DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Laboratorio de datos

Clasificación

Primer Cuatrimestre 2025

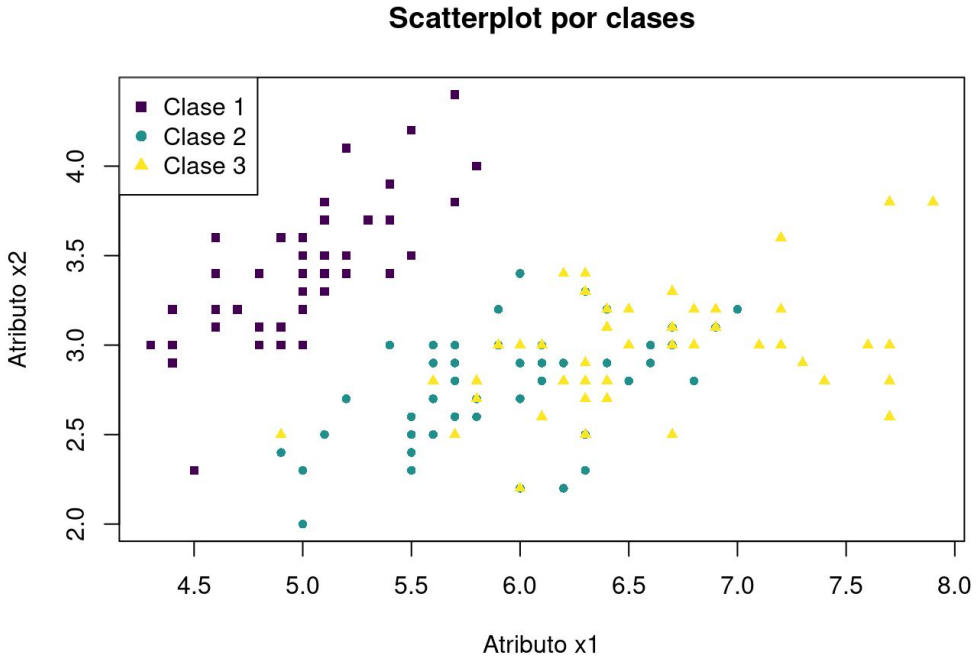
Clase de Paula Perez Bianchi, Viviana Cotik
y Manuela Cerdeiro

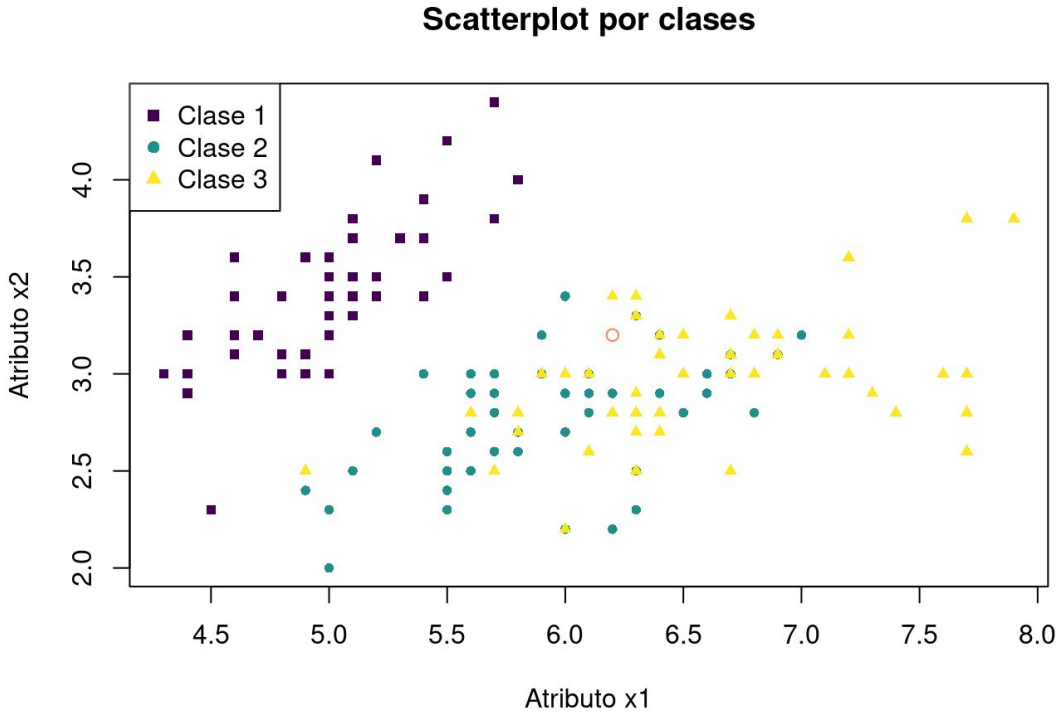
Ejemplo

Tenemos un conjunto de datos con variables *x1, x2, y*.

- Variables explicativas continuas *x1, x2*
- Variable a explicar *y* categórica

Graficamos *x1, x2* en los ejes.
La variable a explicar toma 3 valores: Clase 1, Clase 2, Clase 3 y se representa con símbolos/colores.





¿Qué clase le asignamos a la nueva observación?

Clasificación

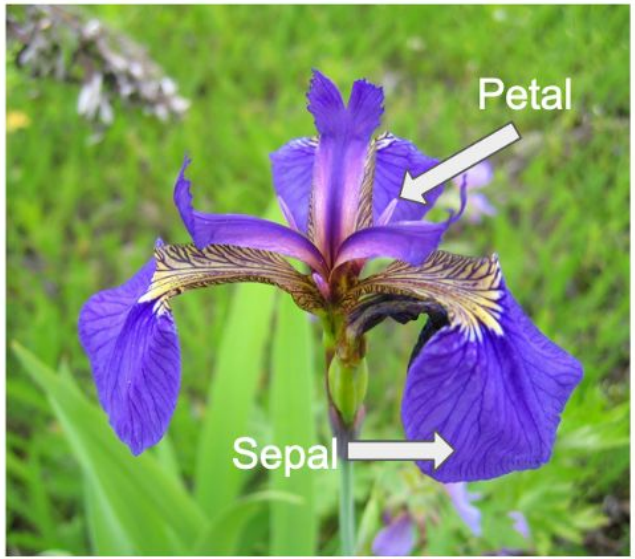
- A partir de los atributos (variables explicativas) queremos poder determinar la etiqueta - la variable categórica Y.
- Aprendizaje supervisado: contamos con un conjunto de entrenamiento en el cual conocemos las etiquetas - valores de la variable Y.
- Evaluación del modelo: medida relacionada con la cantidad de elementos bien o mal clasificados.

Métodos posibles (hay más)

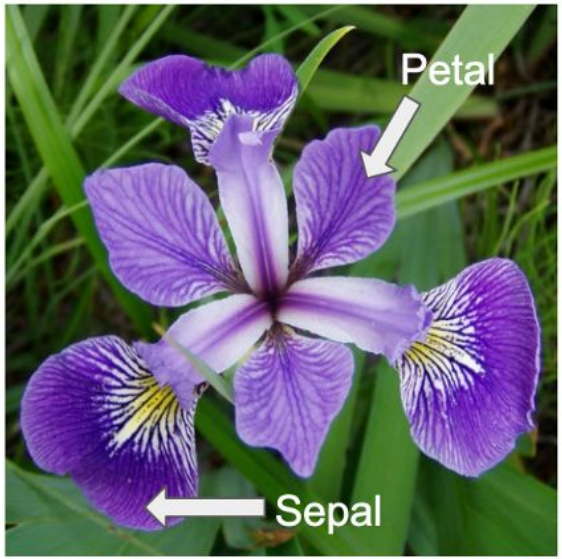
- Umbral ($x_1 > c$)
- Regresión logística
- Árboles de decisión
- Support Vector Machines (SVM)
- K-Nearest Neighbors (KNN)

Dataset de flores - Iris

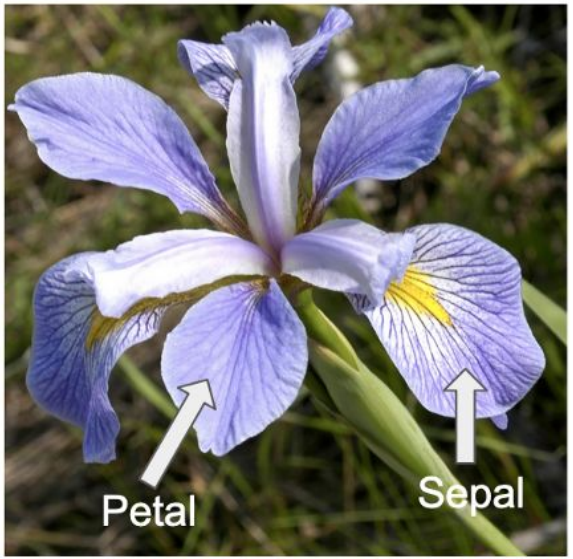
Iris setosa



Iris versicolor



Iris virginica



50 muestras de cada una de tres especies de flores *Iris: setosa, versicolor y virginica*. De cada flor se midieron 4 atributos: largo y ancho del sépalo y del pétalo.

Fisher - 1936

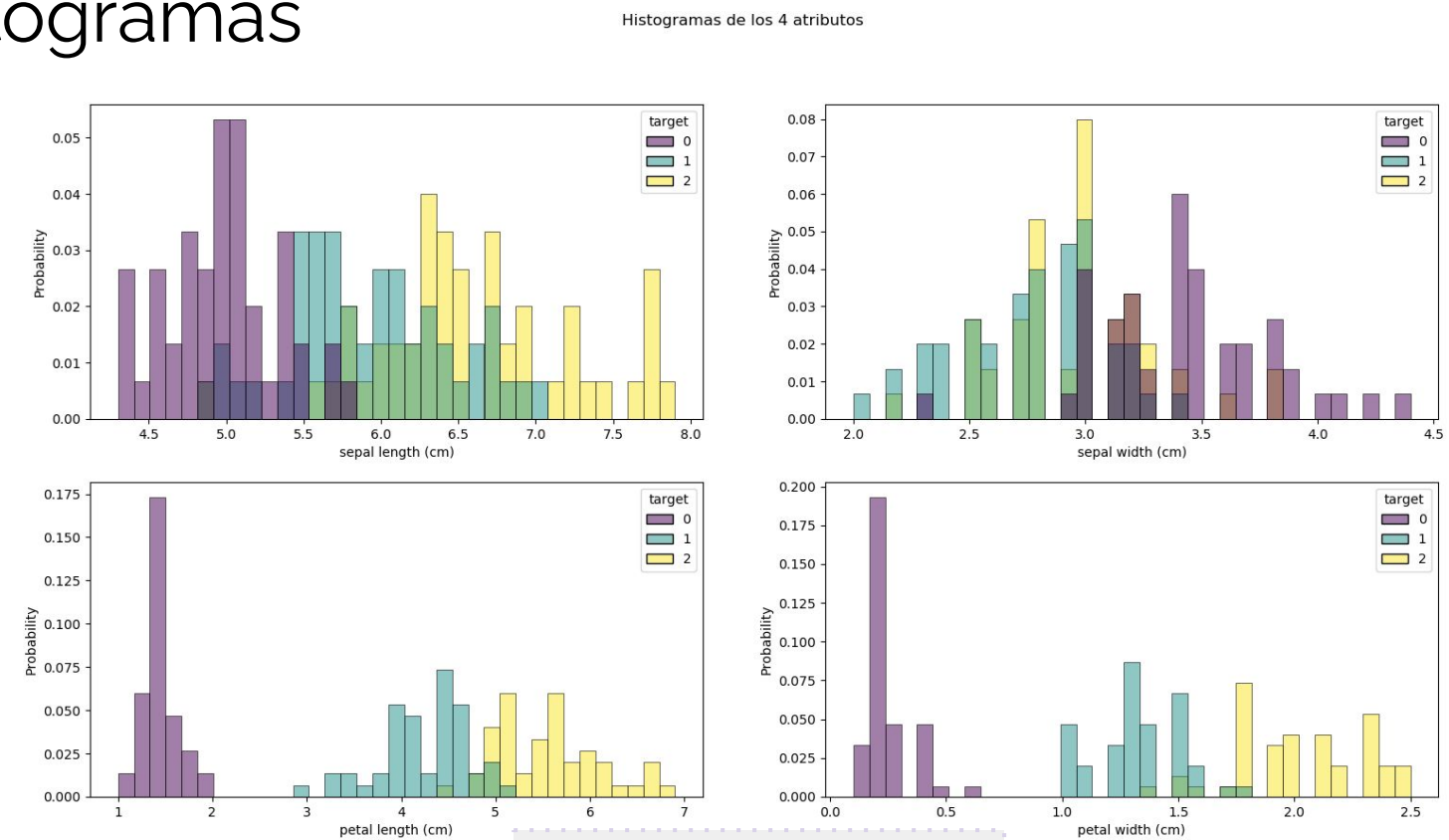
Clasificación de Flores

¿Ante las medidas de una nueva flor, cómo podríamos determinar su especie?

¿Cuál resulta el atributo más importante a la hora de distinguir entre especies?

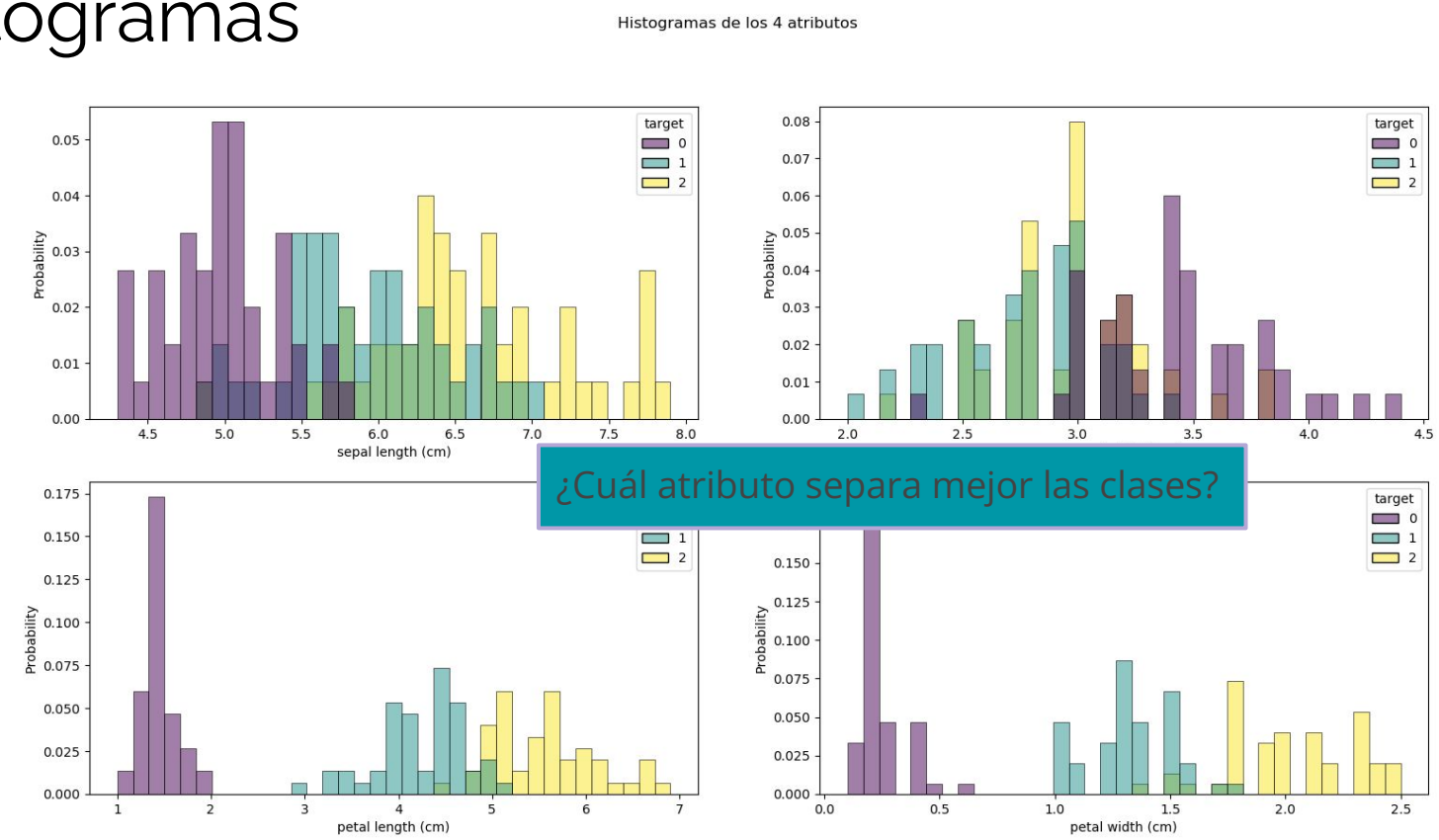
¿Qué herramienta puede ayudarnos a responder esto?

Histogramas



{0: 'setosa', 1: 'versicolor', 2: 'virginica'}

Histogramas



Clasificamos por largo del pétalo

```
def clasificador_iris
```

Clasificamos por largo del pétalo

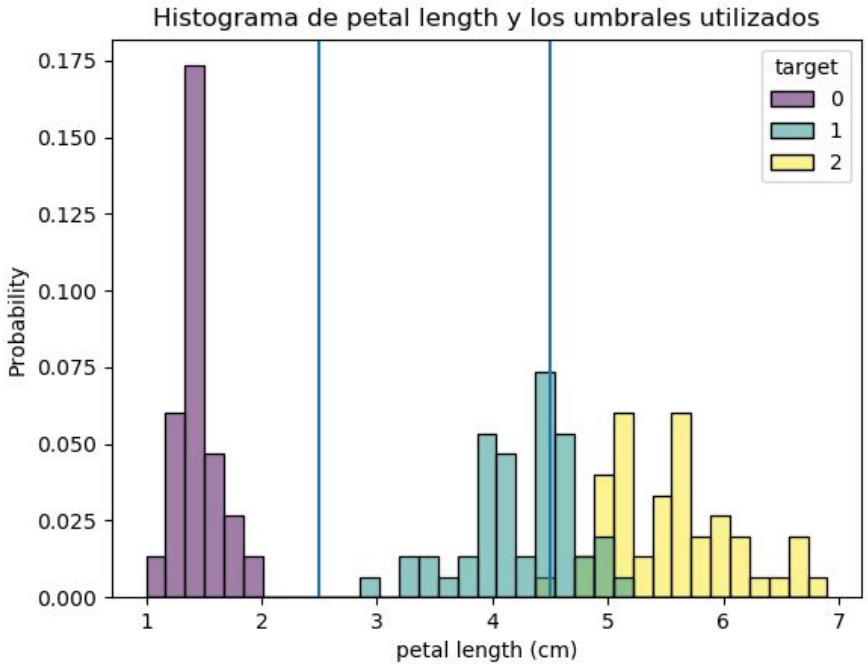
Ahora veamos cómo se comporta este clasificador.

¿Podemos mejorar el clasificador?
¿Cómo comparamos entre dos clasificadores?

Clasificamos por largo del pétalo

Ahora veamos cómo se comporta este clasificador.

Éstas son las líneas de corte.



Medidas para evaluar clasificadores

Matriz de confusión

Para cada clase i, nos fijamos cuántas observaciones de la clase fueron clasificadas en cada clase j.

Esto nos da una matriz cuadrada, con una fila y columna por cada clase.

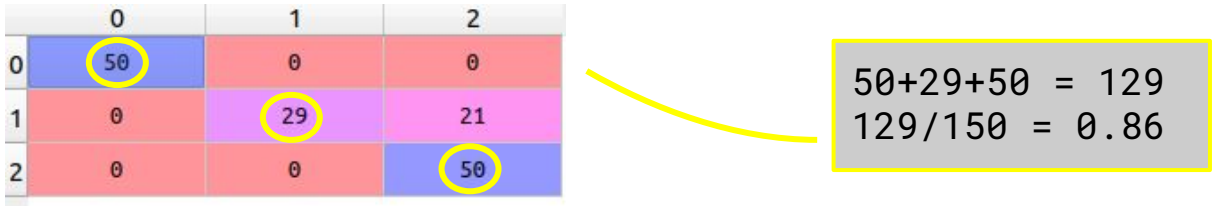
	0	1	2
0	50	0	0
1	0	29	21
2	0	0	50

{0: 'setosa', 1: 'versicolor', 2: 'virginica'}

Medidas para evaluar clasificadores

Exactitud

La exactitud o *accuracy* que es una medida numérica que cuenta la proporción de observaciones *bien* clasificadas.

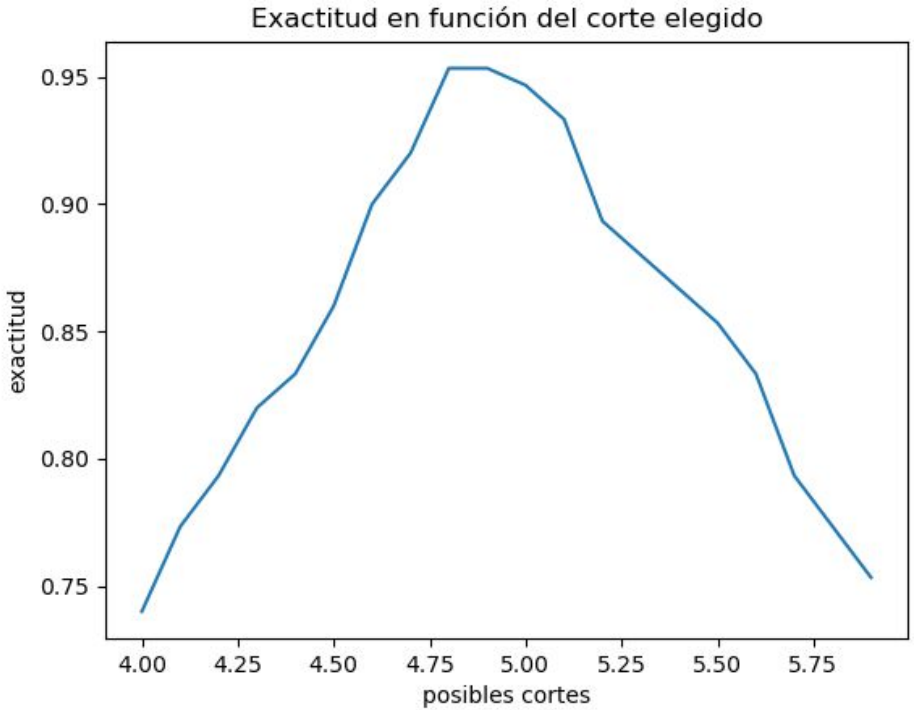


¿Buscamos el mejor clasificador?

Podemos

+ recorrer muchos posibles umbrales dentro de un rango
+ para cada umbral correr el clasificador y evaluarlo
+ comparar los clasificadores para seleccionar el mejor

Comparación de clasificadores



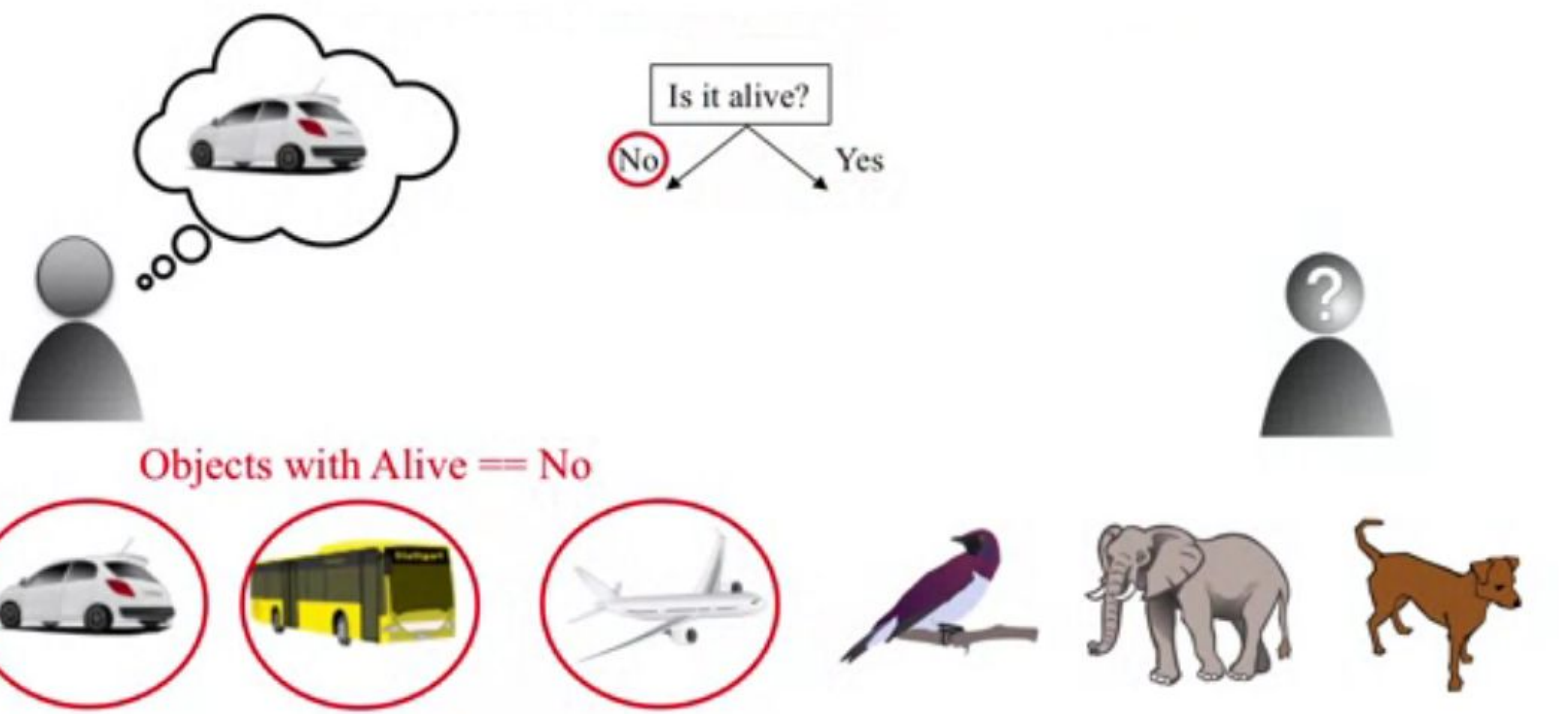

```
corte_selec =
posibles_cortes[np.argmax(exactitudes)]
```

Árboles de decisión

Árboles de decisión

- método para **inferencia inductiva**
- aprenden **reglas if-then** sobre los valores de los atributos. Predicen valor objetivo en función de las reglas.

Árboles de decisión - Ejemplo



Applied Machine Learning with Python. Universidad de Michigan

Árboles de decisión - Ejemplo

nodo: representa pregunta

aristas: representan posibles respuestas

hojas: nodos que representan decisiones

caminos desde la raíz

Árboles de decisión

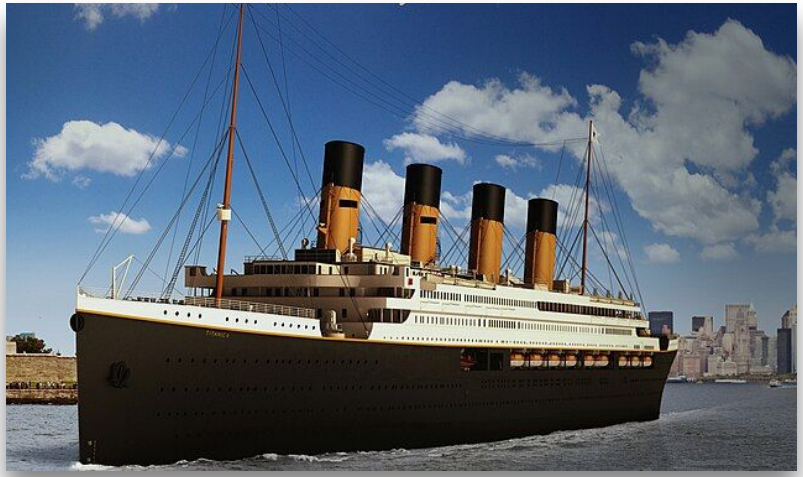
- Modelo de aprendizaje supervisado utilizado principalmente para **clasificación**.
- Se basa en el armado de una **jerarquía de reglas.** Estas las podemos expresar usando una fórmula lógica de ANDs y ORs.
- Es decir que el árbol representa una **disyunción de conjunciones sobre valores de atributos.**

Podemos pensar el armado del árbol como jugar al ¿Quién es quién?.

Los árboles son un ***modelo altamente interpretable***. Es decir que dada una predicción particular podemos entender por qué el modelo la generó. Sólo hay que mirar la rama de la hoja correspondiente a la predicción.

Dataset Titanic

El Titanic se hundió en Abril de 1912.
Tenía la fama se ser “Inhundible” pero chocó con un iceberg y no le alcanzó con la fama.
Como no había suficientes salvavidas 1502 de 2224 pasajeros y personal de a bordo murieron.



Dataset en:
https://www.kaggle.com/competitions/titanic

	PassengerId	Pclass	Sex	Age	SibSp	Parch	Fare
1	2	1	female	38.0	1	0	71.2833
3	4	1	female	35.0	1	0	53.1000
6	7	1	male	54.0	0	0	51.8625
10	11	3	female	4.0	1	1	16.7000
11	12	1	female	58.0	0	0	26.5500

Objetivo: ¿Podemos predecir quienes sobrevivieron al titanic?

Actividad Titanic

Consigna 1: ¿Qué características tenían los pasajeros que sobrevivieron? ¡Exploren los datos y escríbanlas en un papel (o grafíquenlas)!

Algunas preguntas:

- ¿Cuántos pasajeros de primera clase había? ¿Y de segunda y tercera?
- ¿Cuál era la proporción de niños?
- ¿Quiénes les parece que tenían prioridad en los botes de rescate?

Consigna 2: Tenemos los datos de una pasajera,
¿pueden decir si sobrevivió o no?

	Pclass	Sex	Age	SibSp	Parch	Fare	Survived
184	1	female	27.0	1	1	247.5208	???

¿Adivinaron o usaron reglas? ¿Cuáles? ¿Se pueden generalizar?

Consigna 3: Competencia

PassengerId	Pclass	Sex	Age	SibSp	Parch	Fare

Ahora les voy a dar los datos de **10 pasajeros**,
¿Pueden responder cuáles sobrevivieron?

Armar un conjunto de reglas generales, basadas solamente en los atributos indicados.
Generar etiquetas para estos 10 pasajeros.

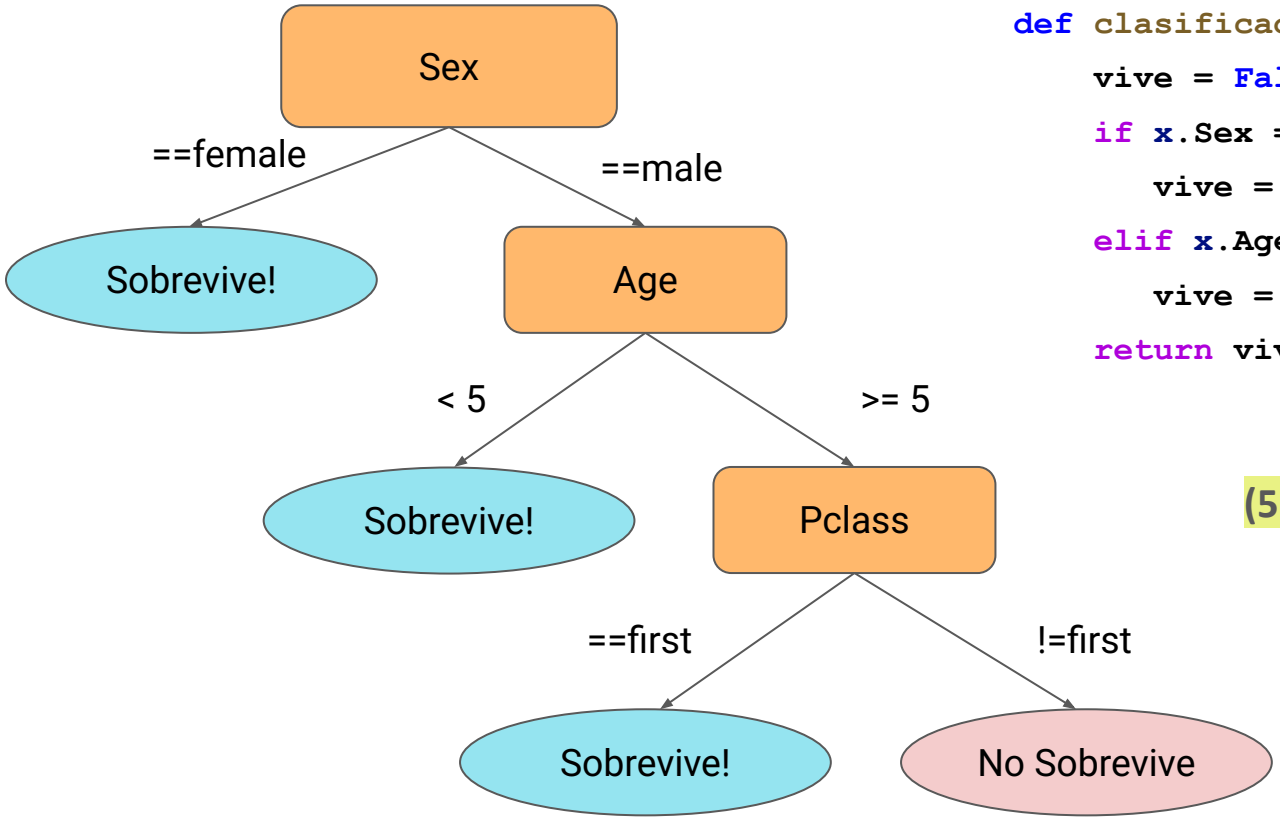
```
def clasificador_titanic(x):
    vive = False
    if REGLA:
        vive = True
    return vive
```

Tienen 10 min!

https://www.online-stopwatch.com/timer/15minute/

🚣 → Sobrevivió (True)
💀 → No sobrevivió (False)

Subir sus predicciones y el código que usaron para sus reglas a un google form.



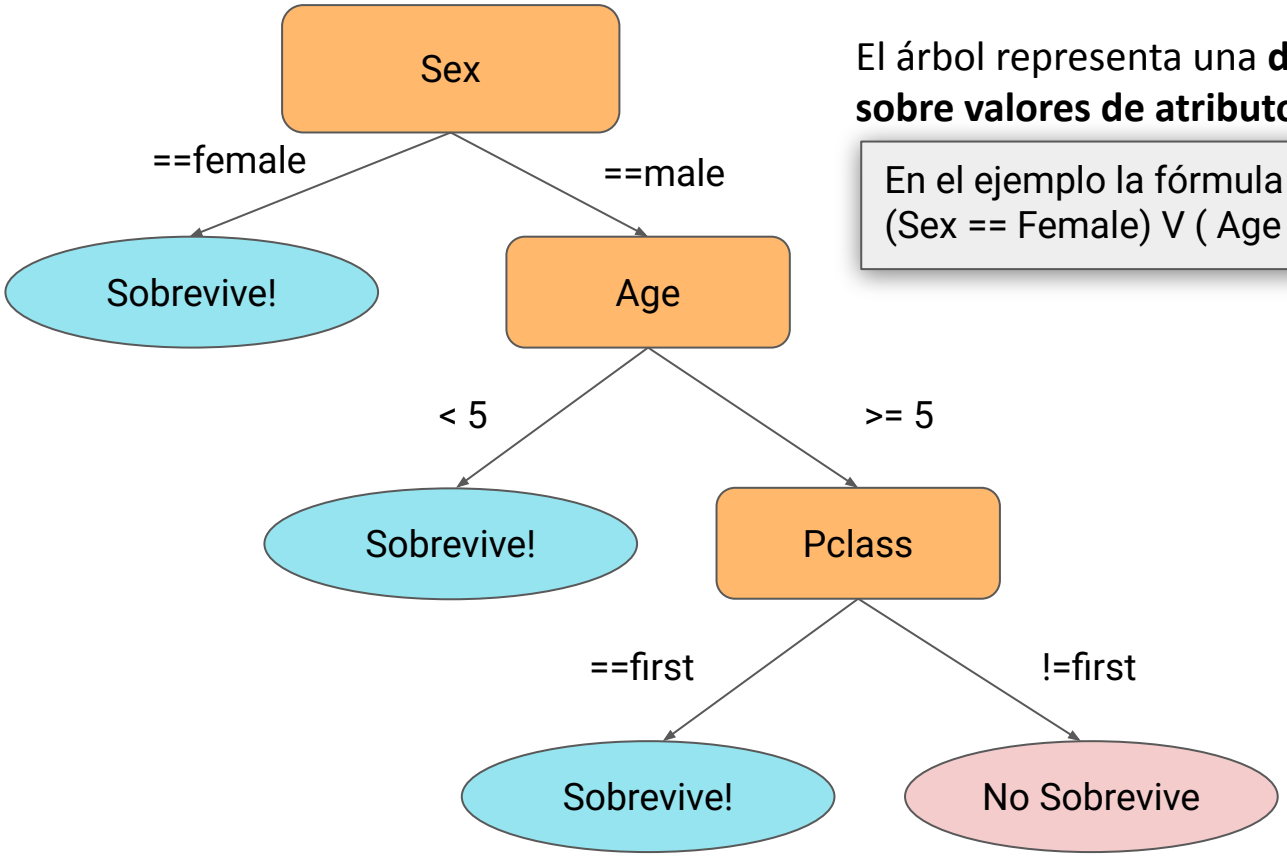

```
def clasificador_titanic(x):
    vive = False
    if x.Sex == "female":
        vive = True
    elif x.Age < 5 or x.Pclass == 1 :
        vive = True
    return vive
```

SCORE : 50%
(5 instancias bien clasificadas)

*Cada nodo interno evalúa un atributo discreto **a***
*Cada rama corresponde a **un valor/umbral** para **a***
*Cada hoja predice un valor de **Y***

Esto que hicimos a ojo se puede formalizar como un método de Aprendizaje Supervisado

Árboles de decisión



El árbol representa una **disyunción de conjunciones sobre valores de atributos.**

En el ejemplo la fórmula quedaría:
(Sex == Female) V (Age < 5) V (Age >= 5 Λ Pclass == 1)

Competencia - Evaluamos los clasificadores.

Vemos respuestas del form.
Evaluamos también con un conjunto de test.
Accuracy - cantidad de errores de cada tipo?

Inducción Top-Down para árboles de decisión

(versión simplificada de ID-3 y C4.5 (Quinlan))

Input: **S** un conjunto de instancias con atributos **A**.

1. Elegimos **a ϵ A**, el atributo que produce el **mejor corte** de **S** para el nodo actual.
2. Para cada valor $\mathbf{v_i}$ posible de **a**, crear un nuevo hijo del nodo actual.
3. Clasificar (repartir) las instancias en los nuevos nodos, según el valor del atributo **a**.

   $$\mathbf{S_i \leftarrow \{ x \mid x \in S \wedge x[a] = v_i\}}$$

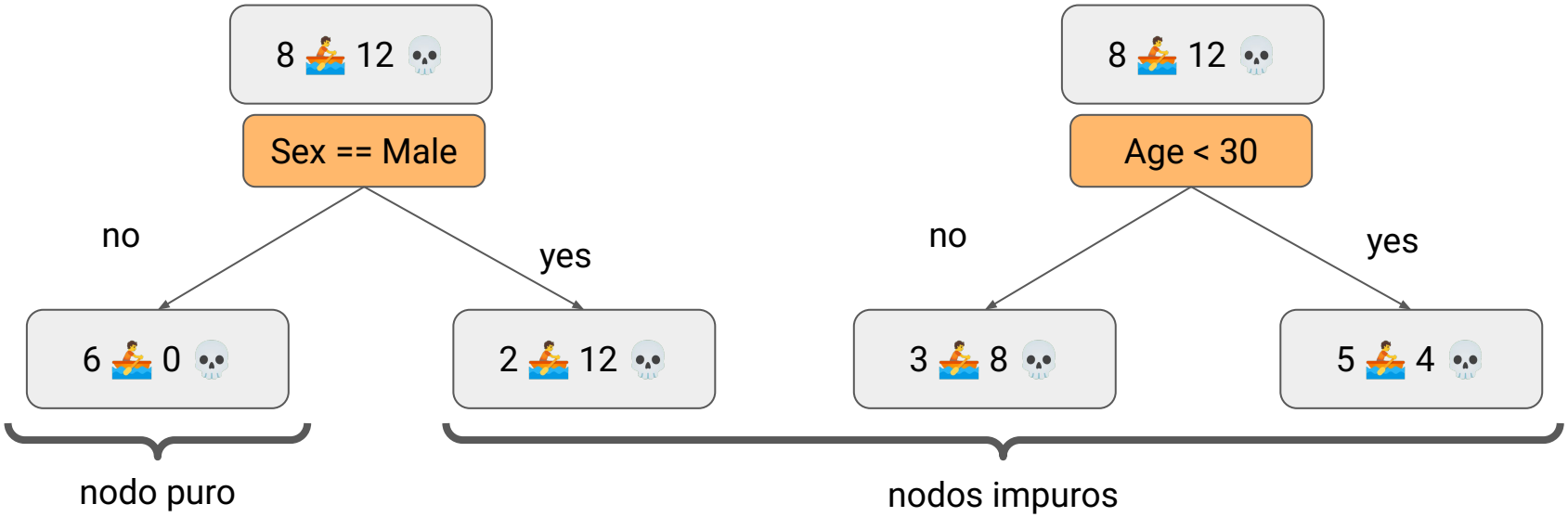
4. Repetir para cada hijo del nodo en el que haya instancias de más de una clase (salvo que se cumpla algún criterio de corte en cuyo caso terminamos).

¿Cómo definimos el mejor corte?
→ Necesitamos métricas!!

Para pensar: ¿Cómo hacemos el paso 2) para atributos continuos? ¿Cómo se definen las regiones del paso 3) ¿en ese caso?

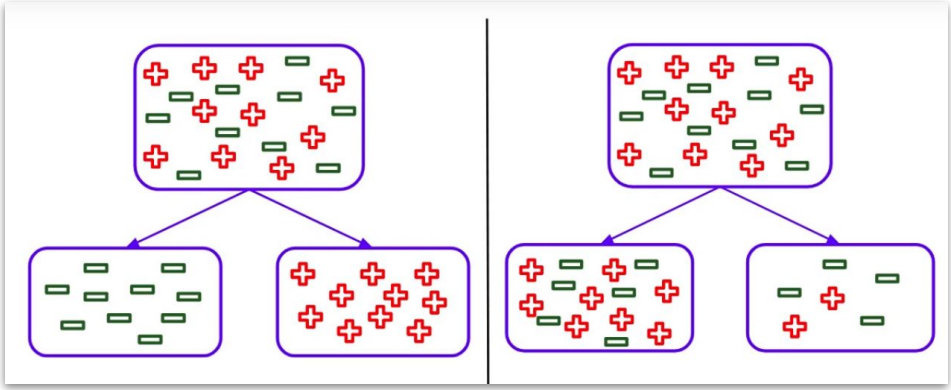
Árboles de decisión - Medidas de impureza

🚣 → Sobrevivió
💀 → No sobrevivió



¿Cuál corte es mejor? ¿Qué lo hace mejor? ¿Cómo medirían esto?

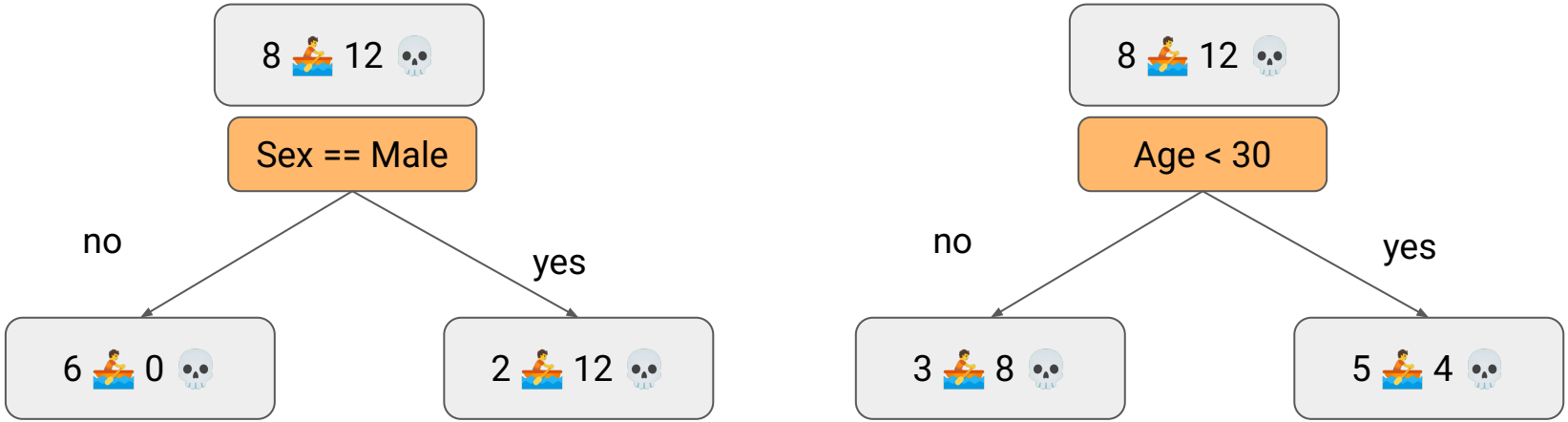
Árboles de decisión - Medidas de impureza



Bajó mucho la impureza
Buena elección de pregunta

No bajó mucho la impureza

Árboles de decisión - Medidas de impureza



$$\Delta M = M(S) - \left(\frac{|S_{\text{sí}}|}{|S|} M(S_{\text{sí}}) + \frac{|S_{\text{no}}|}{|S|} M(S_{\text{no}}) \right)$$

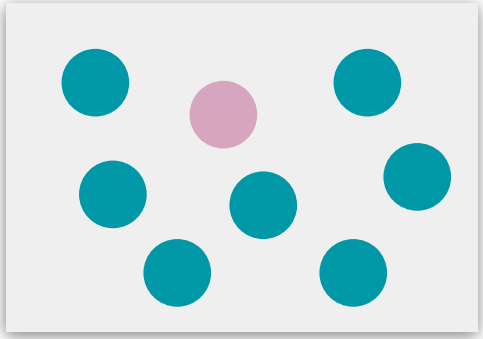
cuánto gano, según la medida **M**, si divido a **S** en regiones con una pregunta

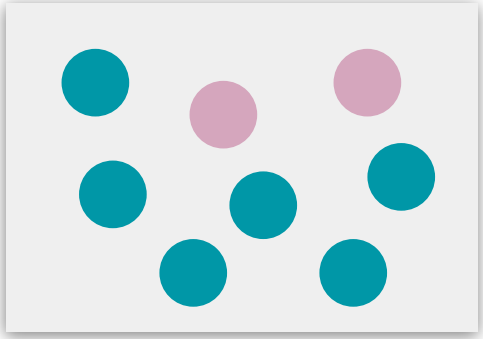
Árboles de decisión - Medidas de impureza

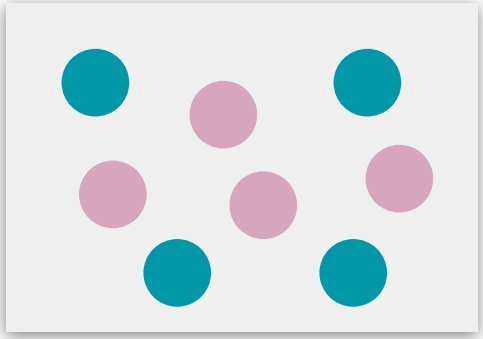
Entropía

Mide la incertidumbre inherente a los posibles resultados de una variable aleatoria.

Voy a sacar una bolita. ¿Cuánta incertidumbre hay?

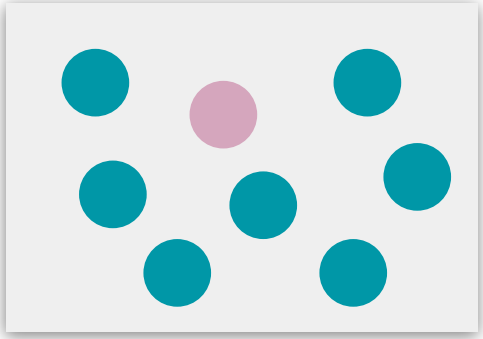




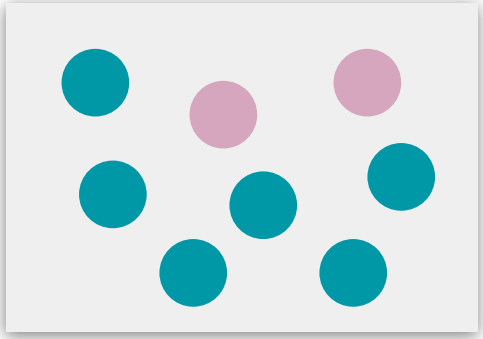


$$H = - \sum p_i \log_2(p_i)$$

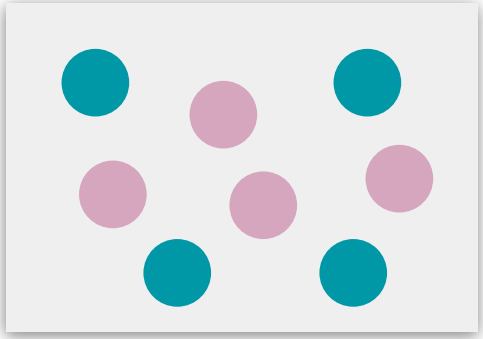
$p_A = 7/8$, $P_R = 1/8$
H = 0.544



$p_A = 6/8$, $P_R = 2/8$
H = 0.811



$p_A = 4/8$, $P_R = 4/8$
H = 1



Árboles de decisión - Medidas de impureza

Info Gain (Ganancia de Información)

Mide cuánta entropía removemos al hacer un corte.
A menor entropía, mayor información.

$$\text{InfoGain} = H(S) - \left(\frac{|S_{\text{sí}}|}{|S|} H(S_{\text{sí}}) + \frac{|S_{\text{no}}|}{|S|} H(S_{\text{no}}) \right)$$

$$H(S) = -\sum_{k \in S} p_S(k) \log_2 p_S(k)$$

Árboles de decisión - Medidas de impureza

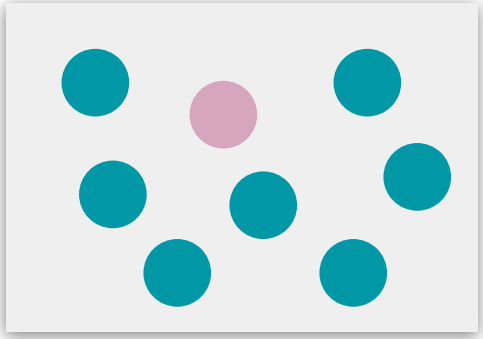
Impureza de Gini

Mide la probabilidad de que una instancia particular sea clasificada erróneamente si esta fuese etiquetada aleatoriamente de acuerdo con la distribución de clases dentro de la región.

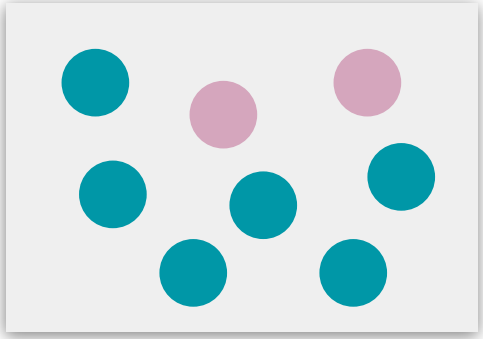
$$G(S) = 1 - \sum_{k \in S} p_S(k)^2$$

$$G = 1 - \sum p_i^2$$

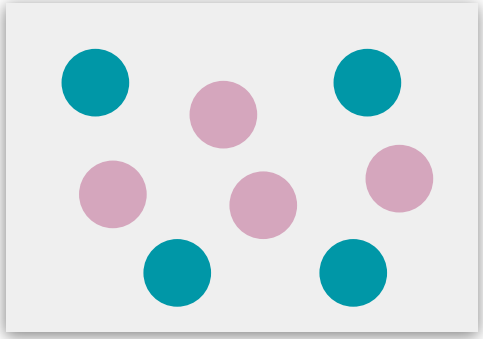
$p_A = 7/8$, $P_R = 1/8$
G = 0,21875



$p_A = 6/8$, $P_R = 2/8$
G = 0,375



$p_A = 4/8$, $P_R = 4/8$
G = 0.5



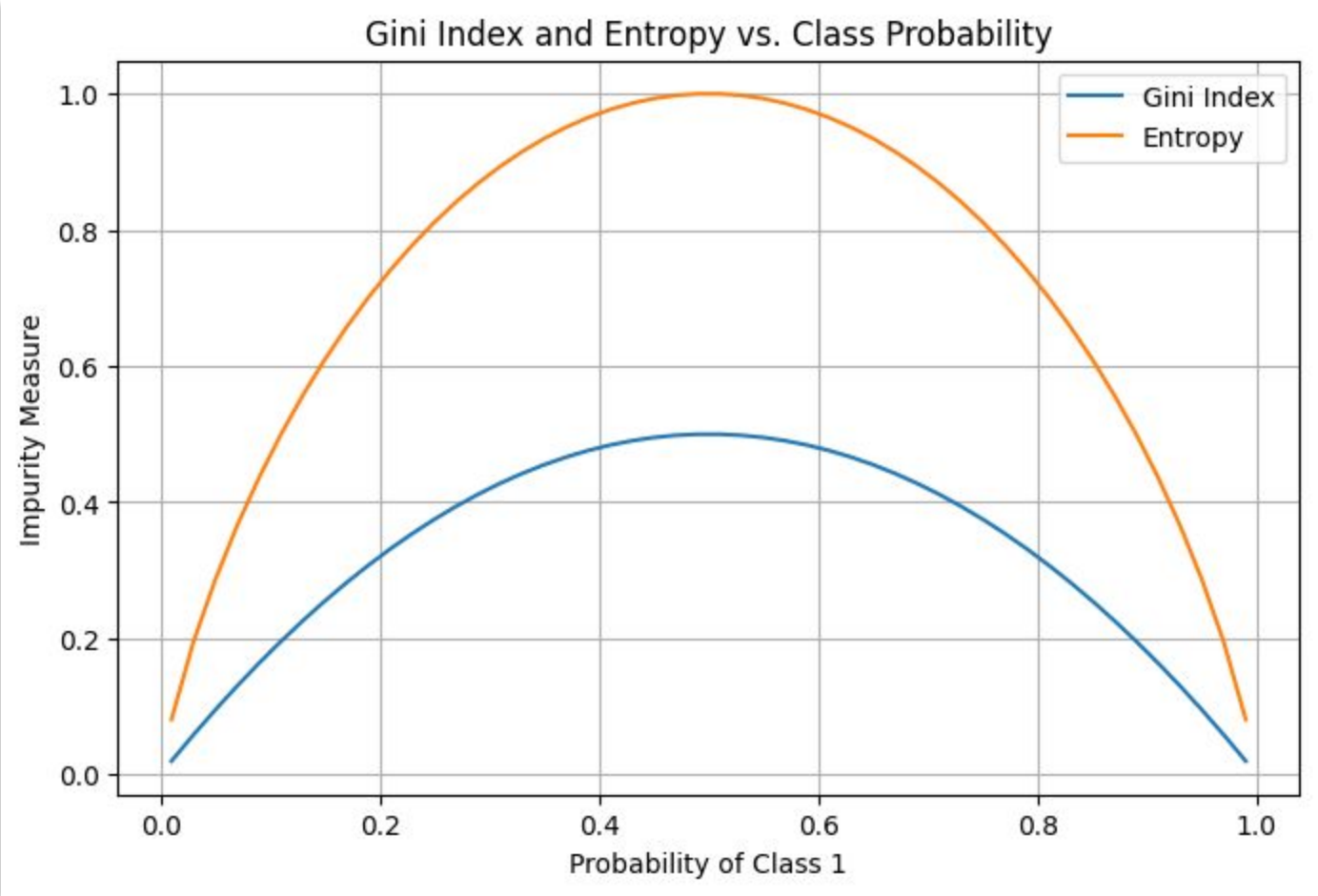
Árboles de decisión - Medidas de impureza

Gini Gain

Mide la probabilidad de que una instancia particular sea clasificada erróneamente si esta fuese etiquetada aleatoriamente de acuerdo con la distribución de clases dentro de la región.

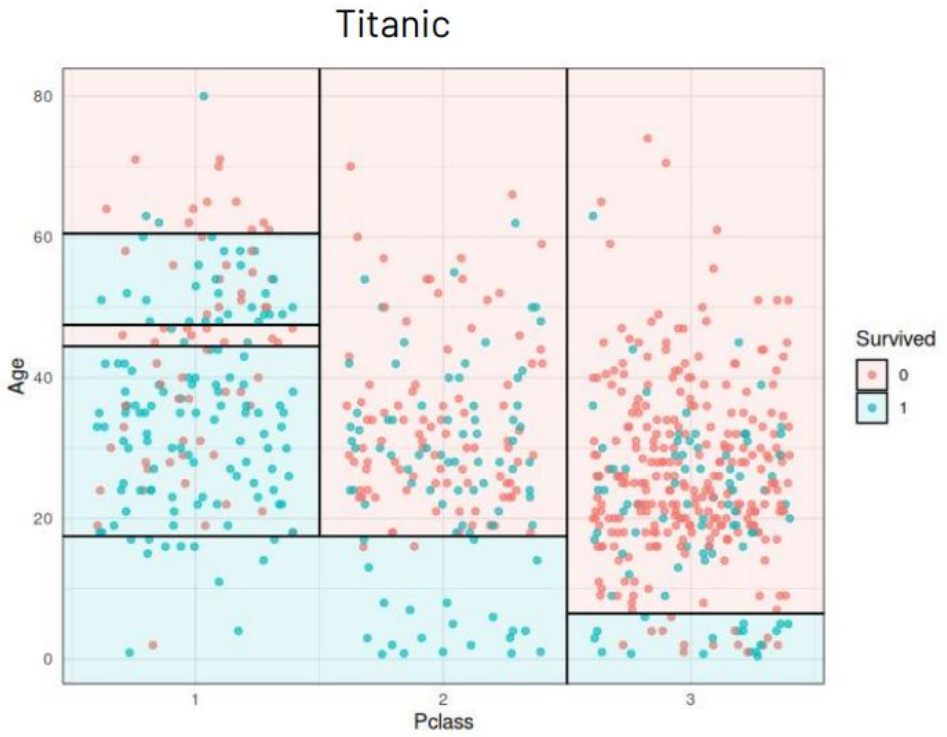
$$\mathrm{GiniGain} = G(S) - \left(\frac{|S_{\text{sí}}|}{|S|} G(S_{\text{sí}}) + \frac{|S_{\text{no}}|}{|S|} G(S_{\text{no}}) \right)$$

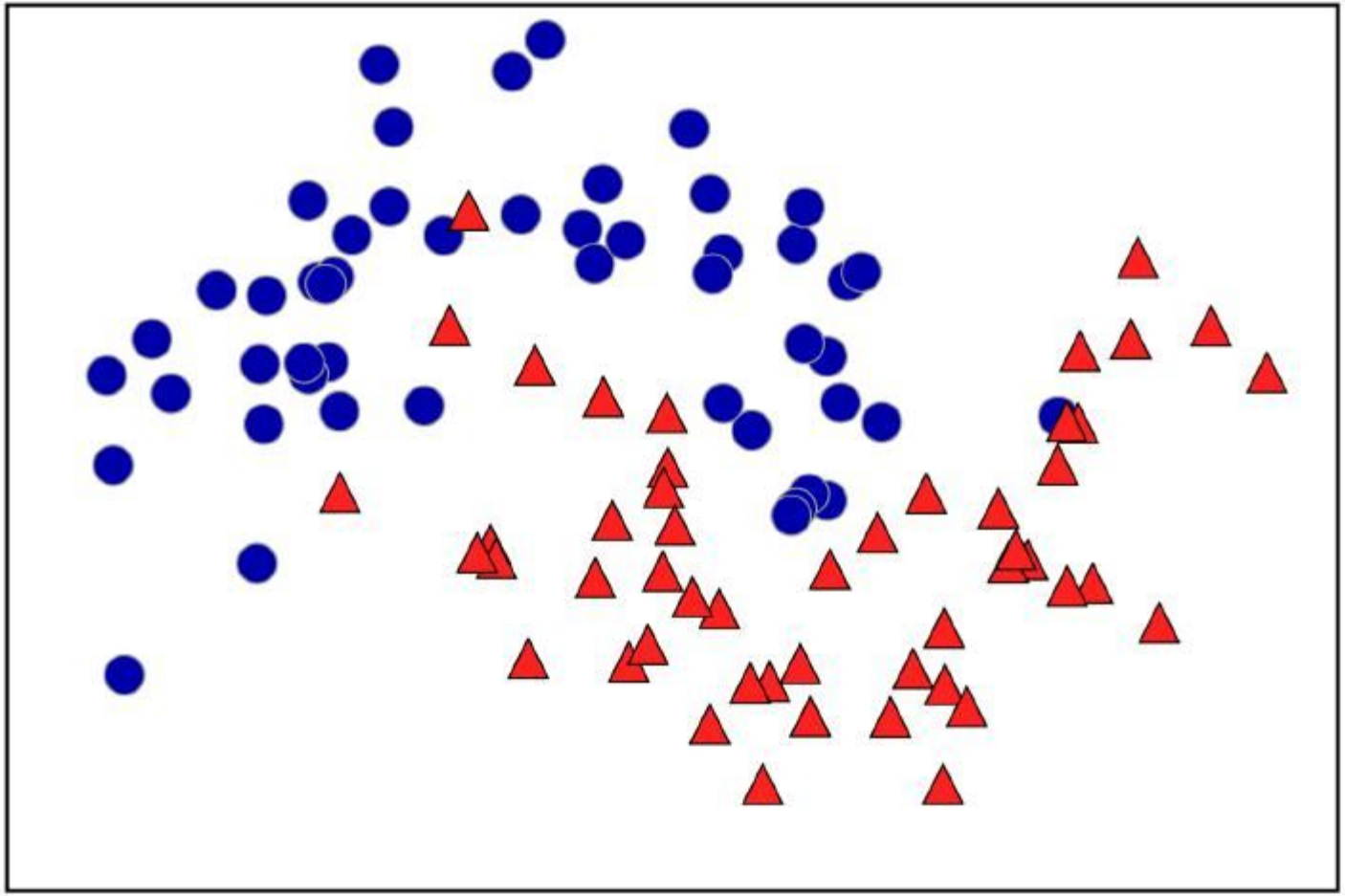
$$G(S) = 1 - \sum_{k \in S} p_S(k)^2$$


Gini Index and Entropy vs. Class Probability
Gini Index
Entropy
Impurity Measure
Probability of Class 1
0.0
0.2
0.4
0.6
0.8
1.0

Fronteras de decisión

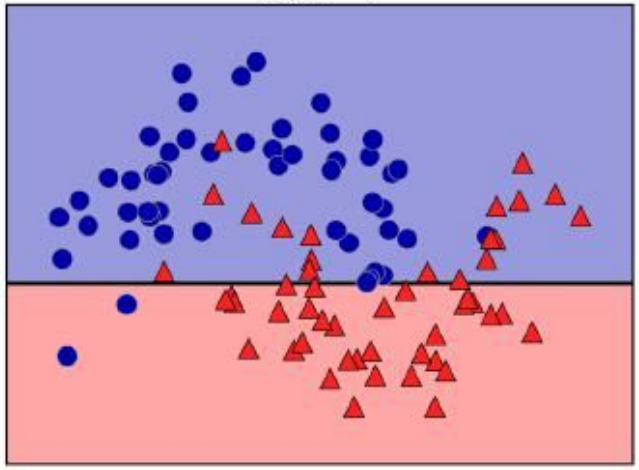
El tipo de regiones de decisión que puede generar un árbol de decisión tienen forma de rectángulos.

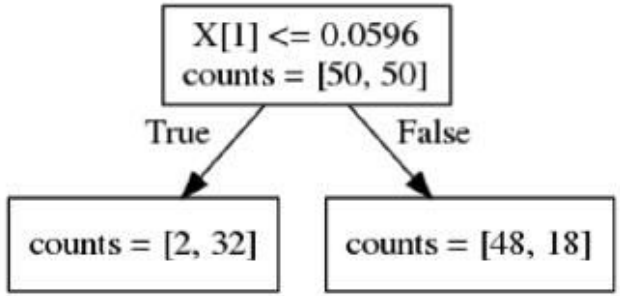




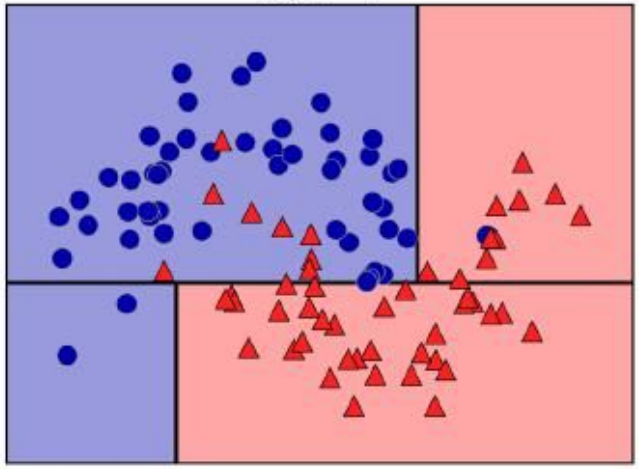
Ejemplo del libro de Müller & Guido

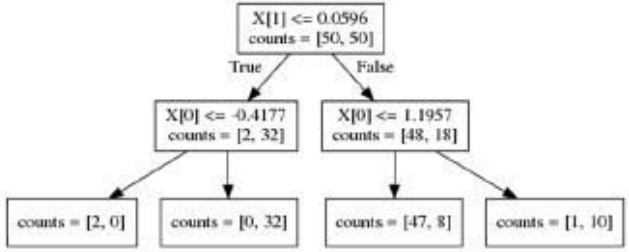
depth = 1

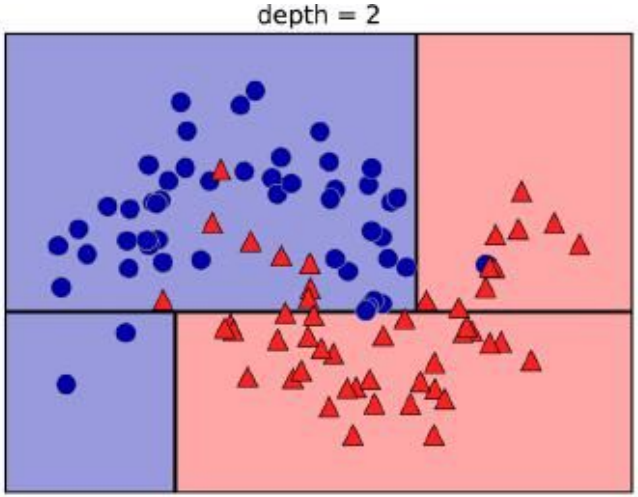


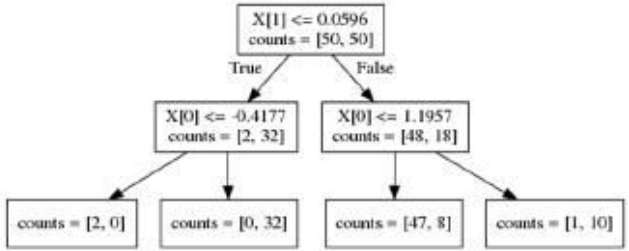


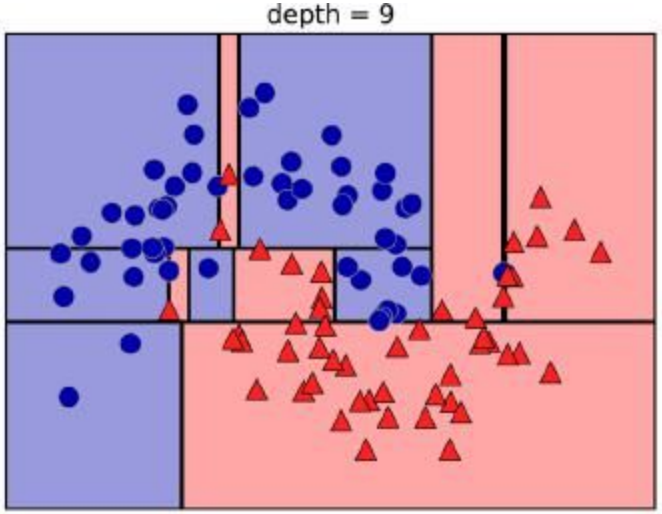
depth = 2

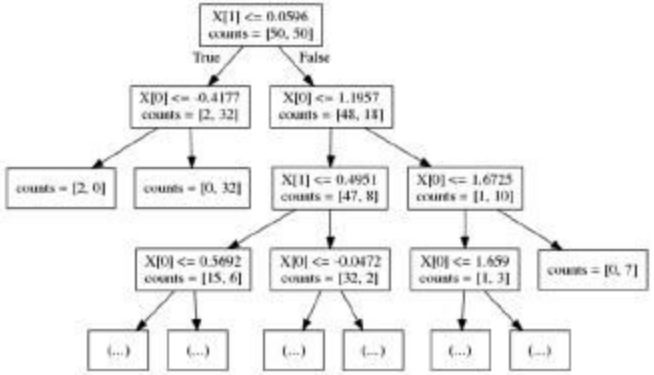





depth = 2


X[1] <= 0.0596
counts = [50, 50]
True
False
X[0] <= -0.4177
counts = [2, 32]
X[0] <= 1.1957
counts = [48, 18]
counts = [2, 0]
counts = [0, 32]
counts = [47, 8]
counts = [1, 10]


depth = 9


X[1] <= 0.0596
counts = [50, 50]
True
False
X[0] <= -0.4177
counts = [2, 32]
X[0] <= 1.1957
counts = [48, 18]
counts = [2, 0]
counts = [0, 32]
X[1] <= 0.4951
counts = [47, 8]
X[0] <= 1.6725
counts = [1, 10]
X[0] <= 0.5692
counts = [15, 6]
X[0] <= -0.0472
counts = [32, 2]
X[0] <= 1.659
counts = [1, 3]
counts = [0, 7]
(...)
(...)
(...)
(...)
(...)
(...)

Sesgo inductivo

Se denomina sesgo inductivo de un algoritmo de aprendizaje automático al conjunto de afirmaciones que el algoritmo utiliza para construir un modelo.

- Incluye:
  - forma de las hipótesis
  - características de funcionamiento del algoritmo (cómo recorre el espacio de hipótesis hasta elegir un único modelo)

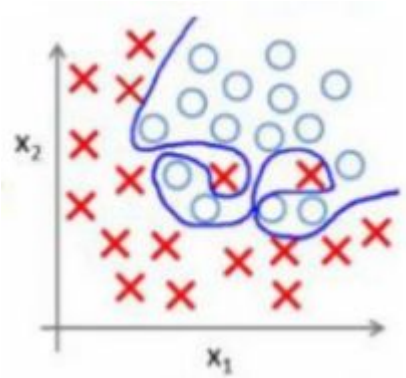
Sesgo inductivo

- El tipo de regiones de decisión que puede generar tienen forma de rectángulos, con cortes paralelos a los ejes.
- Las regiones que exploramos se determinan de manera *Greedy*. En cada paso optamos por mejorar la métrica de manera local. No se revisan decisiones previas. No se consideran todas las combinaciones de regiones posibles.

Overfitting - Sobreajuste

En árboles de decisión, hay sobreajuste cuando el árbol es “demasiado” profundo

¿Qué pasa si hay **descripciones exactas de instancias únicas y aisladas**?



Fernando Berzal, DECSAI, Universidad de Granada

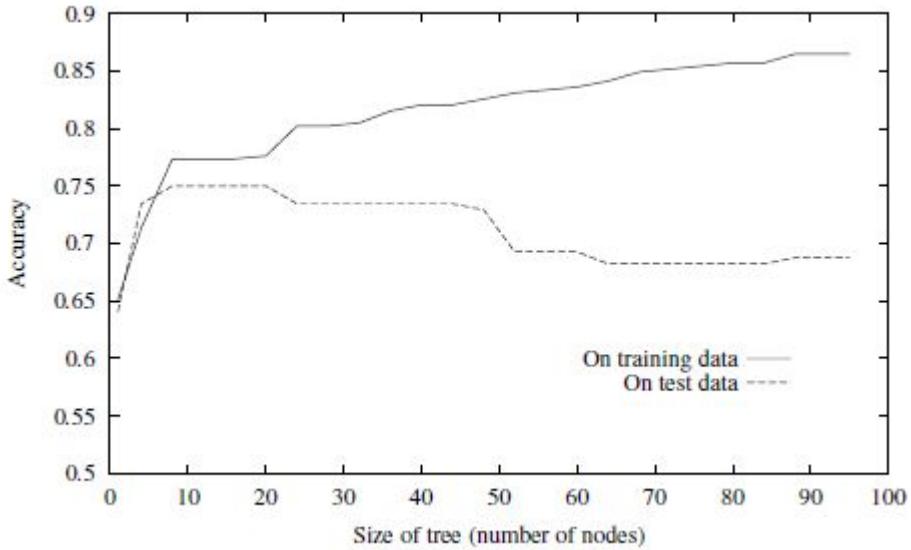
Overfitting

Datos:

- entrenamiento
- test (independiente)

Accuracy (exactitud):

(TP+TN)/(TP+TN+FP+FN)



Machine Learning, Tom M.Mitchel

Overfitting en Árboles de Decisión - Cómo evitarlo

Soluciones:

- **detener crecimiento del árbol** antes de que clasifique perfectamente a los datos
- hacer crecer el **árbol entero, luego podar (post-prune)**

Resumen

- aprendizaje supervisado
- para clasificación y regresión
- fáciles de usar y de entender
- buen método exploratorio para ver qué atributos son importantes
- sesgo, overfitting

Ventajas:

- son altamente interpretables
- fácil visualización
- se pueden usar atributos binarios, categóricos o continuos

Desventajas:

- pueden tener sobreajuste
- suelen necesitarse ensambles de árboles para tener mejor performance

En Python

class `sklearn.tree.`**`DecisionTreeClassifier`**(**, criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, class_weight=None, ccp_alpha=0.0*) [source]

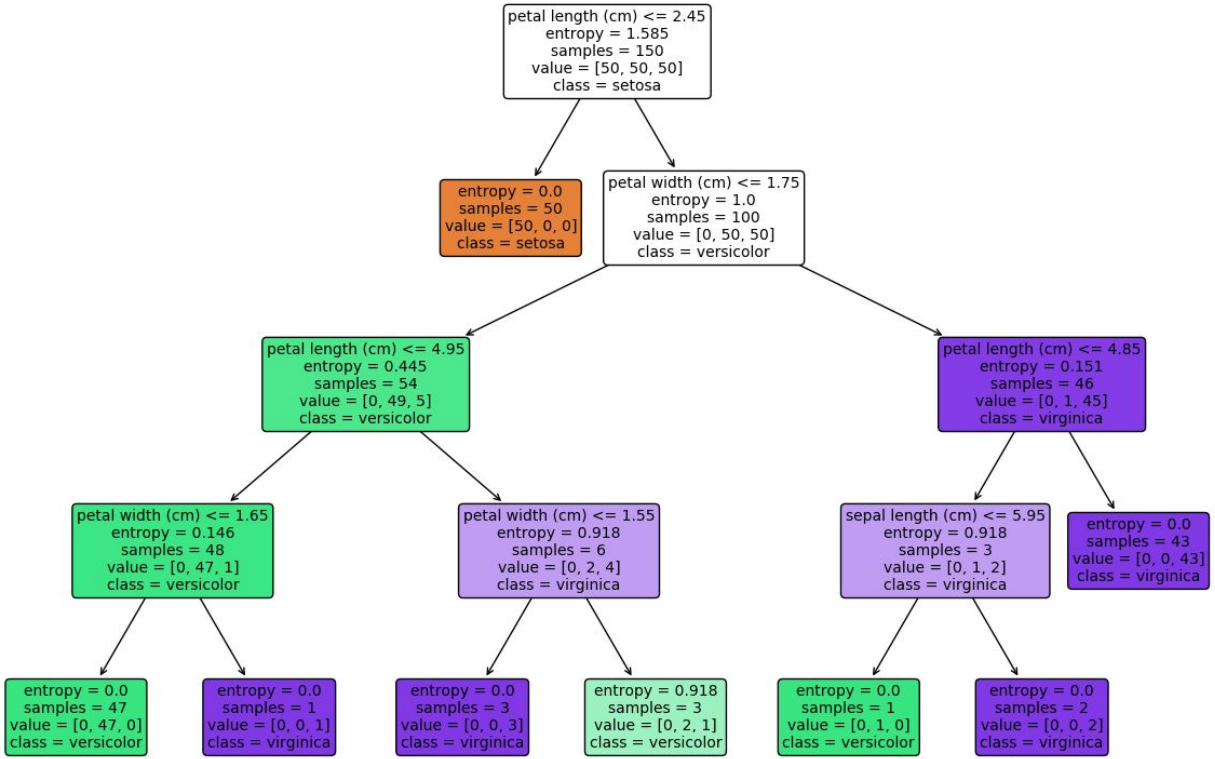
```python
from sklearn.tree import DecisionTreeClassifier

arbol = DecisionTreeClassifier()
arbol.fit(X, y) # Entrenamiento del modelo

prediction = arbol.predict(X) # Generamos las predicciones // llamamos al modelo
```

Hiperparámetros
→ Son los parámetros que nos permiten configurar el modelo.
Por ejemplo, la altura máx del árbol de decisión.

Ejemplo con Iris



Representación del árbol

```
#%% representarlo gráficamente
plt.figure(figsize= [20,10])
tree.plot_tree(clf_info, feature_names = ['altura_tot', 'diametro',
'inclinacio'], class_names = ['Ceibo', 'Eucalipto', 'Jacarandá',
'Pindó'],filled = True, rounded = True, fontsize = 8)

#%% otra forma de ver el arbol
r = tree.export_text(clf_info, feature_names=['altura_tot', 'diametro',
'inclinacio'])
print(r)
#%%
```

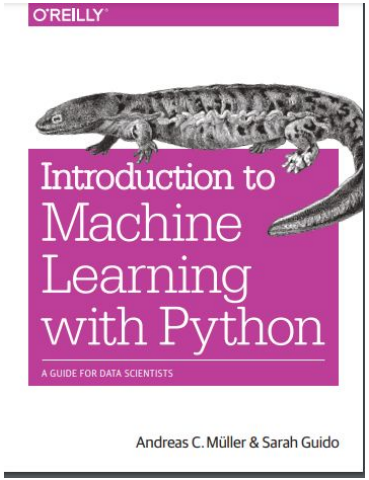
Tarea

Resolver la guía de ejercicios de clasificación con árboles de decisión.

Bibliografía

Libros:

- Introduction to Machine Learning with Python, Müller & Guido
- Machine Learning - Mitchell



Artículos:

- Induction of Decision Trees. Quinlan. http://hunch.net/~coms-4771/quinlan.pdf
- Simplifying Decision Trees. Quinlan. https://www.sciencedirect.com/science/article/pii/S0020737387800536

Recomendación

Película: *The Bit Player.*
Documental sobre Claude Shannon, considerado el padre de la teoría de la información, es quien definió el concepto de entropía de información.