

Aplicación de la minería de datos en la extracción de perfiles de deserción estudiantil*¹

**[Application of data mining in
extracting student dropout profiles]**

**RICARDO TIMARÁN PEREIRA² - ANDRÉS CALDERÓN ROMERO³
JAVIER JIMÉNEZ TOLEDO⁴**

RECIBO: 20.02.2013 - APROBACIÓN: 16.06.2013

Resumen

En este artículo se presentan los primeros resultados del proyecto de investigación cuyo objetivo es detectar patrones de deserción estudiantil a partir de los datos socioeconómicos, académicos, disciplinares e institucionales de los estudiantes de los programas de pregrado de la Universidad de Nariño e Institución Universitaria IUCESMAG, dos instituciones de educación superior de la ciudad de Pasto (Colombia), utilizando técnicas de Minería de Datos. Los resultados obtenidos corresponden a la Universidad de Nariño. Se descubrieron perfiles socioeconómicos y académicos de los

* Modelo para citación de este artículo:

TIMARÁN PEREIRA, Ricardo; CALDERÓN ROMERO, Andrés & JIMÉNEZ TOLEDO, Javier (2013). Aplicación de la minería de datos en la extracción de perfiles de deserción estudiantil. En: Ventana Informática. No. 28 (ene.-jun.). Manizales (Colombia): Facultad de Ciencias e Ingeniería, Universidad de Manizales. p. 31-47. ISSN: 0123-9678

- 1 Artículo de investigación científica y tecnológica proveniente del proyecto *Detección de perfiles de deserción estudiantil con técnicas de minería de datos en los programas de pregrado de la Universidad de Nariño e Institución Universitaria CESMAG*, ejecutado en el periodo 10/2012-04/2013, e inscrito en los grupos de investigación GRIAS de la Universidad de Nariño y Tecnofilia de la IUCESMAG.
- 2 PhD. en Ingeniería, MSc. en Ingeniería, Especialista en Multimedia e Ingeniero de Sistemas y Computación.
Director grupo de investigación GRIAS, Profesor Asociado, Departamento de Sistemas, Facultad de Ingeniería, Universidad de Nariño, Pasto (Colombia). Correo electrónico: ritimar@udenar.edu.co
- 3 MSc. en Geoinformática, Ingeniero de Sistemas. Profesor hora cátedra, Departamento de Sistemas, Facultad de Ingeniería, Universidad de Nariño, Pasto (Colombia). Correo electrónico: aocalderon@udenar.edu.co
- 4 Especialista en Docencia Universitaria, Ingeniero de Sistema. Profesor tiempo completo, Facultad de Ingeniería, Institución Universitaria CESMAG, Pasto (Colombia). Correo electrónico: jajimenez@iucsmag.edu.co

estudiantes que desertan utilizando la técnica de clasificación basada en árboles de decisión. El conocimiento generado permitirá soportar la toma de decisiones eficaces de las directivas universitarias enfocadas a formular políticas y estrategias relacionadas con los programas de retención estudiantil que actualmente se encuentran establecidos.

Palabras clave: *Extracción de Perfiles, Deserción Estudiantil, Minería de Datos, Clasificación, Árboles de Decisión*

Abstract

The first results of the research project that aims to identify patterns of student dropout from socioeconomic, academic, disciplinary and institutional data of students from undergraduate programs at the University of Nariño and IUCESMAG University, two higher education institutions in the city of Pasto (Colombia), using data mining techniques are presented. The results correspond to the University of Nariño. Socioeconomic and academic profiles were discovered of students who drop using classification technique based on decision trees. The knowledge generated will support effective decision-making of university staff focused to develop policies and strategies related to student retention programs that are currently set.

Keywords: *Extraction of Profiles, Student Dropout, Data Mining, Classification, Decision Trees*

Introducción

Los países de América Latina enfrentan desafíos similares en la educación superior, los cuales constituyen el contexto de la deserción estudiantil: financiación, incremento de la cobertura, aseguramiento de la calidad, mejoramiento de la equidad en el acceso y permanencia, mayor articulación con la educación secundaria, diversificación de la oferta para atender distintas dimensiones, intereses y necesidades (ciencia, tecnología, sector productivo, investigación, humanidades, artes, formación integral) y mayor vinculación con el sector laboral y productivo. Según el Instituto para la Educación Superior en América Latina y el Caribe (IESALC), citado por MEN (2006a, 14), Latinoamérica presentó en el año 2003 una cobertura promedio en educación superior del 28.7% y una tasa de deserción estudiantil del 50%.

En Colombia, el sistema educativo cuenta con 277 instituciones de educación superior, de las cuales 81 son públicas y 196 privadas. De

acuerdo al Sistema Nacional de Información de la Educación Superior (SNIES), citado por MEN (2006a, 14), a 2006 la cobertura fue de 26.1%, lo cual equivale a 1.301.728 estudiantes. Uno de los principales problemas que enfrenta el sistema de educación superior colombiano concierne a los altos niveles de deserción estudiantil. Pese a que los últimos años, según MEN (2009, 13), se han caracterizado por aumentos de cobertura e ingreso de estudiantes nuevos, el número de alumnos que logra culminar sus estudios superiores no es alto, dejando entrever que una gran parte de éstos abandona sus estudios, principalmente en los primeros semestres, ya que de cada cien estudiantes que ingresan a una institución de educación superior cerca de la mitad no logra culminar su ciclo académico y obtener la graduación.

Adicionalmente, MEN (2006a, 14) plantea que a 2004, la deserción se estimó en 49%, cuyas causas fueron: limitaciones económicas y financieras, bajo rendimiento académico, desorientación vocacional y profesional y dificultades para adaptarse al ambiente universitario. Es de resaltar que, señala MEN (2006b, 1), la deserción estudiantil conlleva altos costos sociales y económicos que afectan a las familias, los estudiantes, las instituciones y el Estado.

Se entiende por deserción estudiantil, de acuerdo con UPN (2005), al hecho de que un número de estudiantes matriculados no siga la trayectoria normal del programa académico, bien sea por retirarse de ella, por repetir cursos o por retiros temporales. MEN (2009), la define como una situación a la que se enfrenta un estudiante cuando aspira y no logra concluir su proyecto educativo, considerándose como desertor a aquel individuo que siendo estudiante de una institución de educación superior no presenta actividad académica durante dos semestres académicos consecutivos, lo cual equivale a un año de inactividad académica. Esta definición es el que se aplicó en esta investigación.

La minería de datos en la educación no es un tópico nuevo y su estudio y aplicación ha sido muy relevante en los últimos años. El uso de estas técnicas permite, entre otras cosas, predecir cualquier fenómeno dentro del ámbito educativo. De esta forma, utilizando las técnicas que ofrece la minería de datos, se puede predecir, con un porcentaje muy alto de confiabilidad, la probabilidad de desertar de cualquier estudiante, coinciden Valero (2009) y Valero, Salvador & García (2010).

En el entorno internacional se han desarrollado algunos proyectos de investigación aplicando la minería de datos al descubrimiento de patrones de deserción estudiantil:

- Tal como lo señalan Pautsch (2009, 58) y Pautsch, La Red & Cutro (2010), en la Universidad Nacional de Misiones (Argentina) se realizó

una investigación sobre deserción estudiantil utilizando las técnicas de minería de datos. Su objetivo principal fue maximizar la calidad que los modelos tienen para clasificar y agrupar a los estudiantes, de acuerdo a sus características académicas, factores sociales y demográficos, que han desertado de la Carrera Analista en Sistemas de Computación de la Facultad de Ciencias Exactas, Químicas y Naturales analizando los datos de las cohortes entre los años 2000 al 2006.

- De igual manera, según La Red et al. (2010), en la Universidad Nacional del Nordeste (Argentina) se realizó un estudio cuyo objetivo principal fue aplicar técnicas de almacenes de datos y minería de datos basadas en *clustering* para la búsqueda de perfiles de los alumnos de la asignatura Sistemas Operativos de la Licenciatura en Sistemas de Información según su rendimiento académico, situación demográfica y socioeconómica, que permita conocer a priori situaciones potenciales de éxito o de fracaso académico.

- En la Universidad Nacional de la Matanza (Argentina), señalan Sposito et al., (2010), se aplicaron técnicas de minería de datos para evaluar el rendimiento académico y la deserción de los estudiantes del Departamento de Ingeniería e Investigaciones Tecnológicas sobre los datos de los alumnos del periodo 2003 al 2008. La implementación de este proceso se realizó con el software MS SQL Server para la generación de un almacén de datos, el software SPSS para realizar un preprocesamiento de los datos y el software Weka (*Waikato Environment for Knowledge Analysis*) para encontrar un clasificador del rendimiento académico y para detectar los patrones determinantes de la deserción estudiantil.

- Valero, (2009) y Valero, Salvador & García, (2010), señalan que en la Universidad Tecnológica de Izúcar de Matamoros (México) se propuso una investigación para identificar las causas que motivan la deserción de sus estudiantes desde que ingresan. Mediante la técnica de minería de datos clasificación y la herramienta Weka, encontraron relaciones entre atributos académicos que identifican y predicen la probabilidad de deserción y propusieron una herramienta para el tutor que le permite predecir la probabilidad de deserción de cualquier alumno en cualquier momento de su estancia escolar.

En el ámbito colombiano, de acuerdo con Restrepo & López (2008), en la Universidad de La Sabana se realizó un proyecto de investigación donde el objetivo era seleccionar, de una base de datos de estudiantes, los atributos que tuvieran mayor incidencia en la deserción de la Universidad en los últimos cuatro años, con la técnica de minería de datos clasificación por Rough Sets utilizando el paquete ROSE2. De igual manera, Pinzón (2011) presenta la caracterización del perfil

del estudiante desertor de la Escuela de Marketing y Publicidad de la Universidad Sergio Arboleda, utilizando la técnica de minería de datos agrupamiento con el algoritmo *K-means*. Se analizaron las variables demográficas del alumno obtenidas en el registro de última matrícula del mismo semestre de abandono y las causas que lo generaron. Como resultado final, se obtuvieron tres tipos de clúster que para el caso de la investigación, constituyeron perfiles significativos.

En este artículo se presentan los primeros resultados del proyecto de investigación cuyo objetivo es detectar patrones de deserción estudiantil a partir de los datos socioeconómicos, académicos, disciplinarios e institucionales de los estudiantes de los programas de pregrado de dos instituciones colombianas de educación superior, utilizando técnicas de Minería de Datos. Se descubrieron perfiles socioeconómicos y académicos de los estudiantes que desertan utilizando la técnica de clasificación basada en árboles de decisión con la herramienta *Weka*, una de las suites más utilizadas en el área de descubrimiento de conocimiento en los últimos años (García, s.f.). El conocimiento generado permitirá soportar la toma de decisiones eficaces de las directivas universitarias enfocadas a formular políticas y estrategias relacionadas con los programas de retención estudiantil que actualmente se encuentran establecidos.

El resto del artículo se organiza de la siguiente manera. En la sección 1, se presenta los conceptos básicos del proceso de descubrimiento de conocimiento en bases de datos. En la sección 2, se describe la metodología utilizada en la investigación. En la sección 3, se presentan los resultados de la fase de minería de datos y la discusión de resultados y finalmente, en la última sección se presenta las conclusiones y trabajos futuros.

1. Fundamento teórico

1.1 Proceso de descubrimiento de conocimiento en bases de datos

El proceso de extraer conocimiento a partir de grandes volúmenes de datos ha sido reconocido por muchos investigadores como un tópico de investigación clave en los sistemas de bases de datos, y por muchas compañías industriales como una importante área y una oportunidad para obtener mayores ganancias. Fayyad, Piatetsky-Shapiro & Smyth (1996) lo definen como «*El proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y fundamentalmente entendibles al usuario a partir de los datos*».

El Descubrir de Conocimiento en Bases de Datos (DCBD) es básicamente un proceso automático en el que se combinan descubrimiento y análisis. El proceso, de acuerdo con Agrawal & Srikant (1994), Chen, Han & Yu (1996) y Han & Kamber (2001), consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice, tarea que implica generalmente preprocesar los datos, hacer minería de datos (*data mining*) y presentar resultados. El proceso DCBD es interactivo e iterativo, involucra numerosos pasos con la intervención del usuario en la toma de muchas decisiones y se resumen en cinco etapas: Selección de datos, preprocesamiento, transformación de datos, minería de datos (*data mining*) e interpretación. En la figura 1 se muestran estas etapas.

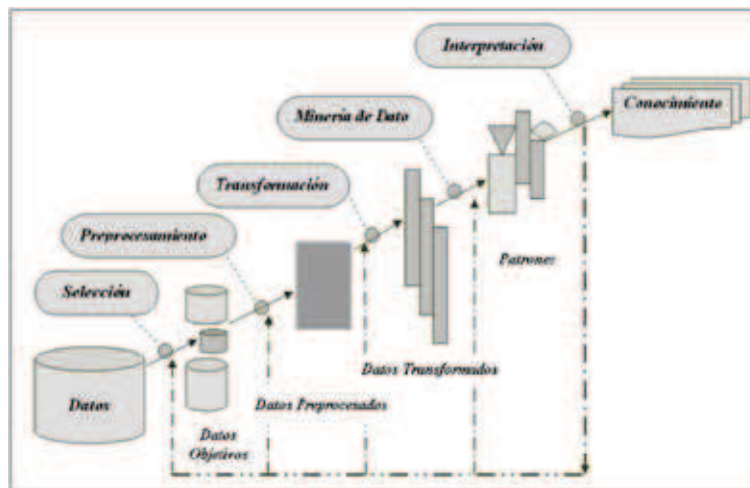


Figura 1. Etapas del proceso de descubrimiento de conocimiento en bases de datos

1.2 Etapa de minería de datos

La minería de datos es la etapa más importante del proceso DCBD, cuyo objetivo es la búsqueda, extracción y descubrimiento de patrones insospechados y de interés. La minería de datos consta de diferentes tareas, cada una de las cuales puede considerarse como un tipo de problema a ser resuelto por un algoritmo de minería de datos, afirman Adamo (2001) y Hernández, Ramírez & Ferri (2005), donde la tarea de clasificación por árboles de decisión es una de ellas.

La clasificación por árboles de decisión es, probablemente, el modelo más utilizado y popular por su simplicidad y facilidad para su entendimiento, de acuerdo con Han & Kamber (2001) y Sattler & Dunemann

(2001). El conocimiento obtenido en el proceso de aprendizaje, según Wang, Iyer & Scott (1998), se representa mediante un árbol en el cual cada nodo interior contiene una pregunta sobre un atributo concreto (con un hijo por cada posible respuesta) y cada hoja del árbol se refiere a una decisión (una clasificación). Durante la etapa de construcción del árbol, en forma recursiva, cada conjunto de datos se divide en subconjuntos de acuerdo a un criterio de particionamiento, con el fin de escoger el atributo que mejor separe los ejemplos restantes en clases individuales. Seleccionar el mejor punto de particionamiento, consideran Sattler & Dunemann (2001), es la parte de la construcción del árbol que mayor tiempo consume.

2. Metodología

Teniendo en cuenta las etapas del proceso DCBD, inicialmente se seleccionaron, de las bases de datos de la Universidad de Nariño los datos socio-económicos, académicos, disciplinares e institucionales de los estudiantes que ingresaron en los años 2004, 2005 y 2006 a los diferentes programas de pregrado, con el fin de hacerles un seguimiento completo hasta el año 2011, determinando si desertaron o no.

Con estos datos se construyó un repositorio de datos utilizando el *SGBD PostgreSQL*. A estos datos se les aplicó las etapas de pre-procesamiento y transformación con el fin de obtener conjuntos de datos limpios y listos para aplicarles las técnicas y los algoritmos de minería de datos. Los primeros resultados se obtuvieron utilizando la técnica de clasificación basada en árboles de decisión con la herramienta libre de minería de datos *Weka*. Finalmente, estos resultados fueron analizados, evaluados e interpretados para determinar la validez del conocimiento obtenido.

2.1 Etapa de selección de datos

El objetivo de esta etapa es obtener las fuentes internas y externas de datos que sirven de base para el proceso de minería de datos. Como fuentes internas, se seleccionaron las bases de datos de Admisiones y Registro Académico. Teniendo en cuenta la ventana de observación de este estudio (2004-2011), en estas bases de datos se encuentra almacenada la información personal y académica de 15.805 estudiantes.

Como fuentes externas principales se seleccionaron datos de la base de datos del Instituto Colombiano para el Fomento de la Educación Superior (ICFES), del Departamento Administrativo Nacional de Estadística (DANE), del Sistema para la Prevención de la Deserción en la

Educación Superior (SPADIES), del Sistema de Identificación de Beneficiarios Potenciales de Programas Sociales (SISBEN) e información de la Registraduría Nacional del Estado Civil Colombiano.

De los 15.805 registros se seleccionaron únicamente los datos de los estudiantes de las cohortes 2004, 2005 y 2006 con los atributos más relevantes para este estudio. Como resultado se obtuvieron 6870 registros y 62 atributos correspondientes a información socioeconómica, académica, disciplinar e institucional. Estos datos fueron almacenados en una base de datos construida con el sistema gestor de base de datos *PostgreSQL*, en la tabla TA62. Esta tabla servirá de base para las subsiguientes etapas del proceso de descubrimiento de patrones de deserción estudiantil.

2.2 Etapa de preprocesamiento de datos

El objetivo de esta etapa es obtener datos limpios, i.e. datos sin valores nulos o anómalos, que permitan obtener patrones de calidad. Por medio de consultas SQL ad-hoc o a través de histogramas, se analizó minuciosamente la calidad de los datos contenidos en cada uno de los atributos de la tabla TA62.

Teniendo en cuenta la relevancia de ciertos atributos para la investigación, los valores nulos de estos atributos fueron actualizados con los valores encontrados en fuentes externas. Por otra parte, los atributos con un alto porcentaje de valores nulos tales como *libreta-militar* (82.45%), *distrito-militar* (84.82%), *idmunicipio-conflicto* (89.57%), *periodo-grado* (89.50%), *padre-vive* (99.47%), *madre-vive* (99.58%), *descripción-laboral* (99.75%) entre otros, fueron eliminados por la imposibilidad de obtener estos valores con las fuentes externas o utilizando técnicas estadísticas como la media, mediana y la moda o derivando sus valores a través de otros.

2.3 Etapa de transformación de datos

El objetivo de esta fase es transformar la fuente de datos en un conjunto listo para aplicar las diferentes técnicas de minería de datos. Con el fin de generar conocimiento acerca de los factores socioeconómicos, académicos, disciplinares e institucionales que pueden incidir en la deserción estudiantil, se seleccionaron de la tabla TA62, los 31 atributos más representativos y con estos se creó la tabla TA31. De estos 31 atributos, se escogieron 18 para analizar el factor socioeconómico y 15 para el factor académico y se crearon las tablas TA18 y TA15 respectivamente. Dado el reducido número de atributos seleccionados para los factores disciplinar e institucional, estos se agregaron a la parte académica del estudiante.

Para facilitar la extracción de patrones, se *discretizaron* los valores numéricos de la tabla TA31 a valores nominales. Este proceso se llevó a cabo utilizando el filtro *discretize* de la herramienta *Weka* con el pará-

metro de frecuencias iguales (*useEqualFrequency*) a 6 valores. Por otra parte se adecuó la tabla TA31 al formato ARFF (*Attribute Relation File Format*) requerido por *Weka* para continuar con la etapa de minería de datos. En la tabla 1 se muestran los atributos de la tabla TA31 con los diferentes valores discretizados en formato ARFF. De acuerdo a esta tabla, los primeros 17 atributos y el atributo 31 de TA31 forman la tabla TA18 y corresponden a los atributos socioeconómicos. Los atributos del 17 al 31 de la tabla TA31 corresponden a los atributos académicos y estructuran la tabla TA15.

2.4 Etapa de minería de datos

El objetivo de la etapa de minería de datos es la búsqueda y descubrimiento de patrones insospechados y de interés aplicando tareas de descubrimiento tales como clasificación, *clustering*, patrones secuenciales, asociaciones entre otras.

La tarea de minería de datos escogida para el proceso de descubrimiento de patrones de deserción estudiantil en la Universidad de Nariño fue clasificación, teniendo en cuenta que con los valores del atributo clase *deserción* se puede construir un modelo de clasificación que determine las características de las estudiantes que desertan o no. Las reglas de clasificación se obtuvieron con la herramienta *Weka* utilizando el algoritmo *J48* que implementa el conocido algoritmo de árboles de decisión *C4.5* (Quinlan, 1993, 81).

Se utilizó el repositorio TA31 para obtener las reglas de clasificación generales que caracterizan a los estudiantes que desertan. Se escogió como clase, el atributo *deserción*. En la figura 2 se muestra el árbol de decisión generado por *Weka*. De igual manera, se utilizaron los conjuntos de datos TA18 y TA15 para determinar, respectivamente, los factores socioeconómicos y académicos que inciden en la deserción estudiantil. Las reglas de clasificación más relevantes se muestran en la sección de resultados.

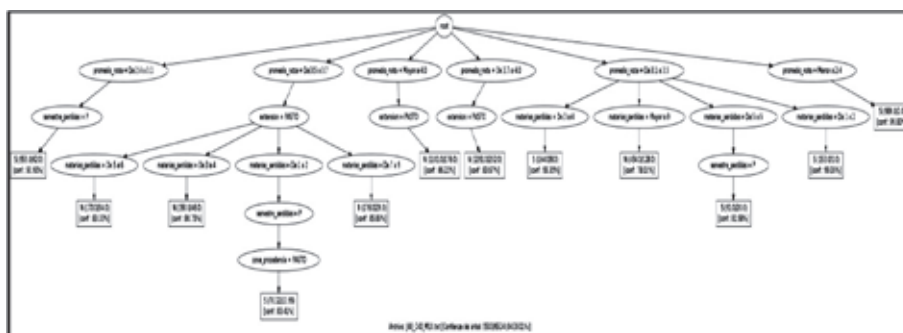


Figura 2. Árbol de decisión para reglas de clasificación generales a partir de TA31

Tabla 1. Atributos tabla TA31 en formato ARFF

No.	Formato arff atributo	Atributos y valores	No.	Formato arff atributo	Atributos y valores
1	@attribute	Genero {m,f}	17	@attribute	Edad_ingreso {'igual a 18','menor a 18','mayor a 22','de 21 a 22','igual a 19','igual a 20'}
2	@attribute	Estado_civil {soltero,casado,separado,'unión libre','madre soltera','viudo,religioso}	18	@attribute	Tipo_colegio {publico,privado}
3	@attribute	Zona_nacimiento {sur, CAPITAL,putumayo,norte,'otras regiones','centro occidental',costa,centro}	19	@attribute	Jornada_colegio {mañana,tarde,completa,noche,sabatina}
4	@attribute	Zona_procedencia { CAPITAL,norte,sur,'centro occidental','centro, costa,putumayo,'otras regiones'}	20	@attribute	lcfes_ponderado {'de 52 a 54','de 50 a 52','de 54 a 58','de 46 a 50','mayor a 58','menor a 46'}
5	@attribute	Regimen_salud {contributivo,subsidiado}	21	@attribute	lcfes_promedio {'de 53 a 56','de 48 a 50','de 46 a 48','de 50 a 53','menor a 46','mayor a 56'}
6	@attribute	Estrato {0,1,2,3,4,5,6,99}	22	@attribute	lcfes_total {'mayor a 475','de 420 a 450','de 450 a 475','de 400 a 420','de 375 a 400','menor a 375'}
7	@attribute	Padre {n,s}	23	@attribute	Extension { CAPITAL,tumaco,tuquerres,ipiales,samaniego,buesaco,'la union','ricaute}
8	@attribute	Ocupacion_padre {varios,'oficiales, operarios, artesanos,industria manufacturera, construccion y mineria','sin ocupacion',hogar,'profesionales universitarios, científicos e intelectuales',pensionados, etc.	24	@attribute	Facultad {'ciencias exactas y naturales','ciencias humanas','ciencias agrícolas','ciencias económicas y administrativas','ciencias de la salud',ingeniería,'ciencias pecuarias',educación,artes,'ingeniería agroindustrial'}

No.	Formato attr	Atributos y valores	No.	Formato attr	Atributos y valores
9	@attribute	Madre {n,s}	25	@attribute	Area_programa {matemáticas y ciencias naturales, ciencias sociales y humanas, agronomía, veterinaria y afines, economía, administración, contaduría y afines, ciencias de la salud, ingeniería, ciencias de la educación, bellas artes}
10	@attribute	Ocupación_madre {trabajadores no calificados, hogar, sin ocupación, trabajadores de los servicios y vendedores, varios, pensionados, profesionales universitarios, etc.	26	@attribute	Promedio_nota {de 2.4 a 3.1, de 3.5 a 3.7, mayor a 4.0, de 3.7 a 4.0, de 3.1 a 3.5, menor a 2.4}
11	@attribute	Tipo_residencia {arrendada o anticresada, propia, propia pagandose por cuotas}	27	@attribute	Materias_perdidas {de 3 a 4, mayor a 9, de 5 a 6, ninguna, de 1 a 2, de 7 a 9}
12	@attribute	Vive_con_familia {s, n}	28	@attribute	Semestre_perdidas {p, m, na, u, ce}
13	@attribute	Hermanos_universidad {n, s}	29	@attribute	Area_materia {formación específica, na, filosofía histórica, pedagogía, componente de fundamentación, formación instrumental, ciencias básicas, paradigmas teóricos, formación investigativa, formación matemática, etc.
14	@attribute	Ingresos_familiares {de 4540000 a 5980000, mayor a 8540000, de 2850000 a 4540000, 5980000 a 8854000, menor a 2850000}	30	@attribute	Veces_perdida {igual a 2, igual a 3, ninguna, igual a 1, igual a 4, mayor a 4}
15	@attribute	Valor_matricula_colegio {de 76639 a 106100, de 60248 a 76639, mayor a 106100, menor a 21550, de 21550 a 44369, de 44369 a 60247}	31	@attribute	Desercion {s,n}
16	@attribute	Valor_matricula {menor a 100259, de 120574 a 158846, de 100259 a 120574, de 234266 a 381504, de 158846 a 234266, mayor a 381504}			

2.5 Etapa de interpretación de datos

En esta etapa se evalúan e interpretan los patrones descubiertos para determinar su calidad y consolidar el conocimiento descubierto e incorporarlo en otro sistema para posteriores acciones o para confrontarlo con conocimiento previamente descubierto. Además, puede incluir la visualización de los patrones extraídos, la remoción de los patrones redundantes o irrelevantes y la traducción de los patrones útiles en términos que sean entendibles para el usuario.

Con el fin de evaluar la calidad y precisión de la predicción de las reglas de clasificación obtenidas se utilizó el método de validación cruzada con 10 pliegues (*n-fold cross validation*). Los resultados de esta etapa se analizan en la siguiente sección.

3. Resultados y discusión

Como resultado de interpretar el árbol de decisión, generado por el algoritmo J48 (figura 2) con el conjunto de datos TA31 se obtuvieron las reglas de clasificación más representativas con una confianza mayor que 80% que se muestran en la tabla 2, donde puede observarse que los factores predominantes en la deserción estudiantil en la Universidad de Nariño son los académicos, especialmente un promedio bajo y el tener materias perdidas en los primeros semestres de la carrera.

Con el fin de determinar los factores socioeconómicos que inciden en la deserción estudiantil, se generaron las reglas de clasificación con una confianza mayor que 80% y con el conjunto de datos TA18. El resultado se muestra en la tabla 3. Para determinar otros factores académicos asociados a la deserción estudiantil, se generaron reglas de clasificación con una confianza mayor que 80%, pero con el conjunto de datos TA15 sin tener en cuenta el atributo *promedio_nota*.

De acuerdo a las reglas de la tabla 3, los factores socioeconómicos que inciden en la deserción estudiantil son el valor de la matrícula mayor que \$381504 y proceder de la zona sur del departamento. El hecho de ser soltero, vivir con la madre y ser de la ciudad capital puede incidir también en la deserción.

Según los resultados, los factores académicos que inciden en la deserción estudiantil, además de un promedio bajo y el tener materias perdidas en los primeros semestres de la carrera, son la facultad a la que pertenece el estudiante y el área a la que pertenece las materias perdidas.

4. Conclusiones y trabajos futuros

Los primeros resultados obtenidos a través de la técnica de clasificación por árboles de decisión indica que esta es capaz de generar modelos consistentes con la realidad observada y el respaldo teórico, basándose únicamente en los datos que se encuentran almacenados en las bases de datos de una de las universidades. Una de las grandes dificultades que se presenta en esta clase de estudios es la mala calidad de los datos que muchas veces, después del proceso de limpieza, hace que se descarten ciertas variables por la imposibilidad de obtener sus valores y que de alguna manera influye en los resultados de la minería de datos.

Se ha obtenido un patrón general de deserción estudiantil determinado por un promedio bajo y el tener materias perdidas en los primeros semestres de la carrera. Se han determinado factores socioeconómicos y académicos asociados a la deserción estudiantil. La evaluación, análisis y utilidad de estos patrones permitirá soportar la toma de decisiones eficaces de las directivas universitarias enfocadas a formular políticas y estrategias relacionadas con los programas de retención estudiantil que actualmente se encuentran establecidos.

Como trabajos futuros están el continuar con el estudio de deserción estudiantil en la universidad estudiada, aplicando otras técnicas de minería de datos tales como asociación y *clustering* con el fin de determinar afinidades, similitudes y relaciones entre los factores socioeconómicos y académicos de las estudiantes que desertan. Para verificar la calidad y precisión del modelo de clasificación obtenido se utilizarán otros clasificadores y se compararán sus resultados. Se aplicará la misma metodología al repositorio de datos de la universidad cooperante en el estudio, para analizar y evaluar los patrones encontrados en ambas instituciones de educación superior.

Agradecimientos

Este proyecto de investigación se financia con recursos del Ministerio de Educación Nacional y con recursos de contrapartida de la Universidad de Nariño y la Institución Universitaria CESMAG.

Tabla 2. Reglas de clasificación más representativas con el conjunto de datos TA31

Reglas	Atributo Clase, Deserta	Soporte	Confianza
promedio_nota = De 2.4 a 3.1 & semestre_perdidas = P	S	0.1559	0.939
promedio_nota = De 3.7 a 4.0 & veces_perdida = 1	N	0.1551	0.8528
promedio_nota = Menor a 2.4	S	0.1519	0.998
promedio_nota = De 3.5 a 3.7 & extension = CAPITAL & materias_perdidas = De 7 a 9	N	0.0314	0.8585
promedio_nota = De 3.1 a 3.5 & materias_perdidas = De 3 a 4	S	0.0264	0.9535
promedio_nota = De 3.5 a 3.7 & materias_perdidas = De 1 a 2 & semestre_perdidas = P	S	0.0227	0.8108
promedio_nota = De 3.1 a 3.5 & materias_perdidas = De 5 a 6 & semestre_perdidas = P	S	0.017	0.8198
promedio_nota = De 3.5 a 3.7 & extension = CAPITAL & materias_perdidas = De 1 a 2 & semestre_perdidas = P & zona_procedencia = CAPITAL	S	0.0129	0.8341

Tabla 3. Reglas de clasificaciones socioeconómicas y académicas

Reglas	Atributo, Clase, Deserta	Soporte	Confianza
Reglas de clasificación socioeconómicas con el conjunto de datos TA18			
valor_matricula = De 158846 a 234266 & vive_con_familia = N	N	0.0933	0.8539
valor_matricula = De 100259 a 120574 & estado_civil = SOLTERO & madre = N & padre = N & hermanos_universidad = N & genero = F	N	0.0337	0.8
valor_matricula = De 158846 a 234266 & vive_con_familia = N & zona_nacimiento = SUR	N	0.0284	0.912
valor_matricula > 381504 & zona_procedencia = SUR	S	0.0242	0.9369
valor_matricula = De 158846 a 234266 & vive_con_familia = N & zona_nacimiento = COSTA	N	0.0224	0.882
valor_matricula = De 158846 a 234266 & vive_con_familia = N & zona_nacimiento = CENTRO OCCIDENTE	N	0.0136	0.908
valor_matricula = De 100259 a 120574 & estado_civil = SOLTERO & madre = S & tipo_residencia = PROPIA & zona_nacimiento = CAPITAL	S	0.0119	0.8071

Reglas	Atributo, Clase, Deserta	Soporte	Confianza
Reglas de clasificación académicas con el conjunto de datos TA15			
extension = CAPITAL & semestre_perdidas = M	N	0.2088	0.8128
extension = CAPITAL & semestre_perdidas = P & facultad = CIENCIAS EXACTAS Y NATURALES	S	0.1032	0.8692
extension = CAPITAL & semestre_perdidas = CE	N	0.0612	0.8672
extension = CAPITAL & semestre_perdidas = NA & facultad = CIENCIAS HUMANAS	N	0.0582	0.8079
extension = CAPITAL & semestre_perdidas = P & area_materia = CIENCIAS BÁSICAS & facultad = CIENCIAS PECUARIAS	S	0.0497	0.8025
extension = TUMACO	S	0.0458	0.8361
extension = IPIALES	S	0.0379	0.9919
extension = CAPITAL & semestre_perdidas = NA & facultad = CIENCIAS ECONÓMICAS Y ADMINISTRATIVAS	N	0.0336	0.863
extension = CAPITAL & semestre_perdidas = P & area_materia = FUNDAMENTACIÓN	S	0.0322	0.8524
extension = CAPITAL & semestre_perdidas = P & facultad = EDUCACIÓN	S	0.0233	0.8092
extension = TUMACO & area_programa = ECONOMÍA ADMINISTRACIÓN CONTADURÍA Y AFINES	S	0.0215	0.8786
extension = CAPITAL & semestre_perdidas = NA & facultad = ARTES	N	0.0207	0.8593
extension = CAPITAL & semestre_perdidas = P & area_materia = PEDAGOGÍA	S	0.019	0.879
extension = CAPITAL & semestre_perdidas = P & area_materia = CIENCIAS BÁSICAS & facultad = CIENCIAS EXACTAS Y NATURALES	S	0.0187	0.8852
extension = CAPITAL & semestre_perdidas = P & area_materia = FORMACIÓN MATEMÁTICA	S	0.0185	0.9587
extension = CAPITAL & semestre_perdidas = P & area_materia = FUNDAMENTACIÓN EN CIENCIAS EXACTAS Y NATURALES	S	0.0179	0.9658
extension = CAPITAL & semestre_perdidas = P & facultad = CIENCIAS DE LA SALUD	S	0.0175	0.807
extension = CAPITAL & semestre_perdidas = NA & facultad = CIENCIAS AGRICOLAS	N	0.0167	0.8716
extension = CAPITAL & semestre_perdidas = P & area_materia = LENGUA EXTRANJERA	S	0.0161	0.8571
extension = TUMACO & semestre_perdidas = P & area_programa = ECONOMÍA ADMINISTRACIÓN CONTADURÍA Y AFINES	S	0.0156	0.9412
extension = CAPITAL & semestre_perdidas = P & area_materia = FORMACIÓN BÁSICA & facultad = ARTES	S	0.0152	0.8283

Referencias bibliográficas

- ADAMO, Jean-Marc (2001). *Data Mining for Association Rules and Sequential Patterns: Sequential and Parallel Algorithms*. New York (USA): Springer-Verlag. 253 p. ISBN: 0-387-95048-6.
- AGRAWAL, Rakesh & SRIKANT, Ramakrishnan (1994). Fast Algorithms for Mining Association Rules. In: 20th International Conference on Very Large Data Bases, VLDB 1994, (12-15/09/1994). Santiago de Chile (Chile): VLDB. Proceedings. p. 487-499. ISBN: 1-55860-153-8.
- CHEN, Ming; HAN, Jiawei & YU, Philip (1996). Data mining: An overview from database perspective. In: IEEE Transactions on Knowledge and Data Engineering. Vol. 8, No. 6 (dic). Los Alamitos (CA, USA): IEEE Computer Society. p. 866-883. ISSN: 1041-4347.
- FAYYAD, Usama; PIATETSKY-SHAPIO, Gregory and SMYTH, Padraic. (1996). The KDD process for extracting useful knowledge from volumes of data. In: Communications of the ACM. Vol. 39, No. 11 (nov). New York (USA): ACM Digital Library. p 27-34. ISSN: 0001-0782.
- GARCÍA MORATE, Diego (s.f.). Manual de Weka [en línea]. Valladolid (España): MetaEmotion, S.L. <<http://www.metaemotion.com/diego.garcia.morate/download/weka.pdf>> 42 p. [consulta: 15/06/ 2012].
- HAN, Jiawei & KAMBER, Micheline (2001). *Data Mining: Concepts and Techniques*. San Francisco (CA, USA): Morgan Kaufmann Publishers, Academic Press. 550 p. ISBN: 1-55860-489-8.
- HERNÁNDEZ, José; RAMÍREZ, María & FERRI, César (2005). *Introducción a la Minería de Datos*. Madrid (España): Pearson Prentice Hall. 656 p. ISBN: 84-205-4091-9.
- LA RED, David, ACOSTA, Julio; CUTRO, Luis; URIBE, Valeria. & RAMBO, Alice (2010). Data Warehouse y Data Mining Aplicados al Estudio del Rendimiento Académico. En: Novena Conferencia Iberoamericana en Sistemas, Cibernética e Informática, CISCi 2010, (29/06-2/07/2010), Orlando (Florida, EE.UU.): International Institute of Informatics and Systemics. Memorias CISCi 2010, Volumen I, p. 289-294. ISBN: 978-1-934272-94-7.
- MEN (2006a). América Latina piensa la deserción. En: Boletín informativo Educación Superior. No 7 (dic). Bogotá (Colombia): Ministerio de Educación Nacional. 20 p. ISSN: 1794-2446.
- MEN (2006b). Deserción estudiantil: prioridad en la agenda. En: Boletín informativo Educación Superior. No 7 (dic). Bogotá (Colombia): Ministerio de Educación Nacional. 20 p. ISSN: 1794-2446.
- MEN (2009). Deserción estudiantil en la educación superior colombiana: metodología de seguimiento, diagnóstico y elementos para su prevención. Bogotá (Colombia): Ministerio de Educación Nacional. 158 p. ISBN: 978-958-691-366-9.
- PAUTSCH, Jesús (2009). *Minería de datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación*. Tesis de grado (Licenciado en Sistemas de Información). Posadas, Misiones (Argentina): Universidad Nacional de Misiones. 193 p.
- PAUTSCH, Jesús; LA RED, David & CUTRO, Luis (2010). Minería de datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación [en línea]. Posadas, Misiones (Argentina): Universidad Nacional de Misiones. <http://www.dataprix.com/files/Analisis%20de%20Desercion%20Univ_0.pdf> [consulta: 18/06/2012].
- PINZÓN, Liza (2011). Aplicando minería de datos al marketing educativo. En: Revista Notas de Marketing. No 1 (jun). Bogotá (Colombia): Universidad Sergio Arboleda, Escuela de Marketing y Publicidad. p 45-61. ISSN: 2248-4930
- QUINLAN, Ross (1993). *C4.5: Programs for Machine Learning*. San Francisco (CA, USA): Morgan Kaufmann Publishers. 299 p. ISBN: 1-55860-238-0.
- RESTREPO, Mauricio & LÓPEZ, Andrés (2008). Uso de la metodología Rough Sets en un modelo de deserción académica. En: XIV Congreso Ibero Latinoamericano de Investigación de Operaciones, CLAIO 2008, (9-12/09/2008), Cartagena (Colombia): Universidad del Norte. Libro de Memorias CLAIO 2008, p. 108-109. Ediciones Uninorte.
- SATTLER, Kai-Uwe. & DUNEMANN, Oliver (2001). SQL Database Primitives for Decision Tree Classifiers. In: The 10th ACM International Conference on Information and Knowledge Management - CIKM, (5-10/11/2001), Atlanta (Georgia, USA): ACM. Proceedings, p. 379-386. ISBN: 1-58113-436-3.
- SPOSITTO, Osvaldo; ETCHEVERRY, Martín; RYCKEBOER, Hugo & BOSSERO, Julio (2010). Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil. En: Novena Conferencia Iberoamericana en Sistemas, Cibernética e

- Informática, CISCi 2010, (29/06-2/07/2010), Orlando (Florida, EE.UU.): International Institute of Informatics and Systemics. Memorias CISCi 2010, Vol. I. ISBN: 978-1-934272-94-7.
- UPN (2005). La deserción estudiantil: reto investigativo y estratégico asumido de forma integral por la UPN [en línea]. En: Encuentro Internacional sobre Deserción en Educación Superior: experiencias significativas (17-18/05/2005) Bogotá (Colombia): Ministerio de Educación Nacional. <http://www.mineduacion.gov.co/1621/articles-85600_Archivo_pdf3.pdf> [consulta: 15/06/ 2012].
- VALERO, Sergio (2009). Aplicación de técnicas de minería de datos para predecir la deserción [en línea]. Izúcar de Matamoros, Puebla (México): Universidad Tecnológica de Izúcar de Matamoros. <<http://www.utim.edu.mx/~svalero/docs/MineriaDesercion.pdf>> [consulta: 10/06/2012].
- VALERO, Sergio; SALVADOR, Alejandro & GARCÍA, Marcela (2010). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos [en línea]. Izúcar de Matamoros, Puebla (México): Universidad Tecnológica de Izúcar de Matamoros. <<http://www.utim.edu.mx/~svalero/docs/e1.pdf>> [consulta: 10/06/2012].
- WANG, Min; IYER, Bala & SCOTT, Jeffrey (1998). Scalable Mining for Classification Rules in Relational Databases. In: International Database Engineering and Application Symposium, IDEAS 98, (08-10/07/1998), Cardiff (Wales, U.K.): IEEE Computer Society. Proceedings, p. 58-67. ISBN: 0-8186-8307-4.