Descubrimiento de perfiles de deserción estudiantil con técnicas de minería de datos

Determinning school dropout profiles using data analysis

Ricardo Timarán Pereira*
Andrés Calderón Romero**
Javier Jiménez Toledo***

Fecha de recepción: 30 de marzo 2013
Fecha de aceptación: 30 de abril de 2013

Resumen

En este artículo se presentan los resultados de un proyecto de investigación, cuyo objetivo fue detectar patrones de deserción estudiantil utilizando técnicas de minería de datos, y tomando como referente datos socioeconómicos, académicos, disciplinares e institucionales de los estudiantes de los programas de pregrado de la Universidad de Nariño y la Institución Universitaria IUCESMAG de la ciudad de Pasto (Colombia). Se construyó así, un repositorio de datos con la información de los estudiantes que ingresaron a la Universidad de Nariño entre el primer semestre de 2004 y el segundo semestre de 2006, con una ventana de observación hasta el 2011. Utilizando técnicas de clasificación y *clustering*, se descubrieron perfiles socioeconómicos y académicos de los estudiantes que desertan. El conocimiento generado permitirá soportar la toma de decisiones eficaces por parte de las directivas universitarias, enfocadas a formular políticas y estrategias relacionadas con los programas de retención estudiantil.

* Departamento de Sistemas Universidad de Nariño. Correo electrónico: ritimar@udenar.edu.co

** Departamento de Sistemas Universidad de Nariño. Correo electrónico: aocalderon@udenar.edu.co

*** Facultad de Ingeniería. Institución Universitaria CESMAG. Pasto, Colombia. Correo electrónico: jajimenez@iucesmag.edu.co.

Palabras clave

Descubrimiento de perfiles, deserción estudiantil, minería de datos, clasificación, agrupamiento.

Abstract

The results obtained at the University of Nariño from 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. A data repository was built with information from students admitted to this university between the first semester of 2004 and second semester of 2006, with an observation window until 2011 Socioeconomic and academic profiles were discovered of students who drop using classification and clustering techniques. 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

Profiles discovery, student dropout, data mining, classification, clustering.

1. Introducción

La educación superior en América Latina presenta altas tasas de deserción estudiantil, especialmente en los primeros semestres académicos, lo cual conlleva a efectos de tipo financiero, académico y social, tanto para las Instituciones de Educación Superior (IES) como para el estudiante, la región, el país y el Estado [4]. Según el Instituto para la Educación Superior en América Latina y el Caribe (IESALC), Latinoamérica presentó en el año 2003 una cobertura promedio en educación superior del 28.7%, y una tasa de deserción estudiantil del 50% [5].

De acuerdo al Sistema Nacional de Información de la Educación Superior (SNIES), a 2006 la cobertura fue de 26.1%, lo cual equivale a 1.301.728 estudiantes [5], reafirmando uno de los principales problemas que enfrenta el sistema de educación superior colombiano, relacionado con los altos niveles de deserción estudiantil. Pese a que los últimos años 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 [6].

Según estadísticas del Ministerio de Educación Nacional, 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 [6]. A 2004, la deserción se estimó en 49%. Como causas del abandono estudiantil se señalaron: limitaciones económicas y financieras, bajo rendimiento académico, desorientación vocacional y profesional y dificultades para adaptarse al ambiente universitario [5].

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 [9]. El Ministerio de Educación Nacional, define la deserción 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 [6], siendo esta la definición 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 nos ofrece la minería de datos, se puede predecir, con un porcentaje muy alto de confiabilidad, la probabilidad de desertar de cualquier estudiante [10,11].

En este artículo se presentan los resultados de un proyecto de investigación, cuyo objetivo fue detectar patrones de deserción estudiantil utilizando técnicas de minería de datos, y tomando como referente datos socioeconómicos, académicos, disciplinares e institucionales de los estudiantes de los programas de pregrado de la Universidad de Nariño y la Institución Universitaria IUCESMAG de la ciudad de Pasto (Colombia),. Se construyó así, un repositorio de datos con la información de los estudiantes que ingresaron a la Universidad de Nariño entre el primer semestre de 2004 y el segundo semestre de 2006, con una ventana de observación hasta el 2011. Utilizando técnicas de clasificación y *clustering* con la herramienta Weka (Waikato Environment for Knowledge Analysis), se descubrieron perfiles socioeconómicos y académicos de los estudiantes que desertan. Weka es una de las suites más utilizadas en el área de descubrimiento de conocimiento en los últimos años [2].

La Universidad de Nariño es una institución pública de educación superior cuya área de influencia es el suroccidente de Colombia, su sede principal se encuentra en la ciudad de Pasto, capital del departamento de Nariño. En ella se forman la mayoría de estudiantes universitarios de la región.

El artículo está organizado por secciones. En la siguiente sección se describe la metodología utilizada en la investigación y cómo esta se desarrolló siguiendo las diferentes etapas del proceso de descubrimiento de conocimiento en bases de datos. En la tercera sección se presentan los resultados de la etapa de minería de datos y la discusión de resultados; en la última sección se presentan las conclusiones y trabajos futuros.

2. Proceso de descubrimiento de perfiles de deserción estudiantil

Teniendo en cuenta las etapas del proceso de descubrimiento de conocimiento en bases de datos, se seleccionaron 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 así el índice de deserción.

Con estos datos se construyó un repositorio de datos utilizando el SGBD PostgreSQL. A

estos datos se les aplicó las etapas de preprocesamiento 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 resultados se obtuvieron utilizando técnicas de clasificación, basadas en árboles de decisión y *clustering* con el algoritmo k-means. Finalmente, estos resultados fueron analizados, evaluados e interpretados para determinar la validez del conocimiento obtenido. Se utilizó la herramienta libre de minería de datos Weka para realizar estos procesos.

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 de la Universidad de Nariño, se seleccionaron las bases de datos *NOTAS* Y *REGISTRO UDENAR* de la Oficina de Control de Admisiones y Registro Académico (OCARA). 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 pertenecientes a 11 facultades.

Como fuentes externas se seleccionó información 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 las bases de datos de UDENAR, se seleccionó únicamente la información de los estudiantes de las cohortes 2004, 2005 y 2006 con los atributos más relevantes para este estudio. Como resultado se obtuvo un repositorio, con información socioeconómica, académica, disciplinar e institucional de los estudiantes de esta universidad. Los datos de los estudiantes de UDENAR fueron almacenados en la base de datos *REPOSITORIOUDENAR*, en la tabla T6870A62, compuesta por 6870 registros y 62 atributos. Esta tabla servirá de base para las subsiguientes etapas del proceso de descubrimiento de patrones de deserción estudiantil. La base de datos REPOSITORIOUDENAR, así como sus tablas fueron construidas con el sistema gestor de base de datos PostgreSQL.

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 T6870A62.

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*, *distrito_militar*, *id municipio_conflicto*, *periodo_grado*, *padre_vive* 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.

Como resultado de esta etapa y 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 T6524A62, por la calidad de los datos y por su importancia para el estudio, 31 atributos, con los cuales se creó la tabla T6870A31. De estos 31 atributos se escogieron 18 para analizar el factor socioeconómico y 14 para el factor académico, se crearon así las tablas T6870A18 y T6870A14 respectivamente. La descripción de estas tablas se muestra en la

Tabla 1: Tablas REPOSITORIOUDENAR

Tabla	Descripción
T6870A31	Tabla que contiene 6870 estudiantes de las cohortes que ingresaron en 2004-2006 y los 31 atributos a considerar en el estudio.
T6870A18	Tabla de los 6870 estudiantes y los atributos a considerar para los factores sociales y económicos.
T6870A14	Tabla de todos los 6870 estudiantes y los 14 atributos a considerar para los factores clínicos.

Fuente: elaboración propia.

tabla 1. Dado el reducido número de atributos seleccionados para los factores disciplinar e institucional, estos se agregaron a la parte académica del estudiante.

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.

Para facilitar la extracción de patrones de deserción, se discretizaron los valores numéricos de la tabla T6870A31 a valores nominales. Este proceso se llevó a cabo utilizando el filtro *discretize* de la herramienta Weka con el parámetro de frecuencias iguales (*useEqualFrequency*) a 6 valores. Por otra parte se adecuaron todas las tablas al formato ARFF (*Atribute Relation File Format*) requerido por Weka para continuar con la etapa de minería de datos.

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. Las tareas de minería de datos aplicadas en el proceso de descubrimiento de patrones de deserción estudiantil en la Universidad de Nariño fueron de clasificación y *clustering*.

La clasificación de datos permite obtener resultados a partir de un proceso de aprendizaje supervisado, por medio del cual se encuentran propiedades comunes entre un conjunto de objetos de una base de datos y se los cataloga en diferentes clases de acuerdo al modelo de clasificación [3]. La técnica de clasificación utilizada fue arboles de decisión. El modelo de clasificación basado en árboles de decisión, es probablemente el más utilizado y popular por su simplicidad y facilidad para entender [8, 12]. Para descubrir patrones de deserción estudiantil, se escogió como clase el atributo deserción que determina si el estudiante deserta 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 [7] con una confianza mínima de 75%.

El *clustering* es un proceso de aprendizaje no supervisado. Agrupa un conjunto de datos (sin un atributo de clase predefinido) basado en el principio de maximizar la similitud intraclase y minimizar la similitud interclase [1]. El análisis de *clustering* ayuda a construir particiones significativas de un gran conjunto de objetos basado en la metodología "divide y conquista", la cual descompone un sistema de gran escala en pequeños componentes, para simplificar el diseño y la implementación. La meta de la segmentación o *clustering* en una base de datos es la partición de ésta en segmentos o *clusters* de registros similares, que comparten un número de propiedades y son considerados homogéneos. Para encontrar similitudes entre los grupos de estudiantes que desertan y no desertan se escogió el *clustering* particional con el algoritmo k-means. Los resultados más relevantes de estas dos tareas de minería de datos se muestran en la sección de resultados.

2.5 Etapa de evaluación 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. Esta etapa 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.

3. Resultados y discusión

3.1 Clasificación

Construyendo un árbol de decisión sobre el conjunto de datos T6870A31 se obtuvieron las reglas de clasificación generales con una confianza mayor que 75 % y soporte mayor que 1.5%. Las reglas con más alta confianza de los estudiantes que desertan se muestran en la tabla 2.

Como se puede observar en la tabla 2, si el promedio de notas es menor que 2.4 el estudiante deserta. El 15% del total de estudiantes (6870) que ingresaron a la Universidad de Nariño entre los años 2004 y 2006 se clasifica de esta manera, y el 30.4 % del total de estudiantes desertores (3366), cumplen con este patrón. Por otra parte si el promedio de notas está entre 2.4 y 3.1 y tiene materias perdidas en los primeros semestres (primero a cuarto semestre) el estudiante deserta. El 14.9% de los 6870 estudiantes que ingresaron en las cohortes estudiadas tienen este perfil, y el 30.6% del total de desertores cumplen este patrón.

En resumen, los factores predominantes en la deserción estudiantil en la Universidad de Nariño son académicos, especialmente un promedio bajo y la pérdida de materias en los primeros semestres de la carrera.

Tabla 2. Reglas de clasificación más representativas con el conjunto de datos T6870A31

Antecedente	Consecuente	Soporte	Confianza	Registro por regla
Promedio nota = De 3.1 a 3.5 & materias perdidas = De 1 a 2	S	2,24	99,35	154
Promedio nota = Menos a 2.4	S	14,88	98,92	1022
Promedio nota = De 3.1 a 3.5 & materias perdidas = De 3 a 4	S	2,52	95,38	173
Promedio nota = De 2.4 a 3.1 & semestre perdidas = P	S	14,99	93,59	1030
Promedio nota = De 3.1 a 3.5 & materias perdidas = De 5 a 6	S	1,62	81,98	111
Promedio nota = De 3.5 a 3.7 & materias perdidas = De 1 a 2 & semestre perdidas = P	S	2,2	80,79	151

Fuente: elaboración propia.

Tabla 3. Reglas de clasificación socioeconómicas con el conjunto de datos T6870A18

Antecedente	Consecuente	Soporte	Confianza	Registro por regla
Valor matricula > 381504 & zona procedencia = SUR	S	2.24	92.02	166
Valor matricula < 100259 & madre = S & genero = M & vive con familia = S & tipo residencia = PROPIA.	S	1.37	83.05	94
Valor matricula entre 100259 y 120574 & estado civil = SOLTERO & madre = S & tipo residencia = PROPIA & zona nacimiento = PASTO	S	1.16	79.94	80
Valor matricula < 100259 & madre = S & genero = M &vive con familia = S & tipo residencia = ARRENDAD O ANTICRESADA.	S	1.81	79.92	124
Valor matricula entre 234266 y 381504 & madre = S	S	4.73	78.15	325
Valor matricula entre 234266 a 381504 & padre = S	S	5.91	77.34	406
Valor matricula entre 234266 a 381504 & madre = N & zona nacimiento = COSTA	S	2.65	77.34	182
Valor matricula > 381504 & zona procedencia = COSTA	S	2.14	76.896	147

Fuente: elaboración propia.

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 75% y con el conjunto de datos T6870A18. Las reglas más representativas de los estudiantes que desertan se muestran en la tabla 3.

De acuerdo a las reglas de la tabla 3, los factores socioeconómicos que inciden en la deserción estudiantil son el alto valor de la matrícula (mayor a $381.504) medio alta (entre $234.266 y $381.504), proceder de la zona sur o de la costa Pacífica del departamento de Nariño (Colombia). El hecho de ser soltero, vivir con la madre y ser de la ciudad de Pasto puede incidir también en la deserción.

Para determinar otros factores académicos asociados a la deserción estudiantil se realizó un proceso de poda de atributos, descartando paulatinamente el campo que determinaba el comportamiento general de las reglas. Según los resultados obtenidos, los factores académicos que inciden en la deserción estudiantil son, además de un promedio bajo y la pérdida de materias en los primeros semestres de la carrera, la facultad a la que pertenece el estudiante, destacando las facultades de Ciencias Exactas y Naturales, Ciencias de la Salud, Educación y Artes. Las materias perdidas están relacionadas con el área de formación matemática, la fundamentación en ciencias exactas y naturales, formación básica, pedagogía y economía, administración, contaduría y afines. En cuanto a la extensión de la Universidad de Nariño, los estudiantes desertan principalmente en las extensiones de Ipiales y Tumaco.

3.2 Clustering

Para llevar a cabo la tarea de *clustering* utilizando el algoritmo K-Means, se particionó

Tabla 4. Particionamiento del conjunto de datos T6870A31

Grupo	Cantidad	Porcentaje
1	3961	58
2	2909	42

Atributo	**Grupo 1**	**Grupo 2**
Genero	F	M
Estrato	2	1
Ingresos familiares	Menor a 285000	De 454000 a 598000
Valor matricula	Menor a 100259	De 2343266 a 381504
Icfes ponderado	De 46 a 50	Menor a 46
Icfes promedio	De 50 a 53	Menor a 46
Facultad	CIENCIAS HUMANAS	CIENCIAS ECONÓMICAS Y ADMINISTRATIVAS
Promedio nota	Mayor a 4.0	De 2.4 a 3.1
Materias perdidas	Ninguna	De 3 a 4
Area materia	NA	CIENCIAS BÁSICAS
Deserción	N	S

Fuente: elaboración propia.

el conjunto de datos T6870A31 en 2 grupos. Los resultados que se muestran en la tabla 4 son únicamente los valores de los atributos predominantes, que no son comunes entre los dos grupos y que los hacen diferentes entre sí. El algoritmo clasificó en el grupo 1 a los estudiantes que no desertan y en el grupo 2 a los que desertan.

Analizando los resultados que se muestran en la tabla 4, el 42 % del total de estudiantes que desertan tienen características similares, relacionadas con el género masculino, el estrato socioeconómico bajo, ingresos familiares bajos un valor de matrícula medio alto, un ponderado y promedio de ICFES bajo, el pertenecer a la facultad de Ciencias Económicas y Administrativas, con un promedio de notas bajo y con un número de materias perdidas, entre 3 y 4, pertenecientes al área de ciencias básicas.

Teniendo en cuenta únicamente los atributos socioeconómicos, el resultado de particionar el conjunto de datos T6780A18 se muestra en la tabla 5, en donde se señala que el 63% del total de estudiantes que desertaron tienen características relacionadas con los factores socioeconómicos, afectados por el hecho de vivir en una residencia propia, tener ingresos familiares anuales medio bajos y un valor de matrícula medio alto.

Teniendo en cuenta únicamente los atributos académicos, el resultado de particionar el conjunto de datos T6780A14 se muestra en la tabla 6.

De acuerdo a la tabla 6, los factores académicos predominantes que caracterizan al 69% del total de estudiantes que desertan están relacionados con un Icfes ponderado y un Icfes promedio por debajo de 50, un Icfes to-

Tabla 5. Particionamiento del conjunto de datos socieconómicos T6870A18

Grupo	Cantidad	Porcentaje
1	4321	63
2	2549	37

Atributo	Grupo 1	Grupo 2
Tipo residencia	PROPIA	ARRENDAD O ANTICRESADA
Ingresos familiares	De 4540000 a 5980000	De 2850000 a 4540000
Valor matricula	De 234266 a 381504	Menor a 100259
Deserción	S	N

Fuente: elaboración propia.

Tabla 6. Particionamiento del conjunto de datos académicos T6870A14

Grupo	Cantidad	Porcentaje
1	4742	69
2	2128	31

Atributo	Grupo 1	Grupo 2
Icfes ponderado	De 46 a 50	De 52 a 54
Icfes promedio	Menor a 46	De 50 a 53
Icfes total	De 420 a 450	De 450 a 475
Facultad	CIENCIAS ECONÓMICAS Y ADMINISTRATIVAS	CIENCIAS HUMANAS
Promedio nota	De 2.4 a 3.1	Mayor a 4.0
Materias perdidas	De 3 a 4	Ninguna
Senestre perdidas	P	NA
Area materia	CIENCIAS BÁSICAS	NA
Veces perdida	Igual a 1	Ninguna
Deserción	S	N

Fuente: elaboración propia.

tal entre 420 y 450, pertenecer a la facultad de Ciencias Económicas y Administrativas, un promedio de notas bajo (2.4 a 3.1), tener materias perdidas entre 3 y 4 en los primeros semestres de la carrera, en el área de las Ciencias Básicas, y haber repetido una materia una vez.

4. Conclusiones y trabajos futuros

Aplicando las técnicas de clasificación y *clustering* sobre los datos de los estudiantes que ingresaron a la Universidad de Nariño entre los años 2004 y 2006 con una ventana de

observación hasta el 2011, se ha obtenido un patrón común de deserción estudiantil, determinado por un promedio bajo y el tener materias perdidas en los primeros semestres de la carrera. Entre los factores socioeconómicos asociados a la deserción estudiantil están el pagar una matrícula alta, que a pesar de que este valor es bajo con relación al valor de las matrículas de otras universidades; es alta para el tipo de estudiantes que ingresan a la única universidad pública de la región, teniendo en cuenta que ellos son de estratos bajos. La zona de procedencia influye también en la deserción estudiantil. Otros factores académicos asociados a la deserción estudiantil son el promedio bajo del Icfes. La evaluación, análisis y utilidad de estos patrones permitirá soportar la toma de decisiones eficaces por parte de las directivas universitarias, las cuales deben estar enfocadas en 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 esta investigación fue 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.

Como trabajos futuros están el continuar con el estudio de deserción estudiantil en la Universidad de Nariño aplicando otras técnicas de minería de datos como asociación, clasificación con otras técnicas, *clustering* demográfico, entre otras; para así evaluar y comparar los patrones obtenidos. Aplicar la misma metodología para el descubrimiento de patrones de deserción estudiantil en la Institución Universitaria IUCESMAG y obtener patrones comunes para las dos IES.

5. Agradecimientos

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

6. Referencias

[1] Chen, M.; Han, J. & Yu, P. Data mining: An overview from database perspective. IEEE Transactions on Knowledge and Data Engineering. Vol. 8, No. 6. pp. 866-883. 1996.

[2] Garcia Morate, D. Manual de Weka. Recuperado de: http://www.metaemotion.com/diego.garcia.morate/download/weka.pdf .

[3] Hernández, O.J.; Ramírez, Q.M. O& Ferri, R.C. Introducción a la minería de datos. Madrid (España): Pearson Prentice Hall, 2005.

[4] Ministerio de Educación Nacional. América Latina piensa la deserción. En: Boletín informativo Educación Superior. No 7, p 14. Bogotá: Ministerio de Educación Nacional, 2006.

[5] Ministerio de Educación Nacional. Deserción estudiantil: prioridad en la agenda. En: Boletín informativo Educación Superior. No 7, p. 1. Bogotá: Ministerio de Educación Nacional, 2006b.

[6] Ministerio de Educación Nacional. Deserción estudiantil en la educación superior colombiana: metodología de seguimiento, diagnóstico y elementos para su prevención. Bogotá: Ministerio de Educación Nacional, 2009.

[7] Quinlan, J. R. C4.5: Programs for Machine Learning. San Francisco: Morgan Kaufmann Publishers. 1993.

[8] Sattler, K. & Dunemann, O. 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, 2001.

[9] Universidad Pedagógica Nacional. La deserción estudiantil: reto investiga-

tivo y estratégico asumido de forma integral por la UPN. En: Encuentro Internacional sobre Deserción en Educación Superior: experiencias significativas. Bogotá: Ministerio de Educación Nacional, 2005. Disponible en: http://www.mineducacion.gov.co/1621/articles-85600_Archivo_pdf3.pdf. Consultado: 15/06/ 2012.

[10] Valero, Sergio. Aplicación de técnicas de minería de datos para predecir la deserción. Izúcar de Matamoros, Puebla: Universidad Tecnológica de Izúcar de Matamoros. 2009. Disponible en: http://www.utim.edu.mx/~svalero/docs/MineriaDesercion.pdf.

[11] Valero, S.; Salvador, A. & García, M. 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: Universidad Tecnológica de Izúcar de Matamoros, 2010. Disponible en: http://www.utim.edu.mx/~svalero/docs/e1.pdf.

[12] Wang, M.; Iyer, B. & Scott, V., J. Scalable mining for classification rules in relational databases. In: International Database Engineering and Application Symposium, IDEAS 98, Cardiff: IEEE Computer Society. p. 58-67. 1998.