Aprendizaje de máquina

Árboles de decisión

Ingeniería

Árboles de decisión

- Los métodos basados en árboles sirven tanto para problemas de regresión como de clasificación.
- Estos implican estratificar o segmentar el espacio predictor en una serie de regiones simples.
- Como el conjunto de reglas de división utilizadas para segmentar el espacio predictivo se puede resumir en un árbol, este tipo de enfoque se conoce como árbol de decisión.
- Dada una base de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

Pros y contras

- Los métodos basados en árboles son simples e interpretables.
- Sin embargo, generalmente no son competitivos con los mejores enfoques de aprendizaje supervisado en términos de precisión de predicción.
- La combinación de una gran cantidad de árboles a menudo puede resultar en mejoras dramáticas en la precisión de la predicción, a expensas de alguna pérdida de interpretación.

Ejemplo: Conjunto de datos

- Predecir si Juan jugará futbol

Ejemplos:
9 Sí / 5 No

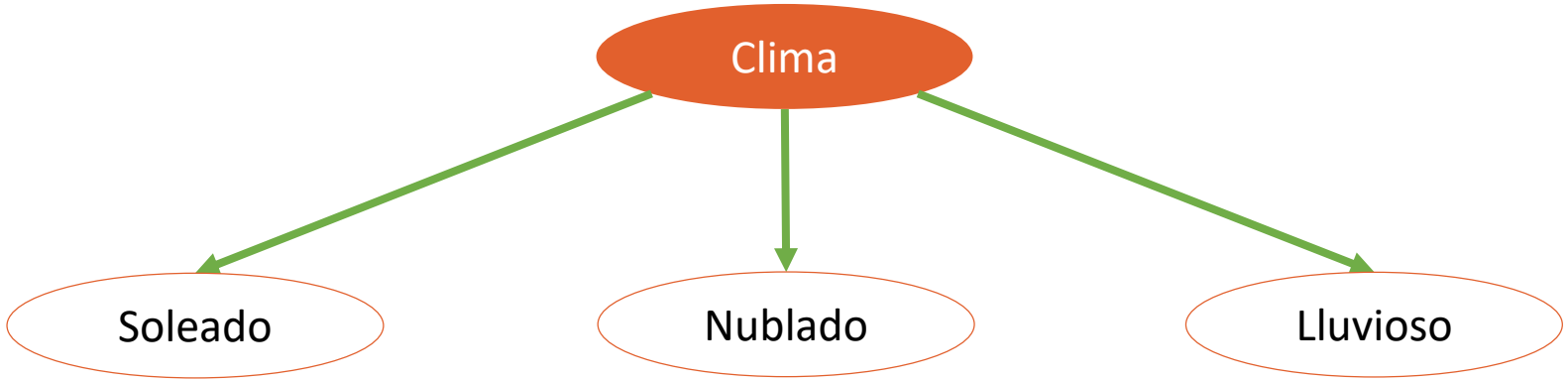
Día	Clima	Humedad	Viento	Jugó
D1	Soleado	Alta	Poco	No
D2	Soleado	Alta	Fuerte	No
D3	Nublado	Alta	Poco	Sí
D4	Lluvioso	Alta	Poco	Sí
D5	Lluvioso	Normal	Poco	Sí
D6	Lluvioso	Normal	Fuerte	No
D7	Nublado	Normal	Fuerte	Sí
D8	Soleado	Alta	Poco	No
D9	Soleado	Normal	Poco	Sí
D10	Lluvioso	Normal	Poco	Sí
D11	Soleado	Normal	Fuerte	Sí
D12	Nublado	Alta	Fuerte	Sí
D13	Nublado	Normal	Poco	Sí
D14	Lluvioso	Alta	Fuerte	No

Ejemplo: Definición del problema

- Si el día 15 está lloviendo, la humedad es alta y hay poco viento, ¿Juan jugará o no?
  - Es difícil adivinar simplemente viendo los datos
  - Podemos seguir una estrategia divide y vencerás
    - Dividir en subconjuntos
    - Si no hay incertidumbre en el subconjunto paramos, si no seguimos dividiendo
    - Verificamos en que subconjunto cae la pregunta que estamos realizando

Ejemplo: Creación del árbol

4 registros clasificados correctamente



Día	Clima	Humedad	Viento
D1	Soleado	Alta	Poco
D2	Soleado	Alta	Fuerte
D8	Soleado	Alta	Poco
D9	Soleado	Normal	Poco
D11	Soleado	Normal	Fuerte

2 Sí / 3 No
Subdividir más

Día	Clima	Humedad	Viento
D3	Nublado	Alta	Poco
D7	Nublado	Normal	Fuerte
D12	Nublado	Alta	Fuerte
D13	Nublado	Normal	Poco

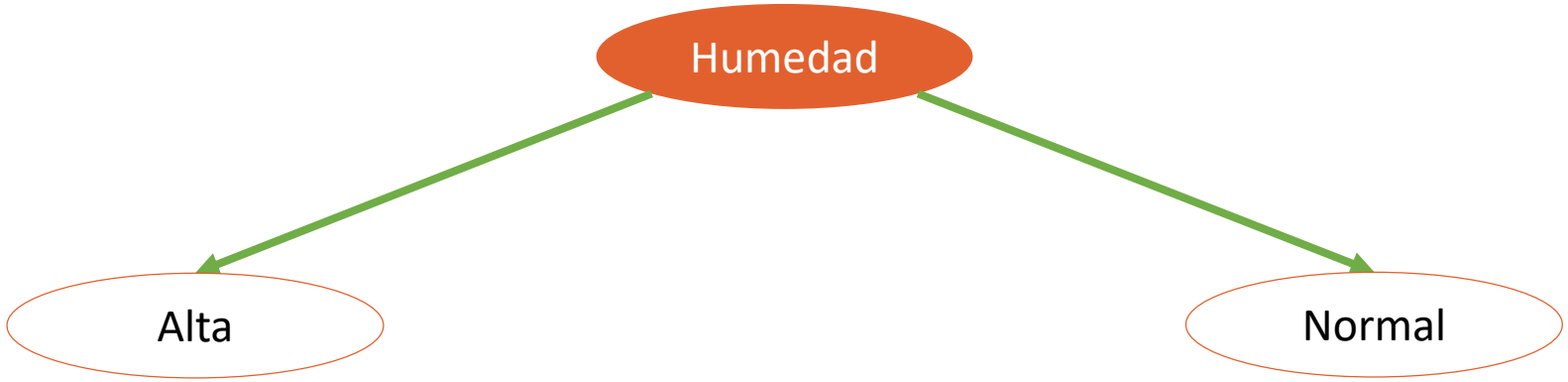
4 Sí / 0 No
Subconjunto puro

Día	Clima	Humedad	Viento
D4	Lluvioso	Alta	Poco
D5	Lluvioso	Normal	Poco
D6	Lluvioso	Normal	Fuerte
D10	Lluvioso	Normal	Poco
D14	Lluvioso	Alta	Fuerte

3 Sí / 2 No
Subdividir más

Ejemplo: Creación del árbol

0 registros clasificados correctamente



Día	Clima	Humedad	Viento
D1	Soleado	Alta	Poco
D2	Soleado	Alta	Fuerte
D3	Nublado	Alta	Poco
D4	Lluvioso	Alta	Poco
D8	Soleado	Alta	Poco
D12	Nublado	Alta	Fuerte
D14	Lluvioso	Alta	Fuerte

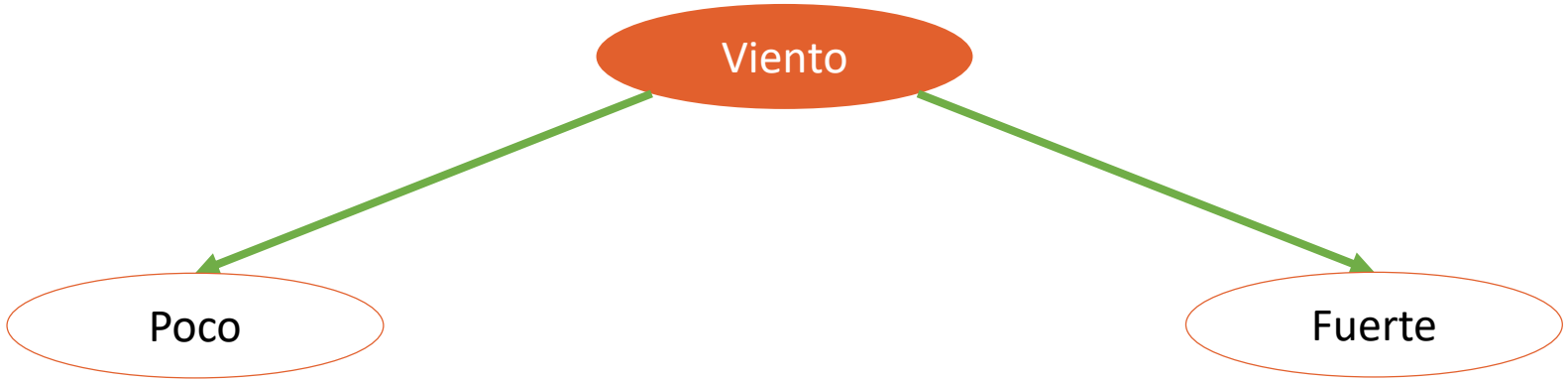
3 Sí / 4 No
Subdividir más

Día	Clima	Humedad	Viento
D5	Lluvioso	Normal	Poco
D6	Lluvioso	Normal	Fuerte
D7	Nublado	Normal	Fuerte
D9	Soleado	Normal	Poco
D10	Lluvioso	Normal	Poco
D11	Soleado	Normal	Fuerte
D13	Nublado	Normal	Poco

6 Sí / 1 No
Subdividir más

Ejemplo: Creación del árbol

0 registros clasificados correctamente



Día	Clima	Humedad	Viento
D1	Soleado	Alta	Poco
D3	Nublado	Alta	Poco
D4	Lluvioso	Alta	Poco
D5	Lluvioso	Normal	Poco
D8	Soleado	Alta	Poco
D9	Soleado	Normal	Poco
D10	Lluvioso	Normal	Poco
D13	Nublado	Normal	Poco

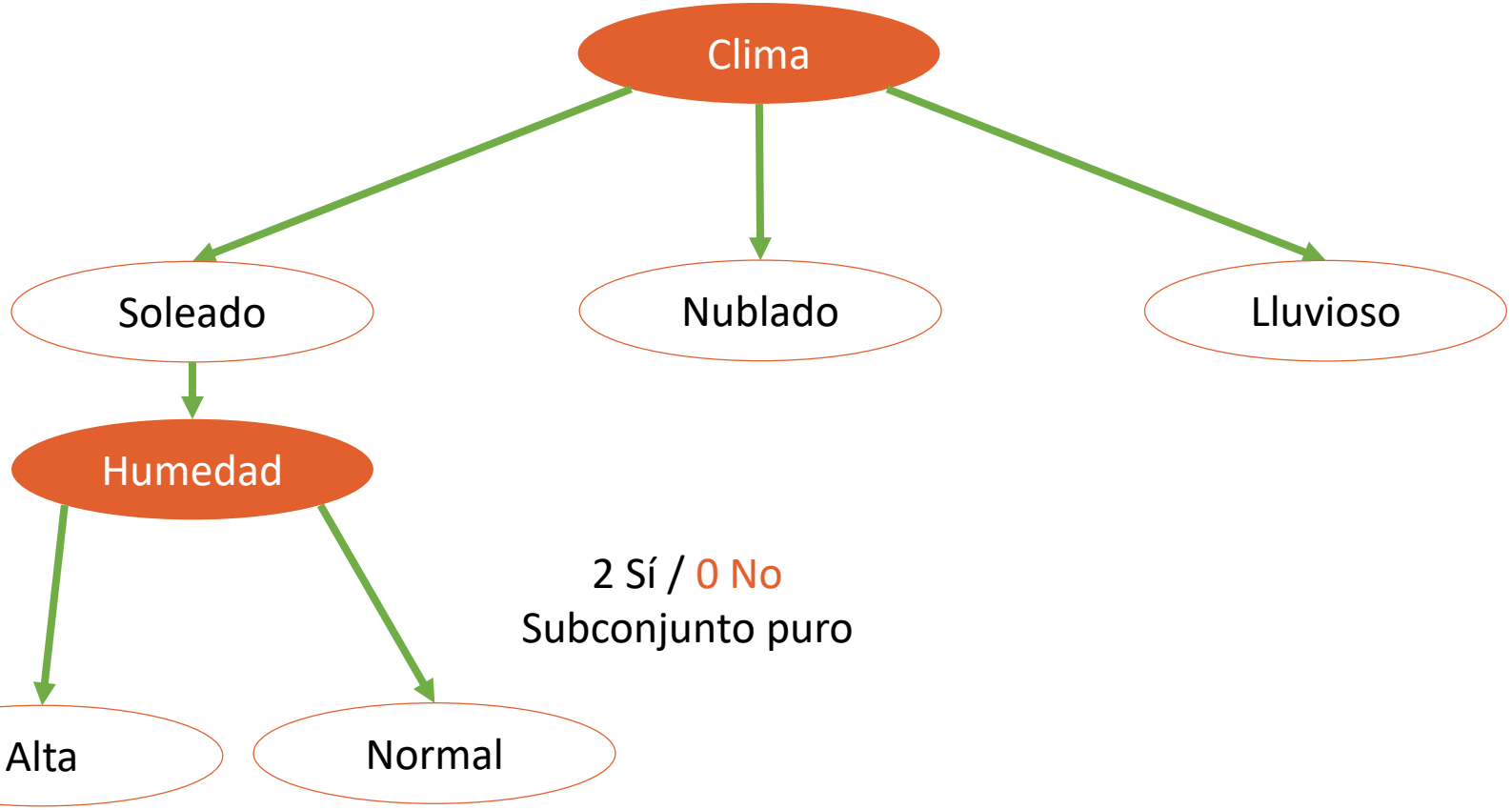
6 Sí / 2 No
Subdividir más

Día	Clima	Humedad	Viento
D2	Soleado	Alta	Fuerte
D6	Lluvioso	Normal	Fuerte
D7	Nublado	Normal	Fuerte
D11	Soleado	Normal	Fuerte
D12	Nublado	Alta	Fuerte
D14	Lluvioso	Alta	Fuerte

3 Sí / 3 No
Subdividir más

Ejemplo: Creación del árbol

5 registros clasificados correctamente



0 Sí / 3 No
Subconjunto puro

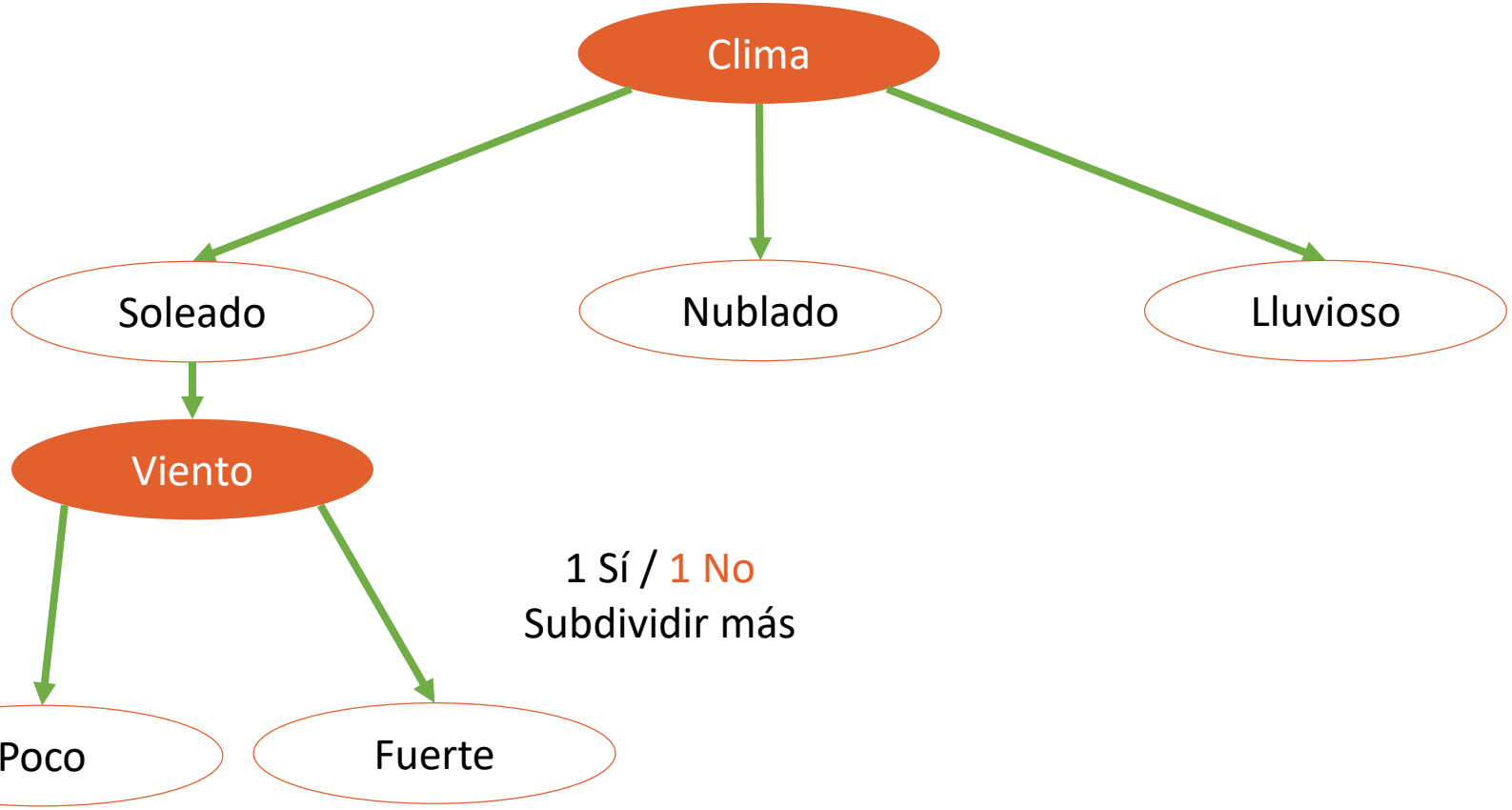
2 Sí / 0 No
Subconjunto puro

Día	Humedad	Viento
D1	Alta	Poco
D2	Alta	Fuerte
D8	Alta	Poco

Día	Humedad	Viento
D9	Normal	Poco
D11	Normal	Fuerte

Ejemplo: Creación del árbol

0 registros clasificados correctamente



1 Sí / 2 No
Subdividir más

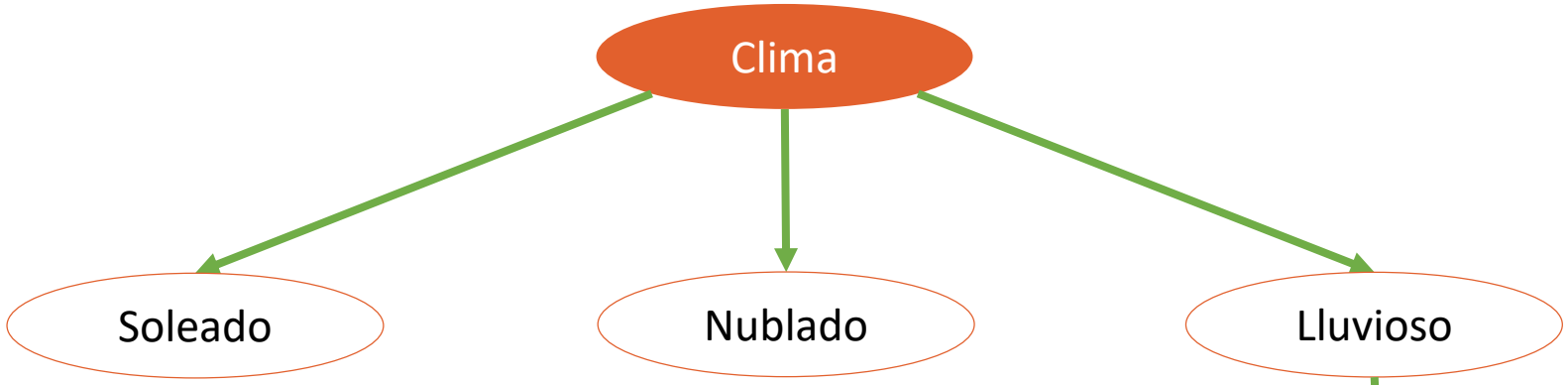
1 Sí / 1 No
Subdividir más

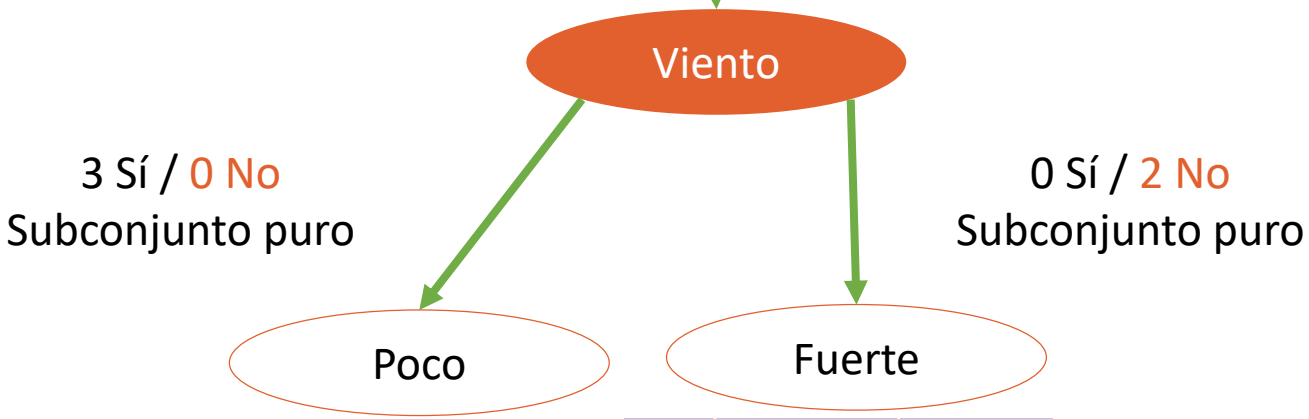
Día	Humedad	Viento
D1	Alta	Poco
D9	Normal	Poco
D8	Alta	Poco

Día	Humedad	Viento
D2	Alta	Fuerte
D11	Normal	Fuerte

Ejemplo: Creación del árbol

5 registros clasificados correctamente



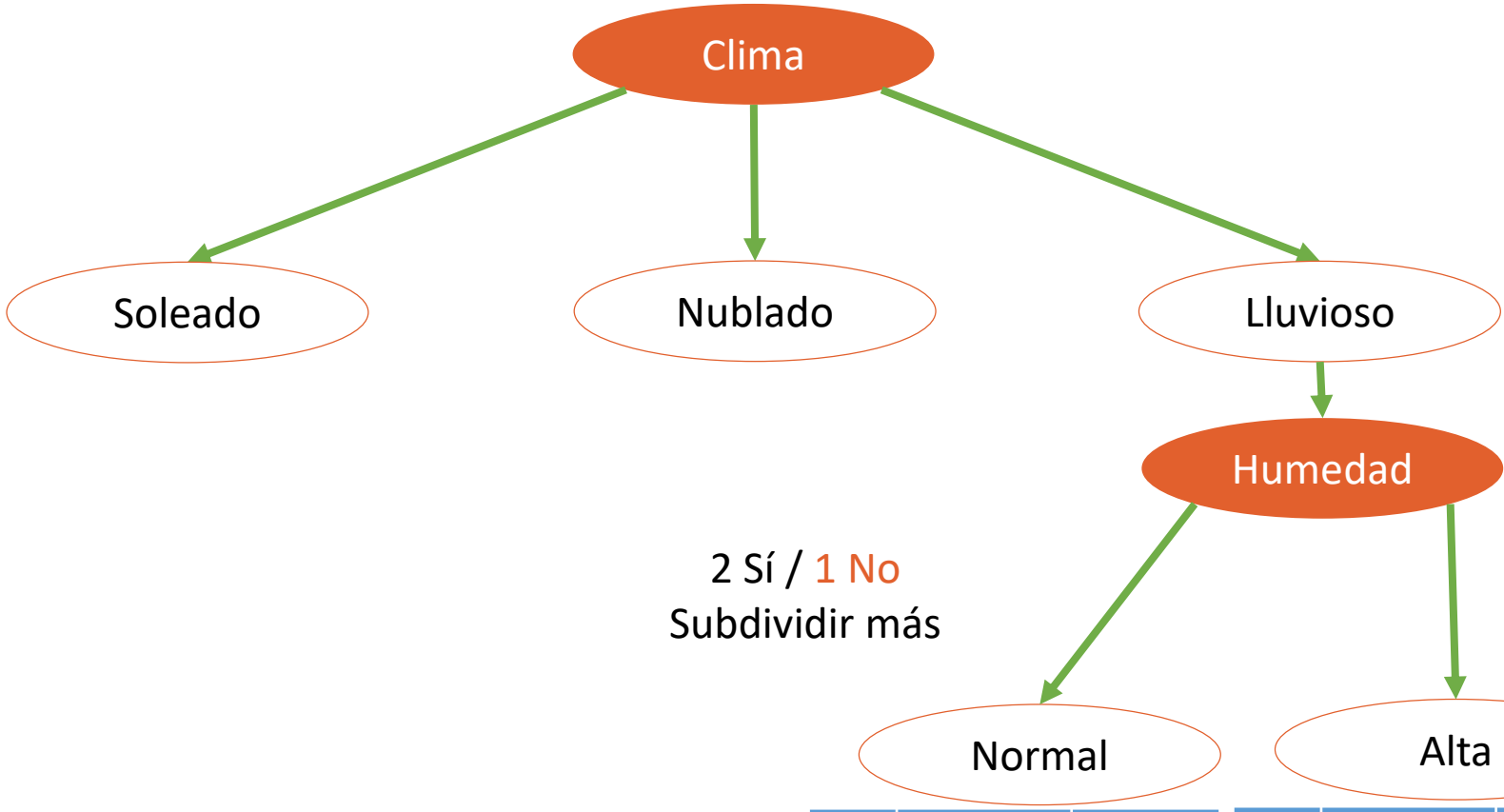


Día	Humedad	Viento
D4	Alta	Poco
D5	Normal	Poco
D10	Normal	Poco

Día	Humedad	Viento
D6	Normal	Fuerte
D14	Alta	Fuerte

Ejemplo: Creación del árbol

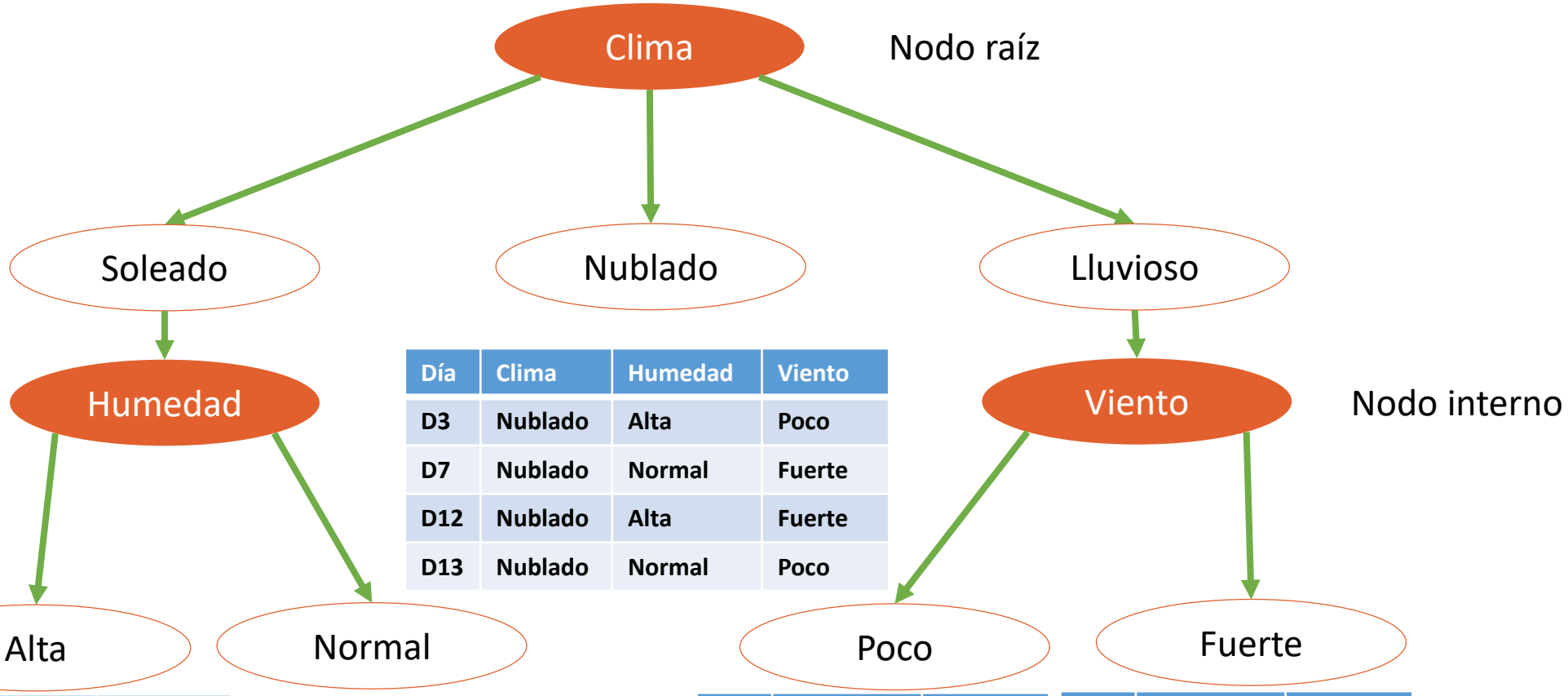
0 registros clasificados correctamente



Día	Humedad	Viento
D5	Normal	Poco
D6	Normal	Fuerte
D10	Normal	Poco

Día	Humedad	Viento
D4	Alta	Poco
D14	Alta	Fuerte

Ejemplo: Creación del árbol



Día	Clima	Humedad	Viento
D3	Nublado	Alta	Poco
D7	Nublado	Normal	Fuerte
D12	Nublado	Alta	Fuerte
D13	Nublado	Normal	Poco

Día	Humedad	Viento
D1	Alta	Poco
D2	Alta	Fuerte
D8	Alta	Poco

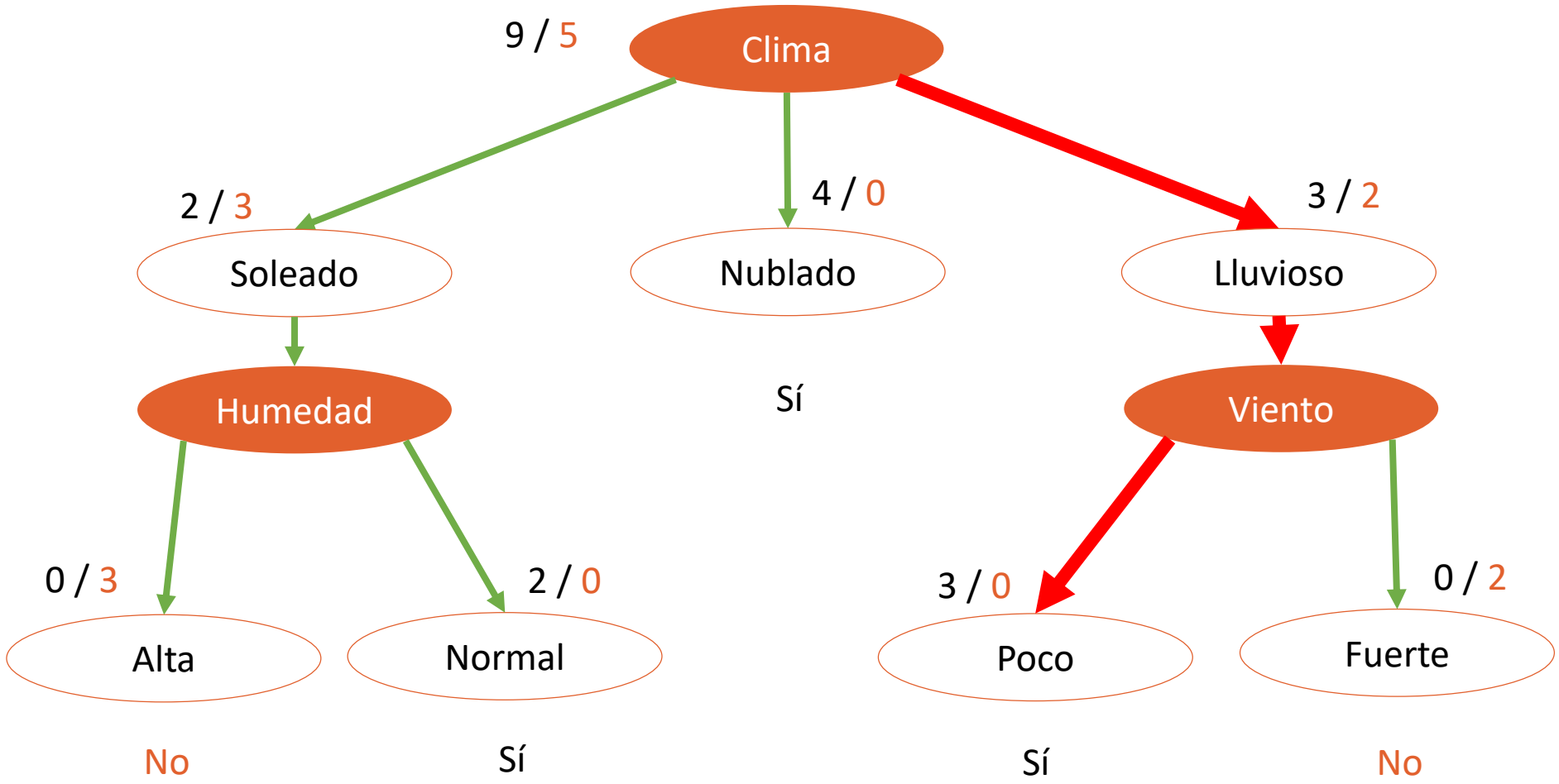
Día	Humedad	Viento
D9	Normal	Poco
D11	Normal	Fuerte

Día	Humedad	Viento
D4	Alta	Poco
D5	Normal	Poco
D10	Normal	Poco

Día	Humedad	Viento
D6	Normal	Fuerte
D14	Alta	Fuerte

Nodo Terminal - Hoja

Ejemplo: Árbol final



Juan si va a jugar

Detalles del proceso de creación del árbol

- Desafortunadamente, es computacionalmente inviable considerar cada posible partición del espacio de características en J regiones.
- Por esta razón, adoptamos un enfoque codicioso de arriba hacia abajo que se conoce como división binaria recursiva.
- El enfoque es de arriba hacia abajo porque comienza en la parte superior del árbol y luego divide sucesivamente el espacio predictor; cada división se indica a través de dos o más nuevas ramas más abajo en el árbol.
- Es codicioso porque en cada paso del proceso de construcción del árbol, la mejor división se realiza en ese paso en particular, en lugar de mirar hacia el futuro y elegir una división que conducirá a un mejor árbol en algún paso posterior.

Métricas para realizar división binaria

- Tasa de error de clasificación

$$E(S) = 1 - \max_k(p_k)$$

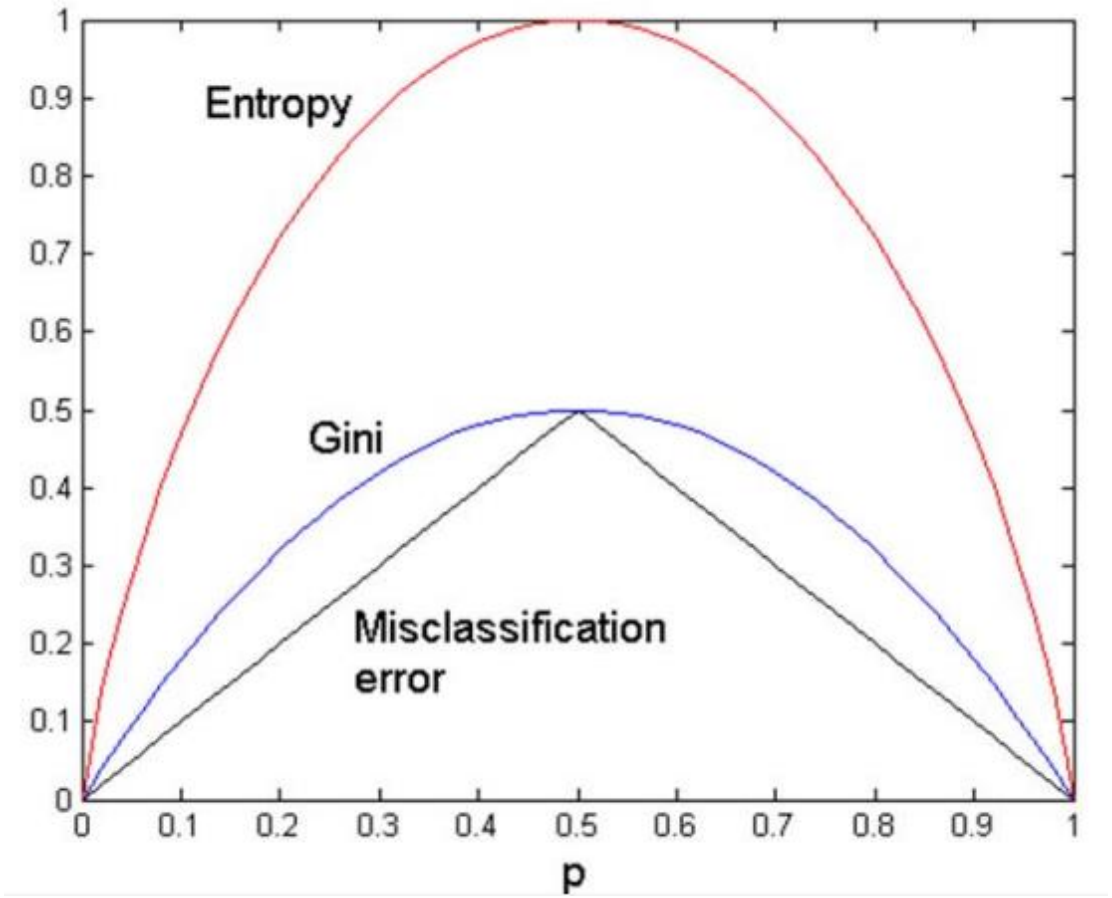
- Entropía

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

- Índice Gini

$$G(S) = \sum_{k=1}^{K} p_k(1 - p_k) = \sum_{k=1}^{K} p_k - \sum_{k=1}^{K} p_k^2 = 1 - \sum_{k=1}^{K} p_k^2$$

Métricas para realizar división binaria



Entropía

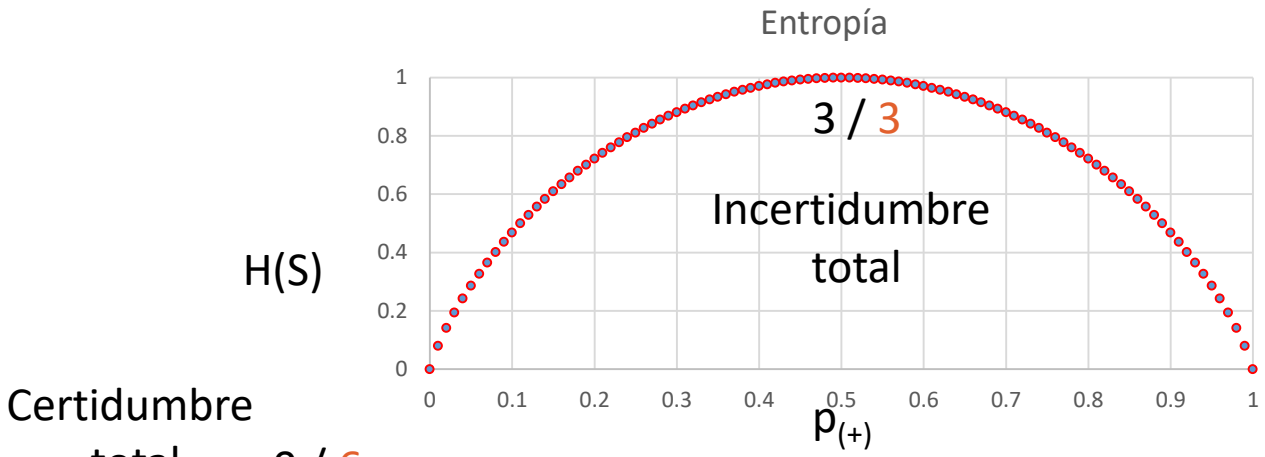
- Buscamos seleccionar el atributo con la mayor ganancia de información
- Este atributo minimiza la información requerida para clasificar los datos en la partición resultante y refleja la menor aleatoriedad o “impureza” en estas particiones

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

p_k es la probabilidad de que un registro arbitrario en S pertenezca a la clase C_k

Entropía

- ¿Cómo defino cuál es el mejor atributo para dividir?
  - Entropía: $H(S) = -p_{(+)}\log_2 p_{(+)} - p_{(-)}\log_2 p_{(-)}$
    - S: subconjunto de ejemplos de entrenamiento
    - $p_{(+)}$| $p_{(-)}$: % de ejemplos en s positivos | negativos



$por\ regla\ de\ L'Hopiltal$

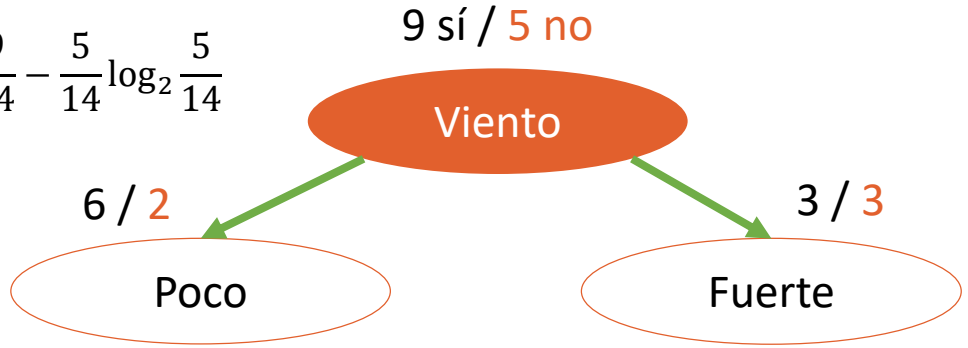
$0 \log_2 0 = 0$

Ganancia de información

- La caída esperada en entropía después de la división

$$IG(S,A) = H(S) - \sum_{v \in V(A)} \frac{|S_v|}{|S|} H(S_v)$$

$$H(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$
$$H(S) = 0.94$$



$$\text{IG}(S, Viento) = H(S) - \frac{8}{14}H(S_{poco}) - \frac{6}{14}H(S_{fuerte})$$
$$\text{IG}(S, Viento) = 0.048$$

$$H(S_{poco}) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}$$
$$H(S_{poco}) = 0.81$$

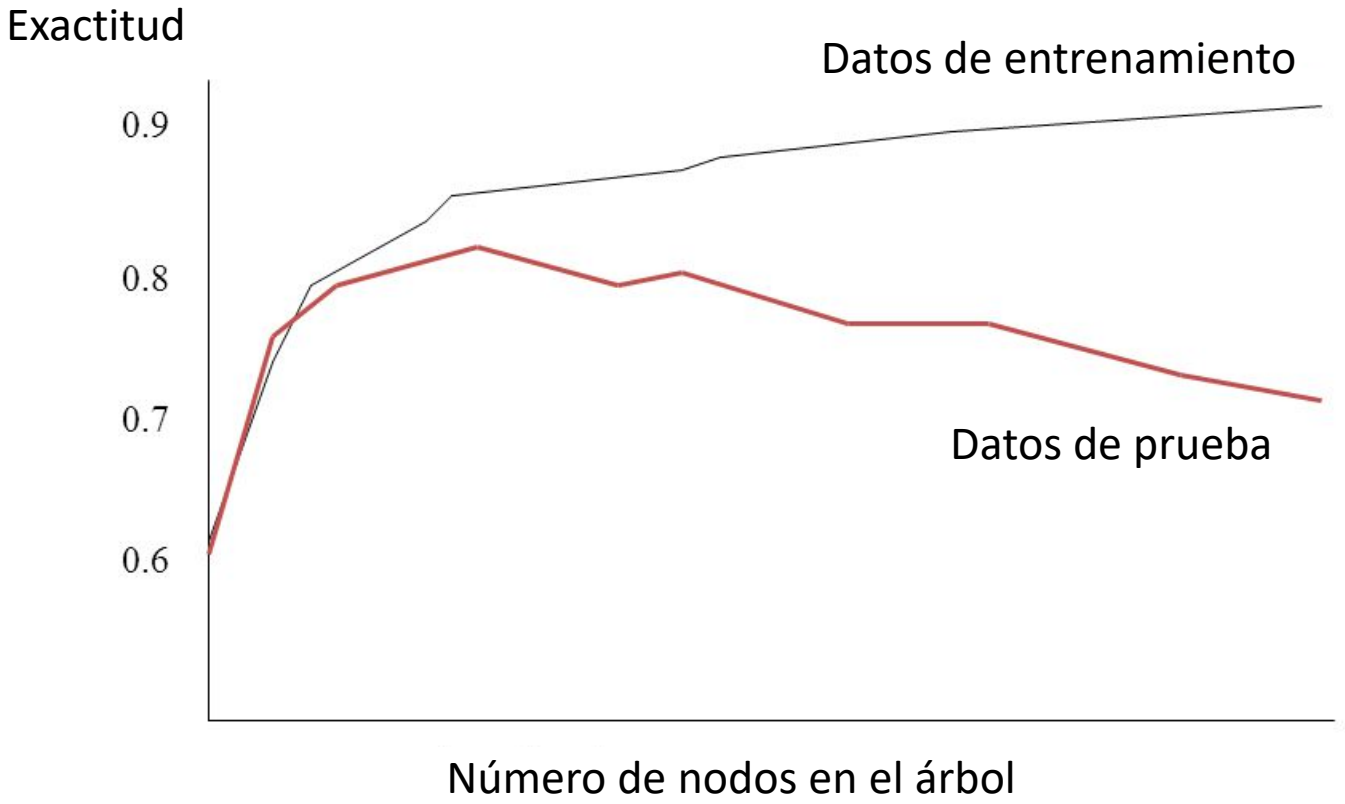
$$H(S_{fuerte}) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$$
$$H(S_{fuerte}) = 1.0$$

Algoritmos

- Algoritmos
  - Ross Quinlan (ID3 Iterative Dichotomiser: 1986), (C4.5: 1993)
  - Breiman et al (CaRT Classification and Regression Trees: 1984)

- ID3
  - Divide(nodo, {ejemplos})
    - A ← el mejor atributo para dividir {ejemplos}
    - Atributo decisor de nodo ← A
    - Para cada valor de A crea un nodo hijo
    - Divide {ejemplos} para cada nodo hijo en subconjuntos
    - Para cada nodo hijo / subconjunto:
      - Si subconjunto es puro: termina
      - Si no: Divide(nodo_hijo,{ subconjunto })

Sobre ajuste en árboles de decisión



Sin incertidumbre se pueden clasificar los datos de entrenamiento perfectamente

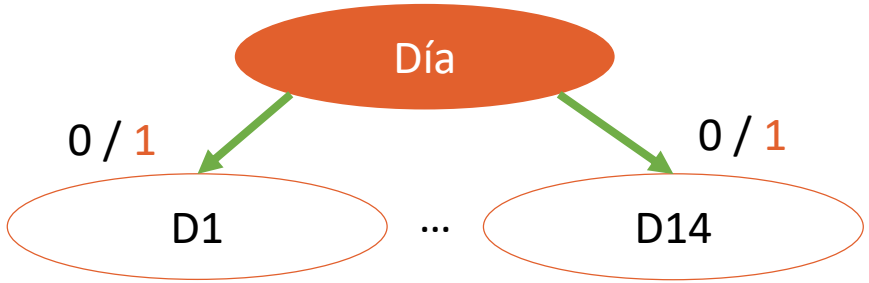
- Continua la división hasta tener un solo registro (singleton)
- Es bueno clasificando lo que ha visto en el pasado, pero no nuevos datos

Sobre ajuste en árboles de decisión

- La solución es no dejar crecer mucho los árboles:
  - Parar la división cuando no sea estadísticamente significativa (no hay atributos predictivos cuya correlación con la clase sea significativa)
  - Dejar que crezca y después se poda
    - Basado en los datos de prueba
    - Para cada nodo
      - Probar como se comporta el árbol sin un nodo y todos sus hijos
      - Medimos desempeño con los datos de prueba
    - Quitamos el nodo que resulte en la mejora más grande
    - Repetimos hasta que la poda empeore el desempeño.
  - Un árbol más pequeño con menos divisiones podría conducir a una menor varianza y una mejor interpretación a costa de un pequeño sesgo.

Tasa de ganancia de información (C4.5)

- Sesgo de la función ganancia de información por seleccionar variables con muchos valores



$$IV(S,A) = -\sum_{v \in V(A)} \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|}$$

$$IGR(S,A) = \frac{IG(S,A)}{IV(S,A)}$$

Variables continuas

- Para variables continuas:
  - Se divide a través de un umbral (temperatura > 72.3)
    - Punto de corte que genera la máxima ganancia
- Ordenamos los valores posibles del atributo A de menor a mayor
- El punto medio entre cada par adyacente se considera como un posible punto de corte
- Se procede a evaluar la entropía para cada corte y se selecciona la mínima

Clasificación multiclase

- Clasificación multiclase:
  - Se predice la clase más frecuente en el nodo que responde a la pregunta
  - Entropía:
  - p_c: % de ejemplos de la clase c en S

$$H(S) = -\sum_{c} p_c \log_2(p_c)$$

CaRT

- Considera cortes binarios óptimos para cada atributo
  - Por ejemplo si el atributo ingreso tiene tres posibles valores {bajo, medio, alto}, tendríamos 6 (2^n-2) subconjuntos
    - El -2 es debido a que se elimina el conjunto que tiene todos los elementos y el nulo dado que no representan corte alguno
  - Se realiza la pregunta si el atributo $A \in S_A$

CaRT

- Utiliza el índice Gini para evaluar la impureza de S

$$G(S) = 1 - \sum_{k=1}^{K} p_k^2$$

p_k es la probabilidad de que un registro arbitrario en S pertenezca a la clase C_k

$$IG(S, A) = G(S) - \frac{|S_1|}{|S|} G(S_1) - \frac{|S_2|}{|S|} G(S_2)$$

Ganancia de información

- La caída esperada en el índice Gini después de la división

$$G(S) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2$$

$$G(S) = 0.459$$

9 sí / 5 no

{soleado, nublado} {lluvioso}

{lluvioso, soleado} {nublado}

{lluvioso, nublado} {soleado}

$$IG(S, \{\text{lluvioso, soleado}\}) = G(S) - \frac{|S\{\text{lluvioso, soleado}\}|}{|S|} G(S_{\{\text{lluvioso, soleado}\}}) - \frac{|S\{\text{nublado}\}|}{|S|} G(S_{\{\text{nublado}\}})$$

$$= 0.459 - \frac{10}{14}\left(1 - \left(\frac{5}{10}\right)^2 - \left(\frac{5}{10}\right)^2\right) - \frac{4}{14}\left(1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2\right)$$

$$= 0.1$$

$$IG(S, \{\text{lluvioso, soleado}\}) = IG(S, \{\text{nublado}\})$$

Regresión

- Dividimos el espacio predictor - es decir, el conjunto de valores posibles para $X_1, X_2, \ldots, X_p$ - en J regiones distintas y no superpuestas, $R_1, R_2, \ldots, R_J$.
- Para cada observación que cae en la región R_J, hacemos la misma predicción, que es simplemente la media de los valores de respuesta para las observaciones de entrenamiento en R_J.
- En teoría las regiones podrían tener cualquier forma. Sin embargo, elegimos dividir el espacio predictor en rectángulos o cajas de alta dimensión, para simplificar y facilitar la interpretación del modelo predictivo resultante.

Métrica para realizar división binaria

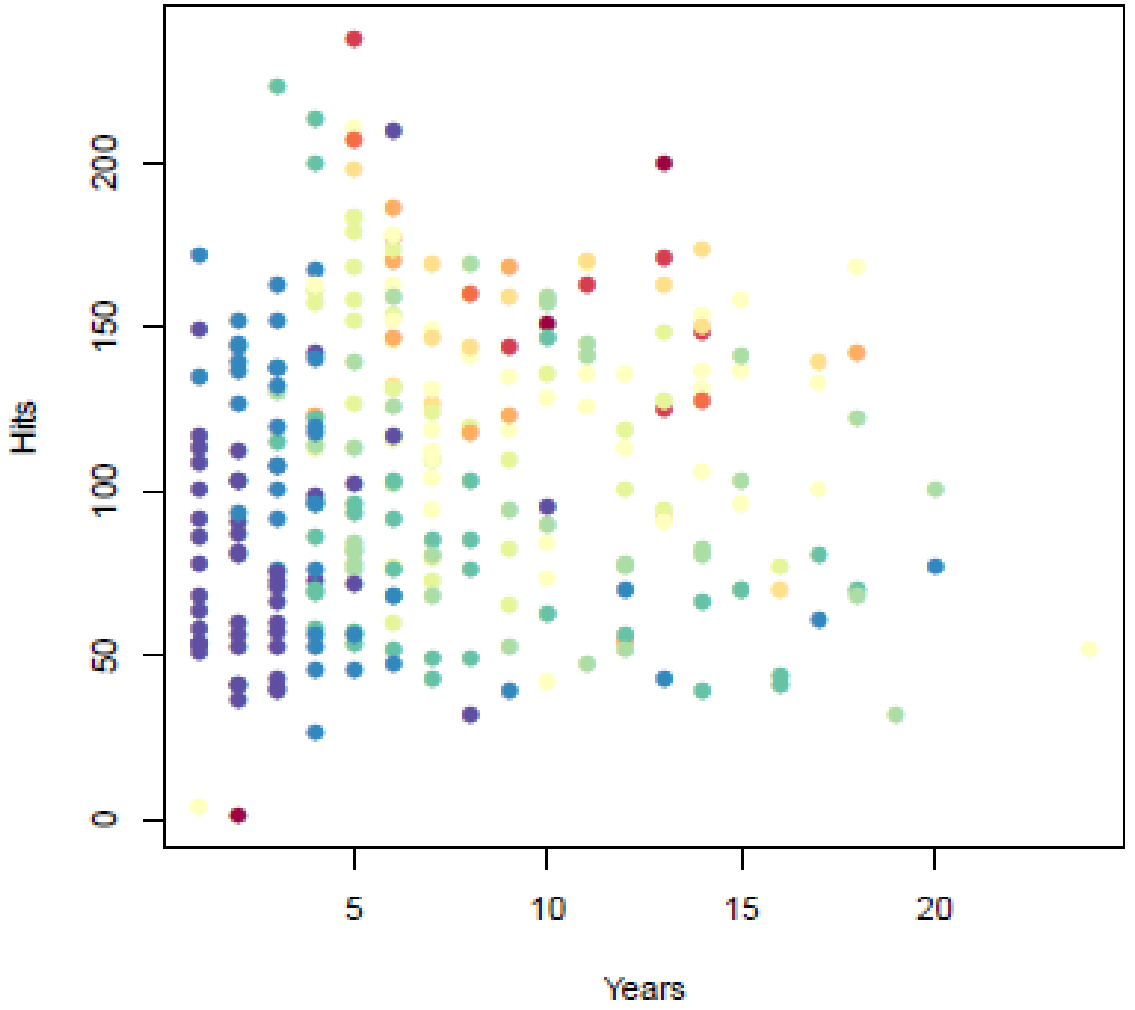
- Suma de residuos al cuadrados (RSS)

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

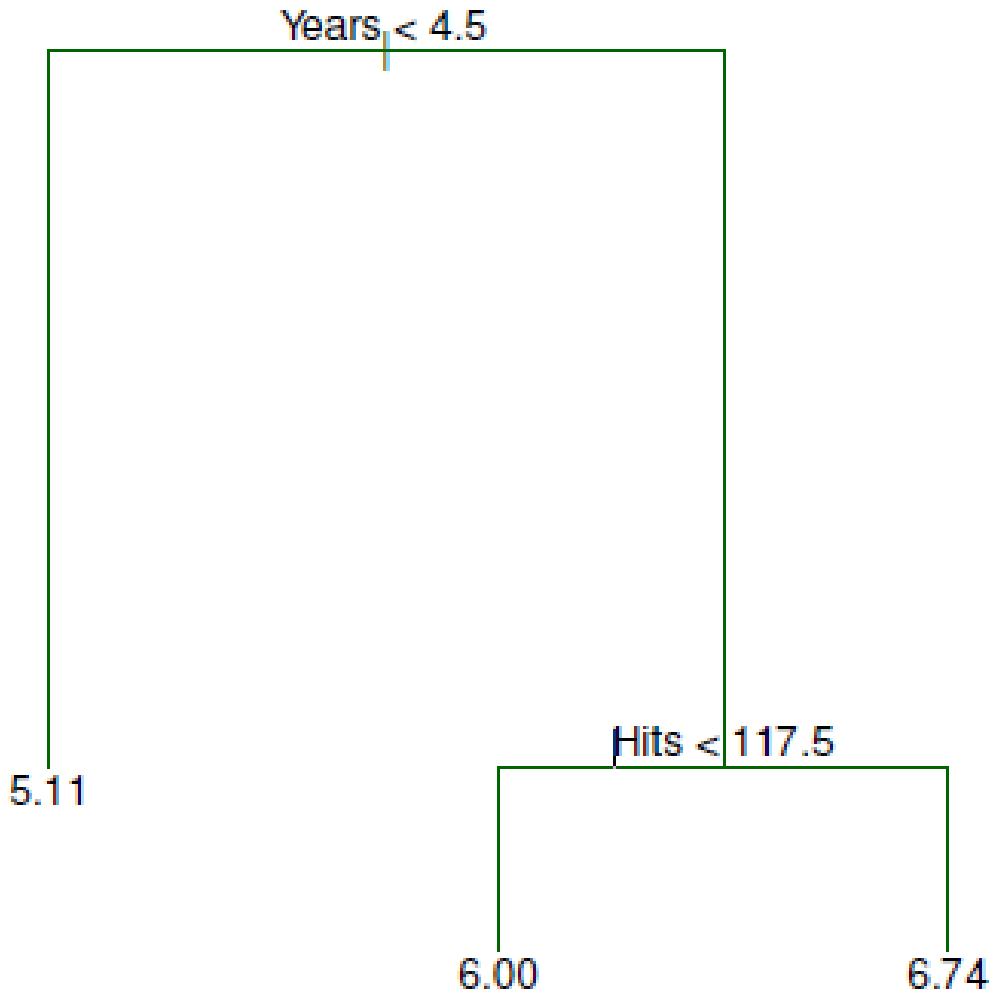
- Donde $\hat{y}_{R_j}$ es la respuesta promedio de las observaciones de entrenamiento dentro de la j-esima caja

Base de salarios en beisbol

- El salario está codificado con colores de bajo (azul, verde) a alto (amarillo, rojo)



Árbol de decisión para ejemplo anterior



Detalles del árbol anterior

- Un árbol de regresión para predecir el logaritmo del salario de un jugador de béisbol, basado en la cantidad de años que ha jugado en las ligas mayores y la cantidad de hits que realizó el año anterior.
- En un nodo interno, la etiqueta (de la forma $X_j < t_k$ indica la rama izquierda que emana de esa división, y la rama derecha corresponde a $X_j \geq t_k$. Por ejemplo, la división en la parte superior del árbol da como resultado dos grandes ramas. La rama de la izquierda corresponde a Años < 4.5, y la rama de la derecha corresponde a Años >= 4.5.
- El árbol tiene dos nodos internos y tres nodos terminales, u hojas. El número en cada hoja es la media de la respuesta para las observaciones que caen allí.

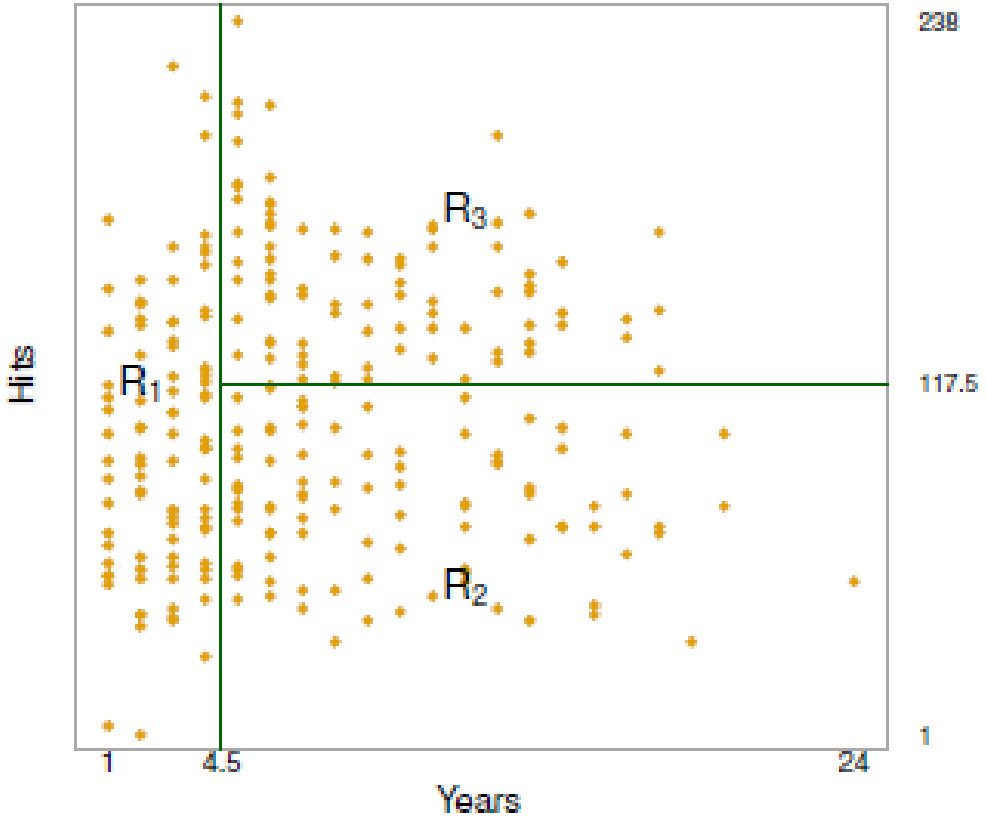
Resultados

- En general, el árbol estratifica o segmenta a los jugadores en tres regiones del espacio predictivo:

$R_1 = \{X | Años < 4.5\}$

$R_2 = \{X | Años \geq 4.5, Hits < 117.5\}$

$R_3 = \{X | Años \geq 4.5, Hits \geq 117.5\}.$



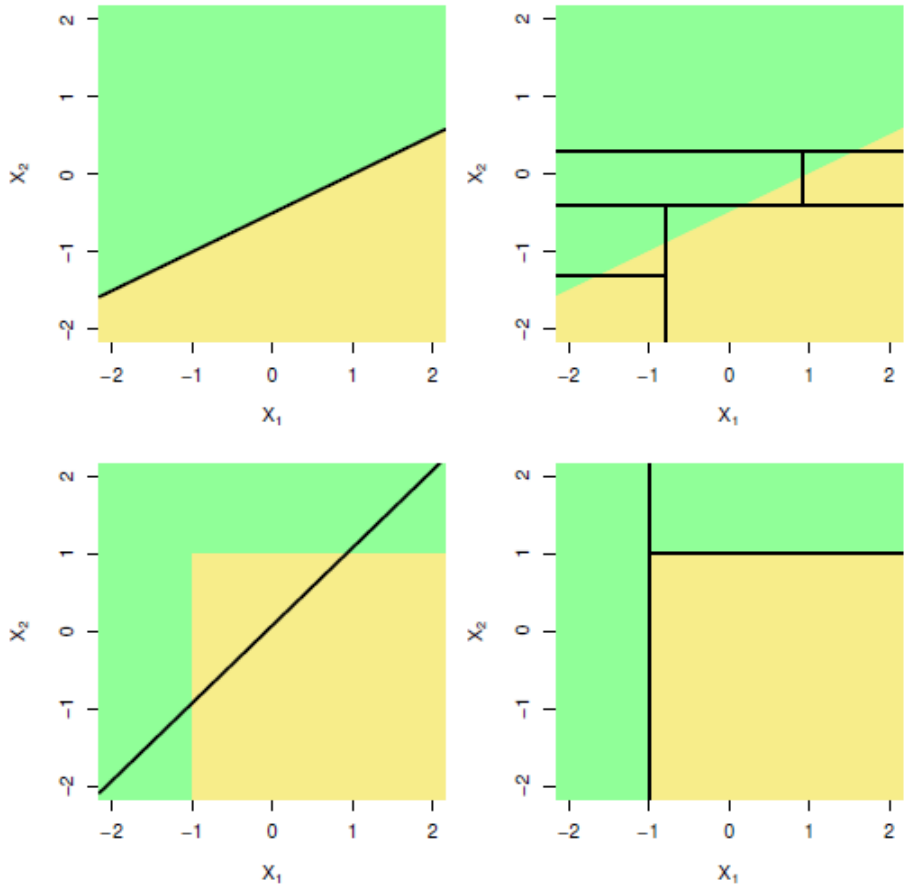
Poda de un árbol de regresión

- Una estrategia es crear un árbol muy grande T_0 y luego podarlo para obtener un subárbol
- Poda por costo de complejidad - también conocido como poda del eslabón más débil - se usa en estos casos
- Consideramos una secuencia de árboles indexados por un parámetro de ajuste no negativo α. Para cada valor de α corresponde un subárbol $T \subset T_0$ tal que

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

- Es lo más pequeño posible. Aquí $|T|$ indica el número de nodos terminales del árbol T, R_m es el rectángulo (es decir, el subconjunto del espacio predictor) correspondiente al m-ésimo nodo terminal, $\hat{y}_{R_m}$ es la media de las observaciones de entrenamiento en R_m

Árboles vs. modelos lineales



Ventajas y desventajas de los árboles

- Los árboles son muy fáciles de explicar a las personas. De hecho, ¡son incluso más fáciles de explicar que la regresión lineal!
- Algunas personas creen que los árboles de decisión reflejan más estrechamente la toma de decisiones humanas que la regresión y enfoques de clasificación vistos anteriormente.
- Los árboles se pueden mostrar gráficamente y pueden ser interpretados fácilmente incluso por alguien no experto (especialmente si son pequeños).
- Los árboles pueden manejar fácilmente predictores cualitativos sin la necesidad de crear variables dummy.
- Desafortunadamente, los árboles generalmente no tienen el mismo nivel de precisión predictiva que algunos de los otros enfoques de regresión y clasificación vistos anteriormente.
- Sin embargo, al agregar muchos árboles de decisión, se puede mejorar sustancialmente el rendimiento predictivo de los árboles.