WEEF 2013 CARTAGENA COLOMBIA Septiembre 24 a 27
World Engineering Education Forum
CENTRO DE CONVENCIONES CARTAGENA DE INDIAS
Innovation in research and engineering education: key factors for global competitiveness
Innovación en investigación y educación en ingeniería: factores claves para la competitividad global

LA MINERÍA DE DATOS COMO UN MÉTODO INNOVADOR PARA LA DETECCIÓN DE PATRONES DE DESERCIÓN ESTUDIANTIL EN PROGRAMAS DE PREGRADO EN INSTITUCIONES DE EDUCACIÓN SUPERIOR

Ricardo Timarán Pereira, Andrés Calderón Romero

Universidad de Nariño
San Juan de Pasto, Colombia

Javier Jiménez Toledo

Institución Universitaria CESMAG
San Juan de Pasto, Colombia

Resumen

En este artículo se presenta uno de los resultados del proyecto de investigación financiado por el Ministerio de Educación Nacional 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 (la primera pública y la segunda privada) de la ciudad de Pasto (Colombia), utilizando técnicas de Minería de Datos. Tomando como metodología las etapas del proceso de descubrimiento de conocimiento en bases de datos, inicialmente se seleccionaron, de las bases de datos de estas instituciones 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 hasta el año 2011, determinando si desertaron o no. Con estos datos se construyó un repositorio de datos utilizando el SGBD PostgreSQL. Este repositorio se pre-procesó y transformó con el fin de obtener un conjunto de datos limpio y listo para aplicarle las 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 libre de minería de datos Weka. Los resultados fueron analizados, evaluados e interpretados para determinar la validez del conocimiento obtenido.

Palabras Clave: detección de patrones; deserción estudiantil; arboles de decisión

Abstract

One of results of the research project financed by the Ministry of National Education aimed 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 (public and private respectively) of the city of Pasto (Colombia), using data mining techniques is presented. Following as methodology the stages of knowledge discovery in databases, initially was selected from the databases of these institutions, the socio-economic, academic, disciplinary and institutional data of students who entered in 2004, 2005 and 2006 to various undergraduate programs, in order to track them until 2011, determining whether or not dropped out. With these data, a data repository was built using the PostgreSQL DBMS. This repository was preprocessed and transformed in order to obtain a clean data set and ready to apply the data mining techniques. Socioeconomic and academic profiles were discovered of students who drop out using classification by decision tree induction with free data mining tool Weka. The results were analyzed, evaluated and interpreted to determine the validity of the knowledge obtained.

Keywords*: detecting patterns; student dropout; decision trees*

1. Introducción

En Latinoamérica, la educación superior presenta altas tasas de deserción estudiantil, especialmente en los primeros semestres académicos, hecho que 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 (MEN, 2006a). 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% (MEN,2006b).

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 (MEN, 2006b). 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 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 (MEN, 2009). 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 (MEN, 2009). 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 (MEN, 2006b).

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 (UPN, 2005). 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 (MEN, 2009). Esta definición fue la 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 (Valero, et al., 2010).

En este artículo se presenta uno de los resultados del proyecto de investigación financiado por el Ministerio de Educación Nacional 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 (IES) de la ciudad de Pasto (Colombia), utilizando técnicas de Minería de Datos.

La Universidad de Nariño (UDENAR) es una institución pública de educación superior cuya área de influencia es el suroccidente de Colombia, cuya sede principal se encuentra en la ciudad de San Juan de Pasto, capital del departamento de Nariño. En ella se encuentra la mayoría de estudiantes universitarios de la región. Por otra parte, la Institución Universitaria CESMAG (IUCESMAG) es una fundación de derecho privado, de utilidad común y sin ánimo de lucro, con personería jurídica, autonomía administrativa y patrimonio independiente. Por su carácter académico es una Institución Universitaria, facultada para adelantar programas de formación en ocupaciones, de carácter operativo e instrumental, programas de formación académica en profesiones o disciplinas y programas de postgrado. La Institución tiene su domicilio principal en la ciudad de San Juan de Pasto, Departamento de Nariño.

El resto del artículo se organiza en secciones. En la siguiente sección se describe la metodología utilizada en la investigación y como se desarrolló la investigación siguiendo las diferentes etapas del proceso de descubrimiento en bases de datos. En la sección 3, se presentan los resultados de la etapa de minería de datos y la discusión de resultados y finalmente, en la última sección se presenta las conclusiones y trabajos futuros.

2. Metodología

Tomando como metodología las etapas del proceso de descubrimiento de conocimiento en bases de datos, inicialmente se seleccionaron, de las bases de datos de estas instituciones 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 hasta el año 2011, determinando si desertaron o no. Con estos datos se construyó un repositorio de datos utilizando el SGBD PostgreSQL. Este repositorio se preprocesó y transformó con el fin de obtener un conjunto de datos limpio y listo para aplicarle las 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 libre de minería de datos Weka. Los resultados fueron analizados, evaluados e interpretados para determinar la validez del conocimiento obtenido. 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.

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 REGISTROUDENAR 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. Por otra parte, para la Institución Universitaria CESMAG, se seleccionaron como fuentes internas las bases de datos SIGA y ZEUS de la Oficina de Admisiones, que almacenan información personal y académica de 5.010 estudiantes, pertenecientes a 5 facultades, bajo la misma ventana de observación de este estudio.

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 las bases de datos de UDENAR e IUCESMAG, 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 dos repositorios, con información socioeconómica, académica, disciplinar e institucional de los estudiantes de las dos IES. 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. Los datos de los estudiantes de la IUCESMAG fueron almacenados en la base de datos REPOSITORIOIUCESMAG en la tabla C1054A62, compuesta por 1054 registros y 62 atributos. Se seleccionaron los mismos 62 atributos para las dos IES con el fin de obtener patrones comunes de deserción estudiantil. Estas tablas servirán de base para las subsiguientes etapas del proceso de descubrimiento de patrones de deserción estudiantil. Las bases de datos REPOSITORIOUDENAR y REPOSITORIOIUCESMAG, 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 las tablas T6870A62 y C1054A62.

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, idmunicipio_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 para la UDENAR de la tabla T6524A62, por la calidad de los datos y por su importancia para el estudio, 31 atributos y con estos se creó la tabla T6870A31. De estos 31 atributos, se escogieron 18 para analizar el

factor socioeconómico y 14 para el factor académico. De igual manera en la IUCESMAG de la tabla C1054A62 se escogieron 28 atributos que formaron la tabla C1054A28 y de estos atributos, 17 para el análisis socioeconómico y 12 para la parte académica del estudiante-. Dado el reducido número de atributos seleccionados para los factores disciplinar e institucional, estos se agregaron a la parte académica del estudiante de cada IES.

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 en las dos IES, se discretizaron los valores numéricos de las tablas T6870A31 y C1054A28 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.

Después de trabajar con los repositorios independientes para cada IES, se procedió a construir un repositorio unificado que integrara ambos conjuntos, con el fin de encontrar patrones que inciden en la deserción estudiantil, tanto en instituciones públicas como privadas. Sin embargo, dado que el conjunto de la Universidad de Nariño posee más registros (6.870 estudiantes) y tres atributos más que el conjunto de la Institución Universitaria CESMAG (1.054 estudiantes y 28 atributos), se procedió a seleccionar una muestra del primer conjunto, con el fin de equipararlo con el tamaño del segundo conjunto y evitar un sesgo en los resultados finales.

Para este proceso, se establecieron distintas estrategias para integrar los conjuntos, pero finalmente se decidió trabajar únicamente con los registros de las facultades comunes entre las dos IES y los 28 atributos del conjunto de datos C1054A28. Como resultado se obtuvo el conjunto de datos U2136A28, que consta de 1.082 registros provenientes de UDENAR y de 1.054 de IUCESMAG, para un total de 2.136 registros y 28 atributos en común. Por otra parte se adecuo el repositorio unificado U2136A28 al formato ARFF (*Atribute Relation File Format*) requerido por Weka para continuar con la etapa de minería de datos. La descripción de los atributos de la tabla U2136A28 se muestra en la tabla 1. Los primeros 16 atributos pertenecen a los datos socioeconómicos del estudiante y los siguientes 11 (atributo 17 al atributo 27) determinan la parte académica del estudiante. El atributo 28 es el atributo clase.

2.4 Etapa de Minería de Datos

El objetivo de esta etapa 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 aplicada al repositorio unificado U2136A28, con el fin de descubrir patrones de deserción estudiantil en la Universidad de Nariño e Institución Universitaria CESMAG, fue clasificación con la técnica de árboles de decisión. Para esta tarea, 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 (Quinlan, 1993) con una confianza mínima de 75%. Las reglas más relevantes se muestran en la sección de resultados.

2.5 Etapa de Interpretación/ Evaluación de Datos

En esta etapa se interpretan y evalúan 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. 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.

Tabla 1. Descripción de los atributos del repositorio U2136A28

No	ATRIBUTO	DESCRIPCIÓN
1	Sexo	Género del estudiante
2	Edad_ingreso	Edad del estudiante al ingresar a la institución.
3	Estrato	Estrato socioeconómico al que pertenece el estudiante
4	Estado_civil	Estado civil del estudiante al ingresar en la institución
5	Régimen_salud	Régimen de salud al que está afiliado el estudiante
6	Zona_nacimiento	Zona del Departamento de Nariño o del país donde nació el estudiante
7	Zona_procedencia	Zona del Departamento de Nariño o del país donde reside el estudiante al ingresar en la institución
8	Padre	Si el estudiante tiene padre o no al momento de ingreso
9	Ocupación_padre	Ocupación del padre
10	Madre	Si el estudiante tiene madre o no al momento de ingreso
11	Ocupación_madre	Ocupación de la madre
12	Hermanos_u	Si el estudiante tiene o no hermanos estudiando en la IES
13	Tipo_residencia	Si el estudiante vive en una residencia propia o arrendada
14	Vive_con_flia	Si el estudiante vive con la familia o no
15	Ingresos-flia	Ingresos del núcleo familiar del estudiante al año
16	valor_matrícula	Valor promedio de la matrícula pagada por el estudiante durante la carrera
17	Tipo_colegio	Si el estudiante terminó el bachillerato en un colegio oficial o privado
18	Jornada_colegio	Jornada de estudios del colegio
19	Icfes_promedio	Promedio de las áreas de la prueba del ICFES.
20	Icfes_total	Puntaje total del ICFES
21	Facultad	Facultad a la que pertenece el estudiante
22	Área_programa	Área a la que pertenece el programa o carrera
23	Promedio_notas	Promedio de notas del estudiante en su carrera
24	Materias_perdidas	No. de materias que ha perdido el estudiante en la carrera
25	Semestre_perdidas	Determina si las materias pérdidas fueron en los primeros semestres (1 a 4), en los del medio (5-7) o en los finales (8-10)
26	Área_materia	Área a la que pertenecen la mayoría de las materias perdidas
27	Veces_perdida	Número de veces que ha perdido una materia
28	Deserción	Atributo clase que determina si el estudiante desertó o no

3. Resultados y Discusión

Como resultado de interpretar el árbol de decisión, generado por el algoritmo J48 con el conjunto de datos U2136A28, se obtuvieron las reglas de clasificación más representativas con una confianza mínima de 75% que se muestran en la tabla 2.

Tabla 2. Reglas de Clasificación con mayor confianza

ANTECEDENTE	CONS.	% SOP.	% CONF.	REG. POR REGLA
promedio_nota = Menor a 2,4	S	18,96	99,75	405
promedio_nota = De 2,4 a 3,1	S	17,88	94,24	382
promedio_nota = De 3,1 a 3,5 & materias_perdidas = De 1 a 2	S	3,42	91,78	73
promedio_nota = De 3,1 a 3,5 & materias_perdidas = De 7 a 9 & vive_con_familia = S	S	1,08	91,67	23
promedio_nota = De 3,1 a 3,5 & materias_perdidas = De 3 a 4 & semestre_perdidas = P	S	2,20	89,36	47
promedio_nota = De 3,1 a 3,5 & materias_perdidas = De 3 a 4	S	3,32	81,69	71
zona_procedencia = SUR & vive_con_familia = S	S	6,26	79,80	134
ingresos_familiares = De 5980000 a 8854000	S	2,32	78,90	50
ingresos_familiares = Mayor a 8854000	S	4,73	77,34	101

Como se puede observar en la tabla 2, los factores predominantes en la deserción estudiantil en las dos IES son los académicos y especialmente si el estudiante tiene un promedio de notas bajo y el tener materias perdidas en los primeros semestres de la carrera. Particularmente si la nota promedio es menor que 2,4 el estudiante deserta. El 19% del total de estudiantes (2.136) que ingresaron a la Universidad de Nariño y la Institución Universitaria CESMAG entre los años 2004 y 2006 se clasifica de esta manera y el 34,8 % del total de estudiantes desertores (1.165), cumplen con este patrón. De igual manera, si el promedio de notas esta entre 2,4 y 3,1 entonces el estudiante deserta. El 18% de los 2.136 estudiantes que ingresaron en las cohortes estudiadas tienen este perfil y el 32,8% del total de desertores cumplen este patrón.

Entre los factores socioeconómicos que inciden en la deserción estudiantil en estas dos IES es el vivir con la familia, proceder de la zona sur del Departamento de Nariño y tener unos ingresos familiares anuales mayores que $5.980.000 COP.

Para determinar otros factores asociados a la deserción estudiantil en ambas IES, se realizó un proceso de poda de atributos, descartando paulatinamente, el campo que determinaba el comportamiento general de las reglas. Como resultado de este proceso se obtuvieron los siguientes factores que pueden incidir en la deserción estudiantil: Pertenecer a la facultad de Ingeniería y Educación, tener un promedio del ICFES bajo (menor que 48), Haber perdido la mayoría de materias en el área de las Ciencias Básicas, Proceder de la Costa Pacífica Nariñense.

4. Conclusiones y Trabajos Futuros

Se ha obtenido un patrón general de deserción estudiantil en las dos IES 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 de Nariño e Institución Universitaria CESMAG 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.

5. Referencias

- MEN (2006a). Deserción estudiantil: prioridad en la agenda. En: Boletín informativo Educación Superior. No 7 (dic). Bogotá (Colombia): Ministerio de Educación Nacional. p 1. ISSN: 1794-2446.
- MEN (2006b). América Latina piensa la deserción. En: Boletín informativo Educación Superior. No 7 (dic). Bogotá (Colombia): Ministerio de Educación Nacional. p 14. 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.
- Quinlan, J.R. (1993). C4.5: Programs for Machine Learning. San Francisco (CA, USA): Morgan Kaufmann Publishers. 299 p. ISBN: 1-55860-238-0.
- UPN (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 (17-18/05/2005) Bogotá (Colombia): Ministerio de Educación Nacional.
- Valero, S., Salvador, A. & Garcia, 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 [en línea]. Izúcar de Matamoros, Puebla (Mexico): Universidad Tecnológica de Izúcar de Matamoros. < http://www.utim.edu.mx/~svalero/docs/e1.pdf> [consulta: 10/06/2012].

Agradecimientos

Este proyecto de investigación fue financiado con recursos del Ministerio de Educación Nacional y con recursos de contrapartida de la Universidad de Nariño e Institución Universitaria CESMAG

Sobre los autores

- **Ricardo Timarán Pereira**: Doctor en Ingeniería, Master of Science en Ingeniería, Especialista en Multimedia e Ingeniero de Sistemas y Computación. Director grupo de investigación GRIAS. Profesor Asociado, Universidad de Nariño. Correo electrónico: ritimar@udenar.edu.co.

- **Andrés Calderón Romero**: Master en Geoinformática, Ingeniero de Sistemas. Profesor hora cátedra, Universidad de Nariño. Correo electrónico: aocalderon@udenar.du.co.
- **Javier Jiménez Toledo**: Especialista en Docencia Universitaria, Ingeniero de Sistemas. Profesor tiempo completo, Institución Universitaria CESMAG. Correo electrónico: jajimenez@iucesmag.edu.co.
