ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN PROFESIONAL

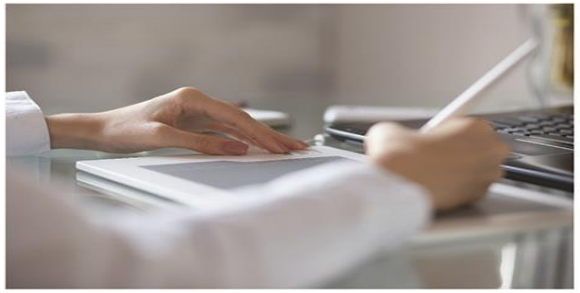
Diplomado en Big Data y Ciencias de Datos

Minería de Datos

Árboles de Decisión

Educación Profesional - Escuela de Ingeniería UC

Sebastián Raveau




Recordemos las técnicas de Minería de Datos

Técnicas Predictivas – Aprendizaje Supervisado

Regresión	ajustar variables/relaciones continuas
Clasificación	ajustar variables/relaciones discretas

Técnicas Descriptivas – Aprendizaje No Supervisado

Clustering	agrupar datos similares
Asociación	identificar patrones y coocurrencias

Árboles de decisión

Los árboles de decisión son métodos de aprendizaje supervisado

El objetivo de un árbol de decisión es identificar de forma automática el valor de una variable de interés (variable endógena) a partir de un conjunto de variables explicativas (variables exógenas)

Árboles de decisión

Existen dos tipos de árboles de decisión:

Árboles de clasificación: cuando la variable de interés es discreta

Árboles de regresión: cuando la variable de interés es continua

Árboles de decisión

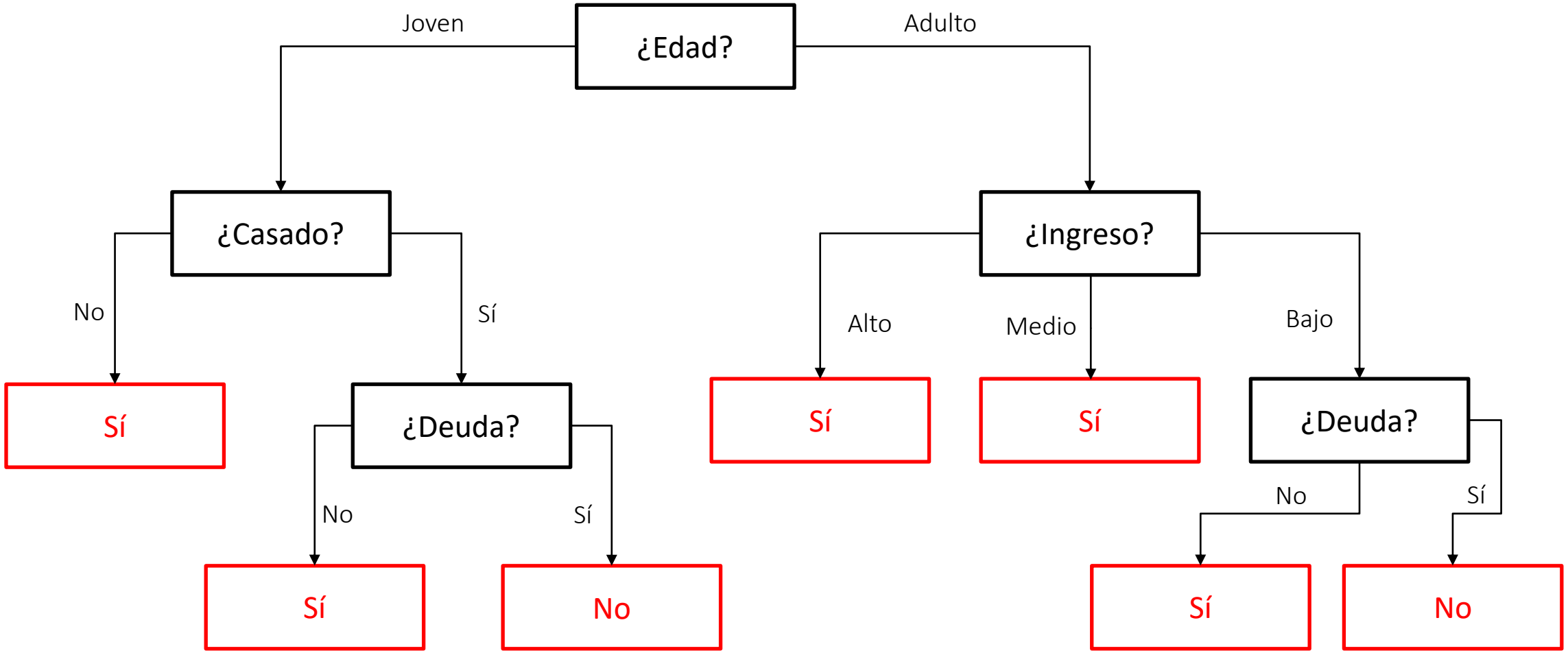
En un árbol de decisión existen dos tipos de nodos:

Los nodos internos de un árbol representan atributos a partir de los cuales se realizan divisiones de los datos (esto incluye la raíz del árbol)

Los nodos terminales de un árbol (también llamados “hojas”) entregan un valor ajustado para la variable de interés

Árboles de decisión

Supongamos que queremos predecir si una persona comprará cierto producto



Árboles de decisión

Los árboles de decisión son iterativos:

En cada nodo escogemos el mejor atributo para dividir

A partir del atributo escogido se construyen ramas a nuevos nodos

Dividimos los registros de acuerdo al atributo para generar dos nuevos nodos

Y repetimos…

Árboles de decisión

En general requieren de dos sets de datos:

Datos de entrenamiento, donde se construye en árbol

Datos de predicción, donde se evalúa el árbol construido

Árboles de clasificación

Ganancia de información

Supongamos un conjunto de datos:

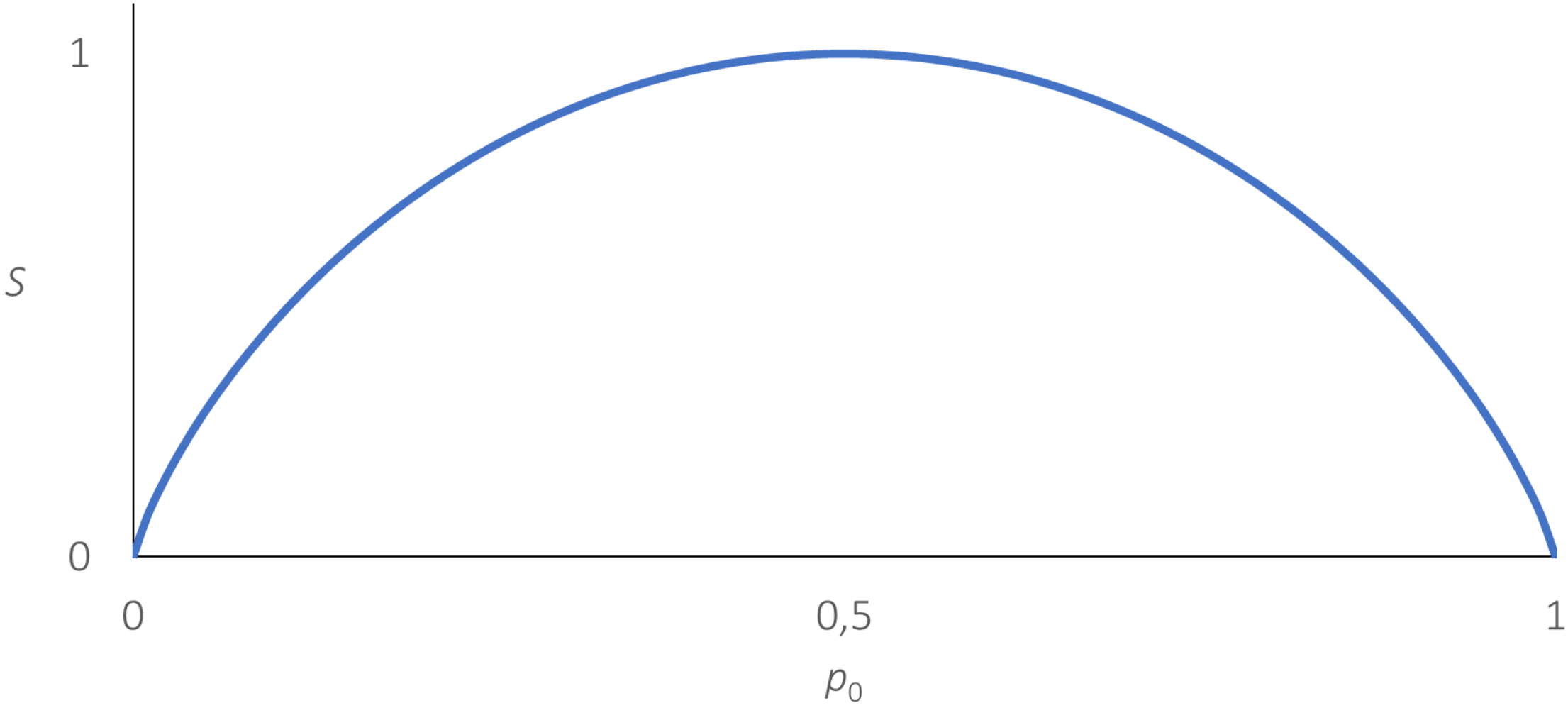
Exógena (predictora)	**Endógena** (a predecir)
B	1
A	1
B	1
B	0
A	0
A	0
A	0

Entropía: medida de homogeneidad u orden

$$S = -p_0 \cdot \log_2(p_0) - p_1 \cdot \log_2(p_1)$$

$$S = -\frac{3}{7} \cdot \log_2\left(\frac{3}{7}\right) - \frac{4}{7} \cdot \log_2\left(\frac{4}{7}\right) = 0{,}985$$

Ganancia de información



Ganancia de información

¿Cuánto ganamos usando nuestra variable endógena para dividir?

Exógena (predictora)	**Endógena** (a predecir)
A	1
A	0
A	0
A	0

Exógena (predictora)	**Endógena** (a predecir)
B	1
B	1
B	0

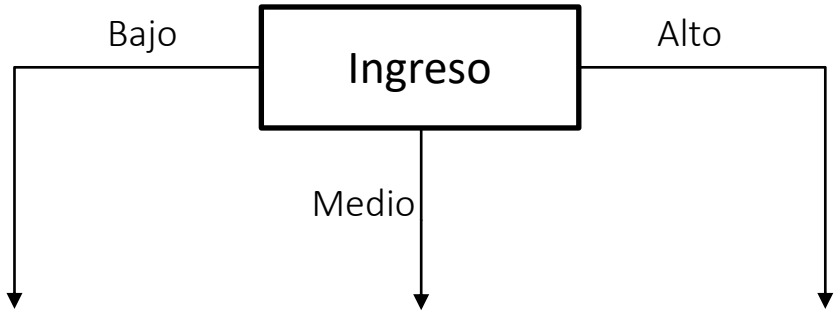
$$S_A = -\frac{1}{4}\cdot\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\cdot\log_2\left(\frac{3}{4}\right) = 0{,}811$$

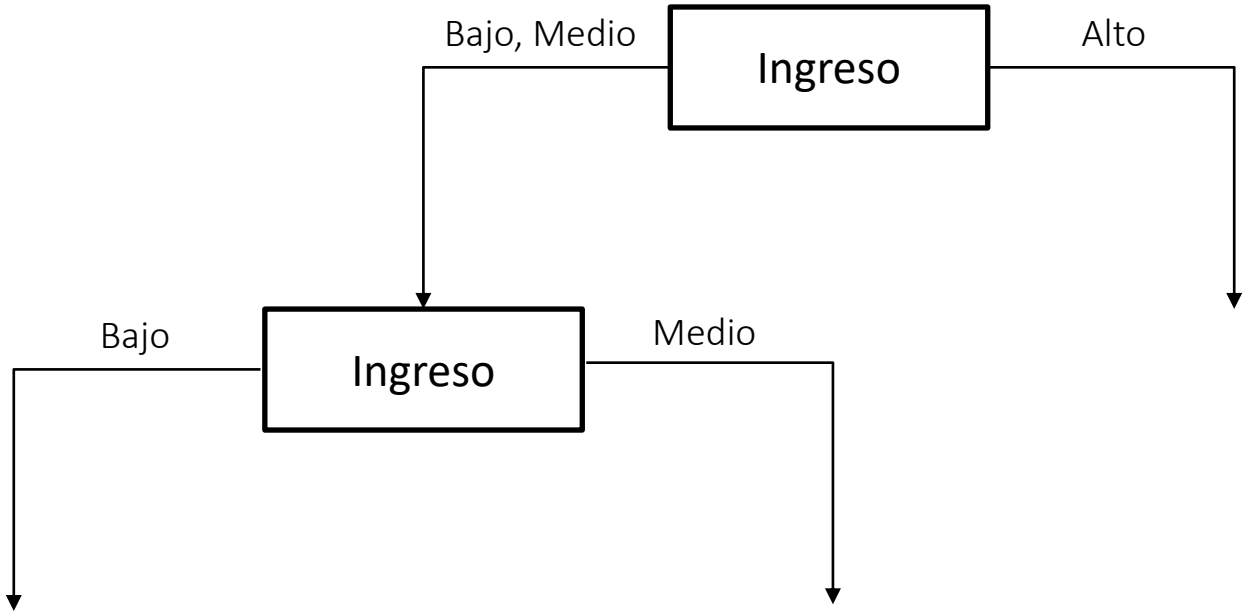
$$S_B = -\frac{2}{3}\cdot\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\cdot\log_2\left(\frac{1}{3}\right) = 0{,}918$$

$$\text{Ganancia} = 0{,}985 - \frac{4}{7}\cdot 0{,}811 - \frac{3}{7}\cdot 0{,}918 = 0{,}128$$

Ganancia de información

Siempre es posible dividir en dos los datos, aún cuando la variable exógena tenga más posibles niveles





Consideremos los siguientes datos

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Nublado	Calor	Alta	Débil	Sí
Soleado	Agradable	Normal	Fuerte	Sí
Nublado	Agradable	Alta	Fuerte	Sí
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Frío	Normal	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Soleado	Calor	Alta	Fuerte	No
Nublado	Calor	Normal	Débil	Sí
Lluvioso	Agradable	Alta	Fuerte	No

Entropía inicial

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Nublado	Calor	Alta	Débil	Sí
Soleado	Agradable	Normal	Fuerte	Sí
Nublado	Agradable	Alta	Fuerte	Sí
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Frío	Normal	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Soleado	Calor	Alta	Fuerte	No
Nublado	Calor	Normal	Débil	Sí
Lluvioso	Agradable	Alta	Fuerte	No

$$S = -\frac{5}{10} \cdot \log_2\left(\frac{5}{10}\right) - \frac{5}{10} \cdot \log_2\left(\frac{5}{10}\right) = 1$$

Clasificando según clima soleado

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Soleado	Agradable	Normal	Fuerte	Sí
Soleado	Calor	Alta	Fuerte	No

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Nublado	Calor	Alta	Débil	Sí
Nublado	Agradable	Alta	Fuerte	Sí
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Frío	Normal	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Nublado	Calor	Normal	Débil	Sí
Lluvioso	Agradable	Alta	Fuerte	No

$$S_1 = -\frac{1}{3}\cdot\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\cdot\log_2\left(\frac{2}{3}\right) = 0{,}918$$

$$S_2 = -\frac{4}{7}\cdot\log_2\left(\frac{4}{7}\right) - \frac{3}{7}\cdot\log_2\left(\frac{3}{7}\right) = 0{,}985$$

$$G = 1 - \frac{3}{10}\cdot 0{,}918 - \frac{7}{10}\cdot 0{,}985 = 0{,}035$$

Clasificando según clima nublado

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Nublado	Calor	Alta	Débil	Sí
Nublado	Agradable	Alta	Fuerte	Sí
Nublado	Calor	Normal	Débil	Sí

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Soleado	Agradable	Normal	Fuerte	Sí
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Frío	Normal	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Soleado	Calor	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Fuerte	No

$$S_1 = -\frac{3}{3}\cdot\log_2\left(\frac{3}{3}\right) - \frac{0}{3}\cdot\log_2\left(\frac{0}{3}\right) = 0$$

$$S_2 = -\frac{2}{7}\cdot\log_2\left(\frac{2}{7}\right) - \frac{5}{7}\cdot\log_2\left(\frac{5}{7}\right) = 0{,}863$$

$$G = 1 - \frac{3}{10}\cdot 0 - \frac{7}{10}\cdot 0{,}863 = 0{,}396$$

Clasificando según clima lluvioso

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Frío	Normal	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Lluvioso	Agradable	Alta	Fuerte	No

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Nublado	Calor	Alta	Débil	Sí
Soleado	Agradable	Normal	Fuerte	Sí
Nublado	Agradable	Alta	Fuerte	Sí
Soleado	Calor	Alta	Fuerte	No
Nublado	Calor	Normal	Débil	Sí

$$S_1 = -\frac{1}{4}\cdot\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\cdot\log_2\left(\frac{3}{4}\right) = 0{,}811$$

$$S_2 = -\frac{4}{6}\cdot\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\cdot\log_2\left(\frac{2}{6}\right) = 0{,}918$$

$$G = 1 - \frac{4}{10}\cdot 0{,}811 - \frac{6}{10}\cdot 0{,}918 = 0{,}125$$

Clasificando según temperatura calurosa

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Nublado	Calor	Alta	Débil	Sí
Soleado	Calor	Alta	Fuerte	No
Nublado	Calor	Normal	Débil	Sí

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Agradable	Normal	Fuerte	Sí
Nublado	Agradable	Alta	Fuerte	Sí
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Frío	Normal	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Lluvioso	Agradable	Alta	Fuerte	No

$$S_1 = -\frac{2}{4}\cdot\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\cdot\log_2\left(\frac{2}{4}\right) = 1$$

$$S_2 = -\frac{3}{6}\cdot\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\cdot\log_2\left(\frac{3}{6}\right) = 1$$

$$G = 1 - \frac{4}{10}\cdot 1 - \frac{6}{10}\cdot 1 = 0$$

Clasificando según temperatura agradable

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Agradable	Normal	Fuerte	Sí
Nublado	Agradable	Alta	Fuerte	Sí
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Lluvioso	Agradable	Alta	Fuerte	No

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Nublado	Calor	Alta	Débil	Sí
Lluvioso	Frío	Normal	Fuerte	No
Soleado	Calor	Alta	Fuerte	No
Nublado	Calor	Normal	Débil	Sí

$$S_1 = -\frac{3}{5}\cdot\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\cdot\log_2\left(\frac{2}{5}\right) = 0{,}971$$

$$S_2 = -\frac{2}{5}\cdot\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\cdot\log_2\left(\frac{3}{5}\right) = 0{,}971$$

$$G = 1 - \frac{5}{10}\cdot 0{,}971 - \frac{5}{10}\cdot 0{,}971 = 0{,}029$$

Clasificando según temperatura fría

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Lluvioso	Frío	Normal	Fuerte	No

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Nublado	Calor	Alta	Débil	Sí
Soleado	Agradable	Normal	Fuerte	Sí
Nublado	Agradable	Alta	Fuerte	Sí
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Soleado	Calor	Alta	Fuerte	No
Nublado	Calor	Normal	Débil	Sí
Lluvioso	Agradable	Alta	Fuerte	No

$$S_1 = -\frac{0}{1}\cdot\log_2\left(\frac{0}{1}\right) - \frac{1}{1}\cdot\log_2\left(\frac{1}{1}\right) = 0$$

$$S_2 = -\frac{5}{9}\cdot\log_2\left(\frac{5}{9}\right) - \frac{4}{9}\cdot\log_2\left(\frac{4}{9}\right) = 0{,}991$$

$$G = 1 - \frac{1}{10}\cdot 0 - \frac{9}{10}\cdot 0{,}991 = 0{,}108$$

Clasificando según humedad

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Agradable	Normal	Fuerte	Sí
Lluvioso	Frío	Normal	Fuerte	No
Nublado	Calor	Normal	Débil	Sí

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Nublado	Calor	Alta	Débil	Sí
Nublado	Agradable	Alta	Fuerte	Sí
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Soleado	Calor	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Fuerte	No

$$S_1 = -\frac{2}{3}\cdot\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\cdot\log_2\left(\frac{1}{3}\right) = 0{,}918$$

$$S_2 = -\frac{3}{7}\cdot\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\cdot\log_2\left(\frac{4}{7}\right) = 0{,}985$$

$$G = 1 - \frac{3}{10}\cdot 0{,}918 - \frac{7}{10}\cdot 0{,}985 = 0{,}035$$

Clasificando según viento

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Nublado	Calor	Alta	Débil	Sí
Lluvioso	Agradable	Alta	Débil	Sí
Nublado	Calor	Normal	Débil	Sí

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Agradable	Normal	Fuerte	Sí
Nublado	Agradable	Alta	Fuerte	Sí
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Frío	Normal	Fuerte	No
Soleado	Calor	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Fuerte	No

$$S_1 = -\frac{3}{4}\cdot\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\cdot\log_2\left(\frac{1}{4}\right) = 0{,}811$$

$$S_2 = -\frac{2}{6}\cdot\log_2\left(\frac{2}{6}\right) - \frac{4}{6}\cdot\log_2\left(\frac{4}{6}\right) = 0{,}918$$

$$G = 1 - \frac{4}{10}\cdot 0{,}811 - \frac{6}{10}\cdot 0{,}918 = 0{,}125$$

Resumiendo

Clima soleado	$G = 0{,}035$
Clima nublado	$G = 0{,}396$
Clima lluvioso	$G = 0{,}125$
Temperatura calurosa	$G = 0$
Temperatura agradable	$G = 0{,}029$
Temperatura fría	$G = 0{,}108$
Humedad	$G = 0{,}035$
Viento	$G = 0{,}125$

Resumiendo

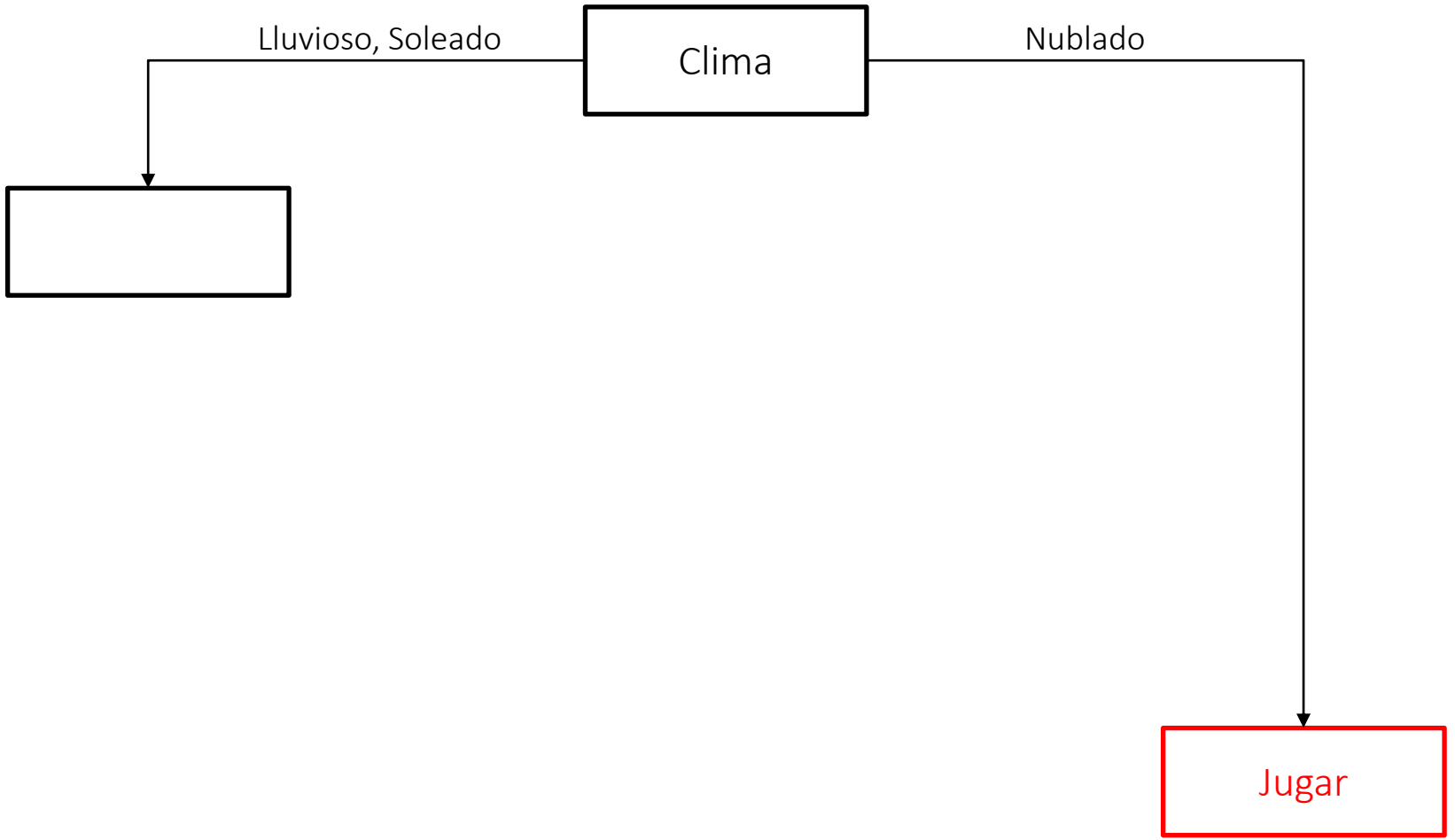
Clima soleado	$G = 0{,}035$
Clima nublado	$G = 0{,}396$
Clima lluvioso	$G = 0{,}125$
Temperatura calurosa	$G = 0$
Temperatura agradable	$G = 0{,}029$
Temperatura fría	$G = 0{,}108$
Humedad	$G = 0{,}035$
Viento	$G = 0{,}125$

elegimos este

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Nublado	Calor	Alta	Débil	Sí
Nublado	Agradable	Alta	Fuerte	Sí
Nublado	Calor	Normal	Débil	Sí

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Soleado	Agradable	Normal	Fuerte	Sí
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Frío	Normal	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Soleado	Calor	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Fuerte	No

Primera clasificación



Entropía del primer nodo

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Soleado	Agradable	Normal	Fuerte	Sí
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Frío	Normal	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Soleado	Calor	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Fuerte	No

$$S = -\frac{2}{7} \cdot \log_2\left(\frac{2}{7}\right) - \frac{5}{7} \cdot \log_2\left(\frac{5}{7}\right) = 0{,}863$$

Clasificando según clima

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Soleado	Agradable	Normal	Fuerte	Sí
Soleado	Calor	Alta	Fuerte	No

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Frío	Normal	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Lluvioso	Agradable	Alta	Fuerte	No

$$S_1 = -\frac{1}{3}\cdot\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\cdot\log_2\left(\frac{2}{3}\right) = 0{,}918$$

$$S_2 = -\frac{1}{4}\cdot\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\cdot\log_2\left(\frac{3}{4}\right) = 0{,}811$$

$$G = 0{,}863 - \frac{3}{7}\cdot 0{,}918 - \frac{4}{7}\cdot 0{,}985 = 0{,}006$$

Clasificando según temperatura calurosa

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Soleado	Calor	Alta	Fuerte	No

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Agradable	Normal	Fuerte	Sí
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Frío	Normal	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Lluvioso	Agradable	Alta	Fuerte	No

$$S_1 = -\frac{0}{2}\cdot\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\cdot\log_2\left(\frac{2}{2}\right) = 0$$

$$S_2 = -\frac{2}{5}\cdot\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\cdot\log_2\left(\frac{3}{5}\right) = 0{,}971$$

$$G = 0{,}863 - \frac{2}{7}\cdot 0 - \frac{5}{7}\cdot 0{,}971 = 0{,}170$$

Clasificando según temperatura agradable

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Agradable	Normal	Fuerte	Sí
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Lluvioso	Agradable	Alta	Fuerte	No

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Lluvioso	Frío	Normal	Fuerte	No
Soleado	Calor	Alta	Fuerte	No

$$S_1 = -\frac{2}{4}\cdot\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\cdot\log_2\left(\frac{2}{4}\right) = 1$$

$$S_2 = -\frac{0}{3}\cdot\log_2\left(\frac{0}{3}\right) - \frac{3}{3}\cdot\log_2\left(\frac{3}{3}\right) = 0$$

$$G = 0{,}863 - \frac{4}{7}\cdot 1 - \frac{3}{7}\cdot 0 = 0{,}291$$

Clasificando según temperatura fría

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Lluvioso	Frío	Normal	Fuerte	No

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Soleado	Agradable	Normal	Fuerte	Sí
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Soleado	Calor	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Fuerte	No

$$S_1 = -\frac{0}{1}\cdot\log_2\left(\frac{0}{1}\right) - \frac{1}{1}\cdot\log_2\left(\frac{1}{1}\right) = 0$$

$$S_2 = -\frac{2}{6}\cdot\log_2\left(\frac{2}{6}\right) - \frac{4}{6}\cdot\log_2\left(\frac{4}{6}\right) = 0{,}918$$

$$G = 0{,}863 - \frac{1}{7}\cdot 0 - \frac{6}{7}\cdot 0{,}918 = 0{,}076$$

Clasificando según humedad

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Agradable	Normal	Fuerte	Sí
Lluvioso	Frío	Normal	Fuerte	No

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Soleado	Calor	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Fuerte	No

$$S_1 = -\frac{1}{2}\cdot\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\cdot\log_2\left(\frac{1}{2}\right) = 1$$

$$S_2 = -\frac{1}{5}\cdot\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\cdot\log_2\left(\frac{4}{5}\right) = 0{,}722$$

$$G = 0{,}863 - \frac{2}{7}\cdot 1 - \frac{5}{7}\cdot 0{,}722 = 0{,}006$$

Clasificando según viento

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Lluvioso	Agradable	Alta	Débil	Sí

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Agradable	Normal	Fuerte	Sí
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Frío	Normal	Fuerte	No
Soleado	Calor	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Fuerte	No

$$S_1 = -\frac{1}{2}\cdot\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\cdot\log_2\left(\frac{1}{2}\right) = 1$$

$$S_2 = -\frac{1}{5}\cdot\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\cdot\log_2\left(\frac{4}{5}\right) = 0{,}722$$

$$G = 0{,}863 - \frac{2}{7}\cdot 1 - \frac{5}{7}\cdot 0{,}722 = 0{,}006$$

Resumiendo

Clima	$G = 0{,}006$
Temperatura calurosa	$G = 0{,}170$
Temperatura agradable	$G = 0{,}291$
Temperatura fría	$G = 0{,}076$
Humedad	$G = 0{,}006$
Viento	$G = 0{,}006$

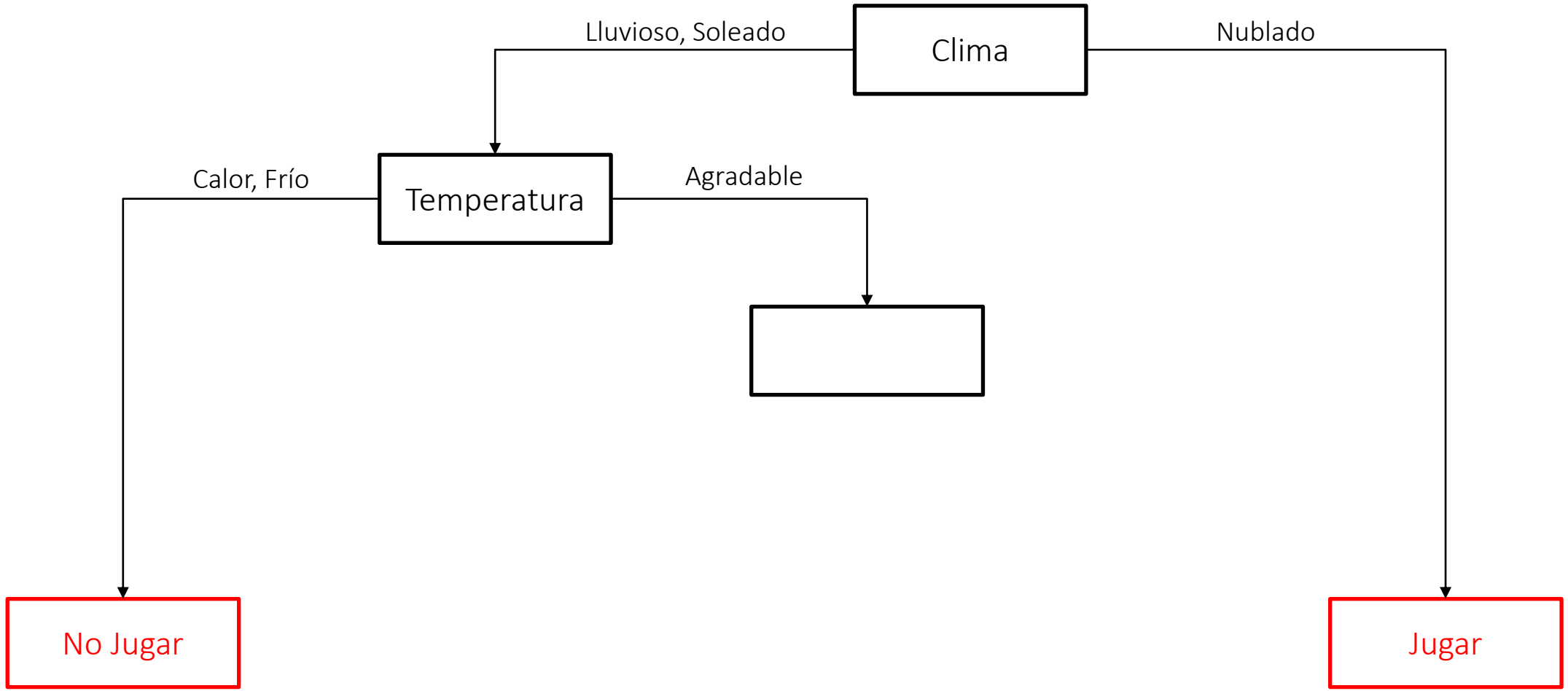
Resumiendo

Clima	$G = 0{,}006$	
Temperatura calurosa	$G = 0{,}170$	
Temperatura agradable	$G = 0{,}291$	elegimos este
Temperatura fría	$G = 0{,}076$	
Humedad	$G = 0{,}006$	
Viento	$G = 0{,}006$	

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Agradable	Normal	Fuerte	Sí
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Lluvioso	Agradable	Alta	Fuerte	No

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
Lluvioso	Frío	Normal	Fuerte	No
Soleado	Calor	Alta	Fuerte	No

Segunda clasificación



Entropía del segundo nodo

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Agradable	Normal	Fuerte	Sí
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Lluvioso	Agradable	Alta	Fuerte	No

$$S = -\frac{2}{4} \cdot \log_2\left(\frac{2}{4}\right) - \frac{2}{4} \cdot \log_2\left(\frac{2}{4}\right) = 1$$

Clasificando según clima o humedad

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Agradable	Normal	Fuerte	Sí

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Lluvioso	Agradable	Alta	Fuerte	No

$$S_1 = -\frac{1}{1}\cdot\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\cdot\log_2\left(\frac{0}{1}\right) = 0$$

$$S_2 = -\frac{1}{3}\cdot\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\cdot\log_2\left(\frac{2}{3}\right) = 0{,}918$$

$$G = 1 - \frac{1}{4}\cdot 0 - \frac{3}{4}\cdot 0{,}918 = 0{,}311$$

Clasificando según viento

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Lluvioso	Agradable	Alta	Débil	Sí

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Agradable	Normal	Fuerte	Sí
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Fuerte	No

$$S_1 = -\frac{1}{1}\cdot\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\cdot\log_2\left(\frac{0}{1}\right) = 0$$

$$S_2 = -\frac{1}{3}\cdot\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\cdot\log_2\left(\frac{2}{3}\right) = 0{,}918$$

$$G = 1 - \frac{1}{4}\cdot 0 - \frac{3}{4}\cdot 0{,}918 = 0{,}311$$

Resumiendo

Clima	$G = 0{,}311$
Humedad	$G = 0{,}311$
Viento	$G = 0{,}311$

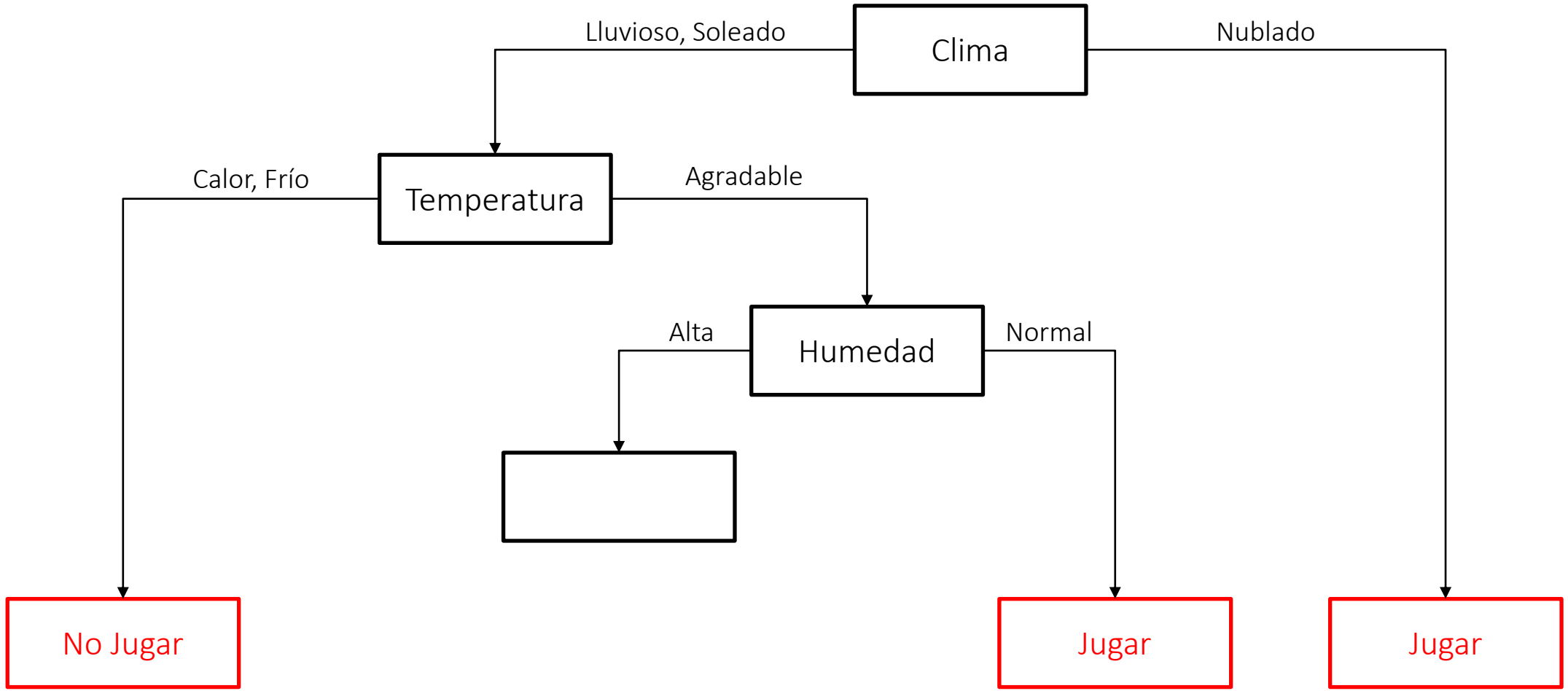
Resumiendo

Clima	$G = 0{,}311$	
Humedad	$G = 0{,}311$	elegimos cualquiera
Viento	$G = 0{,}311$	

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Agradable	Normal	Fuerte	Sí

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Lluvioso	Agradable	Alta	Fuerte	No

Tercera clasificación



Entropía del tercer nodo

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Débil	Sí
Lluvioso	Agradable	Alta	Fuerte	No

$$S = -\frac{1}{3} \cdot \log_2\left(\frac{1}{3}\right) - \frac{2}{3} \cdot \log_2\left(\frac{2}{3}\right) = 0{,}918$$

Clasificando según viento

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Lluvioso	Agradable	Alta	Débil	Sí

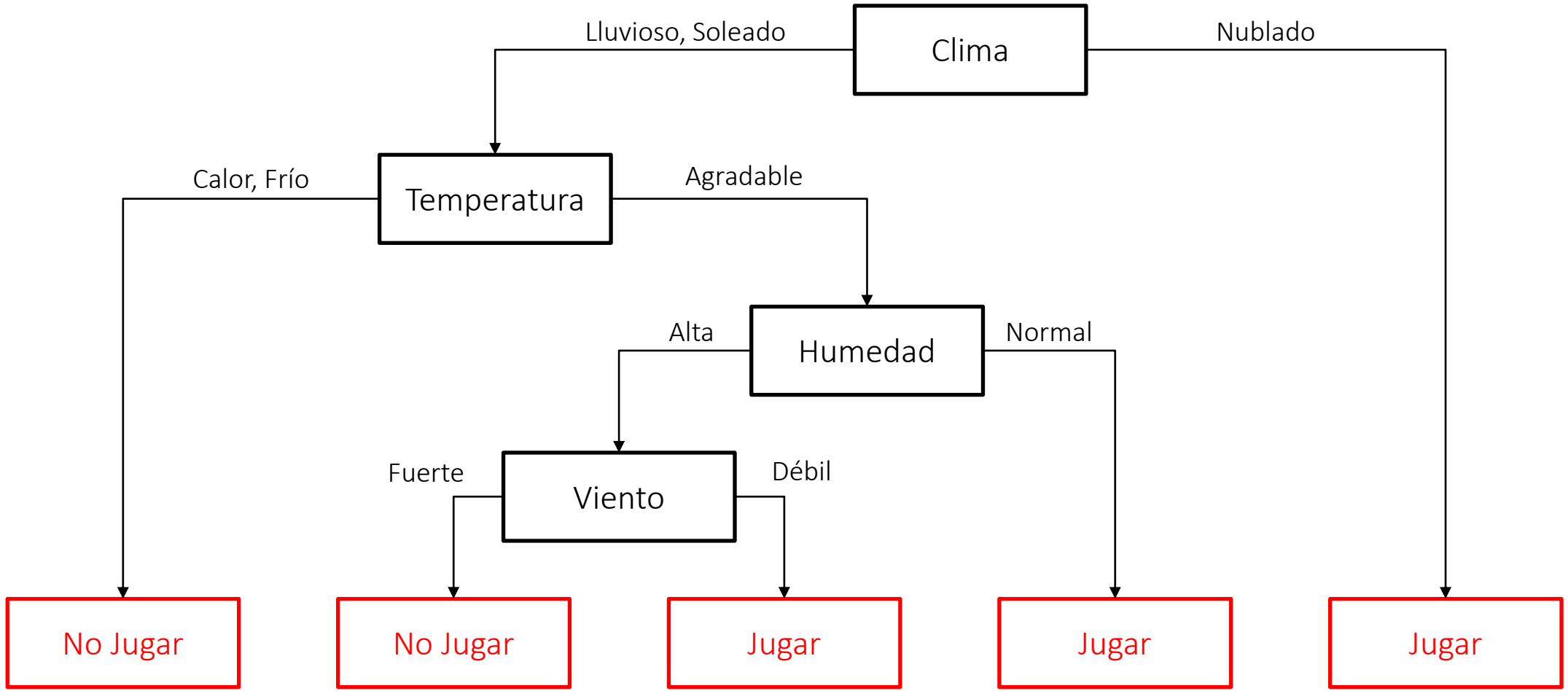
Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Lluvioso	Agradable	Alta	Fuerte	No
Lluvioso	Agradable	Alta	Fuerte	No

$$S_1 = -\frac{1}{1}\cdot\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\cdot\log_2\left(\frac{0}{1}\right) = 0$$

$$S_2 = -\frac{0}{2}\cdot\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\cdot\log_2\left(\frac{2}{2}\right) = 0$$

$$G = 0{,}918 - \frac{1}{3}\cdot 0 - \frac{2}{3}\cdot 0 = 0{,}918$$

Árbol de clasificación resultante



Criterios de parada

¿Cuándo termina el método?

Existen dos criterios de parada estricta:

Cuando todos los registros de un nodo pertenecen a la misma clase

Cuando todos los registros de un nodo poseen iguales atributos

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Fuerte	Sí
Nublado	Calor	Normal	Débil	Sí
Lluvioso	Agradable	Alta	Fuerte	Sí

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Fuerte	No
Soleado	Calor	Alta	Fuerte	Sí
Soleado	Calor	Alta	Fuerte	No

¿Cómo saber cuándo conviene parar de crecer el árbol?

Podemos minimizar el costo de complejidad del árbol:

$$\sum_{i \in \text{Nodos Terminales}} \text{Tasa de Clasificación Errónea en } i + \lambda \cdot \text{Divisiones}$$

Árboles más grandes tendrán menor tasa de clasificación errónea y más divisiones

λ es un parámetro de complejidad que define el modelador

¿Cómo saber cuándo conviene parar de crecer el árbol?

Otros criterios:

Profundidad máxima del árbol

Cantidad mínima de observaciones en un nodo a dividir

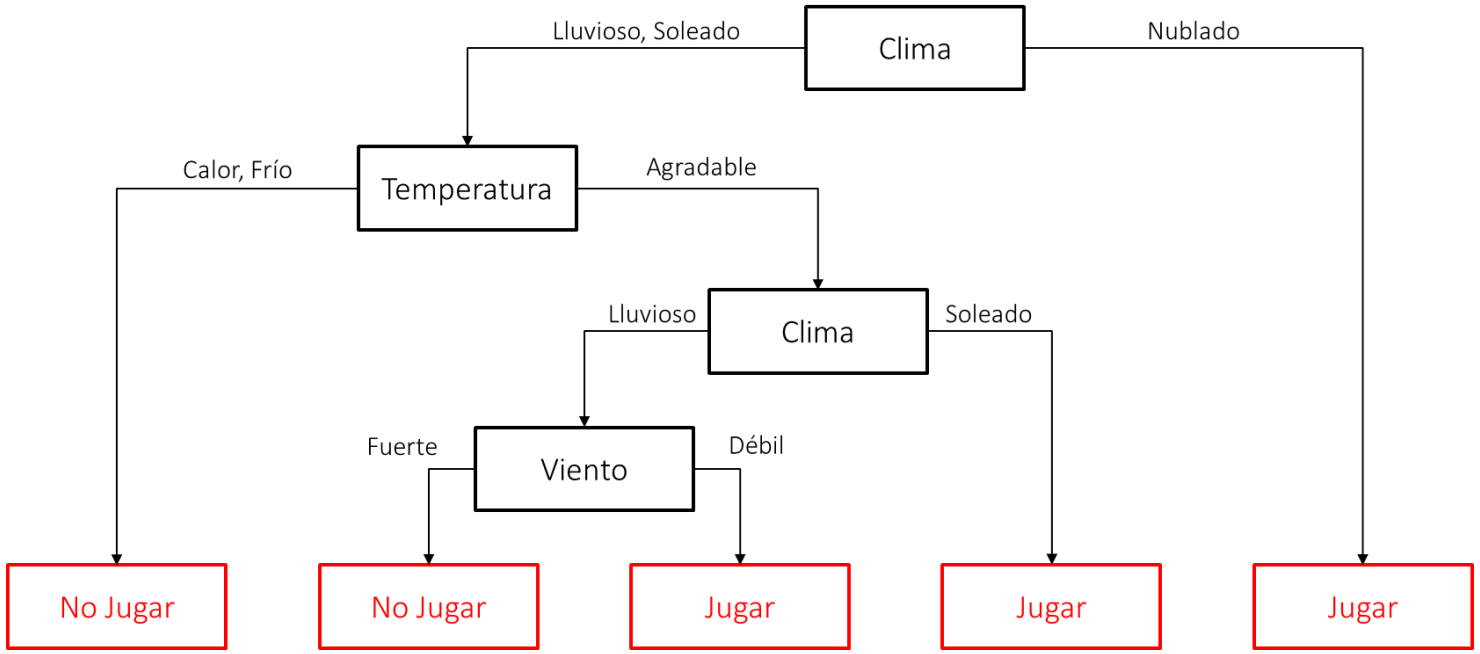
Cantidad mínima de observaciones en un nodo terminal

Aspectos a considerar

Aspectos a considerar

Un atributo puede ser usado más de una vez, incluso dentro de una misma rama

En nuestro ejemplo podríamos haber elegido:



Aspectos a considerar

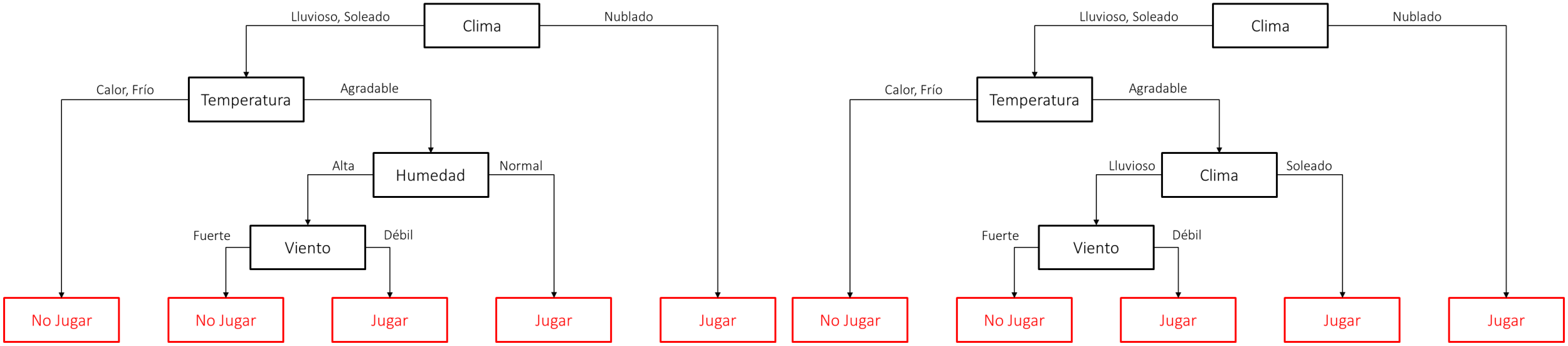
Elecciones entre ramas con la misma ganancia pueden afectar los resultados finales

El método es codicioso (*greedy*), buscando mejoras locales

Divisiones realizadas en nodos superiores no cambiarán en función de divisiones posteriores

Aspectos a considerar

Elecciones entre ramas con la misma ganancia pueden afectar la capacidad predictiva, aún cuando la clasificación dentro de la muestra de entrenamiento sea la misma



Aspectos a considerar

Aún cuando el árbol tenga una profundidad máxima, no siempre será posible reducir la tasa de clasificación errónea de cada nodo terminal a 0

Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
Soleado	Calor	Alta	Débil	No
⋮	⋮	⋮	⋮	⋮
Soleado	Agradable	Normal	Fuerte	Sí
Soleado	Agradable	Normal	Fuerte	Sí
Soleado	Agradable	Normal	Fuerte	No
⋮	⋮	⋮	⋮	⋮
Lluvioso	Agradable	Alta	Fuerte	No

Aspectos a considerar

Existen criterios de clasificación distintos a ganancia de información:

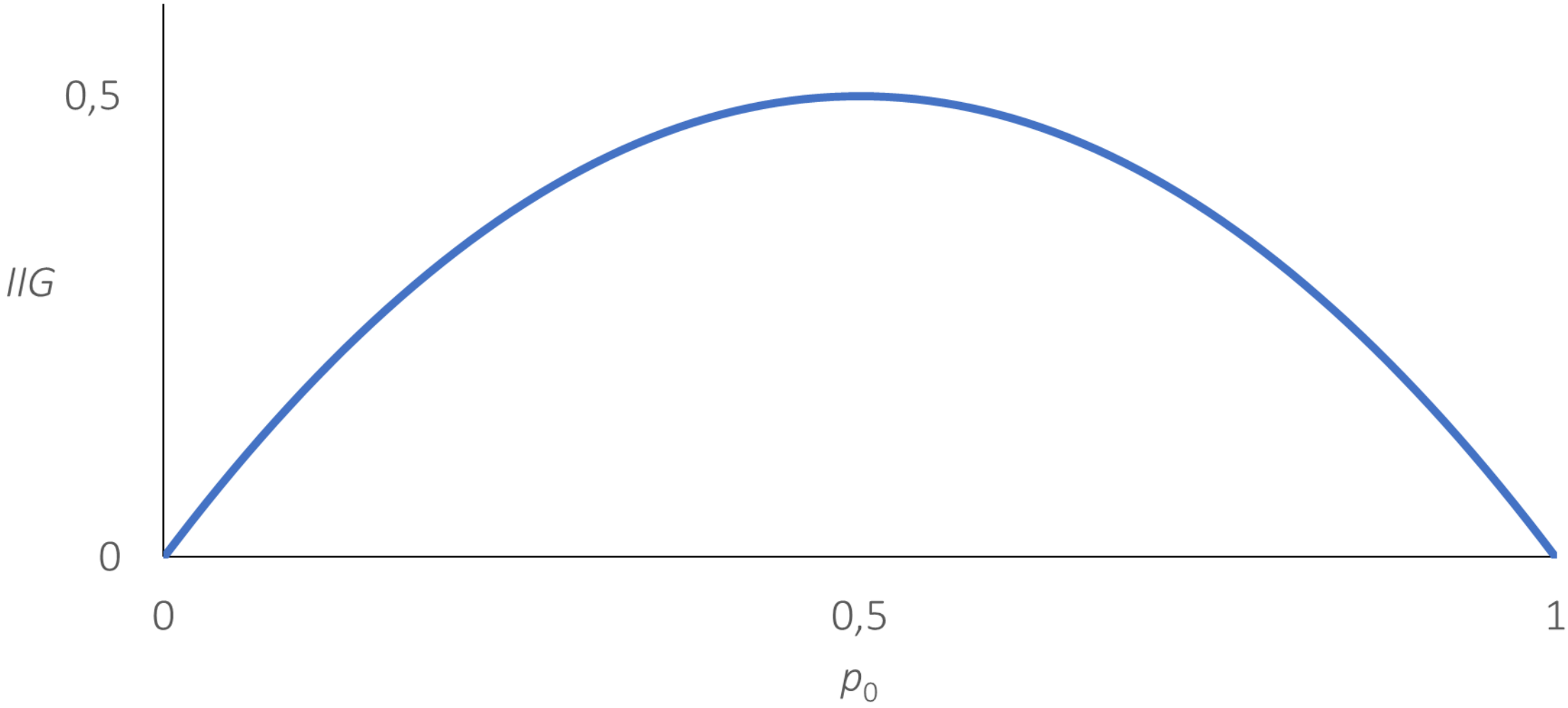
Exógena (predictora)	Endógena (a predecir)
B	1
A	1
B	1
B	0
A	0
A	0
A	0

Índice de Impuridad de Gini

$$IIG = p_0 \cdot \left(1 - p_0\right) + p_1 \cdot \left(1 - p_1\right)$$

$$IIG = \frac{3}{7} \cdot \left(1 - \frac{3}{7}\right) + \frac{4}{7} \cdot \left(1 - \frac{4}{7}\right) = 0{,}490$$

Índice de Impuridad de Gini



Índice de Impuridad de Gini

¿Cuánto ganamos usando nuestra variable endógena para dividir?

Exógena (predictora)	Endógena (a predecir)
A	1
A	0
A	0
A	0

Exógena (predictora)	Endógena (a predecir)
B	1
B	1
B	0

$$IIG_A = \frac{1}{4} \cdot \left(1 - \frac{1}{4}\right) + \frac{3}{4} \cdot \left(1 - \frac{3}{4}\right) = 0{,}375$$

$$IIG_B = \frac{2}{3} \cdot \left(1 - \frac{2}{3}\right) + \frac{1}{3} \cdot \left(1 - \frac{1}{3}\right) = 0{,}444$$

$$\text{Ganancia} = 0{,}490 - \frac{4}{7} \cdot 0{,}375 - \frac{3}{7} \cdot 0{,}444 = 0{,}085$$

Extensión a casos multiclase

Ganancia de información

$$S = -\sum_{k=1}^{K} p_k \cdot \log_2 \left(p_k \right)$$

Índice de Impuridad de Gini

$$IIG = \sum_{k=1}^{K} p_k \cdot \left(1 - p_k \right)$$

Árboles de regresión

Análisis de varianza (ANOVA)

En cada nodo se ajusta un valor “promedio” para la variable continua a analizar

En cada nodo buscamos minimizar la varianza (i.e. suma de los errores cuadráticos) del valor ajustado en cada división

$$\text{Min} \quad SSE = \sum_{i \in R_0} \left(y_i - c_0 \right)^2 + \sum_{i \in R_1} \left(y_i - c_1 \right)^2$$

Análisis de varianza (ANOVA)

Supongamos un nodo con los siguientes registros:

Exógena (predictora)	Endógena (a predecir)
B	10
A	5
B	8
B	4
A	6
A	1
A	2

$$c = \frac{10 + 5 + \ldots + 2}{7} = 5{,}14$$

$$SSE = (10 - 5{,}14)^2 + (5 - 5{,}14)^2 + \ldots + (2 - 5{,}14)^2$$
$$= 60{,}86$$

Análisis de varianza (ANOVA)

¿Cuánto ganamos usando nuestra variable endógena para dividir?

Exógena (predictora)	Endógena (a predecir)
A	5
A	6
A	1
A	2

Exógena (predictora)	Endógena (a predecir)
B	10
B	8
B	4

$c_A = 3{,}5$ $SSE_A = 17$

$c_B = 7{,}33$ $SSE_B = 18{,}67$

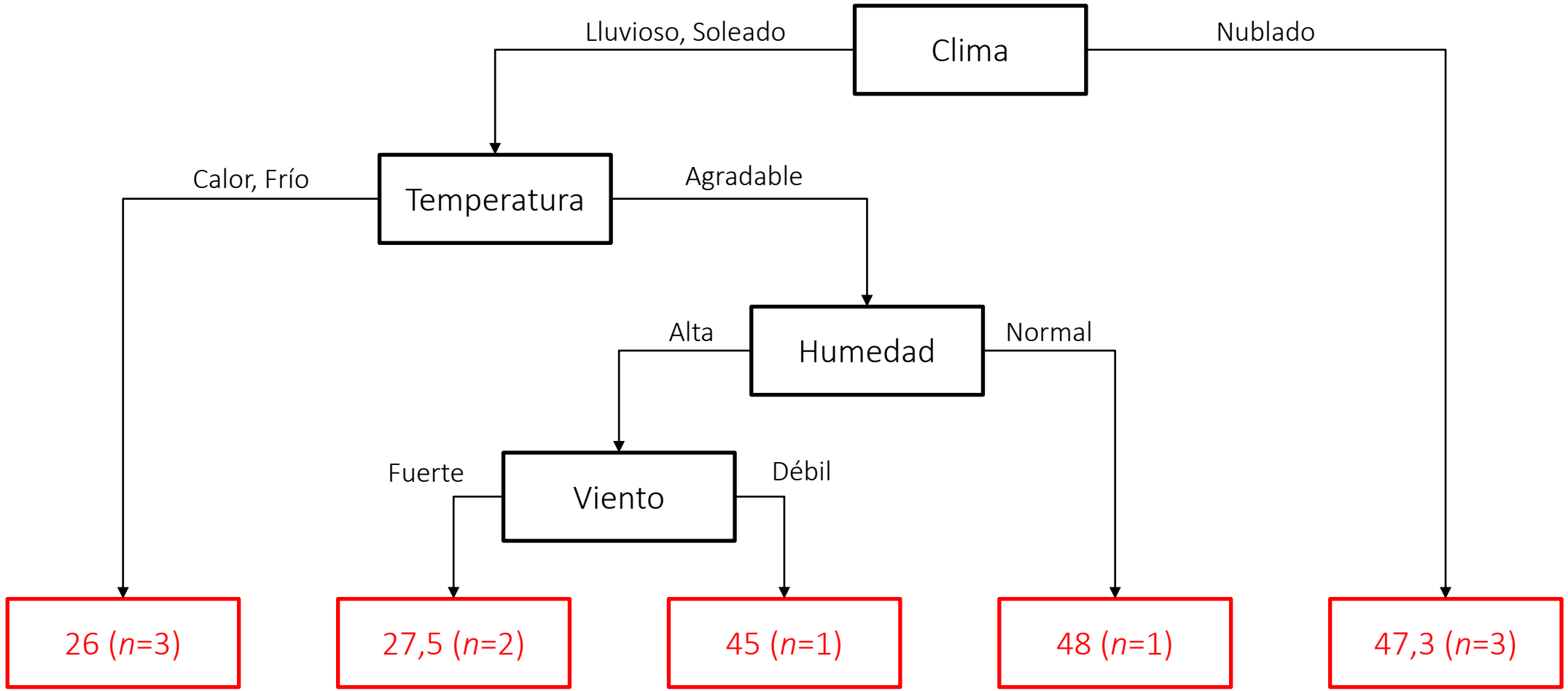
$$SSE = 17 + 18{,}67 = 35{,}67$$

$$\text{Ganancia} = 60{,}86 - 35{,}67 = 25{,}19$$

Consideremos los siguientes datos

Clima	Temperatura	Humedad	Viento	Jugadores
Soleado	Calor	Alta	Débil	25
Nublado	Calor	Alta	Débil	46
Soleado	Agradable	Normal	Fuerte	48
Nublado	Agradable	Alta	Fuerte	52
Lluvioso	Agradable	Alta	Fuerte	25
Lluvioso	Frío	Normal	Fuerte	23
Lluvioso	Agradable	Alta	Débil	45
Soleado	Calor	Alta	Fuerte	30
Nublado	Calor	Normal	Débil	44
Lluvioso	Agradable	Alta	Fuerte	30

Árbol de regresión resultante (aplicando $\lambda^* = 0{,}03$ en R)



Árbol de regresión resultante (aplicando $\lambda^* = 0{,}18$ en R)