Página inicial: 30 Página final: 44
Tipo de artículo: Investigación

EXTRACCIÓN DE PERFILES DE DESERCIÓN ESTUDIANTIL EN LA INSTITUCIÓN UNIVERSITARIA CESMAG [1]

Recibido: noviembre 10 de 2014 **/Revisado:** enero 20 de 2015 **/Aceptado:** abril 24 de 2015
Por: **Ricardo Timarán Pereira**[2] **y Javier Jiménez Toledo**[3]

Para citar este artículo/ To reference this article/ Para citar este artigo
Timarán, R. & Jiménez, J. (enero-junio, 2015). Extracción de perfiles de deserción estudiantil en la Institución Universitaria CESMAG., *Ivestigium IRE: Ciencias Sociales y Humanas,* VI (1), 30-44. doi: http://dx.doi.org/10.15658/CESMAG15.05060103

RESUMEN

El artículo presenta uno de los resultados de un proyecto de investigación cuyo objetivo fue 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 CESMAG, dos instituciones de educación superior de la ciudad de Pasto (Colombia), con la utilización de la técnica de minería de datos clasificación. Los resultados obtenidos corresponden a la I.U.CESMAG. Se descubrieron perfiles socioeconómicos y académicos de los estudiantes que desertan utilizando la técnica de clasificación basada en árboles de decisión. Se obtuvo un patrón general de deserción estudiantil determinado por un promedio de calificaciones bajo y el tener materias perdidas en los primeros semestres de la carrera. La evaluación, análisis y utilidad de estos patrones por parte de las directivas universitarias de la I.U.CESMAG, permitirá soportar la toma de decisiones eficaces, enfocadas a formular políticas y estrategias relacionadas con los programas de retención estudiantil que actualmente se encuentran establecidos.

Palabras clave: Clasificación por árboles de decisión, extracción de perfiles, deserción estudiantil, minería de datos.

[1] El artículo se deriva del proyecto de investigación "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", inscrito en el grupo de investigación aplicada en sistemas GRIAS de la Universidad de Nariño y en el grupo de investigación Tecnófila de la I.U.CESMAG, financiado y avalado por la I.U. CESMAG y Universidad de Nariño, ejecutado en el periodo octubre de 2012 a diciembre de 2013.

[2] Doctor en Ingeniería. Master of Science en Ingeniería. Especialista en Multimedia e Ingeniero de Sistemas y Computación. Director del grupo de investigación GRIAS. Profesor Asociado adscrito al Departamento de Sistemas de la Facultad de Ingeniería de la Universidad de Nariño (Pasto, Colombia). Correo electrónico: ritimar@udenar.edu.co

[3] Especialista en Docencia Universitaria. Ingeniero de Sistemas. Profesor de tiempo completo adscrito a la Facultad de Ingeniería de la Institución Universitaria CESMAG (Pasto, Colombia). Correo electrónico: jajimenez@iucesmag.edu.co

Revista
***INVESTIGIUM IRE:** Ciencias Sociales y Humanas*
Vol. VI. No. 1, Enero-Junio de 2015.
Categoría "C" publindex
ISSN: 2216-1473 e-ISSN: 2357-5239

DATA MINING IN THE EXTRACTION OF PROFILES OF STUDENT DROPOUT IN THE INSTITUTION UNIVERSITY CESMAG

ABSTRACT

This article presents one of the results of the research project whose objective was to detect patterns of drop-out student from socio-economic, academic, disciplinary and institutional data of the undergraduate programs students at the University of Nariño and Institution University Institution CESMAG, two institutions of higher education in the city of Pasto (Colombia), using data mining techniques. The results correspond to the Institution University CESMAG. Socio-economic academic profiles of students who drop out using the technique of classification based on decision trees were discovered. It has obtained a general pattern of drop out student determined by low average and lost areas in the first semesters of the career. The knowledge generated will allow supporting effective decision-making of focused University policies to formulate policies and strategies related to retention student currently established.

Key words: Decision trees, classification, student dropout, extraction of profiles, data mining.

MINERAÇÃO DE DADOS NA EXTRAÇÃO DE PERFIS DE DESERÇÃO DOS ESTUDANTES NA INSTITUCION UNIVERSITARIA CESMAG

RESUMO

Este artigo apresenta um dos resultados do projeto de pesquisa cujo objetivo era detectar padrões de deserção de estudantes a partir dos dados socioeconômicos, acadêmicos, disciplinares e institucionais dos alunos dos programas de graduação da Universidad de Nariño e a Institución Universitaria CESMAG, duas instituições de educação superior da cidade de Pasto (Colômbia), usando técnicas de mineração de dados. Os resultados correspondem à Institución Universitaria CESMAG. Foram descobertos perfis socioeconômicos e acadêmicos de estudantes que abandonam utilizando a técnica de classificação baseada em árvores de decisão. Foi obtido um padrão geral de deserção de estudantes determinado por uma média baixa e de ter perdidas disciplinas nos primeiros semestres da carreira universitária. O conhecimento gerado irá apoiar a tomada de decisões eficazes das diretivas universitárias destinadas a desenvolver políticas e estratégias relacionadas com programas de retenção de alunos que estão atualmente estabelecidos.

Palavras-chave: Árvores de decisão, classificação, deserção de estudantes, remoção de perfis, mineração de dados.

32
Revista
INVESTIGIUM IRE: *Ciencias Sociales y Humanas*
Vol. VI. No. 1, Enero-Junio de 2015.
Categoría "C" publindex
ISSN: 2216-1473 e-ISSN: 2357-5239

INTRODUCCIÓN

Los países de América Latina enfrentan desafíos similares en la educación superior, los cuales constituyen, entre otros, 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 y 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 el Ministerio de Educación Nacional-MEN, 2006a, p. 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 (IES), de las cuales 81 son públicas y 196 privadas. De acuerdo al Sistema Nacional de Información de la Educación Superior (SNIES) (citado en MEN, 2006a, p. 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 el MEN (2009, p. 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, la misma entidad (2006a, p. 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 el MEN (2006b, p. 1), la deserción estudiantil conlleva altos costos sociales y económicos que afectan a las familias, a los estudiantes, a las instituciones y al Estado.

De acuerdo con la Universidad Pedagógica Nacional (UPN, 2005), se entiende por deserción estudiantil, 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. El 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 al ser un estudiante de una institución de educación superior, no presenta actividad formativa durante dos semestres académicos consecutivos,

Revista
***INVESTIGIUM IRE:** Ciencias Sociales y Humanas*
Vol. VI. No. 1, Enero-Junio de 2015.
Categoría "C" publindex
ISSN: 2216-1473 e-ISSN: 2357-5239

lo cual equivale a un año de inactividad en su profesionalización. Esta última definición fue acogida como concepto base para esta investigación.

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

En el entorno internacional se han desarrollado algunos proyectos de investigación con aplicación de la minería de datos al descubrimiento de patrones de deserción estudiantil; un ejemplo es la Universidad Nacional de Misiones (Argentina) donde se realizó una investigación sobre deserción estudiantil utilizando esta técnica. 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 del programa de Analista en Sistemas de Computación de la Facultad de Ciencias Exactas, Químicas y Naturales, con el análisis de los datos de las cohortes entre los años 2000 al 2006 (Pautsch, 2009, p. 58; Pautsch, La Red & Cutro, 2010).

De igual manera, en la Universidad Nacional del Nordeste (Argentina) se realizó un estudio cuyo objetivo principal fue aplicar técnicas de almacenamiento y minería de datos basadas en *clusterin* para la búsqueda de perfiles de los alumnos de la asignatura de 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 (La Red et al., 2010).

En la Universidad Nacional de la Matanza (Argentina) 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 a 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 pre-procesamiento 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 (Spositto et al., 2010).

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 los primeros niveles. 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 (Valero, 2009; Valero, Salvador & García, 2010).

En el ámbito colombiano, en la Universidad de La Sabana se realizó un proyecto de investigación, donde el objetivo fue 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*, con la utilización del paquete *Rose2* (Restrepo & López, 2008).

Pinzón (2011) presenta la caracterización del perfil del estudiante desertor de la Escuela de Marketing y Publicidad de la Universidad Sergio Arboleda, con empleo de 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 *cluster* que para el caso de la investigación, constituyeron perfiles significativos.

En el ámbito regional, en la Universidad de Nariño y en la Institución Universitaria CESMAG, se realizó un proyecto de investigación interinstitu-

cional, financiado por el MEN, cuyo objetivo fue detectar patrones de deserción estudiantil a partir de los datos socioeconómicos, académicos, disciplinares e institucionales de los estudiantes de los programas de pregrado de estas dos instituciones, con el manejo de técnicas de minería de datos. En este artículo se presentan los resultados obtenidos en la I.U.CESMAG. Antes de esta investigación, en estas dos IES no se habían planteado proyectos de investigación que conlleven a analizar la información de la comunidad educativa recopilada en las bases de datos durante los últimos años, en lo relacionado al rendimiento académico y el grado de deserción de los estudiantes con las técnicas de minería de datos, que permitan aplicar estrategias efectivas que ayuden a minimizar estos factores y conlleven al mejoramiento de la calidad educativa en estas universidades.

De acuerdo a la información obtenida en las bases de datos de la mencionada Institución, en el periodo comprendido entre el primer semestre de 2004 y el segundo semestre de 2006, periodo objeto de estudio, ingresaron a los programas profesionales 1.054 estudiantes (véase tabla 1), de los cuales 589 desertaron, dato que corresponde al 55,88% de los ingresos, como se puede observar en la figura 1.

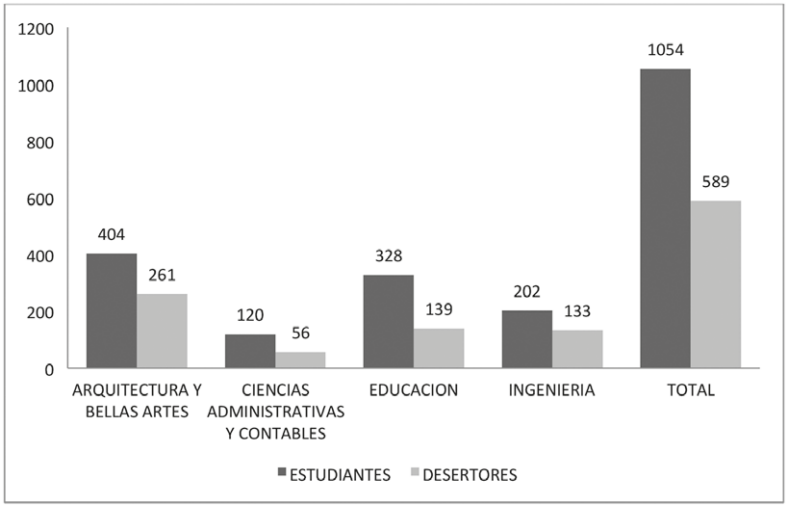

Figura 1. Deserción estudiantil por Facultad en cohortes 2004-2006

Fuente: Base de datos I.U. CESMAG

Tabla 1. Ingreso de estudiantes por Programa en cohortes 2004-2006

Facultad	Programa	No. Estudiantes	%
Arquitectura y Bellas artes	Arquitectura	207	19,64
	Diseño gráfico	197	18,69
Ciencias administrativas y contables	Contaduría pública	120	11,39
Educación	Licenciatura en Educación física	227	21,54
	Licenciatura en Educación preescolar	101	9,58
Ingeniería	Ingeniería de sistemas	202	19,17

Fuente: base de datos I.U. CESMAG

Revista
***INVESTIGIUM IRE:** Ciencias Sociales y Humanas*
Vol. VI. No. 1, Enero-Junio de 2015.
Categoría "C" publindex
ISSN: 2216-1473 e-ISSN: 2357-5239

Seguidamente, el artículo se organiza en secciones. Se describe a continuación la metodología del proceso de descubrimiento de conocimiento en bases de datos. En la sección de resultados se muestran éstos y, también, se interpretan los patrones obtenidos en la etapa de minería de datos; finalmente, en la última sección, se presentan las conclusiones y trabajos futuros.

METODOLOGÍA

Se acogió como metodología las diferentes etapas del proceso de descubrimiento de conocimiento en bases de datos; así, inicialmente se seleccionaron, de las bases de datos de la I.U.CESMAG, la información socio-económica, académica, disciplinar e institucional 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, para determinar si desertaron o no.

Con estos datos se construyó un repositorio 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. Se obtuvieron perfiles de deserción estudiantil con el empleo de la técnica de clasificación basada en árboles de decisión con la herramienta libre de minería de datos *Weka* (García, s.f.). Finalmente, estos resultados fueron analizados, evaluados e interpretados para determinar la validez del conocimiento obtenido.

Etapa de selección de datos

El objetivo de esta etapa fue obtener las fuentes internas y externas de datos que sirvieran de base para el proceso de minería de datos.

Como fuentes internas, se seleccionaron las bases de datos SIGA y ZEUS. La primera almacena información personal y académica de 3.212 estudiantes que ingresaron desde el año 2004 hasta el periodo A de 2009; y la segunda almacena información personal de 1.798 estudiantes que ingresaron a partir del periodo B de 2009 hasta el año 2011, lo cual constituyó la ventana de observación de este estudio. Se hizo una selección inicial de 62 atributos de estas dos bases de datos y, conjuntamente con sus registros, se almacenaron en el repositorio de datos denominado T5010A62, compuesto por 5.010 registros y 62 atributos. En la tabla 2 se muestra el número de estudiantes por facultad.

Tabla 2. Estudiantes por Facultad a 2011

Facultad	Estudiantes	%
Arquitectura y bellas artes	916	18,28
Ciencias administrativas y contables	914	18,24
Ciencias sociales y humanas	1.673	33,39
Educación	918	18,32
Ingeniería	589	11,76
Total	5.010	100

Fuente: base de datos I.U. CESMAG

Como fuentes externas principales, se seleccionaron las bases 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 estos 5010 registros, se escogieron únicamente aquellos datos de los estudiantes que ingresaron en los años 2004, 2005 y 2006, quienes fueron observados hasta el 2011. Como resultados se obtuvieron 1054 registros con 62 atributos, correspondientes a información socioeconómica, académica, disciplinar e institucional. Estos datos fueron almacenados en un nuevo repositorio denominado T1054A62. Con respecto a los 62 atributos, de éstos se seleccionaron inicialmente 29 que los investigadores consideraron más representativos al tener en cuenta un *ranking* de

36
Revista
INVESTIGIUM IRE: *Ciencias Sociales y Humanas*
Vol. VI. No. 1, Enero-Junio de 2015.
Categoría "C" publindex
ISSN: 2216-1473 e-ISSN: 2357-5239

atributos (basado en ganancia de información) y su experiencia como docentes. Finalmente, se obtuvo el conjunto de datos T1054A29, el cual sirvió de base para la siguiente etapa del proceso de descubrimiento de patrones de deserción estudiantil.

Etapa de pre-procesamiento de datos

El objetivo de esta etapa fue obtener datos limpios, i.e. datos sin valores nulos o anómalos que permitieran 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 T1054A29.

Al tener en cuenta la relevancia de ciertos atributos para la investigación, los valores nulos de éstos 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 (94.56%), distrito militar (88.95%), empresa trabajo (99.63%), necesidades educativas (79.80%), vivienda propia (65.77%), fueron eliminados por la imposibilidad de obtener estos valores con las fuentes externas o con la utilización de técnicas estadísticas como la media, mediana y la moda o al derivar sus valores a través de otros. Adicionalmente, se agregaron nuevos atributos obtenidos en fuentes externas y también derivándolos de otros atributos existentes. Estos se describen en la tabla 3.

Como resultado de estos procesos, se obtuvo el repositorio final T1054A28, que se utilizó en las etapas subsiguientes de transformación y minería de datos, cuyos atributos y distintos valores se muestran en la tabla 4.

Etapa de transformación de datos

El objetivo de esta fase fue transformar la fuente de datos en un conjunto listo para aplicar las diferentes técnicas de minería de datos.

Tabla 3. Nuevos atributos adicionados

Atributo	Descripción
Edad ingreso	Edad del estudiante al ingresar a la Institución. Fue calculado a partir de los atributos ya existentes: fecnac y fecing, encontrados en la base de datos SIGA, en los cuales se registran las fechas de nacimiento y de ingreso a la Institución.
Colegio	Colegios de los cuales egresaron los estudiantes. Obtenido de la información almacenada en la oficina de Archivo y Correspondencia.
ICFES promedio	Promedio del puntaje obtenido en las pruebas de Estado Saber 11 del ICFES.
ICFES total	Puntaje total obtenido en las pruebas de estado Saber 11 del ICFES.
Área programa	Áreas que tiene un programa
Promedio notas	Promedio de notas
Materias perdidas	Materias perdidas
Semestres perdidos	Semestres perdidos
Área materia	Materias que corresponde a una área
Veces perdida	Número de veces perdidas una materia
Estrato	Estrato socio económico del estudiante. Creado a partir del campo dir1, que almacena la clasificación de la estratificación socioeconómica por comuna.
Deserción	Campo que denota si el estudiante es desertor o no
Periodo deserción	Periodo en el que deserta cada estudiante
Valor matricula	Registra los valores de la matrícula para cada programa en las diferentes cohortes. Información obtenida en la oficina de Secretaria General de la I.U.CESMAG.

Revista
***INVESTIGIUM IRE:** Ciencias Sociales y Humanas*
Vol. VI. No. 1, Enero-Junio de 2015.
Categoría "C" publindex
ISSN: 2216-1473 e-ISSN: 2357-5239

Para facilitar la extracción de patrones, se *discretizaron* los valores numéricos de la tabla T1054A28 a valores nominales. Este proceso se llevó a cabo con empleo del filtro *discretize* de la herramienta *Weka* con el parámetro de frecuencias iguales (*use Equal Frequency*) a seis valores. Por otra parte, se adecuó la tabla T1054A28 al formato ARFF (*Atribute Relation File Format*) requerido por *Weka* para continuar con la etapa de minería de datos. En la tabla 4 se muestran los atributos del conjunto de datos T1054A28 con los diferentes valores discretizados en formato ARFF. Los primeros 16 atributos corresponden a los socioeconómicos, los 11 siguientes (del 17 al 27) corresponden a atributos académicos, y el último –deserción- corresponde a la clase. Con estos atributos, incluido el atributo clase, se formaron los repositorios T1054A17 y T1054A12, los cuales sirvieron para generar conocimiento acerca de los factores socioeconómicos y académicos que pueden incidir en la deserción estudiantil.

Tabla 4. Atributos del repositorio T1054A28 en formato ARFF

Formato ARFF – T1054A28	
@attribute	género {m, f}
@attribute	Estado civil {soltero, casado, unión libre, separado, madre soltera, viudo, religioso}
@attribute	Zona nacimiento {sur, Pasto, costa, Putumayo, centro, norte, centro occidente, otras regiones}
@attribute	Zona procedencia {Pasto, sur, centro, centro occidente, costa, putumayo, norte, otras regiones}
@attribute	Régimen salud {contributivo, subsidiado}
@attribute	Estrato {0, 1, 2, 3, 4, 5, 6, 7, 9}
@attribute	Padre {n, s}
@attribute	Ocupación padre {varios, oficiales, operarios, artesanos, industria manufacturera, construcción y minería, sin ocupación, hogar, otras}
@attribute	Madre {n, s}
@attribute	Ocupación madre {trabajadores no calificados, hogar, sin ocupación, trabajadores de los servicios y vendedores, pensionados, varios, otros}
@attribute	Tipo residencia {arrendada o anti-cresada, propia, propia pagándose por cuotas}
@attribute	Vive con familia {'s ', 'n '}
@attribute	Hermanos universidad {'n ','s '}
@attribute	Ingresos familiares {de 454.0000 a 598.0000, mayor a 854.0000, de 285.0000 a 454.0000, menor a 285. 0000, 598.0000 a 8.854.000}
@attribute	Valor matrícula {menor a 1.016.800, de 1.016.800 a 1.124.450, de 1.124.450 a 1.230.000, de 1.230.000 a 1.352.850, de 1.352.850 a 1.468.750, mayor a 1.468.750}
@attribute	Edad ingreso {igual a 18, menor a 18, mayor a 22, de 21 a 22, igual a 19, igual a 20}
@attribute	Tipo colegio {público, privado}
@attribute	Jornada colegio {mañana, tarde, completa, noche, sabatina}
@attribute	ICFES promedio {de 53 a 56, de 48 a 50, mayor a 56, menor a 46, de 50 a 53, de 46 a 48}
@attribute	ICFES total {mayor a 475, de 420 a 450, de 400 a 420, de 450 a 475, de 375 a 400, menor a 375}
@attribute	Facultad {Ciencias exactas y naturales, Ciencias de la salud, Ciencias económicas y administrativas, Ciencias humanas, Ciencias agrícolas, otras}
@attribute	Área programa {Matemáticas y ciencias naturales, Ciencias de la salud, Economía, Administración, Contaduría y afines, otras}
@attribute	Promedio nota {de 2.4 a 3.1, de 3.5 a 3.7, mayor a 4.0, menor a 2.4, de 3.1 a 3.5, de 3.7 a 4.0}
@attribute	Materias perdidas {de 3 a 4, mayor a 9, de 5 a 6, ninguna, de 7 a 9, de 1 a 2}

38
Revista
INVESTIGIUM IRE: *Ciencias Sociales y Humanas*
Vol. VI. No. 1, Enero-Junio de 2015.
Categoría "C" publindex
ISSN: 2216-1473 e-ISSN: 2357-5239

Formato ARFF – T1054A28	
@attribute	Semestre perdidas {p , m, na, ce, u }
@attribute	Área materia {formación específica, na, competencias básicas y formación humanística, formación investigativa, otros.}
@attribute	Veces perdida {igual a 2, igual a 3, ninguna, igual a 1, igual a 4, mayor a 4}
@attribute	Deserción {s, n}

Etapa de minería de datos

La minería de datos es la etapa más importante del proceso de *descubrimiento de conocimiento en bases de datos*, cuyo objetivo es la búsqueda, extracción y descubrimiento de patrones insospechados y de interés. Esta 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, como lo afirman Adamo (2001) y Hernández, Ramírez y 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 y Dunemann (2001). El conocimiento obtenido en el proceso de aprendizaje, según Wang, Iyer y 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 y Dunemann (2001), es la parte de la construcción del árbol que mayor tiempo consume.

Por estas razones, se escogió la tarea de minería de datos clasificación por árboles de decisión para el proceso de descubrimiento de patrones de deserción estudiantil en la I.U.CESMAG, al tener en cuenta que con los valores del atributo clase *deserción*, se puede construir un modelo de clasificación que determine las características de los estudiantes que desertan o no. Las reglas de clasificación se obtuvieron con la herramienta *Weka* con utilización del algoritmo que implementa el conocido algoritmo de árboles de decisión (Quinlan, 1993, p. 81). Este algoritmo utiliza el criterio de particionamiento ratio de ganancia (*gain ratio*), que evita que las variables con mayor número de posibles valores, salgan beneficiadas en la selección. Además, el algoritmo incorpora una poda del árbol de clasificación una vez que éste ha sido inducido (Hall et al, 2011). El parámetro más importante que se tuvo en cuenta para la poda, fue el factor de confianza C (*confidence level*), que influye en el tamaño y capacidad de predicción del árbol construido. Cuanto más baja se haga esa probabilidad, se exigirá que la diferencia en los errores de predicción, antes y después de podar, sea más significativa para no podar. El valor por defecto de este factor es del 25% y conforme va bajando este valor, se permiten más operaciones de poda y, por lo tanto, llegar a árboles cada vez más pequeños. Otro parámetro utilizado para variar el tamaño del árbol fue a través de M que especifica el mínimo número de instancias o registros por nodo del árbol.

Se utilizó el repositorio T1054A28 para obtener las reglas de clasificación generales que caracterizan a los estudiantes que desertan. Se escogió como clase, el atributo *deserción*, y se establecieron los factores de poda C=25% y M=10 (1%). En la figura 2 se muestra el árbol de decisión generado por *Weka*. De igual manera, se utilizaron los conjuntos de datos T1054A17 y T1054A12 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.

Revista
***INVESTIGIUM IRE:** Ciencias Sociales y Humanas*
Vol. VI. No. 1, Enero-Junio de 2015.
Categoría "C" publindex
ISSN: 2216-1473 e-ISSN: 2357-5239

Antes de construir el modelo de clasificación, se definió el procedimiento para probar la calidad del modelo y su validez. Según Hall et al. (2011) y Hernández, Ramírez y Ferri (2005), para entrenar y probar un modelo de clasificación, se divide los datos en dos conjuntos: uno de entrenamiento y otro de prueba. Por tal motivo, se utilizó el método de validación cruzada (*cross validation*), por ser la opción por defecto y la más comúnmente utilizada. Este mecanismo permite reducir la dependencia del resultado del experimento en el modo en el cual se realiza la partición. Para este caso particular, se utilizó el método de evaluación validación cruzada con 10 pliegues (*10-fold cross validation*). Este método consiste en dividir el conjunto de entrenamiento en 10 subconjuntos disjuntos de similar tamaño, llamados pliegues (*folds*), de forma aleatoria. Posteriormente, se realizan 10 iteraciones (igual al número de subconjuntos definido), donde en cada una se reserva un subconjunto diferente para el conjunto de prueba y los restantes 10-1 (uniendo todos los datos), para construir el modelo (entrenamiento). En cada iteración se calcula el error de muestra parcial del modelo. Por último, se construye el modelo con todos los datos y se obtiene su error promediando los obtenidos anteriormente en cada una de las iteraciones. Otra ventaja de la validación cruzada, es que la varianza de los 10 errores de muestra parciales, permite estimar la variabilidad del método de aprendizaje con respecto al conjunto de datos.

Por otra parte, se evaluó o estimó el coste del clasificador para el repositorio T1054A28 a través de la matriz de confusión. La matriz de confusión (*confusion matrix*) representa de forma detallada el número de instancias que son predichas por clase. La suma de los registros que se representan en cada fila i, i =1...n, constituyen el número de instancias que realmente pertenecen a la clase i. Similarmente, la sumatoria de los ejemplos o registros en cada columna j, j = 1...n, son las instancias que ha predicho el algoritmo al valor j de la clase. Los valores en la diagonal, son los aciertos, y el resto son los errores de clasificación (ejemplos que pertenecían a la clase i de la fila i y fueron clasificados incorrectamente en otra).

Etapa de interpretación de datos

En esta etapa se interpretan los patrones descubiertos con el fin de 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. Los resultados de esta etapa se analizan en la siguiente sección.

RESULTADOS

Como resultado de interpretar el árbol de decisión, generado por el algoritmo J48 (véase, figura 2), con el conjunto de datos T1054A28 se obtuvieron las reglas de clasificación más representativas, con una confianza mayor que 75%, lo cual se muestran en la tabla 5.

De acuerdo a la tabla 5, si el promedio de notas es menor que 2,4, el estudiante deserta. El 21,92% del total de estudiantes (1.054) que ingresaron a la Institución Universitaria CESMAG entre los años 2004 y 2006, se clasifica de esta manera, y el 39.4% del total de estudiantes desertores (589), cumplen con este patrón. También, si el promedio de notas esta entre 2,4 y 3,1, entonces, el estudiante deserta. El 14,8% de los 1.054 estudiantes que ingresaron en las cohortes estudiadas, tienen este perfil, y el 26,5% del total de desertores cumplen este patrón.

Por consiguiente, los factores predominantes en la deserción estudiantil en la Institución Universitaria CESMAG, son los académicos, especialmente un promedio de calificaciones bajo.

Para determinar otros factores académicos asociados a la deserción estudiantil, se generaron reglas de clasificación con una confianza mayor que el 75%, sin tener en cuenta el atributo *promedio_*

Revista
***INVESTIGIUM IRE:** Ciencias Sociales y Humanas*
Vol. VI. No. 1, Enero-Junio de 2015.
Categoría "C" publindex
ISSN: 2216-1473 e-ISSN: 2357-5239

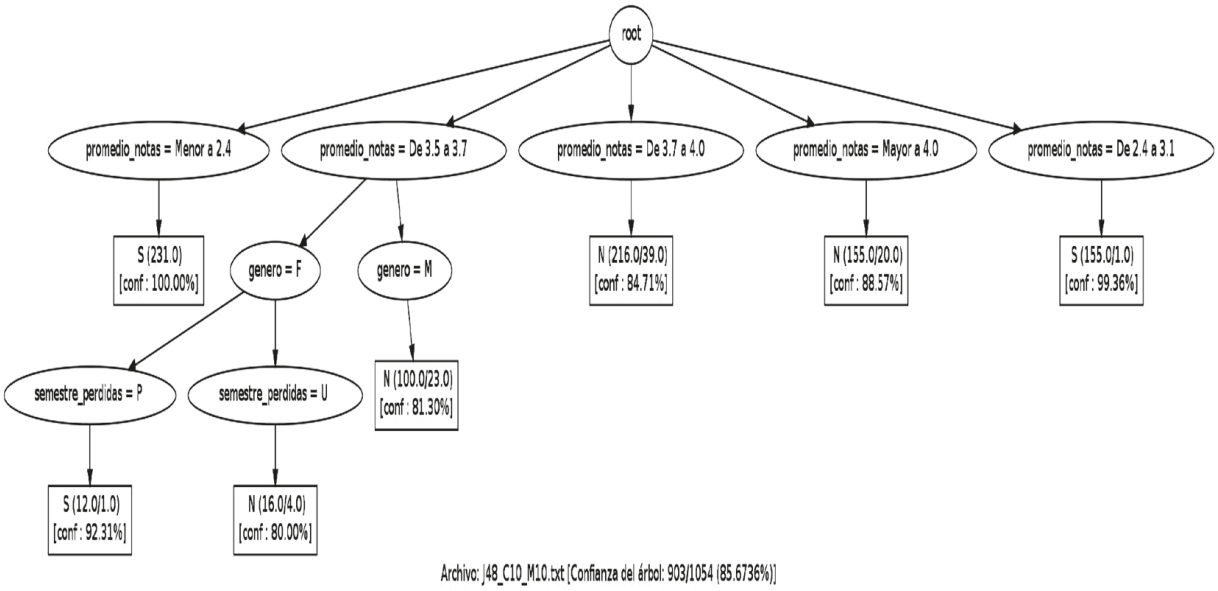

Figura 2. Patrones de deserción estudiantil

Tabla 5. Reglas generales más representativas sobre deserción estudiantil

Regla	Clase deserta	% soporte	% confianza	Registros por regla
Promedio notas menor que 2.4	S	21,92	100,00	231
Promedio notas entre 3.1 y 3.5 y valor matrícula menor que $1.016.800	S	1,80	100,00	19
Promedio notas entre 2.4 y 3.1	S	14,80	99,36	156
Promedio notas entre 3.5 y 3.7, genero = F y semestre perdidas = Primeros semestres	S	1,23	92,31	13
Promedio notas entre 3.1 y 3.5 y materias perdidas entre 1 y 2	S	4,74	88,00	50
Promedio notas entre 3.1 y 3.5 y valor matricula entre $1.016.800 y $1.124.450	S	2,09	81,82	22
Promedio notas entre 3.1 y 3.5 y valor matrícula entre $1.230.000 y $1.352.850	S	2,37	80,00	25
Promedio notas entre 3.1 a 3.5 y materias perdidas entre 7 y 9	S	2,37	80,00	25
Promedio notas entre 3.1 y 3.5 y valor matrícula entre $1.124.450 y $1.230.000	S	1,42	80,00	15
Promedio notas entre 3.5 y 3.7 y valor matrícula menor que $1.016.800	S	1,42	80,00	15
Promedio notas entre 3.1 y 3.5, valor matrícula entre $ 1.352.850 y $1.468.750 y jornada colegio = completa	S	1,27	77,66	13

Revista
***INVESTIGIUM IRE:** Ciencias Sociales y Humanas*
Vol. VI. No. 1, Enero-Junio de 2015.
Categoría "C" publindex
ISSN: 2216-1473 e-ISSN: 2357-5239

nota, para utilizar el repositorio con únicamente los atributos académicos T1054A12. Los resultados obtenidos muestran que el haber perdido materias en los primeros semestres (primero al cuarto), pertenecer a la Facultad de Arquitectura y Bellas Artes, tener perdido entre 5 o más materias diferentes, haber perdido más materias pertenecientes al área de Ciencias básicas como también al área de Competencias básicas y Formación humanística y tener un promedio en el ICFES menor a 46 puntos, son otros factores académicos que inciden en la posible deserción estudiantil.

Por otra parte, con el fin de determinar los factores socioeconómicos asociados a la deserción estudiantil, se generaron otras reglas de clasificación, utilizando el conjunto de datos T1054A17 que contiene sólo atributos socioeconómicos. El parámetro de confianza de la regla es mayor que 75%. Las reglas socioeconómicas más representativas de los estudiantes que desertan se muestran en la tabla 6.

Tabla 6. Reglas socioeconómicas sobre la deserción estudiantil

Regla	Clase deserta	% soporte	% confianza	Registros por regla
Valor matrícula entre $1.352.850 y $1.468.750 y estrato = 1	S	1,19%	87,15%	13
Valor matrícula entre $1.230.000 y $1.352.850	S	19,07%	80,60%	201
Valor matrícula entre $1.016.800 y $1.124.450, zona nacimiento = Pasto				
estrato = 3 y edad ingreso igual a 18 años	S	1,42%	80,00%	15
Valor matrícula entre $1.124.450 y $1.230.000, edad ingreso igual a 18 años y género = M	S	2,75%	79,31%	29
Valor matrícula mayor que $1.468.750				
y zonas nacimiento = Costa Pacífica	S	1,33%	78,57%	14
Valor matrícula menor que $1.016.800	S	22,20%	76,92%	234
Valor matrícula entre $1.352.850 y $1.468.750, estrato = 3 y edad ingreso entre 21 y 22 años	S	1,23%	76,92%	13
Valor matrícula entre $1.016.800 y $1.124.450, zonas nacimiento = Pasto y estrato = 1	S	1,56%	75,70%	16

De acuerdo a la tabla 6, si el valor promedio pagado por matrícula, en el transcurso de la carrera, está entre dos y tres salarios mínimos (valor salario mínimo promedio = $461.878), el estudiante deserta. El 19,07% del total de estudiantes (1.054) que ingresaron a la Institución Universitaria CESMAG entre los años 2004 y 2006, se clasificó de esta manera, y el 34,12 % del total de estudiantes desertores (589), cumplen con este patrón. También, si el valor promedio pagado por matrícula, en el transcurso de la carrera, es menor que dos salarios mínimos, el estudiante deserta. El 22.20% de los 1.054 estudiantes que ingresaron en las cohortes estudiadas, tiene este perfil, y el 40,27% del total de desertores, cumple este patrón.

Por consiguiente, el factor socioeconómico asociado a la deserción estudiantil en la Institución Universitaria estudiada, es el pagar un valor promedio de matrícula, durante la carrera, mayor que dos salarios mínimos. Entre otros factores socioeconómicos asociados a la deserción estudiantil, están: tener 18 o menos años al ingresar a la I.U.CESMAG, pertenecer al estrato 1 o 3, y ser originario de la Costa Pacífica o de la ciudad de Pasto.

42
Revista
INVESTIGIUM IRE: *Ciencias Sociales y Humanas*
Vol. VI. No. 1, Enero-Junio de 2015.
Categoría "C" publindex
ISSN: 2216-1473 e-ISSN: 2357-5239

CONCLUSIONES Y TRABAJOS FUTUROS

En esta investigación se obtuvieron perfiles de deserción estudiantil con la utilización de la tarea de minería de datos clasificación, con la técnica de árboles de decisión, a partir de la información de los estudiantes de las cohortes 2004, 2005 y 2006, almacenada en las bases de datos de la I.U.CESMAG.

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

Una de las grandes dificultades que se presentó en este estudio, fue la mala calidad de los datos, lo cual, en ocasiones, 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; por otra parte, hace más costoso el proceso de descubrimiento de patrones.

Como trabajo futuro está el continuar con el estudio de deserción estudiantil en la Institución Universitaria CESMAG, con aplicación de otras técnicas de minería de datos, tales como asociación *clustering*, y con el fin de determinar afinidades, similitudes y relaciones entre los factores socioeconómicos y académicos de las estudiantes que desertan. Además, solucionar la mala calidad de los datos con la construcción de un mercado de datos que almacene toda la información académica, limpia y transformada de los estudiantes de esta institución.

Revista
***INVESTIGIUM IRE:** Ciencias Sociales y Humanas*
Vol. VI. No. 1, Enero-Junio de 2015.
Categoría "C" publindex
ISSN: 2216-1473 e-ISSN: 2357-5239

REFERENCIAS

Adamo, J. (2001). *Data mining for association rules and sequential patterns: Sequential and parallel algorithms*. New York (USA): Springer-Verlag.

García, D. (s.f.). *Manual de Weka*. Recuperado de <http://www.metaemotion.com/diego.garcia.morate/download/weka.pdf>

Hall, M., Frank, E. and Witten, I. (mayo, 05 de 2011). Practical data mining: [Tutorials]. Recuperado de <http:// www.micai.org/2012/tutorials/Weka%20tutorials%20Spanish.pdf">

Han, J. and Kamber, M. (2001). *Data mining: Concepts and techniques*. San Francisco (CA, USA): Morgan Kaufmann Publishers, Academic Press.

Hernández, J., Ramírez, M. y Ferri, C. (2005). *Introducción a la Minería de Datos*. Madrid (España): Pearson Prentice Hall.

La Red, D., Acosta, J., Cutro, L., Uribe, V. & Rambo, A. (julio, 2010). Data Warehouse y Data Mining aplicados al estudio del rendimiento académico. En *Novena Conferencia Iberoamericana en Sistemas, Cibernética e Informática, CISCI.* International Institute of Informatics and Systemics, Orlando (Florida, EE.UU.).

Ministerio de Educación Nacional (2006a). *América Latina piensa la deserción*. (Boletín informativo Educación Superior. No 7). Bogotá (Colombia): MEN.

Ministerio de Educación Nacional (2006b). *Deserción estudiantil: prioridad en la agenda*. (Boletín informativo Educación Superior. No 7). Bogotá (Colombia): MEN.

Ministerio de Educación Nacional (2009). Bogotá (Colombia): MEN.

Pautsch, J. (2009). *Minería de datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación*. (Tesis de grado Licenciatura). Universidad Nacional de Misiones. Posadas, Misiones (Argentina).

Pautsch, J., La Red, D. y Cutro, L. (2010). Minería de datos aplicada al análisis de la deserción en la carrera de Analista en Sistemas de Computación. Recuperado de <http://www.dataprix.com/files/Analisis%20de%20Desercion%20Univ_0.pdf>

Pinzón, L. (junio, 2011). Aplicando minería de datos al marketing educativo. *Revista Notas de Marketing*, 1, 45-61.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning.* San Francisco (CA, USA): Morgan Kaufmann Publishers.

Restrepo, M. y López, A. (septiembre, 2008). Uso de la metodología Rough Sets en un modelo de deserción académica. En *XIV Congreso Ibero Latinoamericano de Investigación de Operaciones, CLAIO* Universidad del Norte, Cartagena (Colombia).

Sattler, K. & Dunemann, O. (2001). SQL Database Primitives for Decision Tree Classifiers. En *The 10th ACM International Conference on Information and Knowledge Management, CIKM,* Atlanta, Georgia (USA): ACM. Proceedings.

Spositto, O., Etcheverry, M., Ryckeboer, H. & Bossero, J. (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, CISC.I* International Institute of Informatics and Systemics, Orlando (Florida, EE.UU.).

Universidad Pedagógica Nacional (Agosto, 2005). La deserción estudiantil: reto investigativo y estratégico asumido de forma integral por la UPN. En: *Encuentro Internacional sobre Deserción en Educación Superior: experiencias significativas.* Recuperado de <http://www.mineducacion.gov.co/1621/articles-85600_Archivo_pdf3.pdf>

Valero, S. (2009). *Aplicación de técnicas de minería de datos para predecir la deserción*. Recuperado de <http://www.utim.edu.mx/~svalero/docs/MineriaDesercion.pdf>

Valero, S., Salvador, A. & García, M. (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.* Recuperado de http://www.utim.edu.mx/~svalero/docs/e1.pdf

Wang, M., Iyer, B. and Scott, J. (julio, 1998). Scalable Mining for Classification Rules in Relational Databases. In: *International Database Engineering and Application Symposium, IDEAS*. IEEE Computer Society. Proceedings, Cardiff (Wales,U.K.).