Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 155 (2019) 569–574

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Applying Data Mining Techniques to E-Learning and Pedagogical Approaches (ADMEPA)

August 19-21, 2019, Halifax, Canada

Integration of Data Technology for Analyzing University Dropout

Amelec Viloria[a]*, Jholman Garcia Padilla[b], Carlos Vargas-Mercado[c], Hugo Hernández-Palma[d], Nataly Orellano Llinas[e], Monica Arrozola David[f]

[a,b] *Universidad de la Costa, Street 58 # 55 - 66. Barranquilla, Colombia.*
[c] *Corporación Universitaria Latinoamericana, Street 58 #55 -24a, Barranquilla, Colombia.*
[d] *Universidad del Atlántico, Street 30 # 8- 49 ,Puerto Colombia – Colombia*
[e] *Corporación Universitaria Minuto de Dios – UNIMINUTO, Street 53 #74-110 , Barranquilla, Colombia.*
[f] *Universidad Libre Seccional Barranquilla, Street 46 No. 48-170 , Barranquilla, Colombia*

Abstract

Dropout, defined as the abandonment of a career before obtaining the corresponding degree, considering a significant time period to rule out the possibility of return. Higher education students´ dropout generates several issues that affect students and universities. The results obtained from the data provided by the Engineering departments of the University of Mumbai, in India, determine that the variables that best explain a student's dropout are the socioeconomic factors and the income score provided by the University Admission Test (UAT). According to the decision tree technique, it is concluded that the retention is 78.3%. The quality of the classifiers allows to ensure that their predictions are correct, with statistical levels of ROC curve are 76%, 75%, and 83% successful for Bayesian network classifiers, decision tree, and neural network respectively.

Keywords: university retention; university dropout; data mining; education; engineering; Big Data.

* Amelec Viloria. Tel.: +57-304-6238313
E-mail address: aviloria7@cuc.edu.co

1877-0509
10.1016/j.procs.2019.08.079

1. Introduction

The most relevant studies in this discipline are those of [1] and [2]. [1] formulates a theoretical model that explains the processes of interaction between the individual and the institution, which determine the reasons for leaving the university and distinguish those processes that give rise to different forms of dropout behavior. The most important dimensions considered by this model are: pre-entry attributes, goals and commitments, institutional experiences, and personal and regulatory integration. [2] applies a work turnover model to the student dropout. The author points out that a student leaves the University not necessarily because of performance, either it is high or low, but because of other factors which are external to academics. The used variables are the academic factors, psychosocial factors, environmental factors, and socialization factors. It determines which are the most important variables that affect a college student decision to drop out higher education before obtaining the degree.

The referred studies [3], [4], [5] determined that the variables that best explain this phenomenon are the score of the university admission test and the high school qualifications. While [6] reports that the determinants of university retention are recruitment and admission, academic services, curriculum and instruction, student services, and financial aid. These studies use classical statistical techniques to determine their results, without further digging into possible hidden patterns in the data, contributing with a different perspective to the dropout problem.

This research establishes the importance of the variables that affect the university student´s decision to drop out by using data mining techniques, comparing them with the variables identified in the literature review. The correlation between the results obtained in this work and the literature is presented, with a retention of 77.93%, compared to 75.3% reported by [7] for the period 2008-2012 in India. A study in [8] applies a similar method on a case in China showing retention percentages of 81.3% and 81% for J48 and BayesNet algorithms respectively.

2. Method

This work is developed with a descriptive approach for proving the following hypotheses by analyzing quantitative data: (i) The entry academic conditions (middle-school grades, knowledge, UAT score) of the student determine the dropout in the Engineering degree at the University of Mumbai; and (ii) The student's socio-economic conditions determine the decision to drop out of the university studies in Engineering. This research considers dropout in a significant time period for a student to decide to resume higher education studies, equivalent to 3 years, using [7] and [8] as a reference.

This study applies for the years from 2015 to 2018 with a total sample of 19,300 individuals by using data mining techniques. 3 classifiers are created to categorize students between the classes of dropout and non-dropout. 3 algorithms are used: Bayesian networks, neural networks, and decision tree [9], [10], [11], [12], [13], [14], which are assessed for their quality as classifiers and the most important variables in each algorithm that allow to determine whether a student will drop out or not. The KDD (Knowledge Database Discovery) process enables the selection, cleansing, transformation, and projection of data; analysis of data to extract appropriate patterns and proper models; evaluation and interpretation of patterns to turn them into knowledge; consolidation of knowledge by solving potential conflicts with previously extracted knowledge; and making knowledge available for use [9]. The KDD process consists of the following steps: (a) Information Source, (b) Data Preparation, (c) Data Mining, (d) Interpretation and Evaluation, and (e) Knowledge. The variables used in the research are shown in Table 1.

The following software tools are used for the development of this research: SQL Server for database storage, SPSS Statistics for the neural network classifier and decision tree, and Weka for the classifier of Bayesian networks. For the repeatability of the experiment, it is necessary to have at least the same variables used for this analysis, using for the final view: career, UAT scores per test and weighted, NEM (Average student grade in the 4 years of middle education), NEM score (corresponding to the standardized score equivalent to the student's average middle-education score), assigned benefits per year, semi-annual grade average, and academic status.

Table 1. Investigation variables.

Variable	*Normalization method*
Academic status (class variable)	Variables that do not require normalization
Year of career admission	
Student Middle School Year	
Career code to which the student belongs	
Order in the selection list	Method of highs and lows
UAT Weighted Average (University Admission Test)	Series media method
Student application preference	
Average Grade in 4 Years of Middle School (NEM)	
UAT Score for NEM (Standardized Score equivalent to the average student middle-education score)	
Student benefits	
Qualifications per semester for each course taken by the student	

3. Results and Analysis

This section presents the main results of the creation of classifiers: (1) of the Bayesian network classifier construction; (2) the decision tree classifier construction; and (3) classifier construction using neural networks.

3.1 Bayesian network classifier construction

The classifier is developed using the BayesNet algorithm in the Weka software [15]. It results in a graph through which it is possible to characterize the deserting students. Among the results obtained through the Bayesian network, it can be mentioned that the dropout found by this classifier is 33.9% with a retention of 66.1%, which corresponds to the result of the classification for the variable Class (dropout and no-dropout). While 94% of those students who drop out have an in-average score in the UAT, while those who do not dropout are on this average at 87%. Another result from the most important variables to explain dropout is the career application order which indicates that when a student applies to Indian universities he has 10 application options in order of discard, i.e. if the first choice is selected (career, university), the following options are automatically discarded. This preference does not have a significant influence on the decision to drop out as only 9% is in the average preferences, as detected by the algorithm. This may indicate that students who did not place Mumbai University Engineering in the early places drop out of the university. Those dropout students have a conditioned 89% chance of having a middle-school average grade within the sample analyzed for this study.

3.2 Decision tree classifier construction

A classifier is developed using the decision tree algorithm. The following obtained results can be mentioned: for the variable Class (dropout and retention), 21.7% is classified as dropout and 78.3% as retained. While those students who have student benefits (credits, scholarships), in the average or less than this value have an 89.3% chance of remaining in their college career and 10.7% of dropout. On the other hand, those with higher or lesser average (extreme) benefits dropout with a 28.1% chance of leaving of school versus 71.9% who would not. This group of students corresponds to 63.1% of the analyzed students and shows that those who have an economic benefit have a greater chance of staying and finishing their careers.

Those who have an average on their UAT, right in the sample mean, have a 76.7% chance of staying in their careers and 23.3% of dropping out. While those scores at the extremes of the sample (maximum and minimum) have a 25.7% chance of staying in their careers and 74.3% of dropping out, which is assumed at the lower end of this category. Another variable that is at the same level as the UAT average as a good predictor in the decision tree, is the history

test score. Those at the top end of the sample have a 91.1% chance of staying in the career, while those below average in the history test have an 88.8% chance of staying in their career and a 12% chance of dropping out. In the third level of dropout predictors of the decision tree, are those students who have benefits for 3 periods back, standing out those at the extremes of this sample, i.e. those students who have benefits for 3 periods back, highlight those who are at the ends of this sample, i.e. those students who have very high benefits or very small amounts in this period have an 80% chance of dropping out of their higher education studies.

3.3 Classifier construction by using neural networks

A neural network is developed using the Perceptron Multilayer algorithm with an admissible error of 0.0001. In the model architecture, a custom methodology was used for activating the input and output layers, through the hyperbolic tangent function. The neural network is a black box at the level of prediction interpretation, however, the results that the classifier delivers are analyzed with respect to the normalized importance of the variables analyzed as effective classifiers of dropout and/or retention. Regarding the standardized importance of the variables delivered by this analysis the most relevant are mentioned: UAT average, profit in last 3 periods, profit level 0, application weighted average, score of the student´s average teaching grades, and the economic benefits over the past 2 periods. While this result does not match the one obtained in the Naive Bayes classifier, it should be mentioned that both of them use different analysis techniques and must be evaluated in that context (Table 2).

Table 2. Standardized importance, variables neural network

Importance of independent variables		
Variables	Importance	Normalized importance
List_Order_1	.021	16.8%
Average_Weight_1	.088	68.9%
Preference_1	.053	41.5%
NEM_1	.078	61.0%
ScoreEM_1	.084	65.9%
Mathematics_1	.031	24.2%
Language_1	.058	45.7%
History_1	.063	49.4%
Science_1	.064	50.4%
Average_test_1	.128	100.0%
Benefit_n3_1	.110	85.8%
Benefit _n2_1	.080	62.5%
Benefit _n1_1	.051	40.2%
Benefit _n0_1	.089	70.0%

Note that the fifth place is occupied by the EM score, which represents the grade obtained by students in their 4 years of secondary school, transformed into standardized score of the UAT helping the student to apply to the University. To evaluate the quality of the built classifier, Table 3 presents the neural network evaluation parameters. According to the ROC curve, it can be said that the classification was performed correctly in 83% of cases and with a high accuracy for positive cases classified with 73% accuracy and 88% for the ratio of accurately classified negative cases. In both the test and the training of this classifier, 80% of the individuals were correctly labeled, since both tests result in a similar level of correct classification. It implies a good level of accuracy in the results obtained (see Table 3). At a general level, it can be noted that, in terms of the quality of the built classifiers, there is a low probability of having obtained incorrectly classified values according to the ROC curve of the neural network classifiers, decision tree, and Bayesian network (83%, 74% and 76% respectively) (Table 3).

Table 3. Comparison of the classifier evaluation parameters

	Precision	*Recall*	*Tpr*	*Tnr*	*Fpr*	*Fnr*	*Roc Curve*	*F-Measure*	*classified right*	*incorrect classifieds*
Neural network	73%	65%	65%	88%	12%	35%	83%	69%	80%	20%
Decision tree	72%	64%	64%	87%	12%	36%	74%	68%	82%	18%
Bayesian Network	76%	76%	76%	70%	30%	24%	76%	76%	76%	24%

The analysis shows a dropout rate of 33.9% (Bayesian network) and 21.7% (decision tree) for the analyzed Engineering careers. It is not possible to establish that one classifier is better than another, even it can be noted that the three of them are useful for analyzing the problem considering its advantages and disadvantages. The experiment is replicable under similar conditions, replicating the use and type of data used in this research.

4. Conclusions

Three classifiers (Bayesian networks, decision tree and neural networks) were built. Each of these results has its particularities and advantages so they are not fully comparable. However, there is a coincidence on scholarships and credits as student's benefits that are decisive in determining dropout. From the academic perspective, the variable that would best explain the dropout is the UAT score (Table 4).

Table 4. Variables that best classify dropout of students according to each classifier.

Decision tree	*Bayesian Networks*
Benefit Level 0	UAT Average
Benefit Level 3	Benefit level 3
Benefit Level 1	Benefit level 1

The results found in this analysis show that the two hypotheses are positively proven, since both the academic results and the socioeconomic situation influence a student's decision to stay in his or her respective career. Given the high correlation obtained with the literature consulted and the good quality of the models, it is possible to say that managing these variables helps reduce the dropout rates in the university system.

References

[1] Vasquez, C., Torres, M., Viloria, A.: Public policies in science and technology in Latin American countries with universities in the top 100 of web ranking. J. Eng. Appl. Sci. **12**(11), 2963–2965 (2017).

[2] Aguado-López, E., Rogel-Salazar, R., Becerril-García, A., Baca-Zapata, G.: Presencia de universidades en la Red: La brecha digital entre Estados Unidos y el resto del mundo. Revista de Universidad y Sociedad del Conocimiento **6**(1), 1–17 (2009).

[3] Torres-Samuel, M., Vásquez, C., Viloria, A., Lis-Gutiérrez, J.P., Borrero, T.C., Varela, N.: Web Visibility Profiles of Top100 Latin American Universities. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, Springer, Cham, vol **10943**, 1-12 (2018).

[4] Viloria, A., Lis-Gutiérrez, J.P., Gaitán-Angulo, M., Godoy, A.R.M., Moreno, G.C., Kamatkar, S.J.: Methodology for the Design of a Student Pattern Recognition Tool to Facilitate the Teaching – Learning Process Through Knowledge Data Discovery (Big Data). In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, Springer, Cham, vol **10943**,1-12 (2018).

[5] Caicedo, E.J.C., Guerrero, S., López, D.: Propuesta para la construcción de un índice socioeconómico para los estudiantes que presentan las pruebas Saber Pro. Comunicaciones en Estadística, vol. **9**(1), 93-106 (2016).

[6] Mazón, J.N., Trujillo, J., Serrano, M., Piattini, M.: Designing Data Warehouses: From Business Requirement Analysis to Multidimensional Modeling. In Proceedings of the 1st Int. Workshop on Requirements Engineering for Business Need and IT Alignment. Paris, France (2005).

[7] Vásquez, C., Torres-Samuel, M., Viloria, A., Lis-Gutiérrez, J.P., Crissien Borrero, T., Varela, N., Cabrera, D.: Cluster of the

Latin American Universities Top100 According to Webometrics 2017. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, Springer, Cham, vol **10943**, 1-12 (2018).
[8] Haykin, S.: Neural Networks a Comprehensive Foundation. Second Edition. Macmillan College Publishing, Inc. USA. ISBN 9780023527616 (1999).
[9] Jain, A. K., Mao, J., Mohiuddin, K. M.: Artificial neural networks: a tutorial. IEEE Computer 29 (3), 1- 32 (1996).
[10] Sevim, C., Oztekin, A., Bali, O., Gumus, S., Guresen, E.: Developing an early warning system to predict currency crises. European Journal of Operational Research 237(1), 1095-104 (2014).
[11] Sekmen, F., Kurkcu, M.: An Early Warning System for Turkey: The Forecasting of Economic Crisis by Using the Artificial Neural Networks. Asian Economic and Financial Review 4(1), 529-43 (2014).
[12] Singhal, D., Swarup, K.S.: Electricity price forecasting using artificial neural networks. IJEPE 33 (1), 550-55 (2011).
[13] Mombeini, H., Yazdani-Chamzini, A.: Modelling Gold Price via Artificial Neural Network. Journal of Economics, Business and Management 3 (7), 699-703 (2015).
[14] Kulkarni, S., Haidar, I.: Forecasting Model for Crude Oil Price Using Artificial Neural Networks and Commodity Future Prices. International Journal of Computer Science and Information Security 2 (1), 81-89 (2009).
[15] Bontempi, G., Ben Taieb, S., Borgne, Y. A.: Machine learning strategies for time series forecasting. In Lecture Notes in Business Information Processing, ed M.-A. Aufaure., and E. Zimányi, Heidelberg: Springer 138 (1), 70-73 (2013).