Factors to Predict Dropout at the Universities: A case of study in Ecuador

Alban Mayra
Faculty of Computer Science and Computer Systems
Technical University of Cotopaxi
Latacunga, Ecuador
mayra.alban@ utc.edu.ec

David Mauricio
Faculty of Postgraduate of Systems and Informatics
National University San Marcos
Lima, Perú
dmauricios@unmsm.edu.pe

***Abstract*—The dropout at the universities has become a concern in several countries around the world, its high rates generate negative consequences for students and organizations. Based on the analysis of the educative, organizational theories, and the logic reasoning were established 11 factors that influenced in the dropout. This research as an objective to design a model to determine new factors to predict the dropout in which the dimension of analysis were the students, the institutions, the academic context and the social and economic environment. Additionally, trying through the use of Logistical Regression, Decision Tree and Support Vector Machine if the proposed factors are related and or may contribute predicting the dropout at the universities of Ecuador.**

Keywords—university student desertion, factor, data mining

I. INTRODUCTION

The dropout is one of the problems that affect to the majority of the institutions of high education around the world, it turns into a topic that causes controversy in the educative environment in which are involved administrators, professors, and students.

Nowadays, there is high dropout rates in the university education system, this can be evidenced by the data officially presented in the annual reports that Government Agencies worldwide show on the results of academic and administrative management in education [1]. In the India the factor of dropout was 15% in Belgium, the dropout rate was 26, 9% in the United States reached 24%. According to the United Nations Organization in some Latin American countries as Colombia and Ecuador, the dropout overcame 40%, while in Costa Rica the dropout rate reached 50% and in Brazil was 54% approximately.

As a result, university student desertion is related to a swarm of factors that interact with each other [2] and have a positive or negative influence on students' decision to stay or leave the university classroom [3]. Although the empirical evidence on factors that affect dropout is broad, most research has focused on studying the internal or external characteristics of students [4] leaving aside aspects of the academic, social, economic and institutional context in many cases it can affect the university dropout and require more exploration.

In the same sense, there is limited evidence on the application appropriate analytical methods, which allow obtaining the probability of incidence that has specific factors in the desertion of universities, based on the nature of their variables [5].

Based on the need to establish actions that allow the decrease of student desertion in universities, this article aims to try whether the proposed new factors affect the dropout of university students in Ecuador and verify whether these factors were analyzed in the units personal, academic, social, economic and institutional can predict university dropout. The results of the research will help the administrators of the Ecuadorian universities to promote changes in their academic policies and strategies in order to reduce dropout rates.

This article is organized into five sections. In section two shows the literature review. The method applied to the development of this research is in section three, section four explains the experimental process and the discussion of results, and finally, section five presents the conclusions.

II. LITERATURE REVIEW

The abandonment of students in universities is considered a problem that affects higher education institutions [6]. Currently, school dropout rates are analyzed as quality criteria in evaluation and accreditation processes of institutions of higher education [6]. In many cases, these control processes applied by the Government Agencies to universities imply economic and social changes [7] that they affects the students, institutions, and governments [8].

The review of the literature allows describing the determination of factors that influence the desertion of universities. For Willging & Johnson [9] two theories could provide a complete explanation about factors that affect student desertion, one of which is the theory of the integration of the Tinto model, who studies the motivation of the individual and the ability academic impact in the university dropout.

While the second theory is related to the Bean model, where the behavior of the students are analyzed, based on the beliefs and attitudes of the students. Based on the Theory of Tinto, authors such as Hovdlaugen [10], Elfers [11] & Duque [12] examine dropout factors based on motivation, academic behavior, and commitment to student behavior.

978-1-5386-2957-4/18/$31.00 ©2018 IEEE 17-20 April, 2018, Santa Cruz de Tenerife, Canary Islands, Spain
2018 IEEE Global Engineering Education Conference (EDUCON)

In the same sense, Arulampalam [13] determines that the causes of the origin of early school leaving linked to the demographic and academic characteristics of the students. In his study he analyzes data of individual level in students of United Kingdom, the results show that the variation of the qualifications previous to the entrance to the university exerts influence in the probability of desertion of the individuals.

On the other hand, Arulampalam [14] in his research refers to the importance of the analysis of factors related to the social integration of students. The author suggests that the decision taken by students to withdraw from the university is not only influenced by the magnitude of their qualifications but also by the degree of adaptation they have with their social environment.

Although the academic and social factors have considered as excellent predictors of desertion, the factors related to student characteristics, demographics, institutional characteristics and economic aspects were also analyzed in the literature review. Oseguera & Rhee [15] consider that institutional factors also include in desertion. As a result of the research, established that the institutional climate influence dropout rates.

Melguizo [16], evaluated the impact of student loans on low-income students, the objective of the study was to estimate the effects that educational loans can have on academic results to determine if an efficient program of economic aid allows the increase of educational outcomes satisfactory. Within this framework, Bonaldo [17] also studies the desertion based on the demographic characteristics of the students, such as age and marital status in Brazil.

In such a virtue it can be shown that there is a wide range of factors that can influence the decision of students to leave the university classrooms, which have been analyzed in various geographical contexts and related to the application of theories and educational aspects.

On the other hand, it is also essential to assess the ability of the applied techniques to determine factors that allow the identification of students at risk of leaving the university. The main methods of analysis focused on data mining techniques, specifically on statistical methods and artificial intelligence. It is considered necessary not only to determine the factors that influence dropout but also to establish if these factors can be regarded as excellent predictors of dropouts in universities.

III. METHOD

A. Process of selection of new factors

A set of 40 factors related to university student desertion were considered, through the study of literature review, the logical reasoning and the analysis of 65 organizational theories and 12 educational theories. As a result of this research, 11 factors were showed and can be considered as sources of mitigation of university dropout rates. The identified factors were classified into the 5 dimensions: personal, academic, social, institutional and economic.

The personal dimension determines the behavior of the student in his training process. The economic dimension refers to the material comforts that the student must possess for the successful achievement of their studies, while the social dimension focuses on the importance of the student's interaction with their social environment.

On the other hand, the academic dimension is related to the academic development of the student in his formative process. While, the institutional dimension considers the structural and functional characteristics of an institution related to the academic training of the student.

The model to determine new factors of desertion was based on three stages: specification, estimation, and testing.

Specification: identification of the stochastic process, that is, the identification of the appropriate values of entry into the model.

Estimate: process to estimate coefficients of significance and correlation of factors. For the estimation, Logistic Regression was applied as a non-linear parameter estimation method.

Testing: evaluates if the proposed model logically fits the data and is known as the evaluation and verification stage.

The process to determine new dropout factors is presented in Figure 1.

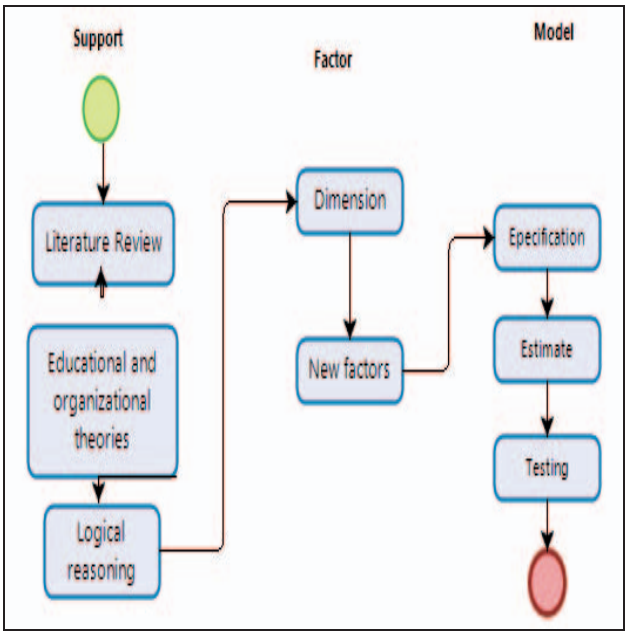

Fig. 1. Process to determine factors of dropout

B. Proposed factors

Six factors were established based on the analysis of educational theories, and 5 factors regarding the logical reasoning, which is presented in Table I.

For the development of the research, the variable explained will be the desertion in the universities, while the new factors were considered as explanatory variables.

TABLE I. FACTORS THAT INFLUENCE IN THE DROPOUT

Dimension	Factor	Support
Academic	Use of specialized software	Processing of the information Theory
	Adaptation to Learning Methodologies	Social Constructivism Theory
Personal	Number of children	Logical reasoning
	Commitment of the teacher to the student	Gestalt Theory
	Perception of the student about the insertion into the labor field	Logical reasoning
Social	Planned and unplanned pregnancies	Psychosocial Theory
	Maleness / Feminism	Logical reasoning
	Bulling	Behavior Theory
	Vices of the student	Logical reasoning
Economic	Commitment for being a firstborn child	Gestalt Theory
Institutional	Ranking of the institution or the carrier	Logical reasoning

IV. EXPERIMENTS AND DISCUSSIONS

A. Dataset

The dataset used is constructed of a survey made to 3000 university students that belong to 18 careers of Technical University of Cotopaxi, of Latacunga city – Ecuador. The face-to-face modality of study, included among the cohorts of admission from April 2011 to September 2016.

We excluded data from students who come from the homologation processes of other institutions of higher education and students who register career changes because these students are not considered for the calculation of permanence and dropout rates according to the evaluation and control bodies of the Ecuador.

B. Data Analysis

The analysis of data based on four stages was performed. First, the integration of the data with the information of the students that contained the socioeconomic cards and the data obtained from the surveys applied to the students.

Second, descriptive statistics are applied to determine the relationship of the factors and the data relationship, regarding directionality through correlation coefficients. It also establishes the condition of causality that refers to the positive relationship between two variables, show Table 2.

Third, the Logistic regression method is applied, since it is an appropriate analysis method for dichotomous variables is [5], used because it exceeds the limitations of the classification and produces estimated probabilities higher than 1. Also, allows establishing the relationship existing between the independent variables with the dependent variable, which in our case is desertion.

Fourth, with the application of artificial intelligence techniques, it is determined if the proposed factors predict or not the desertion in the universities of the Ecuador.

TABLE II. DESCRIPTIVE ESTATISTIC

Variable	Mean	Std. Dev.
CDE_ORIAP	0.889943	0.31301
DES	0.330171	0.470349
B_EE	0.529412	0.499213
NUM_HIJ3	0.829222	0.376374
CDE_DCH	0.891841	0.310631
ADIC_TEL	0.388994	0.487599
INTERNET	0.694497	0.460693
RED_S	0.717268	0.450399
TEC	0.601518	0.489663
HIJO_PRIM	0.620493	0.485341
RANKING	0.713472	0.45221
AMFU_TA	0.637571	0.480778
AMFU_TC	0.677419	0.467538
AMFU_UTEC	0.721063	0.448548
EMB_NP	0.833017	0.37302
FEMINSMO	0.508539	0.500006
MACHISMO	0.559772	0.496493
PEICL_PEP	0.766603	0.423059
All	0.661712	0.473131

C. Experimental results

The Logistic model is used to predict the probability of occurrence of an event in which data are used that fit the logistic curve and must meet two characteristics:

1. The dependent variable must contain binary data. This means that it can have two possible outcomes 1 and 0, which in this case represents the influence or not of the variable on the dropout.

2. Independent variables have no correlation level.

Since the estimated coefficients of the Logit model are logarithmic probability coefficients and cannot be interpreted directly as probabilities.

To calculate the probability of abandonment, the maximum likelihood estimation that is responsible for analysis problems will be applied through an interactive optimization routine that maximizes the function of the likelihood logarithm when the dependent variables are similar. The result

of modeling the dropout factors using the Eviews software is presented in Figure 2.

Through of the Logit method, the probability of incidence was measured to determine the relationship between these factors and the dropout. The value of p (Y = 1) = 0.5 served as a critical point the cut, the advantage of this method is the smoothness of the starting hypotheses that are easily verifiable.

In Figure 2, present the factors use of specialized software, adaptation to learning methodologies, number of children, commitment of the teacher to the student, perception of the student about the insertion into the labor field, planned and unplanned pregnancies, maleness / feminism, bulling, vices of the student, commitment for being a firstborn child, ranking of the institution or the carrier.

Dependent Variable: DES
Method: ML - Binary Logit (Newton-Raphson / Marquardt steps)
Sample: 1 3000
Convergence achieved after 3 iterations
Coefficient covariance computed using observed Hessian

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-1.633.524	0.177515	-9.202.182	0.0000
CDE_ORIAP	-0.389884	0.153656	-2.537.385	0.0112
CDE_DCH	0.164040	0.157624	1.040.706	0.2980
ADIC_TEC	0.314379	0.142823	2.201.176	0.0277
INTERNET	0.304191	0.139106	2.186.750	0.0288
RED_S	-0.557044	0.161871	-3.441.281	0.0006
TEC	0.279078	0.093161	2.995.662	0.0027
HIJO_PRIM	-0.082438	0.094274	-0.874451	0.3819
RANKING	0.095364	0.101578	0.938822	0.3478
AMFU_TA	0.175421	0.113213	1.549.479	0.1213
AMFU_TC	-0.878195	0.118978	-7.381.134	0.0000
AMFU_UTEC	0.343056	0.129480	2.649.498	0.0081
EMB_NP	-0.088607	0.085212	-1.039.852	0.2984
NUM_HIJ23	0.710419	0.107804	6.589.937	0.0000
B_EE	0.070856	0.088935	0.796723	0.4256
FEMINISMO	-0.045459	0.090227	-0.503828	0.6144
PEICL_PEP	0.696718	0.117132	5.948.123	0.0000
McFadden R-squared	0.059436	Mean dependent var		0.331667
S.D. dependent var	0.470890	S.E. of regression		0.454802
Akaike info criterion	1.207.846	Sum squared resid		6.166.039
Schwarz criterion	1.245.886	Log likelihood		-1.792.769
Hannan-Quinn criter.	1.221.529	Deviance		3.585.538
Restr. deviance	3.812.116	Restr. log likelihood		-1.906.058
LR statistic	2.265.778	Avg. log likelihood		-0.597590
Prob(LR statistic)	0.000000			

Fig. 2. Result of the process of modeling the new factors

The result of the experimental process evidence that these factors, they are statistically significant that is to say they are factors that influence the university student desertion.

The representation of the model statistically expresses the following commands:

LS DES C ADIC_TEC AMFU_TA AMFU_TC AMFU_UTEC B_EE CDE_DCH CDE_ORIAP EMB_NP FEMINSMO HIJO_PRIM INTERNET MACHISMO NUM_HIJ3 PEICL_CP RANKING RED_S TEC

The Estimation Equation is represented in the equation (2).

DES = C(1) + C(2)*ADIC_TEC + C(3)*AMFU_TA + C(4)*AMFU_TC + C(5)*AMFU_UTEC + C(6)*B_EE + C(7)*CDE_DCH + C(8)*CDE_ORIAP + C(9)*EMB_NP + C(10)*FEMINSMO + C(11)*HIJO_PRIM + C(12)*INTERNET + C(13)*MACHISMO + C(14)*NUM_HIJ3 + C(15)*PEICL_CP + C(16)*RANKING + C(17)*RED_S + C(18)*TEC (2).

On the other hand, the model has a vectors the independent variables or regressors of X, which are assumed as an influence on the result Y. A typical approximation with a regression of ordinary least squares is incorrect due to the errors of the regression are heteroscedastic and they are not normal; estimated probability results would be nonsense values above 1 or below 0.

For the prediction of the university student desertion, the method of selection of variables CorrelationAttributeEval is established: evaluates the value of an attribute by measuring the correlation Pearson's between it and the class.

TABLE III. FACTORS THAT INFLUENCE IN THE DROPOUT

Weight	Attribute
0.13938	PEICL_Pep_binarized
0.13634	Num_hij3_binarized
0.08518	TEC_binarized
0.07011	Ranking_binarized
0.06872	Feminsmo_binarized
0.05931	AMFU_TC_binarized
0.05564	B_EE_binarized
0.05017	7 Hijo_Prim_binarized
0.04551	Machismo_binarized
0.04518	Internet_binarized
0.04387	Emb_Np_binarized
0.03953	Adic_Tec_binarized
0.03663	CDE_Dch_binarized
0.0341	AMFU_TA_binarized
0.03181	AMFU_UTEC_binarized
0.02385	CDE_Oriap_binarized
0.00175	Red_S_binarized

To the construction of the prediction models was used Logistic Regression, Decision Tree Classifier, Support Vector Machine through the application of cross-validation, Additionally, were considered 10 test for each algorithm.

The first step for each of these methods was to adjust the parameter to the training data set and supervised learning was carried out, based on cross-sectional data obtained through the

analysis of descriptive data. The result of the prediction process is presented in Figure 3.

The validation of the used techniques considers the performance measure accuracy that represents the accuracy of an instrument.

Of these, accuracy is the metric used which is determined by the ratio of true positives (TP) and true negatives (TN) among the total of registers, as is formulated in (2).

$$\frac{TP+TN}{TP+TN+FP+FN} \quad x \ 100 \tag{2}$$

FP is the quantity of false positives and FN the quantity of false negatives.

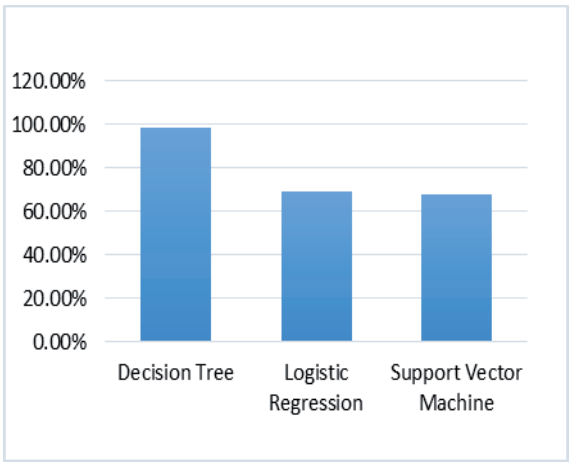

Fig. 3. ccuracy the techniques of prediction

The technique that presents the highest precision is Decision tree classifier with 98 /%. It can be explained because it is an optimal application technique due to its flexibility to treat data of a categorical and numerical nature, monotonous transformations of explanatory variables, ease of interpretation of results and because it presents better precision rates [10].

V. CONCLUSIONS

Given the concern about the dropout of the students, it is essential to determine the causes for which there is still high rates of dropout at the universities. For this reason, it is necessary to implement practical measures to change the vicious circle in which is immersed the universities respect to the dropout, 11 factors were identified to determine the dropout, also the technique that presents more accuracy is Decision Tree Classifier with 98%.

REFERENCES

[1] M. Elias-Andreu, "Los abandonos universitarios: retos ante el Espacio Europeo de Educación Superior," 2008.

[2] C. Henríquez and R. Escobar, "Construcción de un modelo de alerta temprana para la detección de estudiantes en riesgo de deserción de la Universidad Metropolitana de Ciencias de la Educación," *Revista mexicana de investigación educativa,* vol. 21, pp. 1221-1248, 2016.

[3] J.-H. Park and H. J. Choi, "Factors influencing adult learners' decision to drop out or persist in online learning," *Journal of Educational Technology & Society,* vol. 12, 2009.

[4] R. Chen, "Institutional characteristics and college student dropout risks: A multilevel event history analysis," *Research in Higher Education,* vol. 53, pp. 487-505, 2012.

[5] T. Melguizo, F. S. Torres, and H. Jaime, "The association between financial aid availability and the college dropout rates in Colombia," *Higher Education,* vol. 62, pp. 231-247, 2011.

[6] C. Díaz Peralta, "Modelo conceptual para la deserción estudiantil universitaria chilena," *Estudios pedagógicos (Valdivia),* vol. 34, pp. 65-86, 2008.

[7] E. Castaño, S. Gallón, K. Gómez, and J. Vásquez, "Análisis de los factores asociados a la deserción y graduación estudiantil universitaria," *Lecturas de economía,* 2006.

[8] E. Himmel, "Modelos de análisis de la deserción estudiantil en la educación superior," *Revista calidad de la educación,* vol. 17, pp. 91-108, 2002.

[9] P. A. Willging and S. D. Johnson, "Factors that influence students' decision to dropout of online courses," *Journal of Asynchronous Learning Networks,* vol. 13, pp. 115-127, 2009.

[10] E. Hovdhaugen, "Transfer and dropout: Different forms of student departure in Norway," *Studies in Higher Education,* vol. 34, pp. 1-17, 2009.

[11] L. Elffers, "Staying on track: behavioral engagement of at-risk and non-at-risk students in post-secondary vocational education," *European Journal of Psychology of Education,* vol. 28, pp. 545-562, 2013.

[12] L. C. Duque, "A framework for analysing higher education performance: students' satisfaction, perceived learning outcomes, and dropout intentions," *Total Quality Management & Business Excellence,* vol. 25, pp. 1-21, 2014.

[13] W. Arulampalam, R. A. Naylor, and J. P. Smith, "Effects of in-class variation and student rank on the probability of withdrawal: cross-section and time-series analysis for UK university students," *Economics of Education Review,* vol. 24, pp. 251-262, 2005.

[14] W. Arulampalam, R. A. Naylor, and J. P. Smith, "Dropping out of medical school in the UK: explaining the changes over ten years," *Medical Education,* vol. 41, pp. 385-394, 2007.

[15] L. Oseguera and B. S. Rhee, "The influence of institutional retention climates on student persistence to degree completion: A multilevel approach," *Research in Higher Education,* vol. 50, pp. 546-569, 2009.

[16] T. Melguizo, F. Sanchez, and T. Velasco, "Credit for Low-Income Students and Access to and Academic Performance in Higher Education in Colombia: A Regression Discontinuity Approach," *World development,* vol. 80, pp. 61-77, 2016.

[17] L. Bonaldo and L. N. Pereira, "Dropout: Demographic profile of Brazilian university students," *Procedia-Social and Behavioral Sciences,* vol. 228, pp. 138-143, 2016.