LA MINERÍA DE DATOS: ARBOLES DE DECISIÓN Y SU APLICACIÓN EN ESTUDIOS MÉDICOS

Carlos N. Bouza[1]* y Agustín Santiago**
*Universidad de La Habana, Cuba
** Universidad Autónoma de Guerrero, México

ABSTRACT

We present some ideas on main applications of one of the most useful data mining tools: decision (classification) trees in medicine. We show that it can be used to help researchers and managers to provide accurate ideas on the behavior of different issues as hospital and epidemic management. Some applications are presented and decision tree based discussions are developed.

KEY WORDS: Decision trees, CART, CHAID, entropy.

RESUMEN

Presentamos algunas ideas sobre importantes aplicaciones de una de las más utilizadas herramientas de la minería de datos: árboles de decisión (clasificación) en medicina. Mostramos que esta puede ser usada para ayudar a investigadores y funcionarios para obtener ideas precisas sobre el comportamiento de diferentes aspectos del manejo de hospitales y epidemias. Algunas aplicaciones se presentan y se desarrolla una discusión basada en árboles de decisión.

PALABRAS CLAVE: Arboles de decisión, CART, CHAID, entropía.

1. INTRODUCCIÓN

La Minería de Datos comenzó teniendo mucho éxito en los estudios de mercado. Junto con ella se comenzó la introducción de procesos inductivos basados en los árboles de decisión desarrollados en la Teoría de Decisión. El desarrollo de la informática dio pie a tecnologías especializadas como el aprendizaje (Machine Learning) y el reconocimiento de patrones (Pattern Recognition).

Un árbol de decisión (AD) es una representación de una función multivariada y que fue posible utilizar en la vida práctica a partir del desarrollo de las modernas computadoras. El interés por el uso práctico de los AD`s tuvo su origen las necesidades de las ciencias sociales siendo seminal el trabajo de Sonquist-Morgan (1964) el software AID (Automatic Interaction Detection). Este fue uno de los primeros métodos de ajuste de los datos basados en árboles de clasificación. Con ello los AD`s trascendieron, el solo ser una representación ilustrativa en los cursos de toma de decisiones, para convertirse en una herramienta útil y sencilla de utilizar. Estos avances fueron rematados por la obra de Breiman-Friedman-Olshen-Stone (1984)" *Classification and regression trees*". Un método práctico de inducción, para construir AD`s de forma recursiva, fue propuesto. Este ha sido conocido como CART. Quinlan (1986) desarrolló el algoritmo ID3 (*Iterative Dichotomiser 3*) este utiliza la medida de entropía de la información para crear los árboles. Esta fue mejorada y fue denominada C4.5 por su autor Quinlan (1993). Proveniente de la estadística Kass (1980) introdujo un algoritmo recursivo de clasificación no binario, llamado CHAID (Chi-square automatic interaction detection).

Estos métodos permiten superar las deficiencias del AD utilizada en al Teoría Clásica de la Decisión. En la práctica médica puede faltar algún dato de un paciente por rotura o carencia de un equipo, por la imposibilidad de hacerle el examen o por otras causas. Por otra parte esto da pie a un solo AD al usar solo una muestra de entrenamiento. Tal es el caso si la información sobre el paciente carece de alguna variable. El médico deseará tener un AD y usarle antes de tomar su decisión y no perder las posibilidades le brindan los AD`s como herramienta auxiliar. Por otra parte el usuario agradecería poder analizar varios AD`s y que al obtener nueva información el AD mejorara su eficiencia predictiva. Con estos algoritmos se superan tales deficiencias en gran medida.

[1] bouza@matcom.uh.cu

Vale señalar que Cremilleux-Robert (1997) desarrollaron un marco de referencia para usar AD`s en la medicina. Este fue criticado por Kokol et al (1998) fijando las limitaciones que esta técnica presenta al aplicarle en estudioso médicos. Zorman et al (200b) usaron AD`s para considerar en el tratamiento de fracturas usando 2637 casos. Este trabajo se centró en la búsqueda del mejor método de inducción. Consideraron el uso de redes neuronales y AD así como un enfoque basado en criterios evolutivos. Cada uno de ellos tenia alguna deficiencia.

2. MINERIA DE DATOS Y ÁRBOLES DE CLASIFICACIÓN

2.1. **La minería de datos**

La minería de datos (MD) requiere de operaciones que deben ser analizadas por un estadístico, o quien conozca no solo los conceptos sino también sepa interpretar los datos cuando existen cambios. Este es uno de los problemas que plantea la dialéctica del par estadística exploratoria-estadística inferencial. La primera hace análisis de datos, para lo cual la minería de datos es una herramienta. La segunda es una herramienta para sacar conclusiones al hacer minería con los datos generados por un fenómeno. Por ello la MD requiere de una iteración entre las áreas de Computación, Estadística y del área donde se aplica a través de los expertos, que en el caso de la medicina son los médicos.

La MD comenzó a ser utilizada con frecuencia por sus facilidades, transparencia y bajo costo para guardar informaciones. Esta es usualmente clasificada solamente al considerar su rol en el estudio de grandes bases de datos, y la consideran parte de la llamada K*nowledge Discovery in Databases* (búsqueda de conocimientos en bases de datos). Sin embargo es más correcto definir la MD como un proceso, que extrae informaciones esenciales de grandes bases de datos sin requerir de ningún conocimiento previo, para tomar decisiones y aprender sobre el fenómeno.

La MD debe ser considerada como un área multidisciplinaria que relaciona procedimientos, métodos, modelos y técnicas provenientes de la estadística, reconocimiento de patrones, del aprendizaje de máquinas, etc. Teniendo la variopinta presente en la MD se hace la selección de las herramientas, lo que es vital. Las inferencias se hacen a partir de inducciones, yendo de un ejemplo concreto que genera una muestra o población, a modelos generales. Esto es, el objetivo es aprender a clasificar los ítems analizando casos conocidos. Es clara la importancia que tiene tal proceso en medicina para la identificación rápida de las mejores terapias para diferentes dolencias, por ejemplo.

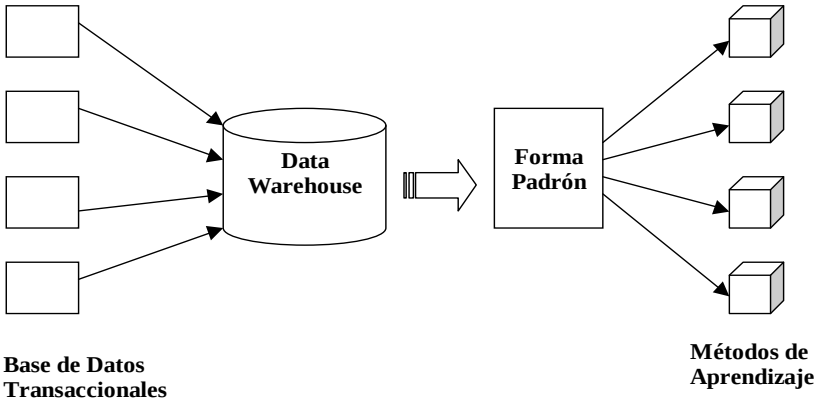

Figura.1. Esquema del proceso de selección acumulación y procesamiento de los datos

Es de gran importancia la potencialidad de la MD para trabajar con el tamaño de las bases de datos a analizar. Como hay varias posible fuentes de datos la aplicación de la MD requiere del llamado "Data Warehouse". Este es un sistema de gerenciamiento de la base de datos relacional. Se utiliza para extraer datos operacionales archivados y dar solución a las inconsistencias entre diferentes formatos de datos y les integra

los datos. Una vez procesados y acumulados en el Data Warehouse los datos no son actualizados ni alterados, simplemente cargados o accesados. Si pensamos en los expedientes de los pacientes y en la información que se puede recabar sobre este de bases de datos en varios centros de salud se ejemplifica la necesidad de usar una Data Warehouse que acopie datos de varias bases, los tipifique y limpie las inconsistencias.

2.2 Los Arboles de Decisión

Los árboles de decisión proveen de una herramienta de clasificación muy potente. Su uso en el manejo de datos la hace ganar en popularidad dadas las posibilidades que brinda y la facilidad con que son comprendidos sus resultados por cualquier usuario. El árbol en si mismo, al ser obtenidos, determinan una regla de decisión. Esta técnica permite:

- Segmentación: establecer que grupos son importantes para clasificar un cierto ítem.
- Clasificación: asignar ítems a uno de los grupos en que está particionada una población.
- Predicción: establecer reglas para hacer predicciones de ciertos eventos.
- Reducción de la dimensión de los datos: Identificar que datos son los importantes para hacer modelos de un fenómeno.
- Identificación-interrelación: identificar que variables y relaciones son importantes para ciertos grupos identificados a partir de analizar los datos.
- Recodificación: discretizar variables o establecer criterios cualitativos perdiendo la menor cantidad posible de información relevante.

En el sector salud la necesidad de hacer nuevas inversiones es constante y en especial en el sector público donde las decisiones envuelven a los políticos. En particular estas inversiones se asocian a la adquisición de nuevos instrumentos, que pueden ser costosos, material gastable o a ampliaciones-reparaciones. Por ellos las decisiones envuelven la demanda de grandes sumas de dinero de las fuentes financieras. Estas decisiones repercuten con un retardo en la calidad del servicio pero sin embargo lo hace de inmediato las finanzas y en las metas a largo plazo fijadas anteriormente. Las inversiones son decisiones de tipo estratégico y se asocia a un grado alto de incertidumbre. Todas estas decisiones de inversión son soportadas por predicciones de los expertos. En los análisis de inversiones es popular el uso del concepto de valor esperado y podemos utilizarles en la confección de árboles de decisión.

El valor esperado representa el promedio a obtener a largo plazo bajo el principio del muestreo repetido. Se asume que hay una medida de probabilidad *P(X)* que permite establecer varios escenarios cuyo resultado es caracterizado por una realización X que puede tomar k posibles valores $x_1, \ldots, x_k$. El valor esperado *E(X)* es

$$E(X) = \sum_{i=1}^{k} x_i P(X = x_i)$$

El árbol de decisión permitirá representar y analizar el resultado de la inversión. Ante la compra de un nuevo equipo la dirección de un hospital tiene como árbol de decisión. Cada decisión lleva a uno de los nodos terminales y se asocia a un costo monetario y/o de prestigio. La probabilidad de transitar cada camino es establecida por ellos. Es claro que de no comprar no se daría el servicio con probabilidad uno y que de hacerlo se dará la mejora con esa misma probabilidad. En el primer caso no hay un costo monetario pero socialmente se vera afectado el prestigio del tercio de salud.

En las evaluaciones se utiliza el razonamiento presente en la programación dinámica de "backward induction": comenzando en un nodo final se regresa al nodo inicial.

Este árbol de decisión es el utilizado en la Teoría clásica de decisión. Estos llevan a algoritmos deficientes para hacer inducciones cuando los datos son incompletos (missing) o con muchos errores (noisy data).

Por otra parte, un médico al evaluar los síntomas de un paciente detecta información a través de dar respuesta a interrogantes, y descarta las posibles enfermedades que le pueden aquejar. En su mente tiene un

árbol de decisión y llega a una conclusión, o varias posibles, al considerar cuan verosímil es el camino que sigue a partir de la respuestas a sus interrogantes. Al dar respuesta a través de evaluar un cuestionario o de análisis el médico traza un camino en un grafo llegando a un nodo final (hoja). Así el interés de un médico es utilizar la información recabada y establecer en que nodos se concentran los posibles trastornos. Similarmente lo podría hacer al analizar un sistema de salud un investigador para detectar que centros asistenciales concentran ciertas propiedades como el nivel de eficiencia, por ejemplo.

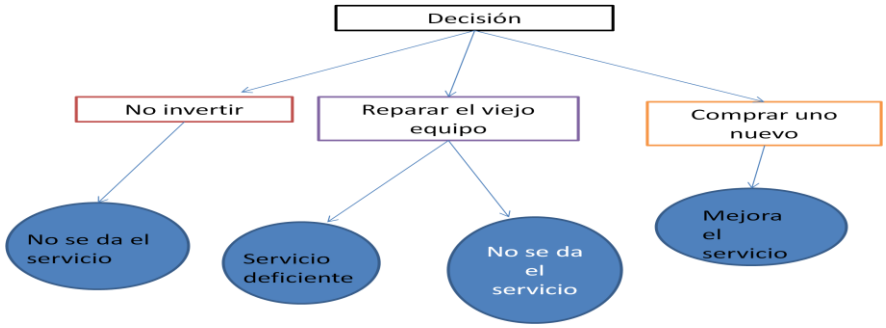

Figura 3. Un Árbol de Decisión para la inversión en un nuevo equipo ante rotura del viejo

En muchos problemas médicos es de interés diseñar el árbol descrito por los usuarios del mismo para considerar el desempeño de un sistema de salud, que puede ir desde un sistema nacional hasta el de un hospital.

Para confeccionar tales árboles se toma una muestra, se observa el camino seguido por los entrevistados y cualquier campaña o plan de desarrollo se centraría en satisfacer los intereses de calidad expresados por las entrevistas. Es claro que algunos caminos serán recorridos con mayor frecuencia permitiendo hacer un análisis estadístico de cómo los entrevistados, que no son sino clientes, evalúan el sistema y sus componentes. Por ejemplo si consideramos el uso de un hospital especializado como un cardiocentro se pueden considerar varios caminos recorridos antes de llega a este. Por ejemplo el camino medico general-cercanía-confianza-rapidez-deficiente diagnóstico puede ser mas frente que policlínico-cercanía-confianza-lentitud-correcto diagnostico para llegar a tratamiento en un cardiocentro. Esto nos llevaría a obtener quizás que la mayor aparte de los que acuden al cuerpo de guardia del cardiocentro sean pacientes que van al policlínico y viven cerca de este. Esto puede llevar a considerar la necesidad de establecer servicios permanentes de cardiología más cercanos en las municipalidades y/o a dar un entrenamiento especializado a los médicos generales.

En medicina la toma de decisiones es harto importante y sistemática. Piense en la elaboración de diagnósticos por ejemplo. En las instituciones avanzadas existen sistemas conocidos como “Decision support systems”, sistemas soporte de las decisiones, que son usados para ayudar al médico a establecer decisiones alternativas, que sean eficientes y fiables, ante situaciones complejas. Estos sistemas son en general capaces de entrenarse con las nuevas informaciones y aprender automáticamente (Automatic Learning). Para tales tareas los árboles de decisión son una herramienta conveniente. Dado su desempeño en muchas aplicaciones los expertos, en las diversas áreas donde se toman regularmente, les utilizan por ser transparente su proceder. El uso de una herramienta que nos lleve a determinar un AD es muy recomendado cuando:

- Estudiamos conceptos del tipo atributo-valor
- La función objetivo se asocia a una VR con valores discretos
- Las descripciones del ítem son disyuntivas
- El conjunto de aprendizaje T tiene errores y/o tiene valores perdidos.

Siguiendo el enfoque de Quinlan (1993) un AD formaliza un mapeo del problema determinando conexiones entre nodos de un árbol donde se exprese el proceso de estudio de la población determinando sub-árboles y hojas. Cada hoja o nodo terminal determina una clase que fija la decisión a tomar. Algunos nodos son denominados de prueba y en ellos se elabora una salida, basada en el análisis de los datos que han entrado en él de los nodos previos con los que se conecta. En sus usos prácticos el AD comienza analizando casos

conocidos. Los ítems son divididos en dos subconjuntos. Uno es usado para determinar el árbol (muestra de entrenamiento) y la otra para evaluar la efectividad del AD (muestra de prueba). Un atributo es fijado para representar la decisión de interés (variable de respuesta, VR). El problema es determinado al fijar para cada ítem un conjunto de atributos que puede ser representado como un vector $\vec{X} \in \Re^k$. Entonces podemos decir que la muestra de entrenamiento es $T=\{x_1,x_2,...,x_m\}$.

Los atributos pueden tomar valores discretos, continuos o cualitativos. En el caso discreto cada valor determina una clase. Si es continua la variable se determinarán intervalos. Las variables cualitativas son discretizadas. En cada paso el conjunto de ítems es subdividido de acuerdo a un cierto criterio en dos o más clases hasta llegar al conjunto de nodos finales. El método usual de partición se hace determinando en $\Re^k$ híper-planos ortogonales al atributo seleccionado. Así el AD divide el espacio en híper-rectángulos y cada uno identifica una decisión. Hay *m-1* posibles particionamientos de *T*. Para hacer una partición se utiliza un criterio de poda que determina cuando detener la segmentación. Si el número de las VS`s es elevado el AD será muy grande dificultando su interpretación.

En el caso de atributos continuos se pueden determinar los intervalos usando no de los siguientes criterios

- Intervalos con la misma amplitud: Se fija el número de clases y se determinan los intervalos.
- Intervalos determinados por percentiles: Se fin el numero de clases y se determinan los intervalos buscando que estos contengan el mismo numero aproximado de ítems.

Siguiendo a Quinlan (1993) podemos fijar la búsqueda de un AD como sigue:

If there are k classes denoted {C1, C2,...,Ck}, and a training set, T, then

if T contains one or more objects which all belong to a single class Cj,, then the decision tree is a leaf identifying class Cj.

- *if T contains no objects, the decision tree is a leaf determined from information other than T.*

if T contains objects that belong to a mixture of classes, then a test is chosen, based on a single attribute, that has one or more mutually exclusive outcomes {O1, O2,..., On}.

T is partitioned into subsets T1, T2,..., Tn, where Ti contains all the objects in T that have outcome O_i of the chosen test.

The same method is applied recursively to each subset of training objects.

Sea $S \subset \Omega$,

$$\Omega = poblacion\ estudiada = \cup_{1 \le i \le n} C_i$$

Al seleccionar una muestra se obtiene una señal sobre la pertenencia de un ítem a una clase. Si

$$f(j_i,S) = número\ de\ objetos\ en\ C_i\ , P_i = \frac{f(j_i,S)}{|S|}$$

Pensando en términos de la Teoría de Información esta señal en términos de bits es *-log2 (f(ji,S) / |S|)*. Considerando *S* la información esperada de la señal es

$$\inf(S) = -log_2\left(\frac{f(C_j,S)}{|S|}\right)$$

Si tenemos una muestra de entrenamiento

$$\inf(T|\Omega) = \sum_{i=1}^{n} \left(\frac{|T_i|}{|T|}\right) \inf(T_i|\Omega)$$

Tenemos entonces una serie de medidas de interés sobre la partición. Nos interesa

g (Ω) = info (T) - info$_\Omega$ (T)=ganancia en información por particionar T de acuerdo Ω.

Si particionamos *T* en n subconjuntos tenemos

$$part - info\ (\Omega)_i = \inf(T|X) = \sum_{i=1}^{n} \left(\frac{|T_i|}{|T|}\right) log_2 \left(\frac{|T_i|}{|T|}\right)$$

La partición genera información que es útil para la clasificación y una medida de la ganancia es *gr* (Ω) = *g*(Ω) / *part-info* $(\Omega)_i$. = *función de impureza*

Esta mide la mezcla de un subconjunto y particiones para disminuir la impureza. Se busca maximizarla. A la larga esta es conlleva el mismo significado que el índice de Gini, usado en economía, Breiman (1984). Esta función toma en cuenta la probabilidad de mal-clasificar una muestra adicional dado el resultado obtenido al usar la muestra de entrenamiento *T*.

Una regla de clasificación asignará a un nodo un nuevo ítem buscando minimizar la tasa de mala clasificación.

Entonces un AD es un gráfico determinado a partir de algún método que modela la toma de decisiones utilizando reglas de fácil comprensión. Los ítems agrupados a partir de los atributos (variables explicativas o de segmentación, VS) son obtenidos segmentando *T*. Al determinar un grupo, se hace un estudio, preferentemente estadístico, de la VR. La homogeneidad de este atributo es realizado para buscar una explicación del efecto de las VS`s sobre la VR. A partir de las VS`s se identifican los miembros de los grupos y se facilita el predecir el valor de la VR.

Esto no es más que un problema de mercadeo con características particulares. Así si entrevistamos a los usuarios del sistema este se enfrenta a varias preguntas. Sus repuestas establecen el movimiento de un nodo a otro del árbol y al final ha trazado un camino a una conclusión que esta representada por uno de los nodos finales. Así el entrevistado es clasificado. En el área de diagnóstico el camino lleva a la evaluación de una regla de decisión. Al analizar los nodos finales se pueden particularizar los síntomas que llevan a ciertas enfermedades y los elementos que le identifican con más frecuencia. El médico podrá, usándolas, tener un grupo de posibles diagnósticos sobre los cuales elaborara los tratamientos alternativos a la dolencia del paciente. Esto es de particular importancia en enfermedades con una gran cantidad d síntomas que son comunes a otras como el Dengue.

En ambos casos presentados el interés es determinar un árbol que sea altamente verosímil. También en ambas un cierto decisor ante condiciones diferentes tiene valoraciones que determinan la toma de decisiones que le llevan a un nodo terminal. Si el árbol diseñado a partir de tomar una muestra es verosímil le dará efectividad a las decisiones que se tomen utilizándoles.

2.3. Los árboles de decisión y las redes neuronales

La mayor parte de los problemas que abordamos haciendo estadística exploratoria tiene conexiones claras con la IA. Sin embargo comúnmente los estadísticos y los especialistas en IA no se comunican efectivamente. Los términos usados por una y otras son diferentes aunque los métodos utilizados sean similares o iguales. En la práctica se desaprovecha todo el arsenal de métodos y herramientas de la estadística por la IA en la que se crean nuevos y en general teóricamente ineficientes métodos para resolver problemas cuya solución ha sido obtenida y parece en simples libros de texto de estadística. La IA se ha centrado en hacer algoritmos eficientes y no en usar modelos inferenciales eficientes.

Para ilustrar la diferencia en la denominación de conceptos usamos la tabla presentada por Lebart (1998), Tabla 1.

Las RN son generalizaciones de clásicos modelos estadísticos que aplican un aprendizaje secuencial y aborda con éxito transformaciones de las variables originales para hacer predicciones y superar los problemas planteados a los estadísticos por los modelos no lineales que caracterizan los fenómenos complejos. Este es una caja negra donde el algoritmo usado utiliza los inputs y brinda outputs. Se espera que el sistema aprenda mejorando sus predicciones. Los métodos de la construcción de AD y la RN son complementarios. La representación mediante un AD es transparente para el médico, y cualquier usuario,

pero no es el caso con RN. Por otra parte la técnica de AD falla cuando hay muchos outliers lo que no ocurre con las RN. Además si usamos AD para aprender el aprendizaje es lento pero las RN lo hacen con rapidez. De ahí el éxito que ha tenido el combinar ambas técnicas. Las propuestas usan un RN la que entrenan y al final del aprendizaje la convierten en un AD. Si el aprendizaje es bueno el AD determinado es mejor que el AD utilizado como dato de entrada para las RN.

Estadística	Inteligencia artificial (Artificial intelligence)
Modelo	Red (network)
Observaciones, individuos, ítems	Ejemplos o patrones (features, inputs, outputs)
Variables independientes, explicativas, regresores	Entradas (inputs)
Variables dependientes, respuestas, regresando	Salidas (ouputs, targets)
Residuos	errores
parámetros	Pesos, coeficiente sinápticos (weights, synapsis)
Estimación, predicción	Entrenamiento , aprendizaje (training, learning)
Criterios de ajuste	Función de error, costo
Regresión. Discriminación	Aprendizaje supervisado
clasificación	Aprendizaje no supervisado

Tabla 1. Equivalencia de términos entre la estadística y la inteligencia artificial.

2.4. Algoritmos Evolutivos

Otra técnica usada contemporáneamente para construir un AD son los algoritmos evolutivos (EVOP). Estos son de uso común en optimización si no hay un algoritmo heurístico eficiente. Una propuesta muy eficiente es la de Cantu-Paz, 2000 y las de Podgorelec and Kokol [2001a; 2001b].

La calidad del AD es medida por

$$LFF = \sum_{i=1}^{k} w_i(1 - acc_i) + \sum_{i=1}^{N} c(w_i) + w_u + nu \quad 1$$

donde

- K= *número de clases de* decisión
- N=numero de nodos de atributos en el árbol
- acc_i = *exactitud de la* clasificación de ítems de una clase de decisión d_i,
- w_i= importancia dada (peso) a clasificar los ítems en la clase de decisión d_i,
- $c(t_i)$=costo de usar el atributo en un nodo t_i,
- nu=numero de nodos no usado
- w_u = *peso de* la presencia de nodos no usado en el árbol.

Será mejor un árbol si LFF es menor . El EVOP busca el árbol con mínimo valor de LFF**.**

3. SU ÉXITO EN ALGUNAS APLICACIONES EN LA MEDICINA

Los AD`s son usados en estudios médicos y en estudio de salud comunitaria desde hace mas de dos décadas. Veremos algunas aplicaciones realizadas usando las herramientas discutidas.

Letourneau et al (1998) tomaron una muestral de dos grupos de enfermeras y desarrollaron un AD para guiar la labor de uno de ellos. Se concluyó que este fue de gran ayuda al comparar la eficiencia del grupo ayudado por el AD y el que no le utilizó. Tsien et al (2000) utilizaron datos obtenidos en Edimburgo y establecieron como AD puede ser de ayuda en la toma rápida de decisiones al hacer las predicción en Sheffield.

Jones (2001) estudiaron el uso de AD`s para fijar las señales que sugieren los efecto secundario de un medicamento.

En el ámbito hospitalario se ha propuesto la técnica de AD para mejorar los sistemas de alarma en las unidades de cuidados intensivos, vea Tsien et al (2000). Bonner (2001) realizó un estudio similar con enfermos mentales.

.

3.1. Estudio del comportamiento del éxito total de los casos de cirugía de urgencia

Se realizó un estudio de 285 casos llegados a urgencias en hospitales de un sistema de salud. Todos ellos debían ser operados de urgencia. Se analizaron las condiciones del hospital en el momento de la intervención, las condiciones bajo las que se efectuó y el éxito total de las mismas. El éxito total se refiere a la no existencia de complicaciones post-operatorias.. Las muertes no fueron consideradas en el estudio. Las variables fueron:

Atributo	Características	del	atributo
Nivel de la urgencia (A_1)	alto	medio	bajo
Estado del salón (A_2)	bueno	regular	--
Nivel de preparación del personal en la guardia quirúrgica			
Cirujano(A_3)	óptimo	aceptable	
Personal auxiliar (A_4)	óptimo	aceptable	
Condiciones físicas del enfermo (A_5)	Bueno	regular	malo

Tabla 2 . Resultados en el estudio de los éxitos en la cirugía de urgencia

Se realizó el estudio de la entropía de la información(EI). La entropía es *I(S)=0,8631*.

Atributo	Ganancia
Nivel de la urgencia (A_1)	0,2728
Estado del salón (A_2)	0
Nivel de preparación del personal en la guardia quirúrgica	
Cirujano(A_3)	0,0150
Personal auxiliar (A_4)	0,2260
Condiciones físicas del enfermo (A_5)	0,0150

Tabla 3. Ganancia en entropía de los atributos

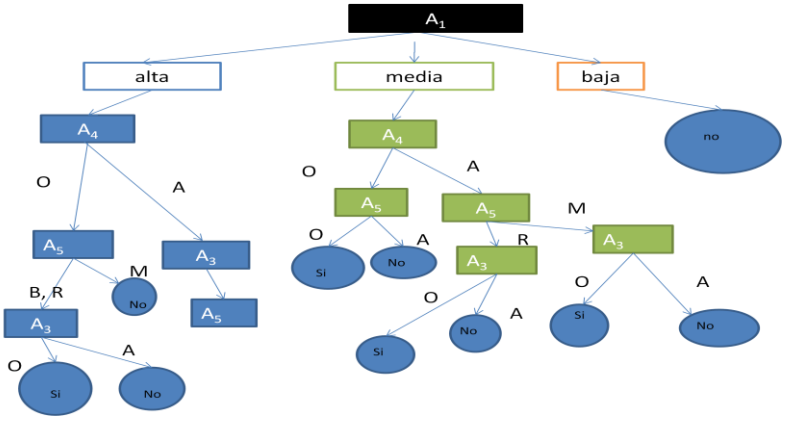

Figura 4. Árbol de Decisión del comportamiento del éxito total de los casos de cirugía de urgencia

Entonces lo más importante para el éxito de una intervención urgente es el nivel de la urgencia. Le sigue el nivel del personal auxiliar del cirujano. Esto debe ser analizado con cuidado, significando que el nivel del cirujano es importante (ser bueno o regular según una evaluación previa). Esto refleja que los cirujanos no se distinguen especialmente en tales casos pero que la intervención depende mucho de la calificación del personal auxiliar (enfermeras, anestesistas etc.).

3.2. Evaluación de un ungüento placebo en pacientes con psoriasis en menos del 10% de su cuerpo.

Se sometieron a un tratamiento con un placebo 360 personas que padecían de psoriasis. El medicamento era un ungüento que venia en dos colores A=Azul y Rojo. Se evaluaron los resultados obteniéndose los resultados en la tabla siguiente:

TABLETA	MEJORARON	NO MEJORARON	SE CONSIDERARON CURADOS		TOTAL
			Recayeron	No Recayeron	
ROJA	53	127	0	0	180
AZUL	0	62	54	64	180
TOTAL	53	189	54	64	360

Tabla 4. Resultados del tratamiento con un ungüento placebo de la psoriasis.

El árbol de clasificación aparece en la Figura 5, en el se nota que los que seleccionaron el ungüento rojo consideraron obtenían mejoría en menos de 4,5 semanas (29,44%) y los del azul en se consterno sin mejoría tras 8,1(34,44%). Los que usaron el ungüento azul se consideraron primeramente curados, no mejorados y fueron dados de alta el 35,55%

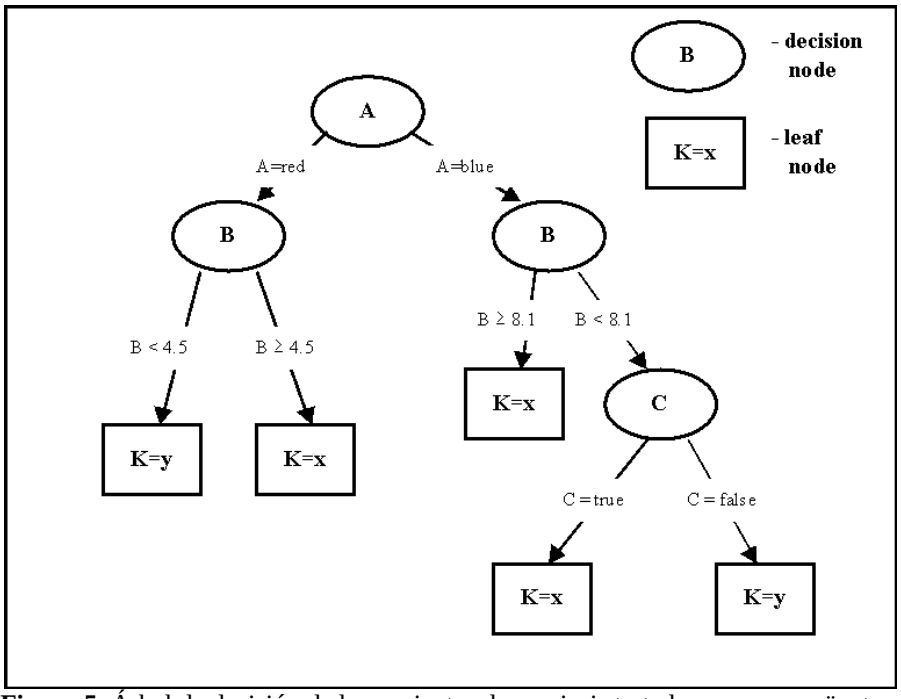

Figura 5. Árbol de decisión de los pacientes de psoriasis tratados con un ungüento placebo.

3.3. Estudio de las condiciones higiénicas y de eficiencia en hospitales quirúrgicos

Se analizaron 6 hospitales clasificados mal valorados por sus pacientes. El interés era clasificarles en no-confiables y confiables. Los resultados de las auditorias llevaron a la Tabla 5.

HOSPITAL	CATEGORÍA	ASISTENCIA PRIMARIA	CIRUGÍA	HIGIENE	CALIFICACIÓN DEL PERSONAL
1	NC	M	B	R	R
2	C	M	B	M	B
3	NC	M	R	B	R
4	NC	B	M	M	R
5	C	M	B	M	R
6	C	MM	MM	M	R

Tabla 5. Resultados de la auditoria a 6 hospitales quirúrgicos

Calculando la entropía de la información(EI) tenemos que la entropía de la asistencia primaria es $I(A_1)$=0,66 pues la EI para los valores son:

$$I_{10} = -\left(\frac{1}{1}\right) log_2(1) - \left(\frac{0}{1}\right) log_2(0) = 0$$

$$I_{11} = -2\left(\frac{2}{4}\right) log_2\left(\frac{2}{4}\right) = 1$$

$$I_{12} = -\left(\frac{1}{1}\right) log_2(1) - \left(\frac{0}{1}\right) log_2(1) = 0$$

Para Cirugía la EI es $I(A_2) = 0,79$ por que:

$$I_{20} = -2\left(\frac{1}{2}\right) log_2\left(\frac{1}{2}\right) = 1$$

$$I_{21} = -\left(\frac{0}{1}\right) log_2(0) - \left(\frac{1}{1}\right) log_2(1) = 0$$

$$I_{22} = -\left(\frac{2}{3}\right) log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right) log_2\left(\frac{1}{3}\right) = 0{,}9183$$

Para Higiene se tiene $I(A3) = 0,54$ ya que

$$I_{30} = -\left(\frac{1}{1}\right) log_2(1) - \left(\frac{0}{1}\right) log_2(0) = 0$$

$$I_{31} = -\left(\frac{1}{4}\right) log_2\left(\frac{1}{4}\right) - \left(\frac{3}{4}\right) log_2\left(\frac{3}{4}\right) = 0{,}88113$$

$$I_{32} = -\left(\frac{1}{1}\right) log_2(1) - \left(\frac{0}{1}\right) log_2(0) = 0$$

Respecto a Calificación del personal $I(A4) = 0,81$ dado que

$$I_{40} = -\left(\frac{0}{1}\right) log_2(0) - \left(\frac{1}{1}\right) log_2(1) = 0$$

$$I_{41} = -\left(\frac{2}{5}\right) log_2\left(\frac{2}{5}\right) - \left(\frac{3}{5}\right) log_2\left(\frac{3}{5}\right) = 0$$

Analizando la higiene tenemos

Higiene buena	Asistencia primaria	Cirugía	Higiene	Calificación del personal	Categoría
Hospital 3	M	R	B	R	NC
Higiene regular					
Hospital 1	M	B	R	R	NC
Higiene mala					
Hospital 2	M	B	M	B	C
Hospital 4	B	M	M	R	NC
Hospital 5	M	B	M	R	C
Hospital 6	MM	MM	M	R	C

Tabla 6. Resultados del análisis de la higiene en la auditoría a 6 hospitales quirúrgicos

La Figura 6 es el AD que genera.

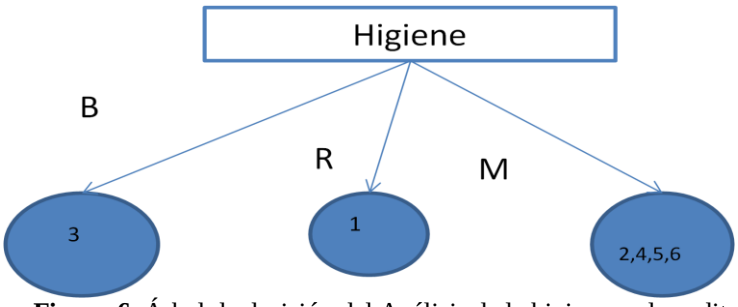

Figura 6. Árbol de decisión del Análisis de la higiene en la auditoría a 6 hospitales quirúrgicos

El análisis de la cirugía nos lleva a la Tabla 7.

Cirugía buena	Asistencia primaria	Cirugía	Higiene	Calificación del personal	Categoría
Hospital 2	M	B	M	B	C
Hospital 5	M	B	M	R	C
Cirugía mala					
Hospital 4	B	M	M	R	NC
Cirugía Muy mala					
Hospital 6	MM	MM	M	R	C

Tabla 7. Resultados del análisis al incluir cirugía junto con la higiene en la auditoría a 6 hospitales quirúrgicos

El correspondiente AD aparece en la figura próxima

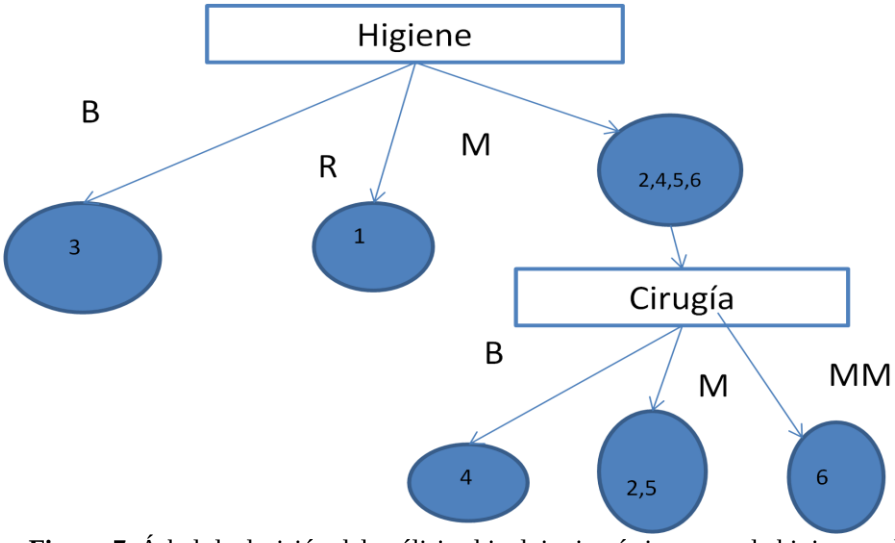

Figura 7. Árbol de decisión del análisis al incluir cirugía junto con la higiene en la auditoría a 6 hospitales quirúrgicos

Vea en la Tabla 8 continuación la inclusión en el análisis de la califica c ion del personal.

Obviamente el peor hospital está entre el 6 y el 5; el mejor el 2 seguido del 1

Calificación del personal Buena	Asistencia primaria	Cirugía	Higiene	Calificación del personal	Categoría
Hospital 2	M	B	M	B	C
Calificación del personal Regular					
Hospital 5	M	B	M	R	C

Tabla 8. Resultados del análisis al incluir calificación del personal además de cirugía e higiene en la auditoría a 6 hospitales quirúrgicos

Entonces se tiene el AD completo del problema, ver Figura 8.

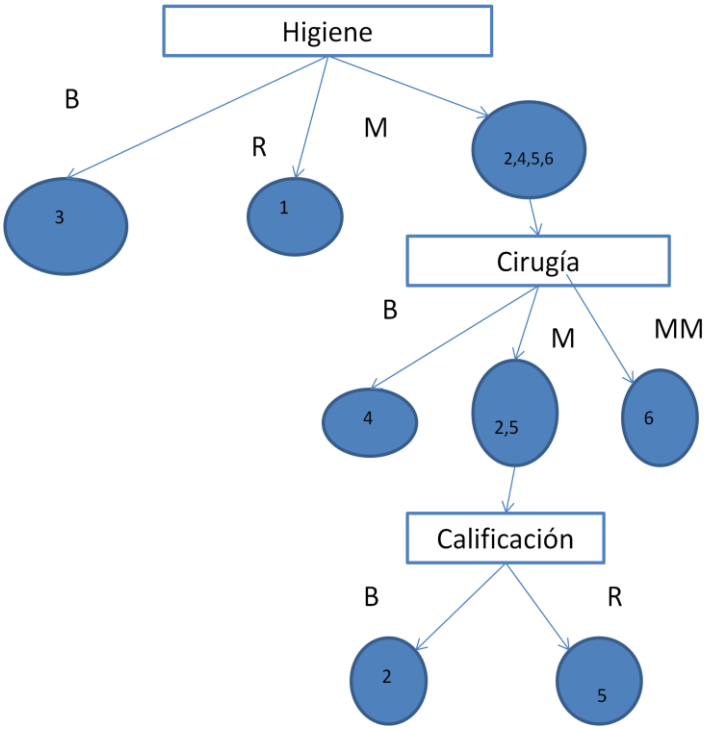

Figura 8. Árbol de decisión del análisis al incluir calificación del personal además de cirugía e higiene en la auditoría a 6 hospitales quirúrgicos.

3.4. Un estudio de condiciones ambientales en viviendas: una campaña contra el agente transmisor del Dengue

Se hizo un estudio en 22.422 viviendas para establecer si estas tenían buenas condiciones (no eran propensas a tener focos de mosquitos) o malas (tenían condiciones para su desarrollo o tenían focos de mosquitos). Se consideró la propiedad sobre la vivienda como la VR y utilizaron como VS`s:

- Nivel de escolaridad mas alto de uno de los miembros de la familia
- Lugar de ubicación
- C=Casa en una ciudad
- D=Departamento en un edificio multifamiliar
- M=Habitación en un Condominio
- S= Vivienda en Área semiurbana
- U=Vivienda en urbana
- V=Vivienda en un villorrio.

Se obtuvo como árbol de decisión el que aparece en la próxima figura.

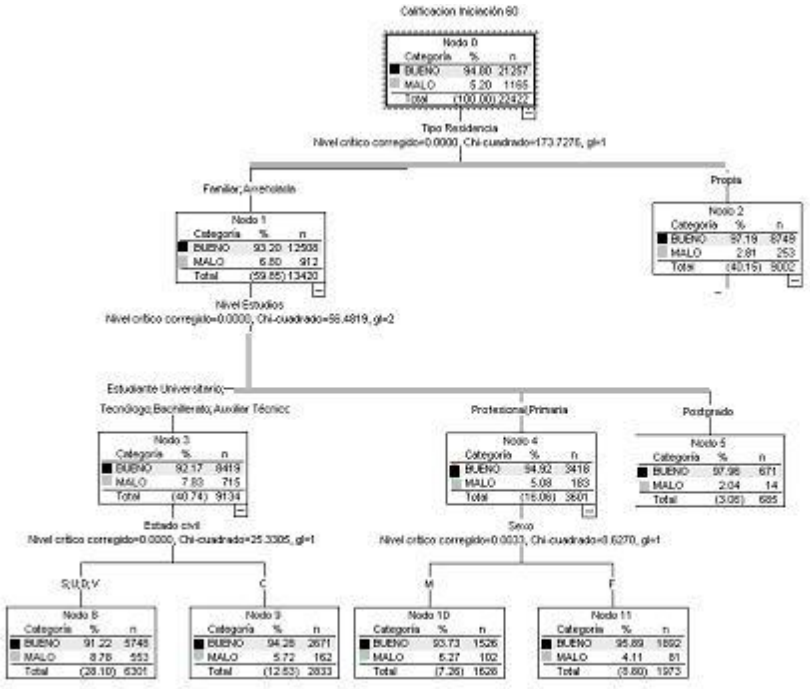

Figura 9. AD de viviendas y su estado ambiental

Este AD nos lleva a detectar que:

- Los catalogados como estudiantes de nivel medio se diferencian las viviendas tipo n S-U-D-V (que no se diferencian entre si) con las del tipo C:
- Los porcientos de catalogadas de Buenos en los dos grupos no pueden aceptarse como similares.
- Para los que tienen primaria concluida ocurre lo mismo con M y F.
- Aquellos que tiene posgrado tienen el mismo comportamiento en todo tipo de ubicación. Los propietarios se diferencia de los que no lo son sin distinción de nivel de escolaridad o ubicación de su vivienda.

Estos resultados permiten enfocar la campaña de eliminación de focos y en hacer una propaganda dirigida a no universitarios sin propiedad.

REFERENCIAS

[1] AUSTIN P. C. (2008): R & S-PLUS produced different classification trees for predicting patient mortality. **J Clin Epidemiol.** 61,1222-12226.

[2] BABIC, S.H., KOKOL, P. & STIGLIC, M.M. (2000): Fuzzy decision trees in the support of Breastfeeding. **Proceedings of the 13th IEEE Symposium on Computer-Based Medical Systems CBMS'2000**, 7-11.

[3] BANERJEE, A.(1994): Initializing neural networks using decision trees. **Proceedings of the International Workshop on Computational Learning Natural learning Systems**, 3-15.

[4] BONNER, G. (2001): Decision making for health care professionals: use of decision trees within the community mental health setting. **Journal of Advanced Nursing**, 35, 349-356.

[5] BREIMAN, L., FRIEDMAN, J.H., OLSEN, R.A. & STONE, C.J. (1984): **Classification and regression trees.** Wadsworth, N. York.

[6] CANTU-PAZ, E. & KAMATH, C.(2000): Using evolutionary algorithms to induce oblique decision trees. **Proceedings of the Genetic and Evolutionary Computation Conference GECCO-2000**, 1053-1060.

[6] CRAVEN, M.W. & SHAVLIK, J.W. (1996): Extracting tree-structured representations of trained networks. **Advances in Neural Information Processing Systems**, 8, MIT Press., Massachusetts.

[7] CRAWFORD, S. (1989): Extensions to the CART algorithm. **International Journal of Man-Machine Studies,** 31, 197-217.

[8] CREMILLEUX, B. & ROBERT, C. (1997): A theoretical framework for decision trees in uncertain domains: Application to medical data sets. **Lecture Notes In Artificial Intelligence,** 1211, 145-156.

[9] DANTCHEV, N. (1996): Therapeutic decision frees in psychiatry. **Encephale-Revue De Psychiatrie Clinique Biologique Et Therapeutique,** 22, 205-214.

[10] JONES, J.K. (2001): The role of data mining technology in the identification of signals of possible adverse drug reactions: Value and limitations. **Current Therapeutic Research-Clinical And Experimental,** 62, 664-672.

[11] KOKOL, P., ZORMAN, M., STIGLIC, M.M. & MALCIC, I. (1998): The limitations of decision trees and automatic learning in real world medical decision making. **Proceedings of the 9th World Congress on Medical Informatics MEDINFO'98,** 52, 529-533.

[12] LEBART, L. (1998): Correspondence analysis, discrimination and neural networks. En **Data Science, Classification and related methods** (Ed. C. Hayashi et al.) Springer, Berlin.

[13] LETOURNEAU, S. & JENSEN, L. (1998): Impact of a decision tree on chronic wound care. **J Wound Ostomy Continence Nurs,** 25, 240-247.

[14] PODGORELEC, V. & KOKOL, P. (2001): Towards more optimal medical diagnosing with evolutionary algorithms. **Journal of Medical Systems**, 25, 195-219.

[15] PODGORELEC, V. & KOKOL, P. (2001): Evolutionary decision forests – decision making with multiple evolutionary constructed decision trees. **Problems in Applied Mathematics and Computational Intelligence,** 97-103.

[16] QUINLAN, J.R. (1986): Induction of decision trees. **Machine Learning**, 1, 81-106.

[17] QUINLAN, J.R. (1987): Simplifying decision trees. **International Journal of Man-machine Studies**, 27, 221-234.

[18] QUINLAN, J.R. (1993): **C4.5: Programs for Machine Learning**. Morgan Kaufmann, San Francisco.

[19] SAUTER, V. L. (2010): **Decision Support Systems for Business Intelligence**, 2nd Edition. Wiley, N. York.

[20] SIERRA, B. (2006): **Aprendizaje Automático**. Ed. Pearson- Prentice Hall, N. York.

[21] SIMS, C.J., MEYN, L., CARUANA, R., RAO, R.B., MITCHELL, T. & KROHN, M. (2000): Predicting cesarean delivery with decision tree models. **American Journal of Obstetrics and Gynecology**, 183, 1198-1206.

1. [22] STATSOFT (2012): **STATISTICA Data Miner** . ww.statsoft.cl/conteudo.php?consultado enero, 2012.

[23] TSIEN, C.L., FRASER, H.S.F., LONG, W.J. & KENNEDY, R.L. (1998): Using classification tree and logistic regression methods to diagnose myocardial infarction. **Proceedings of the 9th World Congress on Medical Informatics MEDINFO'98, 52**, 493-497.

[24] TSIEN, C.L., KOHANE, I.S. & MCINTOSH, N. (2000): Multiple signal integration by decision tree induction to detect artifacts in the neonatal intensive care unit. **Artificial Intelligence In Medicine,** 19, 189-202.

[25] ZORMAN, M., HLEB S. & SPROGAR, M. (1999): Advanced tool for building decision trees MtDecit 2.0. **Proceedings of the International Conference on Artificial Intelligence ICAI-99**.

[26] ZORMAN, M., KOKOL, P. & PODGORELEC, V. (2000a): Medical decision making supported by hybrid decision trees. **Proceedings of the ICSC Symposia on Intelligent Systems & Applications ISA'2000, ICSC**, 56-68.

[27] ZORMAN, M., PODGORELEC, V., KOKOL, P., PETERSON, M. & LANE, J. (2000): Decision tree's induction strategies evaluated on a hard real world problem. **Proceedings of the 13th IEEE Symposium on Computer-Based Medical Systems CBMS'2000**, 19-24.