VNIVERSITAT DE VALÈNCIA

# MASTER DE INGENIERÍA BIOMÉDICA.

# Métodos de ayuda al diagnóstico clínico.

## Tema 6: Árboles de decisión.



Emilio Soria, Antonio José Serrano y José David Martín Dpto Ingeniería Electrónica, ETSE

# Objetivos del tema

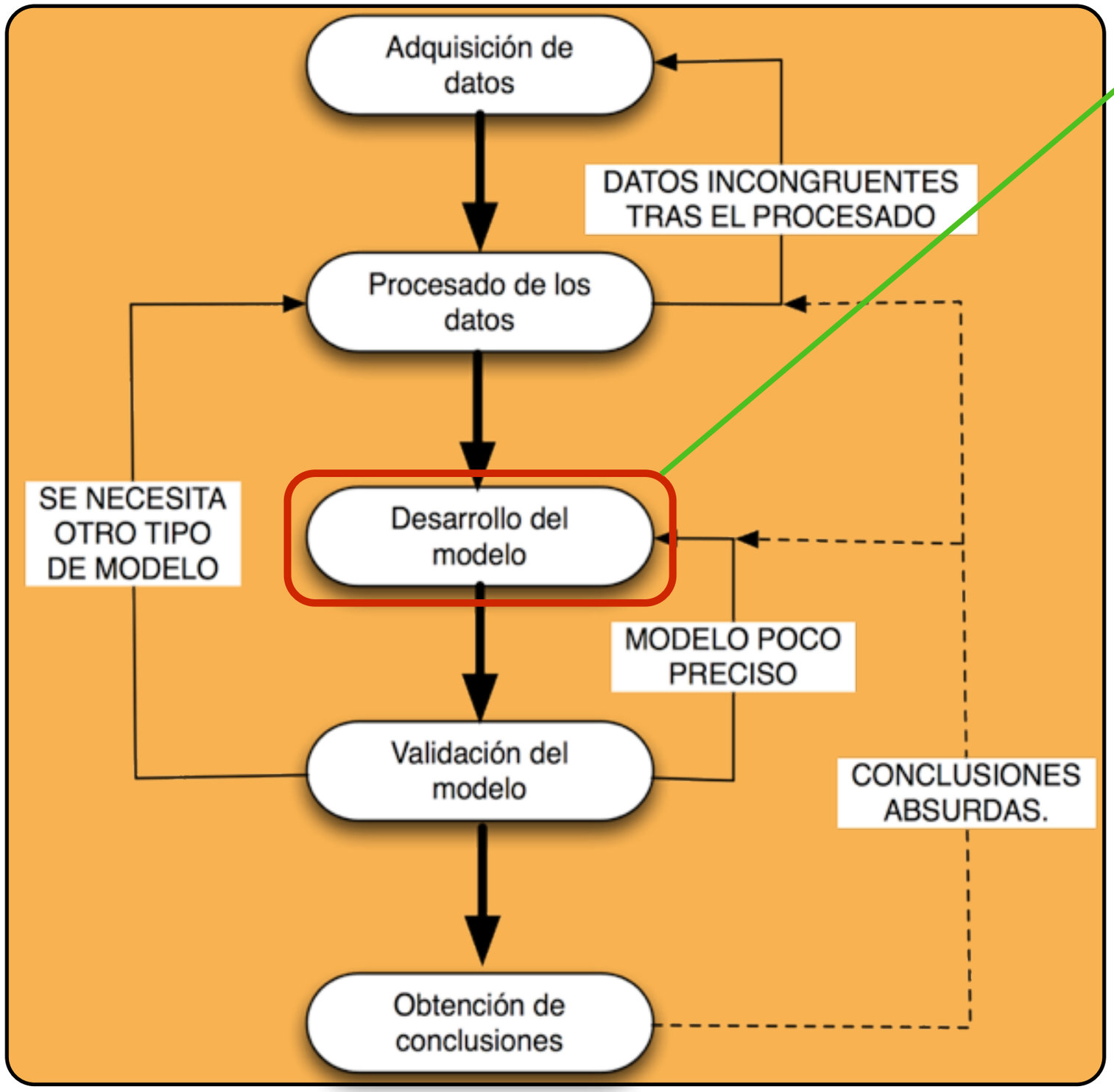
**Conocer en qué consiste un árbol de decisión.**

**Aprender los problemas que pueden surgir al aplicar un árbol de decisión.**

**Conocer las ventajas/inconvenientes frente a otros métodos ya vistos en el curso**

**Aprender a implementar un árbol de decisión.**

# Dónde estamos




Se ha comprobado el funcionamiento de la red neuronal verificándose que funciona mejor que un modelo lineal. Decidimos plantear un árbol de decisión por varias razones

Queremos un sistema cuya forma de clasificar/predecir sea visible

El sistema desarrollado debe permitir la extracción de regla si...entonces de forma directa

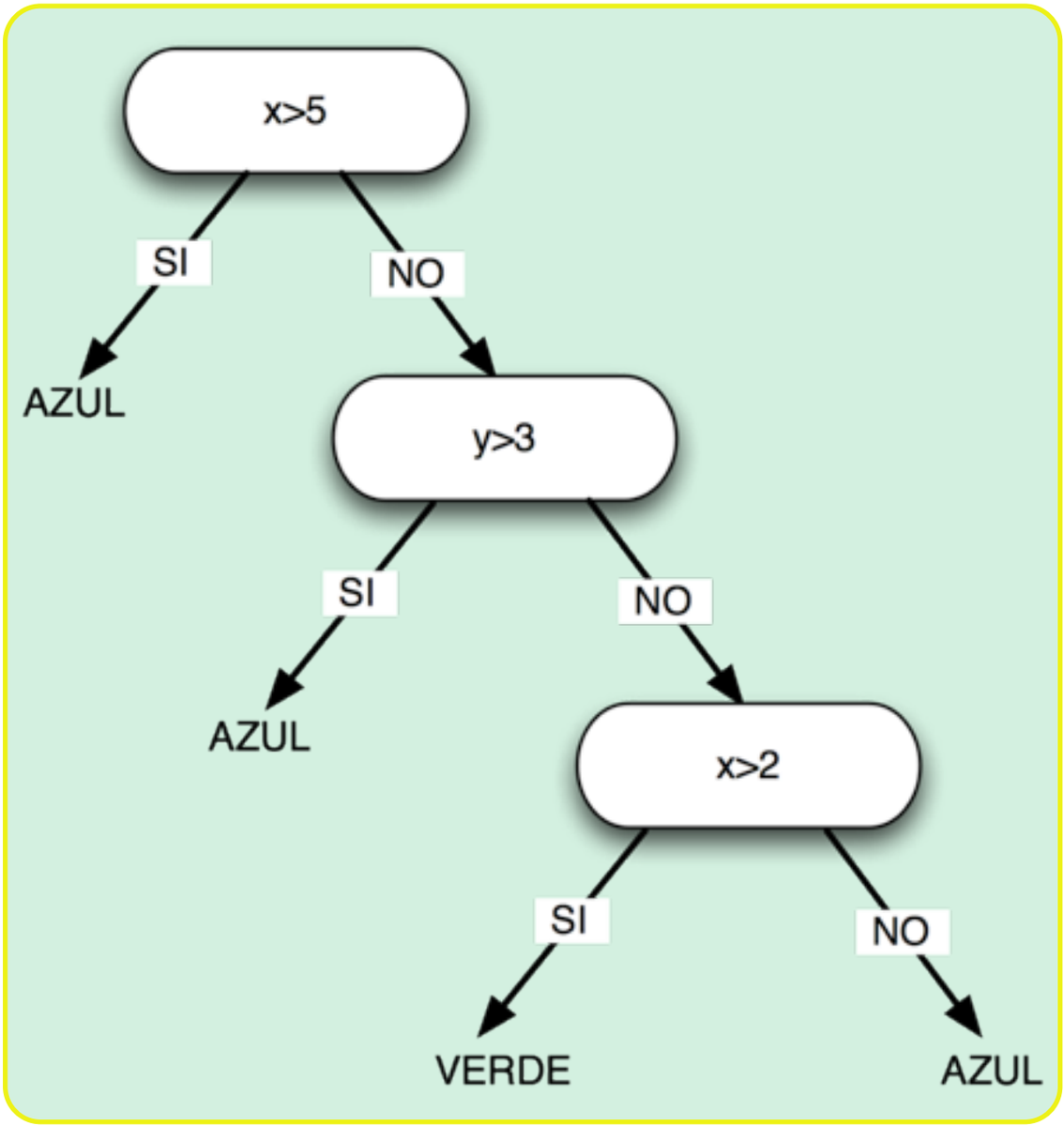
Se tienen un gran número de variables de entrada discretas no continuas.

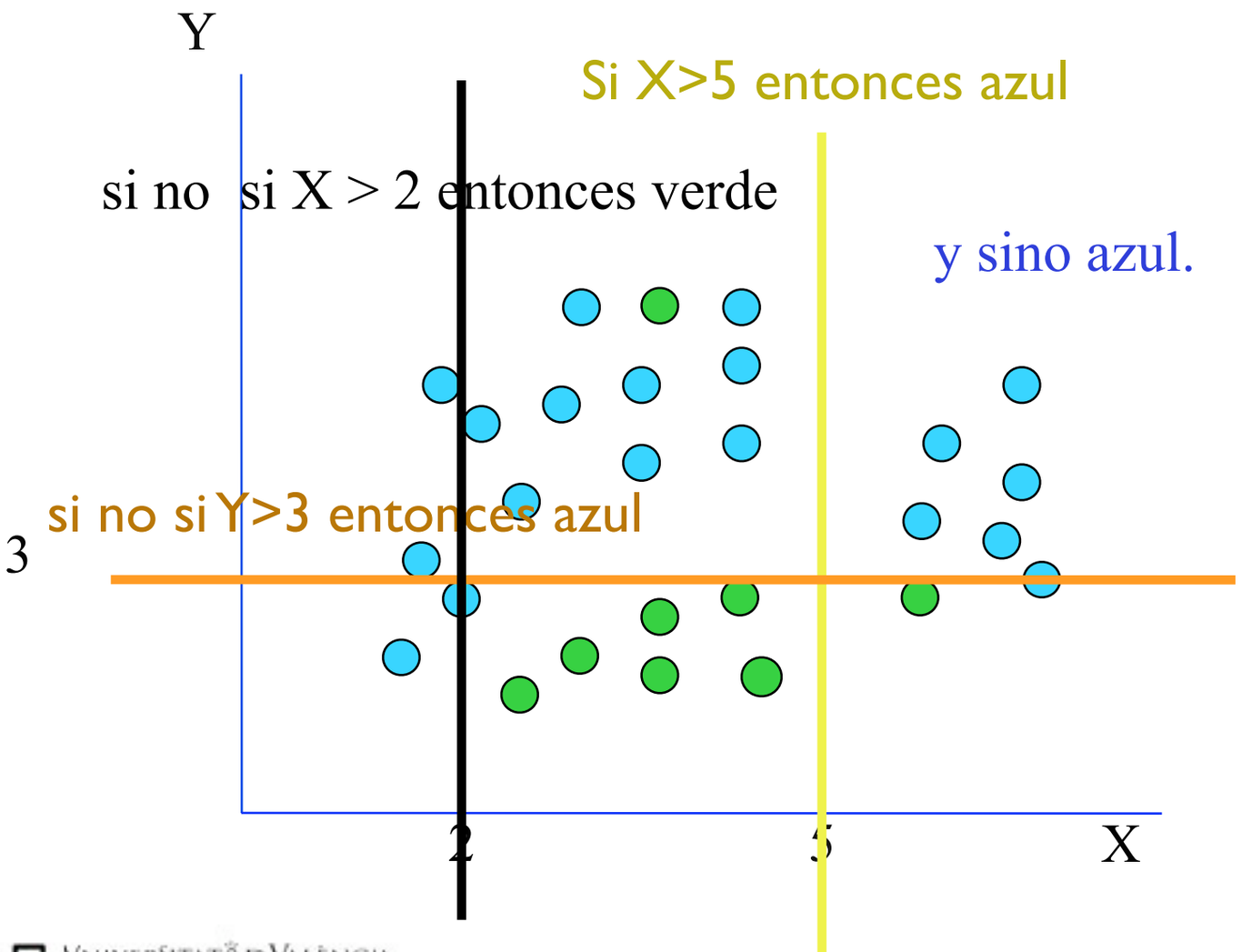
La red neuronal, aunque ofrece bueno resultados es demasiado grande en relación al conjunto de datos

Tenemos la certeza que existen muchos casos especiales dentro del conjunto de datos

# ¿Qué es un árbol de decisión?

Podríamos definir un ***árbol de decisión*** como un sistema que clasifica el vector de entrada en una serie de clases predefinidas usando una serie de preguntas secuenciales. Cada una de estas preguntas hace referencia a una variable de entrada




Aquí hay que preguntarse; ¿qué orden siguen las preguntas?; ¿hasta qué nivel se debe preguntar para que el modelo de decisión tenga sentido?



Emilio Soria, Antonio José Serrano y José David Martín Dpto Ingeniería Electrónica, ETSE
Sistemas de Ayuda a la Decisión Clínica, Curso 2009-2010

# Entropía.

Es la primera piedra en la Teoría de la Información de Shannon, teoría básica e imprescindible para el análisis de sistemas de transmisión/recepción de datos. De una manera intuitiva esta cantidad es directamente proporcional a  a la sorpresa que puede provocar una variable, e inversamente proporcional a la regularidad y redundancia que podamos tener en una variable.  A modo de ejemplo una distribución uniforme presenta la máxima entropía porque todos los valores son igualmente posibles.

$$Entropía = \sum p_i \cdot \log_2\left[1/p_i\right]$$

$$Entropía = -\sum p_i \cdot \log_2\left[p_i\right]$$

La obtención de la entropía en el lanzamiento de una moneda no trucada sería:

$$Entropía = -\left[\frac{1}{2}\cdot \log_2\left(\frac{1}{2}\right)+\frac{1}{2}\cdot \log_2\left(\frac{1}{2}\right)\right]=1$$

Imaginemos que dicha moneda está trucada; tenemos 1/4 posibilidades que salga cara y 3/4 que salga cruz entonces se tiene:

$$Entropía = -\left[\frac{1}{4}\cdot \log_2\left(\frac{1}{2}\right)+\frac{3}{4}\cdot \log_2\left(\frac{3}{4}\right)\right]=0.81$$

Existen otras magnitudes, también se conocen como entropías (por ejemplo la entropía de Renyi). En árboles de decisión otra entropía que se utiliza mucho es el índice de Gini

$$In_{GINI} = \sum_k p_k \cdot (1 - p_k)$$

# Entropía condicionada.

Como ya vimos en el tema de probabilidad el observar un suceso puede modificar la probabilidad de otro suceso si están relacionados de alguna forma. Con el concepto de entropía sucede algo similar apareciendo el concepto de **entropía condicionada**

$$H(Y|X) = \sum_k \Pr ob(X = v_k) \cdot H(Y|X = v_k)$$

| | Probabilidad | $H(Y\|X=v_k)$ |
|---|---|---|
| Matemáticas | 0,5 | 1 |
| Historia | 0,25 | 0 |
| Ciencias | 0,25 | 0 |

| *Asignatura (X)* | *Aprobado(Y)* |
|---|---|
| Matemáticas | Si |
| Historia | No |
| Ciencias | Si |
| Matemáticas | No |
| Matemáticas | No |
| Ciencias | Si |
| Historia | No |
| Matemáticas | Si |

Entropía(Y)=H(Y)=1
(tengo 4 aprobados y 4 suspensos)

$H(Y|X) = 0{,}5 \cdot 1 + 0{,}25 \cdot 0 + 0{,}25 \cdot 0 = 0{,}5$

SE REDUCE LA ENTROPÍA AL CONOCER X.

# Ganancia en información.

La ganancia de información determina el decremento de entropía al conocer el resultado de un suceso

$$GI(Y|X) = H(Y) - H(Y|X)$$

Para los árboles de decisión se va a plantear una modificación del concepto de ganancia de información.

$$GI(S|A) = H(Y) - \sum_{v_k \in V} \frac{|S_{v_k}|}{|S|} H(S_{v_k})$$

De lo que se trata es de determinar las entropías condicionadas PONDERADAS a la proporción de ejemplos que cumplen un determinado atributo (hay que fijarse además de la capacidad de reducir la entropía en la cantidad de ejemplos que van a cada nodo).

| *Asignatura (X)* | *Aprobado(Y)* |
|---|---|
| Matemáticas | Si |
| Historia | No |
| Ciencias | Si |
| Matemáticas | No |
| Matemáticas | No |
| Ciencias | Si |
| Historia | No |
| Matemáticas | Si |

$$H(Apr) = 1$$

En matemáticas tenemos 2 Si y 2 No

$$H(Apr|Mat) = 1$$

$$H(Apr|Historia) = H(Apr|Ciencias) = 0$$

$$GI(Apr|Asig) = 1 - \frac{4}{8} \times 1 - \frac{2}{8} \times 0 - \frac{2}{8} \times 0 = 0.5$$

# Ejemplo clásico (Machine Learning, Tom Mitchell).

| Tiempo | Temperatura | Humedad | Viento | Juega? |
|---|---|---|---|---|
| Soleado | Alta | Alta | No | N |
| Soleado | Alta | Alta | Si | N |
| Nuboso | Alta | Alta | No | P |
| Lluvioso | Media | Alta | No | P |
| Lluvioso | Media | Normal | No | P |
| Lluvioso | Baja | Normal | Si | N |
| Nuboso | Baja | Normal | Si | P |
| Soleado | Media | Alta | No | N |
| Soleado | Baja | Normal | No | P |
| Lluvioso | Media | Normal | No | P |
| Soleado | Media | Normal | Si | P |
| Nuboso | Media | Alta | Si | P |
| Nuboso | Alta | Normal | No | P |
| Lluvioso | Media | Alta | Si | N |

Se tiene el siguiente conjunto de datos con el que se intenta construir un árbol de decisión que, en virtud de las condiciones meteorológicas determine si se juega al tenis (P) o no se juega (N). Hay que determinar las ganancias en información de cada uno de los atributos.

# Cálculo de la ganancia de información (I)

| Tiempo | Juega? |
|---|---|
| Soleado | N |
| Soleado | N |
| Nuboso | P |
| Lluvioso | P |
| Lluvioso | P |
| Lluvioso | N |
| Nuboso | P |
| Soleado | N |
| Soleado | P |
| Lluvioso | P |
| Soleado | P |
| Nuboso | P |
| Nuboso | P |
| Lluvioso | N |

$$Entropía(juego) = -\frac{9}{14}\cdot \log_2\left(\frac{9}{14}\right) - \frac{5}{14}\cdot \log_2\left(\frac{5}{14}\right) = 0.92$$

$$Entropía(soleado) = -\frac{2}{5}\cdot \log_2\left(\frac{2}{5}\right) - \frac{3}{5}\cdot \log_2\left(\frac{3}{5}\right) = 0.97$$

$$Entropía(nuboso) = -\frac{4}{4}\cdot \log_2\left(\frac{4}{4}\right) - \frac{0}{4}\cdot \log_2\left(\frac{0}{4}\right) = 0.00$$

$$Entropía(lluvioso) = -\frac{3}{5}\cdot \log_2\left(\frac{3}{5}\right) - \frac{2}{5}\cdot \log_2\left(\frac{2}{5}\right) = 0.97$$

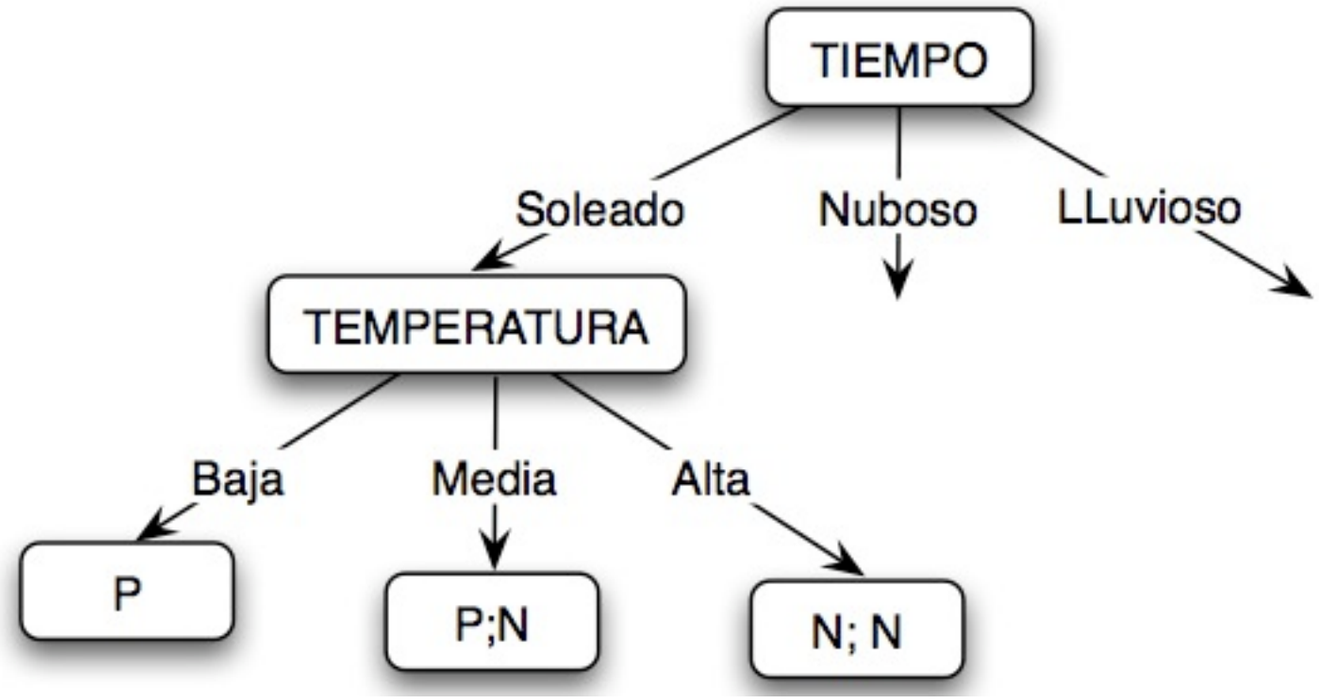
$$Ganancia(tiempo) = 0.92 - \frac{5}{14}\cdot 0.97 - \frac{4}{14}\cdot 0 - \frac{5}{14}\cdot 0.97 = 0.92 - 0.69 = 0.23$$

$$Ganancia(humedad) = 0.15 \quad Ganancia(viento) = 0.05$$

PONEMOS EN LA RAÍZ DEL ÁRBOL EL TIEMPO.

# Cálculo de la ganancia de información (II)

| Tiempo | Temperatura | Juega? |
|---|---|---|
| Soleado | Alta | N |
| Soleado | Alta | N |
| Nuboso | Alta | P |
| Lluvioso | Media | P |
| Lluvioso | Media | P |
| Lluvioso | Baja | N |
| Nuboso | Baja | P |
| Soleado | Media | N |
| Soleado | Baja | P |
| Lluvioso | Media | P |
| Soleado | Media | P |
| Nuboso | Media | P |
| Nuboso | Alta | P |
| Lluvioso | Media | N |




$$Entropía(nodo) = -\frac{2}{5}\cdot log_2\left(\frac{2}{5}\right) - \frac{3}{5}\cdot log_2\left(\frac{3}{5}\right) = 0.97$$

$$Entropía(alta) = Entropía(baja) = 0$$

$$Entropía(media) = 1$$

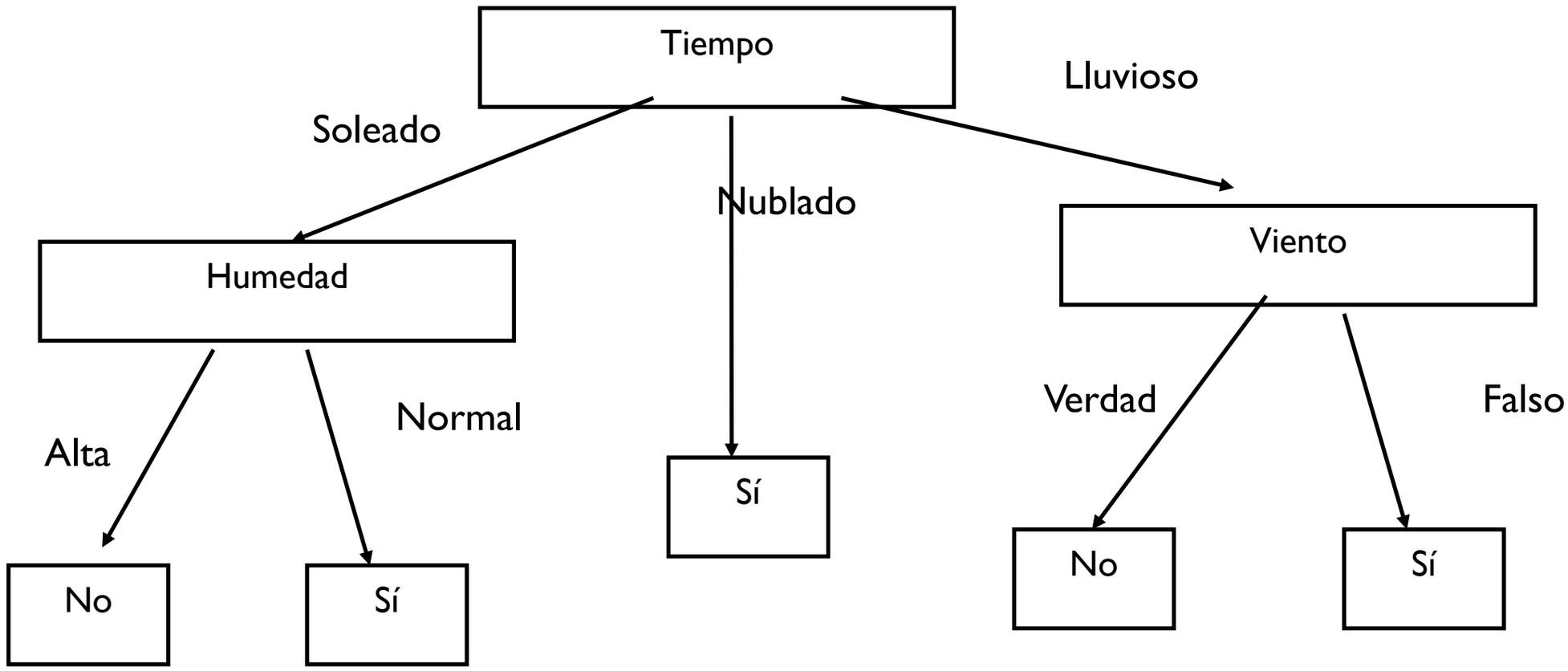
$$Ganancia(temperatura) = 0.97 - \frac{2}{5}\cdot 0 - \frac{2}{5}\cdot 1 - \frac{1}{5}\cdot 0 = 0.97 - 0.4 = 0.57$$

$$Ganancia(humedad) = 0.97 \quad Ganancia(viento) = 0.02$$

Se escoge entonces como siguiente nodo a la humedad

# Cálculo de la ganancia de información (III)

El proceso se repite hasta construir todo el árbol de manera análoga a lo comentado en las anteriores transparencias.



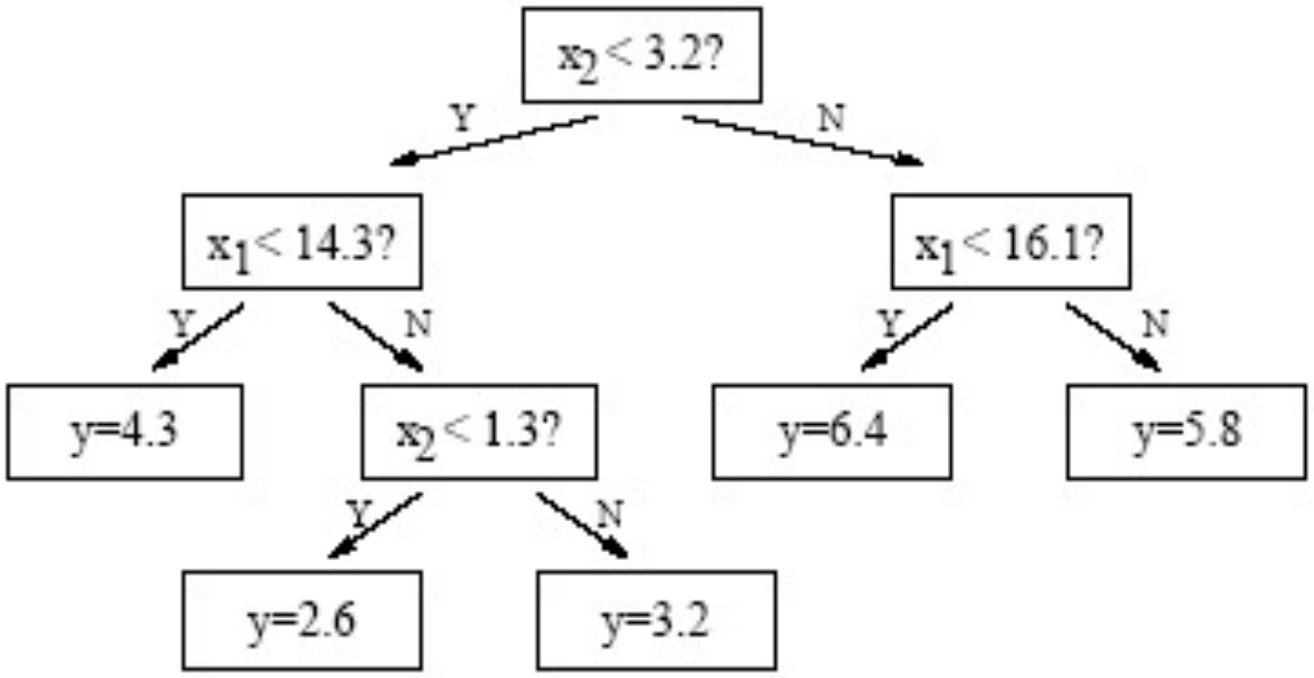

Este árbol proporciona una ayuda para la toma de decisiones de una manera clara y concisa

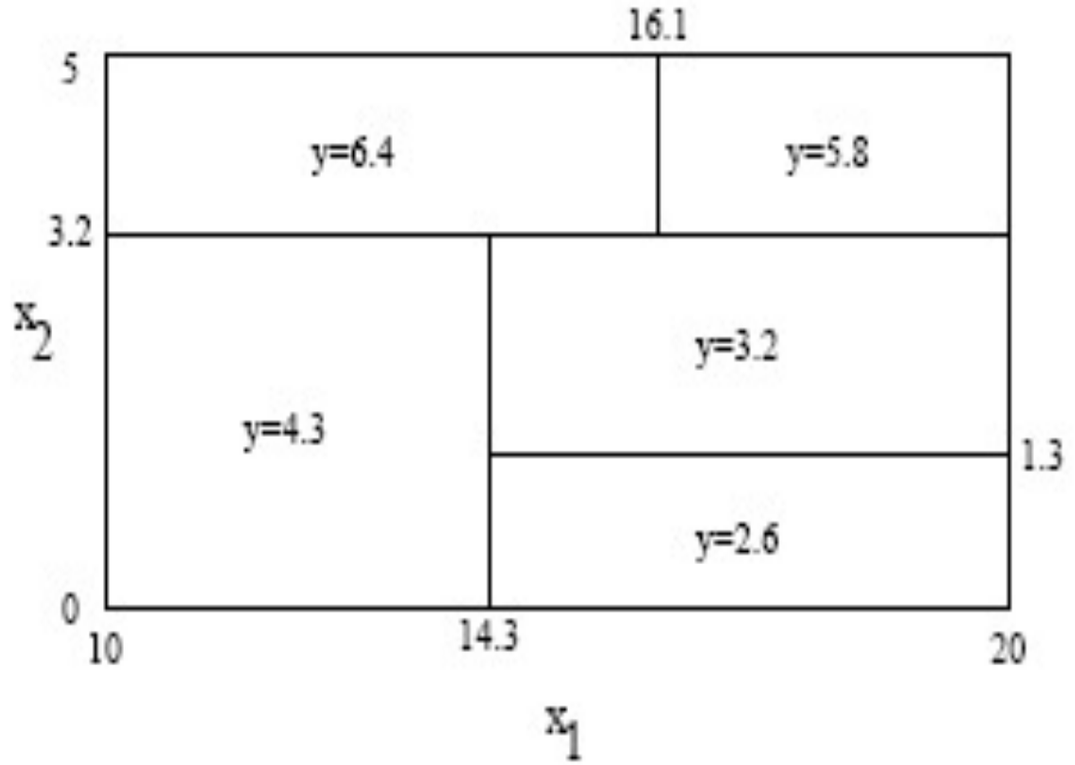
El primer problema que nos encontramos es que, conforme los nodos se dividen , la cantidad de datos utilizados para las siguientes divisiones se hace más pequeño de forma exponencial.

La construcción del árbol asume que todas las variables interactúan (se tienen efectos aditivos) aunque en el problema no intervengan las variables.

# Arboles de regresión

Ahora los nodos finales del árbol contienen valores numéricos (valores predichos).




Cuando las variables son numéricas y no categóricas no tiene sentido utilizar la entropía directamente. Se procede de dos maneras principalmente:

a) Se procede a categorizar las variables numéricas mediante umbrales o algortimos mas sofisticados.

b) Se utiliza la reducción del error cuadrático (o similar) como criterio de separación en vez de la ganancia de información.

$$\sum_{i \in \text{node } k} (y^{(i)} - \bar{y}_k)^2 \quad - \sum_{i \in \text{node } k} \left\{ \begin{array}{ll} (y^{(i)} - \bar{y}_{k_1})^2 & \text{if } x_j^{(i)} < v \\ (y^{(i)} - \bar{y}_{k_2})^2 & \text{if } x_j^{(i)} \geq v \end{array} \right\}$$

Consideramos el error cuadrático medio antes y después de realizar la separación de los datos . El "penalización" de cada conjunto se calcula mediante la suma cuadrado de diferencia entre los valores y el valor medio del conjunto.

# Comentarios sobre los árboles de decisión.

El algoritmo que se ha comentado aquí es el básico (ID3) existiendo muchos más pero se ha escogido ese por su sencillez. Una evolución de ese algoritmo es el C4.5

De igual forma existen árboles que no son de decisión sino que se usan para problemas de regresión. Los más famosos dentro de este grupo son los conocidos como CART.

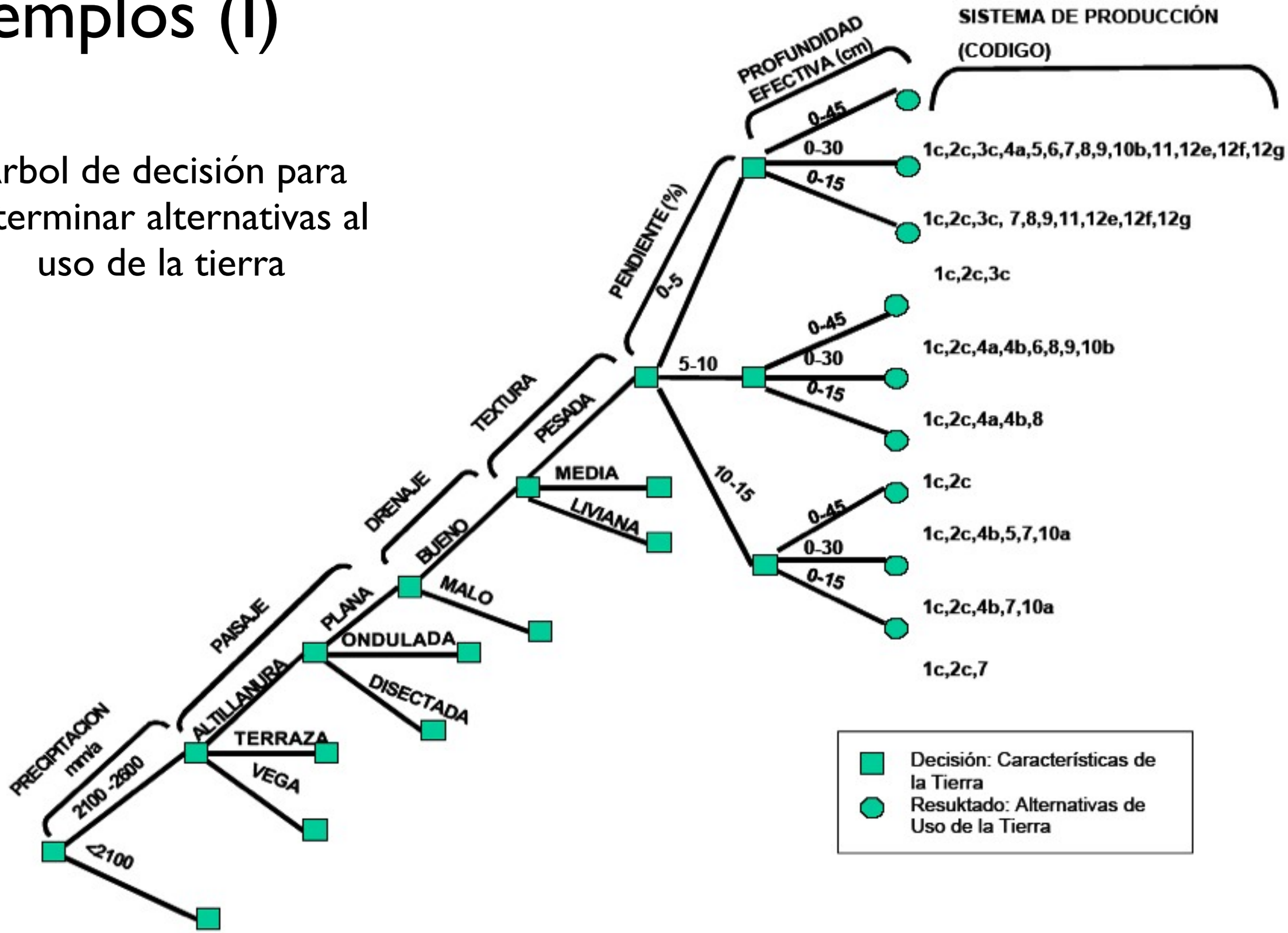
El principal problema de este tipo de modelos es el sobreajuste que se puede cometer. Este sobreajuste se refleja en tener un árbol demasiado profundo, o lo que es lo mismo, especificar para cada nodo último de decisión un patrón de entrada.

Para evitar este problema se plantean algoritmos de poda que, la misión que tienen es eliminar ramas excesivamente profundas y específicas del árbol desarrollado.

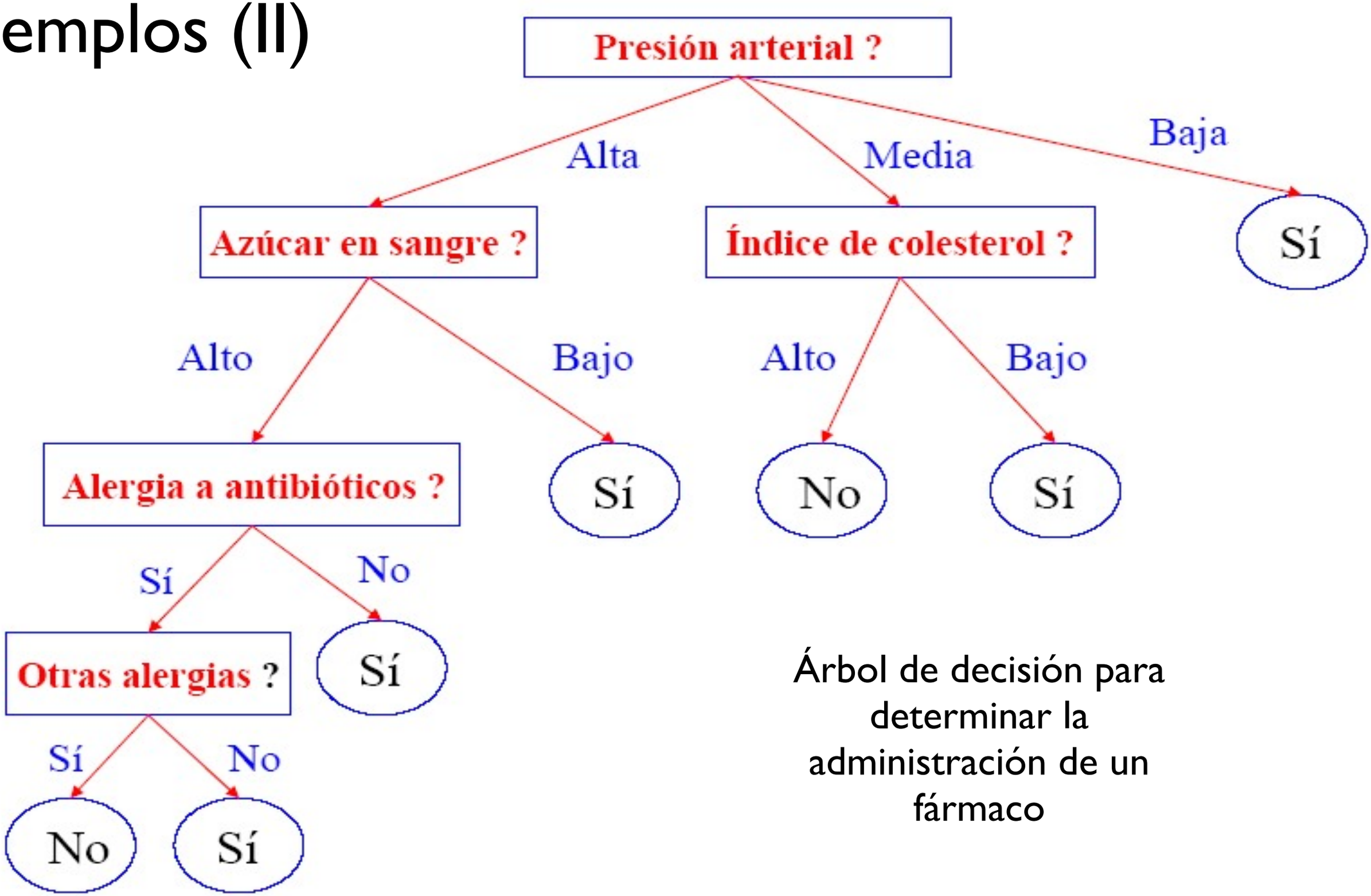
Son modelos muy extendidos en determinados ámbitos del conocimiento existiendo un gran número de paquetes informáticos de libre distribución que los implementan (por ejemplo WEKA o R).

# Ejemplos (I)

Árbol de decisión para determinar alternativas al uso de la tierra

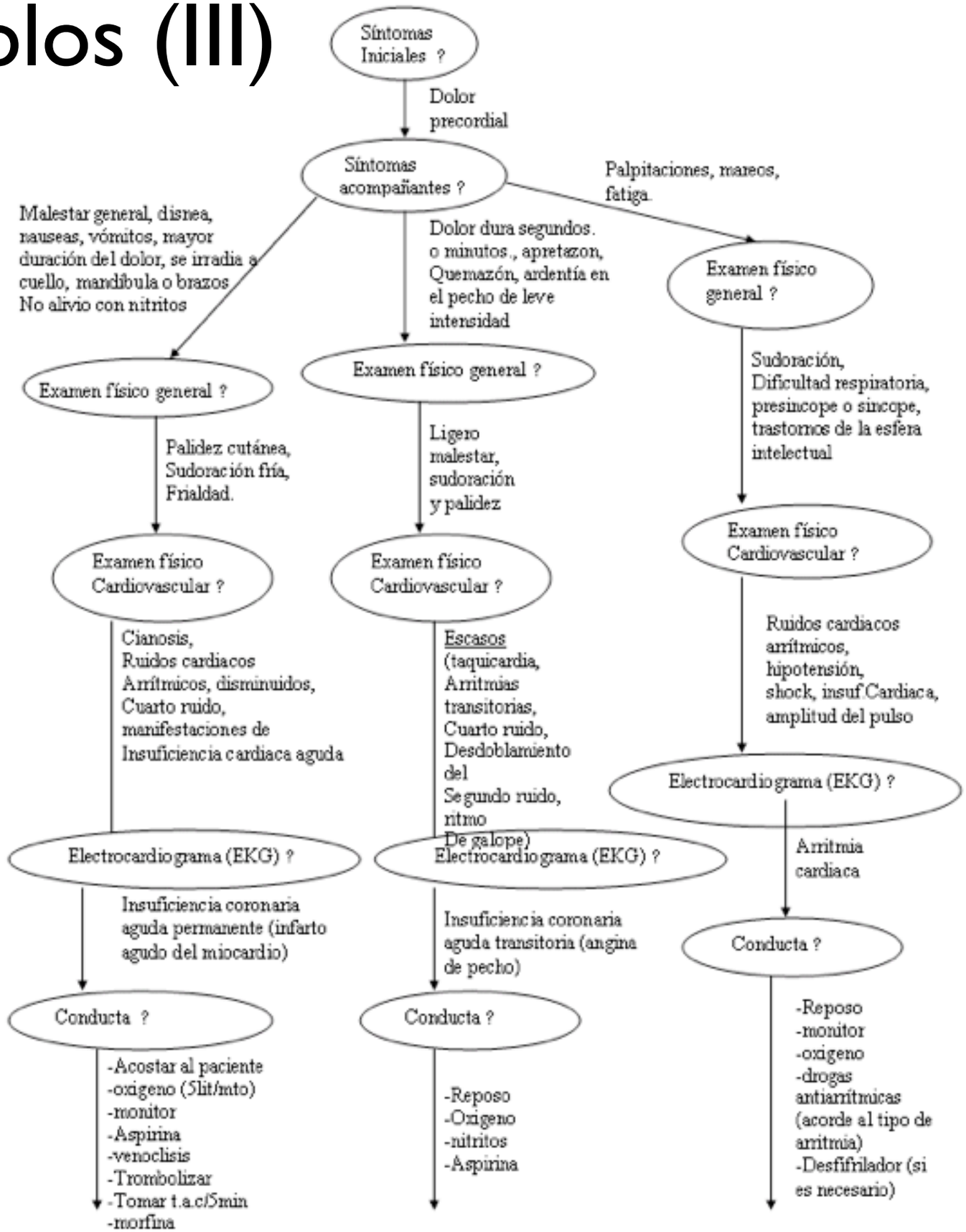


# Ejemplos (II)



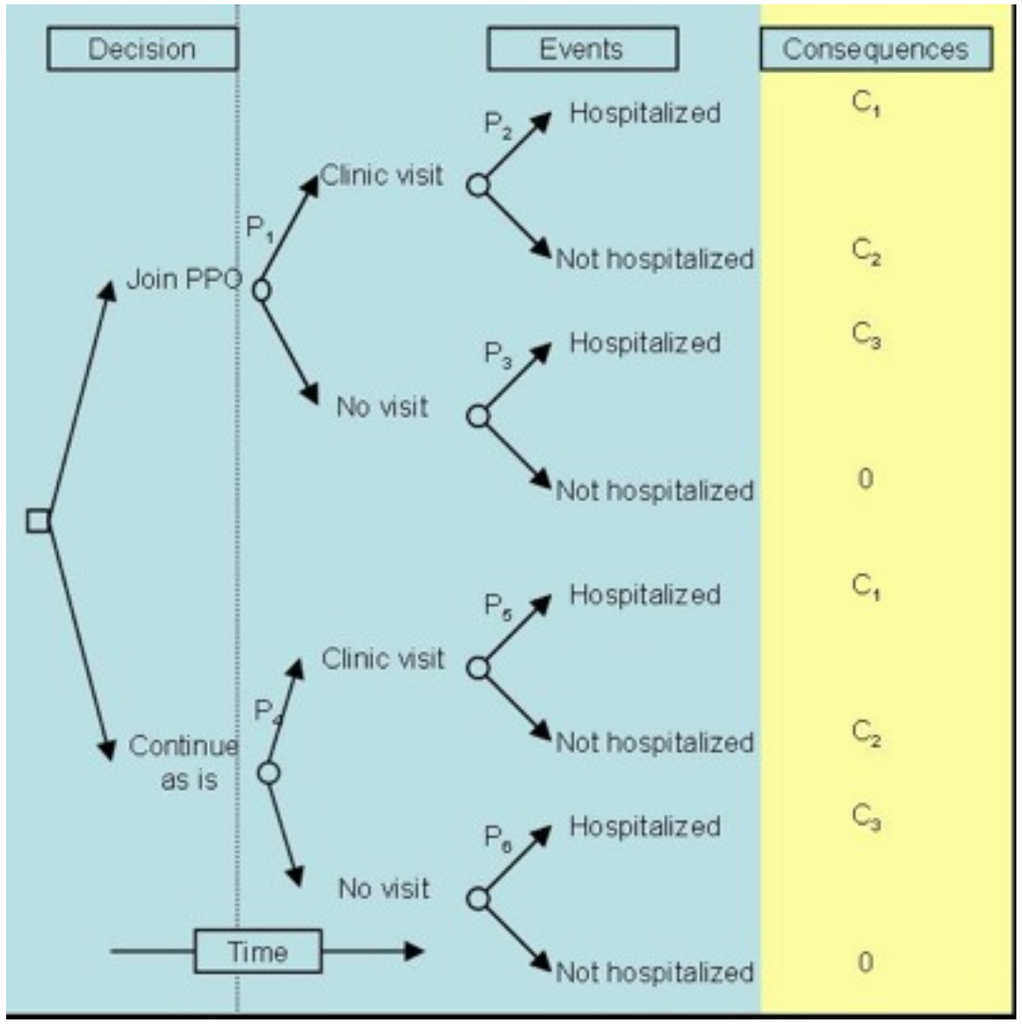

Árbol de decisión para determinar la administración de un fármaco

# Ejemplos (III)




Árbol de decisión o protocolo de actuación médico

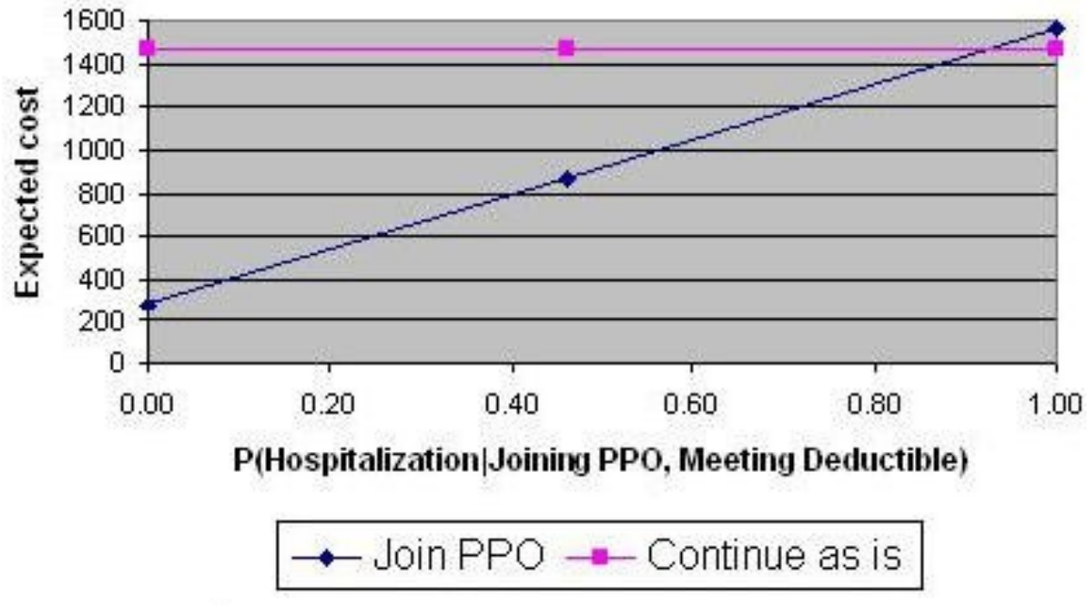
# Toma de decisiones.




Los árboles de decisión son una herramienta para elegir entre varias alternativas. Las decisiones pueden estar afectadas por incertidumbre, coste asociados y utilidad.

Contienen nodos que representan decisiones, nodos que representan situaciones aleatorias y, finalmente, aparecen las consecuencias de las decisiones.

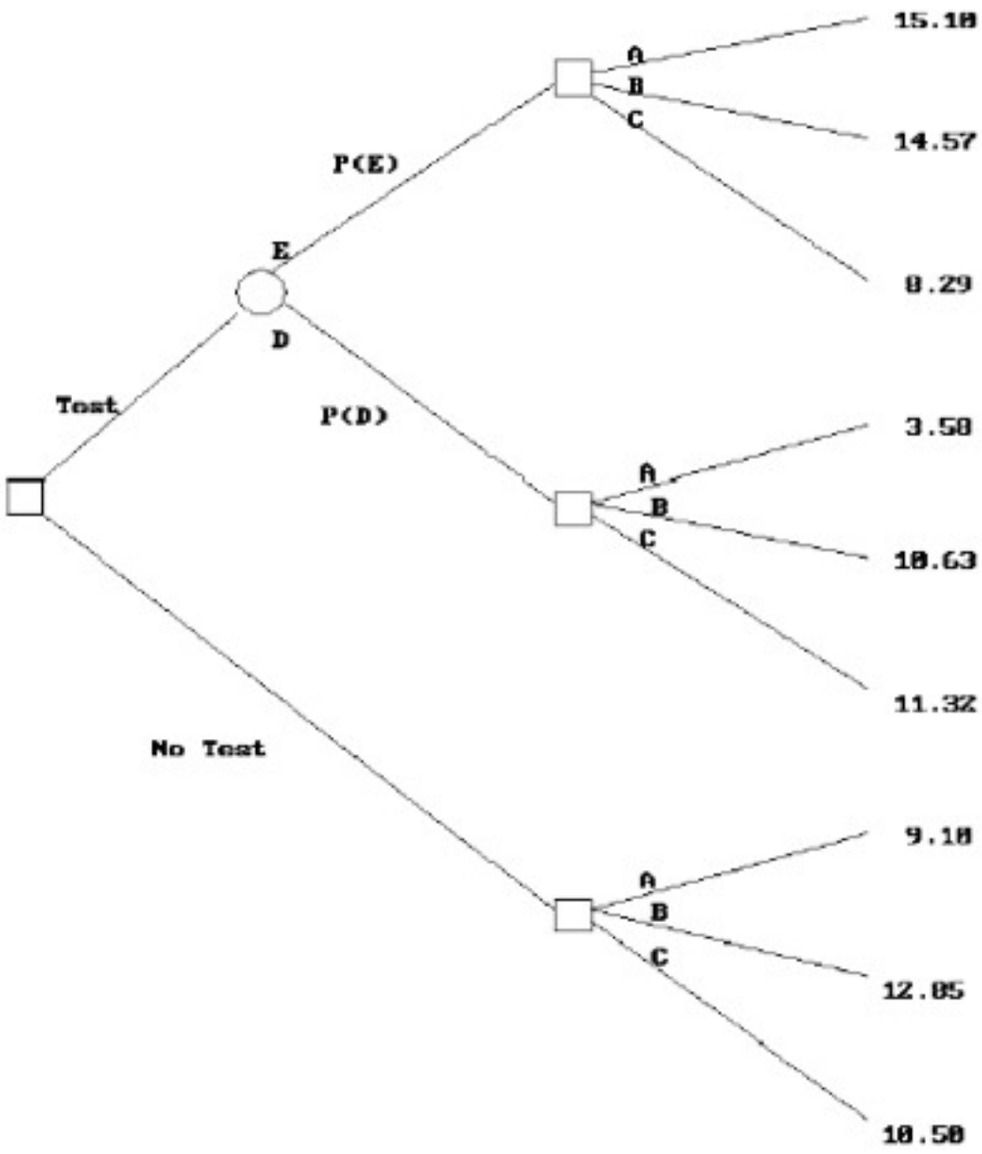
Estas decisiones finales pueden estar asociadas a costes (económicos) o utilidades (otros factores además de los económicos, emocionales, prácticos, etc).

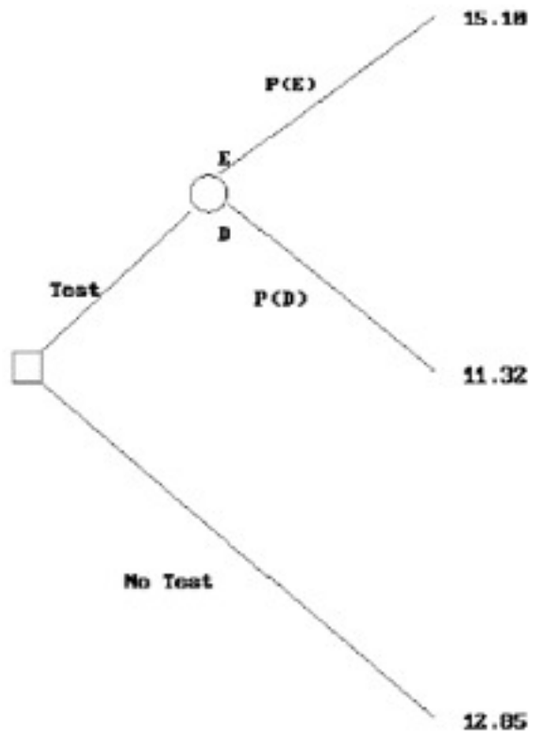


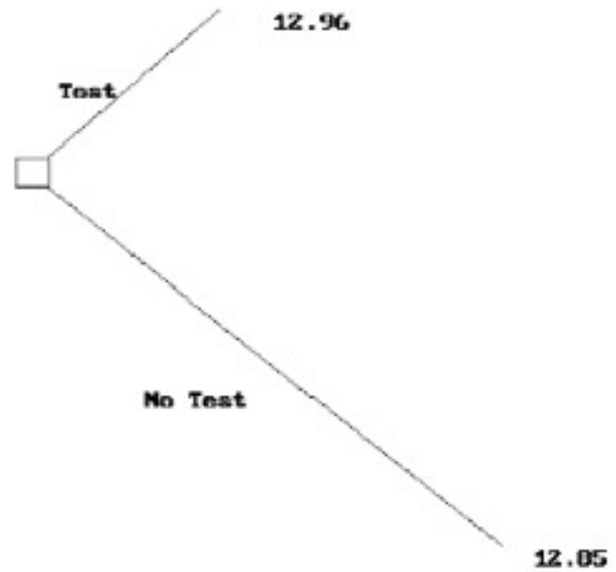

Una manera de mejorar el entendimiento del proceso de toma de decisiones consiste en realizar un análisis de sensibilidad, es decir, realizar cambios en los parámetros hasta que las conclusiones sean afectadas.

# Ejemplo.

Una hospital realiza un test antes de decidir el tratamiento a proporcionar a los pacientes. Existen 3 tipos de pacientes fármacos A, B y C. Un posible árbol para analizar el problema podría se el siguiente:




Finalmente, aplicando los conceptos, de valor esperado es posible reducir el árbol hasta dejar patente cual es la consecuencia (costes o utilidades) de tomar una decisión.

VNIVERSITAT DE VALÈNCIA

# MASTER DE INGENIERÍA BIOMÉDICA.

# Métodos de ayuda al diagnóstico clínico.

## Tema 6: Árboles de decisión.