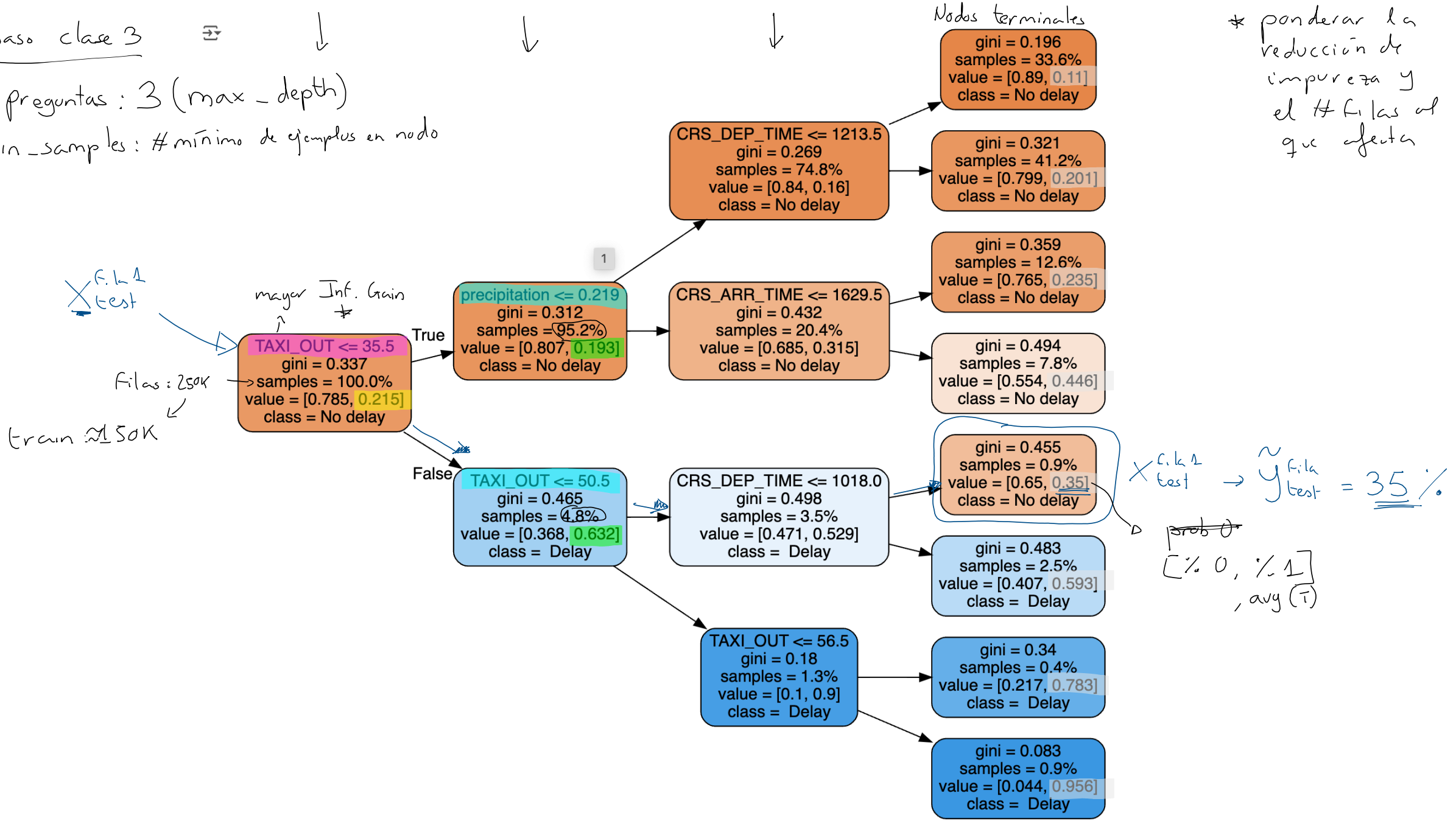
Repaso clase 3

preguntas : 3 (max_depth)

min_samples : # mínimo de ejemplos en nodo

Nodos terminales

* ponderar la reducción de impureza y el # filas al que afecta



Métricas evaluación

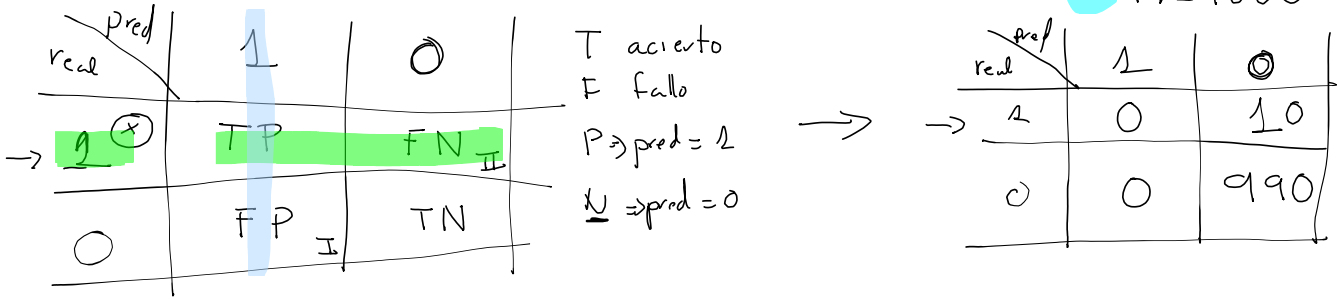
① **Accuracy**: % de aciertos → Pregunta: ¿si acc = 95%, tengo un buen modelo?
[0% - 100%]
↳ es acc train?

Ejemplo: operación fraudulenta
↳ operaciones diarias >>> fraudulentas
↳ avg (T) ≈ 0% ⇒ acc = 99%

$acc_{tr} = 95\% \leftrightarrow acc_{test} = 94\%$ - no overfitting
↳ es un buen modelo? $acc_{val} = 94\%$
↳ es un buen modelo?

Problema: datasets desbalanceados
↳ avg (T) ≈ 0% o ≈ 100%
↳ "modelo tonto" siempre predice 0 ⇒
↳ acc ≈ 100%

② **Confusion matrix**: número de filas en cada combinación predicción (1, 0) / real (1, 0)



n = 1000

real \ pred	1	0
1	0	10
0	0	990

Problema
1. Difícil optimizar con múltiples métricas: TP, TN, FP, FN
2. Los fallos FP, FN pueden distinta importancia

③ F1 Score : media armónica de Precision y Recall

[0 - 1]

✗ ✓

↳ beneficia que ambas métricas se parezcan (mejor $P=0,5$, $R=0,5$ que $P=1$, $R=0$)

Ventajas : . un solo número
. funciona con datasets desbalanceados

$$F1 = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

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

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

Recall (R) / Alcance : del total de (+) reales $= 1$, cuantos identifica mi modelo (pred = 1)

Precision : del total de pred = 1, cuantos realmente lo son

ejemplo : test antígenos

predicción

N
- TN : mitad
- FN : mitad

mucho FN → recall baja

P
- TP : muchos
- FP : pocos

poco FP → Precision alta

Recall y precision se mueven en direcciones opuestas

Problema : hablamos de pred = 1/0, la realidad el modelo da probabilidad

$pred = f(X, th)$

- prob ≥ 0,5 → pred = 1
- prob < 0,5 → pred = 0

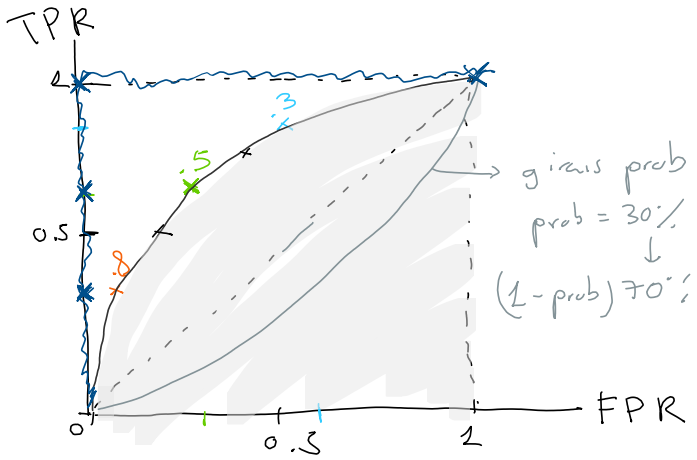
↳ umbral / threshold

⇒

th alto : solo digo pred = 1 si probabilidad muy alta
P ↑ R ↓

th bajos : P ↓ R ↑

(4) ROC, AUC: solucionar el problema del threshold



$$TPR\ (Recall) = \frac{TP}{TP+FN}\ \text{fijo}$$

[0 - 1] ✗ ✓

$$FPR = \frac{FP}{TN+FP}\ \text{fijo}$$

[0 - 1] ✓ ✗

calcular TPR y FPR para diferentes th

- th = 0.5 → TPR = 0.6, FPR = 0.3
- th = 0.3 → TPR = 0.8, FPR = 0.6
- th = 0.8 → TPR = 0.3, FPR = 0.1

AUC [0 - 1] → AUC ≥ 0.75 buen modelo

↳ 0.5 - 1 → no depende del threshold

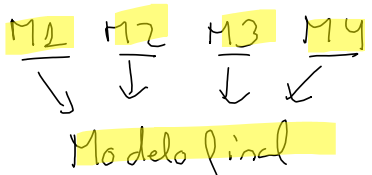
→ el mejor modelo AUC = 1 ("imposible"),

id	real	prob	th = 0,85	th = 0,6	th = 0,45	th = 0,1
1	1	0.9	1	1	1	1
2	1	0.8	0	1	1	1
3	1	0.5	0	0	1	1
4	0	0.4	0	0	0	1
5	0	0.3	0	0	0	1
		TPR	1/3 = 0.3	2/3 = 0.5	3/3 = 1	3/3 = 1
		FPR	0/2 = 0	0/2 = 0	0/2 = 0	2/2 = 1

AUC = 1

Bagging y boosting

Ensembles: combinación de varios modelos intentando predecir la misma tarea con los mismos datos



- Combinación: cada modelo da su probabilidad
  - → Hard voting: avg (pred (1/0))
  - → Soft voting: avg (prob)
  - → stacking: utilizar las probabilidades como atributos de un nuevo modelo

- Modelos deben
  - Precisión: por sí solos sean buenos
  - Variedad: no predecir lo mismo
    - → diferentes algoritmos

→ modificando ligeramente el dataset

→ Bagging + decision tree = Random Forest

→ Boosting + decision tree = XGBoost
↳ ADABoost

Bagging

Dataset original train (20 filas) → réplicas bootstrap

réplica 1 : 2, 5, 1, 2, 3, 7, 8, 4, 9, 8
n = 10
⋮
réplica 50 : 1, 3, 2, 9, 7, 1, 5, 8, 10, 6

mismo algoritmo
decision tree
max_depht = mismo
min_sample

→ M1 ⋮ → M50 → avg (voting) → prob

Random Forest : más preciso que el decision tree, vigilar overfitting
↳ pocas filas < 40K no funciona bien

Boosting : proceso iterativo

DT + Boosting = XGBoost

Train orig → DT → M1 → Y_{train1} vs $\tilde{Y}_{train1}$ = errores 1
train 2 → DT → M2 → Y_{train2} vs $\tilde{Y}_{train2}$ = errores 2
⋮
train 50 → DT → M50 - - - -

average booting → prob

+filas
> 200K
XGBoost
Random forest
50K
Decision tree
- filas

+overfitting
- overfitting

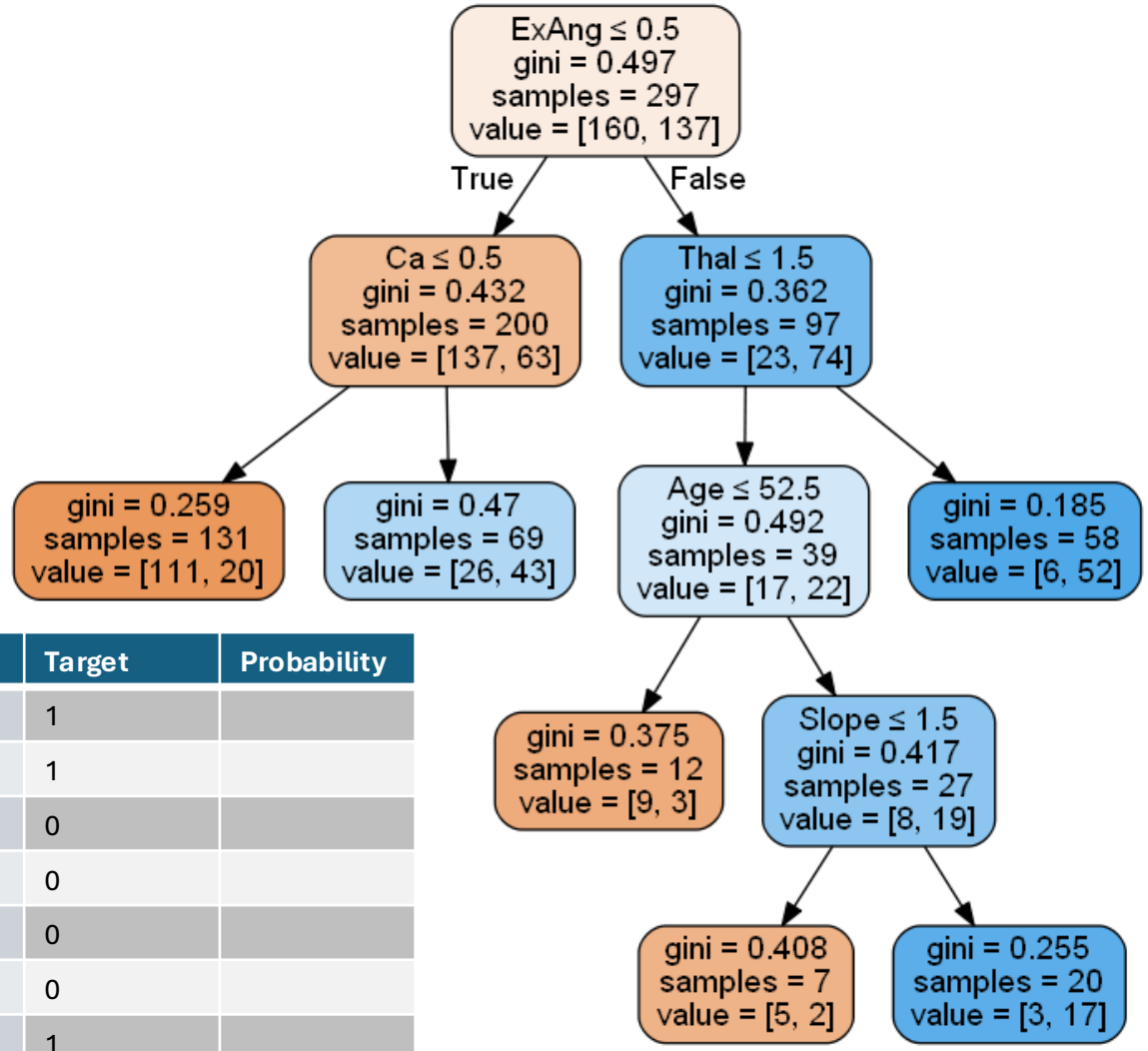
Homework (opcional)

A partir del siguiente árbol, indica:

- Profundidad del árbol
- Min samples split que utilizamos en el entrenamiento
- Con los siguientes datos, dibuja la ROC Curve resultante del modelo.

Pistas:

- Calcula la probabilidad que predice el árbol para cada fila
- Ordena las filas de mayor a menor probabilidad
- Calcula el TPR y FPR para diferentes cortes del threshold (al menos 5 cortes distintos)
- Dibuja estos valores en el gráfico de TPR/FPR



Id fila	ExAng	Ca	Thal	Age	Slope	Target	Probability
1	1	0	8	65	1	1	
2	0	0	5	25	8	1	
3	0	1	2	55	5	0	
4	1	1	1	45	4	0	
5	1	0	0	60	1	0	
6	1	0	0	65	3	0	
7	1	1	3	28	7	1	
8	0	1	0	39	2	1	
9	1	0	0	87	4	1	
10	1	1	1	15	8	0	