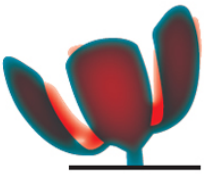

CONGRESO IBEROAMERICANO DE CIENCIA, TECNOLOGÍA, INNOVACIÓN Y EDUCACIÓN

BUENOS AIRES, ARGENTINA
12, 13 Y 14 DE NOVIEMBRE 2014

CONGRESSO IBERO-AMERICANO DE CIÊNCIA, TECNOLOGIA, INOVAÇÃO E EDUCAÇÃO

BUENOS AIRES, ARGENTINA
12, 13 Y 14 DE NOVEMBRO 2014

Detección de Patrones de Deserción Estudiantil en Programas de Pregrado de Instituciones de Educación Superior con CRISP-DM

TIMARAN,R;JIMENEZ,J

ISBN: 978-84-7666-210-6 – Artículo 758

Detección de Patrones de Deserción Estudiantil en Programas de Pregrado de Instituciones de Educación Superior con CRISP-DM

Ricardo Timarán Pereira

Universidad de Nariño, Ciudad Universitaria Torobajo

Pasto-Nariño-Colombia

ritimar@udenar.edu.co

Javier Jiménez Toledo

Institución Universitaria CESMAG

Carrera 20 A No. 14-54

Pasto-Nariño-Colombia

jajimenez@iucesmag.edu.co

Resumen

En este artículo se presentan los resultados del proyecto de investigación cuyo objetivo fue detectar patrones de deserción estudiantil a partir de los datos socioeconómicos y académicos 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), aplicando la metodología para proyectos de minería de datos CRISP-DM. Con este fin, se construyó un repositorio de datos con la información de los estudiantes que ingresaron a estas universidades entre el primer semestre de 2004 y segundo semestre de 2006, con una ventana de observación hasta el 2011. Se descubrieron perfiles socioeconómicos y académicos de los estudiantes que desertaron, utilizando las tareas de minería de datos clasificación, asociación y agrupación. El conocimiento generado permitirá soportar la toma de decisiones eficaces de las directivas universitarias enfocadas a formular políticas y estrategias relacionadas con los programas de retención estudiantil que actualmente se encuentran establecidos.

Palabras clave: Minería de Datos; CRISP-DM; Patrones Deserción Estudiantil.

Abstract

ISBN: 978-84-7666-210-6 – Artículo 758

In this paper, the results of the research project that aim was to discover student dropout patterns from socioeconomic, academic, disciplinary and institutional data of students from undergraduate programs at the University of Nariño and IUCESMAG University, two higher education institutions from San Juan de Pasto city (Colombia), using data mining techniques, are presented. Using the CRISP- DM methodology, a data repository, with information from students admitted to these universities between the first half of 2004 and second half of 2006 with an observation window until 2011, was built. Socioeconomic and academic profiles of students who dropped out were discovered, using classification, association and clustering data mining task. 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: Data Mining; CRISP-DM; Student Dropout Patterns.

1. Introducción

La deserción estudiantil en los programas de pregrado de la gran mayoría de Instituciones de Educación Superior (IES) tanto de Colombia como de Latinoamérica es un problema que tiene un impacto muy amplio en el desarrollo social y económico de un país. América Latina afronta retos parecidos en el área educativa. La financiación, el aumento de la cobertura, el aseguramiento de la calidad, el mejoramiento de la igualdad en el acceso y permanencia, mayor articulación con la educación secundaria, multiplicidad de la oferta para atender distintas necesidades y campos de interés (ciencia, tecnología, investigación, humanidades, artes, formación integral) así como también mayor vinculación con el sector laboral y productivo. Todas estas situaciones establecen el contexto de la deserción estudiantil en la educación superior (MEN, 2006a).

En Colombia, 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, 2012). 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 es la que se aplicó en esta investigación.

En este artículo se presentan los resultados del 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

ISBN: 978-84-7666-210-6 – Artículo 758

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 describen los trabajos relacionados con la aplicación de la minería de datos en la deserción estudiantil. En la sección III se desarrolla la investigación utilizando la metodología CRISP-DM. En la sección IV, 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. Trabajos Relacionados

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, 2009) (Valero et al., 2009).

A nivel Latinoamericano, algunas IES han desarrollado algunos proyectos de investigación con respecto a la deserción estudiantil, orientados a la aplicación de minería de datos para el descubrimiento de patrones y causas de la deserción.

En la Universidad Nacional de Misiones (Argentina) se realizó una investigación sobre deserción estudiantil utilizando las técnicas de minería de datos. Su objetivo principal fue maximizar la calidad que los modelos tienen para clasificar y agrupar a los estudiantes, de acuerdo a sus características académicas, factores sociales y demográficos, que han desertado de la Carrera Analista en Sistemas de Computación de la Facultad de Ciencias Exactas, Químicas y Naturales analizando los datos de las cohortes entre los años 2000 al 2006 (Pautsch, 2009) (Pautsch et al., 2010).

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

ISBN: 978-84-7666-210-6 – Artículo 758

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 al 2008. La implementación de este proceso se realizó con el software MS SQL Server para la generación de un almacén de datos, el software SPSS para realizar un preprocesamiento de los datos y el software Weka (*Waikato Environment for Knowledge Analysis*) para encontrar un clasificador del rendimiento académico y para detectar los patrones determinantes de la deserción estudiantil (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 que ingresan. Mediante la técnica de minería de datos clasificación y la herramienta Weka, encontraron relaciones entre atributos académicos que identifican y predicen la probabilidad de deserción y propusieron una herramienta para el tutor que le permite predecir la probabilidad de deserción de cualquier alumno en cualquier momento de su estancia escolar (Valero, 2009) (Valero et al., 2009).
En Colombia, en la Universidad de La Sabana se realizó un proyecto de investigación donde el objetivo era seleccionar, de una base de datos de estudiantes, los atributos que tuvieran mayor incidencia en la deserción de la Universidad en los últimos cuatro años, con la técnica de minería de datos clasificación por Rough Sets utilizando el paquete ROSE2 (Restrepo et al., 2008).

En la Escuela de Marketing y Publicidad de la Universidad Sergio Arboleda Pinzón se realizó un estudio para caracterizar el perfil del estudiante desertor, utilizando la técnica de minería de datos agrupamiento con el algoritmo K-means. Se analizaron las variables demográficas del alumno obtenidas en el registro de última matrícula del mismo semestre de abandono y las causas que lo generaron. Como resultado final, se obtuvieron tres tipos de clusters que para el caso de la investigación, constituyeron perfiles significativos (Pinzón, 2011).

3. Metodología

CRISP-DM (Cross-Industry Standard Process for Data Mining) es la metodología de referencia más utilizada en el desarrollo de proyectos de minería de datos en los ambientes académico e industrial (Gallardo, 2010). Comprende seis fases: Análisis del problema, análisis de los datos, preparación de los datos, modelado, evaluación y explotación (ver figura 1).

ISBN: 978-84-7666-210-6 – Artículo 758

Figura 1. Fases de la metodología CRISP-DM
Fuente: (Dedalus, 2012)

3.1 Análisis del Problema

En esta fase se requiere comprender con exactitud el problema al cual se le va a dar solución utilizando la minería de datos. Esto permitirá recolectar la información necesaria para interpretar con asertividad los resultados encontrados (Gallardo, 2010).

Según datos proporcionados por la Oficina de Control y Registro Académico (OCARA) de la Universidad de Nariño, en los programas de pregrado de la Universidad de Nariño, en las cohortes 2004-2006 ingresaron 6870 estudiantes, de los cuales, observados hasta el año 2011, desertaron 3366 estudiantes, correspondiente a un índice de deserción estudiantil del 49%. De igual manera, en el periodo comprendido entre el primer semestre del 2004 y el segundo semestre de 2006, en la IUCESMAG ingresaron 1.054 estudiantes, de los cuales, hasta la ventana de observación del 2011, desertaron 589 estudiantes, correspondiente a un índice de deserción del 56%.

El problema de deserción de estas dos IES se convirtió en un problema a resolver con minería de datos con el fin de detectar patrones de deserción estudiantil.

3.2 Análisis de los Datos

En esta fase se realiza la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis (Gallardo, 2010).

Se definieron las fuentes internas y externas de datos de las dos IES con el fin de construir posteriormente un conjunto de datos unificado que sirva de base para aplicar las técnicas de minería de datos.

ISBN: 978-84-7666-210-6 – Artículo 758

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.

3.3 Preparación de los Datos

Una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de minería de datos que se aplicarán. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato (Gallardo, 2010).

Por medio de consultas SQL ad-hoc y 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

ISBN: 978-84-7666-210-6 – Artículo 758

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 fase 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.

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.

Tabla 1. Atributos repositorio unificado 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

ISBN: 978-84-7666-210-6 – Artículo 758

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_matricula	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	Area_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

ISBN: 978-84-7666-210-6 – Artículo 758

		los finales (8-10)
26	Area_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

3.4 Modelado

En esta fase se seleccionan las técnicas de modelado más apropiadas para el proyecto de Minería de Datos.

Se seleccionaron las tareas de minería de datos clasificación, asociación y agrupamiento para descubrir conocimiento sobre deserción estudiantil en las dos IES a partir de los datos del repositorio integrado U2136A28.

Para la tarea de clasificación se utilizó la técnica de árboles 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 (Sattler et al., 2001). La clasificación con árboles de decisión considera clases disjuntas, de forma que el árbol conducirá a una y solo una hoja, asignando una única clase a la predicción. 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).

Para la poda del árbol se tuvo en cuenta el factor de confianza C (*confidence level*), que influye en el tamaño y capacidad de predicción del árbol construido. 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 (Garcia, 2010). Otra forma de variar el tamaño del árbol es a través del parámetro M que especifica el mínimo número de instancias o registros por nodo del árbol (Witten, 2000).

Para evaluar la calidad del modelo, dividiendo el repositorio de datos en dos conjuntos: entrenamiento y prueba, se escogió el método validación cruzada con n pliegues (*n-fold cross validation*). Este método consiste en dividir el conjunto de entrenamiento en n subconjuntos disjuntos de similar tamaño llamados pliegues (folds) de forma aleatoria. El número de subconjuntos se puede introducir en el campo Folds. Posteriormente se realizan n iteraciones (igual al número de subconjuntos definido), donde en cada una se reserva un subconjunto diferente para el conjunto de prueba y los restantes n-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 (Hernández, 2005). En este estudio se utilizó n=10 particiones, que es el valor que comúnmente se usa y que se ha probado que da buenos resultados (Hernández, 2005).

ISBN: 978-84-7666-210-6 – Artículo 758

Para la tarea de Asociación se utilizó el algoritmo Apriori (Agrawal, 1994), implementado en Weka en el paquete *WEKA.associations.Apriori*. Para evaluar las regla de asociación resultantes se utilizaron los parámetros soporte y confianza, dos métricas que permiten conocer la calidad de la regla. El soporte o cobertura de una regla se define como el número de instancias en las que la regla se puede aplicar. La confianza o precisión mide el porcentaje de veces que la regla se cumple cuando se puede aplicar (Hernández, 2005).

Para la tarea de agrupación se utilizó la técnica particional con el algoritmo K-means (Han, 2001), implementado en Weka, como *SimpleKmeans*, en el cual se configura el número de grupos (*NumClusters*) a formar y la semilla (*seed*), que se utiliza en la generación de un número aleatorio, el cual es usado para hacer la asignación inicial de instancias a los grupos. Para evaluar los resultados del agrupamiento, se utilizó el propio conjunto de entrenamiento, (*Use tranining set*), que indica que porcentaje de instancias se van a cada grupo.

3.4.1 Descubrimiento de patrones con Clasificación. Con el fin de detectar patrones de deserción estudiantil confiables utilizando árboles de decisión se generaron 35 árboles variando el factor de confianza C de 0.1 hasta 0.5 con un incremento de 0.1 y el número de instancias por nodo M de 10 en 10 iniciando en 10 hasta 70.

Se analizaron los árboles cuya precisión en el porcentaje de instancias correctamente clasificadas superaban el 75%. Las reglas de clasificación más representativas se muestran en la tabla 2.

3.4.2 Descubrimiento de reglas de Asociación. Con el fin de generar reglas de asociación fuertes (*strong rules*) i.e. reglas que superen el soporte y la confianza mínima, se estableció el soporte mínimo en 3% y la confianza en 80%. Se generaron 1957 reglas, de las cuales se escogieron las reglas con una confianza del 100%. Las ms representativas de acuerdo al soporte se muestran en la tabla 3.

3.4.3 Descubrimiento de Agrupaciones. Con el fin de generar grupos similares entre los registros del conjunto de datos U2136A28 en los cuales se encuentren grupos con estudiantes desertores y grupos con estudiantes no desertores, se configuró el parámetro K del algoritmo K-means en 2, 4 y 6 con una semilla de 100. Analizando los resultados obtenidos, los dos grupos formados con K=2 se escogieron como los más representativos para caracterizar a los estudiantes que desertan y los que no. En la tabla 4 se muestran estos dos grupos.

Tabla 2. Reglas de Clasificación

Antecedente	Consecuente	% soporte	% confianza	No. registros regla
promedio_nota = Menor a 2,4	S	19	99,8	405
promedio_nota = De 2,4 a 3,1	S	17,9	94,2	382

ISBN: 978-84-7666-210-6 – Artículo 758

promedio_nota = De 3,1 a 3,5 & materias_perdidas = De 1 a 2	S	3,42	91,8	73
promedio_nota = De 3,1 a 3,5 & materias_perdidas = De 7 a 9 & vive_con_familia = S	S	1,08	91,7	23
promedio_nota = De 3,1 a 3,5 & materias_perdidas = De 3 a 4 & semestre_perdidas = P	S	2,20	89,4	47
promedio_nota = De 3,1 a 3,5 & materias_perdidas = De 3 a 4	S	3,32	81,7	71
zona_procedencia = SUR & vive_con_familia = S	S	6,26	79,8	134
ingresos_familiares = De 5980000 a 8854000	S	2,32	78,9	50
ingresos_familiares = Mayor a 8854000	S	4,73	77,3	101

Tabla 3. Reglas de Asociación

Antecedente	Consecuente	% Soporte	% Confianza
estado_civil=SOLTERO & promedio_nota=Menor a 2.4 & semestre_perdidas=P & veces_perdida=Igual a 1	S	16,1	100
genero=M & estado_civil=SOLTERO & promedio_nota=Menor a 2.4 & veces_perdida=Igual a 1	S	12,3	100
genero=M & promedio_nota=Menor a 2.4 & semestre_perdidas=P & veces_perdida=Igual a 1	S	12,2	100
tipo_colegio=PUBLICO & promedio_nota=Menor a 2.4 & semestre_perdidas=P &	S	11,3	100

ISBN: 978-84-7666-210-6 – Artículo 758

Antecedente	Consecuente	% Soporte	% Confianza
veces_perdida=Igual a 1			
estado_civil=SOLTERO & tipo_colegio=PUBLICO & promedio_nota=Menor a 2.4 & veces_perdida=Igual a 1	S	11,1	100
estado_civil=SOLTERO & promedio_nota=Menor a 2.4 & semestre_perdidas=P & tipo_universidad=PRIVADA	S	10,7	100
estado_civil=SOLTERO & promedio_nota=Menor a 2.4 & veces_perdida=Igual a 1 & tipo_universidad=PRIVADA	S	10,3	100
promedio_nota=Menor a 2.4 & semestre_perdidas=P & veces_perdida=Igual a 1 & tipo_universidad=PRIVADA	S	10,1	100
zona_nacimiento=PASTO & promedio_nota=Menor a 2.4 & semestre_perdidas=P & veces_perdida=Igual a 1	S	10,1	100
estado_civil=SOLTERO & zona_nacimiento=PASTO & promedio_nota=Menor a 2.4 & veces_perdida=Igual a 1	S	10,0	100
estado_civil=SOLTERO & promedio_nota=De 2.4 a 3.1 & semestre_perdidas=P & veces_perdida=Igual a 1	S	9,0	100
genero=M & estado_civil=SOLTERO & promedio_nota=Menor a 2.4 & tipo_universidad=PRIVADA	S	8,5	100
estado_civil=SOLTERO & icfes_promedio=Menor a 46 & promedio_nota=Menor a 2.4 & semestre_perdidas=P	S	8,4	100

ISBN: 978-84-7666-210-6 – Artículo 758

Antecedente	Consecuente	% Soporte	% Confianza
genero=M & promedio_nota=Menor a 2.4 & semestre_perdidas=P & tipo_universidad=PRIVADA	S	8,4	100
estado_civil=SOLTERO & jornada_colegio=MAÑANA & promedio_nota=Menor a 2.4 & semestre_perdidas=P	S	8,3	100

Tabla 4. Agrupaciones

Atributo	Total (2136)	Grupo1 (1332)	Grupo2 (804)
estrato	2	2	3
edad_ingreso	Menor a 18	Igual a 18	Menor a 18
facultad	INGENIERÍA	INGENIERÍA	ARQUITECTURA Y BELLAS ARTES
area_programa	INGENIERÍA	INGENIERÍA	BELLAS ARTES
promedio_nota	De 3.7 a 4.0	Mayor que 4.0	Menor que 2.4
materias_ perdidas	De 1 a 2	De 1 a 2	De 5 a 6
area_materia	CIENCIAS BÁSICAS	CIENCIAS BÁSICAS	COMPETENCIAS BÁSICAS Y FORMACIÓN HUMANÍSTICA

ISBN: 978-84-7666-210-6 – Artículo 758

tipo_universidad	PUBLICA	PUBLICA	PRIVADA
deserción	S	N	S

3.4 Evaluación

En esta fase 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.

3.4.1 Análisis de patrones de Clasificación. Como se observa en la tabla 2, si el promedio de notas 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 clasifican 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.

De acuerdo a lo anterior, los factores predominantes en la deserción estudiantil en la Universidad de Nariño y la Institución Universitaria CESMAG son los académicos, especialmente un promedio de notas bajo. Otros factores que influyen en la deserción estudiantil son: pertenecer a las facultades de Ingeniería y Educación, haber perdido materias en los primeros semestres, haber perdido entre 3 y 9 materias diferentes, haber perdido materias pertenecientes al área de Ciencias Básicas, tener un promedio de ICFES menor a 48 puntos, proceder de la zona sur del departamento de Nariño o tener ingresos familiares mayores que $5.980.000.

3.4.2 Análisis de patrones de Asociación. Teniendo como referencia la tabla 3, las reglas de asociación más representativas son las siguientes:

Regla 1. El 100% de los estudiantes que desertan son solteros, su promedio de notas es menor que 2.4, han perdido materias en los primeros semestres (1 a 4) y todas las materias las han perdido una sola vez. El 16.1% del total de estudiantes (2.136) que ingresaron a la Universidad de Nariño y la Institución Universitaria CESMAG entre los años 2004 y 2006 cumplen con este patrón.

Regla 2. El 100% de los estudiantes que desertan realizaron su bachillerato en un colegio público, son solteros, su promedio de notas es menor que 2.4, han perdido materias en los primeros semestres (1 a 4) y todas las materias las han perdido una sola vez. El 11.3% del total de estudiantes (2.136) que ingresaron a la Universidad de Nariño y la Institución Universitaria CESMAG entre los años 2004 y 2006 cumplen con este patrón.

Regla 3. El 100% de los estudiantes que desertan son hombres solteros, su promedio de notas es menor que 2.4 y son de una universidad privada, para este caso IUCESMAG. El 8.5% del total de estudiantes (2.136) que ingresaron a la Universidad de Nariño y la Institución Universitaria CESMAG entre los años 2004 y 2006 cumplen con este patrón.

ISBN: 978-84-7666-210-6 – Artículo 758

De acuerdo a los anteriores resultados, dentro de los factores asociados a la deserción estudiantil están el ser soltero, tener un promedio bajo, haber perdido materias en los primeros semestres y provenir de un colegio público.

3.4.3 Análisis de Agrupaciones. Los resultados que se muestran en la tabla IV son únicamente de los atributos, cuyos valores son diferentes entre los dos grupos. El algoritmo K-means clasificó en el grupo 1 a los estudiantes que no desertan y en el grupo 2 a los que desertan.

De acuerdo a las características similares el patrón que determina a los estudiantes que desertan de la Institución Universitaria CESMAG: es: pertenecer a un estrato socioeconómico medio, ser menor de edad, pertenecer a la facultad de Arquitectura y Bellas Artes, de un programa académico que pertenece al área de Bellas Artes, con un promedio de notas menor que 2.4, haber perdido entre 5 y 6 materias del área de Competencias Básicas y Formación Humanística.

Por otra parte, el patrón que determina a los estudiantes que desertan de la Universidad de Nariño es: pertenecer a un estrato socioeconómico bajo, ser menor de edad, pertenecer a la facultad de Ingeniería, de un programa académico que pertenece al área de Ingeniería, con un promedio de notas entre 3.7 y 4.0, haber perdido entre 1 y 2 materias del área de Ciencias Básicas.

3.5 Explotación o Implementación

En esta fase se trata de explotar la potencialidad de los modelos, integrarlos en los procesos de toma de decisión de la organización y difundir informes sobre el conocimiento extraído.

Con esta investigación, tanto la Universidad de Nariño como la Institución Universitaria CESMAG cuenta con un estudio que permite obtener información de calidad sobre factores socio-económicos y académicos que inciden en la deserción estudiantil en los programas de pregrado de las respectivas IES. Los patrones descubiertos permitirán soportar la toma de decisiones eficaces de las directivas universitarias de las dos IES, enfocadas a formular políticas y estrategias relacionadas con los programas de retención estudiantil que actualmente se encuentran establecidos.
Los resultados obtenidos fueron socializados en cada una de las dos IES con la presencia de las directivas universitarias respectivas.

4. Conclusiones y Trabajos Futuros

Los perfiles de deserción estudiantil obtenidos a través de las técnicas de minería de datos: clasificación, asociación y agrupamiento indican que éstas son capaces de generar modelos consistentes con la realidad observada y el respaldo teórico, basándose únicamente en los datos que se encontraron almacenados en las bases de datos de la Universidad de Nariño y de la Institución Universitaria CESMAG, complementados con fuentes externas de datos pertenecientes principalmente a SISBEN, Sistema de Prevención y Análisis de la Deserción en las Instituciones de Educación Superior (SPADIES),Alcaldía Municipal de Pasto (Estratificación), Departamento Administrativo Nacional de Estadística (DANE), Instituto Colombiano

ISBN: 978-84-7666-210-6 – Artículo 758

para el Fomento de la Educación Superior (ICFES) y Registraduría Nacional del Estado Civil Colombiano.

Una de las grandes dificultades que se presentó en esta investigación fue la mala calidad de los datos de la bases de datos de la Universidad de Nariño e Institución Universitaria CESMAG, que hizo que, después del proceso de limpieza de datos, se descartaran ciertas variables por la imposibilidad de obtener sus valores y que de alguna manera influyen en los resultados sobre deserción estudiantil obtenidos.

Se ha obtenido un patrón general de deserción estudiantil común para las dos IES participantes en este proyecto de investigación y es el tener un promedio de notas bajo, el tener materias perdidas en los primeros semestres de la carrera y un puntaje promedio de ICFES bajo.

Se recomienda a las directivas universitarias de las dos IES evaluar, analizar y determinar la utilidad de estos patrones obtenidos en esta investigación para soportar la toma de decisiones eficaces enfocadas a formular políticas y estrategias relacionadas con programas de retención estudiantil.

Como trabajos futuros están el construir un sistema de inteligencia de negocios que cuente con una bodega de datos histórica y limpia, un sistema de análisis multidimensional OLAP, un sistema de minería de datos, visualizadores y generadores de reportes que facilite y provea datos consolidados y de calidad principalmente de las áreas académica, financiera y administrativa que optimice la toma de decisiones y facilite este tipo de estudios tanto en la Universidad de Nariño como en la Institución Universitaria CESMAG.

Agradecimientos

Al Sistema de Investigaciones de la Universidad de Nariño por financiar este estudio.

Referencias

AGRAWAL, R. and SRIKANT, R. (1994). “Fast Algorithms for Mining Association Rules”. *VLDB Conference*. Santiago de Chile.

DAEDALUS. (2012). *Técnicas de modelado predictivo de la contaminación en la ciudad sostenible*. [en línea]. white paper. http://www.daedalus.es/blog/es/whitepaper-tecnicas-de-modelado-predictivo-de-la-contaminacion-en-la-ciudad-sostenible. [fecha de consulta: 05/05/2013].

GALLARDO, J. (2010). *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM.* [en linea]. http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf. [fecha de consulta: 05 /05/2013].

ISBN: 978-84-7666-210-6 – Artículo 758

GARCÍA, M. and ÁLVAREZ, A. (2010). *Análisis de Datos en WEKA –Pruebas de Selectividad*". [en línea]. http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf, [fecha de consulta: 05/05/2013].

HAN, J. and KAMBER, M. (2001). *Data Mining Concepts and Techniques*. San Francisco: Morgan Kaufmann Publishers.

HERNÁNDEZ, J., RAMIREZ, M.J. and FERRI, C. (2005). *Introducción a la Minería de Datos*. España: Pearson Prentice Hall.

LA RED, D., ACOSTA, J.C., CUTRO, L., URIBE, V.E. and RAMBO, A.R. (2010). "Data Warehouse y Data Mining Aplicados al Estudio del Rendimiento Académico". *Novena Conferencia Iberoamericana en Sistemas, Cibernética e Informática, CISCI 2010. Orlando: International Institute of Informatics and Systemics. Memorias CISCI 2010*, Volumen I, pág. 289-294.

MEN. (2006a). "América Latina piensa la deserción". *Boletín informativo Educación Superior*. No 7, pág. 14.

MEN. (2006b). "Deserción estudiantil: prioridad en la agenda". *Boletín informativo Educación Superior*. No 7, pág. 1.

MEN. (2009). *Deserción estudiantil en la educación superior colombiana: metodología de seguimiento, diagnóstico y elementos para su prevención*. Bogotá: Ministerio de Educación Nacional, pág. 68–73.

PAUTSCH, J. (2009). "Minería de datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación". Tesis de grado (Licenciado en Sistemas de Información), Posadas, Misiones (Argentina): Universidad Nacional de Misiones, pág.193.

PAUTSCH, J. LA RES, D. and CUTRO, L. (2010). *Minería de datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación.* [en línea]. Posadas: Universidad Nacional de Misiones. http://www.dataprix.com/files/Analisis%20de%20Desercion%20Univ_0.pdf. [fecha de consulta: 18/06/2012].

PINZÓN, L. (2011). "Aplicando minería de datos al marketing educativo". *Notas de Marketing.* No 1, Bogotá: Universidad Sergio Arboleda, Escuela de Marketing y Publicidad, pág 45-61.

QUINLAN, J.R. (1993). C4.5: *Programs for Machine Learning*. San Francisco: Morgan Kaufmann Publishers, pág. 299.

RESTREPO, M. and LÓPEZ, A. (2008). "Uso de la metodología Rough Sets en un modelo de deserción académica". *XIV Congreso Ibero Latinoamericano de Investigación de Operaciones*, CLAIO, 2008. Cartagena: Universidad del Norte, Libro de Memorias CLAIO 2008, pág.108-109.

SATTLER, K. and DUNEMANN, O. (2001). "SQL Database Primitives for Decision Tree Classifiers". In: *CIKM*, Atlanta, Georgia.

SPOSITTO ,O., ETCHEVERRY, M., RYCKEBOER, H. and BOSSERO, J. (2010). "Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil". *Novena Conferencia Iberoamericana en*

ISBN: 978-84-7666-210-6 – Artículo 758

Sistemas, Cibernética e Informática, CISCI, 2010. Orlando: International Institute of Informatics and Systemics. Memorias CISCI 2010, Volumen I.

UPN. (2012). *La deserción estudiantil: reto investigativo y estratégico asumido de forma integral por la UPN.* Encuentro Internacional sobre Deserción en Educación Superior: experiencias significativas. [en línea]. Bogotá: Ministerio de Educación Nacional. http://www.mineducacion.gov.co/1621/articles-85600_Archivo_pdf3.pdf, [fecha de consulta: 15/06/ 2012].

VALERO, S. (2009). *Aplicación de técnicas de minería de datos para predecir la deserción*. [en línea]. Izúcar de Matamoros: Universidad Tecnológica de Izúcar de Matamoros, 2009. http://www.utim.edu.mx/~svalero/docs/MineriaDesercion.pdf. [fecha de consulta: 10/06/2012].

VALERO, S., SALVADOR, A. and GARCÍA, M. (2009). *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: Universidad Tecnológica de Izúcar de Matamoros. http://www.utim.edu.mx/~svalero/docs/e1.pdf. [fecha de consulta: 10/06/2012].

WITTEN, I. and FRANK, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco: Morgan Kaufmann Publishers, pág. 365.

ISBN: 978-84-7666-210-6 – Artículo 758