International Education Studies; Vol. 12, No. 6; 2019
ISSN 1913-9020 E-ISSN 1913-9039
Published by Canadian Center of Science and Education

Clusters That Determine the Risks to University Desertion in Santander

Omar Millán Delgado[1]

[1] Engineering Department, Environmental Engineer Program. Universidad Libre Seccional Socorro. Colombia

Correspondence: Omar Millán Delgado, Universidad Libre Seccional Socorro. Socorro, Santander, Colombia. Tel: 57-31-6865-6784. E-mail: omar.millan@unilibre.edu.co

Received: October 25, 2018 Accepted: December 3, 2018 Online Published: May 29, 2019

doi:10.5539/ies.v12n6p17 URL: https://doi.org/10.5539/ies.v12n6p17

Abstract

The phenomenon of university desertion is the result of an inefficient system of the entire Colombian national educational structure. Under a predominantly descriptive, quantitative and statistical approach, it is intended to identify the behavior of the desertion in the universities of the department of Santander, with the purpose of knowing their state of control. The information collected in the present study, based on the application of structured instruments, is the input for the design of statistical models, where the dependent variable is highlighted as the intention of desertion risk and the explanatory variables are the rest which are grouped by means of clusters independently and not correlated; and in the end, identify their degree of influence on the proposed mathematical model. This model identifies the variables that most influence the risk of desertion, and the model obtained is the basis for the proposal of strategic activities, to the extent as these obey the variables that most influence the risk of desertion. In this way, the concentration of efforts on specific aspects through this model implementation is expected, and the maximization of effectiveness is achieved through the reduction of desertion rates in the department of Santander.

Keywords: clusters, desertion, mathematical model, regression, variables

1. Introduction

The determinants of desertion, which for this study have been classified initially in Individual, Academic, Institutional and Socioeconomic, have already been addressed in previous studies by authors such as Cabrera et al. (2006), Castaño et al. (2004), Diaz (2008), Ethington (1990), Spady (1971), Tinto (1975), Zamora (2010), among others, who have breed the theoretical, conceptual and methodological bases for the development of this project. In the same way, it is worth mentioning the background of the contribution obtained by the Ministry of National Education-MEN in 2016, through the System for the Prevention of Desertion of Higher Education-SPADIES. This system generates information on variables intrinsic to withdrawal, identifies the levels of risk of it of the students of the Higher Education Institutions-HEI in Colombia, and based on it, establish programs and strategies of institutional and state nature to reduce academic desertion.

With this research, it is proposed to formulate a methodology of a strategic nature that will counteract the desertion of HEIs in the department of Santander - Colombia. In a particular way, it is sought to identify the variables that constitute each of the factors that generate the current desertion levels of these institutions, and to determine the interactions of each of the Individual, Academic, Institutional and Socioeconomic factors that cause this issue in the HEIs, by means of a model that explains their dynamics. In addition to the above, it is intended to formulate strategic activities that are in accordance with the determination and identification of relevant variables, according to the identified model, in order to reduce the desertion levels in the department of Santander.

The present research was developed in municipalities with presence of Higher Education Institutions including universities and their undergraduate programs with accessibility to databases for consultation. The approach is of a mixed type predominantly quantitative and descriptive resorting specific techniques for the collection of information, such as the questionnaire based on probabilistic sampling, as well as in the search of secondary information relevant to the topic. This research is also of a correlational quantitative type, of a non-longitudinal transversal order.

The contribution of this research lies in the identification of the intrinsic relationship of the variables of the factors as determinants of the desertion in the Department of Santander. With this, it is possible to identify how each of them contributes to the student's decision to defect from the Higher Education Institution. It uses statistical

techniques as a theoretical and methodological basis that imprints its scientific relevance. Once the relationship has been identified by means of clusters and the model validated, strategic activities and plans that favor the reduction of dropout rates in the analysis region are formulated.

2. Literature Review

2.1 Academic Desertion

Several authors have managed to define the concept of school dropout as a critical variable for Higher Education Institutions. It refers to the "early abandonment of a program of studies before reaching the degree, and considers a time long enough to rule out the possibility of the student rejoining" (Himmel, 2002, p. 93). In this way, several types of desertion are proposed; the voluntary desertion that is not another thing that the resignation to the academic program or the retirement of the HEI, and the involuntary desertion, due to the institutional decision that makes of the student present his or her resignation. Each type becomes more complex as it is deepened and analyzed (Figure 1).

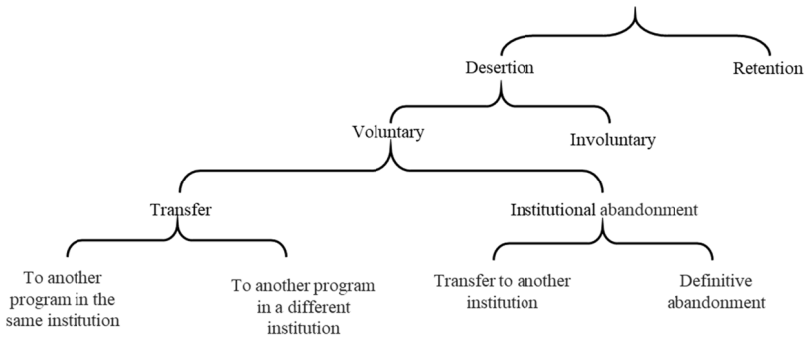

Figure 1. Classification of the desertion (Himmel, 2002)

In another side, and taking the definitions of the authors Giovagnoli (2001), Spady (1971), Tinto (1982), and in the report of the Ministry of National Education (2009), the desertion is defined as that eventuality that must be faced a student when he aspires and fails to complete his educational project; a deserter is one who, being a student of a Higher Education Institution, does not present academic activity during two consecutive academic semesters (MEN, 2009).

Accordingly to Tinto (1989), the definition of student desertion is analyzed according to the different types of abandonment. The analysis depends on the stakeholders and those who are part of the process such as students, official HEI and those responsible for the national education policy. A measure of the dropout rate may be the number of students who leave a Higher Education Institution in a given period, before having obtained the corresponding degree (Colombian Institute for the Promotion of Higher Education-ICFES, 2002). Under this approach, the concept of desertion includes the abandonment of the student from the education system in general and all students who withdraw from a HEI can be defined as deserters (Castaño et al., 2004).

2.2 Multivariate Analysis of Main Components

The numerical analysis of information, which is required to be achieved in some fields of knowledge, is developed through statistical calculations. Nowadays, it is feasible to have methods that provide new possibilities for quantitative treatment which in some way, would not be possible to perform with traditional uni and bivariant procedures. In this way and according to Closas et al. (2013), the methods that gather a series of data analysis are useful to develop studies as for dependence, as well as for interdependence between the different variables which are part of the statistics area of multivariate analysis (Anderson, 1984). The beginnings of statistical techniques of multivariate analysis date from the use of linear regression thanks to its creator Gauss (1809) and then, by Markov in 1900. The most current techniques originated in the decade of the 30's (Closas et al., 2013).

On one side, there are explanatory or dependency techniques which research for the presence of relationships between two or more groups of variables. In case that these groups are classified in dependent and independent variables, the mission of the dependent techniques will be to establish if the independent variables conjunction affects the set of dependent variables in an integrated manner or individually. On the other side, when it is not possible to clarify between this dependency or independence of variables, and what matters in what form and why

the variables are correlated with each other; the statistical methods of interdependence are used (Closas et al., 2013). The interdependence analysis techniques are those that are observed in Figure 2.

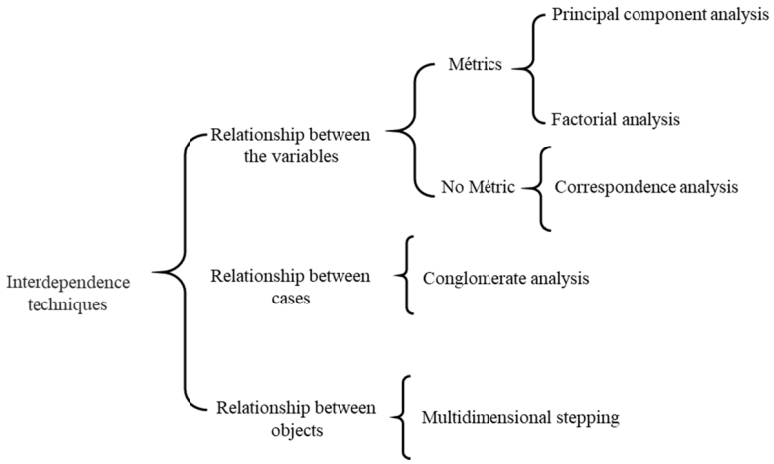

Figure 2. Descriptive or interdependent techniques (Closas et al., 2013)

The Principal Component Analysis (PCA) technique has its origins in orthogonal least-squares adjustments, introduced by K. Pearson. Its use allows:

1) To present in a space of small dimension, observations of a general p-dimensional space. The main components are the first step to identify possible variables, latent or not observed which are generating the variability of the data.
2) To transform the original variables, generally correlated, into new uncorrelated variables, facilitating the interpretation of the data (Peña, 2002).

Citing Castaño (2010), the object of this technique is to: "explain the structure of the covariance matrix of a set of variables by means of a few linear combinations of the original variables "(Castaño, 2010, p. 10). Its objective is to achieve data reduction and facilitate interpretation. In this same way, the PCA is a statistical technique of synthesis of information, or reduction of the dimension (number of variables). By having a data set with many variables, the pursued objective is to reduce them to a smaller number by losing as little information as possible (Hotteling, 1933). A central problem in the analysis of multivariate data, such as that which concerns the problem of desertion, is precisely the reduction of its dimensionality. This is:

"Describe with precision the values of the variables by a small subset r<p of them, decrease or reduce the dimension of the problem at the expense of a small loss of information" (Montanero-Fernandez, 2008, p. 20).

Although the p major components are needed to reproduce all the variability of the system, most of that variability is supported by a small number r of major components. In these cases, the r main components replace the p original variables, thus reducing the original system. As a complement, the new main components will be a linear combination of the original variables, and they will also be independent from each other. A relevant factor in this technique is the interpretation of the factors, since this is not given in empirical form, but is given once the relationship of the factors with the initial variables is clarified; this technique is not entirely easy, if you do not have prior knowledge of the subject in question (Terradez, w.d.).

2.3 Clusters Analysis

Another method of multivariate statistical analysis of great relevance for the purpose of the present research is the Cluster Analysis; this is an exploration tool designed to discover natural groupings (or conglomerates) within a data set. It is very useful when you want to group a small number of elements (Cerpa, Castillo, & Cantillo, 2015, p. 89). The purpose of Cluster Analysis is to classify the variables or elements in almost homogeneous groups, based on a group of classificatory variables. All the variables, without distinction, are analyzed among themselves, whether they are dependent or independent (Arriaza-Balmón, 2000, p. 173). If these classes are grouped in succession in other higher-level classes, the result is a conglomerate structure or clusters, which can be represented graphically by means of dendrograms (Mardia et al., 1979), (Amparo & Grane, 2008). Accordingly to Peña (2002),

the analysis of clusters or conglomerates presents three phases:

1) Data partitioning. Data is available which are suspected of being heterogeneous and are wanted to be divided into a set number of groups, so that each element is linked to a single group; all the elements are classified, and each group is homogeneous.
2) Hierarchy construction. It seeks to organize the elements of a set in a graduated way by its analogy. This implies that the data are arranged in levels, so that the higher levels contain the lower ones.
3) Variable classification. According to Peña, (2002), when many variables are present, an initial exploratory study is recommended, which guides the approach of the formal models, dividing the variables into groups and reducing the dimension.

3. Methodology

3.1 Pilot Survey for the Validation of the Tool

To obtain the validation of the final tool, Cronbach's Alpha statistic was applied. Through the pilot test, which was made as a prerequisite for the validation of the tool, it was possible to apply a total of 32 surveys to university students in different municipalities of the Department of Santander. This statistic generated a value of 0.903, which guarantees a high degree of validity of the tool. The results of the test can be observed in Table 1.

Table 1. Reliability statistics results, on SPSS® base

Reliability statistics	
Cronbach's Alpha	N of elements
0.903	32

The tool, once modified, according to the scales in the questions that were relatively uniform and that showed an adequate level of reliability, was applied to the group of students that generated the sample in the different municipalities identified as having HEI. Based on the results of the surveys that were applied to university students in the municipalities, where academic programs of this level are offered, according to the sampling process, the corresponding coding and transformation were carried out in order to achieve an adequate escalation and with it to be able to develop calculations in the statistical tool SPSS. The standardization of data was a tool used with the purpose of avoiding the sensitivity of the different units of the variables, since variables of scale order of Likert, dichotomous, numerical without scaling in particular were used, which gives the possibility of altering the conformation of clusters and their relationships according to the measure of each variable.

3.2 Sampling

Based on the objective of the research, a type of sampling was developed:

Probability sampling: applicable to students of the universities of the Department of Santander. For this purpose and taking into account that at the moment the Department has a number of 72.702 students in face-to-face mode in public and private universities (MEN, 2016), the calculation of the respective sample was done, according to the results of Table 2.

Table 2. Calculation of the probabilistic sample

Moment (2017)		
Groups of interest	Universities of Santander	
	Population	Sample
Students	72.702	150

Source: Own elaboration, on Excel software base. Statistics for the calculation: Area under the Normal distribution curve of 95% of two-tailed reliability, corresponding to a Z-score of 1.96. N° total population: 72.702. Probability of success of the event (P), equal to the probability of failure of the event (Q), for this type of calculations and based on its ignorance, a value of 50% is chosen (P = Q = 0.5). Estimated error, whose difference between the population parameter and the statistic of the sample is 0.08.

The proportional distribution of students by municipalities is based on the information of students enrolled in the second semester of 2016 that reflects the National System of Higher Education Institutions-SNIES according to

Table 3.

Table 3. Calculation of the probabilistic sample by municipality

Municipality	Enrolled students*	Proportional distribution
Bucaramanga	66.101	135
Barrancabermeja	2316	5
Floridablanca	1.598	3
San Gil	631	3
Socorro	2.056	4
Total	72.702	150

Source: own elaboration base on the information of the SNIES, 2016. *: Base year of the information 2012.

3.3 Cluster Obtainment

The intention of the conformation of clusters lies on being able to determine the form through which, based on the variables (questions of the questionnaire), it is possible to form other groups, which may be the same or different from those proposed by SPADIES; in this case, taking into account that the number of variables is considerably higher (Montanero-Fernandez, 2008), it is probable that clusters or groups are achieved that are different in variables and in number, to which are present in SPADIES (MEN, 2016) . The decision to form clusters depends on where the different groups of variables can be identified. The "cut" or starting line defines them. The intention is to be able to group the variables in clusters that, on the one hand, are not the same variables, but that on the other hand, may present a certain level of discrimination or grouping (Crámer, 1968).

3.4 Measuring Cluster Correlations

Once the median parameter for each cluster was obtained as a non-parametric test, due to the intrinsic variables in each one of them; that is, qualitative binary variables and of Likert scale, among others, and based on the initial model proposed, it has been decided to implement the median as a correlation statistic (Morales-Vallejo, 2013). The usual measures of association between the variables that have been established in the research are the covariance and the correlation. These measures take into account only the linear relationships. The intention of this process is to determine low correlations between the established clusters, thus demonstrating the formation of clusters independently of the others. For this it is proposed that:

Null hypothesis: H_0: There is no association between the pairs of ranges

Alternative hypothesis H_1: There is association between the pairs of ranges

The test is done two-tailed. Significance level: 5%

3.5 Obtainment of the Model for the Identification of Risk of Desertion

Once the clusters with their respective associated variables have been consolidated, and the correlation between them is determined, the questionnaire question is taken as a basis. "What is the probability that you, in the next semester, take the decision of not to continue with your university studies, either partial or definitive?" "Which does it raise the intention of the student's desertion?" For each of the clusters, the medians were built according to the cases. In this way, the question is constituted in the dependent variable of the whole process tending to identify the degree or risk of desertion and the other variables under the calculation of their median for each cluster are constituted in the regression variables. This is how you can determine the model under the following proposal:

$$Risk\ of\ desertion = Median\ (Cluster\ 1 + Cluster\ 2 + Cluster\ 3 + \cdots + Cluster\ 10) \quad (1)$$

Or which is the same:

$$Question\ 15 = Median1 + Median2 + Median3 + \cdots + Median10 \quad (2)$$

This procedure ensures the possibility of identifying the statistic based on the different response options presented by the question in mention. That is, five tables with the respective medians were obtained, based on the five response options of the question in reference. It would be expected that the results of this procedure show some tendency in the behavior that indicate signs of risk in the desertion. The graphs presented below show the tendency of each cluster in each answer option on the question previously stated.

3.6 Correlation Analysis of the Proposed Model

According to the proposed model, it is assumed that:

$$Risk\ of\ desertion = Median\ (Cluster\ 1 + Cluster\ 2 + Cluster\ 3 + \cdots + Cluster\ 10) \quad (3)$$

This premise entails a degree of intrinsic correlation. It is then necessary to determine the existing correlation between the independent variables, that is, the conformed clusters and the dependent variable, question 15 of the applied questionnaire. This stage is required for the mathematical design that is desired to propose. The purpose of the same is to identify a risk interval for each answer option of question 15. For this, it was preceded according to the following methodology:

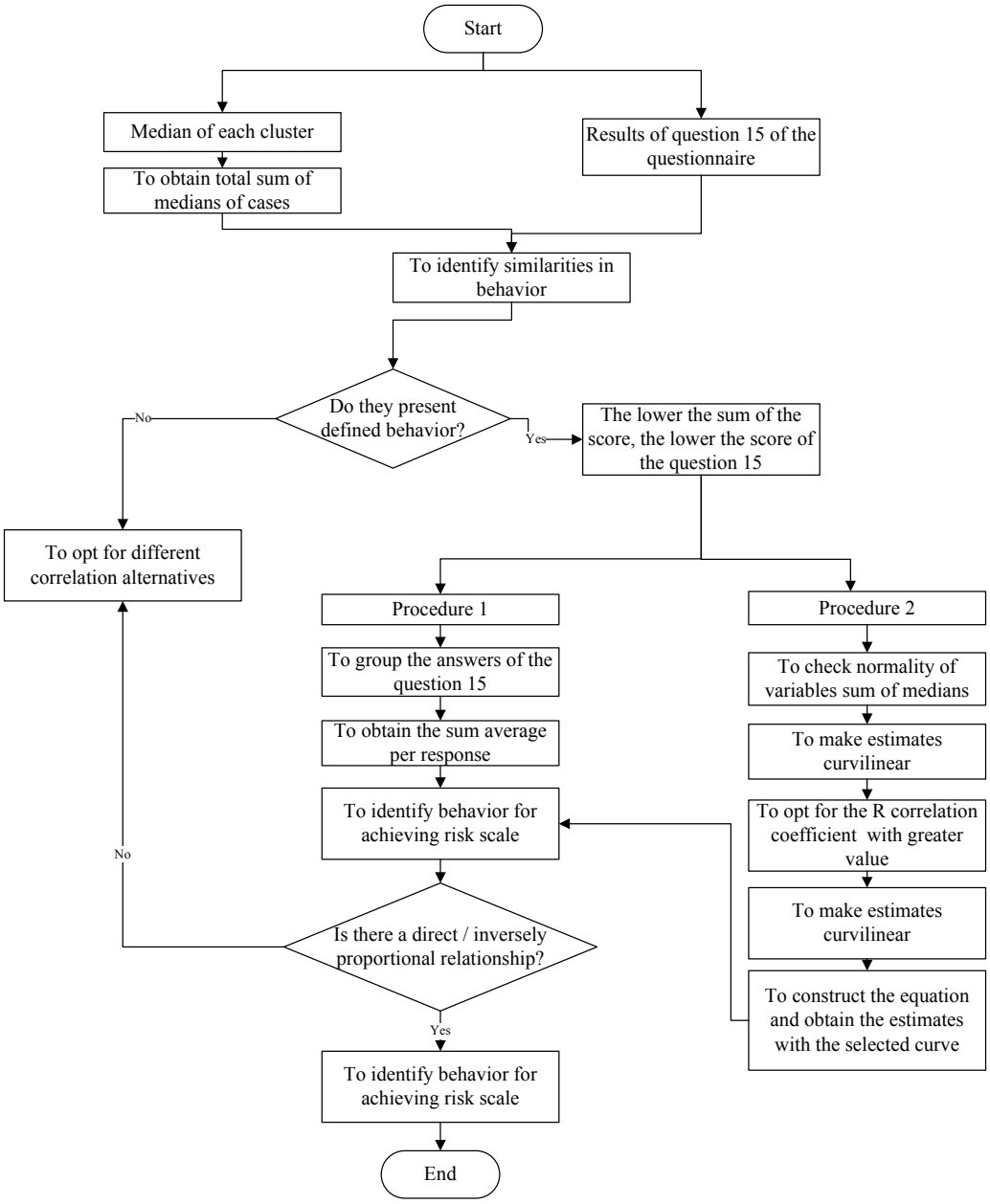

Figure 3. Procedure for the identification of correlation of the proposed desertion model (Own elaboration)

4. Results

4.1 Cluster Conformation

The dendrogram presented below shows the possibility of groupings in variables with a large number among some groups and others with only one variable. On this disjunctive, it was chose to group them according to the cut shown with vertical line present in the Figure 4.

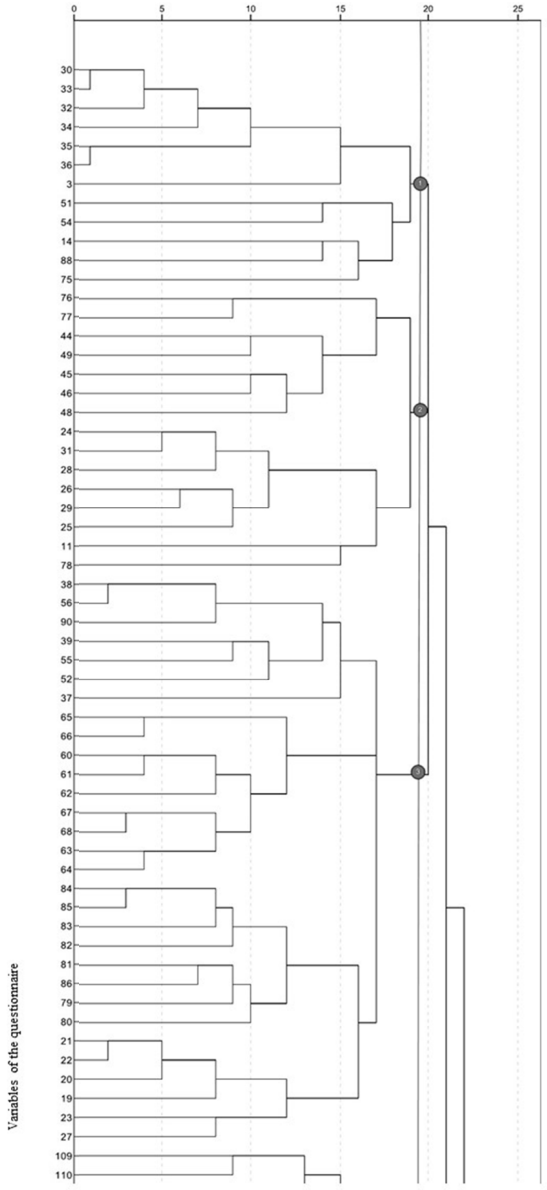

Figure 4. Dendrogram - clusters by variables, on SPSS® base (Own elaboration)

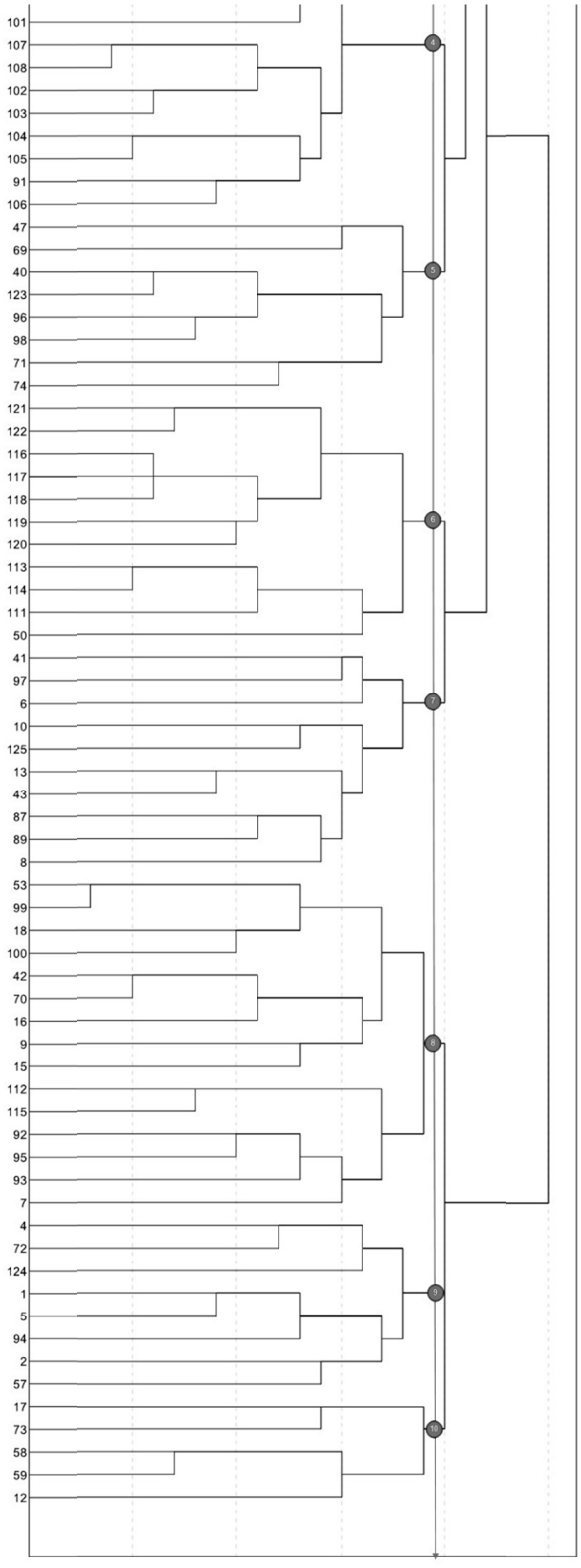

Figure 4. Dendrogram - clusters by variables, on SPSS® base (Own elaboration)

Based on the graph, it is perceived that the groups do not necessarily constitute the four traditional groups that are established at the moment of speaking about university desertion. The large number of established variables, their

correlation and interdependence can generate and in fact generate different clusters to the four traditionally known groups.

Cluster 1 can be explained by the grouping of intrinsic and allusive variables to the process of studying itself; the assessment and recognition of the student in his effort to study the studying habits, financial support received. It is observed that academic, institutional and individual variables are present in it. Cluster 2 takes into account variables such as the student's capacity in their learning process, influential variables in the type of student and the direct costs of study, accommodation, food and transportation. It integrates academic, individual and institutional aspects.

Cluster 3 presents variables that have elements in common and integrate elements such as the student's skills and abilities, satisfaction and integration with career, teacher's influence on performance and intrinsic aspects of the career. They are especially academic and individual variables. Cluster 4 establishes in its group the reasons for deserting the career. Variables of different nature allusive to the probability of deserting school. Cluster 5 involves variables of socioeconomic order (social stratification, income), studies of parents, and financing of the career. In Cluster 6, the reasons for admission to the program are grouped, and consequently, the aspects that may influence the decision to desert. Cluster 7 makes a conglomeration of variables of an individual order (number of people in the home, vocational orientation, orientation to enter the program), socioeconomic order (type of housing, parents' work) and institutional order (institutional support received).

In Cluster 8 variables of academic order are grouped, such as the type of university, last period attended, employment history, subjects studied and not passed, aspects that can influence the desertion, as well as the aspects that influenced the decision of entrance. Cluster 9 contemplates variables of individual order, among them, age, children, spouse, place in ICFES tests, time to work, and mother's work. Cluster 10 integrates the aspects that influence the decision to enter and sustain the career, and academic aspects; grade point average at the end of high school and accumulated in the career.

4.2 Non-Parametric Measures of the Clusters

Based on the criteria for the hypothesis test, the medians of each cluster as a representative statistic and subsequently were determined, and the nonparametric correlation through the Rho or Spearman correlation coefficient was calculated, equivalent to the Pearson correlation coefficient for parametric tests. The results of the correlation are presented in the following matrix (Table 4).

Table 4. Spearman correlations for the 10 medians of the clusters

Rho of Spearman		MED1	MED2	MED3	MED4	MED5	MED6	MED7	MED8	MED9	MED10
MED1	Correlation coefficient	1	0.037	0.321	-0.079	0.069	0.02	0.19	-0.112	0.019	0.083
MED2		0.037	1	-0.02	0.02	-0.018	0.105	0.121	0.018	-0.066	0.228
MED3		0.321	-0.02	1	0.221	-0.007	-0.225	-0.029	-0.118	-0.105	0.032
MED4		-0.079	0.02	0.221	1	-0.056	-0.048	-0.233	0.042	-0.067	-0.137
MED5		0.069	-0.018	-0.007	-0.056	1	0.096	0.063	-0.025	-0.054	0.06
MED6		0.02	0.105	-0.225	-0.048	0.096	1	0.194	0.112	0.066	-0.159
MED7		0.19	0.121	-0.029	-0.233	0.063	0.194	1	0.041	0.223	0.002
MED8		-0.112	0.018	-0.118	0.042	-0.025	0.112	0.041	1	0.02	-0.007
MED9		0.019	-0.066	-0.105	-0.067	-0.054	0.066	0.223	0.02	1	-0.033
MED10		0.083	0.228	0.032	-0.137	0.06	-0.159	0.002	-0.007	-0.033	1

Source: own elaboration with SPSS Statistics® 19 software.

Note. The color code used in the correlation table is related to the following convention (Peña, 2002).

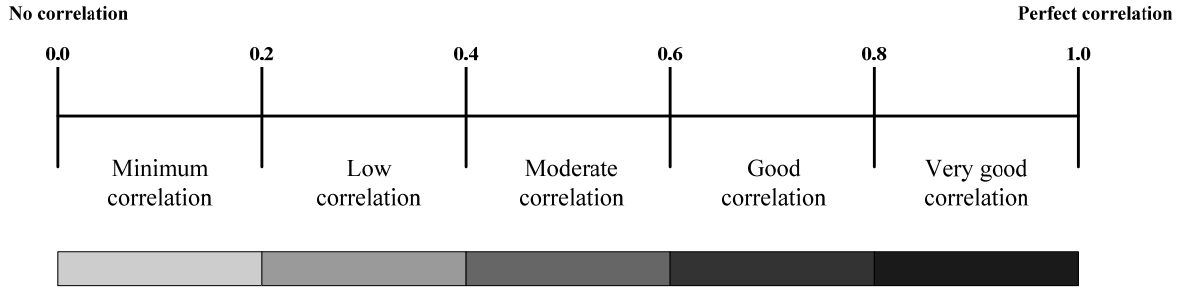

Figure 5. Color code for the Spearman correlation scale

From the previous table it can be deduced that most of the clusters present a correlation that oscillates between minimum and low scales, a fact that favors the process of consolidation of independence of clusters and their respective variables. Based on these results, the statistical method was implemented to achieve the basic mathematical model that identifies the risk levels of the dropout of university students in the Department of Santander.

4.3 Model for Identification of Risk of Desertion

At the moment of implementing procedure 1 of the methodology proposed in item 3.5, figure 3, the sum of the medians of the 10 clusters was generated. Subsequently, the average of the sums of the medians was obtained, which coincided with each of the response options to question 15. For example, for answer option 1, an average of medians with a value of 19.6 was obtained. For the construction of the intervals in the risk scale, the lower limit of the first interval was identified as starting data, based on the minimum value of the sum of averages of the response with option 1; this is 14. As can be seen, when developing the methodology proposed under procedure 1, results were found that do not coincide with the levels of risk sought; this is, for example, cases of desertion risk were found with averages of sum with values of 19.5, in scale 1 (from very low to low), as well as values of 18.5 in medium-high risk scales, what does not coincide with the suggested scales or with a logical order. The detail is presented in the following table.

Table 5. Scales of risk to the desertion under the methodology of sum of averages of the answer options to question 15 of the instrument

		OPTIONS QUEST 15	SUM AVERAGE	RISK SCALE	CRITERIA
Procedure 1	VERY LOW	1	19.6	From 14 to 19.6	VERY LOW – LOW
	LOW	2	19.2		
	MEDIUM	3	19	From 18.4 to 19	MEDIUM-HIGH
	HIGH	4	18.4		
	VERY HIGH	5	20.1	From 20.1 to 25.5	VERY HIGH

Source: Own preparation under the excel software.

As it is observed, the desertion risk ranges in each scale are not entirely consistent with the dynamics that are intended, and it is to construct class intervals that are independent and do not overlap. This fact is presented when using the methodology of sum of averages. To develop procedure 2, the data of the medians, the sum of them, as well as the results of question 15, were taken to SPSS. Initially, its normality was determined as a starting requirement for the subsequent application of a curvilinear model. The Q-Q chart demonstrates this concept (Shapiro, 1968).

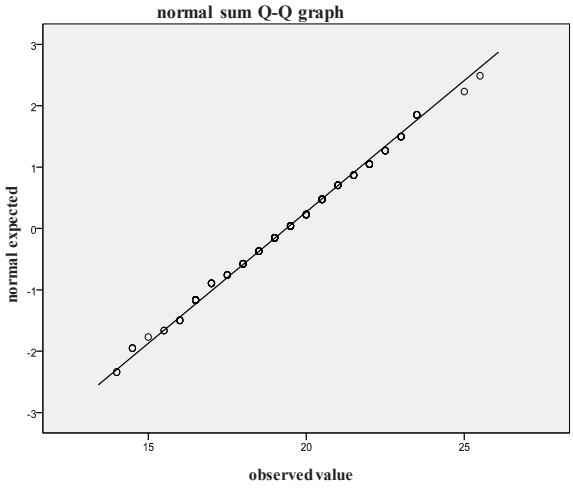

Figure 6. Q-Q graph that shows the normality of the variables in analysis

Once the normality criterion is applied, and eliminating the constants in the equations, the curvilinear regression model adjusted in terms of the best correlation coefficient R2 is the inverse equation model, as shown in the following table.

Table 6. Model summary and parameter estimates

Equation	Model Summary					Parameter estimates		
	R square	F	gl1	gl2	Sig.	b1	b2	b3
Linear	.700	358.715	1	154	.000	.100		
Logarithmic	.711	378.148	1	154	.000	.666		
Reverse	**.712**	**380.663**	**1**	**154**	**.000**	**37.332**		
Quadratic	.709	186.500	2	153	.000	.202	-.005	
Cubic	.709	186.500	2	153	.000	.202	-.005	.000
Compound	.421	112.005	1	154	.000	1.026		
Power	.431	116.547	1	154	.000	.170		
S	.440	120.953	1	154	.000	9.604		
Increase	.421	112.005	1	154	.000	.025		
Exponential	.421	112.005	1	154	.000	.025		
Logistics	.421	112.005	1	154	.000	.975		

Source: own elaboration based on statistical software SPSS.

On this curvilinear model, having an intercept point of Y at the origin, the equation is presented and parameter b1 is estimated. The model and the result is the following:

$$Y = \frac{b_1}{x_1} \tag{4}$$

This is:

$$Y = \frac{37.332}{xi} \tag{5}$$

The variable Y is estimated from the inverse of the variable X that is taken as the sum of the averages of the clusters already identified; b1 according to the model equals 37.332. For the construction of desertion risk levels, a frequency table was initially constructed, with 10 class intervals, each with amplitude equal to one tenth of the total range between the minimum value of 1.464 and the maximum value of 2.67. This presents a range of 1.2, by dividing the 10, you get 0.12. The purpose of this table was to identify some scales with greater probability of finding results of each answer option to question 15, coinciding with the mathematical model, without needing to

divide into ranges of equal magnitude.

Table 7. Frequency table for the identification of ranges of values of risk levels

Equation Model $Y = b/X_i$							
	b = 37.332		Answer options for Question 15				
Intervals	Intervals Range		1	2	3	4	5
1	1.464	1.584	0	0	0	0	2
2	1.585	1.705	14*	4	3	2	2
3	1.706	1.826	19	6	3	0	1
4**	1.827	1.947	9	6	6	0	2
5**	1.948	2.068	17	2	5	1	0
6**	2.069	2.189	10	1	4	2	1
7	2.19	2.31	8	6	2	3	1
8	2.311	2.431	2	2	1	1	0
9	2.432	2.552	0	0	0	0	1
10	2.553	2.673	2	0	2	0	1
Total by response option			**81**	**27**	**26**	**9**	**11**
TOTAL					**154**		

Source: own elaboration from Excel software.

Note. *14 surveyed of option 1 were presented between 1.585 and 1.705; there are three intervals (4, 5 and 6) that overlap response options 1, 2 and 3.

The graphical presentation, result of the previous table, makes the ranges more visible that would probably be part of each risk scale. For answer option 1 and 2 of question 15, intervals 2 through 6 have been taken; with their respective ranges that go from 1.585 to 2.189. Only the first graph is given as an example of the process.

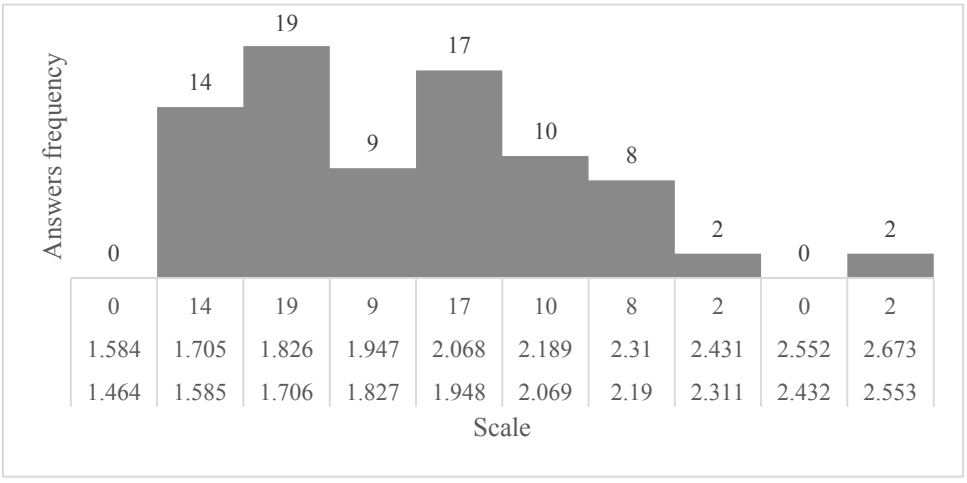

Figure 7. Range of responses on the risk scale for option 1 Question 15

It is important to note that not all responses are added at each level of risk, since the grouping is sought in a majority way; in this sense, the total sum of responses will not necessarily coincide with the 154 cases at all levels of risk. In the answer with option 2, the two responses from interval 2.31 to 2.431 are excluded. In the same way, at the medium risk level, the ranges from 1.827 to 2.31 were taken into account, in order to make relevant and establish the majority of responses for this level of risk. As it is perceived, some answers remain outside of these ranges. In the desertion risk levels considered as high or very high, the same intervals were taken for the latter two. They start with 2.069 to 2.673. 14 responses that are located in the first intervals were excluded, to avoid significant overlap with low risk levels.

Based on the previous table and the graphs, the five intervals were achieved according to the options of question 15, starting with the lowest value obtained in the calculation of the sum of averages. The scale then starts with 1.454 and achieves a maximum value of 2.67, in order to locate cases in a relevant way. The table below shows the results of the risk scale under this methodology.

Table 8. Scales of risk for dropping out, under the inverse equation model

		Options question 15	Risk Scale		Criteria
Procedure 2	VERY LOW	1	1.585	2.189	VERY LOW
	LOW	2	1.585	2.068	LOW
	MEDIUM	3	1.827	2.310	MEDIUM
	HIGH	4	2.069	2.431	HIGH
	VERY HIGH	5	2.069	2.673	VERY HIGH

Own preparation under the excel software.

This procedure shows results more adjusted to the reality of the present study than under the procedure 1 of grouping by averages, according to the proposed methodology.

4.4 Validation of the Model

It is relevant to determine how close is the suggested model to reality, assuming this as the result achieved by each respondent under question 15 of the questionnaire applied to the students of the universities in the Department of Santander. For this, it has been decided to divide the scale of risk to desertion into two parts. The starting point was the lower limit of the interval classified as Medium Risk; that is, 1,827. Values below this reference are considered low or very low risk levels, and higher values are considered high or very high risk levels. The mathematical model identified in procedure 2 must be coincident with the answer options of question 15. There must then be a significant percentage coincident between response options 3, 4 or 5 - high levels of probability of desertion, with the results of the risk scale of the mathematical model. Likewise, there are high percentages of coincidence between the response options 1 and 2, considered as low levels, with the risk scale established in the model. Supported in the xls spreadsheet, the coincidences were identified and the results of the calculation of values for the validation of the model were obtained. The following table summarizes the results of model validation.

Table 9. Results of the validity of the mathematical model based on the matches to the answer options of question 15

Results	Criteria Yes, Question Yes	Criteria No, Question Yes	Criteria Yes, Question No	Criteria No, Question No
Total values	33	27	108	77
%	21%	18%	70%	50%

Source: own elaboration.

Graphically, the construction of a plan was proposed in which the above mentioned results are located in the four quadrants. As shown, according to Figure 8, the matching criteria are representative when it is desired that the response options with values 3, 4 or 5, are consistent with the results of the suggested risk scale. Likewise, the model manages to identify in a greater percentage those students who do not wish to defect, who consequently opted for answers 1 and 2, presenting coincidence with the suggested scale for these levels. Another important aspect to highlight has to do with the total sum of the cases. In the absence of fully delimited scales in each level of risk; when there is overlap between them, it causes cases to arise that can be located in a duplicated manner in each risk scale, which results in values higher than the cases applied in the surveys.

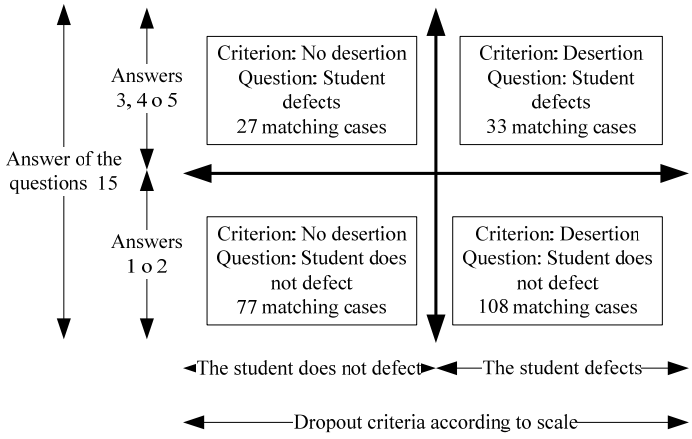

Figure 8. Validity of the mathematical model based on coincidence

5. Discussion

The variables present in the primary information collection instrument were the basis for proposing the clusters as a proposed model. Once the 10 clusters were formed supported by the SPSS software and after the procedure already described in Figure 3, the mathematical model that aims to identify the level of risk and that presented a better adjustment was initially the inverse, and later the cubic model. In the last one, it is identified that cluster 1 affects the behavior of the dependent variable, since there is a type of association due to dependence. The other clusters fail to establish some kind of correlation due to dependency or do not present any association.

In this vein, the proposed model was audacious in that it used a large number of variables, a total of 125, of which one was the dependent variable, and the others were used as returners. The current model used by the SPADIES system does not exceed 30 variables, and does not take into account aspects related to academic relationships and preparations. The sampling process used for the present study had characteristics of being simple random. The search for results closer to reality requires a sampling process, which may be of the stratified or conglomerate type, in such a way that it brings together the characteristics of several population groups from which it is desired to obtain particular information.

The intention of the present research proposal, when including that number of study variables, was to identify its degree of influence on the probability of desertion, under the constitution of clusters. Based on this, the resulting model has only identified the causal correlation of 11 variables that explain risk levels. This may be due to aspects such as the number of elements identified in the initial sample, the same moment, in which the instrument was applied, to the large number of clusters constituted under the closest neighbor relationship, among other aspects. In traditional models, there are other variables that influence the probability of deserting. The 11 variables in cluster 1 explain 72.4% of the variability to desertion, which is highly significant. The variables that are part of the cluster would explain the results of question 15 of the questionnaire. The validity of the model is evidenced in a percentage close to 55%, where it manages to identify coincidences between the intention of desertion with the criterion of the same predictive model (15%), as well as the intention of not defecting coinciding with the criterion of the same model (40%)

The current designed model, part of a transversal approach, where it is intended in a moment of time t, to identify the social phenomenon of desertion through the influence of intrinsic variables and thus be able to establish an approach through correlation statistics of clusters and variables that explain this phenomenon. As a future development, an integration of the present model is suggested, under a longitudinal approach (t, t + 1, t + 2...) of the desertion process, in such a way that the analysis allows following the dependent variable until it occurs the event of interest. This is, to identify through an evolutionary process (Kember, 1989), the previous and subsequent moments, as well as the fluctuations of each variable and cluster identified in the model. The suggestion lies in the same dynamics of the social phenomenon, since the permanence of the student is strongly associated (MEN, 2009); both at the specific moment of his academic career, as well as the variables at different times influence it.

The phenomenon of desertion must be analyzed taking into account both the students who are at a risk of taking the decision, as well as the same deserters, who have already taken the decision to withdraw from the classrooms. In this line, it is prudent to be able to identify them, locate them and obtain from them valuable information that can

give more accurate insights about the variables that led them to take this path. Thus, for future work it is suggested to take into account this group, which would become a primary source of valuable information, which will provide information to lead to predictive models with a higher level of reliability.

References

Amparo, B., & Grane, A. (2008). *100 Problemas Resuestos de Estadística Multivariante*. Madrid: Delta Publicaciones. Retrieved from http://www.listinet.com/bibliografia-comuna/Cdu311-2AC3.pdf

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysi*s (2nd ed.). New York: John Wiley & Sons.

Arriaza-Balmón, M. (2000). *Guía Práctica de Análisis de Datos*. Madrid: Instituto de Investigación y Formación Agraria y Pesquera. https://doi.org/10.16925/greylit.2256

Cabrera, L., Bethencourt, J. T., Álvarez Pérez, P., & Gónzalez, A. (2006). El problema del abandono de los estudios universitarios. *Relieve, 12*(2), 171-203. https://doi.org/10.7203/relieve.12.2.4226

Castaño, E. (2010). *Introducción al Análisis de Datos Multivariados en Ciencias Sociales*. XII Seminario de Estadística Aplicada III Escuela de Verano VII Coloqio regional de Estadística. Medellin, Antioquia, Colombia: Facultad de Ciencias, Universidad Nacional de Colombia. https://doi.org/10.22209/msiu.n3a35

Castaño, E., Gallón, S., Gómez, K., & Vásquez, J. (2004). Deserción estudiantil universitaria: una aplicación de modelos de duración. *Lecturas de Economía*, 39-65. https://doi.org/10.4067/s0718-07642009000500016

Cerpa, W., Castillo, M. P., & Cantillo, S. P. (2015). Análisis multivariado para determinar los factores más relevantes de deserción estudiantil presentes en el programa de Ingeniería Industrial de una Universidad del Caribe colombiano. *Prospect, 13*(1), 8, 86-99. https://doi.org/10.15665/rp.v13i1.363

Closas, A. H., Arriola, E. A., Kuc, C. I., Amarilla, M., & Jovanovich, E. C. (2013). Análisis multivariante, conceptos y aplicaciones en Psicología Educativa y Psicometría. *Enfoques XXV, 1*, 65-92. Retrieved from http://www.redalyc.org/articulo.oa?id=25930006005

Crámer, H. (1968). *Teoría de Probabilidades y sus Aplicaciones*. Madrid: Aguilar.

Diaz, C. (2008). Modelo Conceptual para la Deserción Estudiantil Universitaria Chilena. *Estudios pedagógicos XXXIV,* 65-86. https://doi.org/10.4067/s0718-07052008000200004

Ethington, C. (1990). A psychological model of student persistence. *Research in Higher Education, 31*(3), 279-293. https://doi.org/10.1007/bf00992313

Gauss, C. F. (1809). *Encyclopedia of Statistical Science* (V. 3).

Giovagnoli, P. (2001). *Determinantes de la Deserción y Graduación Universitaria, Una aplicación utilizando modelos de duración1*. La Plata, La Plata, Argentina: Tesis de Maestría. https://doi.org/10.25085/rsea.780104

Himmel, E. (2002). Modelos de Análisis de la deserción estudiantil en la educación superior. Calidad en la educación. *Recuperado de Google académico*, 91-108. https://doi.org/10.31619/caledu.n17.409

Hotteling, H. (1933). Analysis of a Complex of Statistical Variables Into Principal Components. *Journal of Educational Psychology, 24*, 417-441. https://doi.org/10.1037/h0071325

Instituto Colombiano para el Fomento de la Educación Superior-ICFES. (2002). *Estudio de la deserción estudiantil en la educación superior en Colombia, Documento Convenio UN*. Bogotá D.C.: MEN. https://doi.org/10.31619/caledu.n17.409

Kember, D. (1989). A Longitudinal-process Model of Dropout From Distance Education. *Journal of Higher Education, 60*(3), 278-301. https://doi.org/10.2307/1982251

Mardia, K. V., Kent, J. T., & Bibby, J. (1979). *Multivariate Analysis*. London: Academic Press. https://doi.org/10.1002/bimj.4710240520

Ministerio de Educación Nacional. (2009). *Deserción estudiantil en la Educación Superior Colombiana*. Bogotá D.C.: MEN. https://doi.org/10.31619/caledu.n17.409

Ministerio de Educación Nacional-MEN. (2016). *Sistema para la Prevención de la Deserción de la Educación Superior*. https://doi.org/10.1016/j.resu.2015.03.001

Ministerior de Educación Nacional. (2016). *Sistema Nacional de Información de la Educación Superior-SNIES*. https://doi.org/10.1016/j.resu.2016.01.007

Ministerior Nacional de Educación. (2016). SPADIES. *Sistema para la Prevención de la Deserción de la Educación*. Superior-SPADIES. https://doi.org/10.1016/j.resu.2015.03.001

Montanero-Fernandez, J. (2008). *Análisis Multivariante. Extremadura: Espacio Europeo de Educación Superior*. Retrieved from http://matematicas.unex.es/~jmf/Archivos/ANALISIS_MULTIVARIANTE.pdf

Morales-Vallejo, P. (2013). *El Análisis Factorial en la construcción e interpretación de tests, escalas y cuestionarios*. Madrid: Universidad Pontificia Comillas.

Peña, D. (2002). *Análisis de Datos Multivariante*. México D.F.: Mc Graw Hill. Retrieved from https://www.researchgate.net/publication/40944325_Analisis_de_Datos_Multivariantes

Shapiro, S. S., & Wilk, M. B. (1968). A Comparative Study of Various Test of Normality. *JASA, 63*(324), 43-72. http://dx.doi.org/10.2307/2285889

Spady, W. (1971). Dropouts from higher education: toward an empirical model. *Interchance, 2*(3), 38-62. http://dx.doi.org/10.1007/BF02282469

Terradez, M. (w.d.). *Análisis de Componentes Principales*. https://doi.org/10.21134/haaj.v17i2.332.s34

Tinto, V. (1975). Dropout From Higher Education A Theoretical Synthesis of Recent Research. *Journal of Higher Education, 45*(1), 89-125. https://doi.org/10.2307/1170024

Tinto, V. (1989). Definir la deserción una cuestión de perspectiva. *Revista Educación Superior*, 71.

Zamora, R. A. (2010). *Deserción en las instituciones de educación superior del estado de Durango: Modelo y propuesta estratégica*. Puebla, México.