International Journal of Engineering & Technology, 7 (4.36) (2018) 1281-1284

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Decision Trees for the Early Identification of University Students at Risk of Desertion

Mayra Albán[1]*, David Mauricio[2]

[1,2]Technical University of Cotopaxi, Faculty of Computer Science, Ecuador
[2]National University of San Marcos, Artificial Intelligence Group, Perú
**Corresponding author E-mail: mayra.alban@utc.edu.ec*

Abstract

The student's dropout at the universities is a topic that has generated controversy in Higher Education Institutions. It has negative effects which cause problems in the social, academic and economic context of the students. One of the alternatives used to predict the dropout at the universities is the implementation of machine learning techniques such as decision trees, known as prediction models that use logical construction diagrams to characterize the behavior of students and identify early students that at in risk of leaving university. Based on a survey of 3162 students, it was possible to obtain 10 variables that have influence into the dropout, that's why, a CHAID decision tree model is proposed that presents the 97.95% of the accuracy in the prediction of the university students' dropout. The proposed prediction model allows the administrators of the universities developing strategies for effective intervention in order to establish actions that allow students finishing their university careers successful.

Keywords: *Prediction of college desertion, machine learning, decision trees, CHAID.*

1. Introduction

The completion of the university has not always been the norm for the society, in the 1940s, less than half of the US population between the ages of 25 and 29 would have finished the university Ye & Bisway [1]. Although there has existed a concerted effort to close the gap related with the dropout at the universities and decrease its rates, researches that started in 1978 shows that still exists dropout at the universities Abuda & Oda [2] which has caused effects on the economic ambit for Higher Education institutions and governments.

The dropout causes difficulties at the university context [3] and it is considered as an evaluation criterion and an argument of great relevance to incorporate the public policies related with the education at the universities [4]. Although effective systematic changes have been made to solve this problem, students continue to face an educational crisis with the greatest propensity to leave their studies [5]. In addition, dropout generates social consequences in terms of the students' expectations and their families; as well as emotional consequences for the dissonance between the aspirations of young people and their achievements. The important economic consequences for both people and the system as a whole are also considered within this context [6].

The prediction of the university dropout becomes important from some decades ago, where it is started analyzing the factors that influence dropout, combined with the way to predict the risk of dropping out [7]. On the other hand, the prediction of the dropout in the Institutions of Higher Education has been questioned, due to the high rates of dropping out that the institutions still have [8].

In the literature review were found researches such as those of Marquez [9], Herzog [10], Kotsiantis, et.al. [11] the authors establish the models of prediction about dropout through experimental processes that consider methods of machine learning supervised to discover knowledge.

So, if the above problems are maintained, the high error rates in the accuracy of the prediction will continue. For this reason, it is important to establish a model that allows integrating data, variables and appropriate techniques to accurately predict students at risk of dropping out. In addition, it will allow Higher Education Institutions to have an effective tool to make decisions rightly in relation to the dropout of the universities' students.

The research work is divided into five parts. The literature review is in the second one, the method is developed in the third, the results of the experimental process is considered in the fourth and in the last part are presented the conclusions.

2. Literature Review

The university student desertion is a problem that has been widely studied in the literature. It is possible to demonstrate the efforts made by the researchers around the topic of their prediction to try to mitigate the dropout rates and establish strategies that allow the incorporation of strategies for timely decision making. Several works have been identified around the topic under study and are presented in Table 1.

Table 1: Works related to the application of decision trees to predict dropouts in universities

Technique	Resource
Decision tree classifier	[12],[13],[14],[15], [16], [17],[18], [19], [20],[21], [22], [23], [24], [25], [26], [27], [28]

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

3. Materials and Method

3.1. Data set

For the development of the research, a survey is applied to 3162 students enrolled in the face-to-face undergraduate studies of the Engineering Careers of a Public University of Ecuador. The analysis period includes the study cohorts from 2012 to 2017. Through the application of Google Form, the information regarding student behaviour was obtained, specifically information regarding students' behaviour habits. , which can generate dependence and negatively influence the decision to leave the university classrooms.

3.2 Analysis of data

The methodology applied for the application of decision tree models is based on the following stages:

Stage-1 Integration and cleaning of the data, carried out to obtain quality data that allow an adequate prediction process.

Stage-2 Pre-processing of the data, used to determine the normality and consistency of the data.

Stage-3 Extraction of variables, used to determine the variables greater incidence in the prediction model.

Stage-4 Prediction: decision trees are applied to predict the desertion of university students according to the characteristics of the variables of entry into the model.

Stage-4 Evaluation: through evaluation metrics, the reliability level of the proposed model is determined in terms of confidence and reliability of the prediction model.

4. Result and Discussion

Decision trees are techniques that allow decision making based on the use of associated probabilities. They facilitate the level of comprehension regarding the behaviour of the variables that influence the desertion of students in universities.
The dependent variables and the independent variables used as predictor variables are presented in Table 2.

Table 2: Description of the predictor variables

Independent Variable	Description
Red_S	Social networks addiction
Age	Age
Alc	Alcohol addiction
Adic_ apegem	Addiction to emotional attachment
Adic_drug	Addiction to drug
Adic_Tel	Mobile phone addiction
Jueg	Games addiction
Videos	Video games addiction
Adic_com	Shopping addiction
Dependent Variable	
DES	Dropout

Table 3 presents the level of importance of the independent variables for the prediction model.

Table 3: Importance of independent variables

Independent Variable	Importance	Standard Importance
Red_S	0.006	1.000
Age	0.003	0.444
Alc	0.002	0.403
Adic_ apegem	0.001	0.096
Adic_drug	0.001	0.083
Adic_Tel	0.000	0.063
Jueg	0.000	0.006
Videos	0.001	0.096
Adic_com	0.000	0.002

The weights of these variables were determined using the SPSS software. Validation by sampling division corresponds to 80% of the data (2530 cases) for the training sample and 20% (632 cases) for the test sample. Cross-validation 10 times is used to evaluate the results of the statistical analysis and the validation of the proposed model.
The criteria established for the CHAID decision tree correspond to a maximum number of levels 3. The level of significance for the division nodes is equal to 0.05 as is the merging of categories.
For the modelling of the decision tree, binary variables were considered, that is:

0 = students will drop out of college
1 = students will not drop out of college.

For estimating the model, a maximum number of estimates of 100 was considered. And a minimum change value in the expected frequencies of 0.001. The Chi-square statistic used was Person and he made the determination of the values of significance through the Bonferroni method. The summary of the proposed model is presented in Table 4.

Table 4: Summary of the model

Specifications	Growth method	CHAID
	Dependent variable	DES
	Independents Variables	Age, Alc, Jueg, Video, Internet, Red_S, Adic_drug, Adic_Ejer, Adic_Com, Adic_apegem, Adic_Tel
	Validation	Split sample
	Maximum tree depth	3
	Minimum of cases in a subsidiary node	100
	Minimum of cases in a parent node	50
Results	Independent variables included	Red_S, Adic_Tel, Age, Adic_apegem, Adic_Com, Alc
	Number of nodes	15
	Number of terminal nodes	8
	Depth	3

Fig.1 presents the structure of the decision tree and the assignment of the probabilities of the factors used as input variables to the model.

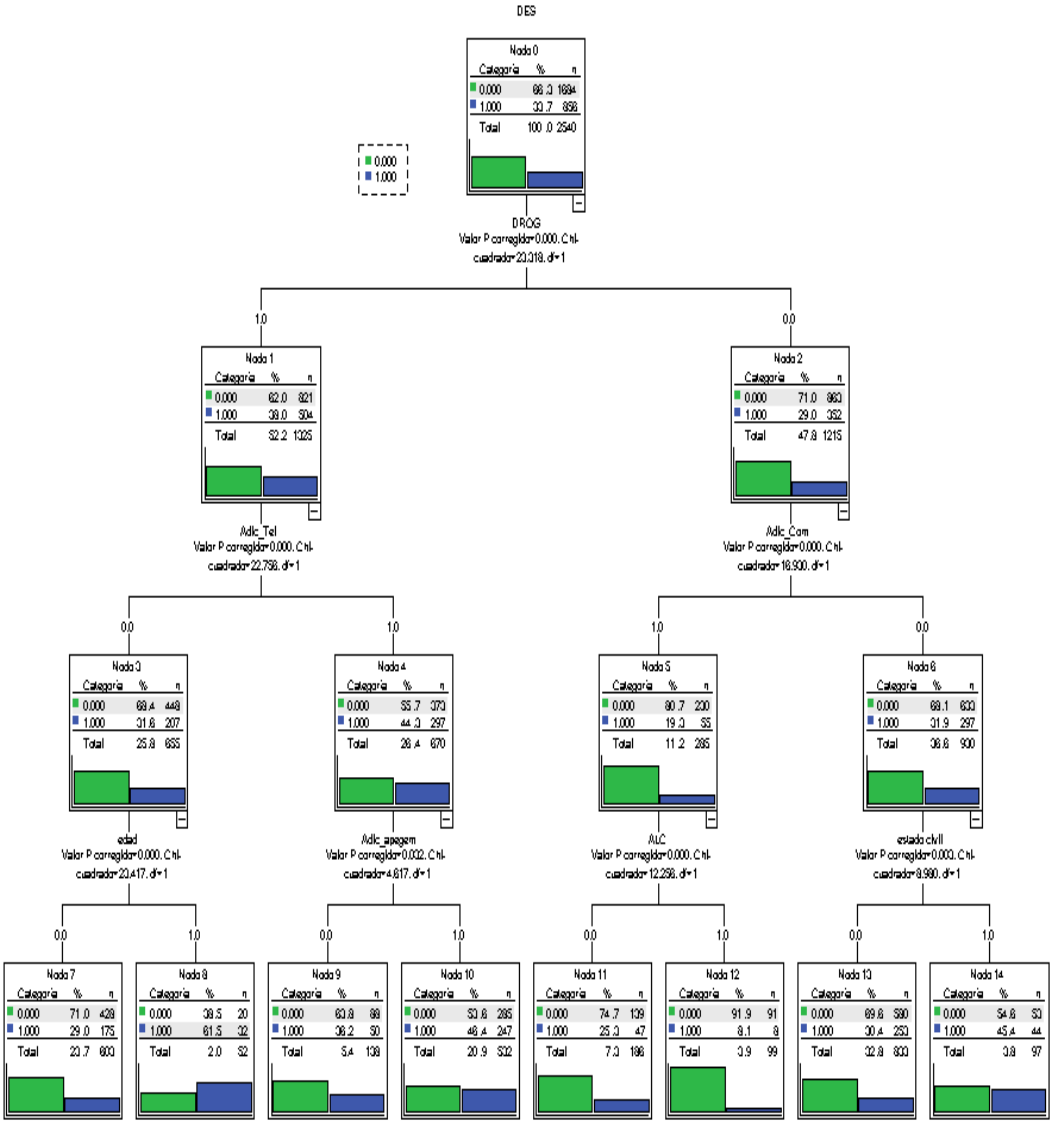


Fig. 1: Decision tree using the CHAID method

A sample of the rules determined to predict desertion through the CHAID method is presented below:

```
/* Node 7 */.
DO IF (SYSMIS (Red_S) OR VALUE (Red_S) NE 0) AND (VALUE (Adic_Tel) EQ 0) AND (SYSMIS (age) OR VALUE (age) NE 1).
COMPUTE nod_001 = 7.
COMPUTE pre_001 = 0.
COMPUTE prb_001 = 0.709784.
END IF.
EXECUTE.

/* Node 8 */.
DO IF (SYSMIS (Red_S) OR VALUE (Red_S) NE 0) AND (VALUE (Adic_Tel) EQ 0) AND (VALUE (age) EQ 1).
COMPUTE nod_001 = 8.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.615385.
END IF.
EXECUTE.

/* Node 9 */.
DO IF (SYSMIS (Adic_drug) OR VALUE (Adic_drug) NE 0) AND (SYSMIS (Adic_Tel) OR VALUE (Adic_Tel) NE 0) AND (VALUE (Adic_apegem) EQ 0).
COMPUTE nod_001 = 9.
COMPUTE pre_001 = 0.
COMPUTE prb_001 = 0.637681.
END IF.
EXECUTE.
```

The results of the experiments performed show a prediction accuracy rate corresponding to 97.75% for the sample used for training and 98.71% for the testing sample. Which means that this model that the proposed model is adequate in relevance and determines the quality and effectiveness of the proposed model.

Table 5: Overall percentage of prediction

Sample	Observed	Predicted	
		0	**1**
Training	0	1664	20
	1	824	32
	Global percentage	97.95%	2.05%
Contrast	0	430	4
	1	184	4
	Global percentage	98.71%	1.29%

5. Conclusions

The students' dropout from the educational system requires a special interest in the Higher Education Institutions, especially in the public sector. This problem causes negative effects such as the student's failure in the achievement of their academic goals. Additionally, it also generates economic losses for the institutions and governments, the decrease in the graduation rate generates big social and institutional problems**.**

The results of the experimental process allow determining that the factors: addiction to the social networks, addiction to the emotional attachment, marital status and age were considered as influencing factors to the dropout process.

An accuracy rate of 97.95% allowed determining that the CHAID method applied to predict dropout at the universities is optimal in terms of quality and effectiveness.

Based on the obtained results, it can be established that the proposed model could be considered as an optimal method to predict dropout in students at the universities.

Also, it can be used by university administrators too as an early warning tool in the detection of students at risk of dropping out. It can also become a support instrument for the application of university policies that allow increasing the rate of student permanence.

References

[1] C. Ye and G. Biswas, "Early prediction of student dropout and performance in MOOCs using higher granularity temporal information," Journal of Learning Analytics, vol. 1, pp. 169-172, 2014.

[2] G. S. Abu-Oda and A. M. El-Halees, "Data mining in higher education: university student dropout case study," International Journal of Data Mining & Knowledge Management Process, vol. 5, p. 15, 2015.

[3] Archambault, M. Janosz, V. Dupéré, M. C. Brault, and M. M. Andrew, "Individual, social, and family factors associated with high school dropout among low-SES youth: Differential effects as a function of immigrant status," British Journal of Educational Psychology, vol. 87, pp. 456-477, 2017.

[4] C. R. Hoyt, "The Impact of the Tax Reform Act of 1986 on Legal Education and Law Faculty," J. Legal Educ., vol. 36, p. 568, 1986.

[5] A. V. D. López, " Strategies to overcome university desertion," Educación y educadores, vol. 7, pp. 177-203, 2004.

[6] C. Vogel, J. Hochberg, S. Hackstein, A. Bockshecker, T. J. Bastiaens, and U. Baumöl, "Dropout in Distance Education and how to Prevent it," in EdMedia+ Innovate Learning, 2018, pp. 1788-1799.

[7] G. M. Alarcon and J. M. Edwards, "Ability and motivation: Assessing individual factors that contribute to university retention," Journal of Educational Psychology, vol. 105, p. 129, 2013.

[8] F. Roso-Bas, A. P. Jiménez, and E. García-Buades, "Emotional variables, dropout and academic performance in Spanish nursing students," Nurse education today, vol. 37, pp. 53-58, 2016.

[9] C. Márquez-Vera, C. R. Morales, and S. V. Soto, "Predicting school failure and dropout by using data mining techniques," IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, vol. 8, pp. 7-14, 2013.

[10] S. Herzog, "Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen," Research in higher education, vol. 46, pp. 883-928, 2005.

[11] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," Emerging artificial intelligence applications in computer engineering, vol. 160, pp. 3-24, 2007.

[12] E. Yukselturk, S. Ozekes, and Y. K. Türel, "Predicting dropout student: an application of data mining methods in an online education program," European Journal of Open, Distance and E-learning, vol. 17, pp. 118-133, 2014.

[13] Y.-H. Hu, C.-L. Lo, and S.-P. Shih, "Developing early warning systems to predict students' online learning performance," Computers in Human Behavior, vol. 36, pp. 469-478, 2014.

[14] D. Thammasiri, D. Delen, P. Meesad, and N. Kasap, "A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition," Expert Systems with Applications, vol. 41, pp. 321-330, 2014.

[15] D. Yasmin, "Application of the classification tree model in predicting learner dropout behaviour in open and distance learning," Distance Education, vol. 34, pp. 218-231, 2013.

[16] M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart, "Predicting MOOC dropout over weeks using machine learning methods," in Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, 2014, pp. 60-65.

[17] M. Tan and P. Shao, "Prediction of student dropout in e-Learning program through the use of machine learning method," International Journal of Emerging Technologies in Learning (iJET), vol. 10, pp. 11-17, 2015.

[18] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting student dropout in higher education," arXiv preprint arXiv:1606.06364, 2016.

[19] W. Xing, X. Chen, J. Stein, and M. Marcinkowski, "Temporal prediction of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization," Computers in Human Behavior, vol. 58, pp. 119-129, 2016.

[20] S. Natek and M. Zwilling, "Student data mining solution–knowledge management system related to higher education institutions," Expert systems with applications, vol. 41, pp. 6400-6407, 2014.

[21] N. Lam-On and T. Boongoen, "Using cluster ensemble to improve classification of student dropout in Thai university," in Soft Computing and Intelligent Systems (SCIS), 2014 Joint 7th International Conference on and Advanced Intelligent Systems (ISIS), 15th International Symposium on, 2014, pp. 452-457.

[22] W. Li, M. Gao, H. Li, Q. Xiong, J. Wen, and Z. Wu, "Dropout prediction in MOOCs using behavior features and multi-view semi-supervised learning," in Neural Networks (IJCNN), 2016 International Joint Conference on, 2016, pp. 3130-3137.

[23] A.-S. Hoffait and M. Schyns, "Early detection of university students with potential difficulties," Decision Support Systems, vol. 101, pp. 1-11, 2017.

[24] A. K. Pal and S. Pal, "Analysis and mining of educational data for predicting the performance of students," International Journal of Electronics Communication and Computer Engineering, vol. 4, pp. 1560-1565, 2013.

[25] M. Fei and D.-Y. Yeung, "Temporal models for predicting student dropout in massive open online courses," in Data Mining Workshop (ICDMW), 2015 IEEE International Conference on, 2015, pp. 256-263.

[26] S. Sultana, S. Khan, and M. A. Abbas, "Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts," International Journal of Electrical Engineering Education, vol. 54, pp. 105-118, 2017.

[27] A. Sangodiah,P. Bleleya, M. Muniandy, L. E. Heng, and C. Ramendran SPR, "Minimizing Student attrition in Higher Learning Institutios in Malaysia Using Support Vector Machine", " Journal of Theoretical & Applied Information Technology, vol. 71, 2015.

[28] M. A. AL-Barrak and M. S. AL-Razgan, "Predicting Student´ Performance trough Classification: A case study," Journal of Theoretical & Applied Information Technology, vol. 75, 2015.