Neural Networks to Predict Dropout at the Universities

Mayra Alban and David Mauricio

***Abstract*—The university student's dropout is a problem that affects the governments, institutions and students. It has negative effects on the high expenditure in the administrative and academic resources. Predicting dropout has become an advantage for university administrators because it allows discovering students that are at risk of dropout as well as develop actions that allow taking decisions in a timely manner. This research presents a neural network approach through the application of multilayer perceptrom algorithms and radial basis function. As input variables to the models, 11 factors were considered, which produce a negative influence in the desertion at the universities; the data was obtained from a survey of 2670 students of a Public University in Ecuador. The results showed that there is no significant difference in the accuracy rates of the proposed models which correspond to 96.3% for multilayer perceptrom and 96.8% for radial basis function. As a conclusion, the studied models could be considered as an optimal option in terms of accuracy and concordance to predict dropout at the universities.**

***Index Terms*—Prediction, university student desertion, neural networks, multilayer perceptrom, radial basis function.**

I. Introduction

Students' desertion at the universities is considered as a problem that affects higher education institutions worldwide [1]. Nowadays, the high dropout rates are considered as possible deficiencies in the undergraduate education system [2]. This can be evidenced in the academic and administrative management reports presented by government agencies worldwide. In the United States, the dropout rate at first and second year reaches 44.8% [3], 15.9% corresponds to the dropout rate in India [4], in countries such as Colombia, Ecuador and Brazil, the dropout rate at the universities exceeds 40% according to the United Nations (UN) in 2016.

As a result, the dropout is a consequence of a set of factors that interact with each other which have a negative influence in students to not finish their university studies successfully [5]. Dropout is considered as the voluntary or involuntary early abandonment of a study program, in which the student discards the successful completion of their studies [6]. This phenomenon is still present in the higher education system and is related to negative effects such as the high economic and social cost that affects the student and the universities [7]. Although there is a large number of studies to try to solve the problem of dropout and determine its causes, there is a limited scientific production that incorporates machine learning algorithms such as neural networks that allows discovering the knowledge based on the nature of the variables obtained through the behavior of the students to know their condition of risk to dropout.

To design actions that allow to decrease the desertion rates at the universities, two models of neural networks, multilayer perceptrom and radial base function are proposed. The results of this research will help universities administrators to promote changes in their academic policies and strategies in order to reduce their dropout rates.

This article is organized in five sections. Section II presents the background, theory and literature review. The materials and methods are presented in Section III. Section IV presents the results and the discussion and in Section V there is the conclusions.

II. Background Theory and Literature Review

Several studies to predict dropout have been identified in the literature review, developed for the early identification of vulnerable and prone students to leave university classrooms. Table I shows some jobs that use neural networks classifier to predict dropouts in universities.

TABLE I: Neural Networks to Predict Dropout at the Universities

Description	References
Prediction of desertion in Turkey, based on data collected in an Information Technology Program.	[8]
Early warning system to predict the dropout of students in Information Literacy and Information Ethics careers.	[9]
Prediction of desertion in Spain	[10]
Prediction of the desertion through characteristics of the academic, financial, demographic performance of the students.	[11]
Prediction of the desertion in Czech Republic, through data collected in students of the career of Applied Informatics	[12]
Methodology to connecting the analysis of learning and educational data mining and solve the problem of dropout at the university	[13]
Proposed a cluster set as a frame of data transformation to identify a prediction of dropout more precise	[14]
Prediction method of the dropout at the university in México	[15]
Prediction of the desertion in Europe through data collected in students of the Faculties of Economic Sciences	[16]

Manuscript received October 20, 2018; revised March 1, 2019. This work was supported in part by Technical University of Cotopaxi, Faculty of Computer Science and Computer Systems. Av. Simón Rodriguez, Latacunga, Ecuador, and National University of San Marcos, Group of Artificial Intelligence, Faculty of Computer Systems. Av. German Aemzaga 375, Lima1, Lima, Perú.

M. Alban, She is with the Faculty of Engineering and Applied Sciences of the Technical University, Cotopaxi (e-mail: mayra.alban@utc.edu.ec).

D. Mauricio is with the National University of San Marcos (e-mail: dmauricios@unmsm.edu.pe).

doi: 10.18178/ijmlc.2019.9.2.779

III. Materials and Method

A. Data Collection

The data set used corresponds to information obtained from 2670 students enrolled in undergraduate studies in the Administrative and Human Sciences Careers of the Public University of Ecuador. The data was collected through an online survey applied through Google Form. The survey was applied to university students from the first to the four academic year. The period of analysis includes the study cohorts from 2014 to 2017.

The information of the students used in the investigation includes demographic and student behavior data, information related to the university education methodologies, data corresponding to the academic processes and socioeconomic information of the students.

B. Result Analysis

For the development of the research, the stages were applied which are presented in Fig. 1.

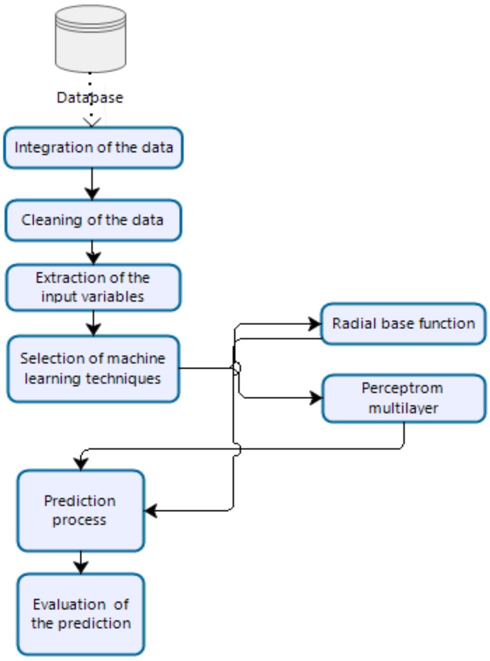

Fig. 1. Process to predict college dropout through neural networks.

As a first step, the integration and cleaning of the data obtained from the survey is established, this step was carried out to determine that there is no information redundancy and blank fields or data that may affect the prediction process. The process of the extraction of factors determines the importance of the variables and are considered as predictors in the input layer of the neural network models.

For the data modeling process, two machine learning techniques were selected based on radial function and multilayer perceptrom, used for being an automatic learning approach for discovering knowledge regarding the behavior of variables with dropout in universities.

The prediction process was used to determine the degree of causality existing between the factors of admission to the models and the desertion. An evaluation of the neural network models was carried out through metrics of precision prediction as sensibility.

IV. Result and Discussion

A. Data Preprocessing

The extraction of the most appropriate factors from the data set was carried out. Which were selected through the determination of weights that establish the importance of the factors based on their attributes, using the ranking method. The definition of the variables used is presented in Table II and the value corresponding to the weights of the variables is presented in Table III.

TABLE II: Definition of Variables

ID	Factor
F01	Limited knowledge about specialized software usage in the university major
F02	Planned and / unplanned pregnancy
F03	Teacher’s commitment to the student
F04	First-born son’s financial commitment to his family
F05	Bullying
F06	Sexism
F07	Students acquired addictions
F08	Student’s number of children
F09	Student’s adaptation to the university learning
F10	Major or institution ranking
F11	Student’s perspective on his or her integration into the labor market
DES	Dropout

TABLE III: Importance of Independent Variables

	Radial base function		Perceptrom multilayer	
	Importance	Standard Importance	Importance	Standard Importance
F3	0,090	0.520	0.0827	0.653
F7	0.080	0.465	0.100	0.796
F1	0.173	1	0.108	0.858
F4	0.104	0.599	0.105	0.830
F10	0.103	0.593	0.078	0.620
F2	0.058	0.339	0.088	0.695
F8	0.093	0.538	0.126	1
F5	0.076	0.438	0.066	0.524
F11	0.060	0.346	0.105	0.837
F9	0.069	0.398	0.0723	0.571
F6	0.088	0.510	0.064	0.512

B. Prediction of University Student Dropout

The objective of the experimental process is to demonstrate that neural networks can be used as high performance algorithms to predict college dropout. Therefore, a comparison was made between two perceptrom multilayer algorithms and the radial basis function.

Secondly, the performance of the applied methods is evaluated to determine the significant differences of the results obtained in terms of the accuracy of the prediction.

Para Montaño [17] neural networks are considered information processing systems based on connections through neurons. It works through the sum of the weights of the factors that are compared as threshold values to obtain an activation or output. Its objective is to select based on the internal parameters of the input layer, the relationships between the connections de las the neurons through the training of the network [18]. Neural networks being one of the most popular methods of machine learning, this has been widely used in education for its ability to recognize patterns in student behavior based on the identified factors.

The weighted sum of the input values for the neural network were considered based on the weights assigned to each of the input functions presented in equation (1), used to define the neurons of the network and considered as the Input parameters of the proposed model.

$$F=(a1x1+a2x2+a3x3\ldots\ldots+b+anxn) \tag{1}$$

where:

a = the weight of the variable
x = neural network entry variable
y = output function of the neuronal
b = bias

For modeling the neural network, binary variables will be considered, that is, 0 to predict that students will drop out of college and 1 to predict that college students will not drop out of college.

On the other hand, for the construction of the first model, multilayer perceptrom was used. The values of the weights are determined from the training sample and the error function of the neural network considered as the mean square error [19]. The training options for Multilayer Perceptron Network are presented below:

```
MLP des (mlevel=n) by F3 F7 F1 F4 F10 F2 F8 F5 F11 F9
F6
Partition training=6 testing=3 holdout=1
Architecture   automatic=yes (minunits=1 maxunits=50)
Criteria training=batch optimization=scaledconjugate
lambdainitial=0.0000005 sigmainitial=0.00005
intervalcenter=0 intervaloffset=0.5 memsize=1000
Print cps networkinfo summary classification solution
importance
Stoppingrules errorsteps= 1 (data=auto) trainingtimer=on
(maxtime=15)    maxepochs=auto    errorchange=1.0e-4
errorratio=0.0010
```

The partition data set corresponds to 60% (1602 cases) of training and 30% (801 cases) for the testing, leaving an additional 10% (267 cases) for the reservation.

In the architecture of the model, it was considered as the minimum number of units of the hidden layer 1 and a maximum value of the hidden layer units 50. For the training process of the neural network, the sigmoid logarithmic function was used. In addition, crossed entropy was used as an error function. The summary of the model using is presented in Table V. The results obtained from the neural network using the SPSS software correspond to 11 factors used in the input layer and defined as independent variables and a dependent variable (DES).

TABLE V: SUMMARY OF THE MODEL PERCEPTROM MULTILAYER

Training	Cross entropy error.	1166.316
	Percentage of incorrect forecasts	0.126
	Stop rule used	1 pasos consecutivos sin disminuci ón del error[a]
	Training time	00:00:01,187
Testing	Cross entropy error.	566.285
	Percentage of incorrect forecasts	0.118
Backup	Percentage of incorrect forecasts	0.134

The number of occupied layers 1a is equal to 9 and the activation function of the hidden layer used was the hyperbolic tangent, while the activation function of the output layer was softmax.

Fig. 2 presents the curve of elevation of the dependent variable (DES), for the prediction model using multilayer perceptron.

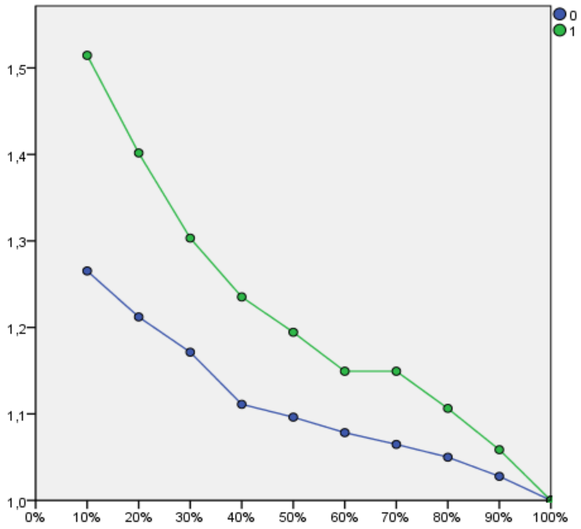

Fig. 2. Elevation curves of the Perceptrom multilayer.

Regarding the second model, the radial base function was used, the data partition set with 70% (1869 cases) of training and 40% (801 cases) for testing. The training options are presented below:

```
Radial basis function network.
RBF des (mlevel=n) by F3 F7 F1 F4 F10 F2 F8 F5 F11 F9
F6
Partition training=7 testing=3 holdout=0
Architecture minunits=auto maxunits=auto
hiddenfunction=nrbf
Criteria overlap=auto
Print cps networkinfo summary classification importance
Missing usermissing=exclude.
```

The 11 factors are used as independent variables for the input layer. In the hidden layer, the number of units is equal to 7a through the softmax activation function.

The data partition set with 70% (2183 cases) of training and 40% (979 cases) for testing. As in the multilayer perceptrom neural network, the 11 factors are used as independent variables for the input layer. In the hidden layer, the number of units is equal to 7a through the softmax activation function.

For the output layer, the dependent variable (DES) was considered again, the number of units equals 2 using the Identity activation function and the sum-of-squares error function. The number of hidden units is determined by the test data criterion, the optimal number of hidden units is the one that produces the smallest error in the test data. The summary of the model used is presented in Table V.

TABLE V: SUMMARY OF THE MODEL OF FUNCTION DE BASE RADIAL

Training	Sum of quadratic errors	485.548
	Percentage of incorrect forecasts	0.116
	Training time	00:00:02,613
Testing	Sum of quadratic errors	210.722[a]
	Percentage of incorrect forecasts	0.105

Fig. 3 presents the curve of elevation of the dependent variable (DES), for the prediction model using multilayer perceptron.

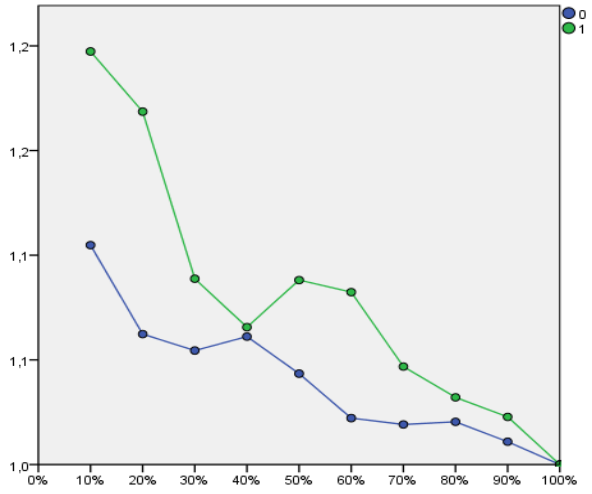

Fig. 3. Elevation curves of the radial base function.

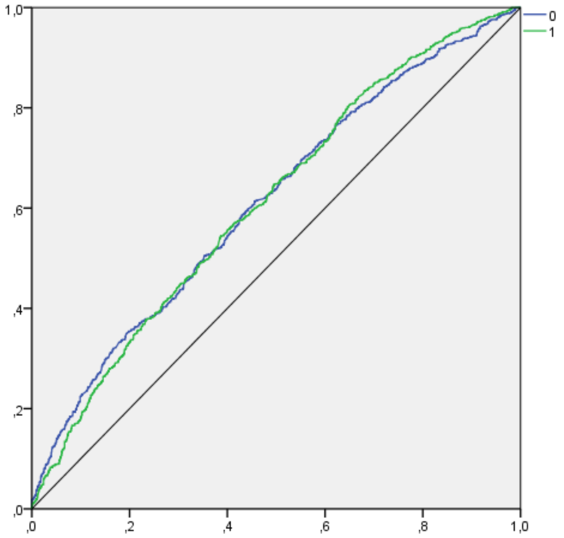

Fig. 4. Sensitivity (Y) vs. Specificity (X).

On the other hand, for evaluate the feasibility of the model the curve is analyzed Roc (receiver operating characteristics) and determine the predictive capacity of prediction models. This type of graphics are created based on pseudo probabilities that use the error of the quadratic sums and the activation function of the output layer. The Fig. 4 shows the visual representation the sensitivity determined in the column (Y) versus the specificity presented in column (X) in relation to the dependent variable DES.

The metric sensitivity is considered as the relation of positive truths while the specificity is the relation of true negatives of the cut points of the neural network [20].

As can be seen, both techniques show an optimal performance in terms of prediction accuracy in universities. There are no significant differences in terms of accuracy between the two models, therefore it can be considered that these can be an option to accurately predict the students at risk of leaving the university classrooms. The results of the prediction process are presented in Fig. 5 and Table VI.

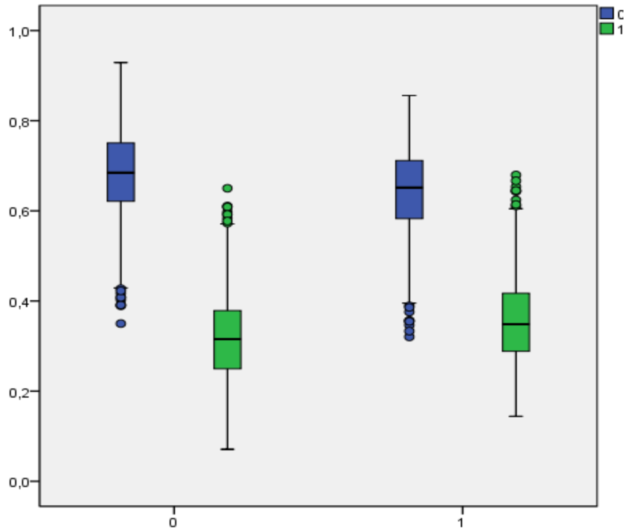

Fig. 5. Pseudo probability predicted

TABLE VI: ACCURACY OF THE PREDICTION MODELS

Neural Networks	Training	Testing
Perceptrom Multilayer	96,3%	98,6%
Function de Base Radial	96,8%	98,1%

V. CONCLUSION

The article presents an application of neural networks through the implementation of multilayer perceptrom algorithms and radial basis function as predictive methods of dropout of students at the universities. 11 factors were considered as variables of income to the prediction models, according to the process of variable selection Teacher's commitment to the student, Students acquired addictions, Limited knowledge about specialized software usage in the university major, were those that have higher index of importance in the proposed models

The results of the analysis showed that the multilayer perceptrom has a prediction rate of 96.3%, while the accuracy rate of the radial base function corresponds to 96.8%. This indicates that models of neural networks are reliable for the identification of students at risk of dropping out of the universities.

These results can help administrators of universities in order to take decision timely and in the implementation of strategies that allow reducing their dropout rates.

REFERENCES

[1] E. Sneyers and K. D. Witte, "The effect of an academic dismissal policy on dropout, graduation rates and student satisfaction. Evidence from the Netherlands," *Studies in Higher Education,* vol. 42, pp. 354-389, 2017.

[2] Á. H. F. Díaz, "Análisis sobre la deserción en la educación superior a distancia y virtual: El caso de la UNAD-Colombia," *Revista de Investigaciones UNAD,* vol. 8, pp. 117-149, 2009.

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

[4] P. S. Wolf, A. David, S. T. Butler-Barnes, and V. Zile-Tamsen, "American Indian/Alaskan Native College Dropout: Recommendations for increasing retention and graduation," *Journal on Race, Inequality, and Social Mobility in America,* vol. 1, p. 1, 2017.

[5] 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.

[6] E. Himmel, "Modelo de análisis de la deserción estudiantil en la educación superior," *Calidad en la Educación*, pp. 91-108, 2018.

[7] K. B. Eckert and R. Suénaga, "Análisis de deserción-permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos," *Formación Universitaria,* vol. 8, pp. 03-12, 2015.

[8] 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.

[9] 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.

[10] J. A. Lara, D. Lizcano, M. A. Martínez, J. Pazos, and T. Riera, "A system for knowledge discovery in e-learning environments within the European Higher Education Area–Application to student data from Open University of Madrid, UDIMA," *Computers & Education,* vol. 72, pp. 23-36, 2014.

[11] 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.

[12] J. Bayer, H. Bydzovská, J. Géryk, T. Obsivac, and L. Popelinsky, "Predicting drop-out from social behaviour of students," *International Educational Data Mining Society,* 2012.

[13] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques," *Computers & Education,* vol. 53, pp. 950-965, 2009.

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

[15] M. A. A. Dewan, F. Lin, and D. Wen, "Predicting Dropout-Prone Students in E-Learning Education System," in *Proc. 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing*, 2015, pp. 1735-1740.

[16] 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.

[17] J. J. Montaño Moreno, *Redes Neuronales Artificiales Aplicadas al anáLisis de Datos*, 2017.

[18] C. Z. Torres, C. A. Ramos, and J. L. Moraga, "Estudio de variables que influyen en la deserción de estudiantes universitarios de primer año, mediante minería de datos," *Ciencia Amazónica:(Iquitos),* vol. 6, pp. 73-84, 2016.

[19] A. I. Al-Omari and C. N. Bouza, "Review of ranked set sampling: modifications and applications," *Investigación Operacional,* vol. 35, pp. 215-235, 2014.

[20] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2016.

Mayra Alban has a degree in computer science and computational systems at the Technical University of Cotopaxi, Ecuador, in 2002, master in production management in 2006 and doctoral aspirant from the National University of San Marcos Lima-Peru. She is a professor of computer science and computational systems at the Faculty of Engineering and Applied Sciences of the Technical University at Cotopaxi. Currently she is the director of the information systems career. He has developed research in the last five years related to the desertion of students in universities and data mining.

David Mauricio has a doctor of science in systems engineering and computing, and master of science in applied mathematics at the Federal University of Rio de Janeiro, Brazil. He also got the bachelor of computing from the National University of San Marcos. He has been a professor at the North Fluminense State University of Brazil from 1994 to 1998. From 1998, he has been a professor at the National University of San Marcos. His areas of interest are in combinatorial optimization, designs and analysis of algorithms, heuristics search, metaheuristics, mathematical programming, expert systems, data mining, artificial intelligence.