisión.
Que modelo más bueno!!!

Métricas: Accuracy

Imagina que tienes pacientes en una consulta y el objetivo es clasificar si tienen diabetes o no. Creamos un modelo que dadas las características e historial de los pacientes clasifica los pacientes en diabéticos y no diabéticos

Calculamos el accuracy: 97% de precisión.
Que modelo más bueno!!!

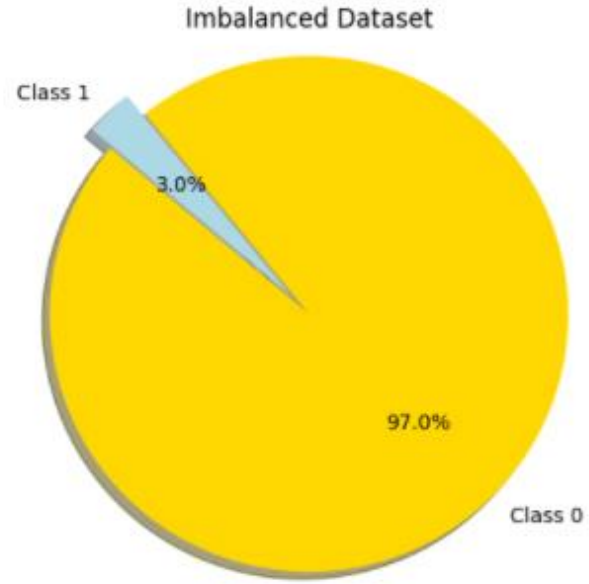


Métricas: Accuracy

Imagina que tienes pacientes en una consulta y el objetivo es clasificar si tienen diabetes o no. Creamos un modelo que dadas las características e historial de los pacientes clasifica los pacientes en diabéticos y no diabéticos

Calculamos el accuracy: 97% de precisión.
Que modelo más bueno!!!

El objetivo del clasificador es que diferencie bien entre las dos clases



Métricas: Accuracy

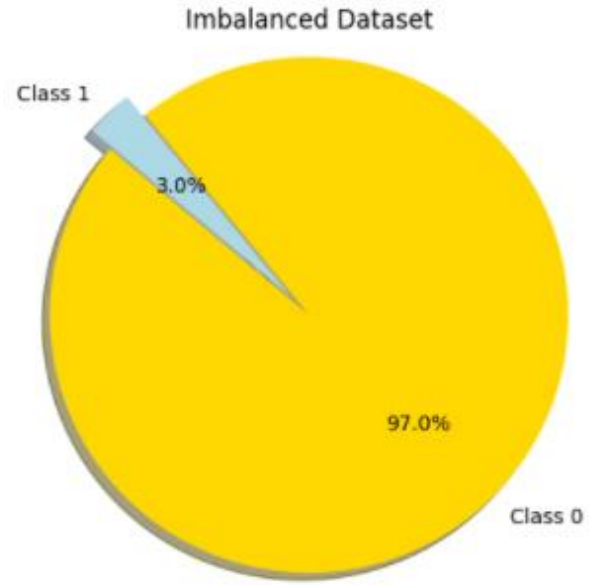
Imagina que tienes pacientes en una consulta y el objetivo es clasificar si tienen diabetes o no. Creamos un modelo que dadas las características e historial de los pacientes clasifica los pacientes en diabéticos y no diabéticos

Calculamos el accuracy: 97% de precisión.
Que modelo más bueno!!!

El objetivo del clasificador es que diferencie bien entre las dos clases

¿Posibles soluciones?

Cambiar la métrica
Conseguir más datos :)
Resampling: o bien ponemos copias de los elementos de la clase desfavorecida, o eliminamos registros de la más poblada
Generar datos sintéticos



Matriz de confusión

Muy útil sobre todo en problemas de clasificación binaria. Vemos en una tabla qué tal se comporta el modelo para cada clase (filas son las clases actuales y columnas las predichas

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Matriz de confusión

Muy útil sobre todo en problemas de clasificación binaria. Vemos en una tabla qué tal se comporta el modelo para cada clase (filas son las clases actuales y columnas las predichas

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Hay que tener claro qué es 1 y que es 0. 1 es la pregunta que queremos resolver en el target. ¿Quién me impaga? ¿Quién sobrevive en el Titanic? ¿Quién da positivo en CV?

0 es si no se da el caso

Por tanto, positivo es 1, y negativo es 0

Matriz de confusión

Muy útil sobre todo en problemas de clasificación binaria. Vemos en una tabla qué tal se comporta el modelo para cada clase (filas son las clases actuales y columnas las predichas

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Hay que tener claro qué es 1 y que es 0. 1 es la pregunta que queremos resolver en el target. ¿Quién me impaga? ¿Quién sobrevive en el Titanic? ¿Quién da positivo en CV?

0 es si no se da el caso

Por tanto, positivo es 1, y negativo es 0

Modelo Felicidad:

EJEMPLO:

Matriz de confusión

Muy útil sobre todo en problemas de clasificación binaria. Vemos en una tabla qué tal se comporta el modelo para cada clase (filas son las clases actuales y columnas las predichas

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Hay que tener claro qué es 1 y que es 0. 1 es la pregunta que queremos resolver en el target. ¿Quién me impaga? ¿Quién sobrevive en el Titanic? ¿Quién da positivo en CV?

0 es si no se da el caso

Por tanto, positivo es 1, y negativo es 0

EJEMPLO:

Modelo Felicidad:

Predicción:

60 felices
40 no felices

Matriz de confusión

Muy útil sobre todo en problemas de clasificación binaria. Vemos en una tabla qué tal se comporta el modelo para cada clase (filas son las clases actuales y columnas las predichas

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Hay que tener claro qué es 1 y que es 0. 1 es la pregunta que queremos resolver en el target. ¿Quién me impaga? ¿Quién sobrevive en el Titanic? ¿Quién da positivo en CV?

0 es si no se da el caso

Por tanto, positivo es 1, y negativo es 0

EJEMPLO:

Modelo Felicidad:

Predicción:	Realidad:
60 felices (1)	45 sí, 15 no
40 no felices (0)	35 no, 5 sí

Matriz de confusión

Muy útil sobre todo en problemas de clasificación binaria. Vemos en una tabla qué tal se comporta el modelo para cada clase (filas son las clases actuales y columnas las predichas

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Hay que tener claro qué es 1 y que es 0. 1 es la pregunta que queremos resolver en el target. ¿Quién me impaga? ¿Quién sobrevive en el Titanic? ¿Quién da positivo en CV?

0 es si no se da el caso

Por tanto, positivo es 1, y negativo es 0

EJEMPLO:

Modelo Felicidad:

Predicción:	Realidad:
60 felices (1)	45 sí, 15 no
40 no felices (0)	35 no, 5 sí

	Predicted 0	Predicted 1
Actual 0	35	15
Actual 1	5	45

Matriz de confusion: Terminología

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

	Predicted 0	Predicted 1
Actual 0	35	15
Actual 1	5	45

Matriz de confusion: Terminología

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

	Predicted 0	Predicted 1
Actual 0	35	15
Actual 1	5	45

True Positive (TP): 45

Matriz de confusion: Terminología

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

	Predicted 0	Predicted 1
Actual 0	35	15
Actual 1	5	45

True Positive (TP): 45

True Negative (TN): 35

Matriz de confusion: Terminología

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

	Predicted 0	Predicted 1
Actual 0	35	15
Actual 1	5	45

True Positive (TP): 45
True Negative (TN): 35
False Positive (FP): 15

Matriz de confusion: Terminología

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

	Predicted 0	Predicted 1
Actual 0	35	15
Actual 1	5	45

True Positive (TP): 45
True Negative (TN): 35
False Positive (FP): 15
False Negative (FN): 5

Accuracy

Los que ha clasificado bien vs todas las muestras a clasificar

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$

$$\text{Accuracy} = \frac{35 + 45}{35 + 15 + 5 + 45} = \frac{80}{100} = 0.8\ (80\%)$$

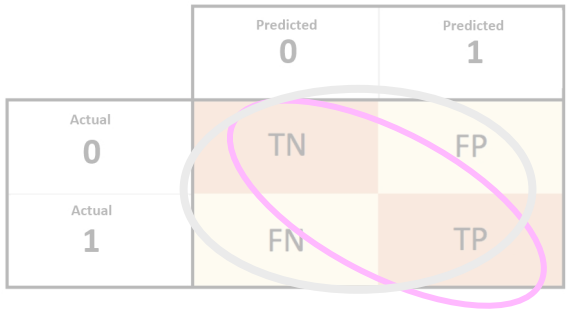
	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

	Predicted 0	Predicted 1
Actual 0	35	15
Actual 1	5	45

Accuracy

Los que ha clasificado bien vs todas las muestras a clasificar

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$



	Predicted 0	Predicted 1
Actual 0	35	15
Actual 1	5	45

Precision

De los que ha predicho como 1, cuántos en realidad ha acertado

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

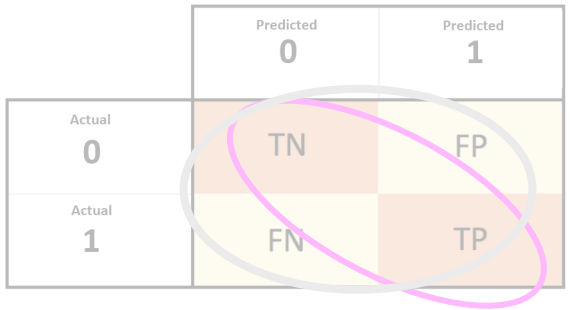
	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

$$Precision = \frac{45}{15+45} = \frac{45}{60} = 0.75\ (75\%)$$

Accuracy

Los que ha clasificado bien vs todas las muestras a clasificar

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total}}$$

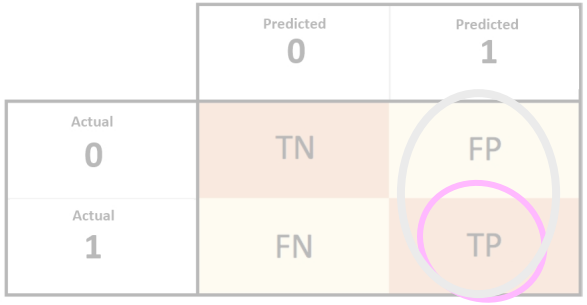


	Predicted 0	Predicted 1
Actual 0	35	15
Actual 1	5	45

Precision

De los que ha predicho como 1, cuántos en realidad ha acertado

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$



$$Precision = \frac{45}{15+45} = \frac{45}{60} = 0.75\ (75\%)$$

Recall o Sensibilidad

Los positivos que he clasificado bien vs todos los positivos que había

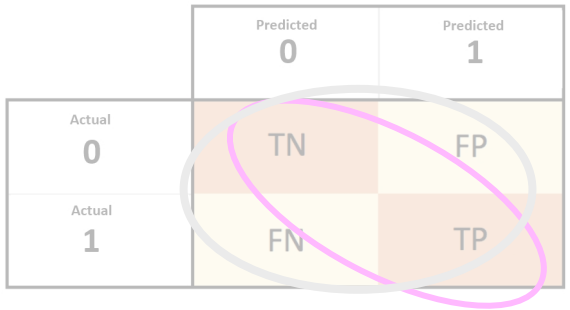
$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Accuracy

Los que ha clasificado bien vs todas las muestras a clasificar

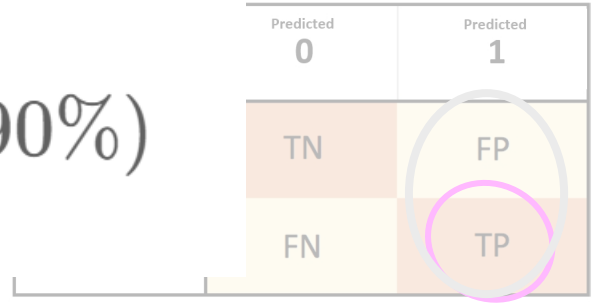
$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total}}$$



	Predicted 0	Predicted 1
Actual 0	35	15
Actual 1	5	45

Precision

$$Recall = \frac{45}{5+45} = \frac{45}{50} = 0.9\ (90\%)$$



$$Precision = \frac{45}{15+45} = \frac{45}{60} = 0.75\ (75\%)$$

Recall o Sensibilidad

Los positivos que he clasificado bien vs todos los positivos que había

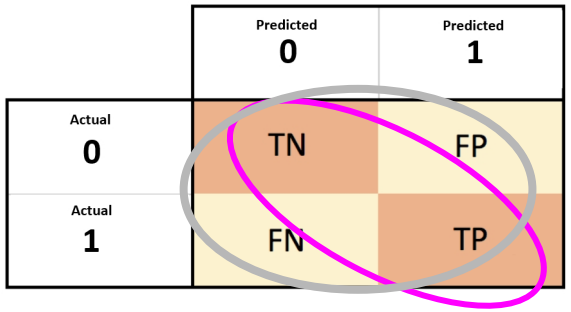
$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Accuracy

Los que ha clasificado bien vs todas las muestras a clasificar

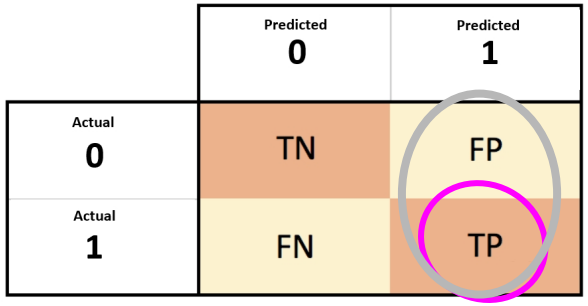
$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total}}$$



Precision

De los que ha predicho como 1, cuántos en realidad ha acertado

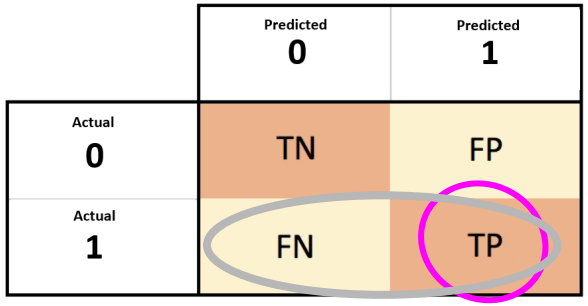
$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$



Recall o Sensibilidad

Los positivos que he clasificado bien vs todos los positivos que había

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$



Escoger métrica

Accuracy

Elegir cuando el problema esté balanceado. NO usar nunca cuando la mayor parte de los datos caiga del lado de una sola clase.

Si intentamos predecir cáncer entre 100 personas, y 5 tienen cáncer. Siendo el modelo muy malo, predecirá todos los casos como no cáncer, y tendrá un accuracy del 95%, cuando está prediciendo muy mal en realidad.

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Escoger métrica

Accuracy

Elegir cuando el problema esté balanceado. NO usar nunca cuando la mayor parte de los datos caiga del lado de una sola clase.

Si intentamos predecir cáncer entre 100 personas, y 5 tienen cáncer. Siendo el modelo muy malo, predecirá todos los casos como no cáncer, y tendrá un accuracy del 95%, cuando está prediciendo muy mal en realidad.

Precision

No me importa que se me escape algún 1, mientras no se me cuele ningún 0 (FP) como si fuese 1. Que cuando prediga como 1, de verdad sea 1. El foco hay que ponerlo en minimizar los FP

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Escoger métrica

Accuracy

Elegir cuando el problema esté balanceado. NO usar nunca cuando la mayor parte de los datos caiga del lado de una sola clase.

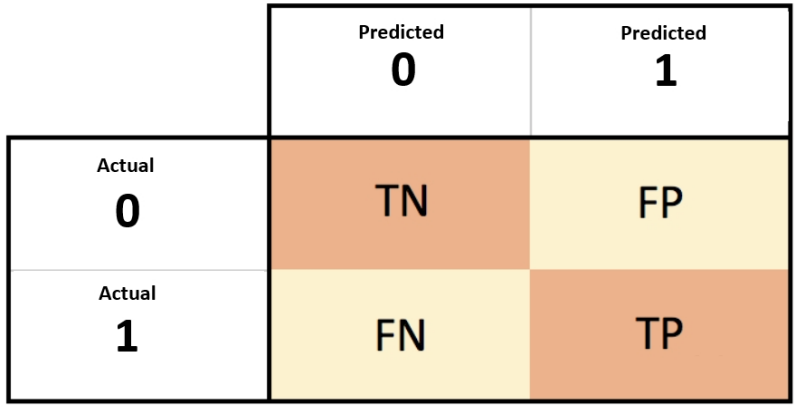
Si intentamos predecir cáncer entre 100 personas, y 5 tienen cáncer. Siendo el modelo muy malo, predecirá todos los casos como no cáncer, y tendrá un accuracy del 95%, cuando está prediciendo muy mal en realidad.

Precision

No me importa que se me escape algún 1, mientras no se me cuele ningún 0 (FP) como si fuese 1. Que cuando prediga como 1, de verdad sea 1. El foco hay que ponerlo en minimizar los FP

Recall

Lo que me importa es que los 1s me los capture bien. No me importa que se me cuele algún 0, pero los 1s no se me pueden escapar. como 0s (FN). Por tanto el objetivo es minimizar los FN



Algunos ejemplos

Algunos ejemplos

Clasificador de videos buenos para niños

No quieres que se te cuele ningun video malo (0) como video bueno (1) -> FP muy bajos ->precisión alta

Por otro lado, no te va a importar perder algún video bueno (1) y clasificarlo como malo -> FN alto -> mal recall

¿Prioridad? Precision

Algunos ejemplos

Clasificador de videos buenos para niños

No quieres que se te cuele ningun video malo (0) como video bueno (1) -> FP muy bajos ->precisión alta

Por otro lado, no te va a importar perder algún video bueno (1) y clasificarlo como malo -> FN alto -> mal recall

¿Prioridad? Precision

Clasificador de ladrones en tienda mediante imágenes

No se me puede escapar ni un ladrón (1), y que se clasifique como no ladrón (0) -> FN bajo -> recall alto

Por otro lado, no me importa clasificar algún cliente como ladrón y realizar registros de vez en cuando -> FP altos -> precisión baja

¿Prioridad? Recall

Algunos ejemplos

Clasificador de videos buenos para niños

No quieres que se te cuele ningun video malo (0) como video bueno (1) -> FP muy bajos ->precisión alta

Por otro lado, no te va a importar perder algún video bueno (1) y clasificarlo como malo -> FN alto -> mal recall

¿Prioridad? Precision

Clasificador de ladrones en tienda mediante imágenes

No se me puede escapar ni un ladrón (1), y que se clasifique como no ladrón (0) -> FN bajo -> recall alto

Por otro lado, no me importa clasificar algún cliente como ladrón y realizar registros de vez en cuando -> FP altos -> precisión baja

¿Prioridad? Recall


THE
BRIDGE