2017 16th IEEE International Conference on Machine Learning and Applications

Aplication of Decision Trees for Detection of Student Dropout Profiles

Ricardo Timarán Pereira
Departamento de Sistemas
Universidad de Nariño
San Juan de Pasto, Colombia
ritimar@udenar.edu.co

Javier Caicedo Zambrano
Departamento de Matemáticas y Estadística
Universidad de Nariño
San Juan de Pasto, Colombia
Jacaza1@udenar.edu.co

***Abstract*— The 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.**

Keywords— Extraction Patterns; Student Dropout; Decision trees.

I. INTRODUCTION

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 [1]. One of the main problems facing the Colombian higher education system concerns the high dropout levels [2]. 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 [3]. 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 [3]. 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 [1]. The dropout carries high social and economic costs that affect families, students, institutions and the State [4].

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 [5], [6].

In this paper, the 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 classification technique based on decision trees are presented.

The remain of the paper is organized as follows: In Section II, the methodology applied in this research is explained in detail. Afterwards, Section III discusses the obtained results and, finally, section IV shows the conclusions and future works.

II. METODOLOGY

Considering the stages of knowledge discovery in databases [7], initially selected 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.

A. 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 15.805 students [8]. 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

0-7695-6321-X/17/31.00 ©2017 IEEE
DOI 10.1109/ICMLA.2017.0-107

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 15.805 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.

B. 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 14 for the academic factor. Similarly, two new tables (T6870A18 and T6870A14) 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. TABLES OF UDENAR_REPOSITORY DATABASE

TABLE	DESCRIPTION
T6870A31	Table containing 6870 students admitted in 2004-2006 with 31 attributes to be considered in the study.
T6870A18	Table of 6870 students and 18 attributes to consider social and economic factors.
T6870A14	Table of 6870 students and 14 attributes to consider for academic factors.

C. Transformation Step

The 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 description of the attributes of T6870A31 in arff format.

TABLE 2. DESCRIPTION OF T6870A31 IN ARFF FORMAT

ARFF FORMAT	
@attribute	gender {m,f}
@attribute	Marital_status {Single, married, 'free union', separated, 'single mother', widower, religious}
@attribute	Birth_area {sur,pasto,costa,putumayo,centro,norte,'centro occidente','otras regiones'}
@attribute	Place_of_provenance {pasto,sur,centro,'centro occidente',costa,putumayo,norte,'otras regiones'}
@attribute	Health_regimen {contributivo,subsidiado}
@attribute	Social_stratum {0,1,2,3,4,5,6,99}
@attribute	father {n,y}
@attribute	Father_occupation{Several, 'officers, workers, craftsmen, manufacturing industry, construction and mining, 'without occupation', household}
@attribute	mother {n,y}
@attribute	Mother_occupation {Unqualified workers, household, 'without occupation', 'service workers and salesmen', pensioners}
@attribute	Type of house {'Leased', own, 'own paid for fees'}
@attribute	Lives_with_the_family{'y ','n '}
@attribute	Brothers_in_the_university {'n ','y '}
@attribute	Family_income {'de 4540000 a 5980000','mayor a 8540000','de 2850000 a 4540000','menor a 2850000','5980000 a 8854000'}
@attribute	School_enrollment_value {'de 76639 a 106100','de 60248 a 76639','menor a 21550','mayor a 106100','de 21550 a 44369','de 44369 a 60247'}
@attribute	University_enrollment_fee {'menor a 100259','de 120574 a 158846','de 158846 a 234266','mayor a 381504','de 100259 a 120574','de 234266 a 381504'}
@attribute	Age_of_admission_to_university {'igual a 18','menor a 18','mayor a 22','de 21 a 22','igual a 19','igual a 20'}
@attribute	Type_college {public, private}
@attribute	School_day {Morning, afternoon, full, night, Saturday}
@attribute	Weighted_ICFES_test {'de 52 a 54','de 50 a 52','de 54 a 58','mayor a 58','menor a 46','de 46 a 50'}
@attribute	Average_ICFES_test{'de 53 a 56','de 48 a 50','mayor a 56','menor a 46','de 50 a 53','de 46 a 48'}
@attribute	Total_ICFES_test {'mayor a 475','de 420 a 450','de 400 a 420','de 450 a 475','de 375 a 400','menor a 375'}
@attribute	campus {pasto,ipiales,tumaco,'la union',ricaurte,tuquerres,samaniego,buesaco}
@attribute	faculty {'Natural sciences', 'health sciences', 'economic and administrative sciences', 'human sciences', 'agricultural sciences'}

ARFF FORMAT	
@attribute	Area_of_the_program {'mathematics and natural sciences', 'health sciences', 'economics, administration, accounting and related',
@attribute	Grade_average {'de 2.4 a 3.1','de 3.5 a 3.7','mayor a 4.0','menor a 2.4','de 3.1 a 3.5','de 3.7 a 4.0'}
@attribute	failed_courses {'de 3 a 4','mayor a 9','de 5 a 6',ninguna,'de 7 a 9','de 1 a 2'}
@attribute	failed_semester {'p ','m ',na,ce,'u '}
@attribute	Area_of_the_course {'formación específica',na,'competencias básicas y formación humanística','formación investigativa', etc}
@attribute	failed_times {'igual a 2','igual a 3',ninguna,'igual a 1','igual a 4','mayor a 4'}
@attribute	dropout {t,f}

D. Data Mining Step

The goal of the data mining step is the search and discovery for unexpected and interesting patterns from data. The data mining task chosen for the process of discovering student dropout patterns was classification using a decision trees technique [9]. 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 [10].

The patterns of student dropout were obtained with the Weka data mining tool (Waikato Environment for Knowledge Analysis), using the J48 algorithm, which implements the algorithm C.45 [9].

T6870A18 repository was used to discover general patterns that affect the student dropout. The dropout attribute was chosen as the class. Similarly, the datasets T6870A18 and T6524A14 were used to determine, respectively, socioeconomic and academic factors that influence student dropout.

E. Evaluation Step

The objective of this final stage is the 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.

III. Results and Discussion

As a result of interpreting the decision tree generated by the algorithm J48 with data from T6524A31, 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 1, 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.

TABLE 3. Most Representative Classification Rules from T6870A31 Dataset

CLASSIFICATION RULES	DROPOUT CLASS	SUPPORT	CONFIDENCE
grade_average between 2.4 and 3.1 & failed_semester = Initial semesters	True	0.1559	0.939
grade_average between 3.7 and 4.0 & failed_times = 1	False	0.1551	0.8528
grade_average less than 2.4	True	0.1519	0.998
grade_average between 3.5 and 3.7 & campus = PASTO & failed_courses between 7 and 9	False	0.0314	0.8585
grade_average between 3.1 and 3.5 & failed_courses between 3 and 4	True	0.0264	0.9535
grade_average between 3.5 and 3.7 & failed_courses between 1 and 2 & failed_semester = Initial semesters	True	0.0227	0.8108
grade_average between 3.1 and 3.5 & failed_courses between 5 and 6 & failed_semester = Initial semesters	True	0.017	0.8198
grade_average between 3.5 and 3.7 & campus = PASTO & failed_courses between 1 and 2 & failed_semester = Initial semesters & place_of_provenance = PASTO	True	0.0129	0.8341

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 university_enrollment_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.

To determine other factors associated with academic dropout, classification rules were generated with T6870A14 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.

IV. 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.

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

ACKNOWLEDGMENT

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

REFERENCES

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

[2] UPN.: 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* ,(consulted: 15/06/ 2012) (2005).

[3] MEN.: 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 (2009).

[4] MEN.: 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 (2006).

[5] Valero, S.: 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* , (consulted: 10/06/2012) (2009).

[6] Valero, S., Salvador, A. & Garcia, 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 [on line]. Izúcar de Matamoros, Puebla (Mexico): Universidad Tecnológica de Izúcar de Matamoros. *http://www.utim.edu.mx/~svalero/docs/e1.pdf*, (consulted: 10/06/2012). (2010).

[7] Han, J & Kamber, M.: Data Mining Concepts and Techniques. San Francisco (CA, USA): Morgan Kaufmann Publishers, Academic Press. 550 p. ISBN: 1-55860-489-8 (2001).

[8] OCARA.: 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 (2011).

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

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