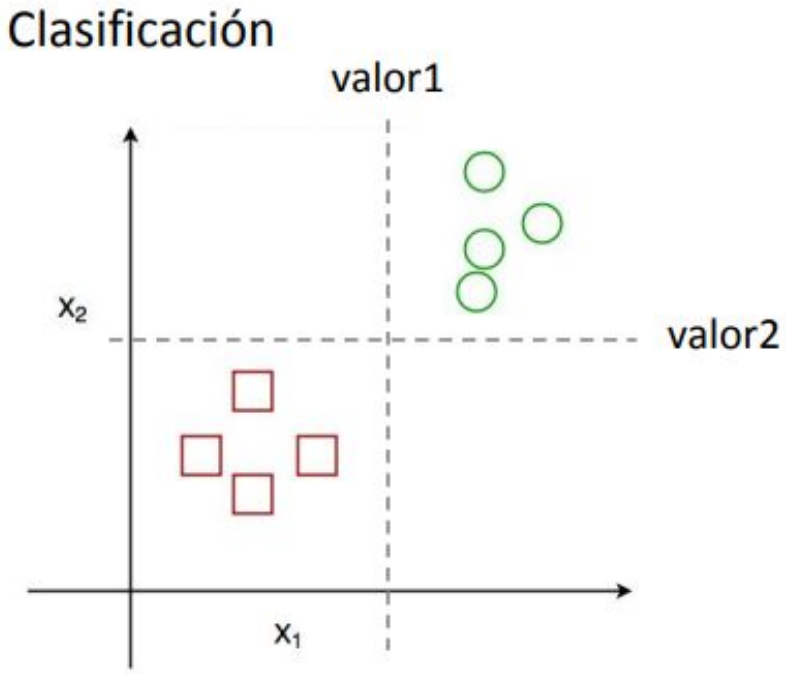
Aprendizaje automático

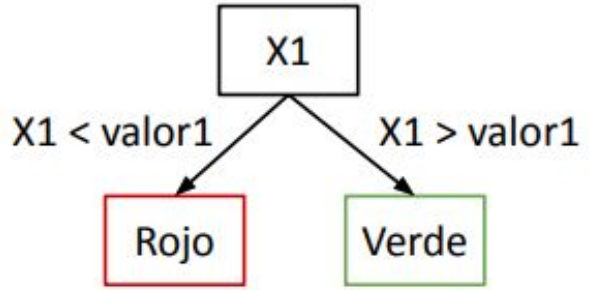
Clase 5

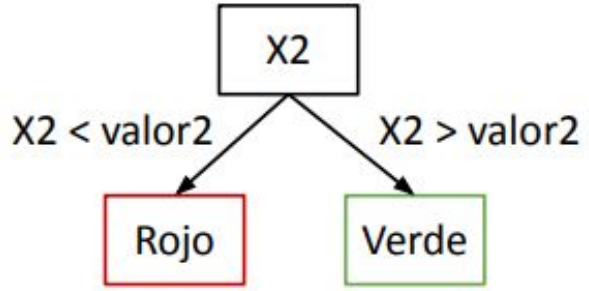
Martin Pustilnik, Iris Sattolo,
Maximiliano Beckel

UNAHUR

Árboles de Decisión

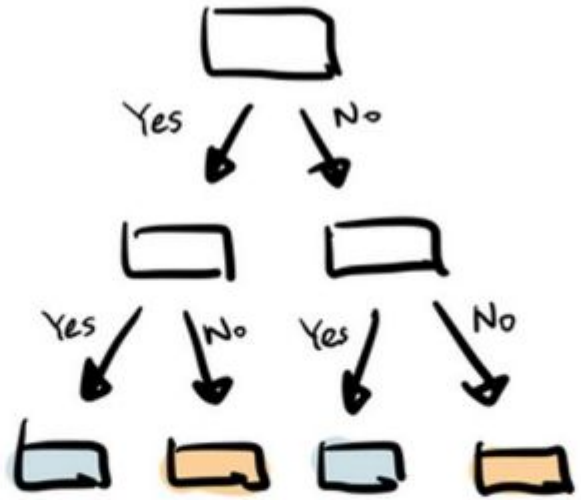






Reglas if-then sobre valores de atributos.
Predicen el objetivo en función de esas reglas.

Árboles de Decisión



Raíz: el nodo desde el cual inicia el árbol

Nodo: representa test sobre un atributo de la instancia

Rama desde un nodo: corresponde a un valor para ese atributo

Hojas: nodos que definen las clases de la decisión

Método de **inferencia inductiva** (busca aproximar una función objetivo).

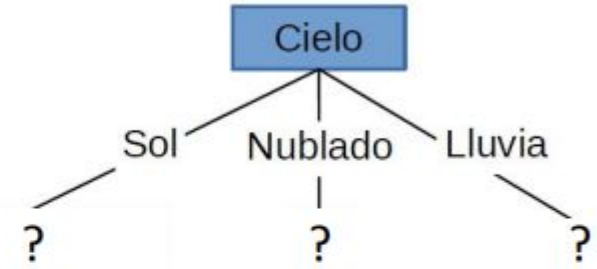
El árbol representa **disyunción de conjunciones** sobre valores de atributos (y/o).

Aprende **reglas if-then** que **reducen localmente el error** con algún criterio.

Árboles de Decisión

Cómo construyo un árbol?

	Atributos				Clase
Instancia	Cielo	Temperatura	Humedad	Viento	Va a correr?
1	sol	calor	alta	débil	No
2	sol	calor	alta	fuerte	No
3	nublado	calor	alta	débil	Sí
4	lluvia	templado	alta	débil	Sí
5	lluvia	frío	normal	débil	Sí
6	lluvia	frío	normal	fuerte	No
7	nublado	frío	normal	fuerte	Sí
8	sol	templado	alta	débil	No
9	sol	frío	normal	débil	Sí
10	lluvia	templado	normal	débil	Sí
11	sol	templado	normal	fuerte	Sí
12	nublado	templado	alta	fuerte	Sí
13	nublado	calor	normal	débil	Sí
14	lluvia	templado	alta	fuerte	No

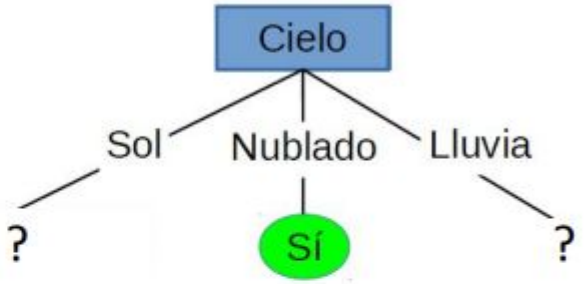


El atributo cielo, es un buen nodo para empezar el árbol, ya que cuando toma el valor “nublado” todas las instancias son de la clase “Sí”

Árboles de Decisión

Cómo construyo un árbol?

	Atributos				Clase
Instancia	Cielo	Temperatura	Humedad	Viento	Va a correr?
1	sol	calor	alta	débil	No
2	sol	calor	alta	fuerte	No
3	nublado	calor	alta	débil	Sí
4	lluvia	templado	alta	débil	Sí
5	lluvia	frío	normal	débil	Sí
6	lluvia	frío	normal	fuerte	No
7	nublado	frío	normal	fuerte	Sí
8	sol	templado	alta	débil	No
9	sol	frío	normal	débil	Sí
10	lluvia	templado	normal	débil	Sí
11	sol	templado	normal	fuerte	Sí
12	nublado	templado	alta	fuerte	Sí
13	nublado	calor	normal	débil	Sí
14	lluvia	templado	alta	fuerte	No

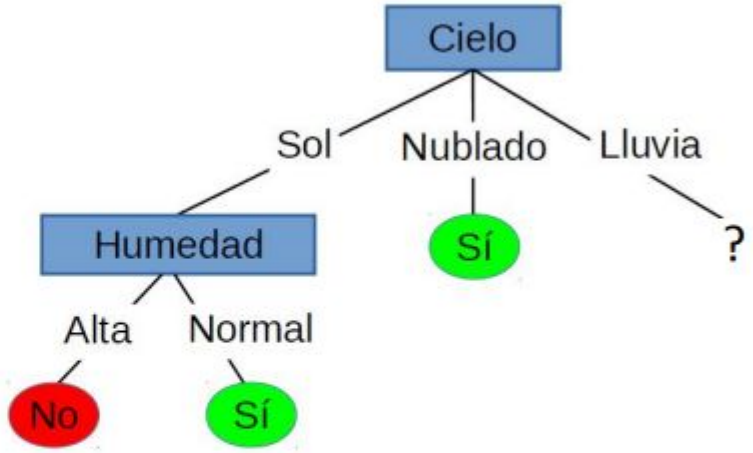


Cuando el valor que toma el atributo cielo es “sol”, algunas instancias son “Sí” y otras “No”. Tengo que buscar si existe algún atributo que me separe bien. **Spoiler**: humedad, que cuando es “alta” es “No” y si es “normal” es “Sí”

Árboles de Decisión

Cómo construyo un árbol?

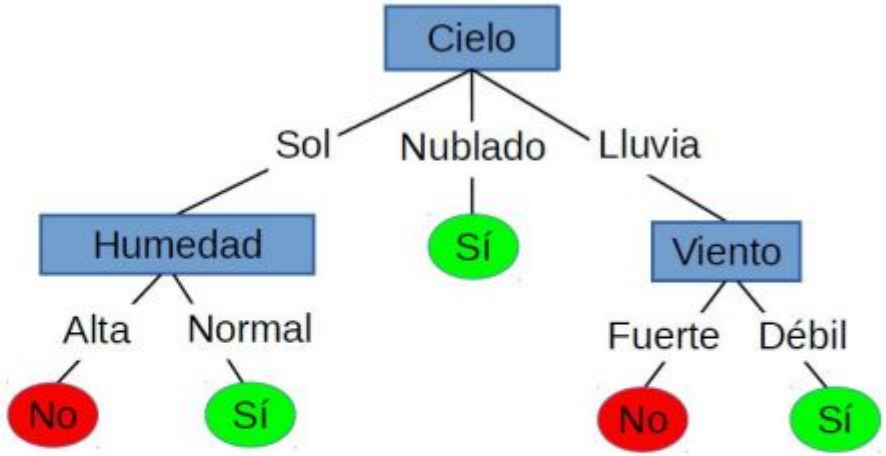
	Atributos				Clase
Instancia	Cielo	Temperatura	Humedad	Viento	Va a correr?
1	sol	calor	alta	débil	No
2	sol	calor	alta	fuerte	No
3	nublado	calor	alta	débil	Sí
4	lluvia	templado	alta	débil	Sí
5	lluvia	frío	normal	débil	Sí
6	lluvia	frío	normal	fuerte	No
7	nublado	frío	normal	fuerte	Sí
8	sol	templado	alta	débil	No
9	sol	frío	normal	débil	Sí
10	lluvia	templado	normal	débil	Sí
11	sol	templado	normal	fuerte	Sí
12	nublado	templado	alta	fuerte	Sí
13	nublado	calor	normal	débil	Sí
14	lluvia	templado	alta	fuerte	No



Busco otro atributo que me separe bien las instancias de cielo “lluvia”. **Spoiler**: es viento, que cuando es “fuerte” es “No” y cuando es “débil” es “Sí”

Árboles de Decisión

	Atributos				Clase
Instancia	Cielo	Temperatura	Humedad	Viento	Va a correr?
1	sol	calor	alta	débil	No
2	sol	calor	alta	fuerte	No
3	nublado	calor	alta	débil	Sí
4	lluvia	templado	alta	débil	Sí
5	lluvia	frío	normal	débil	Sí
6	lluvia	frío	normal	fuerte	No
7	nublado	frío	normal	fuerte	Sí
8	sol	templado	alta	débil	No
9	sol	frío	normal	débil	Sí
10	lluvia	templado	normal	débil	Sí
11	sol	templado	normal	fuerte	Sí
12	nublado	templado	alta	fuerte	Sí
13	nublado	calor	normal	débil	Sí
14	lluvia	templado	alta	fuerte	No



Cada **nodo** interno evalúa un atributo discreto Xi. Cada **rama** corresponde a un valor para ese atributo Xi. Cada **hoja** predice un valor de Y

Árboles de Decisión

Medidas de impureza

Proporción de los datos que están en la hoja m y pertenecen a la clase k.

Medidas de impureza dentro de cada hoja:

Coeficiente de Gini: $$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Entropía: $$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

Si todos los datos dentro de una hoja pertenecen a la misma clase, G = D = 0: la hoja tiene impureza 0.

Se define la impureza de un árbol por el **promedio pesado de las impurezas de cada hoja**, pesado por la fracción de datos en cada hoja.

Árboles de Decisión

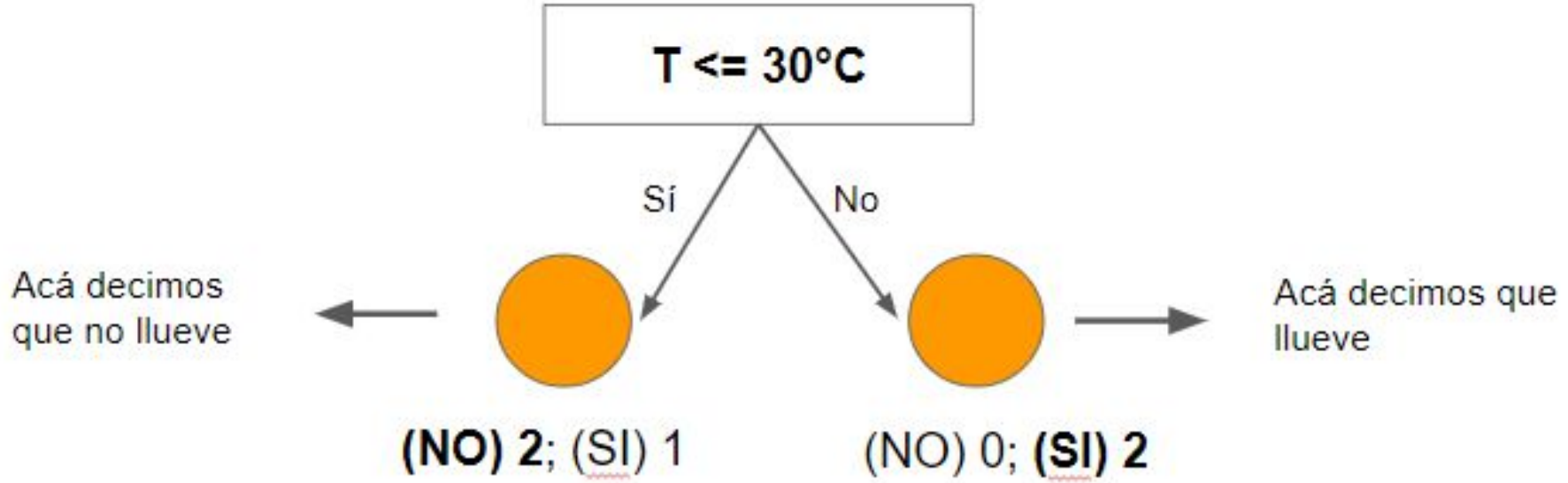
Problemas de clasificación. Algoritmo

- Buscamos el feature y la condición que minimice la impureza del árbol y lo fijamos como raíz.
- Para cada uno de los dos nodos que se desprenden de la raíz buscamos el feature y la condición que me disminuya la impureza en ese subconjunto.
- Así siguiendo hasta que cada dato quede dentro de una hoja pura, o bien hasta que se cumpla algún criterio de convergencia (por ejemplo, hacer crecer el árbol hasta cierta profundidad).

Árboles de Decisión

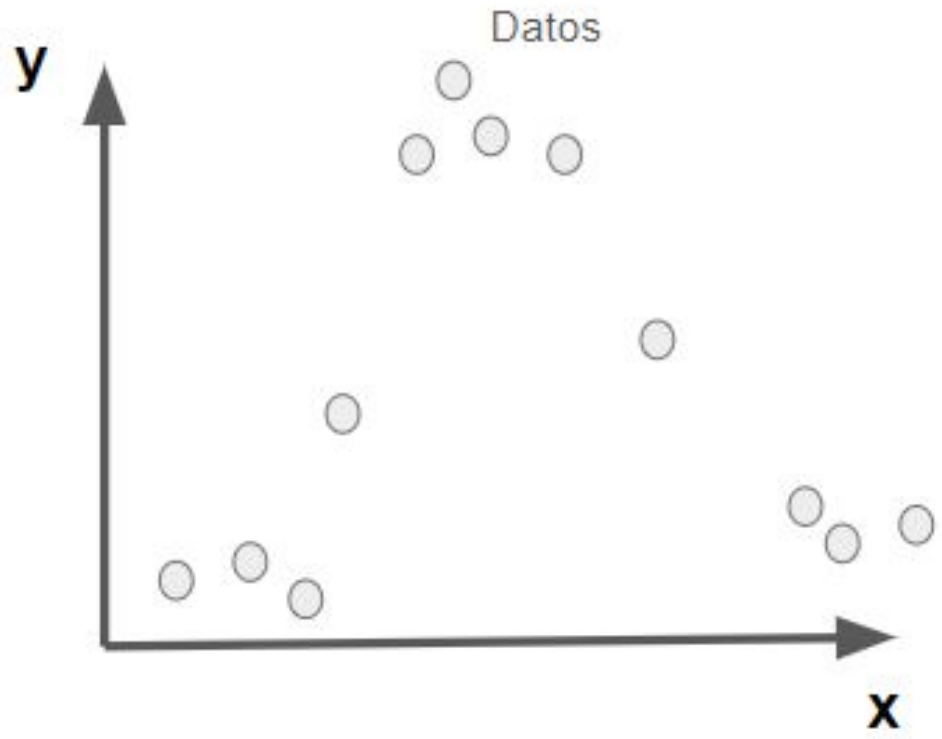
Problemas de clasificación. ¿Cómo predecimos?

La categoría más frecuente dentro de cada hoja (también podríamos dar la probabilidad de que sea de una dada clase en base a la fracción de instancias de cada clase que caigan dentro de cada hoja).



Árboles de Decisión

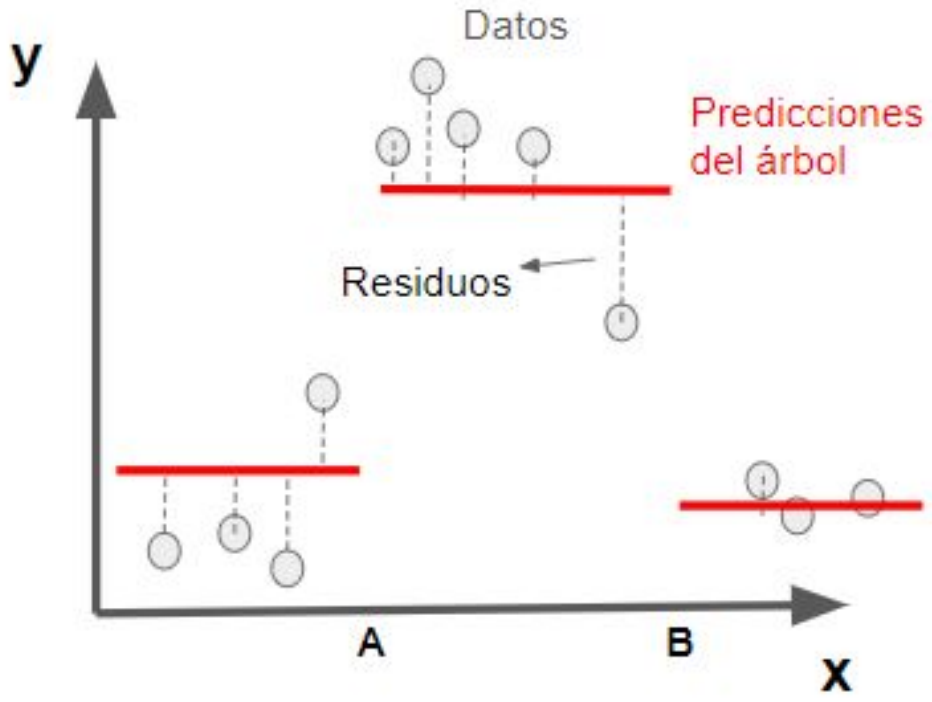
Problemas de regresión



Tenemos una variable numérica y queremos predecir cómo se comporta en base a un conjunto de features. ¿Qué feature y condición elijo?

Árboles de Decisión

Problemas de regresión



¿Cómo sabemos dónde cortar? (En el problema, cómo elegimos A y B?)

Buscamos los cortes que minimicen la suma del cuadrado de los residuos:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

Suma sobre todas las hojas

Predicción para cada hoja (promedio de las instancias dentro).

Árboles de Decisión

Problemas de regresión. Algoritmo

- Buscamos el feature y la condición que minimice la suma del cuadrado de los residuos.
- Para cada uno de los dos nodos que se desprenden de la raíz buscamos el feature y la condición que me disminuya la suma de los cuadrados de los residuos en ese subconjunto.
- Así siguiendo hasta que cada dato quede dentro de una hoja pura, o bien hasta que se cumpla algún criterio de convergencia (por ejemplo, hacer crecer el árbol hasta cierta profundidad).
- Damos como **predicción el promedio de las valores** dentro de cada hoja.

Árboles de Decisión

Ventajas de los árboles de decisión

- Fáciles de interpretar: se asemeja bastante a la forma en la que enfrentamos un problema, más que nada de clasificación.
- No hay que preocuparse por diferencias de escala en datos numéricos
- No hay que hacer one-hot-encoding de features categóricas
- Manejan datos faltantes de una forma natural
- Permite incluir todo tipo de variable: categórica, ordinal, numérica.
- Puede usarse para problemas multiclase y regresión

Árboles de Decisión

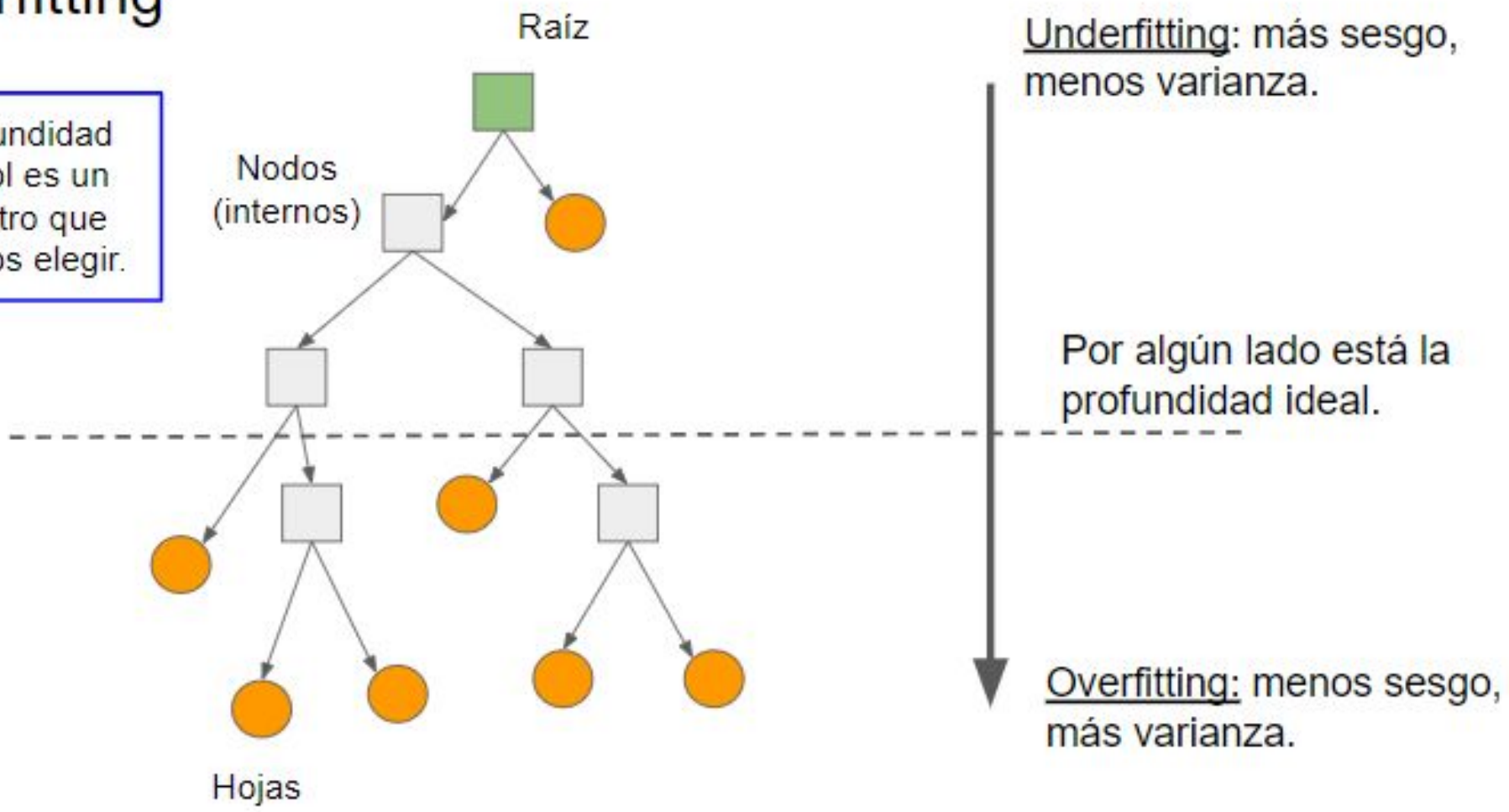
Desventajas de los árboles de decisión

- Muchas veces no son buenos modelos, tienen baja performance.
- Árboles de mucha profundidad tienden a hacer overfitting (puedo irme tan profundo hasta que cada dato esté en hojas puras o con error igual a 0).
- Baja performance si justo arrancamos con un feature muy ruidoso
- Sesgos hacia clases más dominantes (datasets no balanceados)

Árboles de Decisión

Overfitting

La profundidad del árbol es un parámetro que podemos elegir.



Árboles de Decisión

Overfitting. Algunas ideas para evitarlo

- Fijar la profundidad del árbol.
- Fijar la cantidad de hojas (para armar una cantidad fija de grupos de datos).
- Fijar la mínima cantidad de datos que están contenidos dentro de cada hoja (para hacer, por ejemplo, promedios más robustos).
- **Regularización** = *cost complexity pruning*. Penaliza árboles con muchas hojas al buscar minimizar la siguiente función:

$$\sum_{m=1}^{|T|} \sum_{i:\ x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

Constante de regularización

Suma de los residuos al cuadrado

Número de hojas

Árboles de Decisión

Conjunto (ensamble) de clasificadores. Idea

- Entrenar varios modelos distintos que sobre-ajusten (bajo sesgo y mucha varianza). Cada uno de ellos me dan un resultado.

- Promediar varios modelos **reduce la varianza**:
  - Si el problema es de regresión, el resultado final es simplemente el promedio.
  - Si el problema es de clasificación, puedo elegir la clase más frecuente entre todos los modelos (votación).
  - Si el modelo devuelve probabilidades, puedo hacer una votación ponderada.

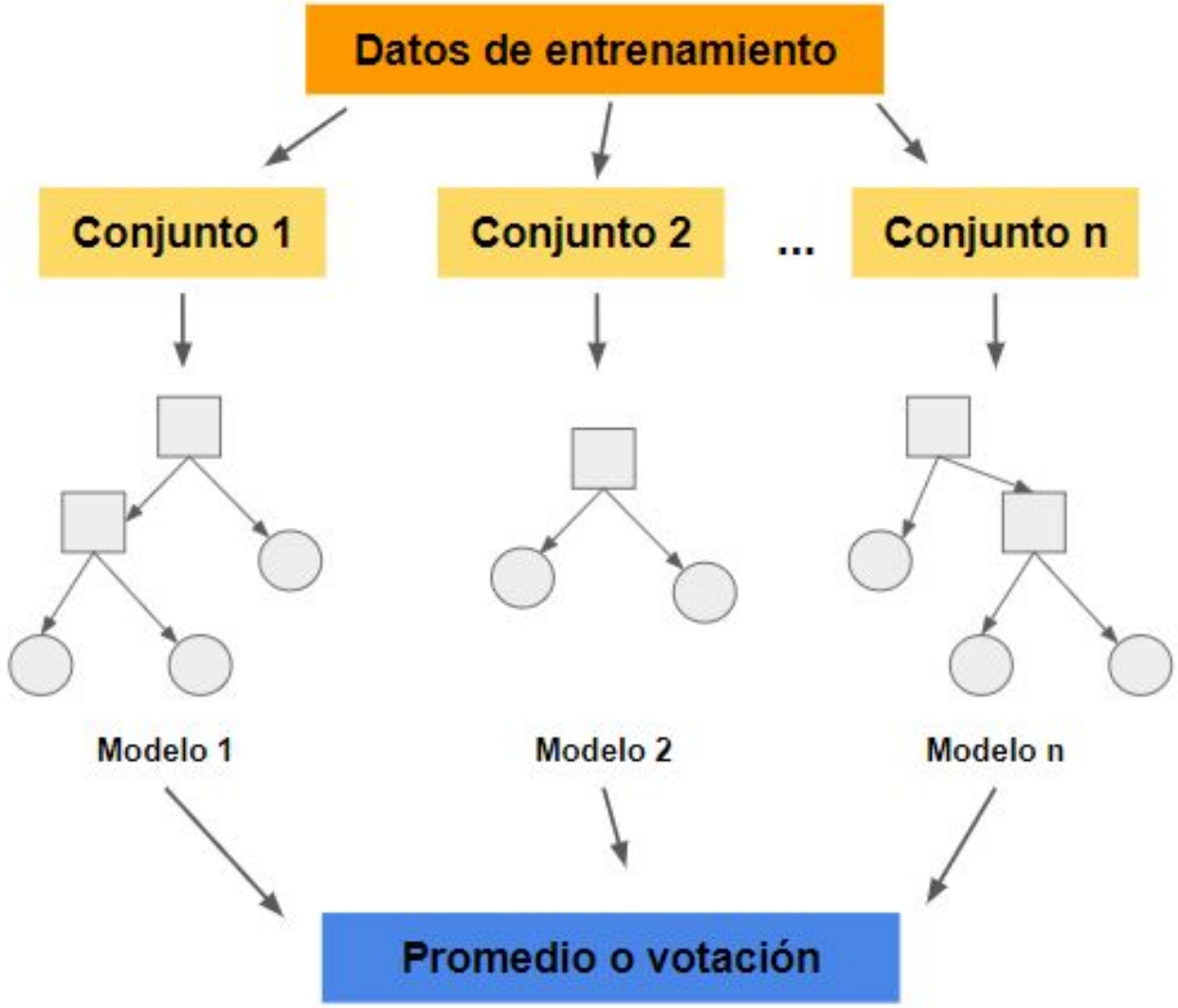
UNAHUR

Árboles de Decisión

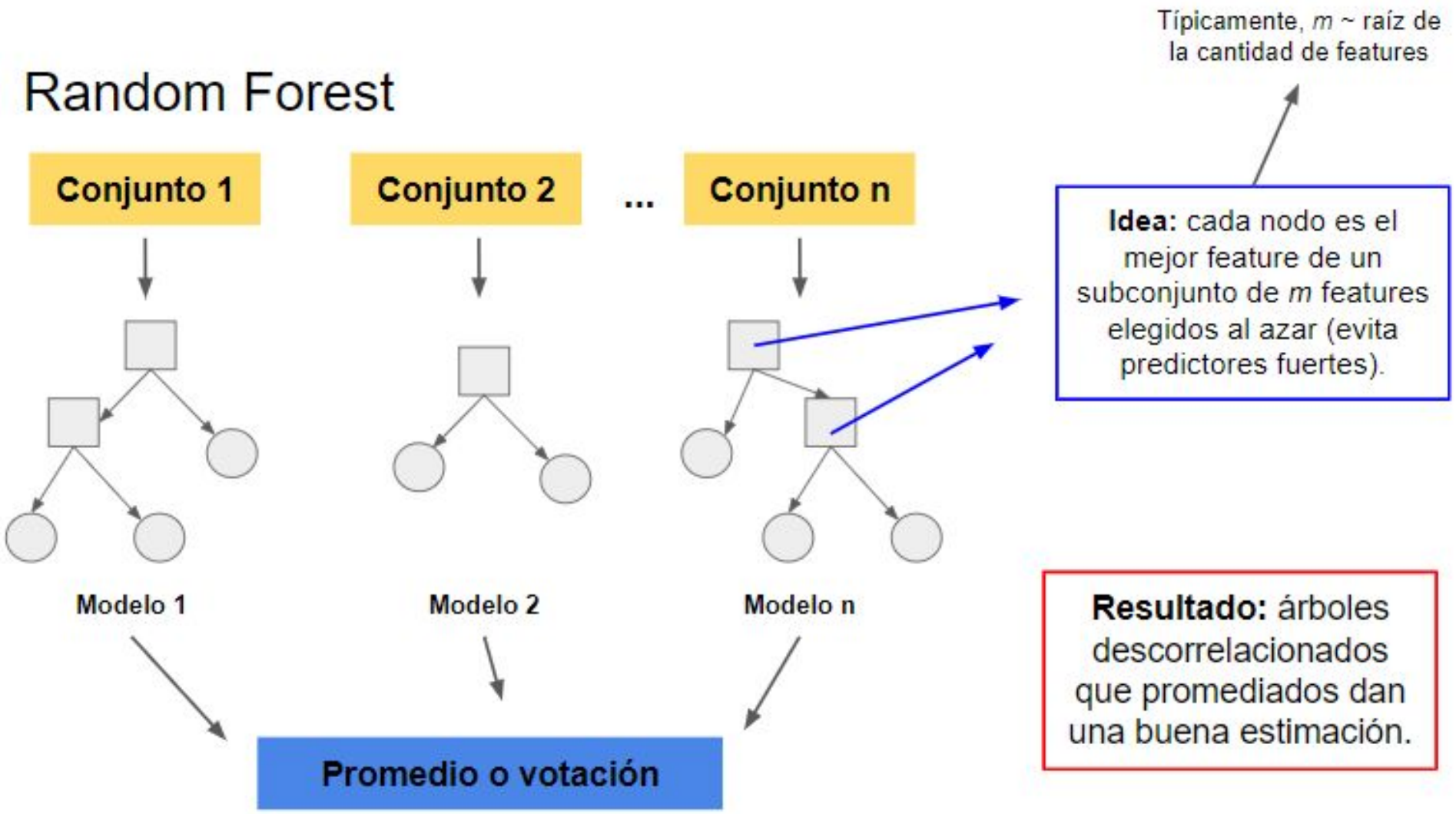
Bagging
(Bootstrap Aggregating)

Idea: entrenar varios árboles sobre conjuntos de datos tomados con **muestras con reemplazo** (bootstraping) de los datos de entrenamiento.

Problema: si hay una variable muy predictora, los árboles van a ser muy parecidos entre sí (estarían muy correlacionados).



Árboles de Decisión



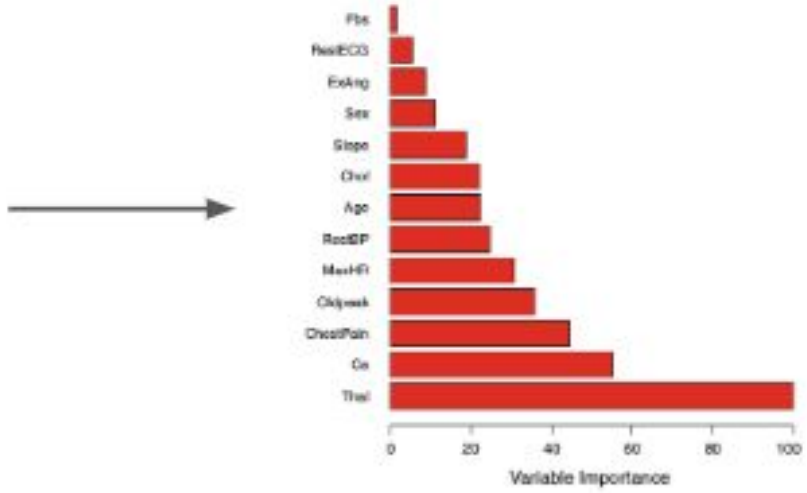
Árboles de Decisión

Bagging y Random Forest. Algunas características

Ventajas: la combinación de diferentes modelos es siempre mucho mejor que un único modelo, podemos esperar una performance mucho más alta.

Desventajas: perdemos interpretabilidad de cómo el ensamble llega al resultado final.

Feature importance: podemos medir en promedio qué tanto una variable reduce el error o la impureza. Esto nos da una idea de qué variable es informativa y cuál no.



Árboles de Decisión

Resumen: regresión y clasificación con árboles

- La idea es encontrar condiciones que separen los datos en grupos donde: haya algunas clases dominantes (clasificación) o el error respecto del promedio sea bajo (problemas de regresión).
- Los árboles de decisión pueden crecer tanto a punto de overfittear, por lo tanto es bueno tener en cuenta todas las técnicas para prevenir esto.
- Mejor que un único árbol es un bosque! Random Forest es un algoritmo mucho más poderoso que los árboles de decisión. El problema es que perdemos interpretabilidad.

UNAHUR