— “Árboles de decisión”

15 Diciembre, 2019; rev. 13 de junio de 2023

Dr. Héctor Julián Selley Rojas, Dra. Elizabeth Guevara Martínez

1. Introducción

Un árbol de decisión es un modelo de predicción que se utiliza en diversas disciplinas como la Inteligencia Artificial, Medicina, Ingeniería, Ciencia de Datos y la Economía entre muchas otras. Los árboles se construyen desde un conjunto de datos, los diagramas resultantes son similares a los sistemas de predicción que se basan en reglas, los cuales sirven para categorizar una serie de condiciones que ocurren en forma sucesiva.

Los árboles de decisión se utilizan en cualquier proceso que implique una toma de decisión, por ejemplo:

- Búsqueda binaria
- Sistemas expertos
- Árboles de juego

Los árboles de decisión son generalmente binarios, lo que significa que pueden tomar dos opciones, aunque es posible que existan árboles de tres o más opciones.

2. Objetivos

- Encontrar un árbol binario que clasifique datos de entrada con una *dispersión* mínima.
- Calcular la eficiencia del proceso de clasificación mediante la *dispersión*.

3. ¿Qué es el árbol de decisión?

Árbol de decisión es una técnica de estructura de datos jerárquicos que se utiliza para la clasificación y regresión de datos. Este método emplea la técnica *divide y vencerás*, mediante la cual encuentra recursivamente la separación por clasificación de los datos de entrada.

Un árbol de decisión es un grafo que consiste en nodos y aristas. Cada nodo puede tener máximo dos aristas, razón por lo que se le denomina como binario. Un árbol de decisión responde una pregunta acerca de los datos y los clasifica de acuerdo con la respuesta de dicha pregunta.

Utilizaremos algunos ejemplos para explicar los árboles de decisión, cómo se definen y construyen.

3.1. Ejemplo 1

La figura 1 muestra un árbol de decisión que mediante una pregunta, cuya respuesta puede ser verdadero o falso, clasifica los datos de entrada en dos grupos.

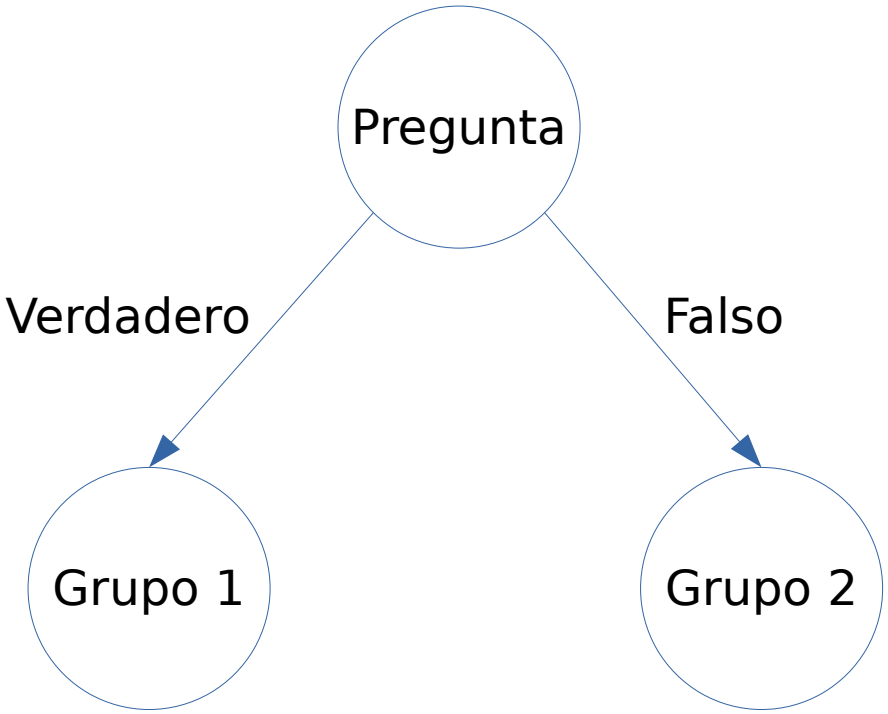


Figura 1: Ejemplo de árbol de decisión.

En los árboles, los **nodos** se representan con círculos o elipses en los cuales se aloja una pregunta, por otro lado, las aristas son la conexión entre ellos a través de la respuesta de la pregunta. Se denomina **rama** al conjunto de al menos dos nodos conectados por una arista.

Imagine que tiene un conjunto de datos que desea clasificar mediante una pregunta cuya respuesta es verdadero o falso, observe el árbol de la figura 1. Esto permite clasificar los datos en dos grupos, uno cuya respuesta fue verdadera y otro cuya respuesta fue falsa. Para un árbol tan pequeño como el de este ejemplo, la separación de los datos es muy limitada por lo que se busca mejorarla empleando más nodos en el árbol, lo que significa un mayor número de categorías. Adicionalmente, la pregunta sólo acepta respuestas absolutas, si se requiere de un rango de respuestas, por ejemplo, un rango de números habría que modificar el árbol.

3.2. Ejemplo 2

Construyamos un árbol con más nodos y ramificaciones, para este ejemplo se clasifica una persona de acuerdo con su edad. Se clasifica a una persona como adulto si su edad es mayor o igual a 18 años, como adolescente si está entre los 12 y 18 años, como niño si está entre los 2 y 12 años y como bebé si es menor a 2 años. El árbol resultante se muestra en la figura 2.

En el árbol resultante de la figura 2 clasifica a las personas de acuerdo con su edad utilizando el criterio antes mencionado. En el árbol las personas han sido clasificadas en los nodos: adulto, adolescente, niño y bebé. A estos nodos, aquellos que tienen flechas que llegan a él pero no salen de él, se les denomina como **nodos terminales** o de decisión. Al nodo inicial, aquel del cual sólo salen flechas de el pero no entran, se le denomina **nodo raíz** o simplemente **raíz**. Los demás simplemente se les denomina como **nodos**.

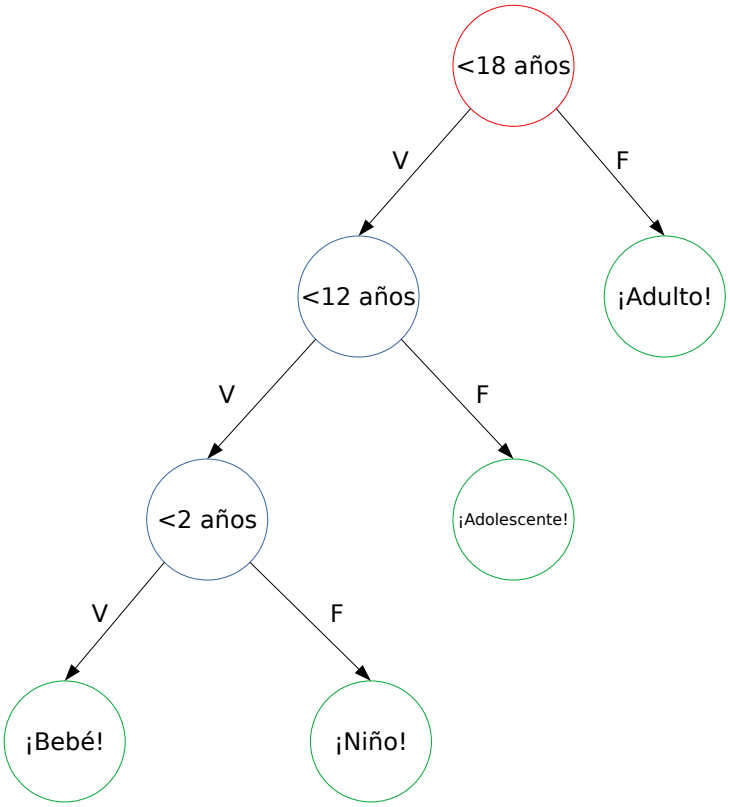


Figura 2: Árbol de decisión con más nodos.

En la figura 2 el nodo en rojo es la raíz, los nodos en verde son terminales y los azules son simplemente nodos.

En el árbol de decisión de la figura 2 clasifica adecuadamente a las personas, dado que una persona sólo tiene una edad, la clasificación es perfecta de esa forma. Imagínese que deseamos clasificar personas de acuerdo con otro criterio, un criterio en el cual la respuesta no será tan específica como la edad o incluso puede que no haya una respuesta. Por ejemplo, imagine que deseamos clasificar personas de acuerdo con su sabor preferido de helado, puede que tenga uno, varios o incluso ninguno. En una situación como ésta, habrá una **impureza** en la clasificación. Más adelante se explicará a través del ejemplo la impureza y cómo se calcula.

3.3. Ejemplo 3

Supongamos que a través de un estudio se obtiene un conjunto de datos acerca de 303 pacientes en los que se sabe si sufren de dolor en el pecho, tienen buena circulación sanguínea, arterias bloqueadas y ataque cardíaco. Se sabe que existe una relación entre estos padecimientos, pero se busca clasificar a los pacientes de la mejor forma posible. Se desea determinar las causas que provocan un ataque cardíaco y clasificar a los pacientes de acuerdo con ello, además un paciente que sufre un ataque cardíaco no presenta necesariamente todos los síntomas. Observe los datos presentados en la tabla 1 para cada uno de los pacientes.

Por otro lado, la tabla 2 muestra la cantidad total de pacientes obtenidos por categoría a través del estudio. Observe que el total de pacientes por síntoma no es igual para todas las categorías, esto es debido a que no se sabe la información completa para todos los pacientes.

Dado que se desea clasificar a los 303 pacientes de acuerdo con el síntoma que les ocasionó un infarto, se necesita determinar mediante cuál de los tres síntomas se debe clasificar en primer lugar. Este primer síntoma con el que se comience la clasificación se convertirá en el nodo raíz del árbol. Por esta razón se analizará cuál de los tres síntomas separa mejor a los pacientes que sufrieron un ataque cardíaco.

En la tabla 3 se muestran los datos totales desglosados por síntoma y si sufrieron o no un ataque cardíaco. Utilizando estos datos se puede llevar a cabo la separación de los pacientes respecto a si sufren

Dolor de pecho	Buena circulación	Arterias bloqueadas	Ataque cardíaco
No	No	No	No
Si	Si	Si	Si
–	Si	No	No
Si	Si	No	No
Si	No	–	Si
No	–	Si	Si
⋮	⋮	⋮	⋮

Tabla 1: Resultados del estudio de cada uno de los pacientes.

	Dolor de pecho	Buena circulación	Arterias bloqueadas	Ataque cardíaco
Si	144	164	123	137
No	159	133	174	160
Total	303	297	297	297

Tabla 2: Resultados totales del estudio de los pacientes.

de dolor de pecho o no. La separación se muestra en la figura 3.

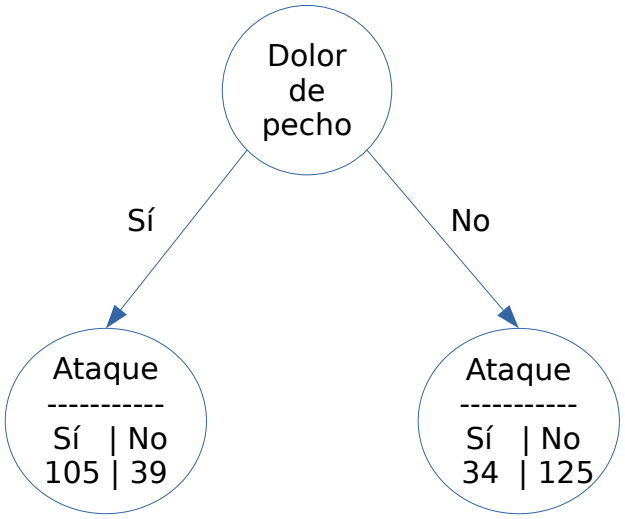


Figura 3: Separación mediante dolor de pecho.

Observe que la separación no es perfecta, dado que en cada rama se encuentran pacientes que han sufrido un ataque y otros que no. Esto es lo que se denomina como **impureza**.

Ataque	Dolor de pecho		Buena circulación		Arterias bloqueadas	
	Si	No	Si	No	Si	No
Si	105	34	37	33	92	45
No	39	125	127	100	31	129
Total	144	159	164	133	123	174

Tabla 3: Resultados Desglosados de los pacientes.

Resulta intuitivo buscar una separación que ocasione una menor impureza, por lo que para medirla resulta indispensable una métrica. Para medir la impureza se utiliza el **índice de impureza de Gini**[1][2], mediante la expresión (1).

$$G = 1 - (\text{Probabilidad Si})^2 - (\text{Probabilidad No})^2 \tag{1}$$

Donde G representa el índice de impureza Gini, *Probabilidad Si* y *Probabilidad No* son la probabilidad de que en un paciente tenga o no tenga dolor de pecho respectivamente. La probabilidad simplemente se calcula mediante el cociente de los pacientes con o sin dolor entre el total de pacientes.

De esta forma, se calcula el índice para cada separación posible y se elige aquella que tenga el menor valor de impureza.

Para calcular el índice de impureza para la separación con respecto a dolor de pecho G_{DP}, se debe analizar a su vez el índice para cada una de las ramas de dicha separación, esto es G_{Si} y G_{No}. Por lo tanto, el cálculo de la impureza para cada caso es el siguiente:

$$\begin{aligned} G_{Si} &= 1 - (\text{Probabilidad Si})^2 - (\text{Probabilidad No})^2 \\ &= 1 - \left(\frac{105}{105+39}\right)^2 - \left(\frac{39}{105+39}\right)^2 \\ &= 0.3949 \end{aligned}$$

$$\begin{aligned} G_{No} &= 1 - (\text{Probabilidad Si})^2 - (\text{Probabilidad No})^2 \\ &= 1 - \left(\frac{34}{34+125}\right)^2 - \left(\frac{125}{34+125}\right)^2 \\ &= 0.3362 \end{aligned}$$

Una vez calculado el índice de impureza de Gini para las dos hojas terminales, se calcula el índice total de impureza al separar los pacientes mediante el dolor de pecho. Sin embargo, dado que ambas hojas no representan la misma cantidad de pacientes se necesita utilizar el promedio ponderado de los índices de impureza para cada rama. Esto está dado en la expresión (2).

$$G = \frac{P_{Si}}{P_{Si}+P_{No}} G_{Si} + \frac{P_{No}}{P_{Si}+P_{No}} G_{No} \tag{2}$$

Donde P_{Si} es la cantidad total de pacientes que sufren dolor de pecho y P_{No} la cantidad de pacientes que no lo sufren.

De esta forma, el índice total de impureza al separar pacientes mediante dolor de pecho G_{DP} es:

$$\begin{aligned} G_{DP} &= \frac{P_{Si}}{P_{Si}+P_{No}} G_{Si} + \frac{P_{No}}{P_{Si}+P_{No}} G_{No} \\ &= \frac{105+39}{105+39+34+125} \times (0.3949) + \frac{34+125}{105+39+34+125} \times (0.3362) \\ &= 0.3641 \end{aligned}$$

Ahora se realiza la separación respecto a la buena circulación de la sangre y todos los cálculos correspondientes para obtener el índice de impureza G_{BC} para la separación mediante buena circulación de la sangre.

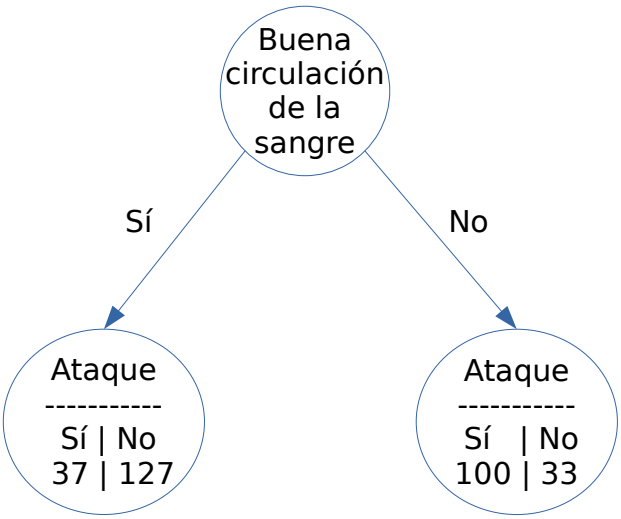


Figura 4: Separación mediante buena circulación de la sangre.

$$
\begin{aligned}
G_{Si} &= 1 - (\text{Probabilidad Si})^2 - (\text{Probabilidad No})^2 \\
&= 1 - \left(\frac{37}{37+127}\right)^2 - \left(\frac{127}{37+127}\right)^2 \\
&= 0.3494
\end{aligned}
$$

$$
\begin{aligned}
G_{No} &= 1 - (\text{Probabilidad Si})^2 - (\text{Probabilidad No})^2 \\
&= 1 - \left(\frac{100}{100+33}\right)^2 - \left(\frac{33}{100+33}\right)^2 \\
&= 0.3731
\end{aligned}
$$

$$
\begin{aligned}
G_{BC} &= \frac{P_{Si}}{P_{Si}+P_{No}} G_{Si} + \frac{P_{No}}{P_{Si}+P_{No}} G_{No} \\
&= \frac{37+127}{37+127+100+33} \times (0.3494) + \frac{100+33}{37+127+100+33} \times (0.3731) \\
&= 0.3600
\end{aligned}
$$

Por último, se realiza la separación mediante las arterias bloqueadas y los cálculos correspondientes para obtener el índice de impureza G_{AB} para la separación mediante las arterias bloqueadas.

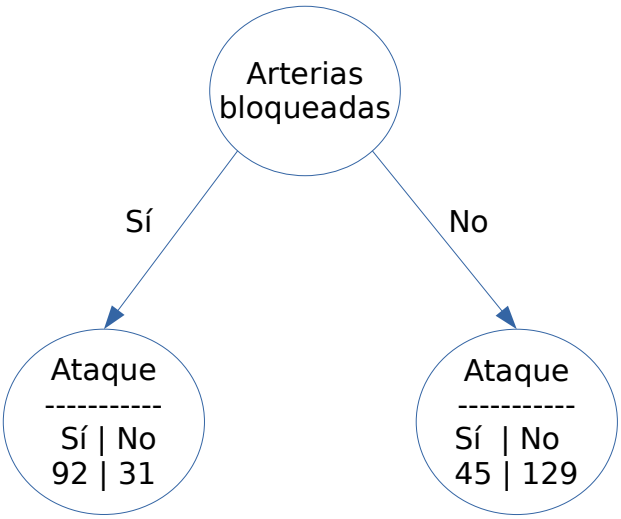


Figura 5: Separación mediante arterias bloqueadas.

$$
\begin{aligned}
G_{Si} &= 1 - (\text{Probabilidad Si})^2 - (\text{Probabilidad No})^2 \\
&= 1 - \left(\frac{92}{92+31}\right)^2 - \left(\frac{31}{92+31}\right)^2 \\
&= 0.3770
\end{aligned}
$$

$$
\begin{aligned}
G_{No} &= 1 - (\text{Probabilidad Si})^2 - (\text{Probabilidad No})^2 \\
&= 1 - \left(\frac{45}{45+129}\right)^2 - \left(\frac{129}{45+129}\right)^2 \\
&= 0.3834
\end{aligned}
$$

$$
\begin{aligned}
G_{AB} &= \frac{P_{Si}}{P_{Si}+P_{No}} G_{Si} + \frac{P_{No}}{P_{Si}+P_{No}} G_{No} \\
&= \frac{92+31}{92+31+45+129} \times (0.3770) + \frac{45+129}{92+31+45+129} \times (0.3834) \\
&= 0.3808
\end{aligned}
$$

La impureza es un efecto no deseado en la separación de pacientes bajo cualquier criterio, por esa razón es que se decidirá entre las posibles separaciones por aquella que tenga un menor valor de impureza. Comparando los valores de impureza $G_{DP} = 0.3641$, $G_{BC} = 0.36$ y $G_{AB} = 0.3808$ para la separación con respecto a dolor en el pecho, buena circulación y arterias bloqueadas respectivamente, se decide por G_{BC} debido a que su valor es el menor de todos. Esto significa que buena circulación se convertirá en el primer nodo, el nodo raíz. El árbol hasta ahora queda como en la figura 4.

Ahora debemos separar los pacientes respecto a dolor de pecho o arterias bloqueadas para cada hoja. Por lo tanto, se repite el procedimiento. En primer lugar, se analiza la rama verdadera (izquierda) del árbol de la figura 4. Observe que el número total de pacientes de esa rama es 164, de los cuales 37 sufrieron ataque y 127 no lo sufrieron.

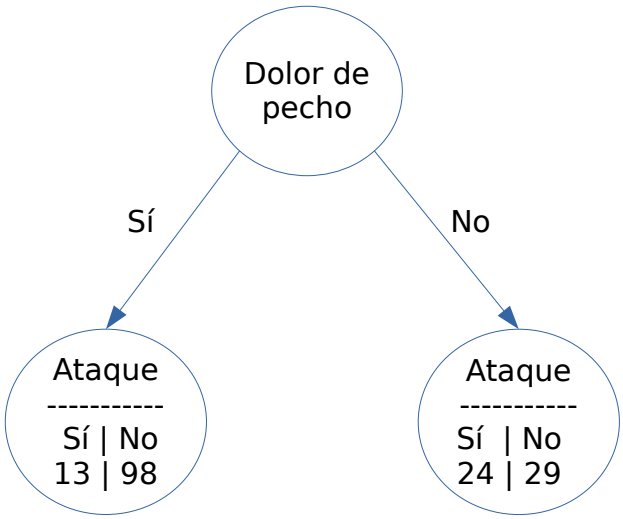


Figura 6: Separación de la rama izquierda del árbol de la figura 4 mediante dolor de pecho.

Mediante esta separación hay 111 pacientes que sufren dolor de pecho de los cuales 13 sufrieron ataque y 98 no, y hay 53 pacientes que no sufren dolor de pecho de los cuales 24 sufrieron un ataque y 29 no. Los cálculos de la impureza para esta separación son los siguientes:

$$G_{Si} = 1 - \left(\frac{13}{13+98}\right)^2 - \left(\frac{98}{13+98}\right)^2 = 0.2068$$
$$G_{No} = 1 - \left(\frac{24}{24+29}\right)^2 - \left(\frac{29}{24+29}\right)^2 = 0.4955$$
$$G_{DP} = \frac{13+98}{13+98+24+29} \times 0.2068 + \frac{24+29}{13+98+24+29} \times 0.4955$$
$$G_{DP} = 0.3001$$

Ahora se realiza la separación mediante arterias bloqueadas. Mediante esta separación de los 164 pacientes 49 padecen de arterias bloqueadas de los cuales 24 de ellos sufrieron ataque y 25 no, mientras que de los 115 restantes que no padecen de arterias bloqueadas 13 sufrieron ataque y 102 no. Observe la figura 7.

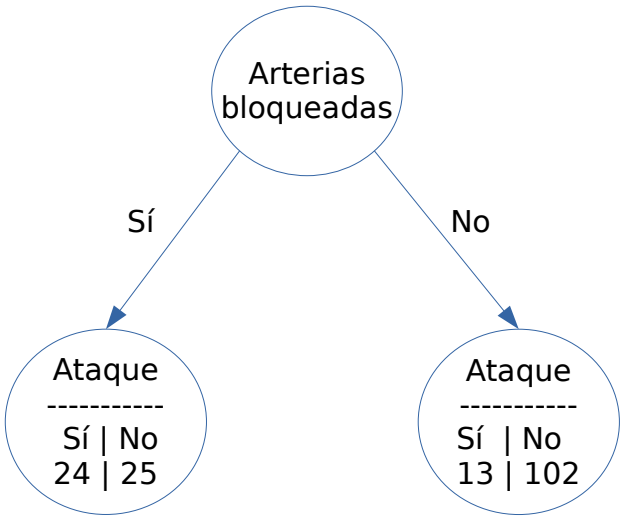


Figura 7: Separación de la rama izquierda del árbol de la figura 4 mediante arterias bloqueadas.

Los cálculos correspondientes son los siguientes:

$$G_{Si} = 1 - \left(\frac{24}{24+25}\right)^2 - \left(\frac{25}{24+25}\right)^2 = 0.4997 \tag{3}$$

$$G_{No} = 1 - \left(\frac{13}{13+102}\right)^2 - \left(\frac{102}{13+102}\right)^2 = 0.2005 \tag{4}$$

$$G_{AB} = \frac{24+25}{24+25+13+102} \times 0.4997 + \frac{13+102}{24+25+13+102} = 0.2899$$

Dado que G_{AB} es menor que G_{DP} se elige arterias bloqueadas para realizar la separación de los pacientes. De esta forma el árbol resultante hasta ahora se muestra en la figura 8.

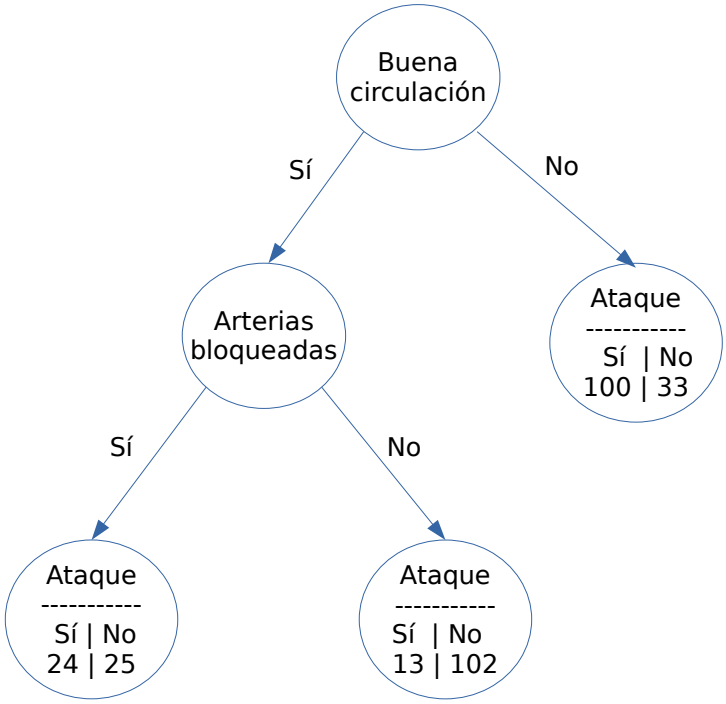


Figura 8: El árbol de decisión con sólo una rama separada mediante dos criterios.

Ahora se debe evaluar el separar los pacientes en la rama izquierda del árbol de la figura 8 mediante el dolor de pecho. Esto se determina calculando la impureza mediante la separación a través del dolor de pecho, si esta impureza resulta menor que la obtenida al separar por arterias bloqueadas entonces se realiza, si no fuera menor entonces no se realiza la separación. Recuerde que el objetivo al separar es obtener la menor impureza posible.

Considere que de los 49 pacientes que sufrieron de un ataque y padecen de arterias bloqueadas, se sabe que 20 padecen de dolor de pecho y 29 no. De los 20 que padecen de dolor de pecho 17 sufrieron un ataque y 3 no, mientras que de los 29 que no padecen dolor de pecho 7 sufrieron un ataque y 22 no. La figura 9 muestra esta clasificación.

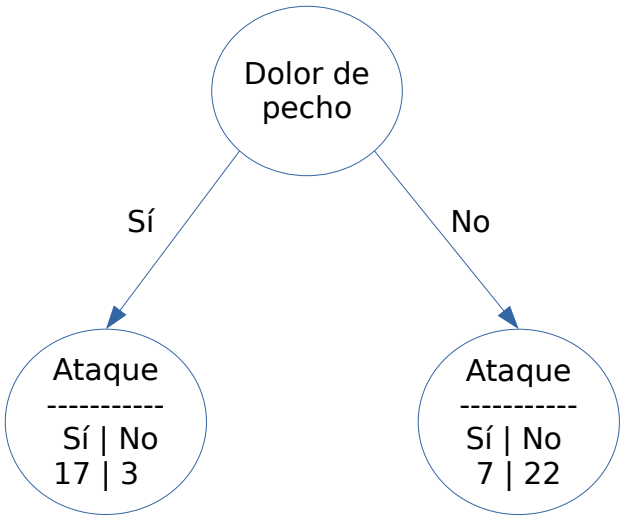


Figura 9: Separación de la rama izquierda afirmativa del árbol de la figura 8 mediante dolor de pecho.

Los cálculos de la impureza de esta separación son los siguientes:

$$G_{Si} = 1 - \left(\frac{17}{17+3}\right)^2 - \left(\frac{3}{17+3}\right)^2 = 0.255$$

$$G_{No} = 1 - \left(\frac{7}{7+22}\right)^2 - \left(\frac{22}{7+22}\right)^2 = 0.3662$$

$$G_{DP} = \frac{17+3}{17+3+7+22} \times 0.255 + \frac{7+22}{17+3+7+22} \times 0.3662 = 0.3208$$

En este punto se debe decidir si los últimos nodos son terminales. Si la impureza G_{DP} de la separación mediante dolor de pecho de la figura 9 es menor que la impureza anterior G calculada en 3 entonces se hace la separación, si no es menor entonces estos nodos son terminales. Dado que $G_{DP} = 0.3208$ es menor que $G_{Si} = 0.4997$ de la expresión 3, se realiza la separación.

Ahora se repiten los cálculos y la comparación con la impureza anterior para la rama negativa, es decir los pacientes que sufren de buena circulación pero no de arterias bloqueadas. De estos 115 pacientes 33 sufren de dolor de pecho y 82 no. De los 33 que sufren dolor de pecho 7 tuvieron un ataque y 26 no, por otro lado, de los 82 que no sufren de dolor de pecho 6 tuvieron un ataque y 76 no. La figura 10 muestra estos datos de la separación.

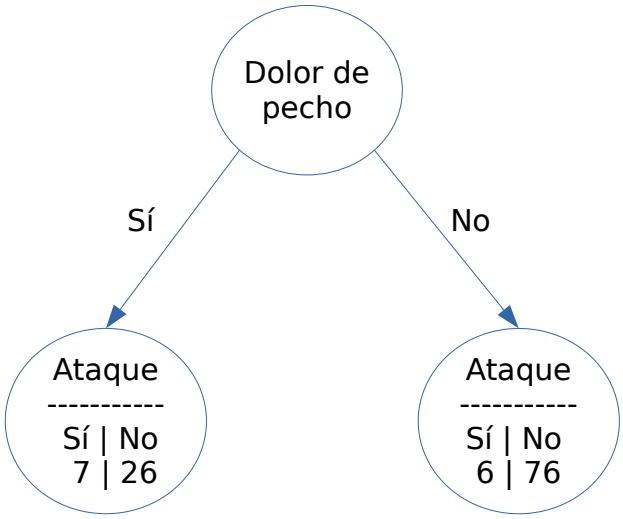


Figura 10: Separación de la rama izquierda negativa del árbol de la figura 8 mediante dolor de pecho.

Los cálculos de la impureza de esta separación son los siguientes:

$$G_{Si} = 1 - \left(\frac{7}{7+26}\right)^2 - \left(\frac{26}{7+26}\right)^2 = 0.3342$$
$$G_{No} = 1 - \left(\frac{6}{6+76}\right)^2 - \left(\frac{76}{6+76}\right)^2 = 0.1356$$
$$G_{DP} = \frac{7+26}{7+26+6+76} \times 0.3342 + \frac{6+76}{7+26+6+76} \times 0.1356 = 0.1926$$

Dado que G_{DP} es menor que la impureza anterior calculada en la expresión 4, también se realiza la separación en esta rama. Por lo tanto, el árbol con las separaciones realizadas en su rama izquierda se muestra en la figura 11.

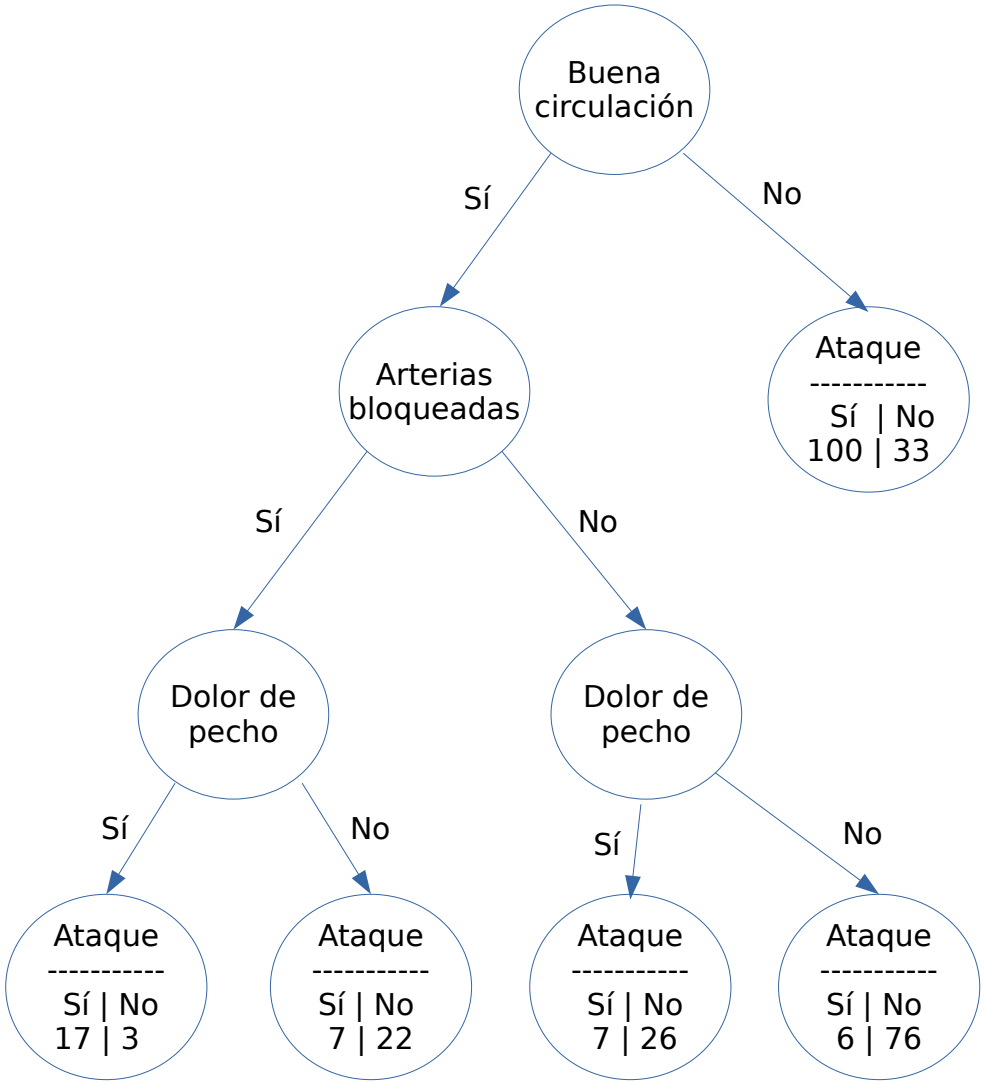


Figura 11: Separación completa de la rama izquierda del árbol.

A continuación solo falta repetir el mismo procedimiento para la rama derecha del árbol de la figura 11, el caso en el que los pacientes no tienen buena circulación. Se debe calcular la impureza para dolor de pecho y arterias bloqueadas, decidir por la menor impureza y si conviene o no la separación por el criterio correspondiente. Realizando los cálculos restantes los cálculos, el árbol de decisión resultante se muestra en la figura 12.

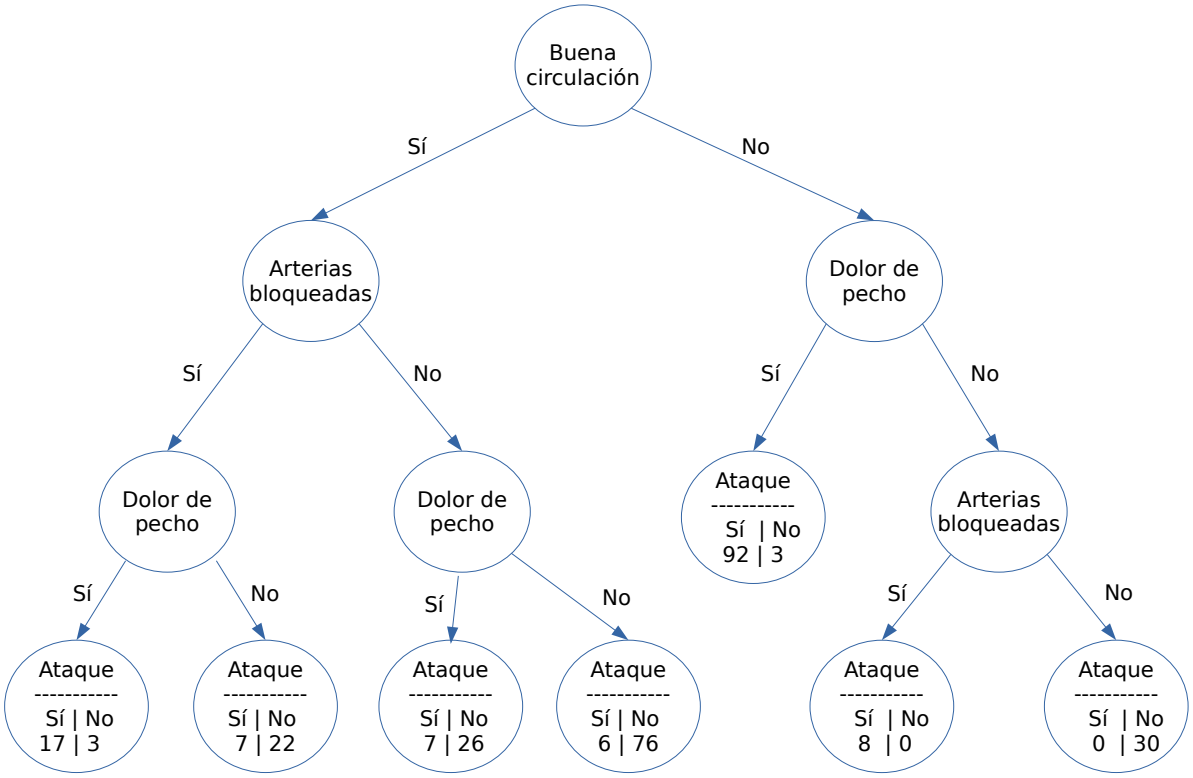


Figura 12: Árbol de decisión final para el ejemplo de los pacientes del estudio de la tabla 1.

4. ¿Cómo funcionan los Árboles de Decisión?

Recapitulando, un árbol de decisión es un método de aprendizaje supervisado no paramétrico que se utiliza para resolver problemas de clasificación y regresión y puede manejar tanto datos categóricos como numéricos.

En el caso más simple del proceso de construcción del árbol, el univariado, un nodo interno se divide de acuerdo con el valor de un solo atributo, a diferencia del caso multivariado en el que se considera la mejor combinación lineal de los atributos para la generación del árbol. Al final, a cada hoja se le asigna una categoría o la probabilidad de que el atributo tenga cierto valor.

Por tanto, un punto crucial en los árboles de decisión es elegir el criterio de división para obtener la mayor homogeneidad de los subconjuntos de acuerdo con la variable o atributo seleccionado y generar un árbol lo menos complejo posible, considerando el número total de nodos, de hojas, la profundidad del árbol y/o el número de atributos utilizados en el árbol.

En este capítulo nos vamos a centrar en el caso univariado y en en el índice de Gini.

Este criterio está basado en el concepto de impureza. La impureza en el proceso de separación es un fenómeno no deseado, que en un caso ideal se desea que sea cero. Un valor de impureza cero significa que la separación no mezcla datos entre categorías.

Existen muchas técnicas para medir la impureza en la separación, una popular es el índice de impureza de Gini[3][4], nombrada en honor a Corrado Gini que desarrolló un “coeficiente Gini” que sirve para medir la desigualdad en economía. Breiman et al.[1] aplicó la idea de Gini para medir la dispersión nominal de datos en el campo de los árboles de decisión.

Por otro lado, el índice de Gini para calcular la impureza de la separación utilizada en la expresión (1), se utiliza de acuerdo a la expresión en Rokach [4] que ha sido utilizada en varios trabajos en Breiman [1] y Gelfand [2]. Dicha expresión se define como

$$Gini(y, S) = 1 - \sum_{c_j \in dom(y)} \left(\frac{|S_{c_j}|}{|S|} \right)^2 .$$

El criterio de evaluación para seleccionar el atributo a_i se define como:

$$\Delta Gini(a_i, S) = Gini(y, S) - \sum_{v_{i,j} \in dom(a_i)} \frac{|S_{v_{i,j}}|}{|S|} Gini(y, S_{v_{i,j}})$$

La impureza de Gini es 0 cuando todos los ejemplos son de la misma clase. Esta métrica se aplica en el algoritmo CART[1].

5. Fundamentos

5.1. Probabilidad

La definición de probabilidad establecida por Laplace [5] para el caso de eventos equiprobables indica que la probabilidad de cualquier evento A es igual al cociente entre el número de resultados favorables de ocurrencia del evento A y el número total de elementos o posibles resultados del espacio muestral E.

$$P(A) = \frac{No.\ de\ casos\ favorables}{No.\ de\ casos\ posibles}$$

5.2. Función de impureza

Dada un variable aleatoria x con k valores discretos distribuidos de acuerdo con $P = (p_1, p_2, \dots p_k)$ una medida de impureza es una función $\varphi : [0, 1]^k$ que satisface las siguientes condiciones:

- $\varphi(P) \leq 0$
- $\varphi(P)$ es mínimo en los puntos $(1, 0, \dots, 0)$, $(0, 1, \dots, 0)$,$(0, 0, \dots, 1)$
- $\varphi(P)$ es máximo en el punto $\left(\frac{1}{k}, \frac{1}{k}, \dots, \frac{1}{k}\right)$
- $\varphi(P)$ simétrica con respecto a los componentes de P, $(p_1, p_2, \dots p_k)$.

Dado un conjunto de entrenamiento S, el vector de probabilidad del atributo objetivo y se define como:

$$P_y(S) = \left(\frac{|S_{c_1}|}{|S|}, \frac{|S_{c_2}|}{|S|}, \dots, \frac{|S_{c_{|dom(y)|}}|}{|S|} \right)$$

donde $dom\,(y)$ es el conjunto de posibles valores del atributo y, $|S_{c_i}|$ es el número de ejemplos en S etiquetados con c_i y $|S|$ es el número total de ejemplos. Para seleccionar el atributo discreto a_i a partir del que se realiza la mejor división de acuerdo con los valores $v_{i,j} \in dom\,(a_i)$, se define la reducción de la impureza como:

$$\Delta\varphi(a_i, S) = \varphi\left(P_y(S)\right) - \sum_{j=1}^{|dom(a_i)|} \frac{|S_{v_{i,j}}|}{|S|} \varphi\left(P_y(S_{v_{i,j}})\right)$$

6. Ejemplo

Dado el conjunto de datos en dos clases de la tabla 4 se realiza la clasificación de los datos mediante un árbol de decisión. Los datos tienen clase **r** y **n** que se muestran azul y rojo respectivamente en la gráfica 13.

a	b	clase
168	141	r
165	143	r
170	143	r
172	145	r
174	145	r
167	147	r
174	147	r
169	149	r
170	150	r
164	151	r
172	151	r
175	152	r
164	153	r
168	154	r
170	156	r
173	157	r
176	159	r
175	162	r
165	151	n
157	153	n
167	156	n
171	156	n
160	155	n
165	150	n
177	161	n
179	162	n
172	163	n
168	160	n
172	164	n
171	165	n
178	165	n
169	166	n
165	168	n
174	168	n
173	169	n
160	143	n

Tabla 4: Conjunto de datos del ejemplo

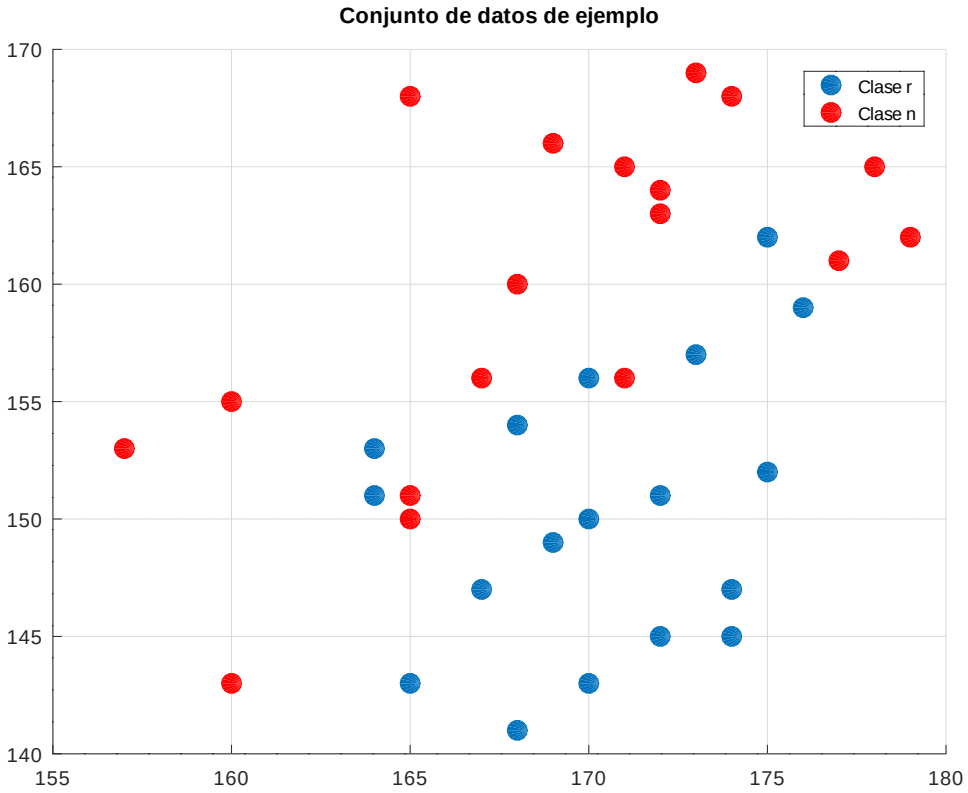


Figura 13: Gráfica del conjunto de datos del ejemplo

En el caso de atributos numéricos, los árboles de decisión se pueden interpretar geométricamente como una colección de hiperplanos ortogonales a los ejes.

Para obtener las divisiones, en este ejemplo de dos dimensiones, se consideran rangos en cada uno de los valores presentes en los atributos A y B. En la tabla 5 se muestran los resultados de utilizar el índice de Gini como medida de impureza.

Se presenta el ejemplo del cálculo de la impureza de Gini para el rango $A < 165$. Acomodando la tabla del conjunto de datos de menor a mayor con respecto a los valores de A se obtiene:

Rango de A	$\Delta Gini$	Rango B	$\Delta Gini$
A < 168	0.484	B < 141	0.5
A < 165	0.497	B < 143	0.486
A < 170	0.494	B < 143	0.486
A < 172	0.498	B < 145	0.484
A < 174	0.498	B < 145	0.484
A < 167	0.481	B < 147	0.456
A < 174	0.498	B < 147	0.456
A < 169	0.494	B < 149	0.42
A < 170	0.494	B < 150	0.399
A < 164	0.455	B < 151	0.411
A < 172	0.498	B < 151	0.411
A < 175	0.5	B < 152	0.396
A < 164	0.455	B < 153	0.371
A < 168	0.484	B < 154	0.375
A < 170	0.494	B < 156	0.375
A < 173	0.498	B < 157	0.396
A < 176	0.484	B < 159	0.365
A < 175	0.5	B < 162	0.377
A < 165	0.497	B < 151	0.411
A < 157	0.5	B < 153	0.371
A < 167	0.481	B < 156	0.375
A < 171	0.498	B < 156	0.375
A < 160	0.486	B < 155	0.346
A < 165	0.497	B < 150	0.399
A < 177	0.455	B < 161	0.353
A < 179	0.486	B < 162	0.377
A < 172	0.498	B < 163	0.357
A < 168	0.484	B < 160	0.326
A < 172	0.498	B < 164	0.379
A < 171	0.498	B < 165	0.4
A < 178	0.471	B < 165	0.4
A < 169	0.494	B < 166	0.438
A < 165	0.497	B < 168	0.455
A < 174	0.498	B < 168	0.455
A < 173	0.498	B < 169	0.486
A < 160	0.486	B < 143	0.486

Tabla 5: Cálculos de la impureza de Gini para encontrar el nodo raíz.

De aquí se puede observar más fácilmente que para el rango A<165, los cálculos son:

$$G_{si} = 1 - \left(\frac{2}{2+3}\right)^2 - \left(\frac{3}{2+3}\right)^2 = 0.48$$
$$G_{no} = 1 - \left(\frac{16}{16+15}\right)^2 - \left(\frac{15}{16+15}\right)^2 = 0.4995$$
$$G_{A<165} = \frac{2+3}{2+3+16+15} \times 0.48 + \frac{16+15}{2+3+16+15} \times 0.4995 = 0.497$$

De esta forma se obtienen todos los resultados de la tabla 5, en donde se observa que de todos estos intervalos el que tiene el menor índice es $B < 160$ con 0.326. Por lo tanto, la primera división se hace en hace en un valor de $B < 160$, en este ejemplo se seleccionó $B = 159.5$, donde se encuentra la primera línea de división. Al continuar este proceso se obtiene el árbol de decisión que se muestra en la figura 14.

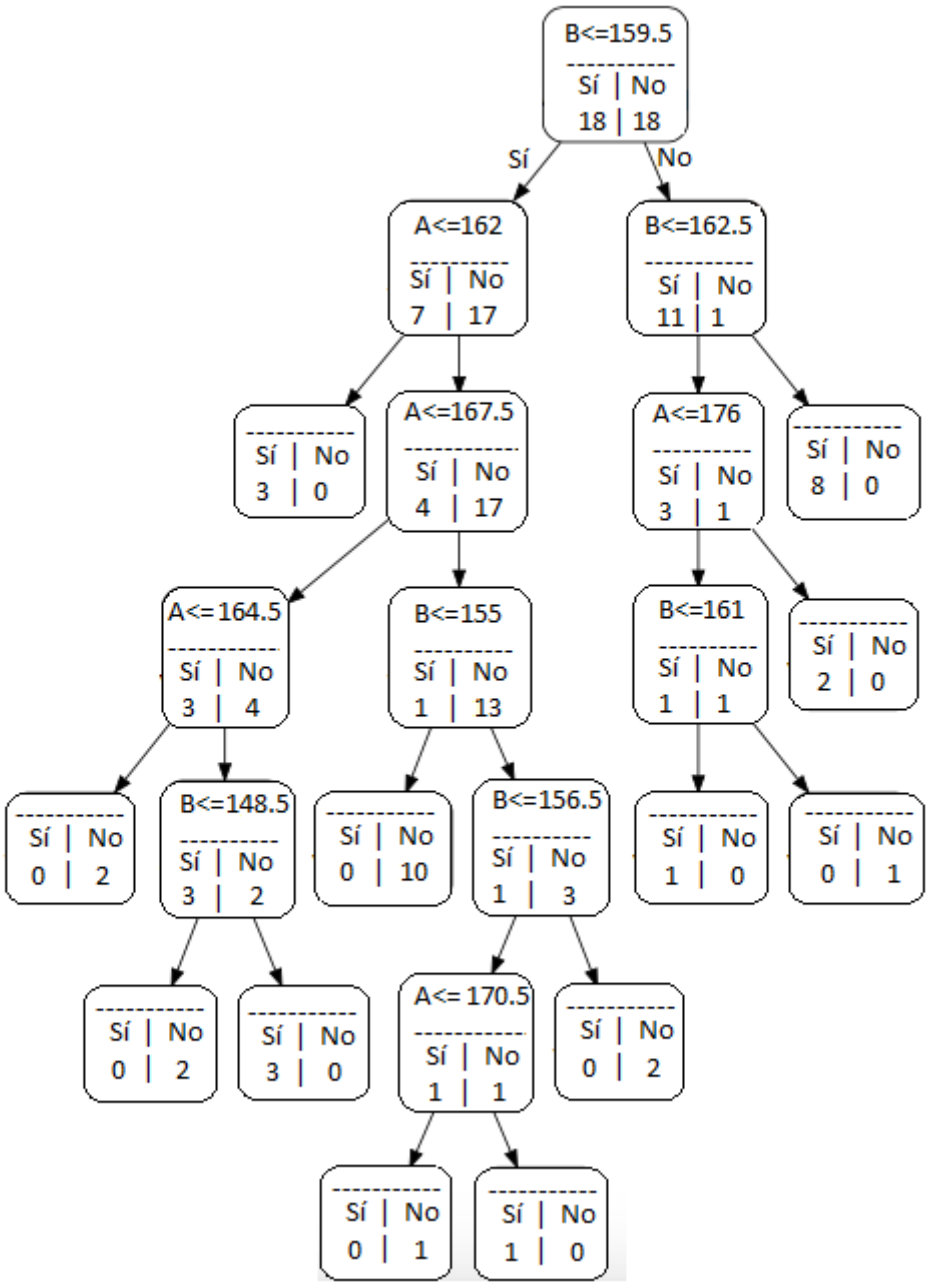


Figura 14: Árbol de decisión que resuelve el ejemplo.

En la figura 15 se muestra la gráfica de las divisiones generadas por el árbol de decisión:

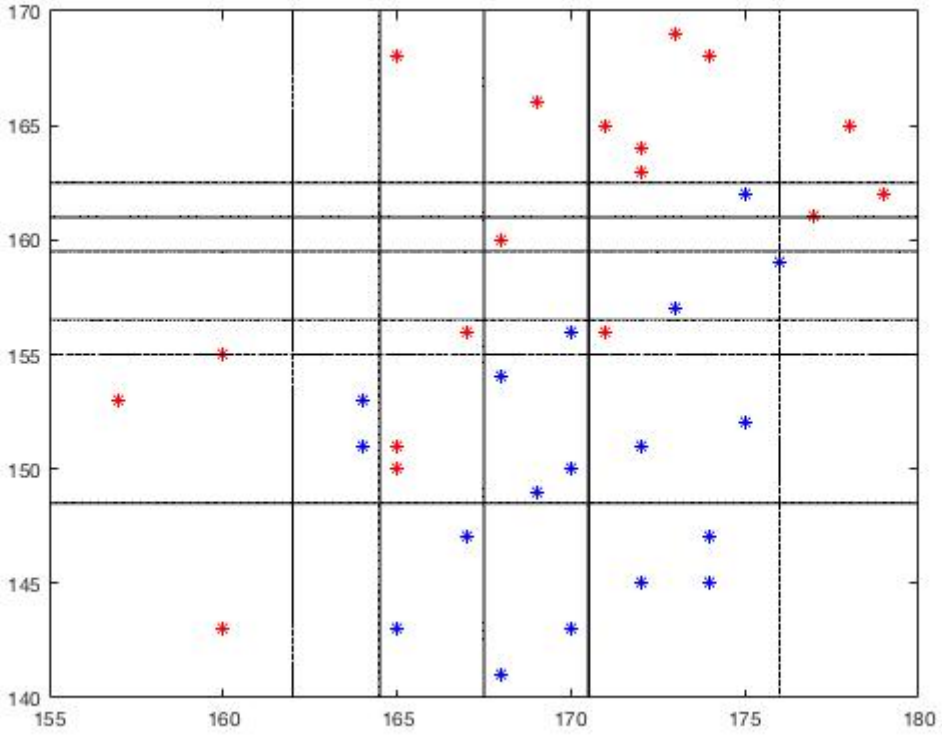


Figura 15: Divisiones generadas por el árbol de decisión.

Una forma de controlar la complejidad del árbol de decisión es mediante el criterio de paro. Algunos criterios de paro que se pueden utilizar son:

1. Todas las instancias del conjunto de entrenamiento pertenecen a una sola clase.
2. Se alcanza la máxima profundidad del árbol establecida.
3. El mejor criterio de división no es mayor que cierto umbral.

En este ejemplo se utilizó el criterio 1, por lo que se obtiene una clasificación perfecta del conjunto de entrenamiento, como se observa en la figura 15, lo que puede generar sobre-entrenamiento.

7. Ejercicios

1. Los datos de la tabla presentan el resultado de un análisis de datos de una campaña publicitaria. En la tabla se muestra la edad, género, exposición a la publicidad del producto y compra. Realice un árbol de decisión para clasificar *compradores* y *no compradores* con respecto a los atributos *edad*, *género* y *exposición*.

Edad	Género	Exposición	Compra
45	Femenino	No	No
23	Femenino	No	No
49	Masculino	No	Si
26	Femenino	Si	Si
65	Masculino	No	Si
17	Masculino	No	No
25	Femenino	Si	Si
37	Femenino	No	Si
35	Masculino	Si	Si
23	Masculino	Si	Si

2. Los datos de la tabla presentan el resultado de un análisis de datos de una campaña publicitaria. En la tabla se muestra la edad, género, exposición a la publicidad del producto y compra. Realice un árbol de decisión para clasificar *compradores* y *no compradores* con respecto a los atributos *edad*, *género* y *exposición*.

Edad	Género	Exposición	Compra
25	Femenino	No	Si
18	Masculino	No	No
34	Femenino	No	Si
32	Femenino	Si	No
56	Masculino	No	Si
76	Masculino	Si	Si
34	Masculino	No	No
28	Femenino	No	Si
20	Masculino	Si	Si
51	Masculino	Si	Si

3. Los datos de la tabla muestran los resultados de personas que contrajeron el virus SARS-COVID19. En ella se muestran padecimientos típicos que se sabe traen complicaciones a la enfermedad y en el peor de los casos la muerte. Realice un árbol de decisión para clasificar si un paciente sobrevive o no con respecto a los atributos *edad*, *diabetes* y *sobrepeso*.

Edad	Diabetes	Sobrepeso	Sobrevive
68	Si	Si	No
62	Si	Si	No
19	No	Si	No
31	No	Si	Si
44	Si	Si	Si
39	No	No	No
45	No	Si	Si
27	Si	No	Si
27	No	No	Si
23	No	No	Si

4. Los datos de la tabla muestran los resultados de personas que contrajeron el virus SARS-COVID19. En ella se muestran padecimientos típicos que se sabe traen complicaciones a la enfermedad y en el peor de los casos la muerte. Realice un árbol de decisión para clasificar si un paciente sobrevive o no con respecto a los atributos *edad*, *diabetes* y *sobrepeso*.

Edad	Diabetes	Sobrepeso	Sobrevive
45	Si	No	No
23	No	No	No
49	Si	Si	Si
75	No	No	Si
18	No	Si	Si
39	Si	No	No
17	No	No	Si
35	Si	No	No
62	No	Si	Si
50	No	No	Si

5. La tabla que se muestra a continuación tiene datos referentes a mediciones de diferentes especies de la flor de iris[6]. Los datos corresponden a el largo y ancho del sépalo y pétalo del iris. Mediante las mediciones se debe clasificar en *setosa*, *versicolor* y *virginica*. Construya un árbol de decisión para realizar la clasificación.

Longitud sépalo	Ancho sépalo	Longitud pétalo	Ancho pétalo	Clase
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
7	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.5	versicolor
5.5	2.3	4	1.3	versicolor
6.5	2.8	4.6	1.5	versicolor
6.3	3.3	6	2.5	virginica
5.8	2.7	5.1	1.9	virginica
7.1	3	5.9	2.1	virginica
6.3	2.9	5.6	1.8	virginica
6.5	3	5.8	2.2	virginica

6. La tabla que se muestra a continuación tiene datos referentes a mediciones de diferentes especies de la flor de iris[6]. Los datos corresponden a el largo y ancho del sépalo y pétalo del iris. Mediante las mediciones se debe clasificar en *setosa*, *versicolor* y *virginica*. Construya un árbol de decisión para realizar la clasificación.

Longitud sépalo	Ancho sépalo	Longitud pétalo	Ancho pétalo	Clase
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.7	2.8	4.5	1.3	versicolor
6.3	3.3	4.7	1.6	versicolor
4.9	2.4	3.3	1	versicolor
6.6	2.9	4.6	1.3	versicolor
5.2	2.7	3.9	1.4	versicolor
7.6	3	6.6	2.1	virginica
4.9	2.5	4.5	1.7	virginica
7.3	2.9	6.3	1.8	virginica
6.7	2.5	5.8	1.8	virginica
7.2	3.6	6.1	2.5	virginica

7. La siguiente tabla muestra los datos obtenidos de un grupo de pacientes para determinar si un tumor en el pecho es cancerígeno. Los datos corresponden al radio (distancia media entre el centro y los puntos del perímetro), textura (desviación estándar de valores en escala de grises), perímetro y área. Con estos datos se clasificó si el tumor es *benigno* o *maligno*. Construya un árbol de decisión para realizar la clasificación.

Radio	Textura	Perímetro	Área	Diagnóstico
14.06	17.18	89.7	609.1	Benigno
10.29	27.61	65.67	321.4	Benigno
10.17	14.88	64.55	311.9	Benigno
15.7	20.31	101.2	766.6	Maligno
17.08	27.15	111.2	930.9	Maligno
12.68	23.84	82.69	499	Maligno
10.26	16.58	65.85	320.8	Benigno
12	28.23	76.77	442.5	Benigno
15.75	20.25	102.6	761.3	Maligno
12.23	19.56	78.54	461	Benigno

8. La siguiente tabla muestra los datos obtenidos de un grupo de pacientes para determinar si un tumor en el pecho es cancerígeno. Los datos corresponden al radio (distancia media entre el centro y los puntos del perímetro), textura (desviación estándar de valores en escala de grises), perímetro y área. Con estos datos se clasificó si el tumor es *benigno* o *maligno*. Construya un árbol de decisión para realizar la clasificación.

Radio	Textura	Perímetro	Área	Diagnóstico
12.96	18.29	84.18	525.2	Benigno
11.45	20.97	73.81	401.5	Benigno
9.56	15.91	60.21	279.6	Benigno
13.43	19.63	85.84	565.4	Maligno
16.78	18.8	109.3	886.3	Maligno
14.29	16.82	90.3	632.6	Benigno
11.46	18.16	73.59	403.1	Benigno
12.67	17.2	81.25	489.9	Benigno
14.53	13.98	93.86	644.2	Benigno
13.3	21.57	85.24	846.1	Benigno

9. Los datos de la tabla muestran las medidas de algunos atributos realizadas a diversos vinos. Los atributos medidos en el vino corresponden a la cantidad de alcohol, ácido málico, ceniza y alcalinidad de la ceniza. Los vinos se clasifican en tres categorias que corresponden a tres tipos de vino. Construya un árbol de decisión para clasificar los vinos con respecto a los tres tipos.

Alcohol	Ácido málico	Ceniza	Alcalinidad ceniza	Tipo
12.77	2.39	2.28	19.5	2
12.84	2.96	2.61	24	2
13.71	1.86	2.36	16.6	0
13.23	3.3	2.28	18.5	2
13.32	3.24	2.38	21.5	2
12.42	2.55	2.27	22	1
11.66	1.88	1.92	16	1
14.12	1.48	2.32	16.8	0
13.28	1.64	2.84	15.5	0
12.16	1.61	2.31	22.8	1

10. Los datos de la tabla muestran las medidas de algunos atributos realizadas a diversos vinos. Los atributos medidos en el vino corresponden a la cantidad de alcohol, ácido málico, ceniza y alcali-

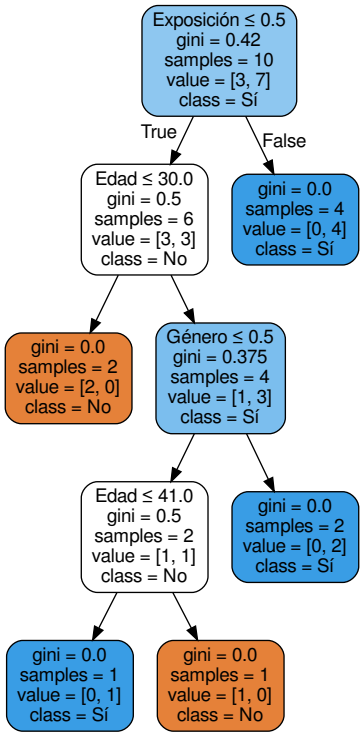
nidad de la ceniza. Los vinos se clasifican en tres categorias que corresponden a tres tipos de vino. Construya un árbol de decisión para clasificar los vinos con respecto a los tres tipos.

Alcohol	Ácido málico	Ceniza	Alcalinidad ceniza	Tipo
12.33	0.99	1.95	14.8	1
13.41	3.84	2.12	18.8	0
13.16	2.36	2.67	18.6	0
12.21	1.19	1.75	16.8	1
12.08	1.83	2.32	18.5	1
13.23	3.3	2.28	18.5	2
12.25	4.72	2.54	21	2
13.48	1.81	2.41	20.5	0
13.08	3.9	2.36	21.5	2
13.34	0.94	2.36	17	1

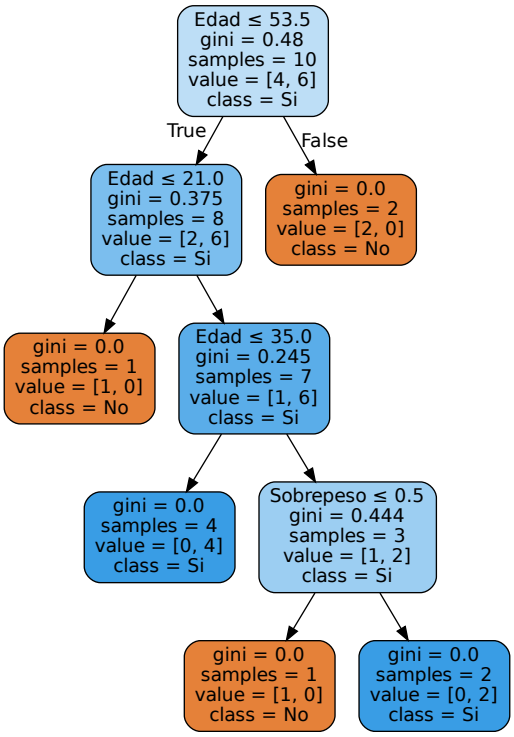
8. Solución

Solución a los ejercicios impares.

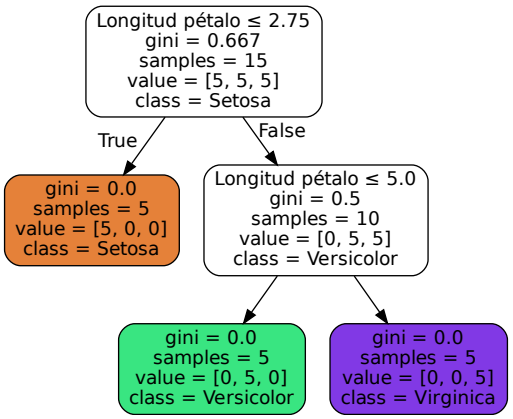
1.



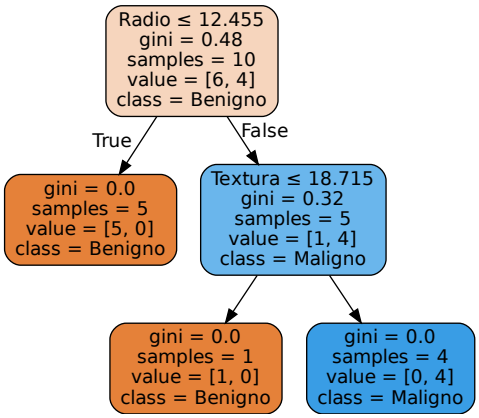
3.


Edad ≤ 53.5
gini = 0.48
samples = 10
value = [4, 6]
class = Si
True
False
Edad ≤ 21.0
gini = 0.375
samples = 8
value = [2, 6]
class = Si
gini = 0.0
samples = 2
value = [2, 0]
class = No
gini = 0.0
samples = 1
value = [1, 0]
class = No
Edad ≤ 35.0
gini = 0.245
samples = 7
value = [1, 6]
class = Si
gini = 0.0
samples = 4
value = [0, 4]
class = Si
Sobrepeso ≤ 0.5
gini = 0.444
samples = 3
value = [1, 2]
class = Si
gini = 0.0
samples = 1
value = [1, 0]
class = No
gini = 0.0
samples = 2
value = [0, 2]
class = Si

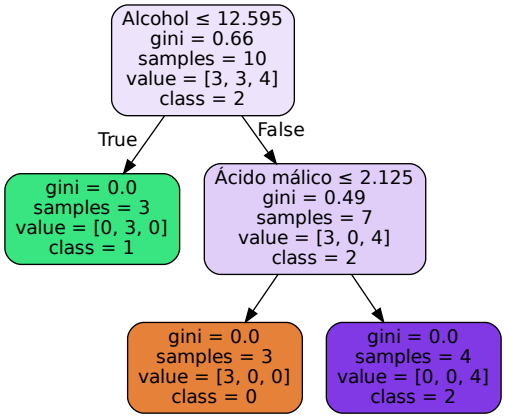
5.


Longitud pétalo ≤ 2.75
gini = 0.667
samples = 15
value = [5, 5, 5]
class = Setosa
True
False
gini = 0.0
samples = 5
value = [5, 0, 0]
class = Setosa
Longitud pétalo ≤ 5.0
gini = 0.5
samples = 10
value = [0, 5, 5]
class = Versicolor
gini = 0.0
samples = 5
value = [0, 5, 0]
class = Versicolor
gini = 0.0
samples = 5
value = [0, 0, 5]
class = Virginica

7.



9.



Referencias

[1] L. Breiman, *Classification and Regression Trees*. CRC Press, 2017.

[2] S. B. Gelfand, C. S. Ravishankar, and E. J. Delp, "An iterative growing and pruning algorithm for classification tree design," *IEEE*, vol. 13, no. 2, pp. 163–174, 1991.

[3] J. Unpingco, *Python for Probability, Statistics, and Machine Learning*. Springer, 2016.

[4] L. Rokach and O. Maimon, *Data Mining with Decision Trees*, vol. 81 of *Series in Machine Perception and Artificial Intelligence*. World Scientific, 2 ed., 2015.

[5] P. S. Laplace, *Teoría analítica de las probabilidades*. Courcier, 3 ed., 1820.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.