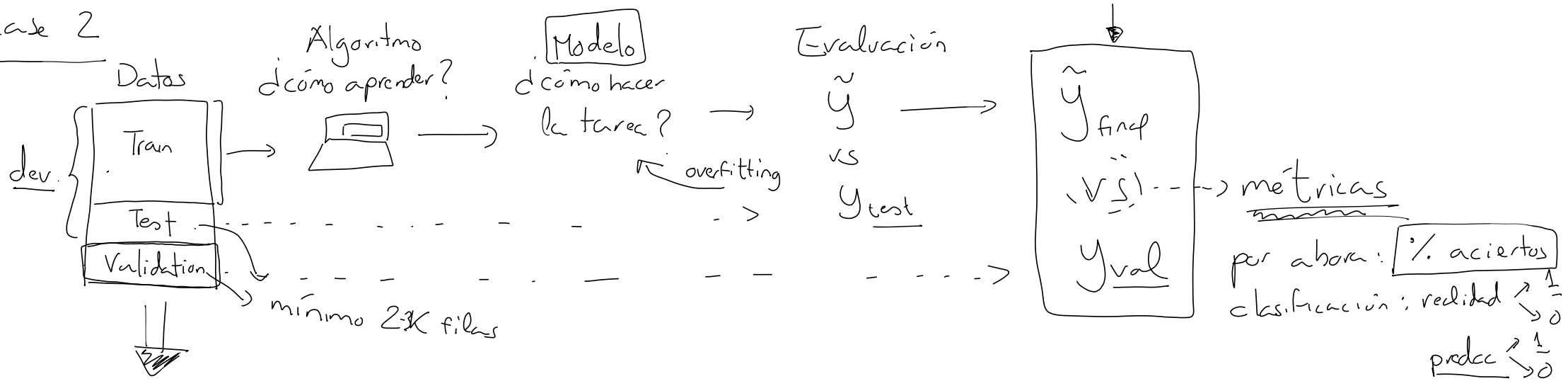
Repaso clase 2



¿Porqué mi modelo puede ir mal?

1. Memorización (overfitting): aprende normas específicas de "Train"
   Comparar el resultado del modelo train vs test (sample aleatoria de "dev")
   ↳ misma distribución

2. Distribución que ve mi máquina es distinta a la realidad
   train vs validation

Data preparation: cambiar todo a número (la "traducción" es clave para dar buena información a la máquina)

1. Nulos

→ según el algoritmos: decision tree (simbolistas) → valor extremo o fuera del rango

Boolean

Bool	avg(target)	#	Pregunta:
True	0,5	50K	I. Nulo = True
→ False	0,3	20K	II. Nulo = False
→ Nulos	0,2	200	III. Nulo = 'Nueva cat'

→

2. Categóricas

3. Fechas: extraer a número todas sus componentes (5/1/2018)

→ dia: 5.
→ mes: 1.
→ año: 2018.

→ semana: 1
→ trimestre: 1
→ año + mes + dia: 20180105
→ boolean: dias festivos

Aprendizaje supervisado

Dataset: Atributos: $X_1, \dots, X_n$ → la máquina busca la mejor combinación de los atributos
target: Y

$$\tilde{y} = f(X_1, \dots, X_n)$$
↑ combinar

¿Cómo combinar los atributos? → lo decide la máquina con el framework algoritmos

Atributo explicativo: Según el valor del atributo X_n, reducimos la incertidumbre con respecto al target.

Decision tree

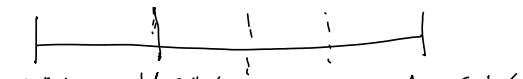
Familia: simbolistas

Tarea: clasificación y regresión

Idea matemática: selección iterativa de atributos, de forma que consiga un subconjunto de datos más puro con respecto al target.

1. Iteración: escoger atributos uno a uno, secuencialmente

avg (T) = 21% → variable: cloud cover % ⇒



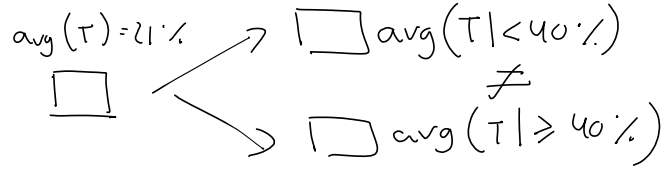
I. elige 1 atributo

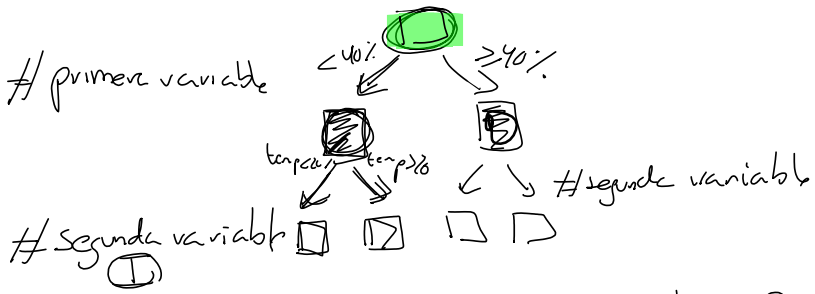
II. Decide un corte aleatorio

III Evalua pureza

III.1 Distintos cortes

III.2 Distintos atributos



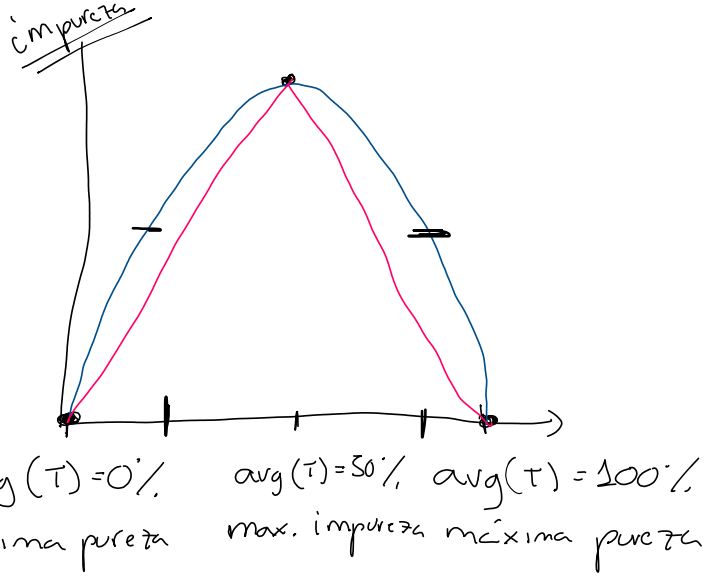


iteración

¿Hasta cuando hace preguntas? → lo decidimos nosotros (fine-tunning)

2. Más puro: grado de parecido del subconjunto con respecto al target

En clasificación



Fórmulas de impureza

- entropía
- gini

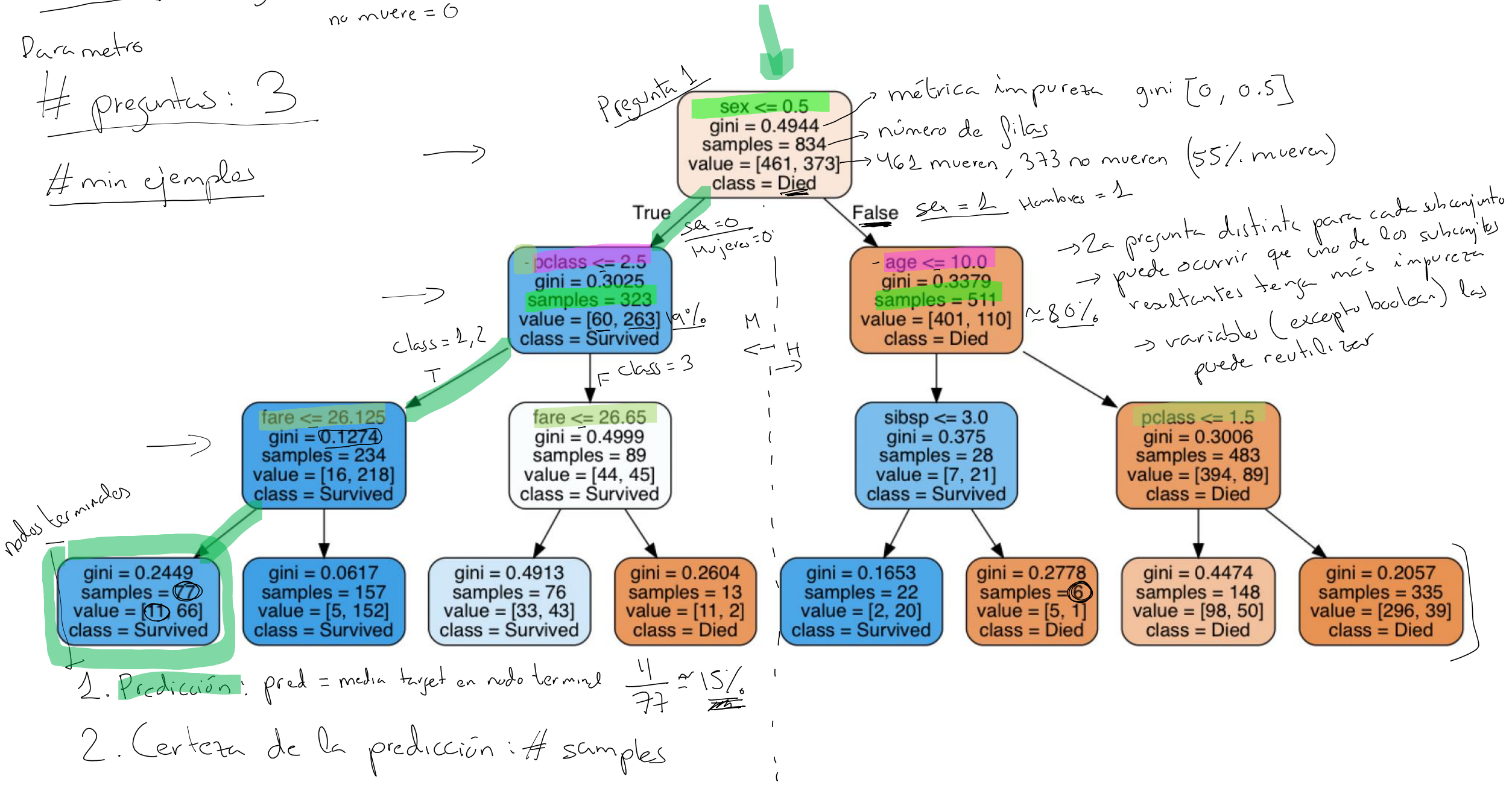
Titanic

target muere = 1
no muere = 0

Parametro

preguntas: 3

min ejemplos



1. Predicción: pred = media target en nodo terminal $\frac{11}{77} \approx 15\%$

2. Certeza de la predicción: # samples