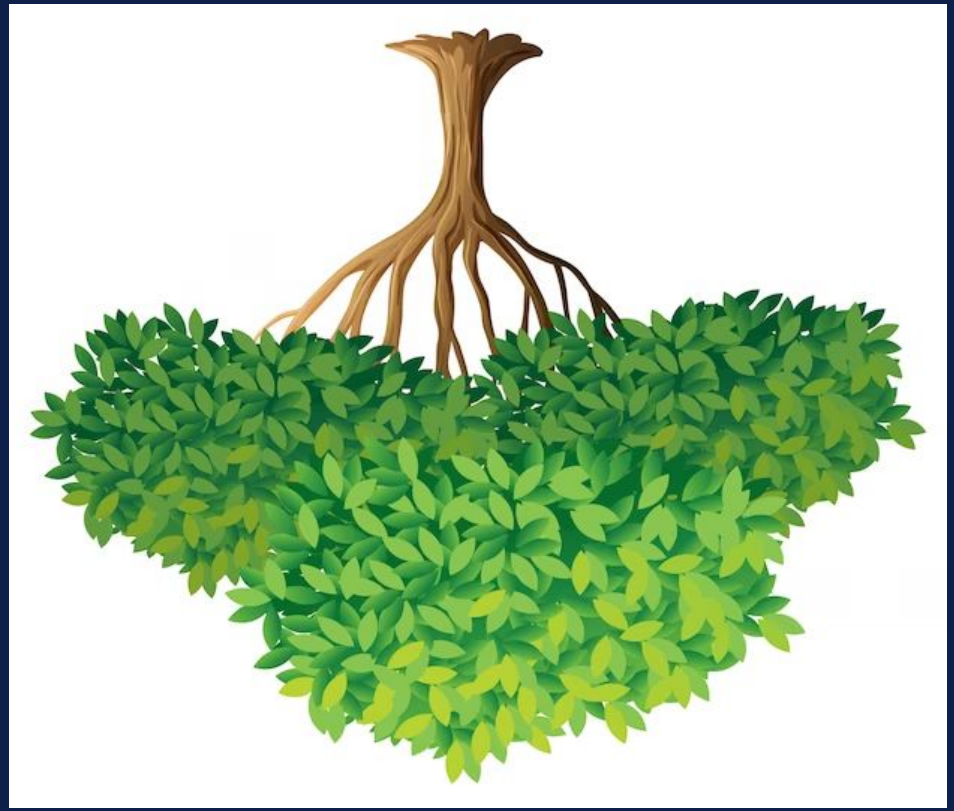
ASTRONOMÍA Y CIENCIA DE DATOS: DE LAS ESTRELLAS A LOS NÚMEROS

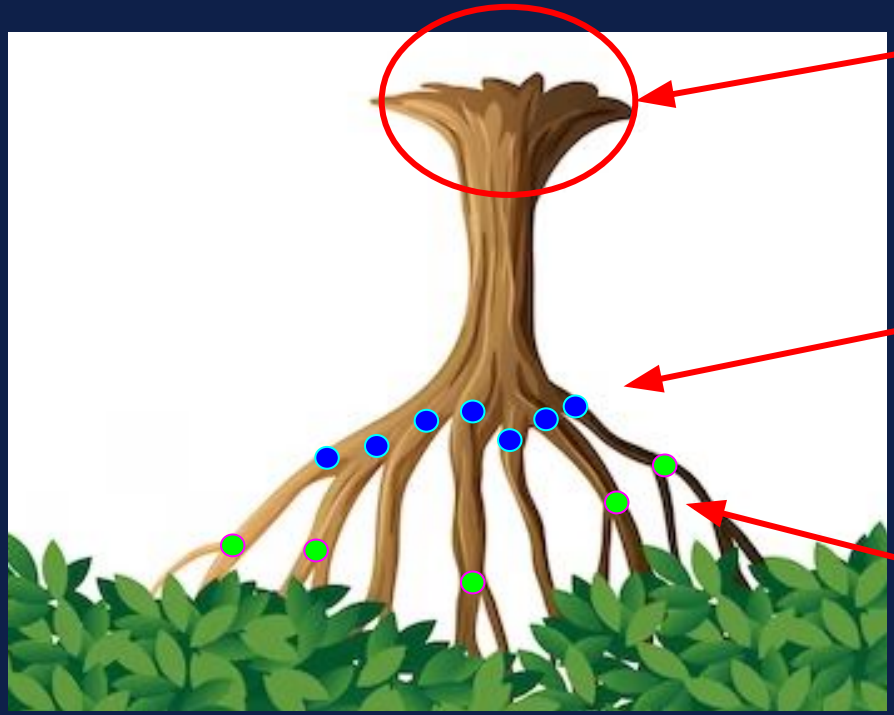
Clase 8: Introducción al Aprendizaje Supervisado II: Random Forest

Árboles de decisión

Árboles de decisión



Árboles de decisión



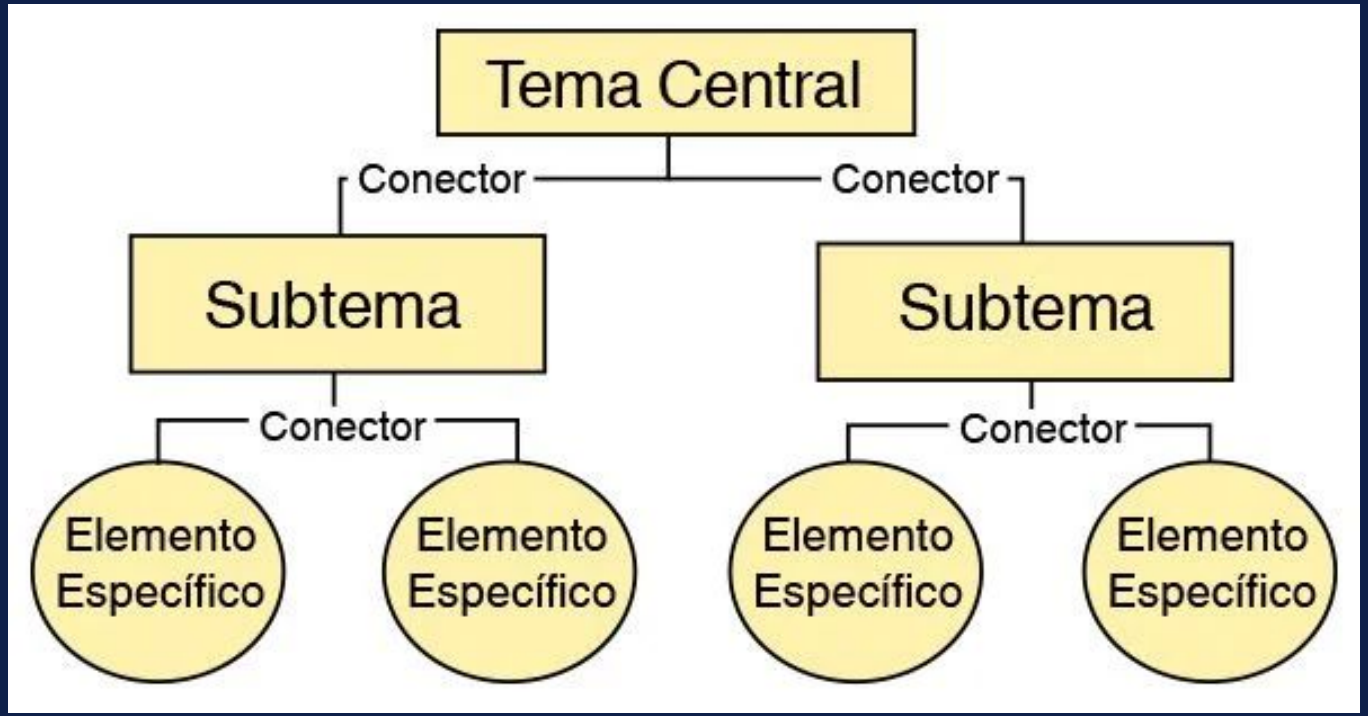
Base del tronco

Bifurcación en ramas principales

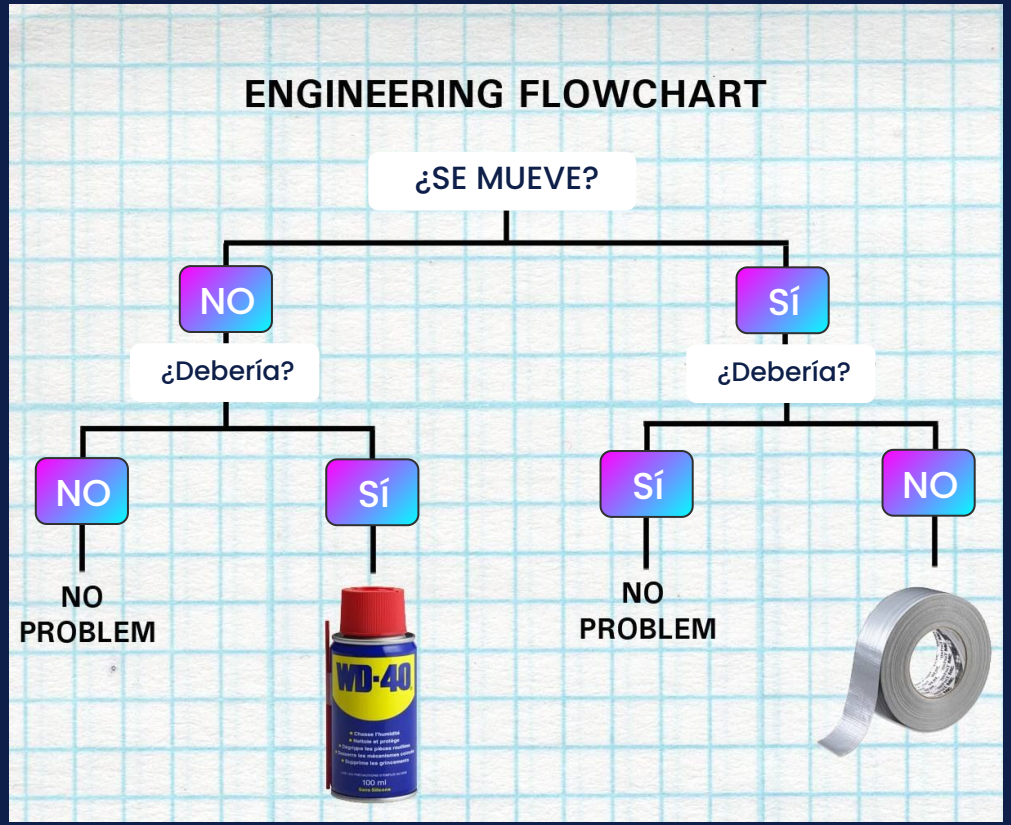
Bifurcación en ramas secundarias

Árboles de decisión

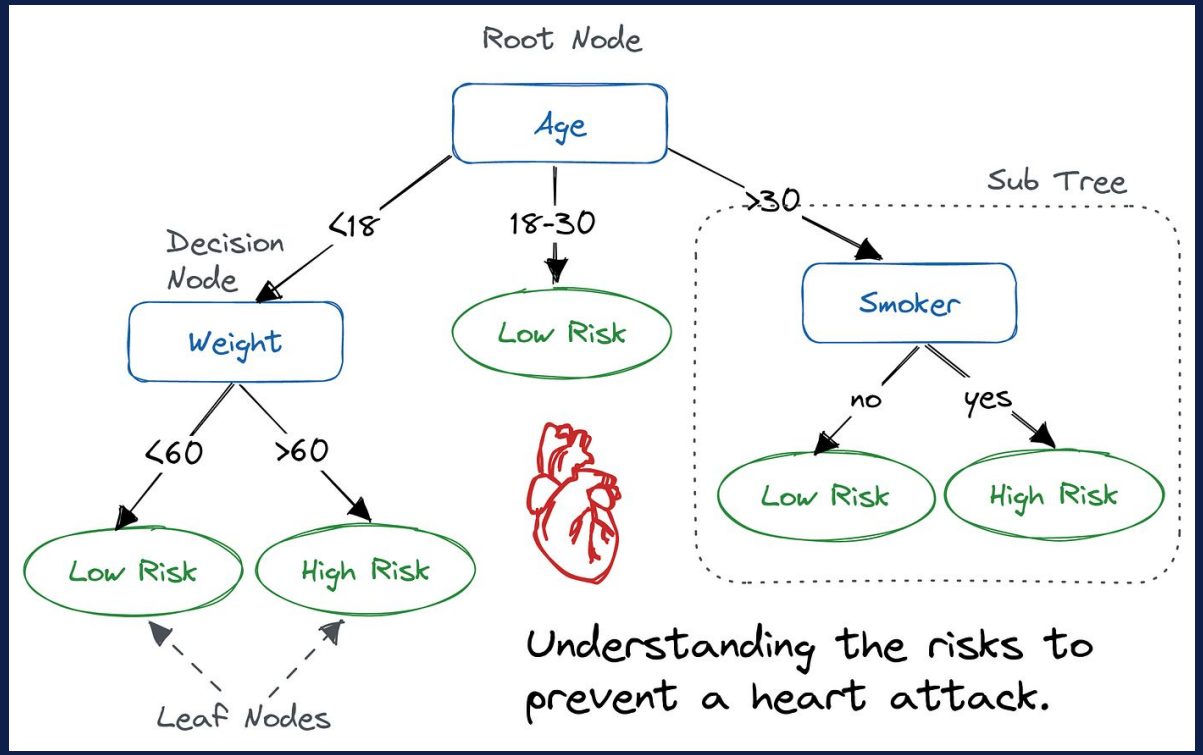
Son de hecho bastante similares a un mapa conceptual



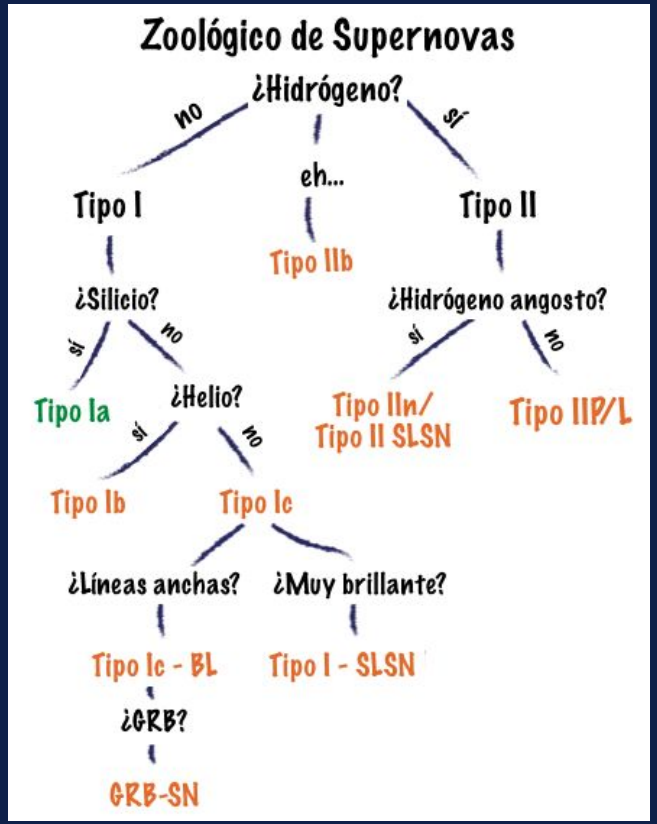
Árboles de decisión



Árboles de decisión



Árboles de decisión



Características de los DT

- ➔ Fáciles de interpretar
- ➔ Rápidos
- ➔ Deterministas → (**SÍ o NO**)
- ➔ Sensible a outliers (casos raros)
- ➔ Comúnmente **se sobreajusta a los datos**.

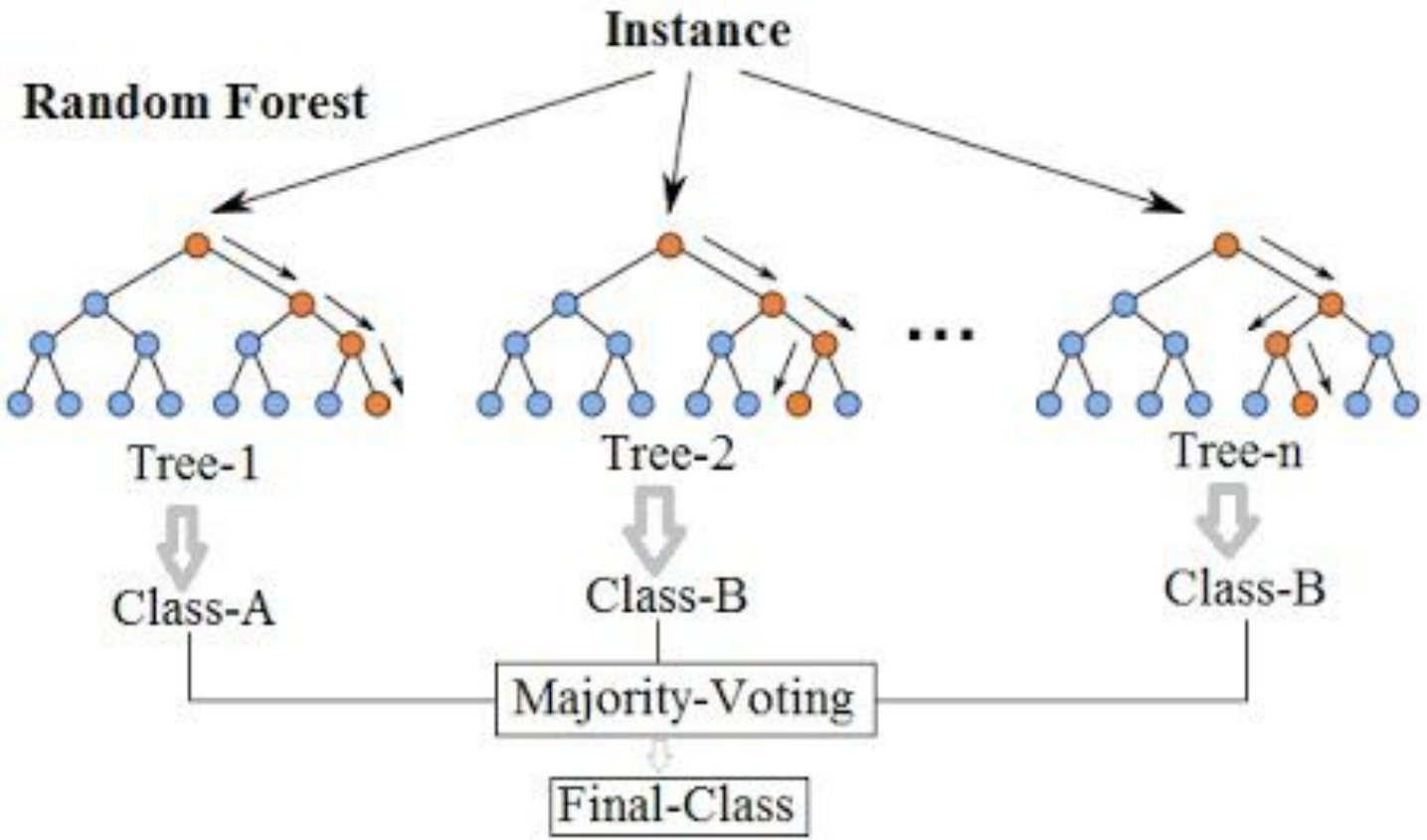
Características de los DT

- ➔ Fáciles de interpretar
- ➔ Rápidos
- ➔ Deterministas → (**SÍ o NO**)
- ➔ Sensible a outliers (casos raros)
- ➔ Comúnmente **se sobreajusta a los datos**.

Cómo podemos compensar los e el u n perder sus ventajas?

CON MÁS ÁRBOLES!

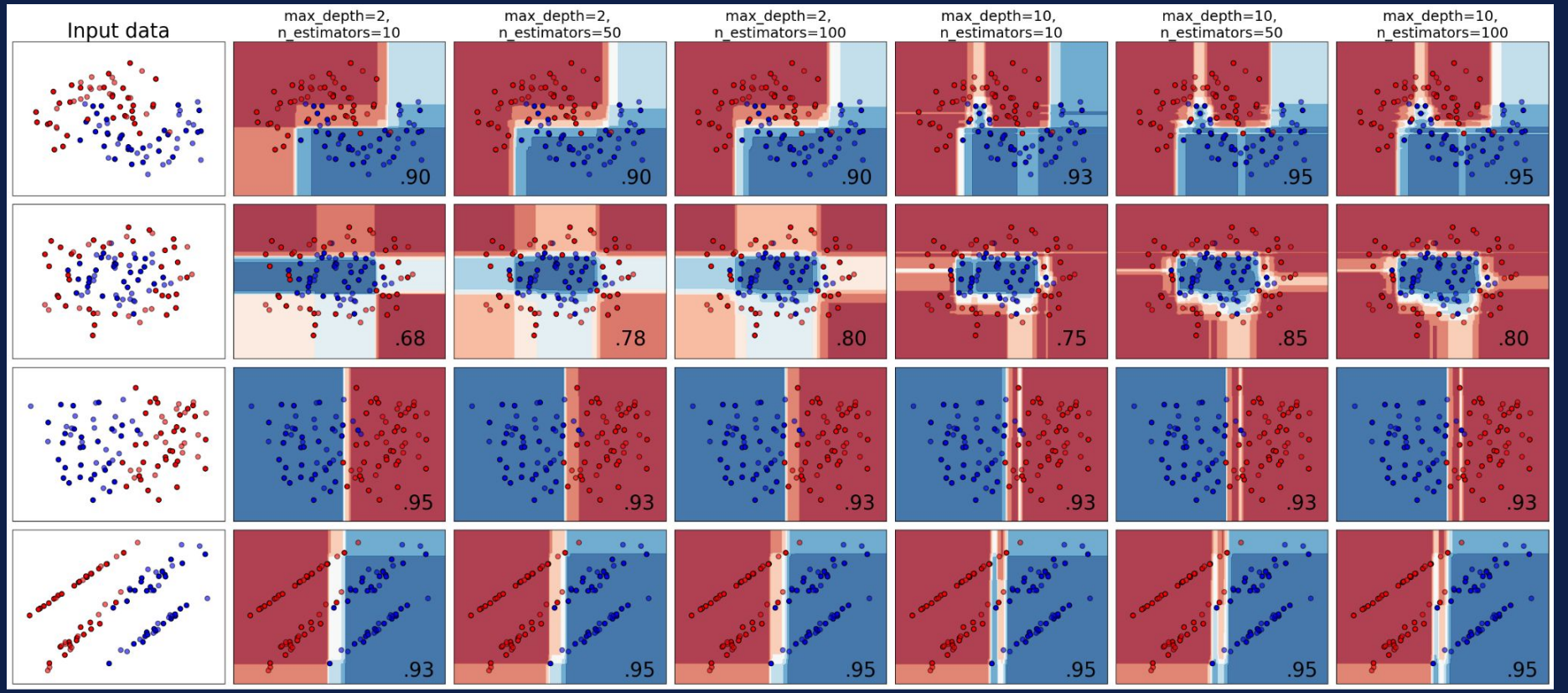
Random Forest Simplified



Random Forest

- ➔ número de árboles (*n_estimators*)
- ➔ Características a considerar (*max_features*)
- ➔ Número maximo de sub-ramas (*max_depth*)

Random Forest



Características de los DT

- ➔ Fáciles de interpretar
- ➔ Rápidos (en entrenamiento y aplicaciones)
- ➔ Deterministas → (**SÍ o NO**)
- ➔ Sensible a outliers (casos raros)
- ➔ Comúnmente **se sobreajusta a los datos**.

Características de RF

- ➔ Se complica la interpretación
- ➔ Rápidos (en entrenamiento y aplicaciones)
- ➔ No Deterministas → (**entrega un porcentaje de confianza,** ej 83%)
- ➔ Más **robusto** con outliers
- ➔ El modelo se vuelve **más general a los datos**.

Árboles v/s Random Forest

	Árboles	Random Forest
Interpretabilidad	✓ ✓	✓ ✗
Precisión	✓ ✗	✓ ✓
Overfitting	✗	✓ ✓
Robustez	✗ ✗	✓ ✓
Velocidad	✓ ✓	✓
Clasificación		
Regresión		

y aún quedan un montón de modelos que no hemos mencionado!

Random Forest

n montón de modelos que no hemos mencionado!

Random Forest

n montón de modelos que no hemos mencionado!

Redes Neuronales

Random Forest

n montón de modelos que no hemos nado!

Super Vector Machine

Redes Neuronales

Random Forest

modelos d

Proceso Gaussiano

Super Vector Machine

nado!

Redes Neuronales

Random Forest

Proceso Gaussiano

modelos c

nado!

Super Vector M

des

nales

QDA y LA

Random Forest

Proceso

modelos

Super Vec

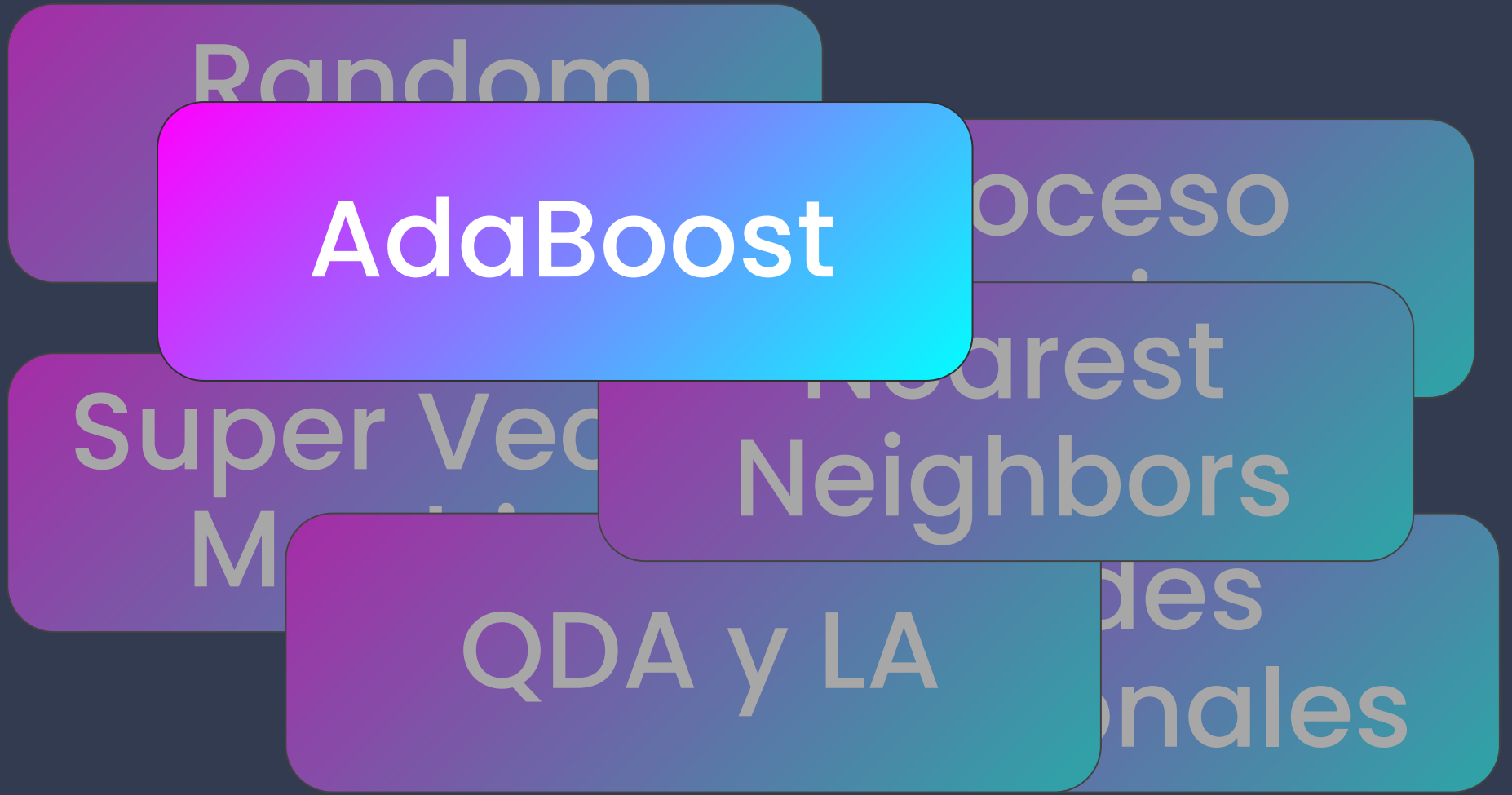
M

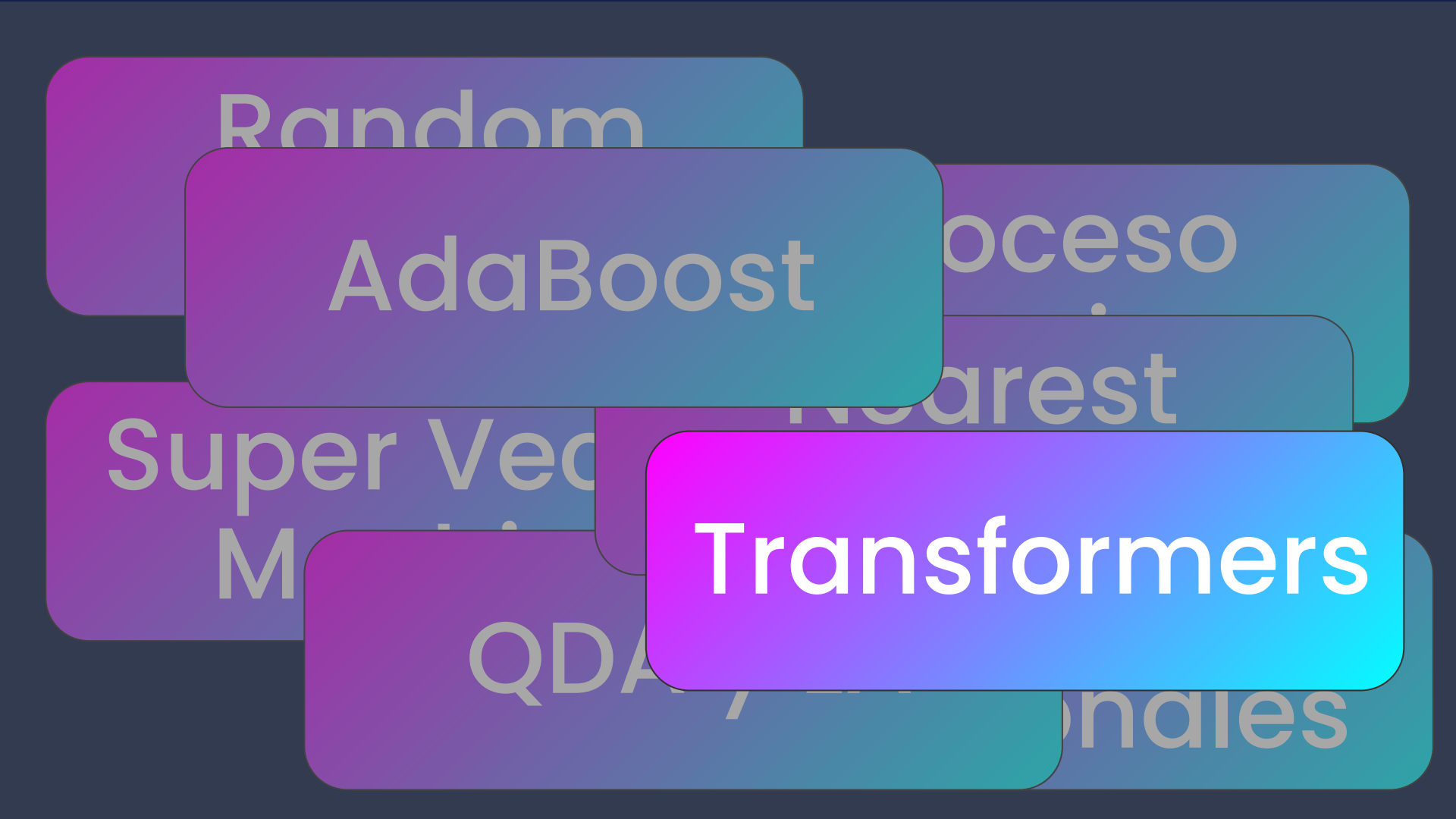
Nearest Neighbors

QDA y LA

des

nales


Random
AdaBoost
oceso
Super Vec
M
Nearest
Neighbors
QDA y LA
des
nales

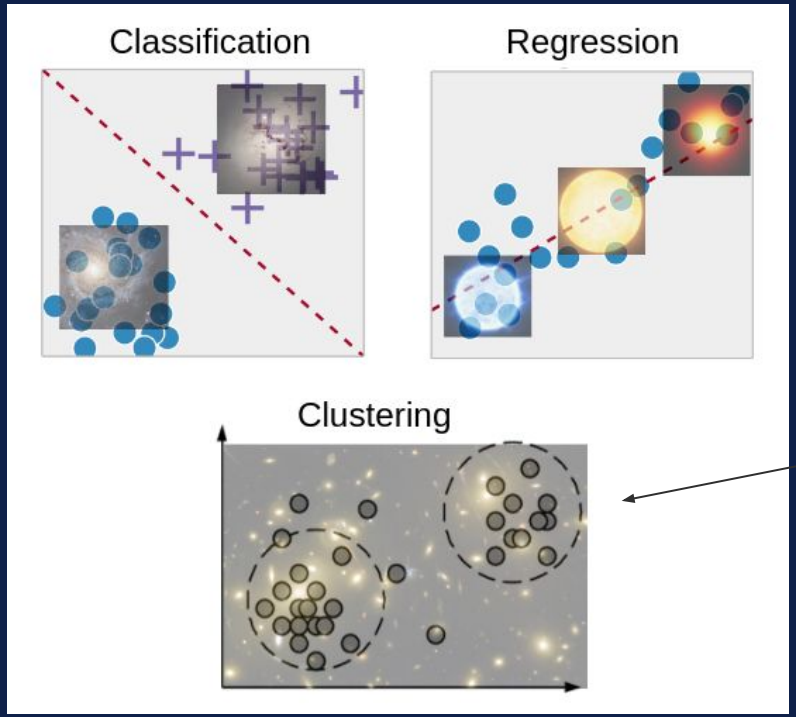

Random
AdaBoost
oceso
Nearest
Super Vec
M
QDA
nales
Transformers

Teniendo tantas opciones, ¿cómo **evaluamos** qué modelo es mejor?

Métricas de evaluación

Dependen del tipo de problema que queremos resolver

Supervisado



No supervisado

Métricas de regresión

Buscamos una buena representación de comportamiento

MAE: $\frac{\sum_{i=1}^{n} |y_i - f(x_i)|}{n}$

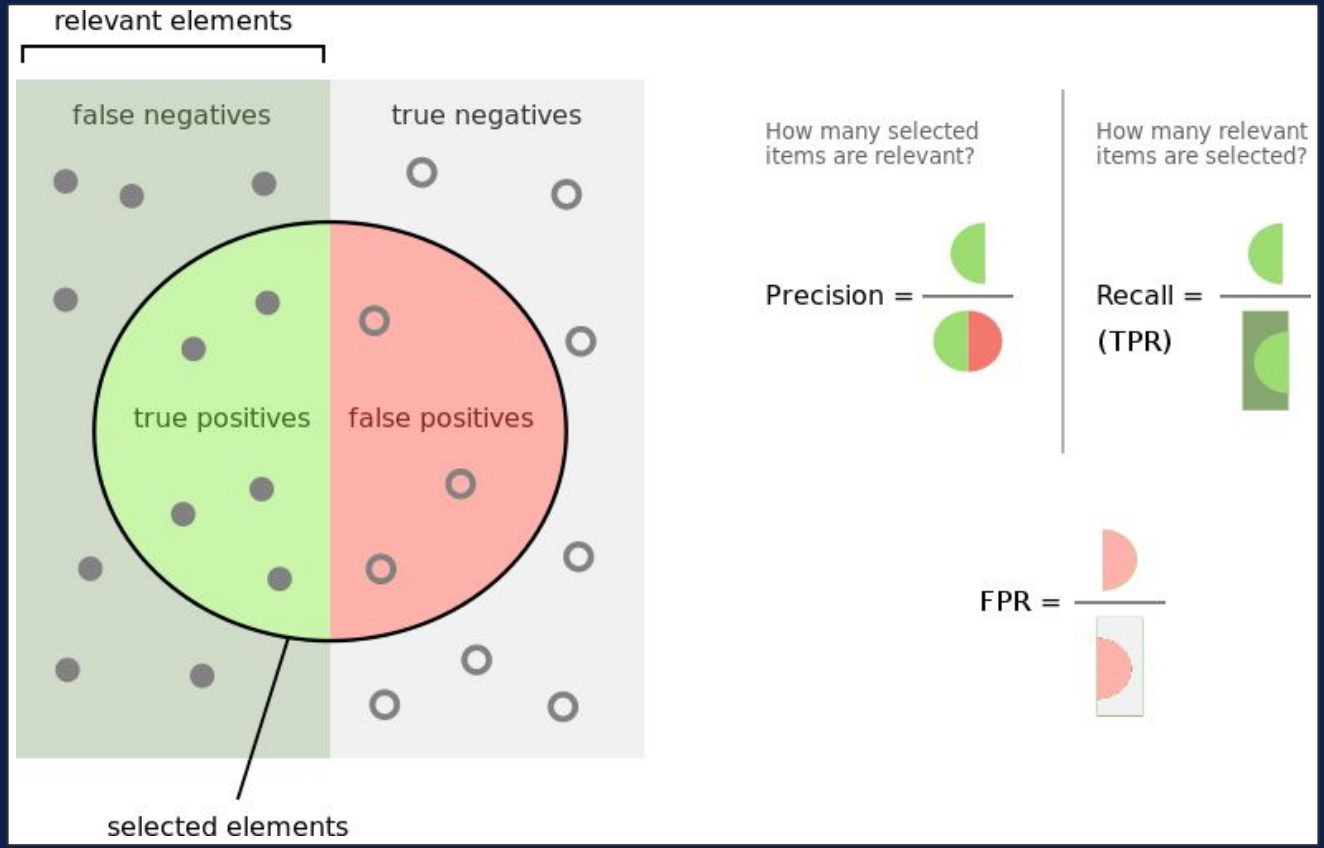
RMSLE: $\sqrt{\frac{\sum_{i=1}^{n} (\log(1+y_i) - \log(1+f(x_i)))^2}{n}}$

MSE: $\frac{\sum_{i=1}^{n} (y_i - f(x_i))^2}{n}$

RMSE: $\sqrt{\frac{\sum_{i=1}^{n} (y_i - f(x_i))^2}{n}}$

¿Cuando se usa cada una?

Métricas de Clasificación



Métricas de Clasificación

Tipos de errores (FP, FN)

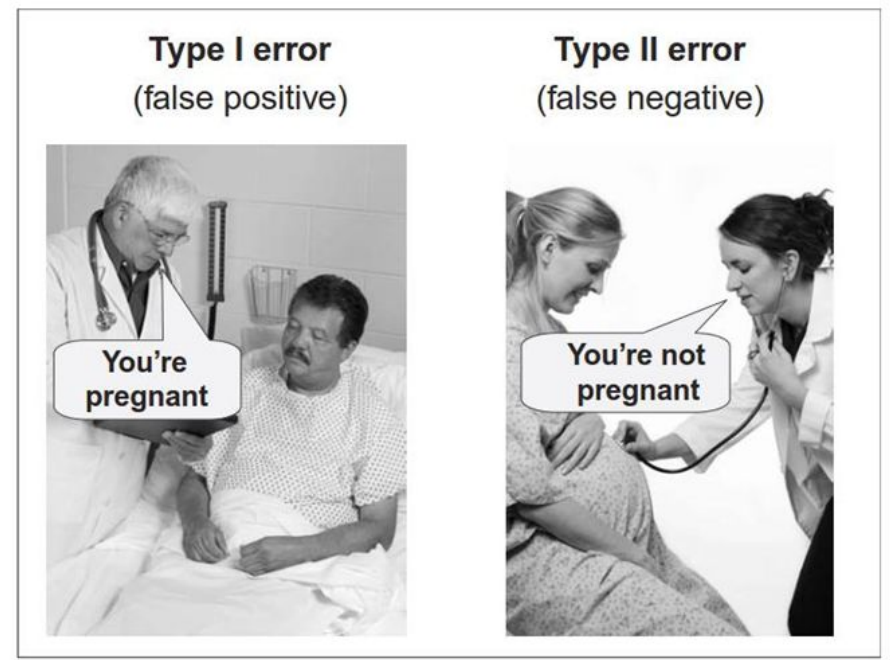
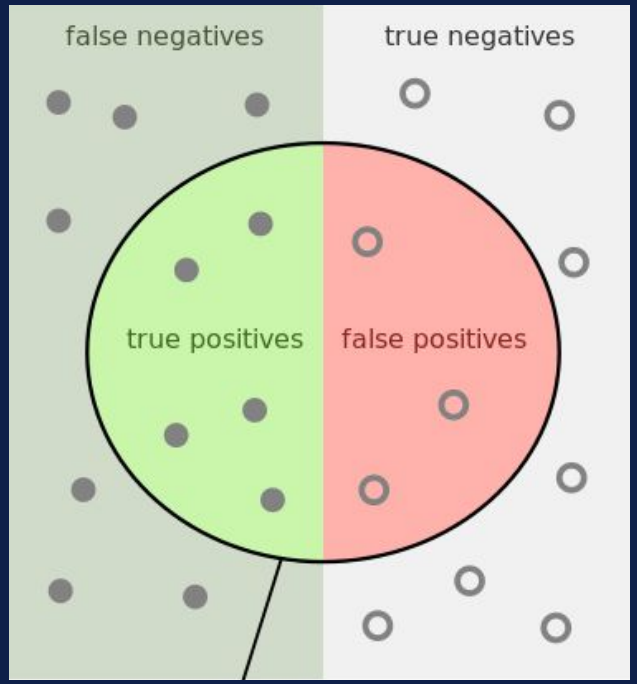


Figure 3.1 Type I and Type II errors

De The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results

Métricas de evaluación



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

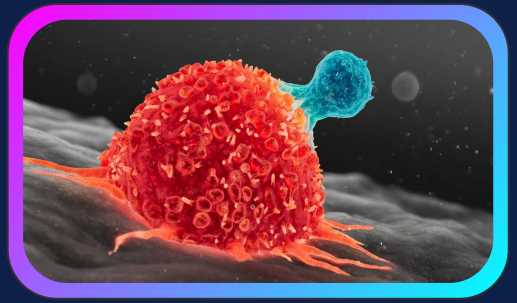
$$Precision = \frac{TP}{TP + FP}$$

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

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Accuracy Paradox

Modelo de clasificación de cáncer



Accuracy: 0.96

	Cáncer	No cáncer
Clasific. cáncer	5	0
Clasific. no cáncer	45	950

Usemos otras métricas...

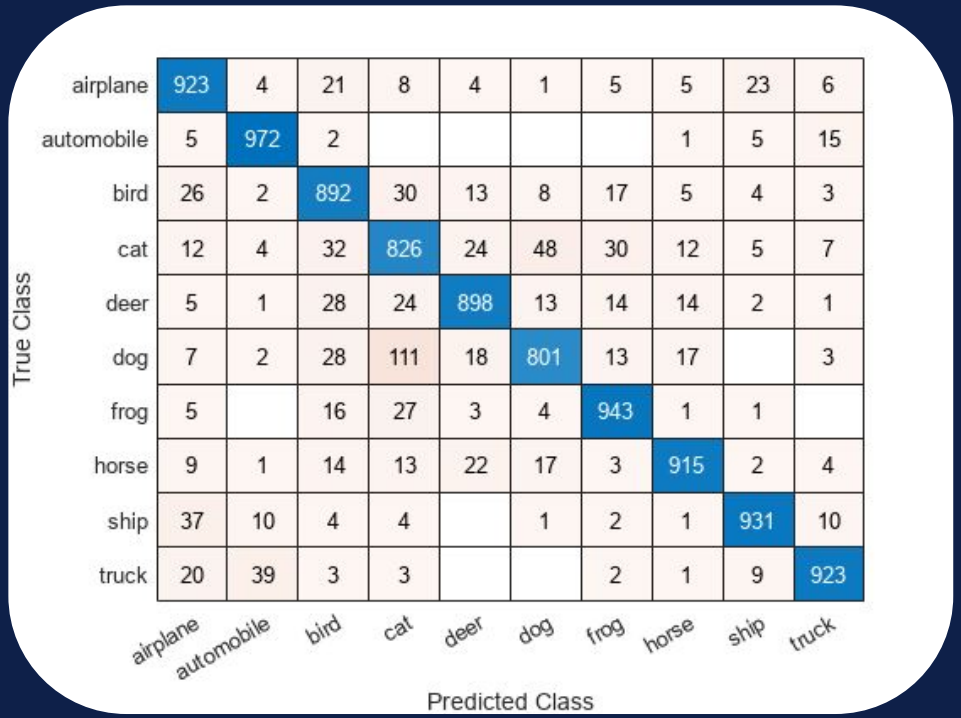
	Cáncer	No cáncer
Clasific. cáncer	5	0
Clasific. no cáncer	45	950

TP = 5
TN = 950
FP = 0
FN = 45

Accuracy = 0.96
Precision = 1
Recall = 0.1
F1-score = 0.182

¿Cómo interpretamos este resultado?

Matriz de confusión

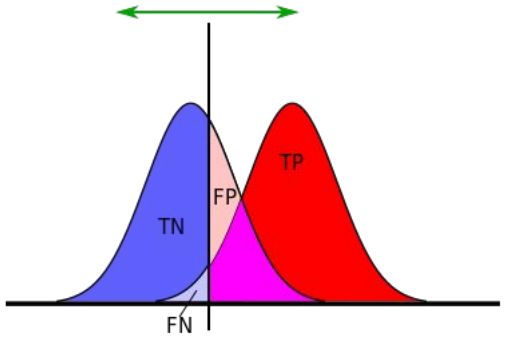


Nos permite visualizar un clasificador multiclase

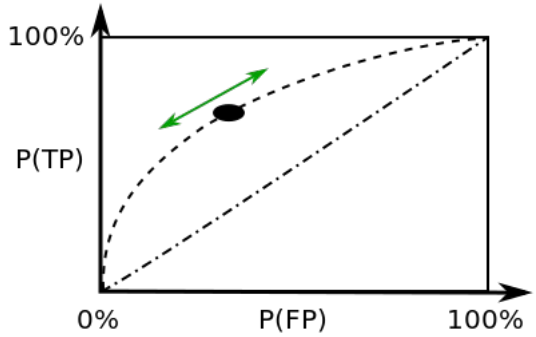
¿Qué esperamos de un buen clasificador?

Métricas de evaluación

ROC Curve

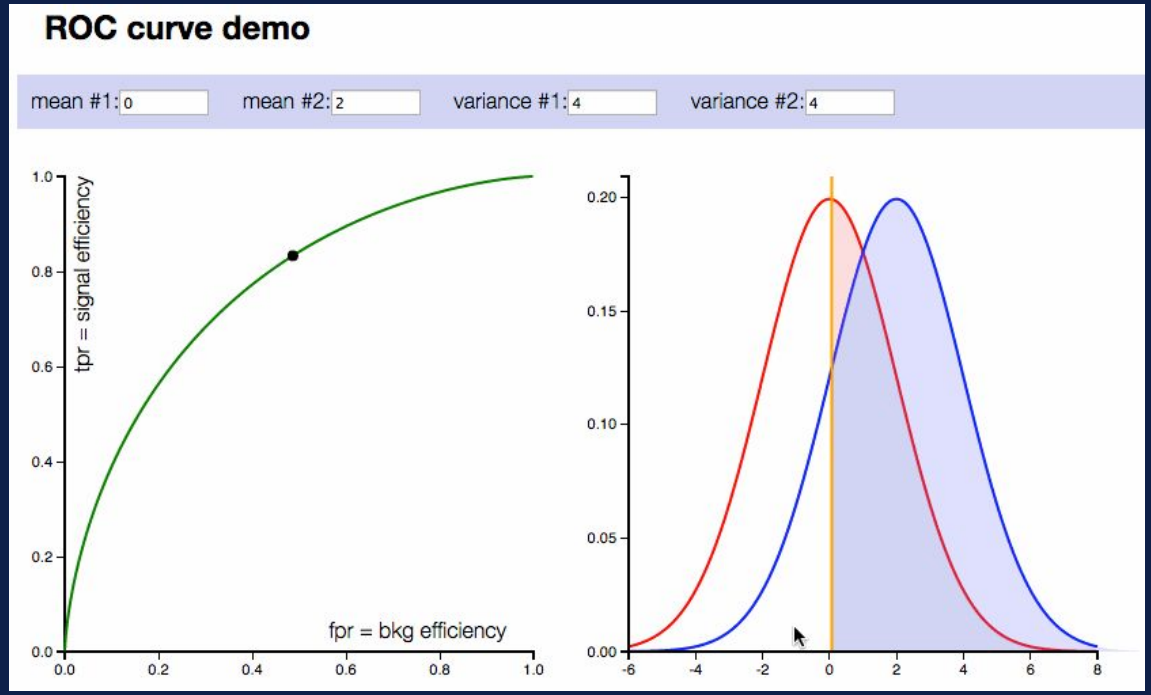


TP	FP
FN	TN



Métricas de evaluación

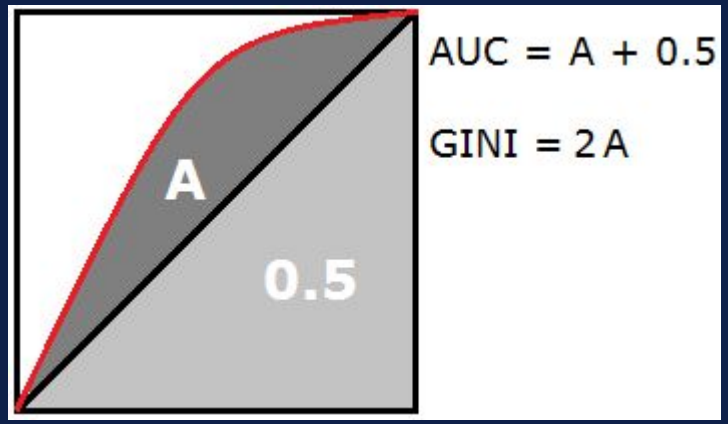
ROC Curve



https://arogozhnikov.github.io/2015/10/05/roc-curve.html

Métricas de evaluación

AUC (Área bajo la curva) y GINI



Comúnmente usada para comparar modelos!

Pero e aparar
y
¡¿l o?!

Ay que lataaa!