SECCIÓN OTROS
REVISTA CENTRO DE ESTUDIOS EN SALUD
Año 9 - VOL 1 - Nº 11 - 2009

LA MINERÍA DE DATOS EN EL DESCUBRIMIENTO DE PERFILES DE DESERCIÓN ESTUDIANTIL EN LA UNIVERSIDAD DE NARIÑO

Ricardo Timarán Pereira[1]

Fecha recepción: Junio - 24/09 Fecha aceptación: Noviembre - 27 / 09

RESUMEN

La deserción estudiantil en los programas de pregrado de la gran mayoría de Instituciones de Educación Superior tanto de Colombia como de Latinoamérica es un problema que tiene un impacto multidimensional en el desarrollo social y económico de un país. A pesar de esto, son muy pocos los estudios que se han realizado en la Universidad de Nariño con respecto a este problema que permitan aplicar estrategias efectivas que ayuden a minimizar este fenómeno y conlleven al mejoramiento de la calidad educativa en la universidad. A través de técnicas de minería de datos es posible predecir cuando un estudiante va a abandonar sus estudios, aplicándolas sobre los datos almacenados en las bases de datos de una institución educativa, con el fin de tomar acciones anticipadas que le permitan disminuir este factor. En este artículo se describe el proceso de descubrimiento de conocimiento que se llevó a cabo en la Universidad de Nariño para determinar en la comunidad universitaria perfiles de bajo rendimiento académico y deserción estudiantil utilizando la base de datos histórica de los estudiantes de pregrado. Este proceso se apoyó con TariyKDD, una herramienta de minería de datos de distribución libre, desarrollada en los laboratorios KDD del Grupo de investigación GRIAS del Departamento de Sistemas de la Facultad de Ingeniería

Palabras clave: Deserción estudiantil; descubrimiento de patrones; minería de datos.

1. Doctor en Ingeniería énfasis Ciencias de la Computación. Director Grupo de Investigación GRIAS. Profesor Asociado Departamento de Sistemas. Facultad de Ingeniería. Universidad de Nariño. San Juan de Pasto, Colombia, ritimar@udenar.edu.co

ABSTRACT

The student desertion in undergraduate programs of most institutions of higher education in both Colombia and Latin America is a problem that has a multidimensional impact on the social and economic development of a country. Despite this, few studies have been conducted at the Nariño University regarding this problem to implement effective strategies to help minimize this phenomenon and lead to improving the quality of education at the university.Through data mining techniques it is possible to predict when a student is going to abandon their studies, applying the data stored in the databases of an educational institution, to take early action to enable it to reduce this factor. This paper describes the knowledge discovery process that is held at the Nariño University in determining the university community profiles low academic performance and student desertion using the historical database of undergraduate students. This process is supported by TariyKDD, a data mining tool for free distribution, developed in the KDD laboratories of Research Group GRIAS of the Department of Systems of Engineering Faculty.

Key words: Student desertion; patterns discovery; data mining

INTRODUCCIÓN

En Colombia, como en Latinoamérica, la educación superior presenta altas tasas de deserción estudiantil, especialmente en los primeros semestres académicos y sin embargo, no se le presta la suficiente atención a este problema, ni tampoco hay políticas formales para enfrentarlo.

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 el o por demorar más tiempo del previsto en finalizar, por repetir cursos o por retiros temporales.[1] La disolución del vínculo que se estipula a través de la matrícula académica, ya sea por parte del estudiante o de la universidad, tiene efectos de tipo financiero, académico y social que implican la pérdida de esfuerzos y recursos en un país como Colombia en donde más de la mitad de los estudiantes que comienzan una carrera universitaria no la concluyen.[2]

La Universidad de Nariño es una institución pública de educación superior con sede principal en la ciudad de San Juan de Pasto, capital del departamento de Nariño y cuya área de influencia es el suroccidente de Colombia. En ella se encuentra la mayoría de estudiantes universitarios de la región. Los estudiantes de educación secundaria aspiran obtener un cupo en ésta, por su calidad educativa, y prestigio de sus egresados. Desafortunadamente, en algunos casos, cuando el estudiante se matricula a un determinado programa, su rendimiento no es el esperado, generando índices de deserción altos y bajo rendimiento académico. Por lo tanto se genera un interrogante acerca de cuáles son las causas que motivan la deserción y/o el bajo rendimiento y qué perfiles tiene este tipo de estudiantes.

De acuerdo a datos obtenidos, en el año 2006 estaban matriculados en los diferentes programas de pregrado 8.136 estudiantes distribuidos en 11 facultades (ver tabla 1), de los cuales el 27.94% ha perdido dos veces la misma asignatura, el 3.44% ha perdido tres veces y el 0.19% cuatro veces. Así mismo el 6.71% tiene un promedio menor a 3.0, el 66.85% tiene un promedio entre 3.0 y 4.0, y el 26.22% tiene un promedio mayor que 4.0 sobre 5.0. Además el promedio del índice de deserción está por encima del 32% en la última corte de los programas.

A pesar de esto, son muy pocos los estudios que se han realizado en la Universidad de Nariño con respecto a este problema que permitan aplicar

estrategias efectivas que ayuden a minimizar este fenómeno y conlleven al mejoramiento de la calidad educativa en la universidad.

Debido al avance de la tecnología en los sistemas computacionales, se hace indispensable y necesaria la utilización de tecnologías informáticas que contribuyan a resolver ciertos problemas que sin su utilización, haría prácticamente imposible el tratamiento de los mismos, brindando soluciones eficientes y sustentadas en la realidad para aplicarlas en el contexto en el que se encuentran. Una de estas tecnologías es la minería de datos, en la que se fundamentó todo el proceso investigativo de este estudio.

A través de técnicas de minería de datos aplicadas a los datos históricos almacenados en las bases de datos de una IES, es posible predecir las características del estudiante que va a abandonar sus estudios o predecir quiénes están propensos a desertar. El determinar perfiles de bajo rendimiento permite predecir qué estudiantes están a punto de desertar. El determinar perfiles de deserción permite predecir qué estudiantes son los que desertan. Teniendo estos perfiles, la IES puede tomar acciones anticipadas que le permitan disminuir este factor.

La minería de datos es 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 y presentar resultados.[3, 4, 5, 6] Este proceso es interactivo e iterativo, involucra numerosos pasos con la intervención del usuario en la toma de muchas decisiones. En la figura 1 se muestran las etapas del proceso de minería de datos.

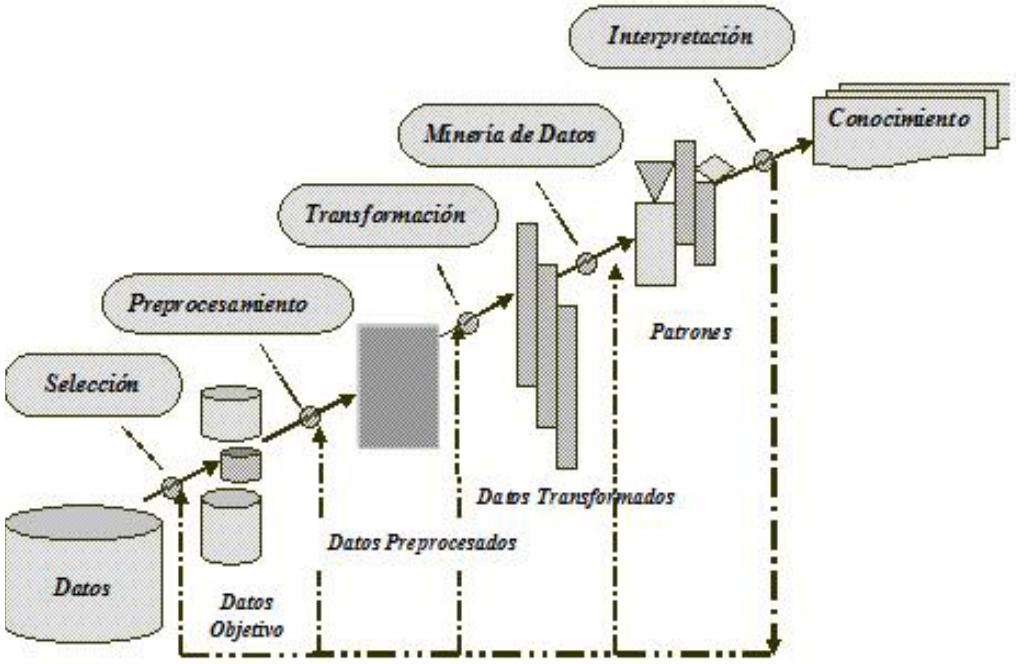
Tabla 1. Número de estudiantes por facultad

Facultades	Nº. Estudiantes
Artes	902
Ciencias Agrícolas	618
Ciencias de la Salud	243
Ciencias Económicas y Administrativas	1408
Ciencias Humanas	1341
Ciencias Naturales y Matemáticas	687
Ciencias Pecuarias	523
Derecho	510
Educación	435
Ingeniería	1188
Ingeniería Agroindustrial	281
Total	**8136**

En este artículo se describe el proceso de minería de datos que se llevó a cabo en la Universidad de Nariño para determinar en la comunidad universitaria perfiles de bajo rendimiento académico y deserción estudiantil utilizando la base de datos histórica de los estudiantes de pregrado, compuesta por información personal y académica de 46.173 estudiantes. Este proceso se apoyó con TariyKDD, una herramienta de minería de datos de distribución libre, desarrollada en los laboratorios KDD del Grupo de investigación GRIAS del Departamento de Sistemas de la Facultad de Ingeniería.[7]

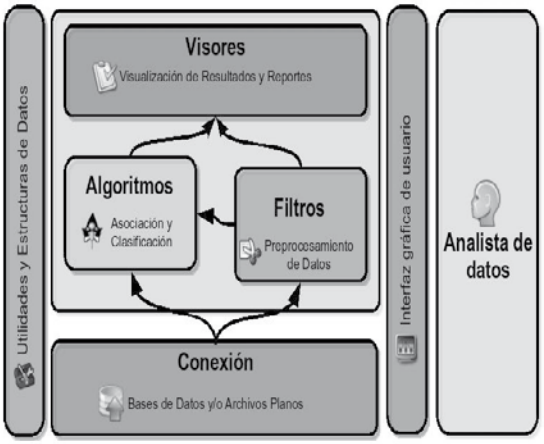
La arquitectura de TariyKDD está compuesta por cuatro módulos: el módulo de conexión que permite la recuperación de datos desde archivos planos y bases de datos relacionales, el módulo de utilidades con clases y librerías comunes, el módulo kernel donde se encuentran los filtros que permiten realizar los procesos de limpieza y transformación de datos, los algoritmos de minería de datos para las tareas de Asociación y Clasificación y los programas de visualización de datos, y el módulo de interfaz gráfica de usuario que facilita la interacción del usuario con la herramienta de una manera amigable.

Figura 1. Etapas del proceso de minería de datos



En TariyKDD se encuentran implementados los algoritmos de minería de datos: Apriori,[8] FPGrowth.[9] y EquipAsso,[10, 11]para la tarea de Asociación y los algoritmos C4.5[12] y Mate-tree [13] para la tarea de Clasificación. La arquitectura de TariyKDD se muestra en la figura 2.

Figura 2. Arquitectura de la herramienta TariyKDD



MATERIALES Y MÉTODOS

Con el objeto de extraer conocimiento a partir de los datos almacenados en las bases de datos de la Universidad de Nariño y determinar patrones de deserción estudiantil se cumplió con cada una de las etapas del proceso de minería de datos:

Etapa de Selección

El objetivo de esta etapa es obtener las fuentes de datos internas y externas que sirven de base para el proceso de minería de datos. Como fuente interna, se seleccionó la base de datos histórica de los estudiantes de la Universidad de Nariño, compuesta por información personal y académica de 46.173 estudiantes. Como fuente externa, se seleccionó la información de los colegios de educación secundaria del país, que se obtuvo con el Ministerio de Educación Nacional de Colombia. Estas fuentes de datos se integraron en la base de datos UDENARDB, construida con el sistema gestor de base de datos PostgreSQL. UDENARDB la componen siete tablas, cuyas descripciones se pueden ver en la tabla 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 ad-hoc sobre la base de datos UDENARDB, se analizó minuciosamente la calidad de los datos contenidos en cada uno de los atributos de las tablas y se seleccionaron los atributos más relevantes para la investigación y aquellos que no contenían valores nulos. Como resultado de esta etapa, quedaron únicamente datos de 20.329 estudiantes, para su posterior análisis. De la tabla alumnos se seleccionaron 19 atributos, de la tabla carreras 4, de la tabla facultades 2, de la tabla materias 3, de la tabla notas 8, de la tabla liquidación 12 y de la tabla colegios 3 atributos. Los atributos seleccionados de las diferentes tablas en su gran mayoría no contenían valores nulos ni anómalos (*outliers*), pero en aquellos casos que se presentaban, estos fueron reemplazados utilizando técnicas estadísticas tales como la media y la moda o derivando sus valores a través de otros como por ejemplo la edad de ingreso del estudiante conocida la fecha de ingreso y la fecha de nacimiento.

Tabla 2. Descripción de tablas de la base de datos UDENARDB

Tablas	Nº. Atributos	Descripción
ALUMNOS	69	Se encuentran todos los datos personales del estudiante.
CARRERAS	10	Se encuentra información de todas las carreras existentes en la Universidad de Nariño
FACULTADES	4	Contiene información de las facultades de la Universidad de Nariño.
MATERIAS	4	Se encuentra toda la información de las materias existentes en el plan académico de cada carrera.
NOTAS	8	Contiene información de las notas por materia de cada estudiante.
LIQUIDACIÓN	27	Se encuentra toda la información financiera del estudiante
COLEGIOS	7	Contiene información de los colegios del país

Etapa de transformación de datos

En la etapa de transformación, se buscan características útiles para representar los datos dependiendo de la meta del proceso de minería de datos. 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. [14]

En esta etapa se construyó el conjunto de datos UDENAR.DAT, integrando los atributos de las diferentes tablas de la base de datos UDENARDB. Se eliminaron los atributos que eran llaves primarias de las tablas, se construyeron nuevos atributos (ver tabla 3) y se discretizaron los atributos continuos, es decir, se transformaron los valores numéricos en atributos discretos o nominales. Algunos de los atributos discretizados se muestran en las tablas 4,5 y 6.

Tabla 3. Descripción de nuevos atributos del conjunto de datos UDENAR.DAT

Atributo	Descripción
Ingresos	Establece un valor real actualizado de ingresos familiares del estudiante. Para ello relaciona los campos ingresos_familiares de la tabla alumnos e ingresos de la tabla liquidación.
Edad	Determina qué edad tiene actualmente el estudiante; para ello se relacionaron los campos fecha nacimiento de la tabla alumnos y la fecha actual.
edad_ing	Establece la edad en la que ingresó el estudiante. Para crearlo se relacionó el campo fecha de ingreso y la fecha de nacimiento del estudiante.
val_matricula	Determina el valor real que paga el estudiante por concepto de matricula financiera; relaciona los valores de los campos nueva_matricula y de nuevos servicios de la tabla liquidación.
Claseal	Determina qué estudiantes han reingresado, se han retirado o no cumplen con ninguna de las condiciones anteriores.
Claserend	Determina la cantidad de materias perdidas por el estudiante.
Clasepromedio	Determina el promedio acumulado del estudiante.

Por otra parte, el conjunto de datos UDENAR.DAT se adecuó al formato ARFF (*Attribute Relation File Format*), utilizado por la herramienta TariyKDD para importar los datos. La estructura del formato ARFF [15] es la siguiente:

- Cabecera: se define el nombre de la relación y su formato es el siguiente: relation <nombre-de-la-relación>
- Declaraciones de los atributos. En esta sección se declaran los atributos que compondrán el archivo arff con su tipo. La sintaxis es la siguiente: @attribute <nombre-del-atributo> <tipo>
- Sección de datos. Se declaran los datos que componen la relación separando entre comas los atributos y con salto de líneas las relaciones.

Tabla 4. Discretización del atributo Edad

Edad	*Valor*	No. Registros
Menores e iguales a 18	A	827
Mayores de 18 y menores que 22	B	3634
Mayores e iguales que 22 y Menores de 26	C	4856
Mayores e iguales que 26	D	11012

Tabla 5. Discretización del atributo Fecha de Ingreso

Fecha Ingreso	Valor	Nº Registros
Antes de 1990	A	1022
Después o igual a 1990 a Menores de 1995	B	4852
Después o igual a 1995 a Menores de 2000	C	5978
Después o igual al 2000 y Menores de 2003	D	5046
Mayores o iguales de 2003	E	3431

Finalmente, se obtuvo el conjunto de datos UDENAR.ARFF con 26 atributos y 20.329 registros, listo para aplicarle las técnicas de minería de datos, utilizando la herramienta TariyKDD, que permita obtener los patrones de bajo rendimiento académico y /o deserción de los estudiantes de la Universidad de Nariño.

Tabla 6. Discretización del atributo Clasepromedio

Clasepromedio	Valor	Nº Registros
Menor a 2	A	2391
Mayor o igual a 2 hasta 3	B	2934
Mayor o igual a 3 hasta 3.5	C	5166
Mayor o igual a 3,5 hasta 4,0	D	6850
Mayor o igual a 4.0 hasta 5.0	E	2988

Etapa de minería de datos

El objetivo de esta etapa es la búsqueda y descubrimiento de patrones insospechados y de interés utilizando diferentes técnicas de descubrimiento tales como clasificación, clustering, patrones secuenciales, asociación entre otras. Para el descubrimiento de patrones de bajo rendimiento académico y deserción estudiantil se utilizaron las tareas de Clasificación y Asociación. Para generar las reglas de clasificación su utilizó el algoritmo C4.5 y para las reglas de Asociación, el algoritmo EquipAsso, disponibles en la herramienta TariyKDD. Los patrones descubiertos se describen en la sección de resultados.[16]

Etapa de interpretación y evaluación de resultados

En esta etapa se interpretan los patro**nes** descubiertos y posiblemente se retorna a los anteriores pasos o 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. Los resultados de esta etapa se analizan en la siguiente sección.

RESULTADOS

Para predecir los perfiles de bajo rendimiento académico, el conjunto de datos UDENAR. ARFF se clasificó escogiendo como clase el atributo *Clasepromedio*. Este atributo indica el rendimiento académico del estudiante basado en el promedio acumulado de las notas hasta el semestre cursado.

Entre las reglas de clasificación más representativas están:

Si el estrato socioeconómico es 2, el ponderado de exámenes de estado ICFES está entre 50 y 70, es del Sur de Nariño, está en primer semestre y pertenece a la facultad de Ciencias Humanas, entonces su rendimiento es Bajo. El 68% con estas características se clasifica de esta manera.

Si la edad de ingreso es menor o igual a 18 años, el estrato socioeconómico es 2, género masculino, el ponderado ICFES está entre 50 y 70, vive con la familia, es del Sur de Nariño, está en primer semestre, está en la facultad de Ciencias Naturales

y Matemáticas, entonces su rendimiento es Bajo. El 67% con estas características se clasifica de esta manera.

Si la edad de ingreso es menor o igual a 18 años, proviene de un colegio privado, el calendario del colegio es septiembre a junio, género femenino, es del Sur de Nariño, está en primer semestre y pertenece a la facultad de Ciencias Naturales y Matemáticas, entonces su rendimiento es Bajo. El 70% con estas características se clasifica de esta manera.

Para predecir los perfiles de deserción estudiantil se escogió como clase el atributo *Clase_al*. Este atributo indica si el estudiante no se ha retirado, ha reingresado o se retiró definitivamente de la Universidad. Entre las reglas de clasificación más representativas están:

Más del 50% de los estudiantes retirados que pertenecen a la facultad de ingeniería, reingresan.

Los estudiantes retirados que pertenecen a las facultades de Ciencias Naturales y Matemáticas y Ciencias Humanas no reingresan.

Entre las reglas de Asociación más representativas, que permiten identificar relaciones no explícitas entre los atributos del conjunto de datos UDENAR.ARFF que involucran bajo rendimiento y deserción están:

El 95% de los estudiantes que tiene promedio bajo está en primer semestre. El 10% de todos los estudiantes, son de primer semestre y tienen promedio bajo.

El 84% de los estudiantes retirados son de estrato socioeconómico 2 y provienen de municipios del Sur de Nariño. El 2.5% de todos los estudiantes, se han retirado, son de estrato 2 y provienen del Sur de Nariño.

El 89 % de los estudiantes retirados son de primer semestre, tienen un ponderado ICFES entre 50 y 70 y proceden del Sur de Nariño. El 2.5% de todos los estudiantes, se han retirado, son de primer semestre, tienen un ponderado ICFES entre 50 y 70 y provienen del Sur de Nariño.

El 88% de los estudiantes retirados, tienen una edad de ingreso menor que 18 años provienen del Sur de Nariño. El 2.5% de todos los estudiantes, se ha retirado, tiene una edad de ingreso menor que 18 años y es del Sur de Nariño.

El 86% de estudiantes retirados terminó su bachillerato en colegios públicos, es de primer semestre y proviene del Sur de Nariño. El 2.5% de todos los estudiantes, se ha retirado, terminó su bachillerato en colegios públicos, es de primer semestre y proviene del Sur de Nariño.

De acuerdo a los resultados obtenidos, la mayoría de los estudiantes de primer semestre, provenientes de la zona sur del departamento de Nariño, de estratos socioeconómicos bajos y matriculados en algún programa de la facultad de Ciencias Naturales y Matemáticas o en la facultad de Ciencias Humanas, presenta un bajo rendimiento académico. Este perfil es similar al perfil de de la mayoría de estudiantes que se retiran. Por otra parte la mayoría de estudiantes que se retiran de estas dos facultades no reingresan, lo que no sucede en la facultad de Ingeniería, donde casi la mayoría de estudiantes retirados reingresan.

CONCLUSIONES Y RECOMENDACIONES

Se han presentado los resultados del primer proyecto de investigación realizado en la Universidad de Nariño, aplicando técnicas de minería de datos para determinar perfiles de bajo rendimiento y deserción estudiantil en sus programas de pregrado.

Dentro de este proyecto, las fases de preprocesamiento y transformación de datos fueron las más costosas en tiempo, debido a la mala calidad de los datos de la base de datos de la población estudiantil utilizada en esta investigación. Se encontraron muchos datos nulos o faltantes y otros redundantes. Además, en el cambio de un semestre a otro se adicionan nuevos atributos y otros se abandonan, lo que corrompe los datos históricos de la base de datos. Esto significó que de 46.173 registros seleccionados, se analizaran solamente 20.329, lo que incidió negativamente en los resultados del estudio.

Es necesario que los patrones de bajo rendimiento y deserción obtenidos, se analicen detenidamente por las directivas de la Universidad de Nariño, con el fin de tomar decisiones y proponer estrategias conducentes a prevenir que estudiantes con estos perfiles deserten o caigan en bajo rendimiento. Específicamente, se debe hacer un seguimiento a los estudiantes de primer semestre que provienen de los municipios de la zona sur de Nariño, que ingresan a programas de las facultades de Ciencias Naturales y Matemáticas e Ingeniería para disminuir el alto grado de deserción. Por otra parte, se recomienda la creación de un Observatorio Académico, la construcción de una bodega de datos y el rediseño del sistema de información de Registro Académico que permita obtener datos de calidad que soporten futuros proyectos encaminados al fortalecimiento de la toma de decisiones con respecto al problema de deserción estudiantil en la Universidad de Nariño.

REFERENCIAS

1. Ministerio de Educación Nacional. La Deserción estudiantil: reto investigativo y estratégico asumido de forma integral por la UPN. [Consultado marzo 16 de 2009]. Disponible en: URL:http//www.mineducacion.gov.co/1621/articles-85600_Archivo_pdf3.pdf.
2. Rojas BM, González DC. Deserción estudiantil en la Universidad de Ibagué. Revista Zona Próxima No 9, Universidad del Norte, Colombia, 2008:70-83.
3. Chen M, Han J, Yu P. Data Mining: An Overview from Database Perspective. IEEE Transactions on Knowledge and Data Engineering; 1996.
4. Imielinski T, Mannila H. A database perspective on knowledge discovery. Communications of the ACM, Vol 39, No. 11, November; 1996.
5. Han J, Kamber M. Data Mining concepts and techniques. San Francisco (CA): Morgan Kaufmann Publishers; 2001.
6. Hernández OJ, Ramírez QM, Ferri RC. Introducción a la minería de datos. Madrid (España): Editorial Pearson Prentice Hall; 2004.
7. Timarán, PR, Calderón, RA, Ramírez, FI, Guevara, F, Alvarado, JC. TariyKDD una herramienta de minería de datos débilmente acoplada con un SGBD. En: Memorias de VII Jornadas Iberoamericanas de Ingeniería de Software e Ingeniería del Conocimiento, Editado por Escuela Superior del Litoral. Guayaquil, Ecuador, 2007: 3-11.
8. Agrawal R, Srikant R. Fast Algorithms for mining association rules. In: proceedings of VLDB Conference. Santiago, Chile, 1994
9. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In: proceedings of. ACM SIGMOD. Dallas (TX); 2000.
10. Timarán PR, Millán M. EquipAsso: un Algoritmo para el descubrimiento de reglas de asociación basado en operadores algebraicos. En: memorias de 4ª Conferencia Iberoamericana en Sistemas, Cibernética e Informática CISCI 2005. Orlando (Florida); 2005, 343-348
11. Timarán, PR, Millán M. EquipAsso: an Algorithm based on new relational algebraic operators for association rules discovery. In: proceedings of the Fourth IASTED International Conference on Computational Intelligence. ACTA Press, Calgary (Canada); 2005.
12. Quinlan JR. C4.5: Programs for machine learning. San Francisco (CA): Morgan Kaufmann Publishers; 1993.
13. Timarán, PR. Mate-tree: un algoritmo para el descubrimiento de reglas de clasificación basado en operadores algebraicos relacionales. En: memorias de 6ª Conferencia Iberoamericana en Sistemas, Cibernética e Informática CISCI 2007. Orlando (Florida); 2007, 196-201m
14. Fayyad U, Piatetsky-Shapiro, G, Smyth P. The KDD process for extracting useful knowledge from

volumes of data. Communications of the ACM, Vol. 39, No 11, November, 1996

15. Witten, IH, Frank, E. Data mining practical machine learning tools and techniques with Java implementations. San Francisco (CA): Morgan Kaufmann Publishers; 2000.
16. Timarán, PR. Detección de patrones de bajo rendimiento académico y deserción estudiantil con técnicas de minería de datos. En: memorias de 8ª Conferencia Iberoamericana en Sistemas, Cibernética e Informática CISCI 2009. Orlando (Florida); 2009, 146-150.