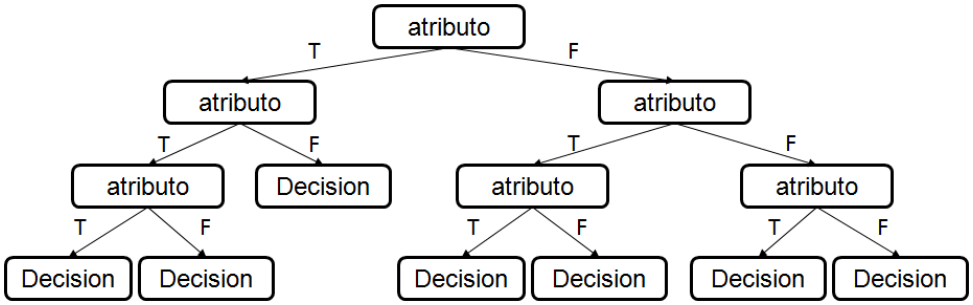
Un árbol de decisión es un modelo de predicción basado en construcciones lógicas que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva para la resolución de un problema, como por ejemplo clasificación binaria de muestras. Cada nodo representa la consulta sobre un estado y según la respuesta se navega el árbol hacia uno de sus hijos. Por otro lado, las hojas representan una solución o decisión final. En el caso de clasificación binaria, los nodos pueden representar atributos específicos y las hojas un estado de pertenecía o no a la clase que se desea clasificar. Es decir, en las hojas se decide si la muestra pertenece o no a la clase buscada.



Creación de un árbol de decisión

Se desea implementar un árbol de decisión que permita clasificar si los pacientes tienen una enfermedad cardiovascular a partir de los otros atributos conocidos en el siguiente dataset compuesto de 10 muestras:

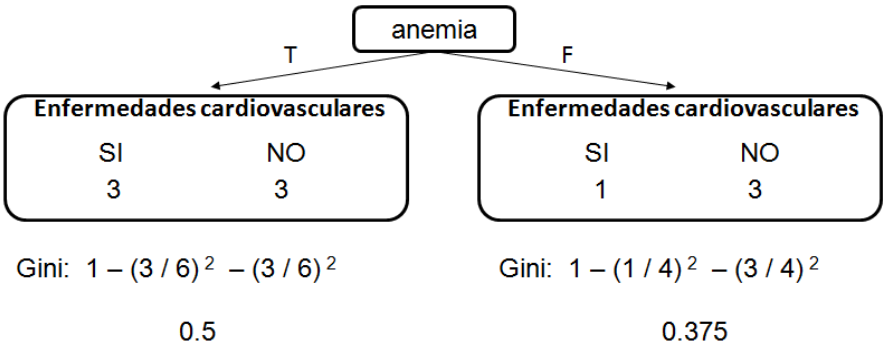
Anemia	Diabetes	Bloqueo de Arterias	Enfermedades cardiovasculares
no	no	no	no
si	no	no	no
si	si	si	si
si	si	no	no
no	no	si	no
si	si	no	si
no	si	si	si
si	si	no	no
no	no	no	no
si	si	si	si

Para crear el árbol de decisión se debe tomar un primer atributo como raíz y, a partir de este, segmentar el dataset según los valores de verdadero o falso. Dependiendo del atributo inicial seleccionado, la división del dataset se verá afectada de mejor o peor manera según el atributo seleccionado, por lo cual se debe seleccionar el atributo que mejor clasifique el objetivo (Enfermedades cardiovasculares). Para determinar el atributo que mejor clasifica el objetivo se pueden utilizar varias métricas, como por ejemplo el factor de impureza **Gini**. El factor de impureza de Gini de una hoja se mide según la siguiente fórmula:

$$1 - (\text{probabilidad SI})^2 - (\text{probabilidad NO})^2$$

Si se toma como raíz el atributo **Anemia**, se obtienen los siguientes valores:

Anemia	Enfermedades cardiovasculares
no	no
si	no
si	si
si	no
no	no
si	si
no	si
si	no
no	no
si	si



Gini: $1 - (3 / 6)^2 - (3 / 6)^2$

0.5

Gini: $1 - (1 / 4)^2 - (3 / 4)^2$

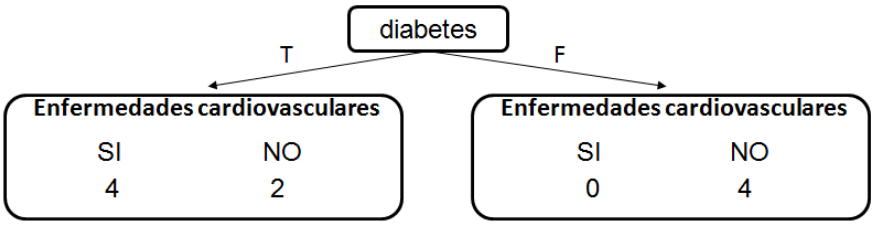
0.375

Para conocer el factor de impureza de Gini del atributo anemia, se requiere sumar ambos factores Gini según la proporción de pacientes de cada hoja. Dado que en la hoja izquierda se tiene 6 pacientes (3 SI, 3 NO), mientras en la hoja derecha se tiene 4 pacientes (1 SI, 3 NO), el valor total de Gini para anemia es:

$$Gini_{anemia} = (6/10) * 0.5 + (4/10) * 0.375$$
$$Gini_{anemia} = 0.45$$

Similarmente, si se toma como raíz el atributo **Diabetes**, se obtiene:

Diabetes	Enfermedades cardiovasculares
no	no
no	no
si	si
si	no
no	no
si	si
si	si
si	no
no	no
si	si



Gini: $1 - (4 / 6)^2 - (2 / 6)^2$

0.444

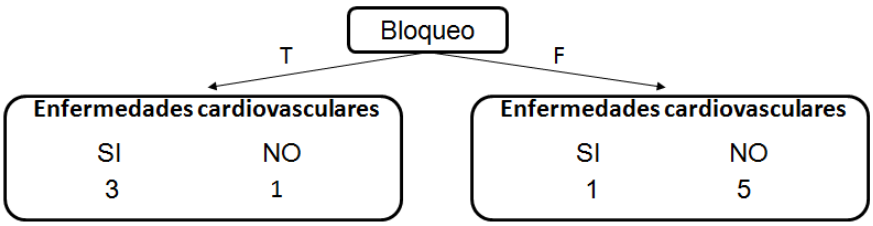
Gini: $1 - (0 / 4)^2 - (4 / 4)^2$

0

$$Gini_{diabetes} = (6/10) * 0.444 + (4/10) * 0$$
$$Gini_{diabetes} = 0.266$$

Si se toma como raíz el atributo **Bloqueo de Arterias**, se obtienen los siguientes valores:

Bloqueo de Arterias	Enfermedades cardiovasculares
no	no
no	no
si	si
no	no
si	no
no	si
si	si
no	no
no	no
si	si



Gini: $1 - (3 / 4)^2 - (1 / 4)^2$

0,375

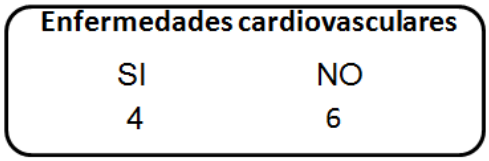
Gini: $1 - (1 / 6)^2 - (5 / 6)^2$

0.2778

$$Gini_{Bloqueo} = (4/10) * 0.375 + (6/10) * 0.2778$$
$$Gini_{Bloqueo} = 0.3166$$

Adicionalmente, se podría tomar la decisión de no dividir el dataset y crear un nodo hoja directamente con los valores objetivo (Enfermedades cardiovasculares). De ser ese el caso, el valor de Gini del nodo sería:

Enfermedades cardiovasculares
no
no
si
no
no
si
si
no
no
si



Gini: $1 - (4 / 10)^2 - (6 / 10)^2$

0,48

Por lo tanto, se puede dividir el dataset usando un nodo con cualquiera de los 3 atributos o no dividirlo, con lo que se obtendrían los siguientes factores de impureza:

$$Gini_{anemia} = 0.45$$
$$Gini_{diabetes} = 0.266$$
$$Gini_{Bloqueo} = 0.316$$
$$Gini_{Hoja} = 0.48$$

Dado que el Gini del atributo **Diabetes** es el menor de todos, se procede a segmentar el dataset con dicho atributo, creando un nodo y dividiendo el dataset en 2 datasets de menor tamaño.

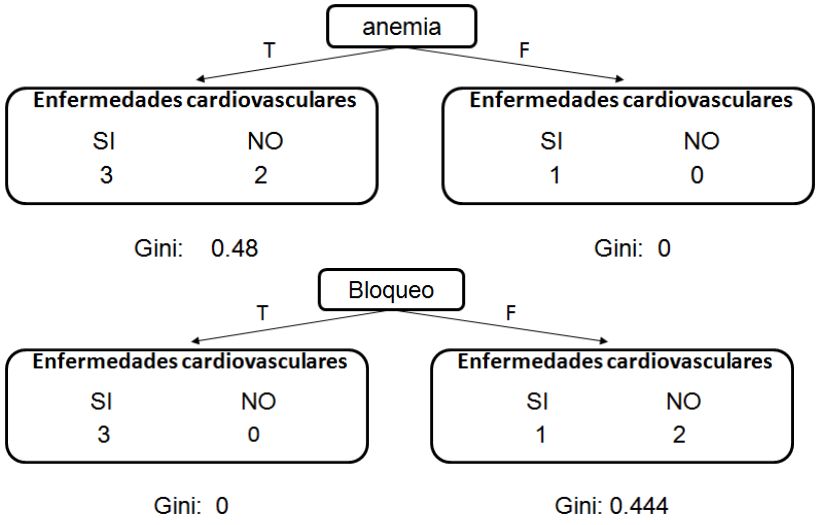
Anemia	Diabetes	Bloqueo de Arterias	Enfermedades cardiovasculares
si	si	si	si
si	si	no	no
si	si	no	si
no	si	si	si
si	si	no	no
si	si	si	si
no	no	no	no
si	no	no	no
no	no	si	no
no	no	no	no

Anemia	Bloqueo de Arterias	Enfermedades cardiovasculares
si	si	si
si	no	no
si	no	si
no	si	si
si	no	no
si	si	si

Anemia	Bloqueo de Arterias	Enfermedades cardiovasculares
no	no	no
si	no	no
no	si	no
no	no	no

Recursivamente se puede seguir dividiendo el dataset del lado izquierdo de la siguiente manera:

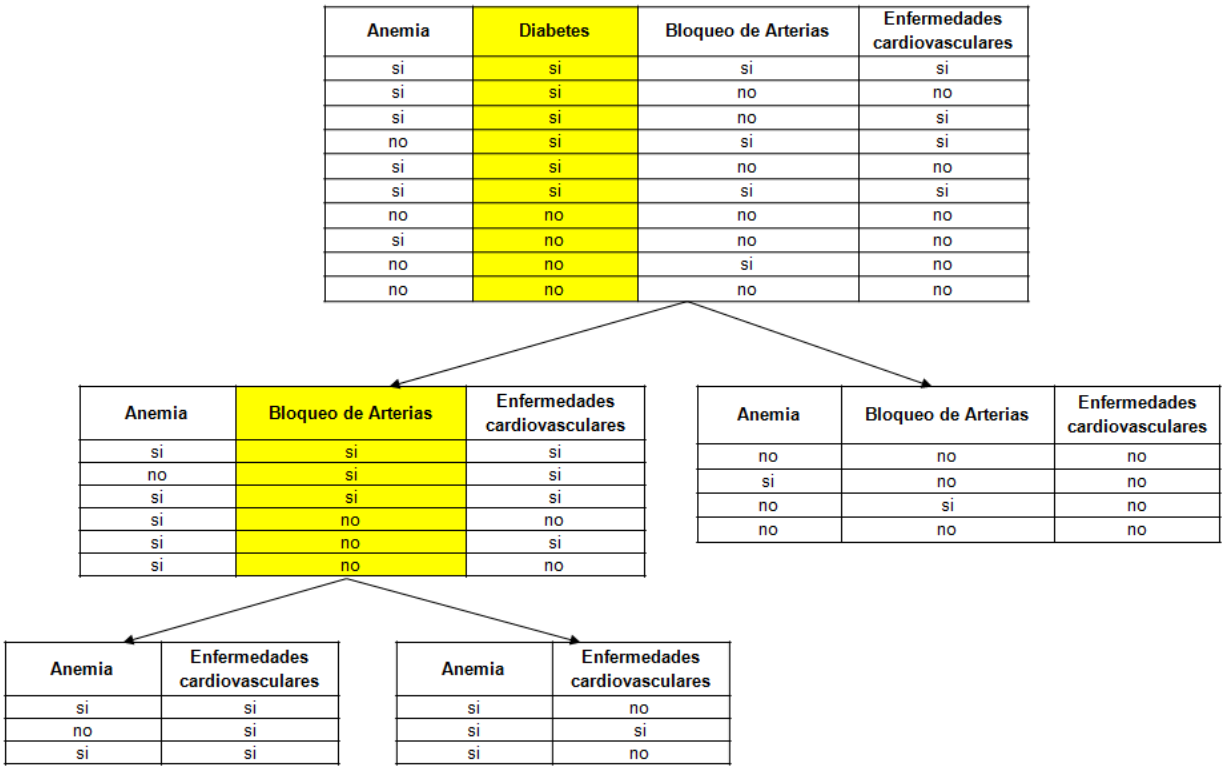
Anemia	Bloqueo de Arterias	Enfermedades cardiovasculares
si	si	si
si	no	no
si	no	si
no	si	si
si	no	no
si	si	si



Obteniendo los siguientes valores:

$Gini_{anemia} = 0.4$
$Gini_{Bloqueo} = 0.222$
$Gini_{Hoja} = 0.444$

Por lo que se elige dividir el dataset usando el atributo Bloqueo:



Anemia	Diabetes	Bloqueo de Arterias	Enfermedades cardiovasculares
si	si	si	si
si	si	no	no
si	si	no	si
no	si	si	si
si	si	no	no
si	si	si	si
no	no	no	no
si	no	no	no
no	no	si	no
no	no	no	no

Anemia	Bloqueo de Arterias	Enfermedades cardiovasculares
si	si	si
no	si	si
si	si	si
si	no	no
si	no	si
si	no	no

Anemia	Bloqueo de Arterias	Enfermedades cardiovasculares
no	no	no
si	no	no
no	si	no
no	no	no

Anemia	Enfermedades cardiovasculares
si	si
no	si
si	si

Anemia	Enfermedades cardiovasculares
si	no
si	si
si	no

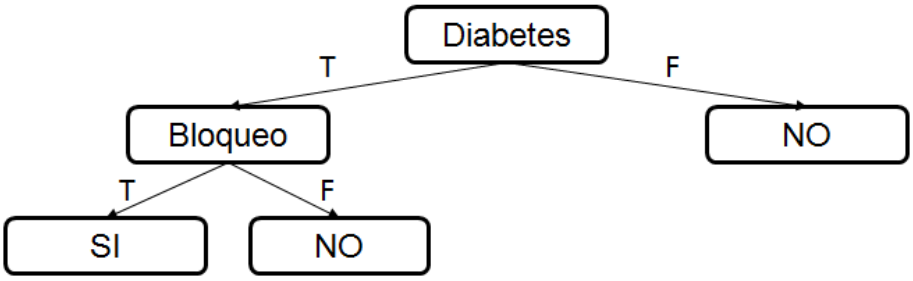
Al continuar recursivamente con los hijos creados se obtienen valores gini de cero o indeterminado, ya que todos los valores de atributo o de objetivo son todos iguales, por lo que no se divide más el dataset.

Regresando recursivamente al lado derecho de la raíz se tiene el siguiente dataset:

Anemia	Bloqueo de Arterias	Enfermedades cardiovasculares
no	no	no
si	no	no
no	si	no
no	no	no

El valor gini de mantener el dataset como hoja es cero, por lo cual no se dividen los datos.

Finalmente se asignan los valores a las hojas según el valor objetivo (Enfermedades cardiovasculares) que tenga más mediciones.



Para medir la eficiencia del árbol de decisión se prueban todas las muestras en el árbol y se tabulan los resultados el árbol. Luego se comparan con el valor original (Arbol vs Enfermedades cardiovasculares) obteniendo un porcentaje de acierto. En el ejemplo mostrado, se puede apreciar que el árbol acierta en 9 de loas 10 muestras, por lo cual su eficiencia es dde 90%

Anemia	Diabetes	Bloqueo de Arterias	Enfermedades cardiovasculares	Árbol
no	no	no	no	no
si	no	no	no	no
si	si	si	si	si
si	si	no	no	no
no	no	si	no	no
si	si	no	si	no
no	si	si	si	si
si	si	no	no	no
no	no	no	no	no
si	si	si	si	si

Porcentaje de acierto = 0.9 = 90%