Extraction Student Dropout Patterns with Data Mining Techniques in Undergraduate Programs

Ricardo Timarán Pereira[1], Andrés Calderón Romero[1] and Javier Jiménez Toledo[2]
[1]*Departamento de Sistemas, Universidad de Nariño, Ciudad Universitaria Torobajo, Pasto, Colombia*
[2]*Programa de Ingeniería de Sistemas, Institución Universitaria CESMAG, Carrera 20ª No 14-54, Pasto, Colombia*

Keywords: Extraction Patterns, Student Dropout, Data Mining.

Abstract: The first results of the research project that aims to identify patterns of student dropout from socioeconomic, academic, disciplinary and institutional data of students from undergraduate programs at the University of Nariño from Pasto city (Colombia), using data mining techniques are presented. Built a data repository with the records of students who were admitted in the period from the first half of 2004 and the second semester of 2006. Three complete cohorts were analyzed with an observation period of six years until 2011. Socioeconomic and academic student dropout profiles were discovered using classification technique based on decision trees. 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.

1 INTRODUCTION

Countries in Latin America face similar challenges regarding to how minimize the impact of student dropout in higher education institutions. Many institutions carry out diverse strategies such as: educational loans, coverage increase, quality assurance, improvement in access equity and permanency, a better articulation with schools, diverse offers to attend different dimensions, interest and needs, and a better inclusion of graduates in the work force. According to UNESCO International Institute for Higher Education in Latin America and the Caribbean (IESALC), countries in Latin America have shown in 2003 an average coverage of higher education of 28.7% and a student dropout rate above of 50% (MEN, 2006a).

In Colombia, the education system has 277 higher education institutions, of which 81 are public and 196 private. According to the National Information System of Higher Education (SNIES) coverage for 2006 was 26.1%, which is equivalent to 1,301,728 students (MEN, 2006a). One of the main problems facing the Colombian higher education system concerns the high dropout levels (UPN, 2005). Although recent years have been characterized by increased coverage and new student enrollment, the number of students which are able to complete their higher education is not high, suggesting that a large part of these abandoned his studies, especially in the first semester (MEN, 2009). According to Ministry of National Education, of every hundred students who enter at a university about half fails to complete their academic year and get graduation (MEN, 2009). In 2004, the dropout was estimated at 49%. As causes of student dropout it can be mentioned: economic and financial constraints, poor academic performance, vocational and professional disorientation and difficulty adjusting to the college environment (MEN, 2006a). The dropout carries high social and economic costs that affect families, students, institutions and the State (MEN, 2006b).

Data mining in education is not a new topic and its study and implementation has been very relevant in recent years. The use of these techniques allows, among other things, to predict any phenomena within the educational environment. Thus, using the techniques offered by data mining, you can predict, with a high percentage of confidence, the probability of dropout of any student (Valero, 2009), (Valero et al., 2010).

This paper describes the process of extracting student dropout patterns from undergraduate programs at the University of Nariño from Pasto city (Colombia) using the classification technique based on decision trees. Considering the stages of knowledge discovery in databases, initially selected

Timarán Pereira R., Calderón Romero A. and Jiménez Toledo J..
Extraction Student Dropout Patterns with Data Mining Techniques in Undergraduate Programs.
DOI: 10.5220/0004543001360142
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval and the International Conference on Knowledge Management and Information Sharing* (KDIR-2013), pages 136-142
ISBN: 978-989-8565-75-4

from the databases of the University of Nariño, the socio-economic, academic, disciplinary and institutional data of students who were admitted between 2004 and 2006 to different undergraduate programs, in order to make a complete follow-up to 2011, determining whether or not dropped out.

With these data, a data repository was built using the PostgreSQL DBMS. These data were applied stages of pre-processing and transformation in order to obtain clean and ready data sets to apply the data mining techniques. The first results were obtained using the technique of classification based on decision trees with the free data mining tool named Weka. Finally, these results were analyzed, evaluated and interpreted to determine the validity of the obtained knowledge. 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.

The remain of the paper is organized as follows: In Section 2 related works are presented. In Section 3, the methodology applied in this research is explained in detail. Afterwards, Section 4 discusses the obtained results and, finally, section 5 shows the conclusions and future works.

2 RELATED WORKS

In Latin America, several research projects using data mining to discover patterns of dropout have been developed. At the National University of Misiones (Argentina) conducted research on student dropout using data mining techniques. Its main objective was to maximize the quality of the models to classify and group students according to their academic characteristics and their social and demographic factors. The study focused at students of the Analyst in Computer Systems Program at the Faculty of Natural Sciences. This research analyzed data from the cohorts from 2000 to 2006 (Pautsch, 2009), (Pautsch et al., 2010).

Similarly, at the National University of the Northeast (Argentina), a study was carried out with the main goal of applying data warehousing and data mining techniques, mainly based on clustering, in order to find dropout profiles. The techniques were applied to students of the Operating Systems Course of the Information Systems Program. As results, students were classified by their academic performance and their demographic and socioeconomic status, allowing a priori knowledge of potential scenarios for academic success or failure (La Red et al., 2010).

At the National University of Matanza (Argentina) were applied data mining techniques to evaluate academic performance and dropout of students from the Department of Engineering and Technological Research on data since 2003 to 2008. The implementation of this process was performed using MS SQL Server software to generate a data warehouse, SPSS software for data preprocessing and Weka (Waikato Environment for Knowledge Analysis) to find a classifier for academic performance to detect patterns of education dropout (Spositto et al., 2010).

At Izúcar Technological University of Matamoros (Mexico), proposed a research to identify the causes for the desertion of its students. Using the technique of classification and Weka as data mining tool, they found relationships between academic attributes that identify and predict the chance of dropout. Additionally, they proposed a tool that allows the tutor to predict the probability of dropping out of any student at any time during your stay at University (Valero, 2009), (Valero et al., 2010).

At the University of Sabana, in Colombia, conducted a research project where the goal was to select, from a database of students, attributes that have the greatest impact on the dropout of the university in the past four years. The study used classification by rough sets as data mining technique using the Rose2 software package (Restrepo and López, 2008).

Pinzon (2011) presents the characterization of student profile defector from the School of Marketing and Advertising at the Sergio Arboleda University (Colombia). The author applied clustering, using the K-means algorithm, as data mining technique. During the study, it was analyzed demographic variables obtained in the last student enrollment record and the causes that generated the withdrawal. The end result is three types of clusters or groups of students. For the case of the research, they were significant profiles.

3 STUDENT DROPOUT PATTERNS DISCOVERY PROCESS

Knowledge Discovery in Databases (KDD) is the extraction of implicit, previously unknown, and potentially useful information from data (Witten and Frank, 2000). KDD is basically an automatic process that combines discovery and analysis. The KDD process is interactive and iterative, involving numerous steps with many decisions made by user. This process usually involves preprocessing the data, make data mining and visualize the results (Agrawal and Srikant, 1994), (Chen et al., 1996), (Piatetsky-Shapiro, Brachman and Khabaza, 1996), (Han and Kamber, 2001).

In the process of discovering student dropout patterns in undergraduate programs, the following steps were performed:

3.1 Selection Step

The main goal of this step is selecting a data set from internal or external sources of data, or focusing on a subset of variables or data samples, on which discovery is to be performed. Internal sources were selected from the *Grades* and *Register-UDENAR* databases from the Admissions and Academic Control Office (OCARA) of the University of Nariño. Given the observation window for this study (2004-2011), these databases store personal and academic information of 15.875 students (OCARA, 2011). As main external sources of data were selected diverse databases from different Colombian Institutions such as: the Colombian Institute for the Development of Higher Education (ICFES), the National Bureau of Statistics (DANE), the Dropout Prevention in Higher Education System (SPADIES), the Potential Beneficiaries of Social Programs Information System (SISBEN) and the Colombian National Registry of Civil Status (all acronyms come from their names in Spanish).

From 15.875 records previously selected only data of students of cohorts 2004, 2005 and 2006 with the attributes most relevant to this study were chosen. The outcome is 6870 records and 62 attributes belonging to socio-economic, academic, and institutional data. These data were stored in a table named T6870A62 in a database called UDENAR_REPOSITORY using PostgreSQL. This table will be the basis for subsequent phases of the dropout patterns discovery process.

3.2 Preprocessing Step

The goal at this stage is to obtain clean data, i.e. data without null or outlier values, in order to retrieve high quality patterns. Through ad-hoc queries or histograms on the T6870A62 table, the quality of the data available for each of its attributes was thoroughly analyzed.

Considering the relevance of certain attributes for this research, null values of these attributes were updated with the values found in external sources. However, the attributes with a high percentage of nulls data (more than 80%), were eliminated by the inability to obtain these values from external sources, using statistical techniques such as mean, median and mode or deriving their values through others.

As result of this stage and in order to generate knowledge about the socioeconomic, academic, disciplinary and institutional factors, the 31 most representative attributes were selected from T6870A62 table. A new table was created and called as T6870A31. From these 31 attributes, 18 were chosen to analyze the socioeconomic factor and 15 for the academic factor. Similarly, two new tables (T6870A18 and T6870A15) were created respectively. A detail description of these new tables is shown in Table 1. Given the small number of selected attributes for disciplinary and institutional factors, these were added in the academic one.

Table 1: New tables of UDENAR-REPOSITORY database.

TABLE	DESCRIPTION
T6870A31	Table containing 6870 students admitted in 2004-2006 and with 31 attributes to be considered in the study.
T6870A18	Table of 6870 students and 18 attributes to consider social and economic factors.
T6870A15	Table of 6870 students and 15 attributes to consider for academic factors.

3.3 Transformation Step

Data transformation includes any process that modifies the form of the data. The aim of this stage is to transform the data source in a dataset ready to apply any of the different techniques of data mining. Among the operations performed to transform the data are: elimination of the least relevant attributes, creation of new attributes by deriving them from others (keeping or replacing these attributes) and / or modification of the type of attributes (using discretization or continuity methods).

In order to facilitate patterns extraction, the numerical values of attributes in table T6870A31 were translated to nominal values. This process (known as discretization) was carried out using the discretize filter in Weka with the equal frequency parameter (useEqualFrequency) set to 6 values. Moreover, the T6870A31 table was adapted to ARFF format (Attribute Relation File Format) required by Weka to continue with the data mining phase. Table 2 shows the attributes of the table T6870A31 with the new discretized values. According to this table, the first 17th attributes together with the attribute 31th, form the T6870A18 table and correspond to the socioeconomic attributes. Similarly, the attributes from 17th to 31st correspond to the academic attributes and conform the T6870A15 table.

Table 2: Attributes T6870A31.

Table 2: Attributes T6870A31. (Cont.)

No	ATTRIBUTES AND VALUES
1	gender {male,female}
2	marital_status {single,married,divorcée,civil union,single mother,widower,religious}
3	place_of_birth {pasto,north,south,west,center,cost,putumayo,other}
4	place_of_provenance {pasto,north,south,west,center,cost,putumayo,other}
5	health_regimen {contributory, subsidized}
6	economic_status {0,1,2,3,4,5,6,99}
7	father_alive {true,false}
8	father_occupation { various, operational, art, industry, building, professional, ...}
9	mother_alive {true,false}
10	mother_occupation { various, operational, art, industry, building, professional, ...}
11	residence_type {leased, owner,mortgage, ...}
12	live_with_family{true, false}
13	siblings_at_university {true, false}
14	family_income{'from 4540000 to 5980000','greater than 8540000','from 2850000 to 4540000','from 5980000 to 8854000','less than 2850000'} *figures in COP.
15	school_tuition_fee {'from 76639 to 106100','from 60248 to 76639','greater than 106100','less than 21550','from 21550 to 44369','from 44369 to 60247'} *figures in COP.
16	college_ tuition_fee {'less than 100259','from 120574 to 158846','from 100259 to 120574','from 234266 to 381504','from 158846 to 234266','greater than 381504'} *figures in COP.
17	age_at_enrollment {'18','less than 18','greater than 22','from 21 to 22','19','20'}
18	school_type {public,private}
19	school_day-part {morning,afternoon,full,night,weekend}
20	icfes_weighted{'from 52 to 54','from 50 to 52','from 54 to 58','from 46 to 50','greater than 58','less than 46'}
21	icfes_average{'from 53 to 56','from 48 to 50','from 46 to 48','from 50 to 53','less than 46','greater than 56'}
22	icfes_total {'greater than 475','from 420 to 450','from 450 to 475','from 400 to 420','from 375 to 400','less than 375'}
23	campus {Pasto,Tumaco,Tuquerres,Ipiales,Samaniego,Buesaco,'La Union',Ricaurte}
24	faculty {'natural sciences','humanities','agricultural sciences','economic sciences','health sciences',engineering, ...}
25	area_program {mathematics,'social sciences',agronomy, economy, accounting, arts, education}
26	grade_average {'from 2.4 to 3.1','from 3.5 to 3.7','greater than 4.0','from 3.7 to 4.0','from 3.1 to 3.5','less than 2.4'}
27	failed_courses {'from 3 to 4','greater than 9','from 5 to 6',none,'from 1 to 2','from 7 to 9'}
28	failed_semester {initial semesters, middle semesters, final semesters, not applicable, elective course}
29	area_course {pedagogy,philosophy,history,basic sciences, statistics, ...}
30	failed_times {'2','3',none,'1','4','greater than 4'}
31	dropout {true, false}

3.4 Data Mining Step

The goal of the data mining step is the search and discovery for unexpected and interesting patterns from data. In this process, intelligent task are applied such as *classification* (Quinlan, 1986), (Wang et al., 1998), (Witten and Frank, 2000), *clustering* (Ng and Han, 1994), (Zhang et al., 1996), *sequential patterns* (Agrawal and Srikant, 1995), *associations* (Agrawal and Srikant, 1994) among others. The data mining task chosen for the process of discovering student dropout patterns was classification using a decision trees technique.

Data classification provides results from a supervised learning process. Data classification is a two-step process. In the first step, a model is built describing a predetermined set of data classes. The input data, also called the training set, consists of multiple examples (records), each having multiple attributes or features and tagged with a special class label. In the second step, the model is used to classify future test data for which the class labels are unknow (Han and Kamber, 2001).

Decision tree classification is the most popular model, because it is simple and easy to understand (Han and Kamber, 2001), (Sattler and Dunemann, 2001), (Timarán and Millan, 2006). A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes. The top-most node in a tree is the root node (Han and Kamber, 2001).

A decision tree classifier is built in two phases: a growth phase and a pruning phase. In the growth phase, the tree is built by recursively partitioning the data until all members belong to the same class. In the pruning phase, many branches are removed with the goal of improving classification accuracy on data (Wang et al., 1998).

The classification rules were obtained with the Weka data mining tool (Waikato Environment for Knowledge Analysis), using the J48 algorithm, which implements the algorithm C.45 (Quinlan, 1993), and utilizing data repositories described in Table 1.

The J48 algorithm is based on the use of the gain ratio criterion. Consequently, those variables with greater number of different values do not have benefit in the selection. Furthermore, the algorithm incorporates a classification tree pruning once it has been induced (Hernández and Lorente, 2009).

T6870A31 repository was used to discover general patterns that affect the student dropout. The *dropout* attribute was chosen as the class. Similarly, the datasets T6870A15 and T6870A18 were used to determine, respectively, socioeconomic and academic factors that influence student dropout.

3.5 Evaluation Step

The objective of this final stage is the evaluation and interpretation of the obtained results in order to consolidate the discovered knowledge with two goals in mind. First, to integrate it into other systems for further action, and second, to compare it with previously discovered knowledge.

The 10-fold cross validation method was used in order to evaluate the quality and prediction accuracy of the discovered patterns. The most relevant classification rules are shown in the following section.

4 RESULTS AND DISCUSSION

As a result of interpreting the decision tree generated by the algorithm J48 with data from T6870A31, the most representative classification rules are shown in Table 3. All of them have a confidence threshold greater than 80%.

According to the rules of Table 3, the predominant factors in the student dropout from the University of Nariño are academics, specially a low average in grades and the number of courses lost in the initial semesters of the program.

In order to determine the socioeconomic factors affecting student dropout, a number of classification rules, with confidence greater than 80%, were generated using the T6870A18 dataset. The results show that the most significant socioeconomic factors affecting student dropout are a tuition fee greater than COP$ 381,504 (around USD$ 212) and a provenance from the south of the department of Nariño (Colombia). The fact of being single, living with mother and be in the city of Pasto may also impact education dropout.

Table 3: Most representative classification rules from T6870A31 dataset.

Classification rules	Dropout class	Support	Confidence
grade_average = from 2.4 to 3.1 & failed_semester = Initial semesters	True	0.1559	0.939
grade_average = from 3.7 to 4.0 & failed_times = 1	False	0.1551	0.8528
grade_average = less than 2.4	True	0.1519	0.998
grade_average = from 3.5 to 3.7 & campus = PASTO & failed_courses = from 7 to 9	False	0.0314	0.8585
grade_average = from 3.1 to 3.5 & failed_courses = from 3 to 4	True	0.0264	0.9535
grade_average = from 3.5 to 3.7 & failed_courses = De 1 to 2 & failed_semester = Initial semesters	True	0.0227	0.8108
grade_average = from 3.1 to 3.5 & failed_courses = De 5 to 6 & failed_semester = Initial semesters	True	0.017	0.8198
grade_average = from 3.5 to 3.7 & campus = PASTO & failed_courses = from 1 to 2 & failed_semester = Initial semesters & place_of_provenance = PASTO	True	0.0129	0.8341

To determine other factors associated with academic dropout, classification rules were generated with T6870A15 dataset with a confidence greater than 80%, without the attribute grade_average. The results shown the factors that influence academic dropout, in addition to a low average of grades and the courses lost in initial semesters, are: the faculty to which the student belongs, specifically the faculties of Natural Sciences, Health Sciences, Education and Arts; Also, the area of the course which was lost, such as the area of mathematical foundations, introduction to natural science, basic training, pedagogy, economics and accounting; and the campus of the University, particularly those located in Ipiales and Tumaco cities.

5 CONCLUSIONS AND FUTURE WORKS

Initial results obtained through the decision tree classification technique indicates that it is able to generate models consistent with observed reality and theoretical support, based only on the data that is stored in the database, for the study case of the University of Nariño. One of the great difficulties faced in these kinds of studies is the poor data quality. Often, when the cleaning process was ended, many variables become useless by the inability to obtain their correct values. Unfortunately, it has a direct influence on the results of data mining.

A set of general patterns for student dropout has been obtained. It is mainly determined by factors such as a low average in grades and the number of courses students have failed at initial semesters. In addition, socioeconomic and academic factors related with student dropout have been identified as well. The assessment, analysis and utility of these patterns will support effective decision-making of university staff focused to develop policies and strategies related to student retention programs that are currently set.

As future works it can be mentioned additional studies of student dropout at the University of Nariño using other data mining techniques such as clustering and association rules in order to determine affinities, similarities and relationships between socioeconomic and academics factors of students who drop out. To verify the quality and accuracy of the rules obtained will be used other classifiers.

Applying the same methodology for student dropout at CESMAG University Institution and analyze and evaluate the patterns found in both higher education institutions.

ACKNOWLEDGEMENTS

This research is funded by the Ministry of Education in Colombia and counterpart funds from the University of Nariño and CESMAG University Institution.

REFERENCES

Adamo, J. M., 2001. Data Mining for Association Rules and Sequential Patterns: *Sequential and Parallel Algorithms. New York* (USA): Springer-Verlag. 253 p. ISBN: 0-387-95048-6.

Agrawal, R., Srikant, R., 1994. Fast Algorithms for Mining Association Rules. In: *20th International Conference on Very Large Data Bases*, VLDB 1994, (12-15/09/1994). Santiago de Chile (Chile): VLDB.Proceedings. p. 487-499. ISBN: 1-55860-153-8.

Agrawal, R., Srikant, R., 1995. Mining Sequential Patterns. In: The *11th International Conference on Data Engineering* ICDE, pp 3-14. Taipei, Taiwan.

Chen, M., Han, J., Yu, P., 1996. Data mining: An overview from database perspective. In: IEEE *Transactions on Knowledge and Data Engineering*. Vol. 8, No. 6 (dic). Los Alamitos (CA, USA): IEEE Computer Society. p. 866-883. ISSN: 1041-4347.

Fayyad, Usama, Piatetsky-Shapiro, Gregory, Smyth, Padrahic, 1996. *The KDD process for extracting useful knowledge from volumes of data. In: Comunications of the ACM*. Vol 39, No. 11 (nov). New York (USA): ACM Digital Library. p 27-34. ISSN: 0001-0782.

Garcia Morate, Diego (s.f.). Manual de Weka (on líne), http://www.metaemotion.com/diego.garcia.morate/download/weka.pdf (consulta: 15/06/ 2012).

Han, Jiawei, Kamber, Micheline, 2001. Data Mining: Concepts and Techniques. San Francisco (CA, USA): Morgan Kaufmann Publishers, Academic Press. 550 p. ISBN: 1-55860-489-8.

Hernández, O. J., Ramírez, Q. M., Ferri, R. C., 2005. Introducción a la Minería de Datos. Madrid (España): Pearson Prentice Hall. 656 p. ISBN: 84-205-4091-9.

Hernández, E., Lorente, R., 2009. Minería de datos aplicada a la detección de Cáncer de Mama. Universidad Carlos III, Madrid. http://www.it.uc3m.es/jvillena/irc/practicas/08-9/14.pdf.

La Red, David, Acosta, J. C., Cutro, Luis, Uribe, V. E., Rambo, A. R., 2010. Data Warehouse y Data Mining Aplicados al Estudio del Rendimiento Académico. In: Novena Conferencia Iberoamericana en Sistemas, Cibernética e Informática, CISCI 2010, (29/06-2/07/2010), Orlando (Florida, USA): *International Institute of Informatics and Systemics. Memorias CISCI 2010*, Volumen I, p. 289-294. ISBN: 978-1-934272-94-7.

MEN, 2006a. América Latina piensa la deserción. *En: Boletín informativo Educación Superior. No 7* (dic). Bogotá (Colombia): Ministerio de Educación Nacional. p 14. ISSN: 1794-2446.

MEN, 2006b. Deserción estudiantil: prioridad en la agenda. *En: Boletín informativo Educación Superior*. No 7 (dic). Bogotá (Colombia): Ministerio de Educación Nacional. p 1. ISSN: 1794-2446.

MEN, 2009. Deserción estudiantil en la educación superior colombiana: *metodología de seguimiento, diagnóstico y elementos para su prevención. Bogotá (Colombia): Ministerio de Educación Nacional*. 158 p. ISBN: 978-958-691-366-9.

OCARA, 2011. Datos de estudiantes matriculados en los programas de pregrado de la Universidad de Nariño en el periodo A de 2004 hasta el periodo A de 2011. *Oficina de Control y Registro Académico de la Universidad de Nariño*. Pasto (Colombia).

Pautsch, Jesús, 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*. 193 p.

Pautsch, Jesús, La Red, David, Cutro, Luis, 2010. Minería de datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación (on líne). Posadas, Misiones (Argentina): Universidad Nacional de Misiones. http://www.dataprix.com/files/Analisis%20de%20Desercion%20Univ_0.pdf. (consulta: 18/06/2012).

Pinzón, Liza, 2011. Aplicando minería de datos al marketing educativo. In: Revista Notas de Marketing. No 1 (jun). Bogotá (Colombia): Universidad Sergio Arboleda, Escuela de Marketing y Publicidad. p 45-61.

Quinlan, J. R., 1986. Induction of Decision Trees. *In: Machine Learning Journal, Vol. 1,* No. 1, Kluwer Academic Publishers, pp. 81-106. Boston, USA.

Quinlan, J. R., 1993. C4.5: Programs for Machine Learning. San Francisco (CA, USA): Morgan Kaufmann Publishers. 299 p. ISBN: 1-55860-238-0.

Restrepo, Mauricio, López, Andrés, 2008. Uso de la metodología Rough Sets en un modelo de deserción académica. In: XIV Congreso Ibero Latinoamericano de Investigación de Operaciones, CLAIO 2008, (9-12/09/2008), Cartagena (Colombia). *Universidad del Norte. Libro de Memorias CLAIO 2008*, p.108-109. Ediciones Uninorte.

Sattler, K., Dunemann, O., 2001. SQL Database Primitives for Decision Tree Classifiers. In: *The 10th ACM International Conference on Information and Knowledge Management* - CIKM, (5-10/11/2001), Atlanta, Georgia (USA): ACM. Proceedings, p. 379-386. ISBN: 1-58113-436-3.

Sposito, Osvaldo, Etcheverry, Martín, Ryckeboer, Hugo, Bossero, Julio, 2010. Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil. In: Novena Conferencia Iberoamericana en Sistemas, Cibernética e Informática, CISCI 2010, (29/06-2/07/2010), Orlando (Florida, EE.UU.). *International Institute of Informatics and Systemics. Memorias CISCI 2010*, Volumen I. ISBN: 978-1-934272-94-7.

Timarán, R., Millán, M., 2006. New Algebraic Operators and SQL Primitives for Mining Classification Rules. In: proceedings of The Five IASTED International Conference on Computational Intelligence (CI 2006), International Association of Science and Technology for Development, ISBN No 0-88986-603-1. San Francisco, USA.

Timarán Pereira, Ricardo, 2009. Una mirada al descubrimiento de conocimiento en bases de datos. In: Ventana Informática. No 20 (ene-jun., 2009). Manizales (Colombia): Centro de Investigaciones, Desarrollo e Innovación, Facultad de Ingeniería, Universidad de Manizales. p 39-58. ISSN: 0123-9678.

UPN, 2005. La deserción estudiantil: reto investigativo y estratégico asumido de forma integral por la UPN (on line). In: Encuentro Internacional sobre Deserción en Educación Superior: experiencias significativas (17-18/05/2005) Bogotá (Colombia): Ministerio de Educación Nacional. http://www.mineducacion.gov.co/1621/ articles-85600_Archivo_pdf3.pdf. (consulta: 15/06/ 2012).

Valero, Sergio, 2009. Aplicación de técnicas de minería de datos para predecir la deserción (on línea). Izúcar de Matamoros, Puebla (Mexico): Universidad Tecnológica de Izúcar de Matamoros. http://www.utim.edu.mx/~svalero/docs/MineriaDesercion.pdf. (consulta: 10/06/2012).

Valero, Sergio, Salvador, Alejandro, García, Marcela, 2010. Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos (on line) Izúcar de Matamoros, Puebla (Mexico): Universidad Tecnológica de Izúcar de Matamoros. http:// www.utim.edu.mx/~svalero/docs/e1.pdf. (consulta: 10/06/2012).

Wang, M.; Iyer, B., Scott, V., J., 1998. Scalable Mining for Classification Rules in Relational Databases. In: *International Database Engineering and Application Symposium,* IDEAS 98, (08-10/07/1998), Cardiff (Wales,U.K.): IEEE Computer Society. Proceedings, p. 58-67. ISBN: 0-8186-8307-4.

Witten, I. and Frank, E., 2000. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers, 365 p, ISBN: 1-55860-552-5. San Francisco, CA, USA.

Zhang, T., Ramakrishnan, R., Livny, M., 1996. BIRCH: An Efficient Data Clustering Method for Very Large Databases. In ACM SIGMOD International Conference on Management of Data, pp. 103-114. Montreal, Canada.