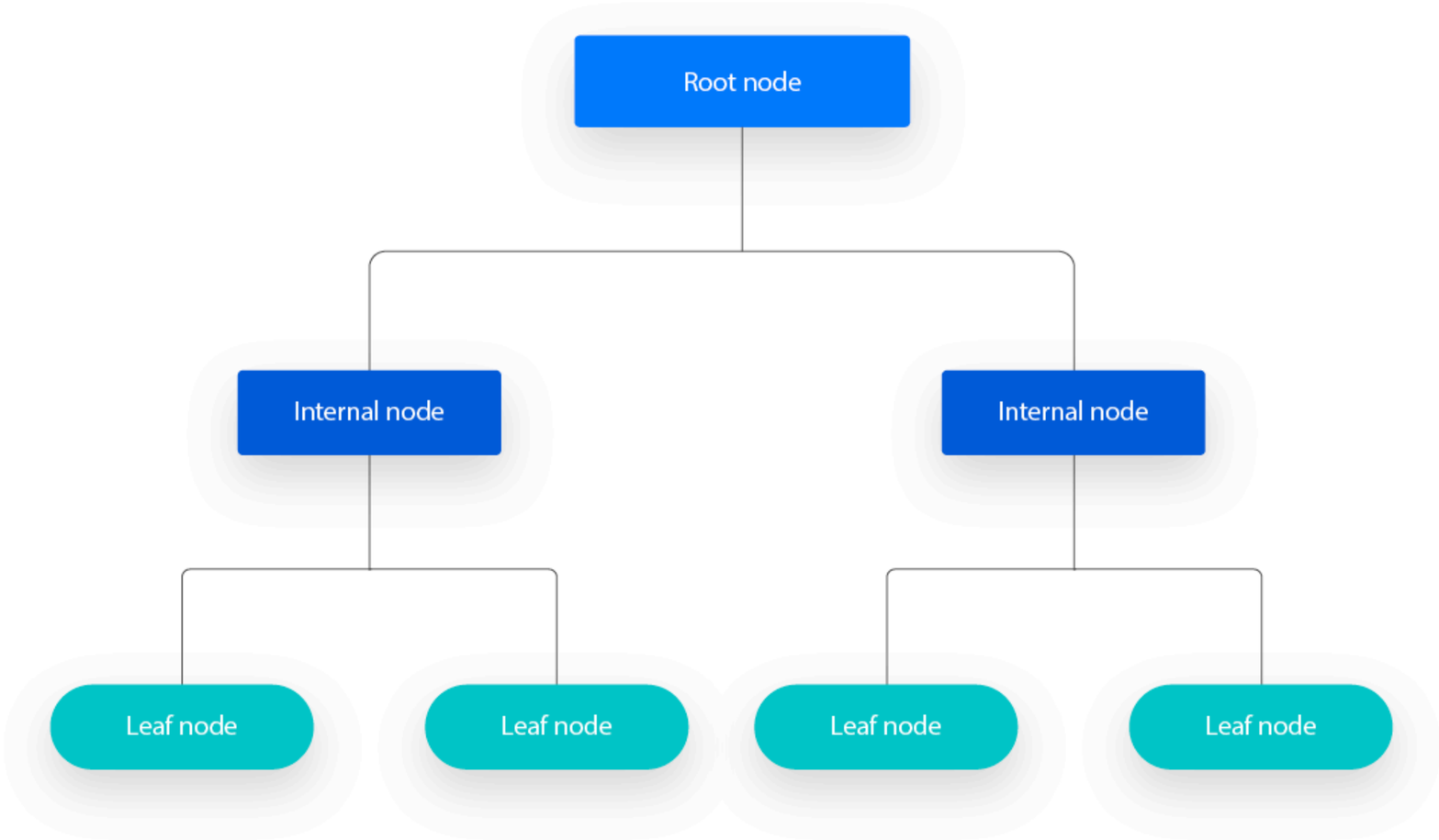
Árboles de Decisión

Licenciatura en Inteligencia Artificial y Ciencia de Datos, CUGDL, Universidad de Guadalajara.

Guadalajara, Jal., agosto de 2025

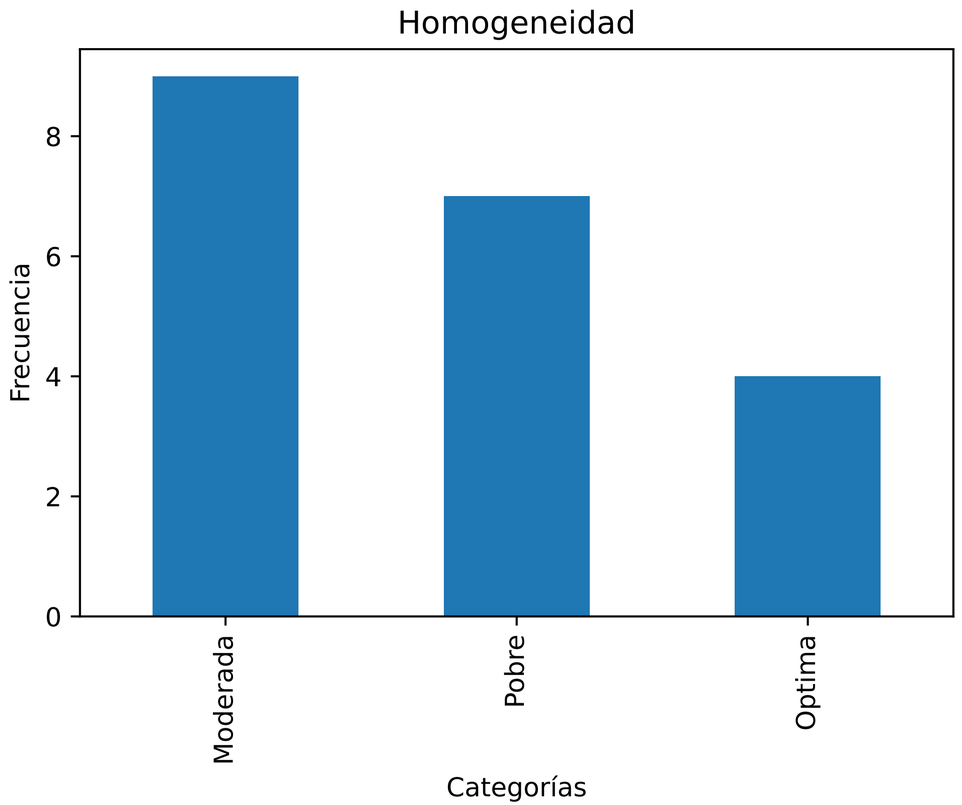
Árbol de decisión

Estructura

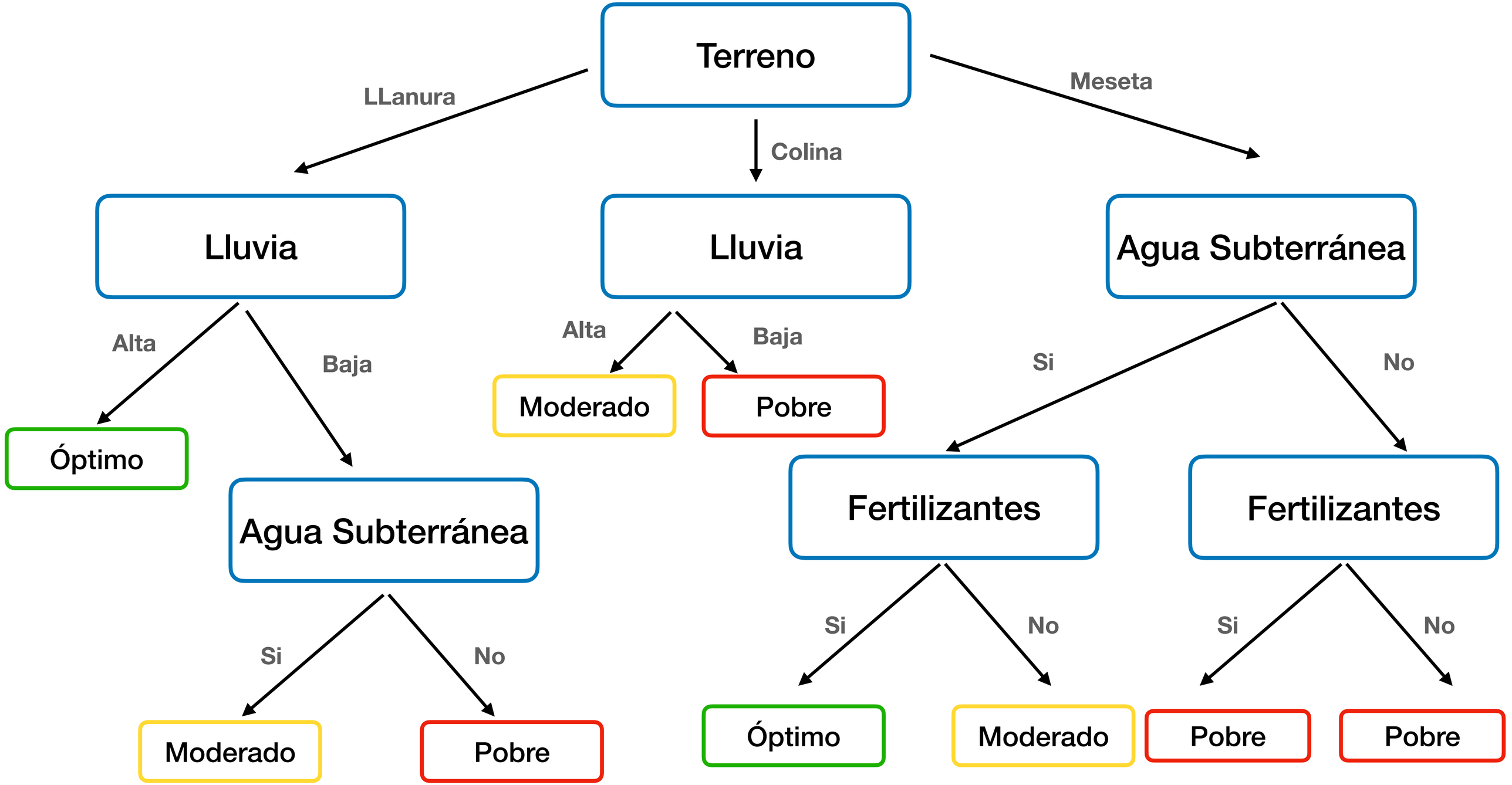


Árbol de decisión

- Estas son las predicciones que queremos llevar a cabo.
- Buscamos parámetros que nos dé una distribución lo más homogénea posible de la variable a predecir.



	Lugar	Lluvia	Terreno	Fertilizantes	Agua_subterranea	Cosecha
0	P1	Altas	Llanura	Si	Si	Optima
1	P2	Bajas	Colinas	No	Si	Pobre
2	P3	Bajas	Meseta	No	Si	Moderada
3	P4	Altas	Meseta	No	Si	Moderada
4	P5	Altas	Llanura	Si	No	Optima
5	P6	Bajas	Colinas	No	No	Pobre
6	P7	Bajas	Meseta	No	No	Pobre
7	P8	Medias	Meseta	No	No	Pobre
8	P9	Altas	Colinas	Si	Si	Moderada
9	P10	Medias	Meseta	Si	Si	Optima
10	P11	Altas	Meseta	Si	No	Optima
11	P12	Medias	Meseta	Si	No	Moderada
12	P13	Altas	Colinas	Si	No	Moderada
13	P14	Bajas	Llanura	Si	Si	Moderada
14	P15	Medias	Llanura	Si	No	Moderada
15	P16	Bajas	Llanura	No	No	Pobre
16	P17	Bajas	Colinas	Si	No	Pobre
17	P18	Medias	Meseta	No	No	Pobre
18	P19	Altas	Llanura	No	Si	Moderada
19	P20	Medias	Colinas	Si	Si	Moderada


Terreno
LLanura
Colina
Meseta
Lluvia
Lluvia
Agua Subterránea
Alta
Baja
Alta
Baja
Si
No
Óptimo
Moderado
Pobre
Agua Subterránea
Fertilizantes
Fertilizantes
Si
No
Si
No
Si
No
Moderado
Pobre
Óptimo
Moderado
Pobre
Pobre

Homogeneidad

Cosecha Lluvia	Moderada	Optima	Pobre
Altas	4	3	0
Bajas	2	0	5
Medias	3	1	2

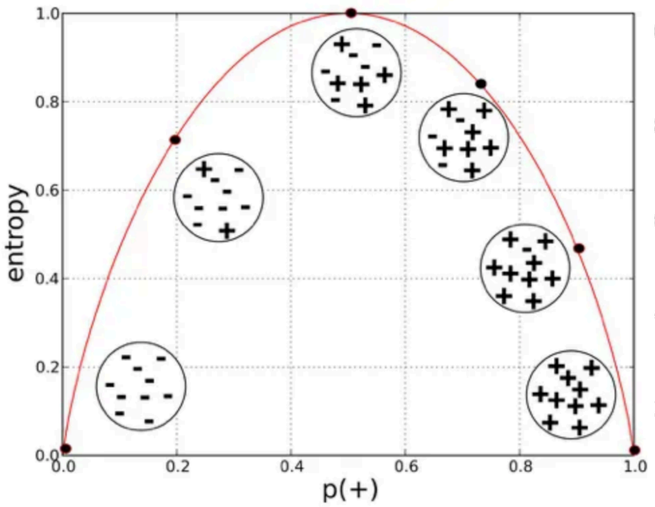
Cosecha Terreno	Moderada	Optima	Pobre
Colinas	3	0	3
Llanura	3	2	1
Meseta	3	2	3

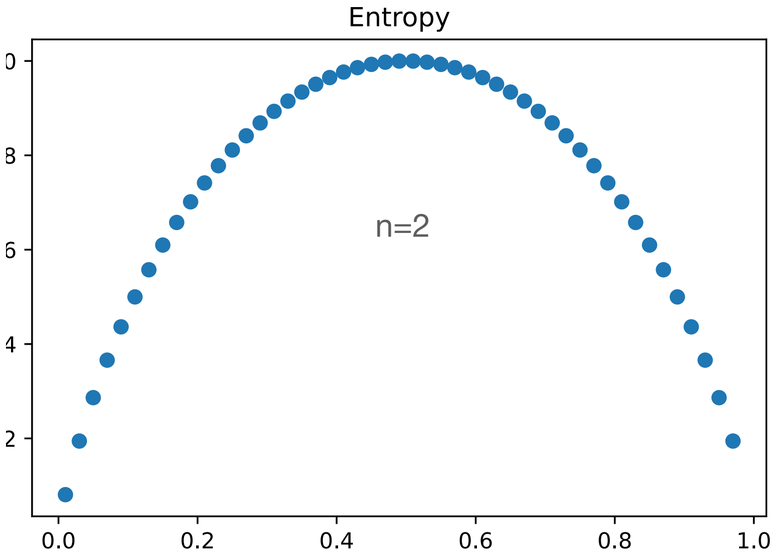
Cosecha Fertilizantes	Moderada	Optima	Pobre
No	3	0	6
Si	6	4	1

Cosecha Agua_subterranea	Moderada	Optima	Pobre
No	3	2	6
Si	6	2	1

Entropía

- $$H(S) = \sum_{i=1}^{n} - p_i \log_2(p_i)$$

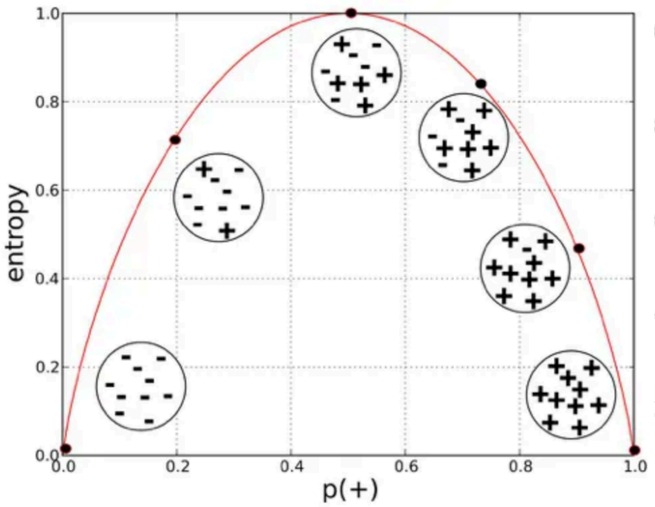


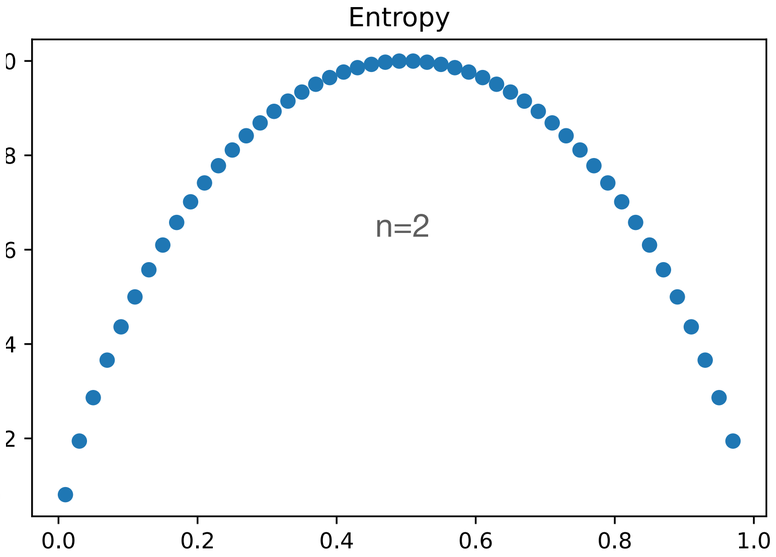


- Donde *n* son las categorías para una variable objetivo a predecir del dataset.
- Cuánto más homogéneos son los nodos de un árbol, se requiere menos información para representarlos.
- La heterogeneidad de un nodo se puede representar con la entropía.
- "Número mínimo de bits para codificar una determinada clasificación".

Entropía

- $$H(S) = \sum_{i=1}^{n} - p_i \log_2(p_i)$$



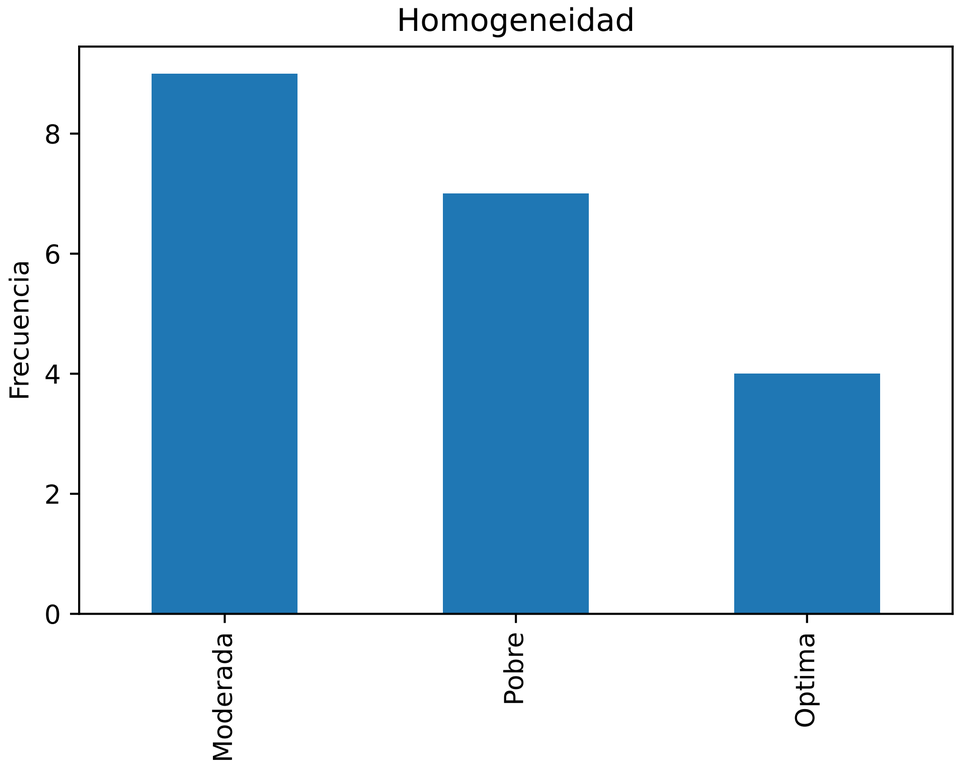


- Una reducción de la entropía se mide como una ganancia de información.
- En árboles de decisión queremos que los nodos tengan mayor ganancia de información.
- La suma se realiza sobre las categorías target.
- La entropía sería cero cuando todas las observaciones son perfectamente homogéneas mientras que tomaría un valor máximo cuando fueran heterogéneas (cuando todo es igual de probable).

Entropía del sistema

Variable target

$$H(S) = -\frac{4}{20} \cdot \log_2(\frac{4}{20}) - \frac{9}{20} \cdot \log_2(\frac{9}{20}) - \frac{7}{20} \cdot \log_2(\frac{7}{20}) = 1.5$$



$$p_{\text{optimo}} = \frac{4}{20}$$

$$p_{\text{moderado}} = \frac{9}{20}$$

$$p_{\text{pobre}} = \frac{7}{20}$$

Ganancia de información

Variable target

$$\Delta H(S, V) = H(S) - \sum_{c \in V} \frac{|V = c|}{V} H(V = c)$$

$$\text{Terreno} = \{\text{llanura}, \text{meseta}, \text{colina}\}$$

Cosecha / Terreno	Moderada	Optima	Pobre
Colinas	3	0	3
Llanura	3	2	1
Meseta	3	2	3

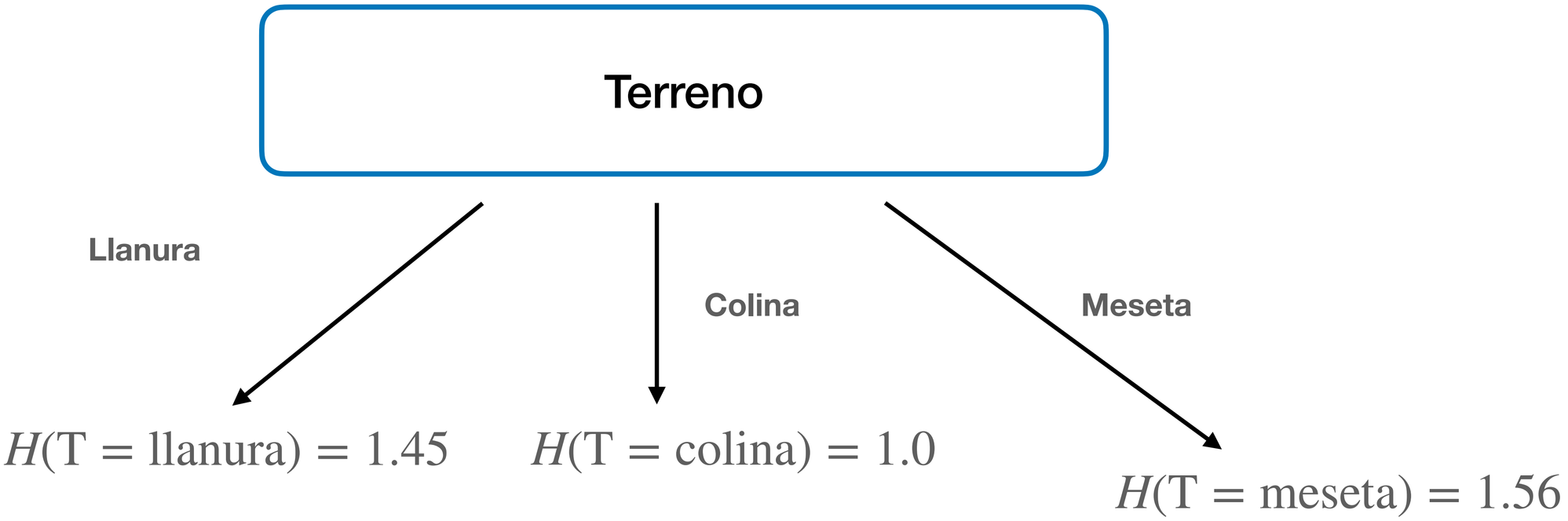
$$H(\text{T} = \text{colina}) = -\left(\frac{0}{6} \cdot \log_2(\frac{0}{6}) + \frac{3}{6} \cdot \log_2(\frac{3}{6}) + \frac{3}{6} \cdot \log_2(\frac{3}{6})\right) = 1.0$$

$$H(\text{T} = \text{llanura}) = -\left(\frac{2}{6} \cdot \log_2(\frac{2}{6}) + \frac{1}{6} \cdot \log_2(\frac{1}{6}) + \frac{3}{6} \cdot \log_2(\frac{3}{6})\right) = 1.45$$

$$H(\text{T} = \text{meseta}) = -\left(\frac{2}{8} \cdot \log_2(\frac{2}{8}) + \frac{3}{8} \cdot \log_2(\frac{3}{8}) + \frac{3}{8} \cdot \log_2(\frac{3}{8})\right) = 1.56$$

Ganancia de información

Variable target



$$\Delta H(S,T) = 1.5 - \left(\frac{6}{20} H(\mathrm{T} = \mathrm{llanura}) + \frac{6}{20} \mathrm{H}(\mathrm{T} = \mathrm{colina}) + \frac{8}{20} \mathrm{H}(\mathrm{T} = \mathrm{meseta}) \right) = 0.14$$

Ganancia de información

Variable target

```
In [101]:  1  1.5 - entropia(table_terreno)

H(T= Colinas ) =  1.0
H(T= Llanura ) =  1.4591479170272448
H(T= Meseta ) =  1.561278124459133
total = 1.3622556248918265

Out[101]: 0.1377443751081735
```

```
In [103]:  1  1.5 - entropia(table_agua)

H(T= No ) =  1.4353713907745331
H(T= Si ) =  1.224394445405986
total = 1.340431765358687

Out[103]: 0.15956823464131298
```

```
In [104]:  1  1.5 - entropia(table_fertilizantes)

H(T= No ) =  0.9182958340544896
H(T= Si ) =  1.3221793455166666
total = 1.140431765358687

Out[104]: 0.35956823464131293
```

```
In [105]:  1  1.5 - entropia(table_lluvia)

H(T= Altas ) =  0.9852281360342515
H(T= Bajas ) =  0.863120568566631
H(T= Medias ) =  1.4591479170272446
total = 1.0846664217184823

Out[105]: 0.41533357828151773
```

- Esto se tiene que hacer para todas las variables restantes.
- La variable que muestre una máxima ganancia de información es la que se elige como nodo.

Árboles de decisión

Método (ID3)

- Se calcula la **entropía** inicial del sistema (basada en la variable de objetivo).
- Se calcula la ganancia de información para cada variable adicional. Se selecciona la variable que regresa la **máxima ganancia de información** y se toma como nuevo nodo de decisión.
- Se repite el paso 2 para cada una de las ramas de los nuevos nodos. El nuevo nodo es identificado como una **hoja**.
- Se comprueba si estos nodos hoja clasifican correctamente la información y si es así, se detiene esta ramificación. Si no es así, se vuelve al paso 2 y se realizan más ramificaciones.

Árboles de regresión

Predicción $\hat{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

(x_1, y_1)

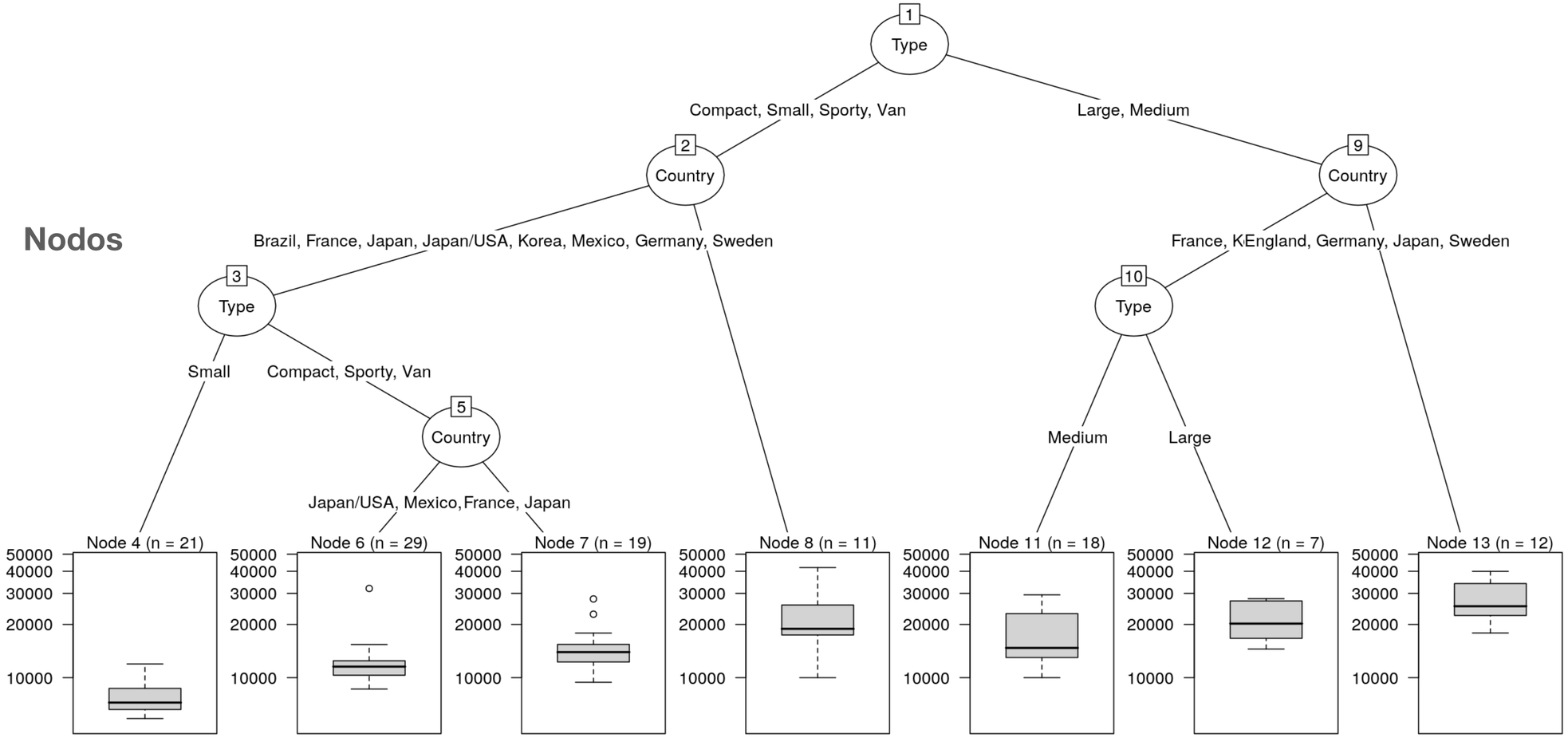
(x_2, y_2)

(x_3, y_3)

$\dots$

(x_n, y_n)

Nodos



Árboles de regresión

Definición

- Los árboles de regresión son un tipo de árboles de decisión en la cual la variable objetivo es de tipo continua en lugar de categóricas.
- Son útiles cuando existen múltiples rasgos en el dataset que interactúan entre ellos, que no sea lineal o logístico.
- Se empieza con un nodo raiz que se va dividiendo en subnodos en base a algún criterio de división. Generalmente se utiliza la reducción máxima de la varianza.

Árboles de regresión

Reducción de la varianza (ejemplo)

Variable	V1	V2	Total
1	2	6	8
0	8	4	12
Total	10	10	20

Por cada variable, en cada nodo tendríamos que,

$$\mu(\text{nodo}) = \frac{8 \cdot 1 + 12 \cdot 0}{20} = 0.4$$

$$\text{var}(\text{nodo}) = \frac{8 \cdot (1 - 0.4)^2 + 12 \cdot (0 - 0.4)^2}{20} = 0.24$$

$$\mu(\text{V1}) = \frac{2 \cdot 1 + 8 \cdot 0}{10} = 0.2$$

$$\text{var}(\text{V1}) = \frac{2 \cdot (1 - 0.2)^2 + 8 \cdot (0 - 0.2)^2}{10} = 0.16$$

$$\mu(\text{V2}) = \frac{6 \cdot 1 + 4 \cdot 0}{10} = 0.6$$

$$\text{var}(\text{V2}) = \frac{6 \cdot (1 - 0.6)^2 + 4 \cdot (0 - 0.6)^2}{10} = 0.24$$

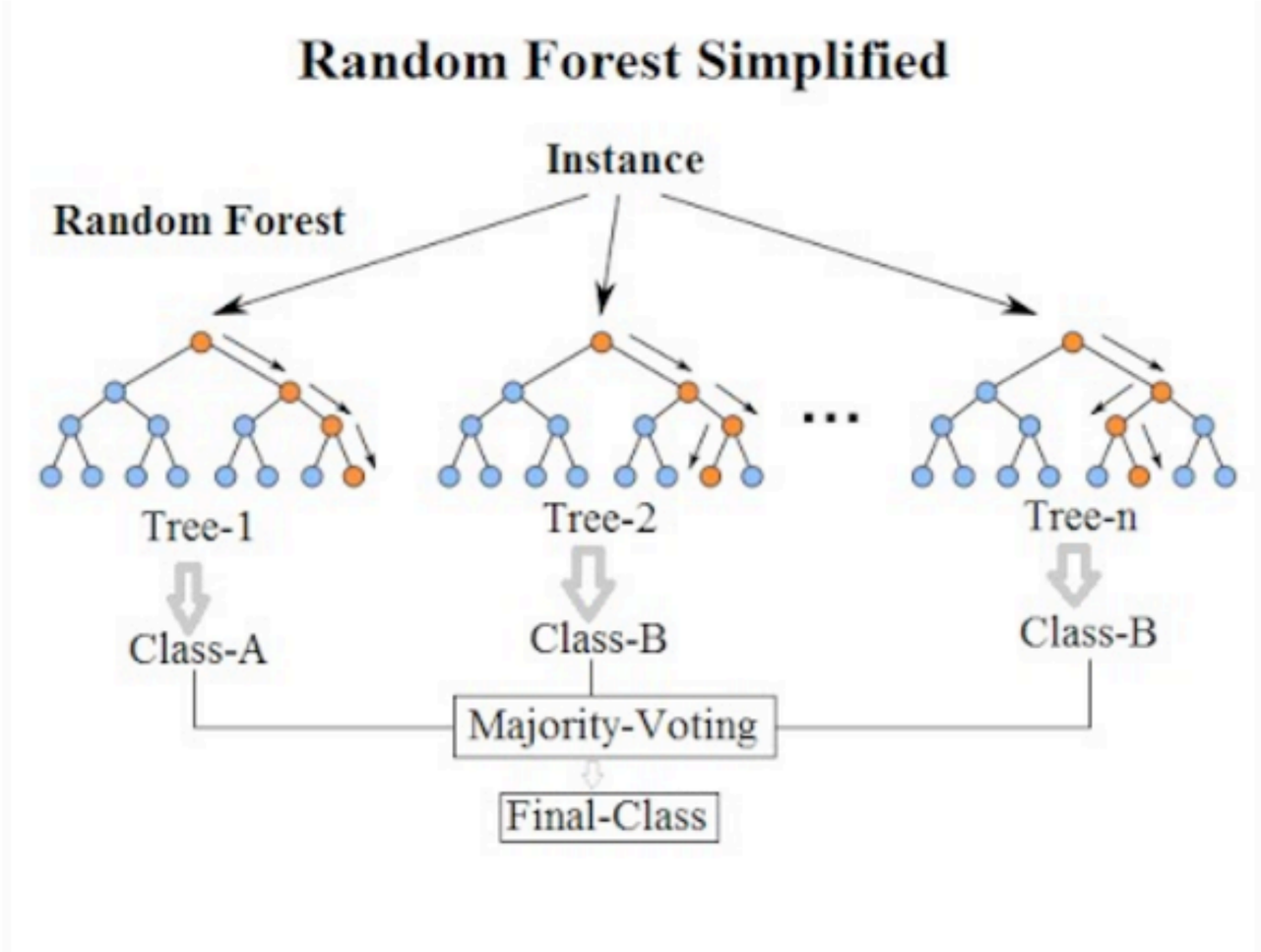
$$\text{Varianza ponderada} = \frac{10}{20} \cdot 0.16 + \frac{10}{20} \cdot 0.24 = 0.20$$

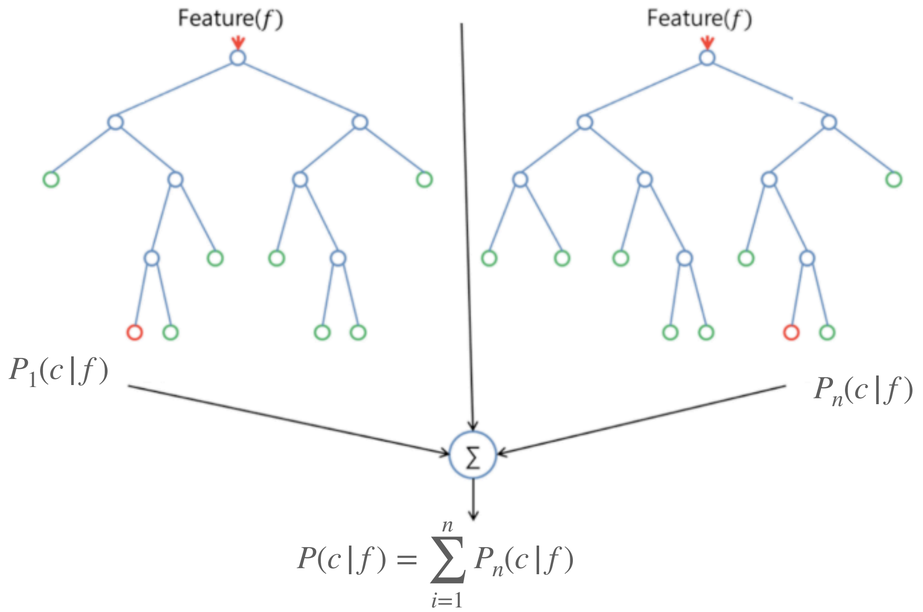
Árboles de regresión

Método

- Se empieza con un nodo que contiene todas las observaciones.
- Se calcula la **media** y **varianza** de la variable target.
- Se calcula la reducción de la varianza para todos los posibles candidatos a variable del siguiente nodo. Se elige la variable que da la r**educción máxima de la varianza** en el nodo.
- Para cada nodo se comprueba si la reducción máxima de la varianza o el número de observaciones dentro de éste cruzan un **umbral**.

Random Forest





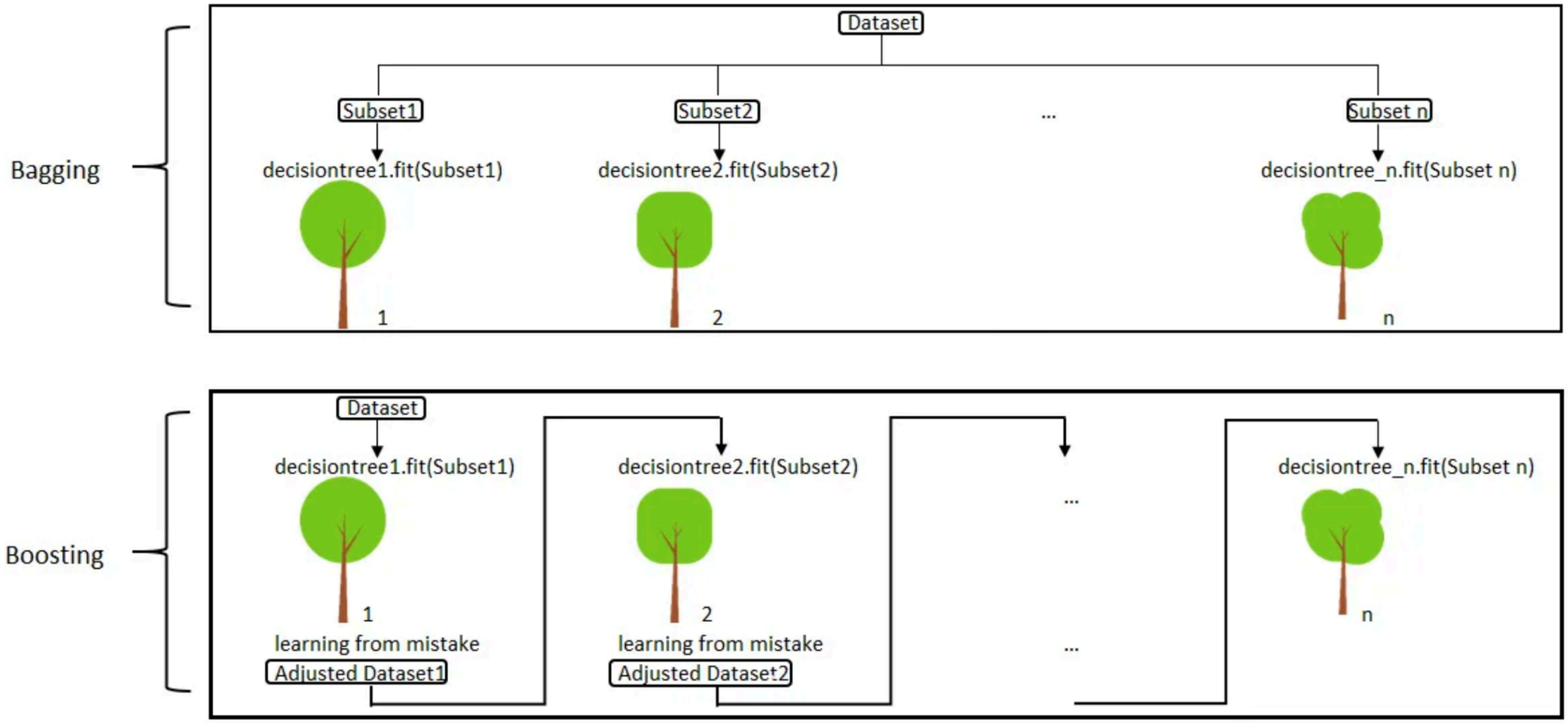
Random Forest

Definición

- Combina el resultado de múltiples árboles de decisión para llegar a un solo resultado.
- Se utiliza tanto para clasificación como regresión.
- Se hacen crecer árboles donde el resultado final sería basado en los siguientes criterios:
- **Métodos de promedio**: se crean varios modelos similares e independientes y se hace un promedio de las predicciones de cada modelo. En el caso de clasificación sería la variable que se repita más veces.
- **Método de boosting**: se entrenan un número de modelos de manera secuencia. Cada modelo "aprende" de los errores del modelo previo.

Árboles de regresión

Métodos



Random Forest

Bagging

- Con n observaciones en el dataset de entrenamiento, con m variables de entrada, se decide hacer crecer S árboles. Cada uno será creado con un dataset de entrenamiento separado. (Por tanto esto no requiere validación cruzada).

- Las n observaciones para cada dataset se toman aleatoriamente con reemplazo del dataset original.

- Cada dataset puede tener observaciones duplicadas y algunas pueden no aparecer nunca como entrenamiento.

Random Forest

Algoritmo

- Se toma una muestra aleatoria simple de tamaño n (con reemplazo).
- Se toma una muestra aleatoria simple de variables predictoras (sin reemplazo).
- Se construye un árbol de regresión con los predictores elegidos (sin podar el árbol).
- Se clasifican las observaciones fuera de la muestra con dicho árbol y se almacena el valor o la clase asignada para cada una.
- Se repiten los pasos 1-4 varias veces para tener un bosque de árboles.
- La predicción final es el promedio de las observaciones de todos los árboles, o en el caso de clasificación, la mayoría de clase en el conjunto de árboles.

Random Forest

Votación