Caracterización de la deserción estudiantil en educación superior con minería de datos

Javier Alejandro Jiménez Toledo[a], Silvio Ricardo Timarán Pereira[b]

[a] Programa de Ingeniería de Sistemas, Institución Universitaria CESMAG, Carrera 20 A No. 14-54 Pasto, Nariño, Colombia.
jajimenez@iucesmag.edu.co

[b] Departamento de Sistemas, Universidad de Nariño, Calle 18 Carrera 50 Pasto, Nariño, Colombia.
ritimar@udenar.edu.co

Resumen. Este artículo presenta los resultados de investigación que a partir de datos socioeconómicos, académicos, disciplinares e institucionales, permitieron la caracterización de perfiles de deserción de los estudiantes de pregrado de la Universidad de Nariño y de la Institución universitaria CESMAG de la ciudad de Pasto (Colombia). La investigación realizada fue de tipo descriptivo bajo el enfoque cuantitativo, aplicando un diseño no experimental. Para ello se tuvo en cuenta la información de los estudiantes que ingresaron a estas universidades en las cohortes 2004, 2005 y 2006 con el fin de hacerles un seguimiento completo hasta el año 2011, construyendo un repositorio de datos para cada institución y un repositorio unificado utilizando el Sistema Gestor de Bases de Datos (SGBD) PostgreSQL. A estos repositorios se les aplicó las etapas de pre procesamiento y transformación con el fin de obtener conjuntos de datos limpios a los cuales se les aplicó técnicas de clasificación basada en árboles de decisión, asociación y clustering, utilizando la herramienta libre de minería de datos Weka. Finalmente, estos resultados fueron analizados, evaluados e interpretados para determinar la validez del conocimiento obtenido cuyo patrón general de deserción estudiantil fue 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.

Palabras Clave: Minería de datos, deserción, técnicas de clasificación, Weka.

1 Introducción

En Colombia, por lo menos el 52% de los estudiantes que comienzan una carrera universitaria no la concluyen [1]. Según el Ministerio de Educación Nacional (MEN), de las promociones de estudiantes que terminaron estudios entre 1999 y el 2004, 48% en promedio finalizaron sus estudios [2], es decir, de cada dos estudiantes que se matriculan en un programa de pregrado, solo uno culmina su carrera. La preocupación es mayor si se tiene en cuenta que el 39,52% de quienes abandonan sus estudios lo tienen que hacer por razones económicas [3].

Para el año 2005, en Colombia, la cobertura en la Educación Superior fue del 21,5% de la población escolar [4] y de esta, más de la mitad de los estudiantes matriculados

abandonan sus estudios sin obtener un título profesional, especialmente durante los primeros semestres. Por otra parte, una buena proporción son estudiantes egresados, pero sin título profesional [5].

De acuerdo con la Universidad Pedagógica Nacional [6], se entiende por deserción estudiantil, al hecho de que un número de estudiantes matriculados no siga la trayectoria normal del programa académico, bien sea por retirarse de ella, por repetir cursos o por retiros temporales. El MEN, la define como una situación a la que se enfrenta un estudiante cuando aspira y no logra concluir su proyecto educativo, considerándose como desertor a aquel individuo que al ser un estudiante de una institución de educación superior, no presenta actividad formativa durante dos semestres académicos consecutivos, lo cual equivale a un año de inactividad en su profesionalización. Esta última definición fue acogida como concepto base para esta investigación [7].

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

Tal como lo señalan Pautsch [8][9] en la Universidad Nacional de Misiones (Argentina) se realizó una investigación sobre deserción estudiantil utilizando técnicas de minería de datos. Su objetivo principal fue de 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, Química y Naturales analizando los datos de las cohortes entre los años 2000 al 2006.

De igual manera, según La Red et al.[10], en la Universidad Nacional del Nordeste (Argentina) se realizó un estudio cuyo objetivo principal fue aplicar técnicas de almacenes de datos y minería de datos basadas en clustering para la búsqueda de perfiles de los alumnos de la asignatura Sistemas Operativos de la Licenciatura en Sistemas de Información según su rendimiento académico, situación demográfica y socioeconómica, que permita conocer a priori situaciones potenciales de éxito o de fracaso académico.

Valero [11] y Valero, Salvador y García [12] señalan que en la universidad Tecnológica de Izúcar de Matamoros (México) se propuso una investigación para identificar las causas que motivan la deserción de sus estudiantes desde que ingresan. Mediante la técnica de minería de datos clasificación y la herramienta Weka, encontraron relaciones entre atributos académicos que identifican y predicen la probabilidad de deserción.

En el ámbito colombiano, de acuerdo con Restrepo y López [13], 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 estudiantes, con la técnica de minería de datos clasificación por Rough Sets utilizando el paquete ROSE2.

De igual manera, Pinzón [14] presenta la caracterización del perfil del estudiante desertor de la Escuela de Marketing y Publicidad de la Universidad Sergio Arboleda, utilizando la técnica de minería de datos agrupamiento con el algoritmo K-means. Para lo cual se analizaron las variables demográficas del alumno obtenidas en el registro de la última matrícula del mismo semestre de abandono y las causas que lo generaron.

Como resultado, se obtuvieron tres tipos de clúster que para el caso de la investigación, construyeron perfiles significativos.

En este artículo se describen las fases del proyecto de investigación cuyo objetivo general 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, utilizando técnicas de Minería de Datos que permitieron formular planes y programas enfocados a la detección temprana de los estudiantes que cumplan estos patrones.

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 Centro de Estudios Superiores María Goretti CESMAG (I.U.CESMAG) es una entidad Católica, de carácter privado, orientada por los principios franciscano-capuchinos y la filosofía personalizante y humanizadora de su fundador, padre Guillermo de Castellana. 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 artículo está organizado en secciones. Se describe a continuación la metodología del proceso de descubrimiento de conocimiento en bases de datos. En desarrollo investigativo se indica el desarrollo de cada fase metodológica. En la sección de resultados y discusión se muestran éstos y, también, se interpretan los patrones obtenidos en la etapa de minería de datos; finalmente, en la última sección, se presentan las conclusiones y trabajos futuros.

2 Metodología

Teniendo en cuenta las etapas del proceso de Descubrimiento de Conocimiento en Bases de Datos, inicialmente se seleccionaron de las bases de datos de estas dos Instituciones de Educación Superior IES, los datos socio-económicos, académicos, disciplinares e institucionales de los estudiantes que ingresaron a los diferentes programas de pregrado a partir del año 2004 hasta el 2011. Posteriormente, se seleccionaron únicamente la información de los estudiantes de las cohortes 2004, 2005 y 2006, con el fin de hacerles un seguimiento completo hasta el año 2011, determinando si desertaron o no.

Con estos datos se construyó un repositorio de datos para cada institución y un repositorio unificado utilizando el Sistema Gestor de Bases de Datos (SGBD) PostgreSQL. A estos repositorios se les aplicó las etapas de pre-procesamiento y transformación con el fin de obtener conjuntos de datos limpios y listos para aplicarles las técnicas de minería de datos. Estas técnicas fueron clasificación basada en árboles

de decisión, asociación y clustering, utilizando la herramienta libre de minería de datos Weka. Finalmente, estos resultados fueron analizados, evaluados e interpretados para determinar la validez del conocimiento obtenido.

2.1 Proceso de descubrimiento de conocimiento en bases de datos

El proceso de extraer conocimiento a partir de grandes volúmenes de datos ha sido reconocido por muchos investigadores como un tópico de investigación clave en los sistemas de bases de datos, y por muchas compañías industriales como una importante área y una oportunidad para obtener mayores ganancias [15]. Fayyad et al. lo definen como "El proceso no trivial de identificación de patrones válidos, novedosos, potencialmente útiles y fundamentalmente entendibles al usuario a partir de los datos" [16]. El Descubrimiento de Conocimiento en Bases de Datos (DCBD) es básicamente un proceso automático en el que se combinan descubrimiento y análisis. El proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice. Esta tarea implica generalmente preprocesar los datos, hacer minería de datos (data mining) y presentar resultados [17] [18] [19].

El proceso de DCBD es interactivo e iterativo, involucra numerosos pasos con la intervención del usuario en la toma de muchas decisiones y se resumen en cinco etapas:

Etapa de Selección. En la etapa de Selección, una vez identificado el conocimiento relevante y prioritario y definidas las metas del proceso DCBD, desde el punto de vista del usuario final, se crea un conjunto de datos objetivo, seleccionando todo el conjunto de datos o una muestra representativa de éste, sobre el cual se va a realizar el proceso de descubrimiento.

Etapa de Preprocesamiento/Limpieza. En la etapa de Preprocesamiento/Limpieza (Data Cleaning) se analiza la calidad de los datos, se aplican operaciones básicas como la remoción de datos ruidosos, se seleccionan estrategias para el manejo de datos desconocidos (missing y empty), datos nulos, datos duplicados y técnicas estadísticas para su reemplazo.

Etapa de Transformación/Reducción. En la etapa de transformación/reducción de datos, se buscan características útiles para representar los datos dependiendo de la meta del proceso. Se utilizan métodos de reducción de dimensiones o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos [16].

Los métodos de reducción de dimensiones pueden simplificar una tabla de una base de datos horizontalmente o verticalmente. La reducción horizontal implica la eliminación de tuplas idénticas como producto de la sustitución del valor de un atributo por otro de alto nivel, en una jerarquía definida de valores categóricos o por la discretización de valores continuos. La reducción vertical implica la eliminación de atributos que son insignificantes o redundantes con respecto al problema. Se utilizan técnicas de reducción tales como agregaciones, compresión de datos, histogramas, segmentación, discretización basada en entropía, muestreo, etc [19].

Etapa de minería de datos. La minería de datos es la etapa más importante del proceso DCBD [20]. El objetivo de esta etapa es la búsqueda, extracción y descubrimiento de patrones insospechados y de interés. La minería de datos consta de diferentes tareas, cada una de las cuales puede considerarse como un tipo de problema a ser resuelto por un algoritmo de minería de datos, afirman Adamo y Hernández, et al. donde las principales tareas son Clasificación, Asociación y Clustering [21][22].

Etapa de Interpretación/Evaluación de Datos. En la etapa de interpretación/evaluación, se interpretan los patrones descubiertos y posiblemente se retorna a las anteriores etapas para posteriores iteraciones. 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. Por otra parte, se consolida el conocimiento descubierto para incorporarlo en otro sistema para posteriores acciones, o simplemente para documentarlo y reportarlo a las partes interesadas, así como también para verificar y resolver conflictos potenciales con el conocimiento previamente descubierto.

2.2 Tareas de minería de datos

Dentro de la minería de datos se encuentran diferentes tipos de tareas, cada una de las cuales puede considerarse como un tipo de problema a ser resuelto por un algoritmo de minería de datos [22]. Entre las tareas de minería de datos más importantes están:

Clasificación. La clasificación de datos es el proceso por medio del cual se encuentra 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 [17].

Este proceso se realiza en dos pasos: en el primer paso se construye un modelo en el cual, cada tupla, de un conjunto de tuplas de la base de datos, tiene una clase conocida (etiqueta), determinada por uno de los atributos de la base de datos, llamado atributo clase. A cada tupla de este conjunto se denomina ejemplo de entrenamiento [19]. En el segundo paso, se usa el modelo para clasificar. Inicialmente, se estima la exactitud del modelo utilizando otro conjunto de tuplas de la base de datos, cuya clase es conocida, denominado conjunto de prueba. Este conjunto es escogido randómicamente y es independiente del conjunto de entrenamiento. A cada tupla de este conjunto se denomina ejemplo de prueba [19].

Se han propuesto varios métodos de clasificación: rough sets, árboles de decisión, redes neuronales, redes bayesianas, algoritmos genéticos entre otros. El modelo de clasificación basado en árboles de decisión, es probablemente el más utilizado y popular por su simplicidad y facilidad para entender [19] [23].

Segmentación o Clustering. El proceso de agrupar objetos físicos o abstractos en clases de objetos similares se llama segmentación o clustering o clasificación no supervisada [18]. Básicamente, el clustering 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.

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. Por heterogeneidad se entiende que los registros en diferentes segmentos no son similares de acuerdo a una medida de similaridad [25].

El algoritmo de clustering, segmenta una base de datos sin ninguna indicación por parte del usuario sobre el tipo de clusters que va a encontrar en la base de datos, por esta razón, se le denomina al método de segmentación o clustering, aprendizaje no supervisado. Algunos de los algoritmos utilizados para clustering son: CLARANS (Clustering Large Applications based upon RANdomized Search) [26] y BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) [27].

Asociación. La tarea de Asociación descubre patrones en forma de reglas, que muestran los hechos que ocurren frecuentemente juntos en un conjunto de datos determinado. El problema fue formulado por Agrawal et al. [28] y a menudo se referencia como el problema de canasta de mercado (market-basket). En este problema, se da un conjunto de ítems y una colección de transacciones que son subconjuntos (canastas) de estos ítems. La tarea es encontrar relaciones entre los ítems de esas canastas para descubrir reglas de asociación que cumplan unas especificaciones mínimas dadas por el usuario, expresadas en forma de soporte y confianza.

Patrones secuenciales. Los patrones secuenciales buscan ocurrencias cronológicas. El problema de descubrimiento de patrones secuenciales se trata en [28]. Se aplica principalmente en el análisis de la canasta de mercado y su objetivo es descubrir en los clientes ciertos comportamientos de compra en el tiempo. El dato de entrada es un conjunto de secuencias, llamado data-secuencia. Cada data-secuencia es una lista de transacciones, donde cada transacción es un conjunto de ítems (literales). Típicamente, hay un tiempo asociado con cada transacción.

3 Desarrollo investigativo

3.1 Proceso de descubrimiento de conocimiento en bases de datos

Etapa de Selección. 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.

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, compuesta por 6870 registros y 62 atributos. Los datos de los estudiantes de la IUCESMAG fueron almacenados en la base de datos REPOSITORIOIUCESMAG, 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.

Etapa de Preprocesamiento/Limpieza. 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.

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 modao 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, por la calidad de los datos y por su importancia para el estudio, 31 atributos y con estos se creó la tabla de atributos generales. De estos 31 atributos, se escogieron 18 para analizar el factor socioeconómico y 14 para el factor académico. De igual manera, en la IUCESMAG se escogieron 28 atributos que formaron la tabla de atributos generales y de estos, 17 para el análisis socioeconómico y 12 para la parte académica del estudiante.

Entre los atributos conjuntos entre las dos universidades para factor socio académico se tomó: género, estado civil, zona de nacimiento, zona de procedencia, régimen de salud, estrato, padre, ocupación padre, madre, ocupación madre, tipo de residencia, vive con familia, hermanos en universidad, ingresos familiares, valor de la matrícula y edad.

Para el factor académico: tipo de colegio, jornada, ICFES promedio, ICFES total, facultad, programa académico, promedio de notas, materias perdidas, semestre perdido y área por materia. 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. La descripción de estas tablas se muestra en la tabla 1.

Tabla 1. Tabla de repositorios

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.
C1054A28	Tabla que contiene 1054 estudiantes de las cohortes que ingresaron en 2004-2006 y los 28 atributos a considerar en el estudio.

Etapa de Transformación/Reducción. 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. Como resultado se obtuvo el conjunto de datos, 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. Los atributos de la tabla U2136A28 se muestra en la tabla 2. 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 2. Atributos repositorio unificado U2136A28

N	Atributo	N	Atributo	N	Atributo
1	Sexo	11	Ocupación_madre	21	Facultad
2	Edad_ingreso	12	Hermanos_u	22	Area_programa
3	Estrato	13	Tipo_residencia	23	Promedio_notas
4	Estado_civil	14	Vive_con_flia	24	Materias_perdidas
5	Régimen_salud	15	Ingresos-flia	25	Semestre_perdidas
6	Zona_nacimiento	16	valor_matricula	26	Area_materia
7	Zona_procedencia	17	Tipo_colegio	27	Veces_perdida
8	Padre	18	Jornada_colegio	28	Deserción
9	Ocupación_padre	19	Icfes_promedio		
10	Madre	20	Icfes_total		

Etapa de minería de datos. Las tareas de minería de datos utilizadas en esta investigación fueron clasificación, asociación y clustering.

La técnica de clasificación utilizada fue arboles de decisión. 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 [29] con una confianza mínima de 75%.

Para la tarea de Asociación se utilizó el algoritmo Apriori [10], implementado en Weka en el paquete WEKA.associations.Apriori.

Para la tarea de agrupación se utilizó la técnica particional con el algoritmo Kmeans [26], 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.

Los resultados más relevantes de estas dos tareas de minería de datos se muestran en la sección de resultados.

Etapa de Interpretación/Evaluación de Datos. Para evaluar la calidad del modelo de clasificación por árboles de decisión, dividiendo el repositorio de datos en dos conjuntos: entrenamiento y prueba, se escogió el método validación cruzada con n pliegues *(n-fold cross validation*) debido a su rendimiento computacional [32]. 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 [22].

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 [22].

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 [30]. 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 [31].

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 [22].

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.

Los resultados de esta etapa se analizan en la siguiente sección.

4 Resultados y discusión

4.1 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 3.

Tabla 3. Reglas de clasificación

Antecedente Regla	Conse-Cuente (deserta)	% Soporte	% Confian-za	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
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_procedenci=SUR&vive_con_familia=S	S	6,26	79,8	134
ingresos_familiares = De 5980000 a	S	2,32	78,9	50
8854000 ingresos_familiares = Mayor a 8854000	S	4,73	77,3	101

Como se puede observar en la tabla 3, 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.

4.2 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 4 donde 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.

Tabla 4. Reglas de asociación

Antecedente Regla	**Consecuente (deserta)**	**% Soporte**	**% Confianza**
estado_civil=SOLTERO & promedio_nota=Menor a 2.4 & semestre_perdidas=P & veces_perdida=1	S	16,1	100
genero=M & estado_civil=SOLTERO & promedio_nota=Menor a 2.4 &veces_perdida= 1	S	12,3	100
genero=M & promedio_nota=Menor a 2.4 & semestre_perdidas=P & veces_perdida=1	S	12,2	100
tipo_colegio=PUBLICO & promedio_nota=Menor a 2.4 & semestre_perdidas=P	S	11,3	100
estado_civil=SOLTERO & colegio=PUBLICO & promedio_nota=Menor a 2.4 & perdida= 1	S	11,1	100
estado_civil=SOLTERO& promedio_nota=Menor a 2.4 & semestre_perdidas=P& universidad=PRIVADA	S	10,7	100
estado_civil=SOLTERO & promedio_nota =Menor a 2.4& perdida=1 & universidad=PRIVADA	S	10,3	100
promedio_nota=Menor a 2.4 & semestre_perdidas=P & perdida= 1 & universidad=PRIVADA	S	10,1	100

Antecedente Regla	Conse-cuente (deserta)	% Sopor-te	% Con-fianza
zona_nacimiento=PASTO & promedio_nota= Menor a 2.4 & semestre_perdidas=P& perdida=1	S	10,1	100
estado_civil=SOLTERO & zona_nacimiento=PASTO & promedio_nota=Menor a 2.4 & perdida=1	S	10,0	100
estado_civil=SOLTERO & promedio_nota=De 2.4 a 3.1 & semestre_perdidas=P & perdida=1	S	9,0	100
genero=M & estado_civil=SOLTERO & promedio_nota=Menor a 2.4 & 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

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 las dos universidades entre los años 2004 y 2006 cumplen 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 las dos universidades entre los años 2004 y 2006 cumplen con este patrón.

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.

4.3 Agrupación

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 5 se muestran estos dos grupos.

Los resultados que se muestran en la tabla 5 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.

Tabla 5. 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
Tipo_universidad	Publica	Publica	Privada
Deserción	S	N	S

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.

5 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 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

Este trabajo fue financiado por la Universidad de Nariño y la Institución Universitaria CESMAG y desarrollado por investigadores del grupo de investigación GRIAS y TECNOFILIA respectivamente.

Referencias

1. Ministerio de Educación Nacional, MEN: Deserción estudiantil en la educación superior colombiana: metodología de seguimiento, diagnóstico y elementos para su pevención. Men. 68 a 73 (2006).
2. Ibid.
3. Ministerio de Educación Nacional, MEN: Diagnóstico de la Deserción Estudiantil en Colombia: Educación Superior. Bol. Inf. 7. 3 a 5 (2006)
4. Ministerio de Educación Nacional: Educación Superior en Colombia: MEN, Bogotá (2007)
5. Rojas B. M., González D. C.: Deserción estudiantil en la Universidad de Ibagué: Zona Próxima, 9, 70 a 83 (2008)
6. Universidad Pedagógica Nacional, UPN: La deserción estudiantil: reto investigativo y estratégico asumido de forma integral por la UPN (2004)
7. Ministerio de Educación Nacional. Bogotá (Colombia): MEN. (2009)
8. Pautsch, J.: Minería de datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación: Tesis de grado (Licenciado en Sistemas de Información). Posadas, Misiones (Argentina): Universidad Nacional de Misiones. 193, (2009)
9. Pautsch, J., La Red, D., Cutro, L. (2010). Minería de datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación, http://www.dataprix.com/files/Analisis%20de%20Desercion%20Univ_0.pdf
10. La Red, D., Acosta, J., Cutro, L., Uribe, V., Rambo, A.: Data Warehouse y Data Mining Aplicados al Estudio del Rendimiento Académico. En: Novena Conferencia

Iberoamericana en Sistemas, Cibernética e Informática, CISCI 2010, Orlando (Florida, EE.UU.), 1, 289 a 294 (2010)
11. Valero, S.: Aplicación de técnicas de minería de datos para predecir la deserción. En: Universidad Tecnológica de Izúcar de Matamoros. http://www.utim.edu.mx/~svalero/docs/MineriaDesercion.pdf
12. 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: Universidad Tecnológica de Izúcar de Matamoros. http://www.utim.edu.mx/~svalero/docs/e1.pdf
13. Restrepo, M., López, A.: Uso de la metodología Rough Sets en un modelo de deserción académica. En: XIV Congreso Ibero Latinoamericano de Investigación de Operaciones, CLAIO, Cartagena (Colombia): Universidad del Norte. Libro de Memorias, 108 a 109 (2008)
14. Pinzón, L.: Aplicando minería de datos al marketing educativo. En: Revista Notas de Marketing. 1, 45 a 61 (2011)
15. Timarán Pereira, R.: Una mirada al descubrimiento de conocimiento en bases de datos: Ventana Informática. 20, 39 a 58 (2009)
16. Fayyad, U., Piatetsky Shapiro, G., and Smyth, P.: The KDD process for extracting useful knowledge from volumes of data: Comunications of the ACM. 39, 11 , 27 a 34. (1996)
17. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. En: 20th International Conference on Very Large Data Bases, VLDB 1994, Santiago de Chile. pp. 487 a 499, (1994)
18. Chen, M., Han, J., YU, P.: Data mining: An overview from database perspective. In: IEEE Transactions on Knowledge and Data Engineering. Vol. 8, No. 6 . Los Alamitos (CA, USA): (1996).
19. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, Academic Press. San Francisco (2001)
20. Imielinski, T., Manila, H.: A database perspective on knowlegde discovery communications: Association for Computing Machinery, 39, 11 (1996).
21. Adamo, J. M: Data Mining for Association Rules and Sequential Patterns: Sequential and Parallel Algorithms. New York (2001)
22. Hernández, O. J., Ramírez, Q. M. y Ferri, R.C.: Introducción a la Minería de Datos. Pearson Prentice Hall, Madrid (2005)
23. Sattler, K., Dunemann, O.: SQL Database Primitives for Decision Tree Classifiers. In: The 10th ACM International Conference on Information and Knowledge Management - CIKM, Atlanta, Georgia (USA): ACM. Proceedings, pp. 379 a 386. (2001)
24. 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 (Wales, U.K.): IEEE Computer Society. Proceedings, pp. 58 a 67 (1998)
25. Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., Zanasi, A.: Discovering Data Mining from Concept to Implementation, Prentice Hall PTR (1997)
26. Ng, R., Han, J.: Efficient and Effective Clustering Method for Spatial Data Mining, VLDB Conference, Santiago de Chile (1994)
27. Zhang, T., Ramakrishnan, R., Livny, M.: BIRCH: An Efficient Data Clustering Method for Very Large Databases, ACM SI (1996)

28. Agrawal, R., Srikant R.: Mining Sequential Patterns. En: Proceedings of the 11th International Conference on Data Engineering. (1995)
29. Quinlan, J. R.: Programs for Machine Learning. San Francisco: Morgan Kaufmann Publishers. (1993)
30. García, M., Álvarez, A.: Análisis de Datos en WEKA –Pruebas de Selectividad, http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf
31. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. p. 365, San Francisco (2000).
32. Arlot, S., Celisse, A.: A survey of cross-validation procedures for model selection. In: Statistics Surveys, 4, 40 a 79 (2010)