TÉCNICAS DE CLASIFICACIÓN Y ANÁLISIS DE REPRESENTACION DEL CONOCIMIENTO PARA PROBLEMAS DE DIAGNÓSTICO

TECHNIQUES CLASSIFICATION AND ANALYSIS OF PROBLEMS OF REPRESENTATION OF DIAGNOSIS CONOCIMIENTOPARA

RESUMEN

El diagnostico medico es un proceso fundamental para la identificación de enfermedades, por eso en esta investigación se analizan algunas metodologías de clasificación de datos basados en patrones como son los arboles de decisiones, clúster y Análisis de componentes principales, estas técnicas de clasificación nos ayudan a la toma decisiones de una manera más eficiente y temprana.

PALABRAS CLAVES: Análisis de Componentes Principales, Arboles de Decisión, Arboles de regresión, Clúster, Data minig, factores de riesgo cardiovascular, inflamación, lipoproteínas de alta densidad, lipoproteínas de baja densidad, lipoproteínas de muy baja densidad, redes neuronales

ABSTRACT

The medical diagnosis is a fundamental process for the identification of diseases, so this research explores some methods of data classification based on patterns such as decision trees, cluster and principal components analysis, the technical classification helps us make decisions more efficiently and earlier.

KEYWORDS: Principal Component Analysis, decision trees, regression trees, Cluster, Data Mining, cardiovascular risk factors, inflammation, high-density lipoprotein, low density lipoprotein, very low density lipoproteins, neural networks

GUILLERMO ROBERTO SOLARTE MARTINEZ
Ingeniero Sistemas, Msc
Profesor Auxiliar
Universidad Tecnológica de Pereira
roberto@utp.edu.co

CARLOS ALBERTO OCAMPO S.
Ing. de Sistemas y Computación,
Esp. Auditoria de Sistemas
Mte Ciencias Computacionales, ©
Profesor Auxiliar
Universidad Tecnológica de Pereira
caos@utp.edu.co

1. INTRODUCCIÓN

Hoy en día se manejan grandes cantidades de información, esto hace que sea inevitable el uso de herramientas robustas para garantizar la fiabilidad del sistema, a pesar de esto, dentro de esta inmensa cantidad de datos todavía existe una enorme masa de información oculta, de gran importancia, a la que no se puede acceder por técnicas clásicas de recuperación de información. El descubrimiento de esta información oculta es posible gracias a la *Minería de Datos o Data Mining*, que nos permite la elaboración de resultados apropiados, pues durante la misma se aplica el algoritmo automático encargado de extraer el conocimiento inherente a los datos. Sin embargo, esta etapa se ve intervenida en gran medida por la calidad de los datos que llegan para su análisis desde la etapa previa.

En la medicina y en la investigación científica, el diagnóstico constituye una parte primordial, siendo una fase previa para la aplicación de las terapias o tratamientos médicos. Diagnosticar es equivalente a clasificar a un sujeto en una patología concreta. Cuando hablamos de clasificar a un sujeto en un grupo determinado, a partir de los valores de una serie de parámetros medidos u observados.

Dataming [1]. Utiliza varias técnicas de clasificación algunas de ellas son los arboles de decisiones, clúster, redes neuronales y métodos estadístico

Los arboles de clasificación o de decisión, se caracterizan por su sencillez, su campo de acción abarca diversas áreas como: el diagnostico medico, juegos, predicciones meteorológicas y control de calidad, este es el modelo de aprendizaje inductivo supervisado no paramétrico más utilizado. Como forma de representación del conocimiento.

Los arboles de decisión trabajan con dos opciones de entradas las cuales pueden ser un objeto o una situación descrita por medio de un conjunto de atributos y a partir de esto devuelve una respuesta que al final es una decisión que es tomada a partir de las entradas. Los valores que pueden tomar las entradas y las salidas

Fecha de Recepción: 8 de julio de 2009
Fecha de Aceptación: 22 de julio de 2009

pueden ser valores discretos o continuos. Se utilizan más los valores discretos por simplicidad, cuando se utilizan valores discretos en las funciones de una aplicación se denomina clasificación y cuando se utilizan los continuos se denomina regresión[3].

Una de las utilidades de los arboles de decisión es que se pueden descubrir patrones en los datos, estos datos se recogen y se organizan en modelos que se utilizaran posteriormente, los modelos pueden ser descritos como gráficos o árboles (los árboles son los gráficos en los que cualquiera de los dos nodos están conectados exactamente un camino)

Arboles de Clasificación:

- El Algoritmo C5.0 Incorpora:

- Ponderación de errores de clasificación y los costos.
- Generación y combinación de múltiples modelos para mejorar la precisión.
- Realiza la selección de los atributos más útiles y los utiliza para generar el modelo.

El análisis de clúster es una técnica que básicamente consiste en agrupar un conjunto de observaciones en un número dado de clúster, este agrupamiento se basa de la distancia o similitudes entre las observaciones.

La obtención de dichos clúster [5] depende del criterio o distancia considerados; así, por ejemplo, una baraja de cartas españolas se podría dividir de distintos modos: en cuatro clústers (los cuatro palos), en ocho clústers (los cuatro palos y según sean figuras o números),

2. CONTENIDO

En este trabajo de investigación se realizara una comparación de las técnicas de clasificación estadística, Análisis de Componentes Principales, Clúster y Arboles de Decisión, la cual nos va ayudar a determinar ¿cuál es la mejor opción para la búsqueda de patrones en enfermedades complejas?

Empezamos por explicar la técnica estadística, que es el Análisis de Componentes Principales, para solucionar este problema de clasificación se utiliza el programa SPSS 1.3 para Windows, para este análisis de datos.

2.1 ACP, pertenece a unas series de técnicas estadísticas multivariantes, eminentemente descriptivas, se utiliza en grandes masas de datos, su principal objetivo es reducir la dimensionalidad de los datos, transformando el conjunto de **p** variables originales en otro conjunto de **q** variables incorrelacionadas llamadas componentes principales.

Este análisis nos permite trabajar con dos opciones: Usar la matriz de correlaciones o bien, la matriz de covarianzas. En la primera opción se le está dando la misma importancia a todas y a cada una de las variables; esto puede ser conveniente cuando se considera que todas las variables son igualmente relevantes. La segunda opción se puede utilizar cuando todas las variables tengan las mismas unidades de medida y considerando también su grado de variabilidad

Una de las formas de utilizarla es realizar unas combinaciones lineales de las variables originales a estos componentes, de tal manera que se ordenen en función del porcentaje de varianza. En este sentido, el primer componente será el más importante porque es el que explica el mayor porcentaje de la varianza de los datos.

Además este estudio se realiza en el espacio de las variables y, en forma dual, en el espacio de los individuos. Se acostumbra a representar gráficamente los puntos-variables y los puntos-individuos tomando como ejes de coordenadas los componentes. A veces, puede facilitar la interpretación de los resultados, al observar la similar ubicación de los puntos en los planos respectivos. Aunque el plano de puntos-variables no se superpone al plano de puntos-individuos, es de gran utilidad "interpretar" la cercanía de un grupo de puntos-individuos, a ciertas variables.

Abreviaturas: ACP. Análisis de Componentes Principales. *CV* Cardiovascular. *EC:* Enfermedad coronaria. *HDL:* Lipoproteínas de alta densidad. *IAM:* Infarto agudo del miocardio. *IMT:* Grosor íntima/media. *IVUS:* Ultrasonido intravascular. *LDL:* Lipoproteínas de baja densidad. *Lp(a):* Lipoproteína (a). *Lp-PLA$_2$:* Fosfolipasa A_2 asociada a las lipoproteínas. *LPL:* Lipoprotein lipasa. *SM:* Síndrome metabólico. *PCR-us:* Proteína C-reactiva ultra sensible. *RMN:* Resonancia magnética nuclear. *TG:* Triglicéridos. *VLDL:* Lipoproteínas de muy baja densidad. ACP análisis de componentes principales.

2.2 Base de Datos Lípidos

La base de datos estudiada corresponde a un estudio entre 100 individuos a los cuales se les tomo 16 muestras diferentes de centro del Saludcoop EPS de municipio de santa rosa-Chinchina. , las variables estudiadas nos ayudan a determinar la salud cardiovascular.

La edad de los individuos oscila entre los 19 y 40 años. La muestra está compuesta por 25 mujeres y 75 hombres. El colesterol esta compuesto por el LDL, VDL y el HDL. La presión arterial se mide: presión arterial sistólica y la presión arterial diastólica.

2.3 Componentes principales

El análisis de componentes principales se utiliza para reducir variables, para identificar un número pequeño de componentes que expliquen la mayoría de la varianza total observada. Usando el SPSS tenemos:
Correlación entre las variables; como p–valor 0,000 se puede concluir que existe correlación significativa entre las variables.

Niveles de colesterol total	
Menos de 200 mg/dL	Nivel "deseable" que le expone a menos riesgo de enfermedades del corazón
200–239 mg/Dl	Límite alto. Un nivel de colesterol de 200 mg/dL o más aumenta el riesgo
240 mg/dL y más	Colesterol "alto". Una persona con ese nivel tiene más del doble de riesgo que una persona con nivel deseable

Figura 1. Niveles de colesterol total

Niveles de colesterol LDL	
Menos de 100 mg/Dl	Óptimo
100–129 mg/dL	Cerca o por encima del valor óptimo
130–159 mg/dL	Límite alto
160–189 mg/dL	Alto
190 mg/dL y más	Muy alto

Figura 2. Niveles de colesterol LDL

Niveles de colesterol HDL	
Menos de 40 mg/dL (hombres) Menos de 50 mg/dL (mujeres)	Colesterol HDL bajo, este nivel aumenta el riesgo de enfermedad cardiovascular.
60 mg/dL y más	Colesterol HDL alto (óptimo). Este nivel reduce el riesgo de enfermedad cardiovascular

Figura 3. Niveles de colesterol HDL

Niveles de Triglicéridos	
Menos de 150mg/dL	Normal
150-199 mg/dL	Límite Alto
200-499 mg/dL	Alto
500 mg/dL o más	Muy Alto

Figura 4. Niveles de Triglicéridos

Figura 5. KMO y Prueba de Bartlett

KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		.256
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	1805.337
	gl	66
	Sig.	.000

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.
Edad	100	19.0	40.0	24.170	3.2785
Peso	100	107.0	234.0	158.640	27.7949
Colesterol	100	115.0	285.0	190.820	35.2060
Triglicéridos	100	43.0	480.0	97.350	59.5065
HDL	100	26.0	71.0	45.180	9.9518
LDL	100	72.1	223.2	144.087	32.5861
Peso_Ideal	100	70.9	152.0	100.397	13.4257
Altura	100	59.0	80.0	69.510	4.0414
Grasa_piel	100	4.0	42.0	18.050	8.0332
SystolicBP	100	100.0	138.0	123.330	6.7809
DiastolicBp	100	60.0	144.0	77.750	9.6634
Eje_R/min	100	.0	360.0	80.450	77.7853
N válido (según lista	100				

Figura 6. Estadísticos Descriptivos

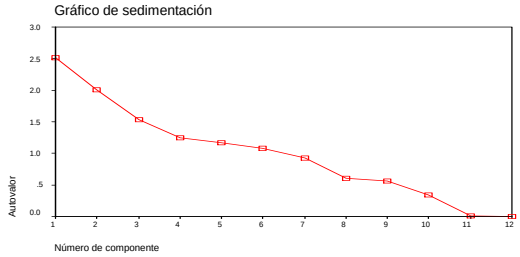

Figura 7. Segmentación

Varianza total explicada						
	Auto valores iníciales			Sumas de las saturaciones al cuadrado de la extracción		
#	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	2.510	20.91	20.9	2.51	20.91	20.913
2	2.010	16.74	37.6	2.01	16.74	37.660
3	1.538	12.81	50.4	1.53	12.81	50.474
4	1.245	10.37	60.8	1.24	10.37	60.849
5	1.171	9.756	70.6	1.17	9.756	70.605
6	1.077	8.972	79.5	1.07	8.972	79.577
7	.927	7.725	87.3			
8	.602	5.021	92.3			
9	.562	4.685	97.0			
10	.345	2.877	99.8			
11	1.379E-02	.115	100.			
12	2.594E-07	2.162E-06	100.000			
Método de extracción: Análisis de Componentes principales.						

Figura 8. Grafica de varianza total explicada

Las primeras seis variables explican el 79.58% de la varianza total, en el gráfico de sedimentación se presentan seis valores propios mayores que uno, con lo que las seis variables resumirán al resto representándolas de forma coherente; es decir serán seis componentes principales que resumen toda la información.

Observando la matriz de componentes principales y el gráfico de componentes se ve que la primera componente esta correlacionada fuerte y positivamente con las variables Colesterol, LDL, Triglicéridos y Edad y negativamente con las variables Altura y Eje_R/min, esto significa que estas seis variables son las que más aportan a este componente. Este componente nos está indicado que riesgo cardiovascular.

En el componente dos, las variables más correlacionadas positivamente son Peso, Altura y SystolicBP y negativamente con el Peso_Ideal. Por lo cual a este componente lo podemos llamar actividad del corazón. El tercer componente las variables que se relaciona fuerte y positivamente con Peso_Ideal y Grasa_piel, dada la naturaleza de estas variables, podríamos considerar a este componente indica cuerpo ideal. El cuarto componente fuerte y positivamente a las variables DiastolicBp y SystolicBP, al cual podemos denominar como corazón sano. El quinto componente relaciona fuerte y positivamente las variables Eje_R/min y Peso_Ideal, estas variables nos indican sobre las personas que están en forma. El componente seis, relaciona fuerte y positivamente las variables HDL, Altura y Edad, que nos pueden indicar bajo riesgo de enfermedad cardiovascular.

Matriz de componentes(a)

	Componente					
	1	2	3	4	5	6
Edad	.423	-2.796E-02	.204	-.213	2.879E-02	.346
Peso	8.764E-02	.809	.301	-.141	.413	.160
Colesterol	.917	-.107	-.292	-1.025E-02	8.630E-02	9.594E-02
Triglicéridos	.602	.327	-4.917E-02	-.142	-.346	-.416
HDL	.248	-.565	3.514E-02	.170	.127	.669
LDL	.898	4.705E-02	-.324	-5.888E-02	6.440E-02	-8.839E-02
Peso Ideal	.310	-6.763E-02	.724	.171	.454	-.258
Altura	-.144	.814	-.260	-.254	6.812E-02	.390
Grasa piel	.216	-.141	.630	-.392	-7.896E-02	-4.308E-04
SystolicBP	6.533E-02	.374	.279	.598	-.318	.112
DiastolicBp	.285	.296	9.445E-02	.647	-.201	.107
Eje_R/min	-6.391E-02	9.795E-04	-.369	.324	.703	-.228

Método de extracción: Análisis de componentes principales. a 6 componentes extraídos

Figura 9. Matriz de Componentes

La matriz de coeficientes para el cálculo de las puntuaciones en las componentes:
C1= 0,168 Edad + 0,035 Peso + 0,365 Colesterol + 0,240 Triglicéridos + 0,099 HDL + 0,368 LDL + 0,124 Peso_Ideal – 0,057 Altura + 0,086 Grasa_Piel + 0,026 SystolicBP + 0,114 DiastolicBp - 0,025 Eje_R/min.

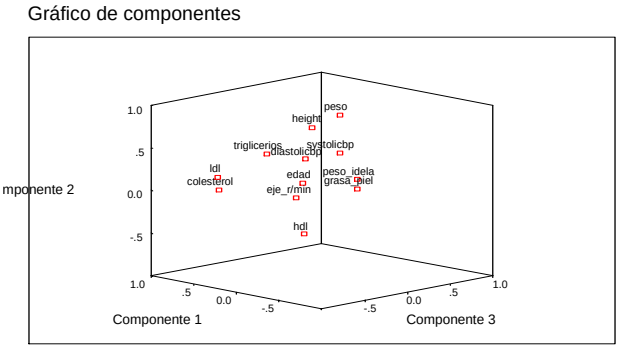

Figura 10. Gráficos de Componentes

C2= -0,14 Edad + 0,403 Peso - 0,53 Colesterol + 0,163 Triglicéridos - 0,281 HDL + 0,023 LDL -0,34 Peso_Ideal + 0,405 Altura - 0,070 Grasa_Piel + 0,186 SystolicBP + 0,147 DiastolicBp - 0,000 Eje_R/min

C3= 0,133 Edad + 0,196 Peso - 0,190 Colesterol - 0,032 Triglicéridos + 0,023 HDL - 0,211 LDL - 0,471 Peso_Ideal - 0,169 Altura + 0,410 Grasa_Piel + 0,181 SystolicBP + 0,061 DiastolicBp - 0,240 Eje_R/min

C4= -0,171 Edad - 0,113 Peso - 0,008 Colesterol - 0,114 Triglicéridos + 0,137 HDL - 0,047 LDL +0,137 Peso_Ideal - 0,204 Altura + 0,315 Grasa_Piel - 0,480 SystolicBP - 0,519 DiastolicBp + 0,260 Eje_R/min.

C5= 0,025 Edad + 0,352 Peso + 0,074 Colesterol - 0,295 Triglicéridos + 0,109 HDL + 0,055 LDL + 0,388 Peso_Ideal + 0,058 Altura - 0,067 Grasa_Piel - 0,272 SystolicBP - 0,172 DiastolicBp + 0,600 Eje_R/min

C6= -0,321 Edad + 0,149 Peso + 0,089 Colesterol - 0,386 Triglicéridos + 0,621 HDL - 0,082 LDL - 0,240 Peso_Ideal + 0,362 Altura + 0,000 Grasa_Piel + 0,104 SystolicBP + 0,099 DiastolicBp - 0,211 Eje_R/min.

Analizando las puntuaciones que se han guardado como fact1_1, fact1_2, fact1_3, fact1_4, fact1_5 y fact1_6. Observamos que:

Significados de los componentes**:**

	Variables	Nombre
C1	Colesterol, LDL, Triglicéridos	Riesgo cardiovascular alto
C2	Peso, Altura y HDL	Índice de masa muscular
C3	Peso Ideal, Grasa Piel	Porcentaje de masa corporal
C4	Presión diástolica y sistólica	Presión arterial
C5	Ejercicio, Peso Ideal	Corazón sano
C6	HDL.	Protector de riesgo cardiovascular

Figura11. Significado de los componentes

2.4 Análisis de Conglomerados Cluster

Permite agrupar variables, en función de la similaridad existente entre ellos. Usamos los conglomerados jerárquicos. Los datos dentro de cada conglomerado, son similares entre sí (alta homogeneidad interna) y diferentes a los objetos de los otros conglomerados (alta heterogeneidad externa).

Método de aglomeración: Vecino más cercano

a Distancia euclídea usada

Matriz de distancias

	Archivo matricial de entrada											
Caso	Edad	Peso	olesterol	gliceric	HDL	LDL	eso_ide	Altura	rasa_pi	ystolic	Eastolic	je_R/m
Edad	.000	3.392	2.296	2.501	2.685	2.647	2.969	4.367	3.501	3.737	4.373	4.763
Peso	3.392	.000	4.254	3.263	5.799	3.729	1.420	7.808	3.490	2.836	3.097	3.838
Colesto	2.296	4.254	.000	0.956	1.336	2.755	3.313	4.635	3.626	4.515	2.822	3.822
Triglic	2.501	3.263	0.956	.000	5.855	0.162	3.792	3.796	3.361	2.975	3.030	4.831
HDL	2.685	5.799	1.336	5.855	.000	3.484	3.670	5.740	3.248	4.374	3.874	4.196
LDL	2.647	3.729	2.755	0.162	3.484	.000	3.386	4.161	3.859	4.489	2.815	3.739
Peso_i	2.969	1.420	3.313	3.792	3.670	3.386	.000	6.748	1.810	3.365	3.354	3.804
Altura	4.367	7.808	4.635	3.796	5.740	4.161	6.748	.000	4.994	3.467	3.615	4.085
Grasa_	3.501	3.490	3.626	3.361	3.248	3.859	1.810	4.994	.000	4.564	4.333	5.880
Systoli	3.737	2.836	4.515	2.975	4.374	4.489	3.365	3.467	4.564	.000	1.636	4.568
Diasto	4.373	3.097	2.822	3.030	3.874	2.815	3.354	3.615	4.333	1.636	.000	4.269
Eje_R/	4.763	3.838	3.822	4.831	4.196	3.739	3.804	4.085	5.880	4.568	4.269	.000

Figura12. Matriz de distancia

Si elegimos los tres primeros grupos que se forman encontramos que las primeras agrupaciones son: Colesterol, LDL, Triglicéridos y HDL, Sístole y diástole, Peso, Altura, Peso Ideal y Grasa en la piel.

Método de aglomeración: Vecino más cercano

a Distancia Coeficiente Pearson

Matriz de distancias

	Archivo matricial de entrada											
Caso	Edad	Peso	olesterol	gliceric	HDL	LDL	eso_ide	Altura	rasa_pi	ystolic	Diastolic	Eje_R/mi
Edad	.000	.094	.236	.211	.187	.192	.151	-.042	.079	.047	-.043	-.101
Peso	.094	.000	-.026	.112	-.261	.048	.341	.692	.081	.168	.134	.033
Coleste	.236	-.026	.000	.394	.351	.962	.105	-.082	.062	-.064	.170	.035
Triglice	.211	.112	.394	.000	-.270	.478	.039	.039	.098	.150	.143	-.111
HDL	.187	-.261	.351	-.270	.000	.082	.056	-.251	.114	-.044	.028	-.018
LDL	.192	.048	.962	.478	.082	.000	.095	-.013	.030	-.060	.171	.047
Peso_ic	.151	.341	.105	.039	.056	.095	.000	-.417	.296	.098	.099	.038
Altura	-.042	.692	-.082	.039	-.251	-.013	-.417	.000	-.136	.084	.064	-.002
Grasa_p	.079	.081	.062	.098	.114	.030	.296	-.136	.000	-.071	-.038	-.274
Systolic	.047	.168	-.064	.150	-.044	-.060	.098	.084	-.071	.000	.316	-.072
Diastoli	-.043	.134	.170	.143	.028	.171	.099	.064	-.038	.316	.000	-.028
Eje_R/r	-.101	.033	.035	-.111	-.018	.047	.038	-.002	-.274	-.072	-.028	.000

Figura13. Matriz de distancia Person

Diagrama de témpanos vertical

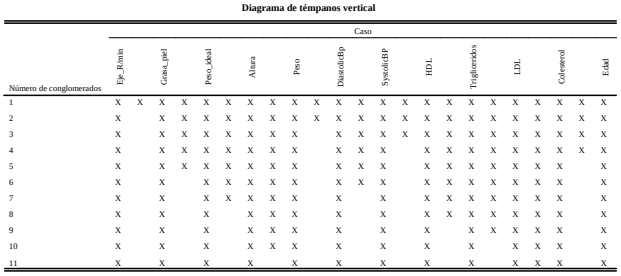

Figura14. Diagrama de témpanos verticales

Si elegimos los tres primeros grupos que se forman encontramos que las primeras agrupaciones son: Colesterol, LDL, Triglicéridos y HDL, Sístole y diástole, Peso, Altura y Peso Ideal.

2.5 Árbol de Decisión

Es una técnica de predicción que se emplea en el campo de inteligencia artificial [2]., donde a partir de una base de datos se construyen diagramas de construcción lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren en forma repetitiva para solución de un problema.

Por lo general en esta técnica de construcción de arboles de clasificación, se toma la próxima partición de manera optima en el conjunto del árbol, esto evita la confusión combinatoria en cuanto numero de decisiones futuras a considerar, por eso hay que elegir la medida justa a optimizar en cada corte, para facilitar las próximas divisiones.

Los paso a seguir de esta técnica son los siguientes :

1. Aprendizaje: Consisten en la construcción del árbol a partir de un espacio muestral X, este paso es el más complejo, y de él depende el resultado final.

2. Clasificación: en este paso se realiza el etiquetado de un patrón W, independiente del conjunto de aprendizaje, donde se trata de responder a los cuestionamientos asociados a nodos interiores, utilizando un parámetro de patrón W, este proceso se repite desde la raíz , hasta alcanzar una hoja, siendo el camino impuesto por los resultado de cada evaluación.

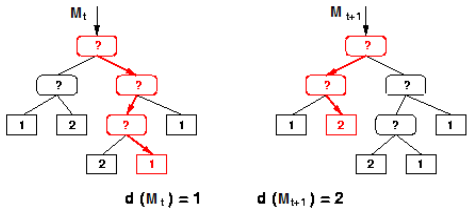

Figura 15. Árbol de Decisión

LDL, Triglicéridos, Peso y Edad. El problema consiste en decidir si una persona es propensa a tener Riesgo Cardiovascular, tomando en cuenta los siguiente parámetros de evaluación: LDL, Triglicéridos, Peso y Edad. Considerando un Conjunto de aprendizaje en el que los patrones están compuestos por atributos categóricos y la clase cierta asociada es Si o No algunos de estos prototipos serán:

{ LDL = optimo , Edad= joven
Triglicéridos=Normal, Si}

{ LDL = optimo , Edad= adulto
Triglicéridos=Limite alto, No}

{ LDL = alto , Edad= adulto
Triglicéridos=Limite alto, Si}

{ LDL = muy alto , Edad= adulto
Triglicéridos=Limite alto, Si}

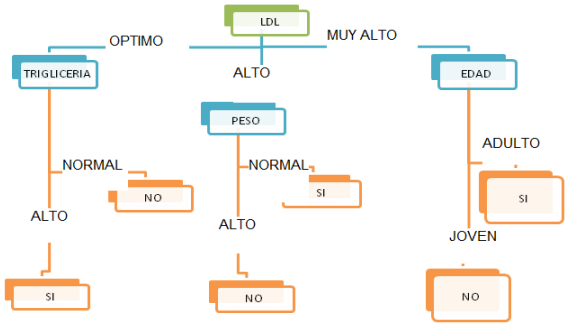

Figura 17. Árbol de Decisión1

La regla para toma de decisiones de riesgo cardio-vascular seria:
(LDL =Normal ∧ Edad=adulto) ∨ (Triglicéridos= alto) ∨ (LDL =alta ∧ Triglicéridos= Limite alto)
La regla para toma de decisiones de no tener riesgo cardio vascular seria:

(LDL =Normal ∧ Edad=joven) ∨ (Triglicéridos= normal) ∨ (LDL =optimo ∧ Triglicéridos= normal)

Para el árbol presentado, solo se requiere la evaluación de un máximo de dos atributos para llegar a tomar decisiones

3. Conclusiones

-Las técnica de categorización que utiliza clasificadores basados en patrones permiten catalogar el objeto o sujeto con una calidad competitiva con un nivel superior, técnica de clúster y análisis de componentes principales, logrando así un modelo más legible y de fácil interpretación.

-La técnica de arboles de decisiones nos permite trabajar con patrones, sin embargo su eficiencia esta estrechamente relacionada con la calidad de los patrones que utilicemos.
Una de las limitaciones que se observo en la técnica que utiliza patrones es que desde sus inicios está limitado por el costo computacional de las búsquedas exhaustivas.
-Las técnicas estadísticas son fundamentales a la hora de validar hipótesis y analizar datos, por lo cual la estadística desempeña un papel muy importante, para cuantificar adecuadamente la incertidumbre
-La utilización de estas técnicas de clasificación o diagnostico nos ayudan a ratificar e identificar prematuramente los datos más significativos de las muestras tomadas a los pacientes y garantizar la oportuna intervención terapéutica preventiva para disminuir la mortalidad.
-El diagnóstico clínico sigue requiriendo la intervención del médico, ya que no existe el consenso de un patrón oro para confirmar su
diagnóstico, elementos indispensables en la práctica médica general, la investigación científica, los ensayos clínicos y la información epidemiológica.

4. Recomendaciones

1. Crear un modelo hibrido entre arboles decisiones y algoritmos genéticos, para que las búsquedas sean más inteligentes y eficientes.
2. Incluir búsquedas inteligentes de parámetros como Tabú.

5. BIBLIOGRAFÍA

[1]. Ian H. Witten and Eibe Frank. Data Mining, Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers. Second Edition, 2005.
[2]. Britos, P., Hossian, A., García-Martinez, R. y Sierra, E. 2005. Minería de Datos Basada en Sistemas Inteligentes. Nueva Librería.
[3]. Breiman, L., Friedman, J., and Olshen, R.: Classification and Regression Trees. Wadsworth International Group, 1984.
[4] Duda, R. y Hart, P., (1973) Pattern Classification and Scene Analysis, John Wiley & Sons

Enlaces

[1].http://www.cua.uam.mx/files/cuerpoAcademicoBBC.pdf[Consulta: 10 de noviembre de 2008]
[2].http://drestmont.googlepages.com/drmsemI.pdf [Consulta: 8 de noviembre de 2008]
[3].http://www.unedbizkaia.es/archivos/practicas/ambientales/biologia/practica4parte1.pdf [Consulta: 8 de noviembre de 2008]