Arboles de decisiones en el diagnóstico de enfermedades cardiovasculares

Decision trees in the diagnosis of cardiovascular diseases.

Guillermo Roberto Solarte Martínez[1], José A. Soto Mejía[2]
Universidad Tecnológica de Pereira, Pereira, Colombia
gsolarte294@gmail.com
jomejia@utp.edu.co

***Resumen*— En este artículo se presenta una descripción de los árboles de decisión y del algoritmo ID3 (Induction Decision tree) para determinar si se debe o no aplicar fármacos a paciente con enfermedades cardiovasculares.**
En esta investigación se demuestra empíricamente que es posible diagnosticar la necesidad de administrar fármacos en pacientes con síntomas de enfermedad cardiovascular, usando las variables presión arterial, índice de colesterol, azúcar en la sangre, alergias a antibióticos y otras alergias, mediante la utilización de árboles de decisión con el algoritmo ID3 (Induction Decision tree) implementado en el lenguaje Java.

***Palabras clave*— Arboles de decisión, Sistema Integrado de Gestión, Bases de Datos diagnóstico, enfermedades cardiovasculares.**

***Abstract*— In this paper is presented a description of a decision trees and the ID3 algorithm to determine whether or not to apply drugs to patients with cardiovascular diseases. It is also empirically shown that is possible to diagnose the necessity of administering drugs to patients with cardiovascular diseases based on arterial pressure, cholesterol index, level of sugar and other allergies by means of decision trees and the ID3 algorithm implemented in Java language.**

***Key Word* — Decision trees, Integrated Management Systems, Databases, Clinic History, diagnosis, cardiovascular disease.**

I. INTRODUCCIÓN

La Minería de Datos es una tecnología nueva de gran importancia que permite la integración de un conjunto de áreas [1] (estadística, inteligencia artificial, matemáticas, biología y medicina), además ayuda a identificar información oculta significativa que se encuentra en grandes volúmenes de datos [2], cuyo objetivo específico es que dicha información encontrada sirva de base para la toma de decisiones de acuerdo al caso de estudio.

En la segunda sección de este artículo se realiza una descripción de las técnicas de minería de datos, arboles de decisión y el algoritmo ID3. Posteriormente se demuestra empíricamente que es posible diagnosticar la necesidad de administrar fármacos en pacientes con síntomas de enfermedad cardiovascular, usando las variables presión arterial, índice de colesterol, azúcar en la sangre, alergias a antibióticos y otras alergias, mediante la utilización de árboles de decisión evaluados con el algoritmo ID3 (Induction Decision tree) utilizando una aplicación en java realizada por los autores.

II. TÉCNICAS DE MINERÍA DE DATOS

Los modelos de Minería de Datos se pueden aplicar en:
-Predicción de ventas.
-Clasificación y estratificación de Clientes.
-Determinar relaciones entre productos que generalmente se venden juntos.
-Buscar secuencias en el orden en que los clientes agregan productos a una canasta de compra.
-Diagnóstico médico.

Dentro de las principales técnicas de *Minería de Datos*[3] se encuentran:

- Técnicas de inferencia estadística.
- Visualización.
- Razonamiento basado en memoria.
- Detección de conglomerados.
- Análisis de vínculos.
- Árboles de decisión.
- Redes neuronales.
- Algoritmos genéticos.

A. Árboles de decisión

Los arboles de decisión es una de las técnica de aprendizaje inductivo supervisado no paramétrico, se utiliza para la predicción y se emplea en el campo de inteligencia artificial, donde a partir de una base de datos se construyen diagramas de construcción lógica, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una

[1] M.sc, Ingeniero de Sistemas
Fecha de Recepción: 25 de Agosto de 2011
Fecha de Aceptación: 12 de Noviembre de 2011

[2] Ph.d , Profesor Titular, Facultad de ingeniería Industrial

serie de condiciones que ocurren en forma repetitiva para la solución de un problema.

Propiedades de los Arboles de decisión.

Una de las propiedades de esta técnica es que permite una organización eficiente de un conjunto de datos, debido a que los árboles son construidos a partir de la evaluación del primer nodo (raíz) y de acuerdo a su evaluación o valor tomado se va descendiendo en las ramas hasta llegar al final del camino (hojas del árbol), donde las hojas representan clases y el nodo raíz representa todos los patrones de entrenamiento los cuales se han de dividir en clases. Los sistemas que implementan arboles de decisión tales como ID3 son muy utilizados en lo que se refiere a la extracción de reglas de dominio. Este método (ID3) se construye a partir del método de Hunt. La heurística de Hunt, consiste escoger la característica más discriminante del conjunto X, luego realizar divisiones recursivas del conjunto X, en varios subconjuntos disyuntos de acuerdo a un atributo seleccionado.

A. El algoritmo ID3

Una de las dificultades que se presenta al realizar el proceso de construcción de un árbol de decisión es escoger el atributo más apropiado. Este atributo debe ubicarse en la raíz del árbol para lo cual se debe realizar una prueba estadística a cada uno de los atributos que permite determinar que tan acertado se están clasificando los ejemplos de entrenamiento. Una vez se obtiene el atributo más apropiado, se selecciona y se utiliza como nodo prueba en la raíz del árbol, luego para cada uno de los otros atributos se procede a generar un nuevo descendiente. Los datos de entrenamiento son divididos y asignados al nodo descendiente adecuado, es decir, se organizan las ramas de acuerdo al valor que toma cada atributo. Este procedimiento se realiza recursivamente en cada nodo descendiente, utilizando los datos de entrenamiento correspondientes.

A continuación se describe el pseudocódigo del algoritmo ID3.

ID3 (Ejemplos,AtributoClasificador,Atributos)

Ejemplos son los datos de entrenamiento. *AtributoClasificador* es el atributo cuyo valor va a ser predecido por el árbol y que toma valores positivos o negativos. *Atributos* es una lista con otros atributos que pueden ser ensayados o candidatos a ser elegidos para ser la raíz de este árbol.

Inicio

- Crear un nodo *raíz* para el árbol.
- Si todos los ejemplos son positivos, regrese el nodo *raíz* tipo hoja con etiqueta positiva
- Si todos los ejemplos son negativos, regrese el nodo *raíz* tipo hoja con etiqueta negativa
- Si *Atributos* esta vacio, regrese el nodo *raíz* tipo hoja, con etiqueta igual al valor más común (la moda) del *Atributo Clasificador*.

De lo contrario- A ← el atributo de *Atributos* que mejor clasifique los ejemplos.

- Etiquetar el nodo raíz con el nombre de A (nodo tipo rama)
- Para cada posible valor *vi* de *A*
- Adicionar una nueva rama al nodo *raíz* para la prueba *A = vi*
- Hacer *Ejemplos vi* ← El subconjunto de *Ejemplos* donde *A = vi*
- *Si Ejemplos vi* está vacio
- Bajo esta nueva rama, adicionar un nodo hoja con etiqueta igual al valor más común (la moda) del *Atributo Clasificador*.

De lo contrario

- Bajo esta nueva rama, adicionar el subárbol:
ID3(*Ejemplos vi Atributo Clasificador* , *Atributos-{A}*)
- Regresar el nodo *raíz*

Fin

Código 1.0 Pseudocódigo del Algoritmo ide3 [3]

Para decidir qué atributo es el más apropiado a usar en cada nodo del árbol se utiliza una propiedad estadísticas llamada ganancia de información, que mide que tan bien clasifica ese atributo a los datos de entrenamiento. Así que elije el nodo del árbol que tenga mayor ganancia de información y luego expande sus ramas utilizando la misma metodología. La ganancia de información es una diferencia de entropías. El concepto de entropía se basa en la teoría de la información Esta teoría fue desarrollada inicialmente por Claude Shannon[4] a mediados del siglo XX. Vamos a aclarar el concepto de entropía usando el conjunto de datos utilizado en esta investigación. A un conjunto de pacientes con enfermedad cardiovascular, de acuerdo a un concepto médico, se les administra un fármaco según sea el valor de la presión, el azúcar en sangre, índice de colesterol y otras alergias (ver tabla 1).

Paciente	Presión	Azúcar en la sangre	Índice de colesterol	Alergias a antibióticos	Otras alergias	Administrar fármacos
1	Alta	Alto	Alto	NO	NO	SI
2	Alta	Alto	Alto	SI	NO	SI
3	Baja	Alto	Bajo	NO	NO	SI
4	Media	Alto	Alto	NO	SI	NO
5	Media	Bajo	Alto	SI	SI	NO
6	Baja	Bajo	Alto	SI	SI	SI
7	Alta	Bajo	Alto	SI	N0	SI
8	Alta	Bajo	Bajo	NO	SI	SI
9	Alta	Alto	Bajo	SI	SI	NO
10	Baja	Bajo	Alto	SI	SI	SI
11	Media	Bajo	Alto	SI	SI	SI
12	Alta	Bajo	Bajo	SI	SI	NO
13	Baja	Alto	Alto	SI	SI	SI
14	Baja	Alto	Bajo	NO	NO	SI

Tabla 1. Administración de fármacos Fuente: autores

[3] Tesis de maestría, UTP. Reinel Arias Montoya. Detección Temprana De Fallas en La Red De Internet Banda Ancha Aplicando Minería De Datos /Oct/2010

[4] http://www.nyu.edu/pages/linguistics/courses/v610003/shan.htm

En el grupo de datos S (vea tabla 1) que contiene valores positivos o negativos sobre una variable dicotómica para calcular la entropía de S relativa a su clasificación booleana se debe definir:

$P_{p;}$ es la probabilidad de que las respuestas sean positivas según el conjunto S.

$P_{n;}$ es la probabilidad de que las respuestas sean negativas según el conjunto S.

$(P_n = 1 - P_P)$ X son los datos de los pacientes. Se puede observar que de los 14 resultados, 10 tienen resultados positivos y 4 tienen resultados negativos. La probabilidad de cada resultado es:

$$P_P = \frac{10}{14} = 0{,}71 \quad y\ que \quad P_n = \frac{4}{14} = 0{,}28$$

La entropía de X se define con base a las probabilidades anteriores, así:

$$H(S) = -P_P log_2 P_P - P_n log_2 P_n \qquad (1)$$

Según la ecuación (1), la entropía del conjunto de los 14 datos respecto a la variable "Administrar fármacos" se calcula de la siguiente manera:

Si la entropía toma un valor de cero es cuando todos los miembros pertenecen a una misma clase ya sea negativa o positiva debido que $log(1)_2 = 0$.

Por tal motivo la entropía se encuentra siempre en un intervalo de cero a uno, alcanzando a un máximo cuando esta proporción es de 0,5 es decir existe una máxima aleatoriedad.

B. Concepto de ganancia de información

$$H(S) = \sum_{i=1}^{C} - P_i log_2(P_i) = -\frac{10}{14} log_2 \frac{10}{14} - \frac{4}{14} log_2 \frac{4}{14} = 0{,}86$$

Como se dijo anteriormente la entropía es una medida de desorden e impureza en un conjunto de datos. Para la clasificación de los datos se utiliza una medida llamada ganancia de información, esta medida reduce la entropía

$$H(A,S) \equiv \sum_{V \in Valores(A)} \frac{|S_v|}{|S|} = H(S_v) \qquad (2)$$

obtenida al realizar la división de los datos en los subconjuntos de entrenamiento.

Donde S es un grupo de muestras clasificadas en C clases, A son los atributos y S_v es un subconjunto de S

Valores. "A" es una lista de los posibles valores de cada atributo. La fórmula de ganancia de información se define como:

$$G(S, A) \equiv H(S) - H(S, A) \qquad (3)$$

En la anterior expresión el primer término H(S) corresponde a la entropía de S, el segundo término corresponde al valor esperado de la entropía después de que S ha sido particionado de acuerdo al atributo A.

Como podemos observar el segundo término de la fórmula de ganancia no es más que la sumatoria de entropías de cada subconjunto S_v, ponderado por la fracción $\frac{|S_v|}{|S|}$

Tablas de contingencia

Para facilitar los cálculos anteriores se usan tablas de contingencia. La tabla 2, de contingencia para la presión arterial se obtiene a partir de los datos de la tabla 1.

	Alto	Medio	Bajo	Total
Si	4	2	5	10
No	2	1	0	4
Total	6	3	5	14

Tabla 2 Respuestas de presión arterial
Fuente: autores

Para calcular la entropía del conjunto de datos, $H(S)$, se procede a calcular la entropía de cada uno de los valores de A, es decir, la entropía de Presión arterial (alta, media y baja),

$$H(S_{PA=alta}) = \sum_{i=1}^{C} - P_i log_2(P_i) = -\frac{4}{6} log_2 \frac{4}{6} - \frac{2}{6} log_2 \frac{2}{6} = 0{,}918$$

utilizando la tabla 2 de contingencia.

Luego se realiza de igual manera el cálculo de la entropía para la presión arterial media y baja.

Una vez se obtengan los tres valores se procede al cálculo de la ganancia de información con ese atributo utilizando la formula (3)

De la misma manera se realizan las tablas de contingencia para cada uno de los otros atributos mostrados en la tabla 1, y a partir de ellas se realizan los

$$G(S, PA) = 0{,}86 - \frac{6}{14} 0.918 - \frac{5}{14} 0 - \frac{3}{14} 0.918 = 0.272787$$

cálculos de entropía y de ganancia de información para cada una de ellos. Los resultados de ganancia de información de los todos los atributos se muestran en la tabla 3

Atributos	**Ganancia de Información**
Presión Arterial	*G (S, PA)= 0.2727*
Azúcar en Sangre	G (S, AZ) 0.0
Índice de Colesterol	G (S, IC) = 0.0207
Otras Alergias	G (S, OA) = 0.0195
Alergias a antibióticos	G (S, AA) =0.01495

Tabla3. Ganancia de Información para todos los atributos. Fuente: autores.

Ensamblaje Del Árbol

Como se observa, de la tabla 3, el atributo que se debe seleccionar como nodo raíz es la "Presión arterial" ya que de acuerdo con la medida de ganancia de información (0.2727) es el más adecuado para ser nodo inicial (raíz) creando así tres ramas.

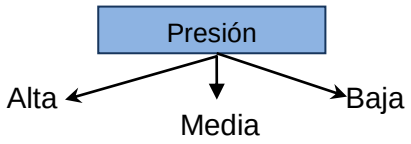

Figura 1.Selección del mejor atributo **"**Presión arterial"

A continuación se debe aplicar la misma técnica en cada uno de los nuevos nodos creados, pero en cada nodo creado solo se usa un subconjunto de los datos como se observa en las siguientes tablas 4 a 6.

Paciente	Presión Arterial	Azúcar en la sangre	Índice de colesterol	Alergias a antibióticos	Otras alergias	Administrar fármacos
1	Alta	Alto	Alto	NO	NO	SI
2	Alta	Alto	Alto	SI	NO	SI
7	Alta	Bajo	Alto	SI	N0	SI
S	Alta	Bajo	Bajo	NO	SI	SI
9	Alta	Alto	Bajo	SI	SI	NO
12	Alta	Bajo	Bajo	SI	SI	NO

Tabla 4.Datos de entrenamiento presión arterial alta

Paciente	Presión	Azúcar en la sangre	Índice de colesterol	Alergias a antibióticos	Otras alergias	Administrar fármacos
4	Media	Alto	Alto	NO	SI	NO
5	Media	Bajo	Alto	SI	SI	NO
11	Media	Bajo	Alto	SI	SI	SI

Tabla 5. Entrenamiento presión arterial media

Paciente	Presión	Azúcar en la sangre	Índice de colesterol	Alergias a antibióticos	Otras alergias	Administrar fármacos
3	Baja	Alto	Bajo	NO	NO	SI
6	Baja	Bajo	Alto	SI	SI	SI
10	Baja	Bajo	Alto	SI	SI	SI
13	Baja	Alto	Alto	SI	SI	SI
14	Baja	Alto	Bajo	NO	NO	SI

Tabla 6. Datos de entrenamiento presión arterial baja

Si observamos el comportamiento de los datos en la tablas anteriores podemos deducir que para el atributo "Presión Arterial = baja" la recursión tiende a terminar, ya que la variable "administrar fármacos" es positiva en todos los casos, por lo tanto este atributo queda apuntado a una hoja llamada baja con un valor de "si". Sin embargo las otras dos ramas restantes quedaran en evaluación recursiva, dividiendo el espacio de búsqueda y reduciendo el número de datos de entrenamiento. Igualmente se realiza el mismo procedimiento recursivo con los demás atributos hasta formar el árbol.

El método de ganancia de información en casos extremos presenta dificultades debido a que genera la misma cantidad de reglas como elementos tiene el conjunto de entrenamiento.
Para evitar esta dificultad se utiliza la medida de la "*proporción de ganancia de información*" propuesta por Quinlan[4]. La *proporción de ganancia* penaliza los atributos que tienen demasiados valores, incorporando un nuevo término llamado *información de la división*, el cual es sensitivo a qué tan amplia y uniformemente el atributo separa los datos**.**

$$\mathbf{Idiv(S,\,A)} \equiv -\sum_{i=1}^{C}\left|\frac{S_i}{S}\right|\boldsymbol{log}_2\left|\frac{S_i}{S}\right| \qquad (4)$$

Aquí S_1 hasta S_c son los c subconjuntos resultantes de particionar S de acuerdo al atributo A que tiene c valores distintos. La medida de *proporción de ganancia* se define en términos de la medida de ganancia en la ec. (5).

$$\mathbf{Pgan(S,A)} \equiv \frac{G(S,A)}{Idv(S,A)} \qquad (5)$$

Los resultados obtenidos para la *proporción de ganancia de información* de todos los atributos se muestran en la tabla 7.

Atributos	**Ganancia de Información**
Presión Arterial	*Pgan (S, PA)= 0.4620*
Azúcar en Sangre	*Pgan* (S, AZ) 0.0
Índice de Colesterol	*Pgan* (S, IC) = 0.18481
Otras Alergias	*Pgan* (S, OA) =0.01439
Alergias a antibióticos	*Pgan* (S, AA) =0.01495

Tabla 7. Tabla de Ganancia de Información
Fuente: autores

III. RESULTADOS DE LA EVALUACIÓN

En esta sección se presentan los resultados obtenidos al correr la aplicación en Java implementada por los autores de esta investigación para el mismo conjunto de datos mostrados en la tabla 1. Una vez que se tiene la matriz de ocurrencias se procede a formas las tablas de contingencias. A partir de cada tabla de contingencia se procede a realizar los cálculos de entropía. A continuación se muestra el fragmento de código para el cálculo de entropías para cada una de las tablas de contingencia previamente creadas. El siguiente es el código en java para el cálculo de la entropía.

```
public double calcular_entropia(double pp, double pn){
        double logbase2pp,logbase2pn,entropia=0;
        aux=pp+pn;
    if (aux!=0)
    {   pp = pp / aux;
        pn=pn/aux;
         logbase2pp=Math.log(pp) / Math.log(2);
         logbase2pn=Math.log(pn) / Math.log(2);
         entropia=((-1)*(pp*logbase2pp))-(pn*logbase2pn);
         }
    return entropia;
        }
```

Código 2.0 Calcula Entropía

El fragmento de código de *ganancia de información* y *proporción de ganancia*, se presenta a continuación.

```
public void gana(int e[][],int a, double u[]) throws ArithmeticException
{ double pp=0,sum=0;  int i,j=0,w=0,ps=0; double v4[]=
  new double [a]; v4[0]=0;
  double en=0,total=0,my=0.0,ac=0;
 mat=getNombresColumnas(); // visualiza los nombres de
 for (i = 0; i <a; i++)
          { if (j==0){ pp = e[j][i];
                    if(pp!=0)
                    {pp = pp / aux; en =pp*u[i];
                    sum=sum+en;}} }
     total = q - sum;    ac=total/sum;
     System.out.print("\n H(S)          "+ en);
     System.out.print("\n G(S,)          "+ total);
     System.out.print("\n Idiv =>          "+sum);
     System.out.print("\n pgan=G(S,)/Idiv    "+ ac);
     sum=0; total=0; en=0;pp=0;
}
```

Código3.0 De Ganancia de Información

Los resultados generados por la aplicación en Java resultado de esta investigación son:

- Entropía,
- Ganancia de información,
- Proporción de ganancia, y
- generación de las reglas de clasificación.

Tabla de Contingencia de Presión Arterial

6	3	5
4	2	0
2	1	5

Calculo de Entropía
[0] = 0.9182958340544896
[1] = 0.9182958340544896
[2] = 0.

Calculo de Ganancia de Información y proporción de ganancia
H(S) = 0.19677767872596202
G(S,) = 0.272787532388745
Idiv = 0.590333036177886
Pgan = 0.4620909142319209

Tabla de Contingencia de Azúcar en la Sangre

7	7
5	5
2	2

Calculo de Entropía
[0] = 0.863120568566631
[2] = 0.863120568566631

Calculo de Ganancia de Información y proporción de ganancia
H(S) = 0.4315602842833155
G(S,) = 0.0
Idiv = 0.863120568566631
Pgan = 0.0

Tabla de Contingencia Índice de Colesterol

7	3
2	2
9	5

Calculo de Entropía
[0] = 0.7219280948873623
[2] = 0.9182958340544896

Calculo de Ganancia de Información y proporción de ganancia
H(S) = 0.19199255746094904
G(S) = 0.20703137867809185
Idiv = 0.6560891898885391
Pgan = 0.31555371109416347

Tabla de Contingencia Alergia A Antibióticos

7	3
2	2
9	5

Calculo de Entropía
[0] = 0.72
[2] = 0.9181

Calculo de Ganancia de Información y proporción de ganancia
H(S) = 0.2578314624597723
G(S,) = 0.01495606992897247
Idiv = 0.8481644986376585
Pgan = 0.017633454303964922

Tabla de Contingencia Otras Alergias

7	3
2	2
9	5

Calculo de entropía
[0] = 0.7642045065086203
[2] = 0.9709505944546686

Calculo de Ganancia de Información y proporción de ganancia
H(S) = 0.2578314624597723
G(S,) = 0.01495606992897247
Idiv = 0.8949517866414867
Pgan = 0.035567522798417635

Atributos	Ganancia de información
Presión Arterial	*G (S, PA)= 0.272877*
Azúcar en Sangre	G (S, AZ) 0.0
Índice de Colesterol	G (S, IC) =0.207031
Otras Alergias	G (S, OA) = 0.01495
Alergias a antibióticos	G (S, AA) =0.01495

Tabla 8. Resumen de resultados de Ganancia de Información.

Atributos	Proporción de ganancia de Información
Presión Arterial	Pgan = 0.4620
Azúcar en Sangre	Pgan = 0.0
Índice de Colesterol	Pgan = 0.3155
Otras Alergias	Pgan = 0.0176
Alergias a antibióticos	Pgan = 0.0176

Tabla 9. Resumen resultados de la Proporción de Ganancia de Información

Generación de la reglas.

La aplicación implementada genera el siguiente grado de reglas de clasificación para la toma de decisiones [5].

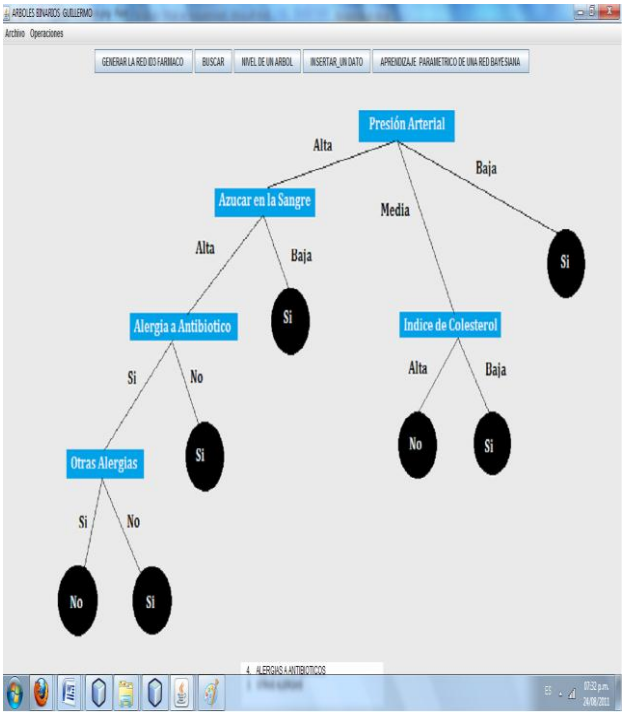

Figura 2. Generación de reglas de decisión con el algoritmo ide3 implementado.

De la figura 2 se deduce que para administrar fármacos a un paciente debemos tomar en cuenta las siguientes reglas generadas.
Se puede administrar fármacos a un paciente SI:

- ✓ Si tiene presión alta y si no es alérgico a antibiótico
- ✓ Si tiene presión alta y si es alérgico a antibiótico pero no tiene otras alergias
- ✓ Si tiene presión alta y si es alérgico a antibiótico y a otras alergias, pero tiene azúcar baja en la sangre
- ✓ Si tiene presión media y su índice de colesterol es bajo
- ✓ Si tiene presión arterial baja

NO se puede administrar fármacos a un paciente:

- ✓ Si tiene presión media e índice de colesterol alto.
- ✓ Si tiene presión alta y si es alérgico a antibiótico y a otras alergias, y además tiene azúcar alta en la sangre.

IV. CONCLUSIONES

Se observa que los resultados mostrados en la tabla 8, resumen de ganancia de información y tabla 9, resumen de proporción de ganancia de información que los valores generados son muy similares utilizando las dos técnicas mencionadas. En ambos métodos la variable que tiene mayor *ganancia* y *proporción de ganancia* es presión arterial, así que con ambos enfoques el nodo raíz es la Presión arterial creando así tres ramas Alta, Media, Baja.

-Se demostró empíricamente que es posible diagnosticar la administración o no, de fármacos en pacientes con síntomas de enfermedad cardiovascular, usando las variables presión arterial, índice de colesterol, azúcar en la sangre, alergias a antibióticos y otras alergias mediante la utilización de árboles de decisión.

-La técnica de árbol de decisión conjuntamente con el algoritmo ID3 entrega un conjunto de reglas entendibles que le permiten al médico ó al tomador de decisión hacerlo de manera rápida.

REFERENCIAS

[1] Giudici, Paolo. (2003). Applied Data Mining Stadistical Methods for Bussines and Industry. Chichester. Jhon Wiley & Sons, Inc. 364p

[2] Campell, Mary. base IV Guía de Autoenseñanza. España. Editorial McGraw Hill Interamericana. 1990. pp110/111,121/122,16,169, 179-191/192.. *(4 Mar 2009).*

[3] Larose, Daniel T. (2005). Discovering Knowledge in Data an Introduction to Data Mining. Hoboken, New Jersey. Jhon Wiley & Sons, Inc Publication. 222p.

[4] QUINLAN, J. R. Induction of Decision Trees. Machine Learning. P 81-106, 1986 QUINLAN, J. R. C4.5: Programs for machine learning. Morgan Kaufmann Publishers,1993.

[5] Cohen Karen Daniel. (1996). Sistemas de información para la toma de decisiones. México. McGraw-Hill. 243p. *(4 Mar 2009.*