Árboles de decisión

Héctor Selley

Universidad Anáhuac México

21 de junio de 2023

Contenido

1. ¿Qué es un árbol de decisión?
2. Ejemplos
   - Ejemplo 1
   - Ejemplo 2
   - Ejemplo 3
3. Bibliografía

¿Qué es un árbol de decisión?

¿Qué es un árbol de decisión?

- Un árbol de decisión es un modelo de predicción.

¿Qué es un árbol de decisión?

- Un árbol de decisión es un modelo de predicción.
- Se utiliza en diversas disciplinas como la Inteligencia Artificial, Medicina, Ingeniería, Ciencia de Datos y la Economía, entre muchas otras.

¿Qué es un árbol de decisión?

- Un árbol de decisión es un modelo de predicción.
- 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

¿Qué es un árbol de decisión?

- Un árbol de decisión es un modelo de predicción.
- 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

¿Qué es un árbol de decisión?

- Un árbol de decisión es un modelo de predicción.
- 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
- Sirven para categorizar una serie de condiciones que ocurren en forma sucesiva.

¿Qué es un árbol de decisión?

Los árboles de decisión se utilizan en cualquier proceso que implique una toma de decisión

¿Qué es un árbol de decisión?

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

¿Qué es un árbol de decisión?

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

- Búsqueda binaria

¿Qué es un árbol de decisión?

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

- Búsqueda binaria
- Sistemas expertos

¿Qué es un árbol de decisión?

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

- Búsqueda binaria
- Sistemas expertos
- Árboles de juego

¿Qué es un árbol de decisión?

- Los árboles de decisión son generalmente binarios

¿Qué es un árbol de decisión?

- Los árboles de decisión son generalmente binarios
- Significa que pueden tomar dos opciones

¿Qué es un árbol de decisión?

- Los árboles de decisión son generalmente binarios
- Significa que pueden tomar dos opciones
- Aunque es posible que existan árboles de tres o más opciones.

¿Para qué sirve un árbol de decisión?

Objetivos del árbol de decisión:

¿Para qué sirve un árbol de decisión?

Objetivos del árbol de decisión:

- Encontrar un árbol binario que clasifique datos de entrada con una *dispersión* mínima.

¿Para qué sirve un árbol de decisión?

Objetivos del árbol de decisión:

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

¿Qué es un árbol de decisión?

Árbol de decisión

Árbol de decisión es una técnica de estructura de datos jerárquicos que se utiliza para la clasificación y regresión de datos.

¿Qué es un árbol de decisión?

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.

¿Qué es un árbol de decisión?

- Un árbol de decisión es un grafo que consiste en nodos y aristas.

¿Qué es un árbol de decisión?

- 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.

¿Qué es un árbol de decisión?

- 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.

¿Qué es un árbol de decisión?

- 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.

Contenido

1. ¿Qué es un árbol de decisión?
2. Ejemplos
   - Ejemplo 1
   - Ejemplo 2
   - Ejemplo 3
3. Bibliografía

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.

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.

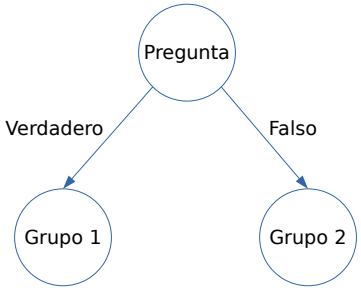


Figura: Ejemplo de árbol de decisión.

Ejemplo 1

- En los árboles, los **nodos** se representan con círculos o elipses en los cuales se aloja una pregunta

Ejemplo 1

- En los árboles, los **nodos** se representan con círculos o elipses en los cuales se aloja una pregunta
- 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.

Ejemplo 1

- Imagine que tiene un conjunto de datos que desea clasificar mediante una pregunta cuya respuesta es verdadero o falso. (figura 1)

Ejemplo 1

- Imagine que tiene un conjunto de datos que desea clasificar mediante una pregunta cuya respuesta es verdadero o falso. (figura 1)
- Esto permite clasificar los datos en dos grupos, uno cuya respuesta fue verdadera y otro cuya respuesta fue falsa.

Ejemplo 1

- Imagine que tiene un conjunto de datos que desea clasificar mediante una pregunta cuya respuesta es verdadero o falso. (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.

Ejemplo 1

- Imagine que tiene un conjunto de datos que desea clasificar mediante una pregunta cuya respuesta es verdadero o falso. (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.

Ejemplo 2

Descripción del problema

Construyamos un árbol con más nodos y ramificaciones, para este ejemplo se clasifica una persona de acuerdo con su edad.

Ejemplo 2

Descripción del problema

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.

Ejemplo 2

Descripción del problema

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.

Ejemplo 2

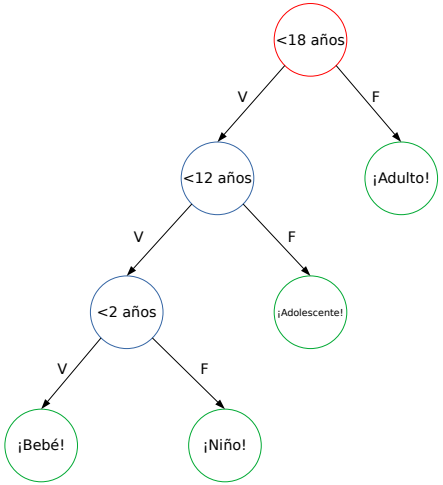


Figura: Árbol de decisión con más nodos.

Ejemplo 2

- En el árbol resultante de la figura 2 clasifica a las personas de acuerdo con su edad utilizando el criterio antes mencionado.

Ejemplo 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.

Ejemplo 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 los nodos que tienen flechas que llegan a él pero no salen de él, se les denomina como **nodos terminales** o de decisión.

Ejemplo 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 los nodos que tienen flechas que llegan a él pero no salen de él, se les denomina como **nodos terminales** o de decisión.
- Al nodo inicial del que sólo salen flechas de el pero no entran, se le denomina **nodo raíz** o simplemente **raíz**.

Ejemplo 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 los nodos que tienen flechas que llegan a él pero no salen de él, se les denomina como **nodos terminales** o de decisión.
- Al nodo inicial del que sólo salen flechas de el pero no entran, se le denomina **nodo raíz** o simplemente **raíz**.
- Los demás simplemente se les denomina como **nodos**.

Ejemplo 2

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

Ejemplo 2

- 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.

Ejemplo 2

- 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.

Ejemplo 2

- 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.

Ejemplo 2

- 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.

Ejemplo 3

Descripción del problema

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.

Ejemplo 3

Descripción del problema

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.

Ejemplo 3

Descripción del problema

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.

Ejemplo 3

Descripción del problema

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.

Ejemplo 3.

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: Resultados del estudio de cada uno de los pacientes.

Ejemplo 3

- La tabla 2 muestra la cantidad total de pacientes obtenidos por categoría a través del estudio.

Ejemplo 3

- 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

Ejemplo 3

- 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.

Ejemplo 3

- 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.

	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: Resultados totales del estudio de los pacientes.

Ejemplo 3

- Dado que se desea clasificar a los 303 pacientes de acuerdo con el síntoma que les ocasionó un infarto

Ejemplo 3

- 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.

Ejemplo 3

- 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.

Ejemplo 3

- 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.

Ejemplo 3

- En la tabla 3 se muestran los datos totales desglosados por síntoma y si sufrieron o no un ataque cardíaco.

Ejemplo 3

- 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 de dolor de pecho o no.

Ejemplo 3

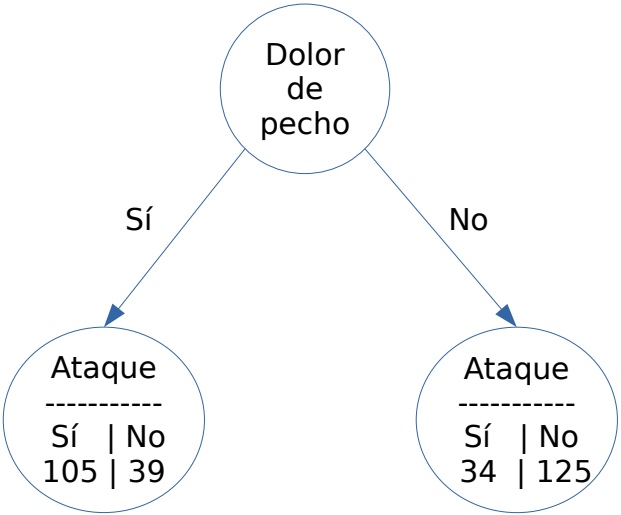
- En la tabla 3 se muestran los datos totales desglosados por síntoma y si sufrieron o no un ataque cardíaco.
- Utilizando estos datos se puede llevar a cabo la separación de los pacientes respecto a si sufren de dolor de pecho o no.
- La separación se muestra en la figura 3.

Ejemplo 3

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: Resultados Desglosados de los pacientes.

Ejemplo 3



Ejemplo 3

- Observe que la separación no es perfecta, dado que en cada rama se encuentran pacientes que han sufrido un ataque y otros que no.

Ejemplo 3

- 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**.

Ejemplo 3

- 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**.
- Resulta intuitivo buscar una separación que ocasione una menor impureza

Ejemplo 3

- 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**.
- Resulta intuitivo buscar una separación que ocasione una menor impureza
- Por lo que para medirla resulta indispensable una métrica.

Ejemplo 3

- 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**.
- 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).

Ejemplo 3

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

- Donde G representa el índice de impureza Gini

Ejemplo 3

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

- 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.

Ejemplo 3

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

- 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.

Ejemplo 3

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

- 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.

Ejemplo 3

- 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

Ejemplo 3

- 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}.

Ejemplo 3

- 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:

Ejemplo 3

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

Ejemplo 3

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

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

Ejemplo 3

- 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.

Ejemplo 3

- 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.

Ejemplo 3

- 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).

Ejemplo 3

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

- 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.

Ejemplo 3

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

Ejemplo 3

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

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

Ejemplo 3

- 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.

Ejemplo 3

- 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.

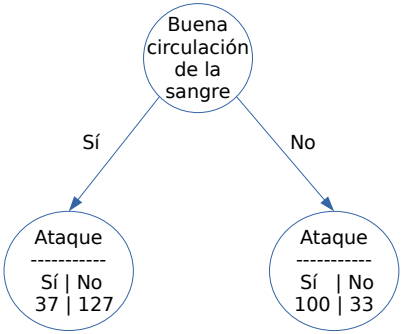


Figura: Separación mediante buena circulación de la sangre.

Ejemplo 3

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

Ejemplo 3

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

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

Ejemplo 3

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

Ejemplo 3

- 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.

Ejemplo 3

- 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.

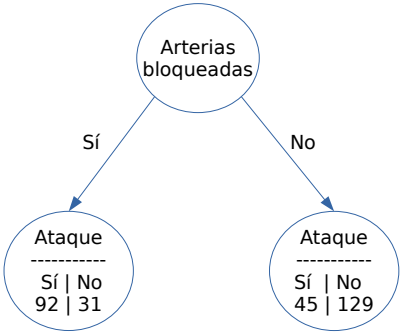


Figura: Separación mediante arterias bloqueadas.

Ejemplo 3

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

Ejemplo 3

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

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

Ejemplo 3

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

Ejemplo 3

- La impureza es un efecto no deseado en la separación de pacientes bajo cualquier criterio

Ejemplo 3

- 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.

Ejemplo 3

- 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:

Ejemplo 3

- 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:
  - $G_{DP} = 0.3641$
  - $G_{BC} = 0.36$
  - $G_{AB} = 0.3808$
- Se decide por G_{BC} debido a que su valor es el menor de todos.

Ejemplo 3

- 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:
  - $G_{DP} = 0.3641$
  - $G_{BC} = 0.36$
  - $G_{AB} = 0.3808$
- 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**.

Ejemplo 3

- 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:
  - $G_{DP} = 0.3641$
  - $G_{BC} = 0.36$
  - $G_{AB} = 0.3808$
- 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.

Ejemplo 3

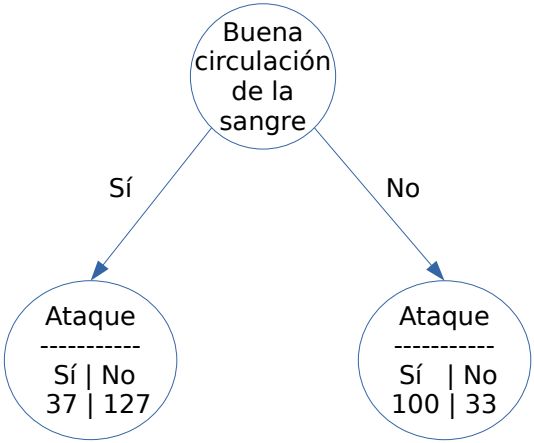


Figura: Separación mediante buena circulación de la sangre.

Ejemplo 3

- Ahora debemos separar los pacientes respecto a dolor de pecho o arterias bloqueadas para cada hoja.

Ejemplo 3

- Ahora debemos separar los pacientes respecto a dolor de pecho o arterias bloqueadas para cada hoja.
- En primer lugar, se analiza la rama verdadera (izquierda) del árbol de la figura 4.

Ejemplo 3

- Ahora debemos separar los pacientes respecto a dolor de pecho o arterias bloqueadas para cada hoja.
- 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.

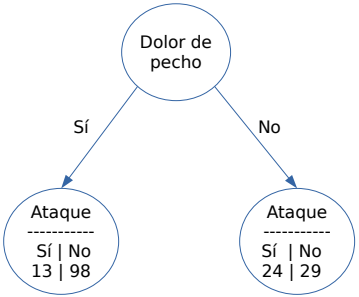


Figura: Separación de la rama izquierda del árbol de la figura 4 mediante dolor de pecho.

Ejemplo 3

- 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.

Ejemplo 3

- 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:

Ejemplo 3

- 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:

$$G_{Si} = 1 - \left(\frac{13}{13+98}\right)^2 - \left(\frac{98}{13+98}\right)^2 = 0.2068$$

$$G_{No} = 1 - \left(\frac{24}{24+29}\right)^2 - \left(\frac{29}{24+29}\right)^2 = 0.4955$$

$$G_{DP} = \frac{13+98}{13+98+24+29} \times 0.2068 + \frac{24+29}{13+98+24+29} \times 0.4955$$

$$G_{DP} = 0.3001$$

Ejemplo 3

- Ahora se realiza la separación mediante arterias bloqueadas.

Ejemplo 3

- 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

Ejemplo 3

- 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.

Ejemplo 3

- 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 8.

Ejemplo 3

- 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 8.

Ejemplo 3

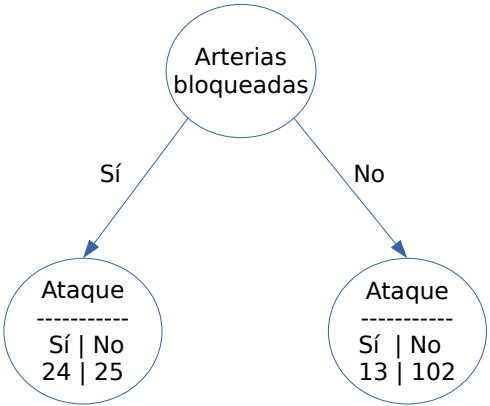


Figura: Separación de la rama izquierda del árbol de la figura 4 mediante arterias bloqueadas.

Ejemplo 3

Los cálculos correspondientes son los siguientes:

Ejemplo 3

Los cálculos correspondientes son los siguientes:

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

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

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

Ejemplo 3

- Dado que G_{AB} es menor que G_{DP} se elige arterias bloqueadas para realizar la separación de los pacientes.

Ejemplo 3

- 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 9.

Ejemplo 3

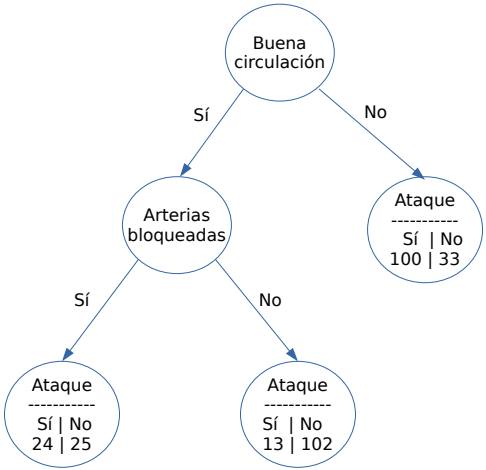


Figura: El árbol de decisión con sólo una rama separada mediante dos criterios.

Ejemplo 3

- Ahora se debe evaluar el separar los pacientes en la rama izquierda del árbol de la figura 9 mediante el dolor de pecho.

Ejemplo 3

- Ahora se debe evaluar el separar los pacientes en la rama izquierda del árbol de la figura 9 mediante el dolor de pecho.
- Esto se determina calculando la impureza mediante la separación a través del dolor de pecho

Ejemplo 3

- Ahora se debe evaluar el separar los pacientes en la rama izquierda del árbol de la figura 9 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

Ejemplo 3

- Ahora se debe evaluar el separar los pacientes en la rama izquierda del árbol de la figura 9 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.

Ejemplo 3

- Ahora se debe evaluar el separar los pacientes en la rama izquierda del árbol de la figura 9 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.

Ejemplo 3

- 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.

Ejemplo 3

- 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.

Ejemplo 3

- 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.

Ejemplo 3

- 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 10 muestra esta clasificación.

Ejemplo 3

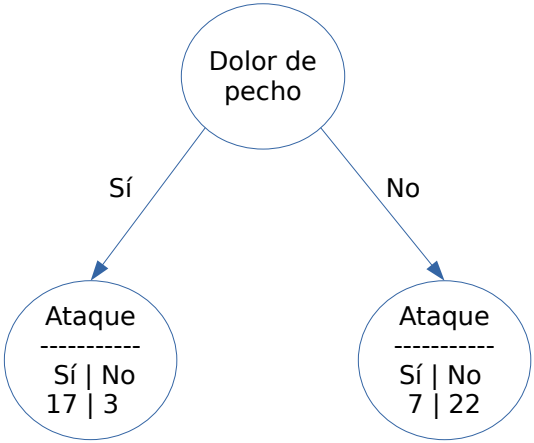


Figura: Separación de la rama izquierda afirmativa del árbol de la figura 9 mediante dolor de pecho.

Ejemplo 3

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

Ejemplo 3

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

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

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

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

Ejemplo 3

- En este punto se debe decidir si los últimos nodos son terminales.

Ejemplo 3

- 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 10 es menor que la impureza anterior G calculada en 3 entonces se hace la separación, si no es menor entonces estos nodos son terminales.

Ejemplo 3

- 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 10 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.

Ejemplo 3

- 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.

Ejemplo 3

- 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.

Ejemplo 3

- 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

Ejemplo 3

- 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.

Ejemplo 3

- 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 11 muestra estos datos de la separación.

Ejemplo 3

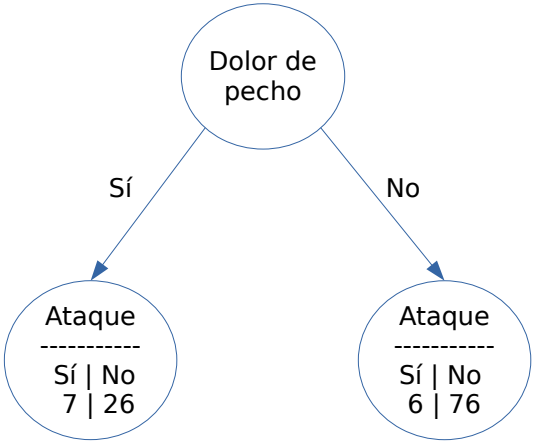


Figura: Separación de la rama izquierda negativa del árbol de la figura 9 mediante dolor de pecho.

Ejemplo 3

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

$$G_{Si} = 1 - \left(\frac{7}{7+26}\right)^2 - \left(\frac{26}{7+26}\right)^2 = 0.3342$$

$$G_{No} = 1 - \left(\frac{6}{6+76}\right)^2 - \left(\frac{76}{6+76}\right)^2 = 0.1356$$

$$G_{DP} = \frac{7+26}{7+26+6+76} \times 0.3342 + \frac{6+76}{7+26+6+76} \times 0.1356 = 0.1926$$

Ejemplo 3

- Dado que G_{DP} es menor que la impureza anterior calculada en la expresión 4, también se realiza la separación en esta rama.

Ejemplo 3

- 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 12.

Ejemplo 3

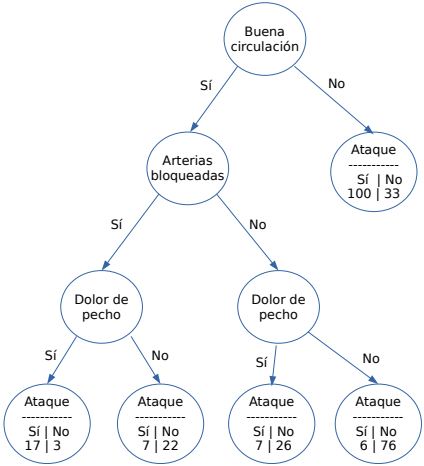


Figura: Separación completa de la rama izquierda del árbol.

Ejemplo 3

- A continuación solo falta repetir el mismo procedimiento para la rama derecha del árbol de la figura 12, el caso en el que los pacientes no tienen buena circulación.

Ejemplo 3

- A continuación solo falta repetir el mismo procedimiento para la rama derecha del árbol de la figura 12, 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.

Ejemplo 3

- A continuación solo falta repetir el mismo procedimiento para la rama derecha del árbol de la figura 12, 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 13.

Ejemplo 3

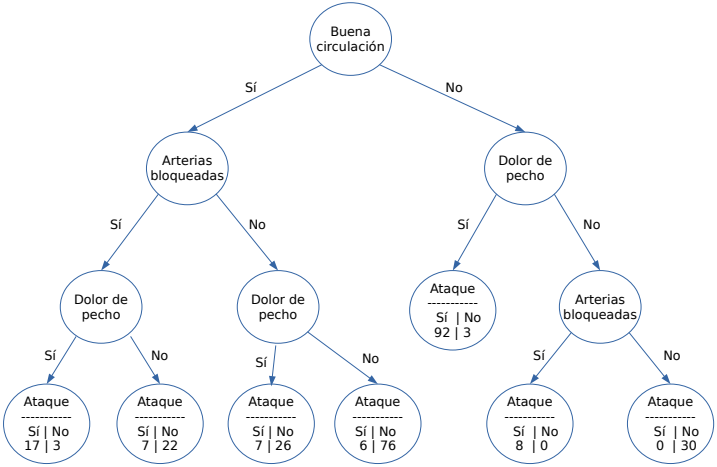


Figura: Árbol de decisión final para el ejemplo de los pacientes del estudio de la tabla 1.

Contenido

1. ¿Qué es un árbol de decisión?
2. Ejemplos
   - Ejemplo 1
   - Ejemplo 2
   - Ejemplo 3
3. Bibliografía

References I

L. Breiman.
Classification and Regression Trees.
CRC Press, 2017.

S. B. Gelfand, C. S. Ravishankar, and E. J. Delp.
An iterative growing and pruning algorithm for classification tree design.
IEEE, 13(2):163–174, 1991.

P. S. Laplace.
Teoría analítica de las probabilidades.
Courcier, 3 edition, 1820.

References II

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, 12:2825–2830, 2011.

Lior Rokach and Oded Maimon.
Data Mining with Decision Trees, volume 81 of *Series in Machine Perception and Artificial Intelligence*.
World Scientific, 2 edition, 2015.

José Unpingco.
Python for Probability, Statistics, and Machine Learning.
Springer, 2016.