FACES
Facultad de Ciencias Económicas y Sociales

Año 12 Nº 26 mayo-agosto 2006

Facultad de Ciencias Económicas y Sociales
Universidad Nacional de Mar del Plata

CENTRO DE DOCUMENTACIÓN
Instituto de Investigaciones
Facultad de Ciencias Económicas y Sociales
Universidad Nacional de Mar del Plata
cendocu@mdp.edu.ar
http://eco.mdp.edu.ar/cendocu/

ECONOMÍA
FACES, año 12 , N° 26, Mayo / Agosto de 2006

Aspirantes a una institución de educación superior. Caracterización mediante Árboles

Higher education institution´s candidates. Characterization through trees.

Paulino E. Mallo[1]
María A. Artola[2]
Alicia Zanfrillo[3]
Mariano Morettini[3]
Marcelo J. Galante[3]
Mariano E. Pascual[3]
Adrián R. Busetto[3]

RESUMEN / SUMMARY

La toma de decisiones en las Instituciones de Educación Superior (IES) requiere de información que soporte la generación de políticas y la planeación de estrategias. En la gestión académica, específicamente para el área de ingreso, el aporte de las técnicas de minería de datos a las herramientas estadísticas convencionales resulta enriquecedor pues amplía el análisis y proporciona una visión más comprensiva para las actividades del sector.

Con el propósito de suministrar información estratégica para un área relevante del quehacer institucional como es el ingreso a la universidad, foco de atención en los últimos tiempos, se implementa una técnica de minería de datos, un modelo de agrupamiento con la detección de características significativas, a través del uso de árboles.

La aplicación del modelo se realiza sobre los datos de los aspirantes al Ingreso 2006 de la Facultad de Ciencias Económicas y Sociales (FCEyS) de la

[1] Director del Grupo de Investigación Modelos de Decisión Bajo Riesgo e Incertidumbre (GIMoDeBRI)
[2] Codirectora del GIMoDeBRI
[3] Docente Investigador integrante del GIMoDeBRI
Dirección electrónica: paulinomallo@speedy.com.ar

Universidad Nacional de Mar del Plata (UNMDP), inscriptos para las carreras de Contador Público (CP), Licenciatura en Administración (LA), Licenciatura en Economía (LE) y Licenciatura en Turismo (LT), correspondientes al plan de estudios 2005.

Decision making in higher education institutions (HEI) requires information that supports policies generation and strategies planing . In academic management, specially in ´Admission department´, the contribution of data mining techniques to the conventional statistics tools becomes fulfilling as it enlarges the analysis and gives a more comprehensive vision to that sector activities.

With the aim of giving strategic information to a relevant department of the institution, as the ´Admission´ one, focus of attraction in the last years, a data mining technique is implemented, a joint model with detection of significant characteristics through the use of trees.

The application of the model is based on the data of candidates in 2006 of Facultad de Ciencias Económicas y Sociales -FCEyS- of the Universidad Nacional de Mar del Plata -UNMDP- to the career of Accountancy, Business Administration, Economics and Tourism, plan 2005.

PALABRAS CLAVE / *KEY WORDS*:

Árboles - ingreso universitario - educación superior -minería

Trees - university admisión - higher education - mining

INTRODUCCIÓN

El ingreso constituye hoy en día un tema importante en la agenda universitaria. Ya sea desde las áreas de ingreso de las unidades académicas, de la unidad central de la Universidad o desde la Secretaría de Políticas Universitarias, se realizan acciones concretas para la articulación con el nivel de Educación Superior relacionadas con el financiamiento de proyectos de articulación.

Numerosos estudios se han llevado a cabo a fin de establecer las causas del fracaso de los aspirantes en las pruebas de admisión. Así, la deficiencia en el nivel secundario o polimodal y el nivel socio-económico y educativo de sus padres son factores determinantes en su paso a la universidad

El análisis de las causas de la lentificación, el atraso en el avance de la

carrera y el abandono presenta dos tipos de factores: exógenos y endógenos. En el estudio del ingreso universitario, se mencionan los factores exógenos a la institución universitaria que resultan más relevantes: género, edad, residencia, nivel socio-económico, nivel educativo de los padres, condición de actividad económica del estudiante, formación académica previa y las aspiraciones y motivaciones individuales (García de Fanelli, 2005: 72-75).

La evaluación del desempeño de los aspirantes de la IES se instrumenta, por lo general, utilizando técnicas descriptivas univariadas o bivariadas. La distribución de aspirantes en función de sus calificaciones en las pruebas de admisión por carrera, por grupo etáreo, por sexo, por tipo de establecimiento y por titulación secundaria o polimodal son las presentaciones tradicionales para determinar la distribución del rendimiento según las variables clasificatorias disponibles.

La búsqueda de correlación o asociación entre las variables del análisis constituye la práctica usual en este tipo de problemática. En la práctica, dichas variables pueden no presentar la correspondencia deseada; es en este tipo de situaciones donde las técnicas de minería de datos resultan de utilidad para la obtención de información a partir de un conjunto de variables independientes. (Luan, 2001: 6). El empleo de otras técnicas en el análisis del proceso de admisión a los estudios de nivel superior enriquece entonces la información obtenida anteriormente, con la adición de la selección de las variables más determinantes según la configuración presentada y su importancia o jerarquía en el proceso.

Con la incorporación de la minería de datos al conjunto de herramientas de análisis para las problemáticas de la Educación Superior, se intenta responder a interrogantes del tipo: ¿por qué fracasan los alumnos en su examen de admisión? O bien: ¿cuál es el perfil del alumno que ingresa efectivamente a la unidad académica? (Luan, 2001: 3).

MÉTODO DE ANÁLISIS: MINERÍA DE DATOS

La Minería de datos consiste en una búsqueda de relaciones y patrones de comportamiento sin que existan hipótesis a priori sobre tales aspectos. Estas herramientas, de múltiples orígenes, permiten descubrir modelos o tendencias interesantes en importantes volúmenes de datos. Se parte de un análisis denominado "hipótesis cero" a fin de encontrar patrones de

comportamiento, relaciones ocultas entre los datos a analizar.

Árboles de decisión

Un árbol de decisión se representa por medio de un conjunto de condiciones lógicas del tipo "si-entonces" extraídas de forma inductiva de los datos de análisis. Las reglas se encuentran organizadas jerárquicamente de modo tal que el seguimiento de la decisión final se cumple recorriendo el árbol desde su raíz, nodo inicial, hasta cada una de las hojas, nodos finales (Orallo *et al.,* 2004).

Los algoritmos frecuentemente utilizados para la generación de reglas basadas en árboles de decisión son el ID3 (Quinlan, *1979: 1620-1628*), CHAID (Kass, *1980: 29*), el C&RT (Breiman *et al., 1984*) y el C4.5 (Quinlan, *1993*), entre otros.

El algoritmo CHAID *Chi-squared Automatic Interaction Detector* realiza particiones no lineales óptimas para cada variable explicativa o predictiva a partir de la elaboración de tablas de contingencia basadas en el cálculo del estadístico X^2, test que determina las diferencias entre los valores observados y los valores esperados en los perfiles marginales de las categorías de las variables. Los datos son divididos en función de la variable que realiza la mejor clasificación o agrupamiento, esto es aquella con mayor valor de X^2. Cada nuevo grupo, nodo, obtenido es analizado para generar nuevas divisiones hasta el cumplimiento de alguna de las reglas de parada. La misma variable puede ser empleada en distintos niveles del árbol.

Una ventaja que ofrece el CHAID respecto de otros algoritmos es la construcción de árboles no binarios, es decir que puede presentar más de dos nodos por rama del árbol, con una o más categorías de la variable. Otra ventaja es la clasificación o agrupamiento en nodos mutuamente excluyentes, de modo tal que el árbol define una única respuesta a partir de las probabilidades de pertenencia al nodo.

Validación del árbol de decisión

A fin de generalizar los resultados extraídos del árbol de decisión en forma de reglas, se realiza una evaluación del modelo inductivo obtenido. Dicha evaluación radica en la determinación del número de casos clasificados o agrupados correctamente.

Una de las formas más difundidas es la división de los datos de análisis en dos grupos por métodos de selección aleatoria: un subgrupo se utiliza para el entrenamiento, obtención del modelo en forma de reglas, y el otro subgrupo es el de testeo, donde se comprueba el modelo generado. Si la capacidad predictiva del modelo en el grupo de entrenamiento es similar al obtenido en el grupo de testeo, se debería poder formular una generalización del modelo con el propósito de aplicarlo a los próximos procesos decisorios, referidos a la problemática de análisis a través de datos similares a los presentados.

Explicación de la técnica utilizada

A partir de los datos existentes, correspondientes al ingreso de los aspirantes a la institución, se buscan patrones de agrupamiento en la condición obtenida frente a las pruebas de admisión a los efectos de determinar las variables más contributivas del análisis.

Los datos utilizados en la matriz de datos 1374 casos válidos para la construcción del modelo corresponden a los aspirantes inscriptos para el Ingreso 2006 a la FCEyS de la UNMDP, en las diferentes carreras (CP, LA, LE y LT), en el plan de estudios 2005.

Se seleccionaron las variables consideradas como las más significativas y con la confiabilidad requerida para el estudio. Al conjunto de variables seleccionadas se le aplicaron los análisis estadísticos tradicionales de asociación y correlación. Sobre este conjunto de variables se emplea la técnica de árboles de decisión y una vez validado el modelo por el entrenamiento y testeo, se determina en un proceso iterativo las mejores variables de agrupamiento junto con el menor porcentaje de error.

Con el propósito de caracterizar a los ingresantes a partir de los datos disponibles del proceso de admisión, se utiliza un modelo de árbol de decisión que proporciona la jerarquía de variables relevantes en el análisis y su incidencia en el mismo, en un estudio de tipo exploratorio o descriptivo.

La utilización de árboles de agrupamiento en el examen de los datos a través del algoritmo CHAID se efectúa mediante el software estadístico SPSS, versión de evaluación 13.0.

Variables utilizadas

Las variables seleccionadas para el modelo son:

Carrera: CP, LA, LE y LT.

Sexo: Masculino o Femenino.

Tipo secundario/polimodal: tipo de institución donde el aspirante cursó sus estudios secundarios o de polimodal. Valores: nacional, provincial, municipal, privado no religioso, privado religioso, instituto militar, seminario, extranjero y otros.

Años egreso secundario/polimodal: cantidad de años transcurridos desde la fecha de egreso de sus estudios secundarios o de polimodal. Valores: 0 a 30 años.

Título secundario: título otorgado por la entidad en que el aspirante cursó sus estudios secundarios o de polimodal. Valores: Bachiller, Perito Mercantil, Técnico, Maestro y otros.

Condición: estado del aspirante al aprobar, desaprobar o estar exceptuado de rendir los exámenes previstos para su admisión en la institución. Valores: ingresante, no ingresante y exceptuado.

Las variables que corresponden a país, provincia, partido y ciudad de procedencia no se utilizan en el análisis debido a las inconsistencias detectadas en los valores de las distintas categorías de las mismas.

RESULTADOS OBTENIDOS MEDIANTE UN ANÁLISIS DESCRIPTIVO

Análisis estadístico

Como se desprende de los datos que nos muestra la Tabla 1, el porcentaje de ingresantes es del 25,1%, el 7,3% corresponde a los exceptuados del examen de admisión y el 67,6% a los no ingresantes. La excepción al examen de admisión se efectúa cuando el aspirante acredita su titulación de nivel secundario o polimodal del Colegio Nacional Arturo Illia, dependiente de la UNMDP, es uno de los tres mejores promedios correspondientes a establecimientos secundarios o polimodales públicos nacionales o bien cuando ha aprobado una asignatura de otra carrera de una institución universitaria.

Tabla 1: CONDICIÓN

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Ingresante	345	25,1	25,1	25,1
	No ingresante	929	67,6	67,6	92,7
	Exceptuado	100	7,3	7,3	100,0
	Total	1374	100,0	100,0	

Modelo de árbol de agrupamiento

Con el propósito de obtener las variables que resultan más determinantes en la caracterización del perfil de los aspirantes, se propone la utilización de un modelo de árbol de agrupamiento que selecciona las variables más productivas para el análisis y muestra la proporción de cada categoría de la variable seleccionada como objetivo.

En la Tabla 2 se presentan las especificaciones y los resultados del modelo propuesto. En las especificaciones se selecciona la variable dependiente, Condición y las variables independientes: Carrera, Sexo, Años egreso, Tipo secundario/polimodal y Título secundario. Para la determinación óptima de las particiones se optó por el algoritmo CHAID, dada su aplicación en la selección de atributos significativos. Se recortaron 50 casos como mínimo para el establecimiento de nodos secundarios y 100 para los nodos principales. El método de validación utilizado es el de entrenamiento y testeo sobre la totalidad de los casos validados, *SPLITSAMPLE*.

Los resultados que se exhiben en la Tabla 2 son las variables más significativas del análisis y su orden de importancia: Años egreso secundario, Tipo secundario/polimodal y Carrera.

Tabla 2: MODEL SUMMARY

Especificaciones	Método de crecimiento	CHAID
	Variable dependiente	Condición
	Variables independientes	Carrera, Sexo, Años egreso, Tipo Sec./Pol., Título Sec./Pol.
	Validación	SPLITSAMPLE
	Máxima profundidad de árbol	3
	Casos mínimos en nodo principal	100
	Casos mínimos en nodo secundario	50
Resultados	Variables independientes incluidas	Años egreso, Tipo Sec./Pol., Carrera
	Número de nodos	8
	Número de nodos terminales	5
	Profundidad	3

El modelo expuesto en la Figura 1 presenta las variables y categorías de las mismas, que resultaron más discriminantes en el análisis de la caracterización de los aspirantes. Se visualiza el árbol con una tabla de frecuencias en cada nodo, con el número de casos, valor relativo y absoluto para cada categoría de la variable dependiente. La categoría dentro de la tabla de frecuencias, que aparece resaltada, es la de mayor valor en ese nodo.

El diagrama del árbol de agrupamiento ofrece en el nodo raíz la variable dependiente Condición, con el 70,1% de No ingresantes, el 23,5% de Ingresantes y el 6,4% de Exceptuados.

Se puede observar que la variable más discriminante en la construcción del árbol es Años de egreso secundario/polimodal hasta un año de la fecha de egreso, el 70,1%, entre uno y tres años inclusive, el 12,4% y para más de tres años, el 17,5%. Esta variable constituye la única considerada en los casos en que los años transcurridos sean entre uno y tres y para más de tres años. Entre estos dos nodos terminales, la diferencia radica básicamente en la proporción de exceptuados y de ingresantes pues el porcentaje de no ingresantes es similar, 78,7% y 77,6% respectivamente. En el nodo que consigna entre uno y tres años transcurridos, los exceptuados corresponden al 7,9% y los ingresantes al 13,5% mientras que en el nodo con más de tres años transcurridos, los exceptuados corresponden al 20,8% y los ingresantes al

1,6%.

La segunda variable discriminante resulta Tipo de secundario/polimodal únicamente para el nodo que corresponde hasta un año de la fecha de egreso. La apertura en nodos hijos es de dos de similares proporciones: los valores de nacional, municipal y provincial público para un nodo, 36,4% y los valores de privado religioso y privado no religioso privado para el otro, 33,7%. El nodo correspondiente a público presenta los mayores porcentajes de No ingresantes en relación con el nodo de privado, 75,8% y 56,8% respectivamente. El nodo público expone escasos valores de Exceptuados, 1,7% y 3,5% para privado. El nodo privado ofrece los mayores porcentajes de Ingresantes, 41,5% respecto de 20,8%.

La tercera variable en orden discriminante es Carrera para el nodo privado. En este nodo la apertura es también de dos nodos hijos: CP con el 16,6% y LA, LE y LT con el 17,1%. La diferencia entre estos dos nodos se halla, según la carrera elegida, en el porcentaje de ingresantes para CP el 56,3% y en el nodo con los valores de LA, LE y LT el 27%. El porcentaje de exceptuados es similar para los dos nodos terminales.

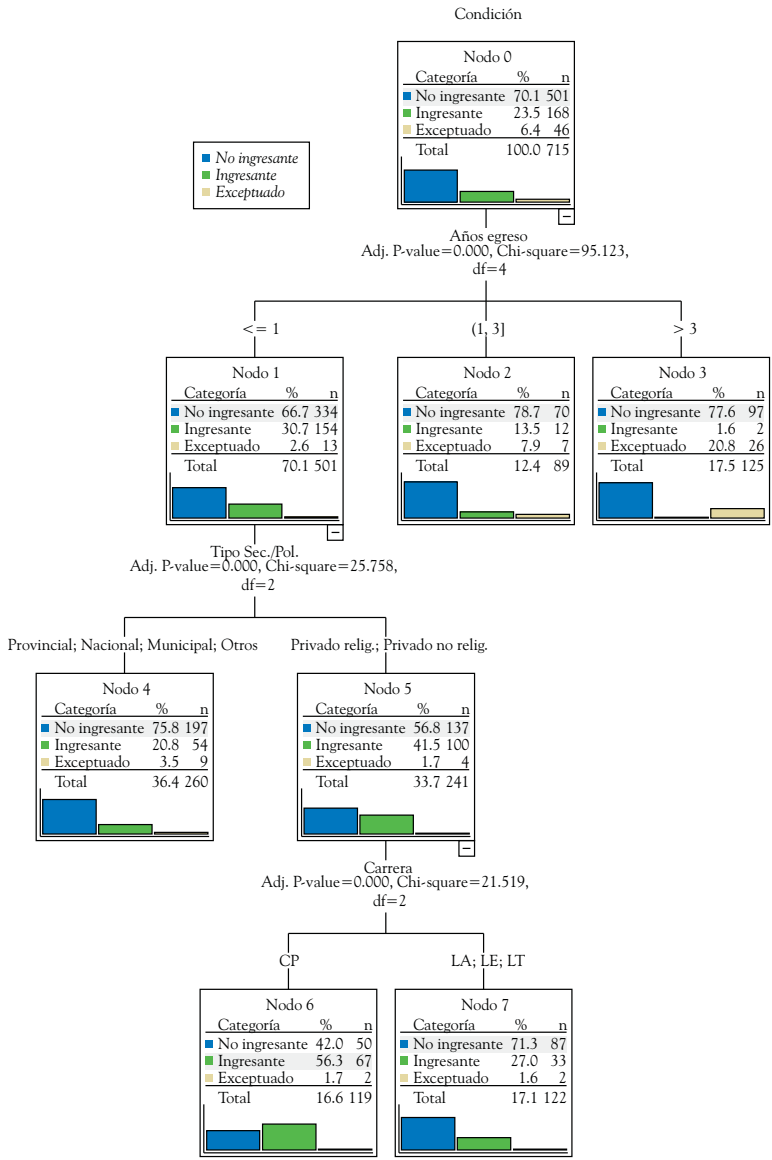

Figura 1: DIAGRAMA DE ÁRBOL DE AGRUPAMIENTO

En la Tabla 3 se exhibe el riesgo estimado, es decir, el porcentaje de error (27,6%) en que se puede incurrir en el agrupamiento de la condicionalidad de los aspirantes, de acuerdo con las características analizadas.

Tabla 3: RIESGO

Muestra	Estimación	Desviación Error
Entrenamiento	,276	,017
Contraste	,332	,018

Método de crecimiento: CHAID
Variable dependiente: Condición

En la Tabla 4 se observan los nodos terminales ordenados en forma descendente con el número de casos que contienen y el porcentaje que representan en relación con el total de casos estudiados tanto para el entrenamiento como para el testeo.

Tabla 4: GANANCIAS PARA LOS NODOS

		Nodo		Ganancia			
Muestra	Nodo	N	Porcentaje	N	Porcentaje	Respuesta	Índice
Entrenamiento	6	119	16,6%	67	39,9%	56,3%	239,6%
	7	122	17,1%	33	19,6%	27,0%	115,1%
	4	260	36,4%	54	32,1%	20,8%	88,4%
	2	89	12,4%	12	7,1%	13,5%	57,4%
	3	125	17,5%	2	1,2%	1,6%	6,8%
Contraste	6	102	15,5%	57	32,2%	55,9%	208,1%
	7	121	18,4%	44	24,9%	36,4%	135,4%
	4	236	35,8%	55	31,1%	23,3%	86,8%
	2	82	12,4%	10	5,6%	12,2%	45,4%
	3	118	17,9%	11	6,2%	9,3%	34,7%

Método de crecimiento: CHAID
Variable dependiente: Condición

En la Tabla 4 se presentan los casos concernientes a cada nodo terminal en la columna Nodo y el número de casos correspondientes a la categoría objetivo seleccionada, condición de ingresantes en la columna de Ganancia. En la columna Respuesta se consigna el porcentaje de casos en el nodo de la categoría objetivo seleccionada respecto del total de casos. El valor Índice establece cuán lejos se encuentra la categoría de la variable observada en relación con la esperada. Los valores bajos atinentes a la columna Índice, menores al 100%, representan la menor concentración de casos que corresponden a la categoría seleccionada. Los valores mayores al 100% significan que hay más casos de la categoría seleccionada que los pertenecientes a todo el porcentaje de dicha categoría.

CONCLUSIONES

Los procesos decisorios de alto nivel de las IES requieren de una etapa de diagnóstico que aporte información sustantiva para la formulación de políticas o el planeamiento estratégico. Esta información sustantiva comprende la identificación de características o comportamientos tácitos o no evidentes que enriquezcan y faciliten la comprensión del escenario donde se desarrollan dichas instituciones.

La interpretación del examen realizado permite identificar cuál es la jerarquía de las variables seleccionadas, en orden de importancia, y la contribución de cada una de ellas a la configuración de un escenario donde se estudian algunas características de los aspirantes frente a los resultados de su prueba de admisión.

En este contexto de análisis, se verifica que la cantidad de años transcurridos desde la fecha de egreso del secundario constituye la mejor característica para determinar la condición del aspirante. Para aquellos aspirantes que recién egresan hasta un año la característica más destacada es el tipo de establecimiento de sus estudios secundarios/polimodal. Para los alumnos que provienen de establecimientos privados la característica más significativa resulta la carrera elegida.

El porcentaje de ingresantes es mayor para aquellos que recién finalizan sus estudios de secundario o polimodal. Esto se mantiene para los aspirantes que proceden de instituciones privadas y en gran medida para los que eligen la carrera de Contador Público.

De la interpretación de los resultados del análisis surge que las políticas de la unidad académica deberán estar orientadas hacia los procesos de articulación con las instituciones de Enseñanza Media, tal como se encuentran en este momento, propiciando el interés de los alumnos por la continuidad de sus estudios en un futuro próximo y facilitando los entornos de aprendizaje necesarios para superar las dificultades detectadas en la aprobación de las pruebas de admisión.

La utilización de la herramienta propuesta, de amplia difusión en el ámbito empresario, posibilita un avance en el estudio de las variables más relevantes del ingreso universitario y la exclusión de aquellas que no resultan productivas. Permite determinar, además, el nivel de significatividad de las variables de examen, descartando algunas que tradicionalmente han prevalecido en años anteriores.

La facilidad de interpretación de los árboles de decisión así como la identificación de variables relevantes ofrecen esta herramienta como un instrumento de importancia para los procesos decisorios.

Bibliografía

Breiman, L.; Firedman, J. H.; Olshen, R. A.; y Stone, C. J. (1984) *Classification and regression trees. Monterrey, CA. Wadsworth and Brooks-Cole.* 358 pp.

Kass, G. V. (1980). *An exploratory technique for investigating large quantities of categorical data. Applied Statistics*, 29, Nro, 2,pp.119-127.

Fayyad, Usama, ed.; Piatetsky-Shaphiro, Gregory, ed.; Smyth, Padhraic, ed.; Uthurusamy, Ramasamy, ed. (1996). *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA: MIT Press.661 pp

García de Fanelli, Ana M. (2005). Indicadores y estrategias en relación con la graduación y el abandono universitario. In: La agenda universitaria: propuestas de políticas públicas para la Argentina. Carlos Marquís, comp. Escuela de Educación Superior. Universidad de Palermo. Pp.65-89

Luan, Jing. (2001). "Data Mining as Driven by Knowledge Management in Higher Education. Persistence Clustering And Prediction". [en línea]. In: *Keynote for SPSS Public Conference, UCSF.* <http://www.cabrillo.edu/services/pro/oir_reports/UCSFpaper.pdf>

[Consulta: 21 mar. 2006].

Orallo, José Hernández; Ramírez Quintana, María José; Ferri Ramírez, César (2004). *Introducción a la minería de datos.* Madrid: Pearson Educación.680 pp

Quinlan, J. R. (1979) Discovering rules by induction from large collection of examples. In: Expert Systems in the Microelectronic Age, ed. Michie Edinburgh: Edinburgh University Press, 168-201 pp.

Quinlan, J. R. (1993) C4.5: Programs for machine learning. San Mateo, CA: Morgan Kauffman Publishers, 302 pp.

SPSS (2001). "The SPSS C&RT Component: *A decision tree component enabling more effective classification and prediction of target variables*" [en línea]. In: *White Paper: Technical Report.* <http://www.spss.com/downloads/Papers.cfm?List=all&Name=all>. [Consulta: 21 mar. 2006], disponible bajo registro.

-------- (1999). "AnswerTree Algorithm Summary" [en línea]. In: *White Paper: Technical Report.* <http://www.spss.com/downloads/Papers.cfm?List=all&Name=all>. [Consulta: 21 mar. 2006], disponible bajo registro.