Applying CRISP-DM in a KDD Process for the Analysis of Student Attrition

Luis Fernando Castro R.(✉), Esperanza Espitia P., and Andrés Felipe Montilla

Universidad del Quindío, Armenia, Colombia
{Lufer, eespitia, afmontilla}@uniquindio.edu.co

Abstract. The data mining techniques are focused mainly on supporting to the decision makers in a specific organization. The student attrition is a common phenomenon that worries to public and private universities, which are economically and socially affected. Several studies have addressed this issue, however they have focused mainly on academic, social, demographic, and economic aspects. In this paper, we propose a method for analyzing the academic desertion by providing a view of this problematic from a KDD (knowledge discovery in databases) perspective and using techniques for identifying behavior patterns of the students. Unlike other proposals we also considerer variables provided by the BADyG test. This proposal is important because it will provide a support to the decision-making and the creation of action plans by higher education institutions to reduce the high rate of student attrition.

Keywords: Data mining · Student attrition · KDD · Patterns · CRISP-DM Analysis

1 Introduction

Nowadays, the student attrition is a problematic that worries to the public and private universities. The causes of this problem have not been accurately identified; besides, the way of using the data associated with students in order to generate useful information that can serve to face this problem is a challenge. According to [8], one of the main difficulties facing the current educational system is the desertion, its accumulated value reaches levels of 45% at the national level; one of its main causes is the desertion of academic type. Some authors [16] claim that the main problem facing the Colombian higher education system is the high levels of student attrition. They argue that the number of students that completed their higher education is very low, perceiving that most of these abandoned his studies in the first semester. And, that half the number of matriculated students in a higher education institution cannot complete their academic cycle. Finally, they say that the student attrition in 2004 was estimated at 49% by exhibing the following causes: economic and financial constraints, low academic performance, vocational and professional disorientation and difficulties for adapting to university environment. According to [1], only one of every two students who enroll in an undergraduate program completes his career. The concern is greater if you consider that 39.52% of who drop out his studies argue economic reasons.

J. E. Serrano C. and J. C. Martínez-Santos (Eds.): CCC 2018, CCIS 885, pp. 386–401, 2018.
https://doi.org/10.1007/978-3-319-98998-3_30

The increase in attrition rates has become a problem that has been of interest to higher education institutions and the educational authorities. This problematic has important socio-economic consequences. The loss of students causes serious problems to the universities, since it makes their sources of income are unstable. In addition, student desertion can compromise the future of a country in the medium and long term because the accumulation of scientific and technological knowledge is one of the factors that determine the socio-economic development of a nation [3]. According to [12], school failure is in any case a catastrophe, absolutely desolating on the moral, human and social level, which very often generates exclusions that will mark to young people during their adult life. Besides, desertion and abandonment produce uprooting, loneliness, absence of rites, lack of routines and loss of negotiating capacity with others, social loneliness". Then, the authors explain that desertion is a problem of the educational system related with the environments of the same, such as educational environments, family situations, environmental and cultural requirements that directly affect to the deserter.

Several studies based on KDD [6] have been developed. In [8] the authors use KDD in the extraction of knowledge from the database of the academic information system (SIA) of the University of Caldas in order to calculate indicators of academic performance. In [9] useful knowledge is generated in order to find possible causes of the student attrition problem of the autonomous University of Manizales from the large amounts of academic information in the academic registry office. The work presented in [1] contribute to the decision-making to reduce student attrition levels in the undergraduate programs of Mariana University applying the KDD process and based on a unified data repository with the socio-economic, personal, academic and institutional information of the students. Other study [2] makes an analysis of the student attrition in the system engineering program of the Simon Bolivar University. There, the causes of desertion are grouped into variables based on information from the academic registry office in order to establish patterns and support the decision making using the data mining techniques and the tool for automatic learning and data mining called WEKA. In this paper we propose a method for analyzing the academic desertion by using techniques, tools and methodologies based on KDD. Unlike the previous works, we are focused on the students of the system and computer engineering program at the University of Quindío. Besides, we include a set of variables provided by the statistical department, which are the result of a proof that assess different cognitive aspects related to the students. These variables are added to the information provided by others information systems such as register office and planning department. The data were provided by three sources: information concerning to the SPADIES taken from the planning and development office, BADyG test provided by the department of statistics and the personal and academic information of the student, taken from the register and control office of the University of Quindío

According to the problematic of the student attrition, some authors [12] say that is obligation of the educational institutions, especially the universities, to establish academic, administrative and adjustment mechanisms to the university life of its students so that they overcome the difficulties of the academic programs and culminate successfully their careers. In addition to the large volume of data available in the higher education institutions related with its students, in [6] the authors propose that a new

generation of computational techniques and tools is required to support the extraction of useful knowledge from the rapidly growing volumes of data. These techniques and tools are related with an emerging field of knowledge discovery in databases (KDD) and data mining. The information that can be obtained from academic databases will serve to answer to such questions as: Which are the causes of the student's retention in the university? Why do students desert? According to [13] automatic data mining techniques can be applied to solve these questions and to facilitate the development of strategies for improving the academic processes and the educational programs. The data mining can offer for the problematic of the student attrition a great variety of statistical and computational methods to investigate the existence of relations and behavior patterns of students in the first year of their university career [10].

The paper is organized as follows: Sect. 2 describes the main concepts to understand the proposal. In Sect. 3 we discuss some related work. In Sect. 4 we describe the proposal together with methodologies and tools used. And finally, in Sect. 5 some conclusions are presented.

2 Theoretical Framework

2.1 Student Attrition

According to [12], student attrition must be understood as the definitive abandonment of the classrooms for different reasons and the non-continuity in the academic formation for each person who begins his studies, hopeful in ending happily the university studies.

2.2 KDD

The term KDD refers to the overall process of discovering useful knowledge from data. Besides, this process focuses on the search for data patterns that are valid, novel, potentially useful and understandable [6].

2.3 BADyG

The term BADyG refers to a test for assessing different cognitive related to the subjects. The theoretical foundation of this test is that the intelligence is composed by a set of differentiated capacities instead of a single capacity [17, 18]. In Table 1 the variables provided by the BADyG test are showed.

2.4 Data Mining

Data mining is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules, allowing for example, a corporation to improve its marketing, sales, and customer support operations through a better understanding of its customers [11].

2.5 CRISP-DM

CRISP-DM is a methodology widely used for analyzing large volumes of data and discover valuable information. The CRISP-DM methodology, according to [19], consists of six phases: understanding the business, understanding the data, preparing the data, modeling, evaluation and implementation.

2.6 Rapid Miner

Rapid Miner is an open-source software (OSS). Rapid Miner provides functionality data preprocessing and visualization, predictive analytics and statistical modeling, in addition of data mining. There are many OSS in the market but the user interface and workflow of Rapid Miner [20] makes it different from other tools.

Table 1. Evaluated variables in the BADyG test. Source: [17, 18]

Variable	Mean
CI	Coefficient of intelligence
IG	General intelligence
RL	Logic reasoning
Rv	Analog relations
Rn	Numerical series
Re	Logical matrices
Sv	Complete sentences
Sn	Numerical problems
Se	Fit figures
Ma	Auditory memory
Mv	Visual memory
De	Discrimination of differences
RA	Speed
EF	Effectiveness

3 Related Work

Several works uses data mining techniques to identify behavioral patterns that are useful for a particular context. In [14] propose collecting statistical data in order to identify behavioral patterns about travelers visiting the city of Pereira. Also describe their habits and classify those habits regarding trends. This work was carried out by using the CRISP-DM methodology, the KDD process, and the tool called RapidMiner. The work presented in [7] generates useful knowledge from historical data analysis of successful and failed projects on an IT outsourcing company by applying data mining techniques, in order to obtain patterns for allowing the organization to make decisions and guide software projects towards the achievement of success. This work was done by using the CRISP-DM methodology, the KDD process and the SPSS (statistical

product and service solutions) tool. In [15] the authors use the data mining techniques to address the problem of student attrition at the Universidad distrital – Francisco José de Caldas in order to determine the causes that lead to desertion the students of the university. This work is intended to generate a decision tree model by implementing the J48 algorithm through the use of the WEKA tool for identifying such causes; also, they use the CRISP-DM methodology to development the work. Finally, the work presented in [5] analyze the academic information related to the academic results of the student and their interaction with the university. The authors identify factors that influence the student desertion of the computer science career at Gastón Dachary University in Argentina. This work apply data mining techniques and use classifications algorithms such as decision trees, Bayesians networks and rules. The work is developed under the KDD process and the WEKA tool.

These studies use the same technology in the same context of our proposal. However they have focused mainly on academic, social, demographic, and economic aspects. Unlike such proposals we also considerer variables provided by the BADyG test.

4 Our Proposal

This proposal consist of a method to analyze the academic desertion, starting from the data provided by the University of Quindío related to the students and making a particular focus on the faculty of engineering. For this proposal, we intend to work with the CRISP-DM methodology, the KDD process, especially in the data mining stage and the tool called RapidMiner. The goal of this proposal is to identify some behavioral patterns and relationships between a large numbers of variables of the students of the University of Quindío. Thus, this work can support decision-making and the creation of plans focused on addressing related problems with desertion in the engineering faculty of the University of Quindío. According to the CRISP-DM methodology, the proposal will handle five phases: understanding the domain, understanding the data, preparation of data, modeling, and analysis and evaluation, the Fig. 1 illustrates this proposal.

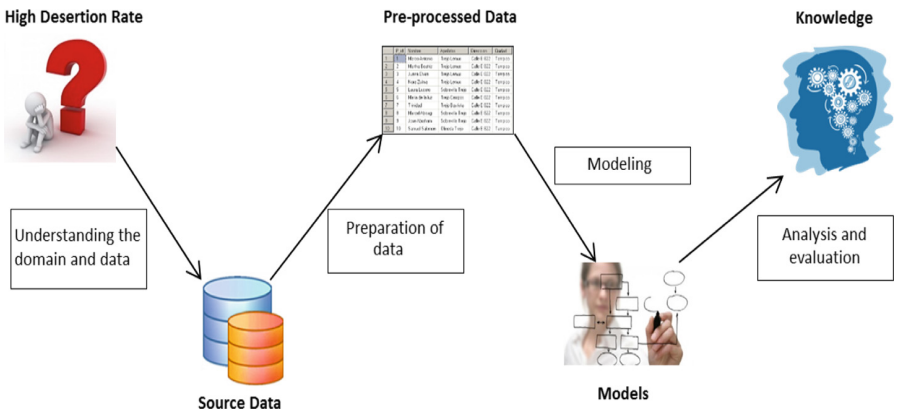

Fig. 1. Phases of the proposal.

4.1 Understanding the Domain

The development of this phase known in CRIS-DM as "understanding the domain" allows us to obtain a higher level of understanding of the problem related to the case study. In this case the task consists of consulting researches and previous work related to the problem of desertion at the University of Quindío as well as at national and international level. Documentation related to knowledge discovery techniques and their application to desertion themes is also reviewed.

4.2 Understanding the Data

In this phase of CRISP-DM named as "understanding the data" the methodology propose criteria for selecting the data. Such data are obtained and explored in order to identify the elements for allowing the determination of the quality of the same. In this case the following criteria were defined: student´s data provided by the diverse departments at the University of Quindío responsible for collecting and storing the information of the students. Such departments were control and register office, planning department, and statistical department. The source data were provided by three sources: information concerning to the SPADIES taken from the planning and development office, BADyG test provided by the department of statistics and the personal and academic information of the student, taken from the register and control office of the University of Quindío. BADyG test was performed by the department of statistics at the University of Quindío [18]. A sample of the results obtained can be consulted in Figs. 2, 3, and 4.

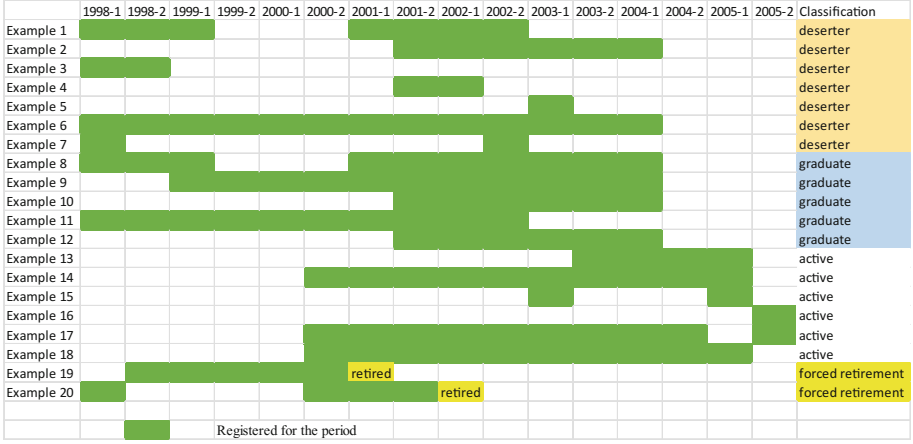

Fig. 2. Structure of the information provided by the SPADIES system. Source: [4]

4.3 Preparation of the Data

This phase named "Preparation of the data" in CRISP-DM consists of organizing and debugging the data obtained. This phase is composed by four general tasks and four

ESTP_ID	IG	RL	Rv	Rn	Re	Sv	Sn	Se	Ma	Mv	De	RA	EF
412345	97	47	16	15	16	15	12	23	21	18	15	143	68
412340	104	54	20	13	21	18	9	23	20	9	23	164	63
413207	59	32	17	8	7	12	0	15	20	13	11	136	43
415589	67	37	12	14	11	8	7	15	8	3	21	100	67
412350	46	24	12	4	8	11	4	7	10	9	17	120	38
412338	95	48	15	17	16	11	12	24	17	10	26	118	81
412334	73	44	14	16	14	12	2	15	23	12	25	104	70
412379	114	66	18	23	25	14	11	23	24	18	24	135	84
412352	57	26	12	8	6	12	7	12	12	11	22	127	45
412335	74	40	18	10	12	16	7	11	9	12	20	166	45
413208	85	50	23	15	12	18	7	10	21	10	17	143	59
412384	82	43	15	17	11	17	10	12	23	24	21	170	48
412456	101	66	21	22	23	13	15	7	18	19	15	135	75
412360	68	43	11	15	17	5	1	19	24	5	21	111	61
412336	71	36	13	13	10	10	7	18	26	17	18	114	62
412455	81	43	12	16	15	11	10	17	17	15	22	99	82
412344	26	8	3	1	4	8	2	8	8	12	14	92	28
412337	90	51	15	17	19	11	10	18	22	18	18	98	92
412347	77	39	18	8	13	17	2	19	17	11	28	159	48
415233	97	54	16	23	15	13	16	14	15	22	16	150	65
415232	65	33	8	16	9	12	8	12	12	17	24	127	51
412339	77	47	15	12	20	9	10	11	13	16	22	115	67
412351	71	38	17	9	12	14	4	15	14	16	20	129	55
412342	83	42	12	14	16	11	9	21	12	8	23	100	83
412454	60	34	12	11	11	10	7	9	6	13	16	112	54
412333	96	48	12	17	19	17	6	25	27	27	22	124	77
415308	58	23	11	0	12	15	6	14	26	14	17	93	62
412373	65	33	8	11	14	15	4	13	21	15	19	115	57
412459	85	53	18	14	21	12	5	15	20	15	22	140	61
412458	50	30	10	11	9	8	5	7	13	12	14	109	46

Fig. 3. Structure of the information provided by the statistical department of the University of Quindío related to BADyG tests. [18]

INGENIERIA DE SISTEMAS Y COMPUTAC	408321	CC	1922950644	INGENIERIA DE SISTEMAS Y COMPUTACION	BOGOTÁ D.C.	RISTIAM DANIEL SABOGAL RODRIGUE	M	29/09/1996	1				SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412333	TI	9712109963	INGENIERIA DE SISTEMAS Y COMPUTACION	ARMENIA	RISTIAN MAURICIO ACOSTA PERDON	M	21/01/1997	1	1	1	1	SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412334	TI	9903608090	INGENIERIA DE SISTEMAS Y COMPUTACION	0	DENILSON ANDRÉ AGUDELO TORRES	M	19/03/1999	1	1	1	1	SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412335	CC	1097405947	INGENIERIA DE SISTEMAS Y COMPUTACION	CALARCA	JUAN DAVID ALARCON RIVERA	M	22/09/1997	1	1	1	1	SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412336	TI	1002245303	INGENIERIA DE SISTEMAS Y COMPUTACION	ARMENIA	JENERI STIVEN ALVAREZ ACOSTA	M	26/02/1998	1	1	1	1	SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412337	CC	1094960957	INGENIERIA DE SISTEMAS Y COMPUTACION	ARMENIA	SEBASTIAN CAMILO ANTTURY SANCH	M	01/02/1997	1	1	1	1	SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412338	CC	1053854523	INGENIERIA DE SISTEMAS Y COMPUTACION	MANIZALES	FELIPE ARCE GIRALDO	M	29/09/1996	1	1	1	1	SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412339	TI	9710712970	INGENIERIA DE SISTEMAS Y COMPUTACION	ARMENIA	MARIA CAMILA ARCILA BERRIO	F	27/09/1997	1				SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412340	CC	1094943652	INGENIERIA DE SISTEMAS Y COMPUTACION	ARMENIA	JUAN DAVID ARISTIZABAL LOPEZ	M	09/09/1994	1	1			SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412341	CC	1097404979	INGENIERIA DE SISTEMAS Y COMPUTACION	CALARCA	JORES FELIPE BENAVIDES HERNANE	M	25/02/1996	1	1			SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412342	TI	9707059884	INGENIERIA DE SISTEMAS Y COMPUTACION	ARMENIA	DARWIN JESUS BONILLA RASCOS	M	09/07/1997	1	1	1	1	SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412344	CC	1097404520	INGENIERIA DE SISTEMAS Y COMPUTACION	CALARCA	JAVIER ANDRES CALDERON OSPINA	M	29/09/1996	1				SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412345	TI	9709269528	INGENIERIA DE SISTEMAS Y COMPUTACION	MONTENEGRO	YEFERSON CARDONA GIRALDO	M	24/09/1997	1				SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412346	CC	1097403389	INGENIERIA DE SISTEMAS Y COMPUTACION	CALARCA	KISE JEFERSON CASALLAS SANCHE	M	13/09/1995	1	1			SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412347	TI	9705236021	INGENIERIA DE SISTEMAS Y COMPUTACION	ARMENIA	DANIEL CASTRILLON MEJIA	M	29/05/1997	1				SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412348	CC	1097723953	INGENIERIA DE SISTEMAS Y COMPUTACION	MONTENEGRO	SEBASTIAN CEBALLOS ARIAS	M	19/09/1996	1	1	1	1	0
INGENIERIA DE SISTEMAS Y COMPUTAC	412349	TI	9707409840	INGENIERIA DE SISTEMAS Y COMPUTACION	CIRCASIA	JHON ALEJANDRO CELIS LOAIZA	M	04/07/1997	1				SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412350	TI	9711417265	INGENIERIA DE SISTEMAS Y COMPUTACION	QUIMBAYA	WILSON ANDRES CHIQUITO VILLADA	M	04/11/1997	1	1			SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412351	CC	1094022021	INGENIERIA DE SISTEMAS Y COMPUTACION	CIRCASIA	LUZ EDILIA DIAZ CORREA	F	02/11/1993	1				CASADO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412352	CC	1097403265	INGENIERIA DE SISTEMAS Y COMPUTACION	CALARCA	JORDAN CAMILO LOPEZ QUIENDO	M	22/07/1995	1				SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412353	CC	1026593160	INGENIERIA DE SISTEMAS Y COMPUTACION	ARMENIA	OSCAR ENRIQUE GARCÉS QUINTERO	M	14/04/1991	1	1	1	1	SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412354	CC	1094944979	INGENIERIA DE SISTEMAS Y COMPUTACION	ARMENIA	OCTAVIO ANDRES GOMEZ CORRALES	M	19/03/1994	1				SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412355	CC	1094984955	INGENIERIA DE SISTEMAS Y COMPUTACION	ARMENIA	DHAN STEVEN GONZALEZ VELASQUE	M	19/09/1996	1	1			UNION LIBRE
INGENIERIA DE SISTEMAS Y COMPUTAC	412356	CC	1006318949	INGENIERIA DE SISTEMAS Y COMPUTACION	CIRCASIA	VIVIANA GRIMALDO RODRIGUEZ	M	14/01/1997	1	1			SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412357	CC	1097726952	INGENIERIA DE SISTEMAS Y COMPUTACION	MONTENEGRO	BRANDON SMITH LASSO OREGO	M	24/09/1996	1				UNION LIBRE
INGENIERIA DE SISTEMAS Y COMPUTAC	412359	CC	1094003630	INGENIERIA DE SISTEMAS Y COMPUTACION	ARMENIA	JOLMAN ANDRES LOPEZ GARCIA	M	05/09/1992	1				UNION LIBRE
INGENIERIA DE SISTEMAS Y COMPUTAC	412360	TI	9708304888	INGENIERIA DE SISTEMAS Y COMPUTACION	LA TEBAIDA	LUIS DANIEL MARTINEZ RIOS	M	13/08/1997	1	1	1	1	SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412361	TI	9807256690	INGENIERIA DE SISTEMAS Y COMPUTACION	ARMENIA	MICHAEL ESTIVEN MENDIETA MEJIA	M	12/07/1998	1				SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412362	CC	1097039873	INGENIERIA DE SISTEMAS Y COMPUTACION	QUIMBAYA	FABIAN OROZCO GIRALDO	M	26/03/1997	1				SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412363	CC	1061794581	INGENIERIA DE SISTEMAS Y COMPUTACION	LEIVA	CAMILO STIVEN ORTIZ GOMEZ	M	09/10/1996	1	1	1	1	SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412365	TI	9707609940	INGENIERIA DE SISTEMAS Y COMPUTACION	ARMENIA	NICOLAS PALACIOS RIOS	M	18/07/1997	1	1	1	1	SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412366	TI	9804302642	INGENIERIA DE SISTEMAS Y COMPUTACION	MONTENEGRO	GUSTAVO ALBERTO PAVA RUIZ	M	24/04/1998	1	1			SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412367	TI	9805059905	INGENIERIA DE SISTEMAS Y COMPUTACION	CALARCA	ELMER ALEJANDRO PEÑA GUEVARA	M	19/05/1998	1	1			SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412368	CC	1097399984	INGENIERIA DE SISTEMAS Y COMPUTACION	ARMENIA	JOSE LUIS PEÑA PATIÑO	M	22/05/1993	1				SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412369	CC	1053625754	INGENIERIA DE SISTEMAS Y COMPUTACION	TUMACO	CARLOS EDUARDO QUIÑONES	M	27/09/1993	1	1	1	1	SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412370	TI	1002957389	INGENIERIA DE SISTEMAS Y COMPUTACION	CALARCA	NICOLAS STEPH CARVAJAL	M	05/02/1998	1	1			SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412371	CC	1093312081	INGENIERIA DE SISTEMAS Y COMPUTACION	CIRCASIA	MISAEL ANGEL RESTREPO SANCHEZ	M	18/12/1996	1	1			SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412372	CC	1097405217	INGENIERIA DE SISTEMAS Y COMPUTACION	CALARCA	JUAN ESTEBAN REYES ORTIZ	M	06/03/1997	1	1			SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412373	CC	1094958981	INGENIERIA DE SISTEMAS Y COMPUTACION	ARMENIA	DIEGO ALEXANDER PINCON MARIN	M	27/09/1996	1	1			SOLTERO(A)
INGENIERIA DE SISTEMAS Y COMPUTAC	412374	TI	9710801380	INGENIERIA DE SISTEMAS Y COMPUTACION	CALARCA	JUAN DAVID ROMERO GARCIA	M	08/10/1997	1	1	1	1	SOLTERO(A)

Fig. 4. Structure of the information provided by the planning and development department of the University of Quindío related to the personal and academic information of the students.

specific tasks. The task 1 allows us to determine the data will be included in the study and to exclude those with low quality. In this case the data related to the name and address of the students were excluded. Such data were classified as sensible information. The task 2 performs the cleaning process of the data starting from the selection criteria previously defined. The task 3 consists of the structuration of the data. In this step de tables for storing de data were structured. Finally, the task 4 allows us the integration of the data.

In previous phases we can see several sets of data which were provide by diverse and heterogeneous sources. So, the information presents great difficulties for its

adequate integration. For example, the Table 2 shows the structure and the possible values related to the data provided by the planning and development department.

Table 2. Data structure provided by the planning and development department.

Attribute	Value
CIVIL STATUS	SINGLE
ARMED_CONFLICT_VICTIM	Not
ES_DISPLACED	Not
ES_DISABLED	Not
STRATUM	4
SECURITY_REGIME	Contributory (E.P.S)
SISBEN_CATEGORY	Don't apply
AMOUNT_FAMILY_GROUP	4
FAMILY_NUCLEUS	two parents and brothers
CONTRIBUTE_TO_THE_FAMILY	Not
INCOME_QUANTITY	BETWEEN 0–1 MINIMUM SALARY
FINANCING_SOURCE	INCOME OF YOUR PARENTS AND/OR FAMILIES, EXCEPT HUSBAND OR PARTNER
WORK	Not
MOTIVATION	Because I'm interested in the comprehensive education offered by universities

As we can see such data as well as the data provided by the others dependencies, present several inconsistencies related to heterogeneous structure, missing values, duplicated records, and redundant information, among others.

In general all data provided by the different sources should be cleaned and integrated. All this process was carried out by using a set of macros in Microsoft Excel ®. Such macros were developed by the authors. Particularly, the information provided by such heterogeneous sources was imported in several excel tables. Then, we use several intermediate dynamic tables where the partial results were recorded. So, each of these tables were generated with Microsoft Excel ® by using macros, formulas and SQL sentences. Finally, the data resulting of the previous tasks were combined and integrated.

4.4 Modeling

In this phase the modeling techniques are selected and applied and their parameters are calibrated to optimal values. In this case we use the classification tree for relating the previously selected attributes. Such attributes were selected taking into account the level of incidence on the decision about the decision to desert or not to desert: BADyG test parameters, genre, age, marital status, victim of conflict, displaced, disabled, and stratum, etc. Such parameters can be consulted in Fig. 5a, b.

(a)

ESTP_ID	IG	RL	Rv	Rn	Re	Sv	Sn	Se	Ma	Mv	De	RA	EF
412345	97	47	16	15	16	15	12	23	21	18	15	143	68
412340	104	54	20	13	21	18	9	23	20	9	23	164	63
413207	59	32	17	8	7	12	0	15	20	13	11	136	43
415589	67	37	12	14	11	8	7	15	8	3	21	100	67
412350	46	24	12	4	8	11	4	7	10	9	17	120	38
412338	95	48	15	17	16	11	12	24	17	10	26	118	81
412334	73	44	14	16	14	12	2	15	23	12	25	104	70
412379	114	66	18	23	25	14	11	23	24	18	24	135	84
412352	57	26	12	8	6	12	7	12	12	11	22	127	45
412335	74	40	18	10	12	16	7	11	9	12	20	166	45
413208	85	50	23	15	12	18	7	10	21	10	17	143	59
412384	82	43	15	17	11	17	10	12	23	24	21	170	48
412456	101	66	21	22	23	13	15	7	18	19	15	135	75
412360	68	43	11	15	17	5	1	19	24	5	21	111	61
412336	71	36	13	13	10	10	7	18	26	17	18	114	62
412455	81	43	12	16	15	11	10	17	17	15	22	99	82
412344	26	8	3	1	4	8	2	8	8	12	14	92	28
412337	90	51	15	17	19	11	10	18	22	18	18	98	92
412347	77	39	18	8	13	17	2	19	17	11	28	159	48
415233	97	54	16	23	15	13	16	14	15	22	16	150	65
415232	65	33	8	16	9	12	8	12	12	17	24	127	51
412339	77	47	15	12	20	9	10	11	13	16	22	115	67
412351	71	38	17	9	12	14	4	15	14	16	20	129	55
412342	83	42	12	14	16	11	9	21	12	8	23	100	83
412454	60	34	12	11	11	10	7	9	6	13	16	112	54
412333	96	48	12	17	19	17	6	25	27	27	22	124	77
415308	58	23	11	0	12	15	6	14	26	14	17	93	62
412373	65	33	8	11	14	15	4	13	21	15	19	115	57
412459	85	53	18	14	21	12	5	15	20	15	22	140	61
412458	50	30	10	11	9	8	5	7	13	12	14	109	46

(b)

ESTP_ID	ESTADO_CIV	VICTIMA_COMFLI	ES_DESPLAZA	ES_DISCAPACIT	ESTRATO
412345	SOLTERO(A)	NO	NO	NO	2
412340	SOLTERO(A)	NO	NO	NO	3
413207	SOLTERO(A)	NO	NO	NO	2
415589	SOLTERO(A)	NO	NO	NO	1
412350	SOLTERO(A)	NO	NO	NO	1
412338	SOLTERO(A)	NO	NO	NO	6
412334	SOLTERO(A)	NO	NO	NO	1
412379	SOLTERO(A)	NO	NO	NO	3
412352	SOLTERO(A)	NO	NO	NO	3
412335	SOLTERO(A)	NO	NO	NO	2
413208	SOLTERO(A)	NO	NO	NO	1
412384	SOLTERO(A)	NO	NO	NO	2
412456	SOLTERO(A)	NO	NO	NO	2
412360	SOLTERO(A)	NO	NO	SI	1
412336	SOLTERO(A)	NO	NO	NO	1
412455	SOLTERO(A)	NO	NO	NO	2
412344	SOLTERO(A)	NO	NO	NO	3
412337	SOLTERO(A)	SI	NO	SI	2
412347	SOLTERO(A)	NO	NO	NO	1
415233	SOLTERO(A)	NO	NO	NO	2
415232	SOLTERO(A)	NO	NO	NO	3
412339	SOLTERO(A)	NO	NO	SI	1
412351	CASADO(A)	NO	NO	NO	1
412342	SOLTERO(A)	NO	NO	NO	3
412454	SOLTERO(A)	NO	NO	NO	2
412333	SOLTERO(A)	NO	NO	NO	2
415308	SOLTERO(A)	NO	NO	NO	3
412373	SOLTERO(A)	NO	NO	NO	1
412459	SOLTERO(A)	NO	NO	NO	2
412458	SOLTERO(A)	NO	NO	NO	1

Fig. 5. (a) Consolidated final data. (b) Consolidated final data.

4.5 Analysis and Evaluation

At this phase the model (or models) obtained are more thoroughly evaluated and the steps executed to construct the model are reviewed. Consequently, we present the results once the information was modeled, analyzed and evaluated. Such results are illustrated in Figs. 6, 7, 8, 9, 10 and 11.

Civil Engineering The decision tree in Fig. 6 shows that the attributes and their relationships which encourage a student for deserting from civil engineering program are the following:

Age and number of people composing the home. According to analyzed data for this academic program, the students with age between 19 and 21 years old, besides their family group has 3 or more members.

Regarding BADyG test, the most significant attributes which encourage a student for deserting from this academic program, are the following:

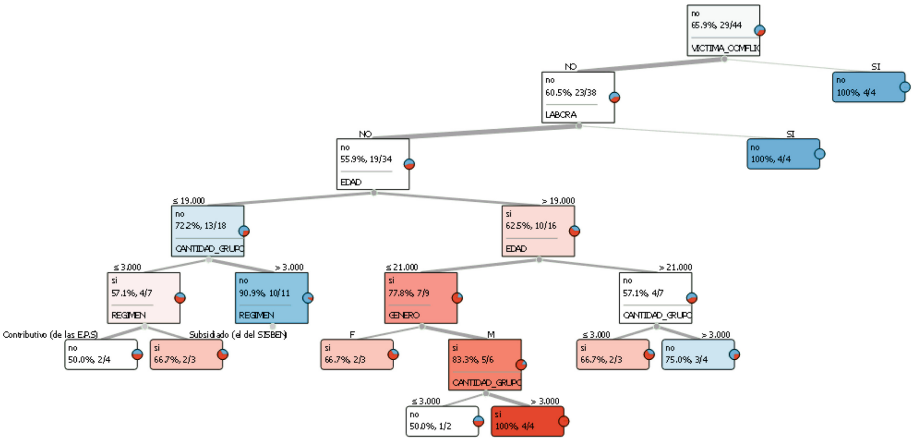

Fig. 6. Civil engineering.

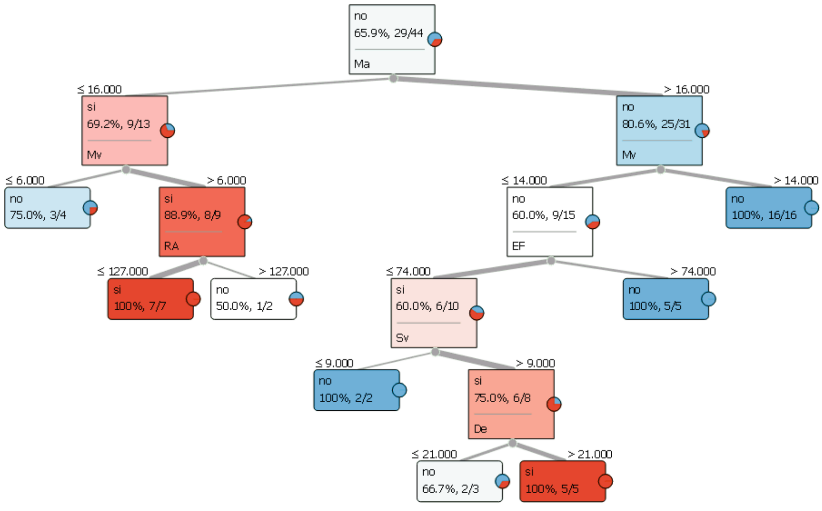

Fig. 7. BADyG test at the civil engineering program

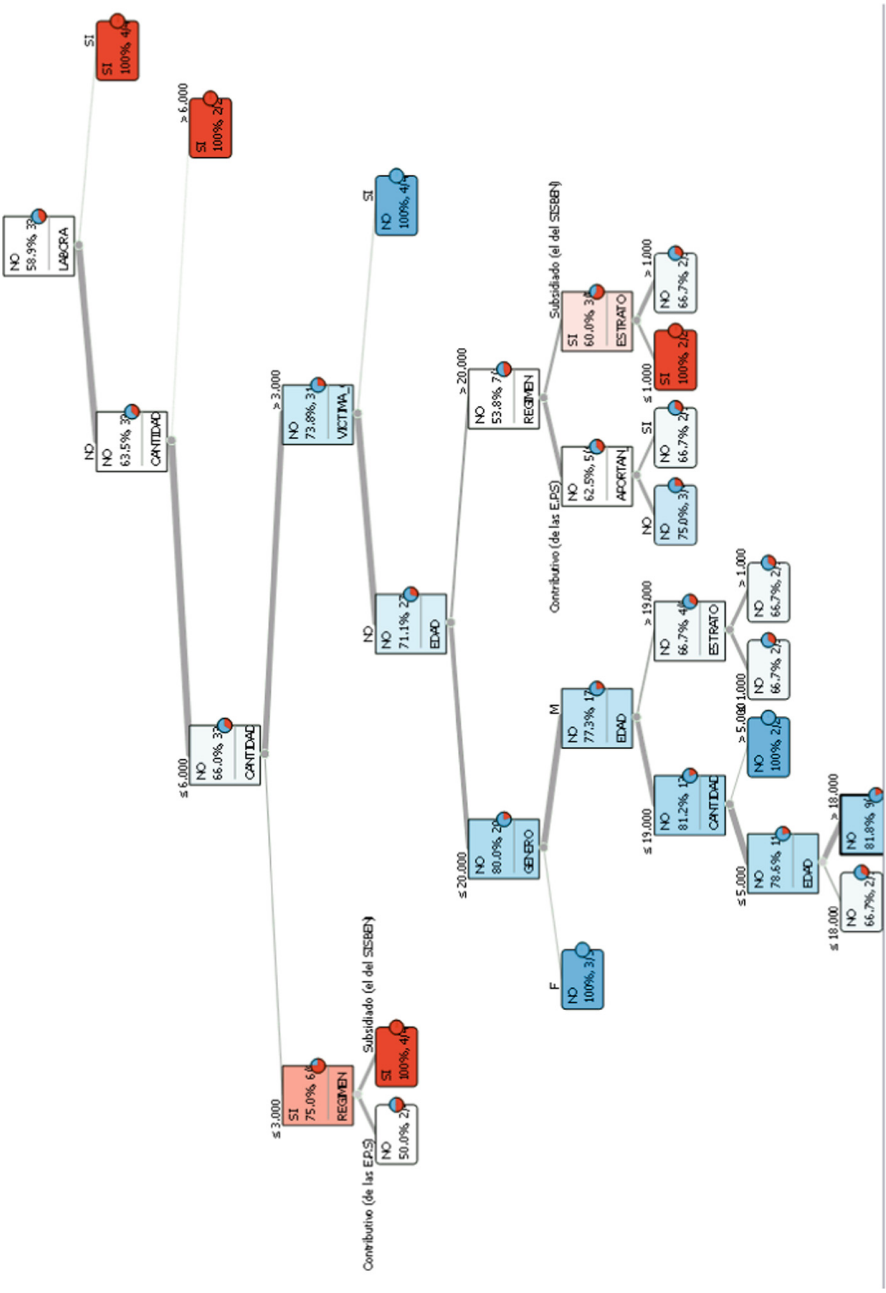

Fig. 8. Electronic engineering

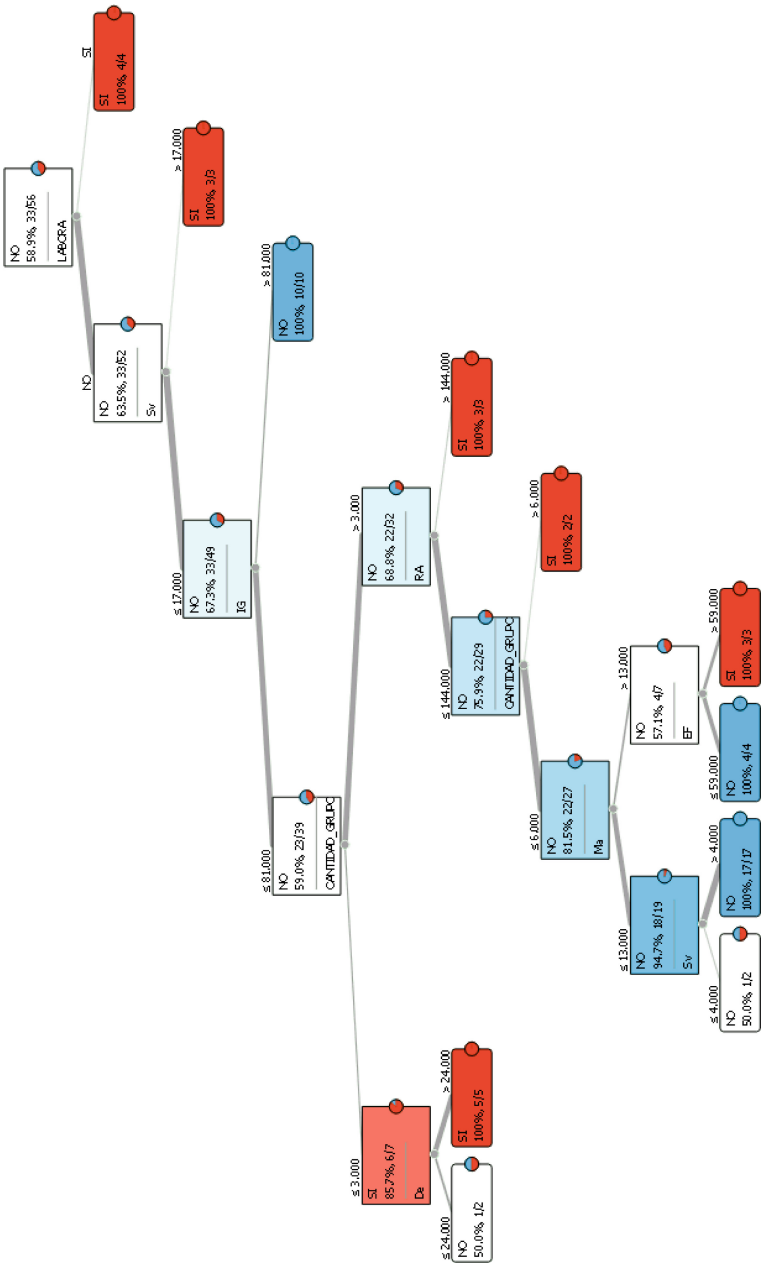

Fig. 9. BADyG test at the electronic engineering program

MV (Visual memory): this aspect is one of the most relevant since it represents 69.2% percentage of those who deserted and had low score on this aspect.

RA (Speed): the speed to process information is other relevant aspect. Such aspect is a complement of MV, which makes it very difficult for students to learn on this academic program.

Reading comprehension also plays an important role in the determination to desert or not desert. As can be seen, the SV (Complete sentences) and DE (Discrimination of differences) aspects have a high percentage of students who deserted at this program.

Electronic engineering. The decision tree in Fig. 8 shows that the attributes and their relationships which encourage a student for deserting from electronic engineering program are the following:

A significant weight is associated to the aspect determining if the student work or not. According to the data results provided by the decision tree, the students who study and work at the same time have a high probability for deserting. On the other hand, students who do not work depend on the attributes: number of family members, age, and stratum. Thus, students with stratum 1 and older than 20 trend to desert on the first 3 semesters.

Regarding BADyG test, the most significant attributes which encourage a student for deserting from this academic program, are the following:

SV (Complete sentences) and RA (Speed) demonstrate that students have problems with the reading comprehension, which represents a high probability for deserting.

Systems and Computing Engineering. The decision tree in Fig. 10 shows that the attributes and their relationships which encourage a student for deserting from systems and computing engineering program are the following:

The number of family members, age, regime, contribute or not, and stratum. According to the Fig. 8, those students who desert economically contribute and have between 4 and 6 family members. Another interesting situation is related to the students who have between 1 and 3 family members and their regime is contributory (EPS, Empresa Prestadora de Salud). The same happens with the students who have ages between 19 and 22 years old, have between 1 and 4 family members and have stratum 1. Finally, students who are under 19 years desert when the stratum is 2 and have between 1 and 3 family members.

Regarding BADyG test, the most significant attributes which encourage a student for deserting from this academic program, are the following:

Students who have a score under 68 in EF (Effectiveness), a score under 19 in Mv (Visual memory), a score under 74 in IG (General intelligence), and have between 1 and 6 family members have a high probability for deserting. Likewise, students with high probability for deserting have a score under 57 in EF (Effectiveness), a score under 6 in Sn (Numerical problems), a score under 74 in IG (General intelligence), a score greater than 6 in Rv (Analog relationships), a score under 68 in Mv (Visual memory), and a number of family members between 1 and 6.

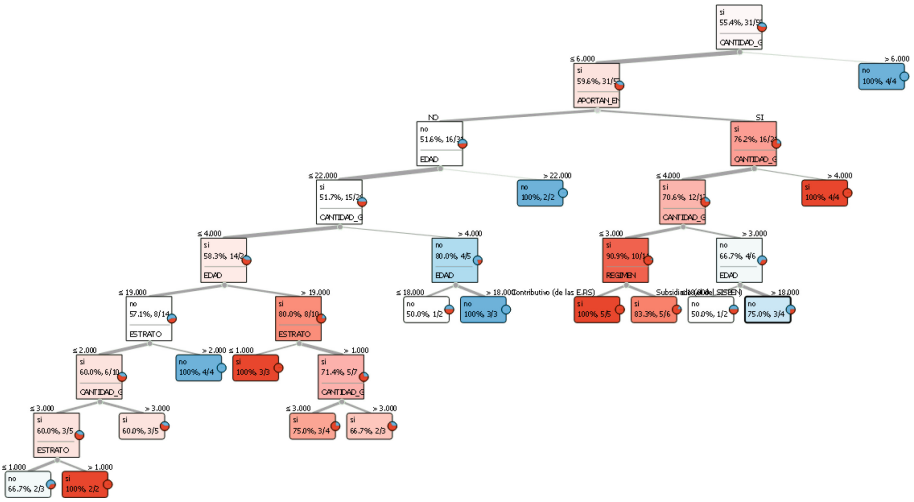

Fig. 10. Systems and computing engineering

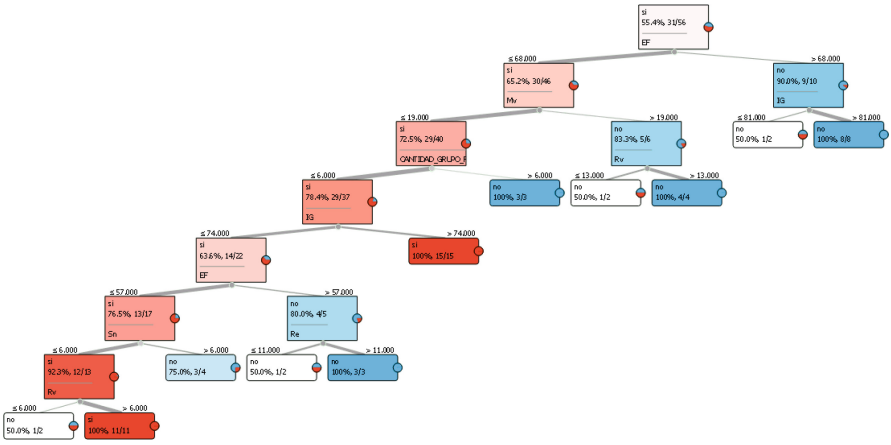

Fig. 11. BADyG test at the systems and computing engineering program

5 Conclusions

In this paper a method for analyzing the student attrition in the faculty of engineering at the University of Quindío was proposed. Thus, an explanation of the process was described by using the phases provided by the methodology CRISP-DM. This methodology favors the obtaining of results because it allows us the complete understanding of the problem in terms of the business and its meaning in terms of data mining.

The advantage of this work is that it takes into account a group of aspects related to intelligence tests. This was achieved by analyzing the data obtained from the BADyG (Battery of General and Differential Aptitudes) tests. The results of this work reflect the need of the university to establish action plans that improve the performance of students. Such as extracurricular advice and accompaniment to strengthen the flaws found. This strategy can be applied even in other universities.

Some inconveniences were found with the application of university policies to determine the status of a student deserter. Which were not very clear or according to reality. Consequently, as future work it is necessary to develop additional studies to improve the classification of deserter students.

References

1. Argote, I., Jiménez, R.: Detección de patrones de deserción en los programas de pregrado de la Universidad Mariana de San juan de Pasto, aplicando el proceso de KDD y su implementación en modelos matemátios de predicción. In: Cuarta Conferencia Latinoamericana Sobre El Abandono En La Educación Superior, pp. 1–7 (2016)
2. Azoumana, K.: Análisis de la deserción estudiantil en la Universidad Simón Bolívar, facultad Ingeniería de Sistemas, con técnicas de minería de datos. Pensamiento americano, 41–51 (2013)
3. Castaño, E., Gallón, S., Vásquez, J.: Análisis de los factores asociados a la deserción estudiantil en la Educación Superior: un estudio de caso. Revista de educación **345**, 255–280 (2008)
4. Cruz Ordiérez, D.I., Ortega Calpa, J.F.: Análisis de la deserción estudiantil en la facultad de Ciencias Exactas de la Universidad de Nariño utilizando el sistema SPADIES (Tesis de pregrado). Univerisdad de Nariño, San Juan de Pasto, Colombia (2008)
5. Eckert, K., Suénaga, R.: Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos. Formación universitaria, pp. 03–12 (2015)
6. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The KDD process for extracting useful knowledge from volumes of data. Commun. ACM **39**, 27–34 (1996)
7. Gallego Gallego, M.: Descubrimiento de conocimiento en una empresa de outsourcing de TI de la ciudad de Medellín, aplicacndo técnicas de minería de datos que permita identificar potencialidades en el éxito de los proyectos de desarrollo de software. (Tesis de maestría). Universidad Autónoma de Manizales, Manizales, Colombia (2014)
8. González Cardona, J.C.: Sistema de apoyo para la acreditación de la calidad de programas academicos de la universidad de caldas, aplicando técnicas en minería de datos (Tesis de maestría). Universidad Autónoma de Manizales, Manizales, Colombia (2011)
9. Gutierrez, J.E.: Descubrimiento De Conocimientos En La Base De Datos Académica De La Universidad Autónoma De Manizales Aplicando Redes Neuronales (Tesis de maestría). Universidad Autónoma de Manizales, Manizales, Colombia (2012)
10. Cáceres, J.H.: Descubrimiento de conocimiento en la base de datos académica de una institución de educación superior usando redes neuronales. Vector, 7–19
11. Linoff, G., Berry, M.: Why and What is Data Mining? En Data Mining Techniques, pp. 1–10. Wiley (2004, 2011)
12. Páramo, G.J., Correa, C.A.: Deserción estudiantil universitaria Conceptualización. Revista Universidad EAFIT **114**, 65–78 (2012)

13. Salazar, A., Gosálbez, J., Bosch, I.: A case study of knowledge discovery on academic achievement, student desertion and student retention. Inf. Technol.: Res. Educ. **2004**, 150–154 (2004)
14. Toro, D.F.S.: Tendencias y características de los viajeros que visitan la ciudad de Pereira por medio de técnicas de minería de datos (Tesis de maestría). Universidad Autónoma de Manizales, Manizales, Colombia (2015)
15. Sotomonte Castro, J.E., Rodríguez Rodríguez, C.C., Montenegro Marín, C.E., Gaona García, P.A., Castellanos, J.G.: Hacia la construcción de un modelo predictivo de deserción académica basado en técnicas de minería de datos. Revista Científica **3**, 35–48 (2016)
16. Pereira, R.T., Toledo, J.J.: extracción de perfiles de deserción estudiantil en la institución universitaria cesmag. Ivestigium IRE: Ciencias Sociales y Humanas, **6**, 30–44 (2015)
17. Yuste, C.: Batería de Aptitudes Diferenciales y Generales. CEPE, Madrid (2001)
18. Diana, G.M., García González, M.D., Hurtado Tobón, L.H., Sánchez Botero, C.E.: Estudio sobre la Deserción Estudiantil en la University of Quindío. Armenia, Quindío: Grupo de Investigación y Asesoría en Estadística (2007)
19. Chapman, P., et al.: CRISP-DM 1.0. Step-by-step Data Mining Guide (2000)
20. Santhanakumar, M., Columbus, C.C.: Web usage based analysis of web pages using rapidminer. Wseas Trans. Comput. **14**, 455–464 (2015)