Multilayer Neural Networks for Predicting Academic Dropout at the National University of Santa - Peru

Hugo Esteban Caselli Gismondi
Ingeniería de Sistemas e Informática
Universidad Nacional del Santa
Chimbote, Perú
hcasellig@uns.edu.pe ORCID: 0000-0002-2812-6727

Luis Vladimir Urrelo Huiman
Ingeniería de Computación y Sistemas
Universidad Privada Antenor Orrego
Trujillo, Perú
lurreloh@upao.edu.pe ORCID: 0000-0003-1523-2640

2021 International Symposium on Accreditation of Engineering and Computing Education (ICACIT) | 978-1-6654-0719-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICACIT53544.2021.9612507

***Abstract*—Investigations have applied the Machine Learning to predict whether a college student culminate or not his studies, however, in each scenario the factors that influence student graduation are multiples, then: how predict defection students at the Universidad Nacional of Santa - Peru with a precision greater than 90%? In the present research a model based on Multilayer Neural Networks was trained to predict the academic dropout at the School of Engineering from the aforementioned university, to Neural Networks Multilayer of 6 layers, it provided a model with an accuracy of the 98.97% in the training set, which is satisfactory in relation to alternative models they worked in 15 different experiments and which were compared with classification algorithms obtained in the service AutoAI of IBM Watson Studio that returned to classifier XGB as the best predictor with an accuracy of 87.1%.**

Keywords – Artificial Intelligence, Neural Network, Data Mining in Education, College Graduation, Academic Follow-Up

I. Introduction

One of the problems of universities is the neglect on the part of the student, but, which are the factors that directly and indirectly influence to predict the desertion and take action prior corresponding at the National University of Santa (UNS) - Peru? To predict similar problems in universities, authors applied Machine Learning approach based on classification algorithms as the algorithm Support Vector Machine (SVM), Forest Random (RF), Gaussian Processes (GPs) and Machines Boltzmann Deep (DBMs) [1]. On the other hand, numerous factors can affect a student's ability to graduate: college preparation and student support services provided by a university; analyzing the impact of these factors on graduation rates [2]. For Whitlock [3] the associated factors to graduate were related to the factors: institutional and academic characteristics, student aptitude and student community, which applied to predictive methods: Logistic Regression, Decision Trees, Random Forest, Artificial Neural Networks (ANNs) and SVM, allowed to predict it. Vijayalakshmi and Venkatachalapathy [4] studied the consequences of low performance in students who drop out of the degree, for this they applied the techniques of hidden two-layer ANNs using the ReLu activation function and an output layer using the Softmax activation function, on the base of three classes of prediction of exit from abandonment such as: low, medium and high, obtaining an accuracy of 85% in the training set. Therefore, it is management performance and neglect of students reviewed the Faculty of Engineering of the UNS and applied ANNs N layers to predict the factors that has connection with the desertion of students seeking to improve academic track with prediction to determine if the student completes their studies, graduates, titles or drops out.

II. Problematic Reality

A. *Reality of the National University of Santa - Peru*

From the beginning of the Systems Engineering and Computer Science degree at the UNS, to 2018 there have been 1,522 new entrants, of which only 32% have managed to obtain a bachelor's degree and 23% a Professional Title. Based on what has been shown, we can ask many questions: Do the students not respond to the demands of the subjects and / or their professional career? o Are teachers too demanding for the level of preparation of the students? Isolated questions that do not allow us to visualize the context of why these scenarios are really happening.

B. *Scope of the research*

The scope of this research is to propose a predictive model for the academic monitoring of the UNS student and to be able to predict the achievement of obtaining the bachelor's degree, the professional title, simply graduating or maintaining the status of student due to temporary or permanent abandonment.

C. *Formulation of the problem*

Given the complexity of the academic monitoring of university students that allows, through their academic performance and the various factors that influence it, to be able to determine with greater certainty who is more likely to: complete their studies, graduate, obtain their titles or, failing that, abandon the career and not getting the corresponding diploma, taking into account the characteristics of their socio-economic and academic data, to adopt the appropriate corrective, our formulation of the problem arises.

How to predict the dropout of students at the National University of Santa Peru with an accuracy greater than 90%?

III. Research Objectives

A. *Objective General*

Train a model to Predict Academic Dropout in the Faculty of Engineering of the National University of Santa-Peru.

B. *Specific Objectives*

(1) Analyze existing data analytics project methodologies. (2) Perform intake and preprocessing data related to academic dropout of students from the Faculty of Engineering of the National University of Santa. (3) Identify the technique and generate a model to predict the academic dropout indicators, seeking a precision greater than 87.1% to pass the Auto AI

978-1-6654-0719-9/21/$31.00 ©2021 IEEE

experiment with the XGB classifier. (4) Evaluate the results of the analysis of the success indicators of the case studied.

IV. Theoretical Framework

A. Student dropout

In Latin America, through the Latin American conference on dropout in higher education (CLABES), incidence is made in the study of dropout in education in relation to the quality of teaching in universities [5]. In the Management Guide Permanence Student Higher Education [6], the factors influencing the decision to the student for permanence or abandonment were analyzed. Hellas in [7] shows also the characteristics predictive investigated more frequent in relation to values for the case of the retention / dropout prognosis, the most common being: performance in pre-subjects, subjects and high school, as well as demographic data, gender, age, family and personality data [8,9].

B. Machine Learning

Machine Learning is based on techniques of Artificial Intelligence (AI) [10], supported by data science, related with mathematics, statistics, Natural Language Processing and Deep Learning [11]. The application of Machine Learning (ML) in education is distributed mostly in the prediction of student performance with 54.55%, follows in application regarding the retention of students with 22.08%, on a smaller scale, the graduation of students is addressed with 15.58% and in the end it is used for the evaluation of students with 7.79%, the above in 77 cases studied [12].

C. Artificial Neural Networks (ANNs)

A ANNs multilayers consisting of several layers fully connected nodes. The nodes of the input layer correspond to the input data set. Nodes interlayers used a logistic function (sigmoid, Relu), while layer nodes final output using a function Softmax to support multi-classification. The number of nodes in the output layer must match the number of predictable alternatives [13,14].

D. Data analytics methodology

There are different frameworks and methodologies from data mining to perform data analysis, such as: KDD, SEMMA, CATALYST and CRISP-DM, but considering that, for predictive modeling of university student dropout, we need to cover aspects of the understanding. In the university environment, a framework shown in Fig. 1 was proposed:

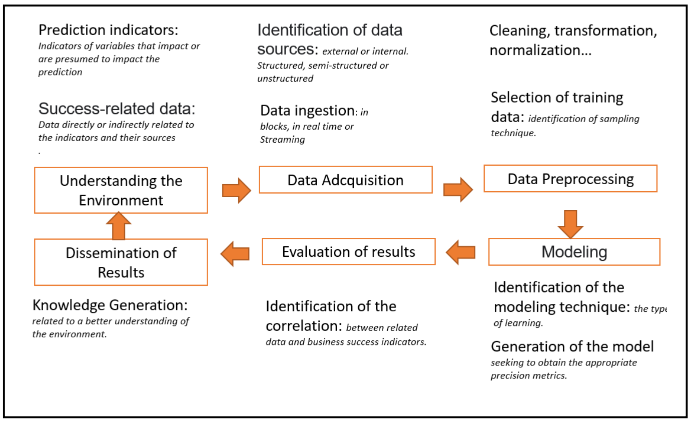

Prediction indicators: Indicators of variables that impact or are presumed to impact the prediction
Identification of data sources: external or internal. Structured, semi-structured or unstructured
Cleaning, transformation, normalization...
Success-related data: Data directly or indirectly related to the indicators and their sources
Data ingestion: in blocks, in real time or Streaming
Selection of training data: identification of sampling technique.
Understanding the Environment
Data Adcquisition
Data Preprocessing
Dissemination of Results
Evaluation of results
Modeling
Knowledge Generation: related to a better understanding of the environment.
Identification of the correlation: between related data and business success indicators.
Identification of the modeling technique: the type of learning.
Generation of the model seeking to obtain the appropriate precision metrics.

Fig. 1. Model proposed to deploy the proposed analytical predictive of dropout of students from the National University of Santa.

E. Research Hypothesis

A model based on Multilayer Neural Networks allows predicting the dropout level of university students at UNS with an accuracy level greater than 87.1%.

V. Methodology

The methodology applied in the present research followed the steps: (1) Understanding of the university environment and characteristics that can influence on the attrition university. (2) Acquisition and preprocessing of data history of the 2004-2018 semesters of the Faculty of Engineering of the UNS. (3) Identification of the modeling technique to predict the dropout of University students through experimentation with Python and Auto AI from IBM Watson. (4) Generation of the model to predict the dropout level of students from the Faculty of Engineering of the UNS.

Population: students of UNS, show: students of four Professional Schools of the Faculty of Engineering Systems and Informatics (EPISI), Energy and Physics (IPSS), Agroindustrial (EPIA) and Civil (EPIC), between the semesters 2004-2018.

VI. Results

A. Understanding the problem environment

To understand the environment of the problem, the information of the university welfare systems was analyzed: Socio-economic data, with characteristics such as services and the system of the Degrees and Titles office, anonymized from its database, with characteristics related to the state graduation, consolidating the coverage in TABLE I below.

TABLE I: MASTER DATASET

Number	Characteristic	Records	Type of data
1 (*)	Code	767	String
2	Sex	767	Int64
3	Mobile	767	Int64
4	Economic dependency	767	Int64
5	Condition of work responsible family	767	Int64
6	Total family income	767	category
7	Place of origin	767	Int64
8	Light	767	Int64
9	Water	767	Int64
10	Drain	767	Int64
11	Telephone	767	Int64
12	Cable	767	Int64
13	Internet	767	Int64
14	Averages	767	Float64
15	Number of semesters	767	Float64
16	Student	767	Int64
17	Graduated	767	Int64
18	Graduate	767	Int64
19	Titled	767	Int64

(*) The code is excluded for purposes of assessment

B. Acquisition and pre-processing of historical data from the semesters 2004 to 2018 of the U NS

The different sources mentioned above were loaded and preprocessed with Python and the Pandas library, where were working with the socio-economic data: the multivalued services feature was separated, in each of its services separately, the date of birth feature was eliminated because it was not relevant to the study, it was detected that the characteristic called student works had more than 90% of missing values, so this characteristic was discarded, in the same way, the characteristics: place of birth, marital status, type of dwelling, times who applied, housing material ; then and readjusted to the following Boolean values as sex (have) cell to 0 and 1; the characteristics: economic dependency and work condition, as well as, head of family and place of origin, of categorical types, were normalized to numerical values. On the other hand, in the records of graduated and graduates, the characteristics of the date of diploma and Professional school were eliminated.

C. Modeling techniques to predict the dropout of university students

Given the problems of the case study, a model was raised ANNs Multilayer, in order to determine that, given certain academic characteristics and socioeconomic be verified to the condition of students, graduates are able to obtain a bachelor's degree or title, said Multilayer ANNs were compared with the classification models generated by the AutoAI service of IBM Watson Studio. The initial model, in March year 2020, consists of 14 features, one hidden layer of 14 nodes and an output layer with four nodes, which coincide with the four classes motif prediction. Rectified Linear Unit (ReLU) linear activation was used for the hidden layer and Softmax activation for the output layer. Maintaining the structure of the previous model, in order to select the model that provides the best precision, we experimented with 14 additional models that varied in the number of hidden layers and number of nodes in each hidden layer, but both the input layer (14 features) and the output layer (4 prediction classes) remained unchanged, as shown in Fig. 2.

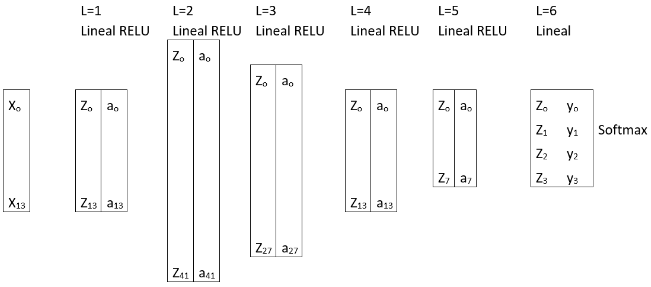
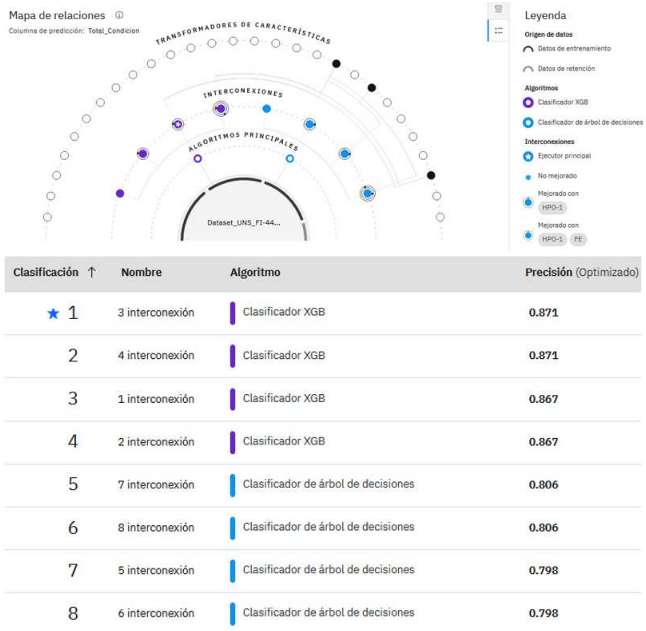


Fig. 2. Architecture of the six layer Neural Network.

In one first test with the service AutoAI of IBM Watson Studio and with one dataset from a professional school of UNS 767 records were identified to the algorithms: classifier XGB and classifier LGBM as the most accurate. Then, we worked with a data volume of 3529 records, from four Professional Schools, which made AutoAI identify the algorithms: XGB and the decision tree classifier, as the most accurate as shown in Fig. 3. Where the algorithms have 4 and interconnections made 28 transformations of characteristics, generating an accuracy level of 87.1%.



Clasificación ↑	Nombre	Algoritmo	Precisión (Optimizado)
★ 1	3 interconexión	Clasificador XGB	0.871
2	4 interconexión	Clasificador XGB	0.871
3	1 interconexión	Clasificador XGB	0.867
4	2 interconexión	Clasificador XGB	0.867
5	7 interconexión	Clasificador de árbol de decisiones	0.806
6	8 interconexión	Clasificador de árbol de decisiones	0.806
7	5 interconexión	Clasificador de árbol de decisiones	0.798
8	6 interconexión	Clasificador de árbol de decisiones	0.798

Fig. 3. AutoAI experiments with 3529 records.

D. Generation of the model to predict the level of dropout students

The implementation of the model was carried out with the Python Machine Learning libraries: Scikit-learn, for Deep Learning Tensorflow was used, for the preprocessing of the Scipy data. To reduce the loss, the following was used: Softmax cross entropy with logits, the model was configured with a learning ratio of 0.0001, with mini batches of size 32 and initially it was worked with 1500 iterations, and in the backpropagation the Adam optimizer was considered.

In a first batch of experiments, these were carried out with a dataset from the Professional School of Systems Engineering of the UNS with 767 records. 80% was considered for the training set and the remaining 20% for the test set. Even though the training set of experiment 13 (6 layers) has an accuracy of 82.87%, which is relatively high, for the test set it has too low an accuracy of 64.94%. Under this circumstance, the increased number of iterations of training model. In the second set of experiments, the periods of 1500 to 12000. was varied were performing intermediate tests, up to limit for improvement, presented in TABLE II.

TABLE II: PRECISION IN PREDICTION OF THE MULTILAYER NUERONAL NETWORKS ALGORITHM WITH 12000 ITERATIONS.

Exp.	Layers-Nodes	Set of Training	Set of Proof
14	28,56,42,28,14,8,4	99.47%	79.18%
13	56,42,28,14,8,4	98.97%	81.73%
12	42,28,28,14,14,4	96.95%	77.05%
10	42, 28, 14,8,4	96.00%	78.61%
11	14, 42, 28, 14,8,4	93.59%	80.31%
05	28, 20, 14,8,4	92.84%	83.00%
09	42, 14,10,4	92.49%	82.01%
07	42, 14,8,4	91.82%	81.16%
06	28, 14, 14,8,4	91.57%	83.14%
03	28, 14,8,4	90.19%	83.85%

When reviewing the results, we found that in all cases there has been an improvement in the accuracy in the training set.

VII. Discussion of Results

Regarding the characteristics that affect student persistence or dropout, we selected 12 demographic and 02 academics and we agree with Hellas [7] and Musso [15]. Considering also that multiple variables can influence a prediction of student performance, such as the high school average, gender, the courses that have a prerequisite [16,17,18], or the days of absence of the students [4], as well as, their age [19].

The proposed model can predict whether the student can complete studies, obtain a bachelor's degree, obtain a professional degree, or drop out and remain in the condition of a student without an end date. The evaluation of the predictive model, as did Chanlekha and Niramitranon in [18] is made with the prediction performance, plus the metric recall, f1-score, the loss and confusion matrix, using the findings, the precision metric.

The historical data shows the percentages low graduation and degree at the UNS, with the prediction model would increase the graduation rate in the range from 28.89% to 50.13%, and for titles improvement would be in the range of 40.37% to 58.47%.

Conclusions

A framework based on SEMMA, CATALYST and CRISP-DM was proposed, considering the understanding phase of the university student dropout environment.

It was determined that the variables that offered the greatest influence on the prediction were the academic variables number of semesters and averages, and the demographic variables were the internet and the cell phone.

The model implemented ANNs from 02 layers to 07 layers and combinations of numbers of neurons per layer, additionally we evaluated against the results of the XGB Sorter and tree classifier decisions, with best results obtained with the model of ANNs 6 layers with the sequence of nodes (56, 42, 28, 14, 8, 4), with which it is possible to identify both students with a high risk of dropping out who will not be able to complete and obtain the corresponding diploma, with a precision 81.73% in the test set and 98.97% in the training set.

The proposed prediction model manages to predict dropout and improve the academic follow-up of the students of the Professional Engineering Schools of the UNS, in ranges that go from 28.89% to 58.47% per school either graduation or degree, with this we will have certainty of those who could fail, but it is necessary to implement a student retention and graduation management system.

References

[1] C. G. Nespereira, E. Elhariri, N. El-Bendary, A. F. Vilas, and R. P. D. Redondo, "Machine Learning Based Classification Approach for Predicting Students Performance in Blended Learning," *Adv. Intell. Syst. Comput.*, vol. 407, pp. 47–56, 2016, doi: 10.1007/978-3-319-26690-9_5.

[2] T. Ojha, G. L. Heileman, M. Martinez-Ramon, and A. Slim, "Prediction of graduation delay based on student performance," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2017-May, pp. 3454–3460, Jun. 2017, doi: 10.1109/IJCNN.2017.7966290.

[3] J. L. Whitlock, "Using Data Science and Predictive Analytics to Understand 4-Year University Student Churn," 2018, Accessed: Jul. 13, 2021. [Online]. Available: https://dc.etsu.edu/etd/3356.

[4] V. Vijayalakshmi and K. Venkatachalapathy, "Comparison of Predicting Student's Performance using Machine Learning Algorithms," *Int. J. Intell. Syst. Appl.*, vol. 11, no. 12, pp. 34–45, 2019, doi: 10.5815/ijisa.2019.12.04.

[5] CLABES, "CLABES 2019 - Universidad del Rosario," 2019. https://www.urosario.edu.co/CLABES/inicio/ (accessed Jul. 13, 2021).

[6] Ministerio de Educación Nacional, *Guía Para La Implementación del modelo de gestión de permanencia y graduación estudiantil en instituciones de Educación Superior*. 2015.

[7] A. Hellas *et al.*, "Predicting academic performance: A systematic literature review," *Annu. Conf. Innov. Technol. Comput. Sci. Educ. ITiCSE*, pp. 175–199, Jul. 2018, doi: 10.1145/3293881.3295783.

[8] A. Salvador Blanco, Laurentino; García-Valcárcel Muñóz-Repiso, "El rendimiento académico en la universidad de Cantabria: abandono y retraso en los estudios - Publicaciones - Ministerio de Educación y Formación Profesional," 1988. https://bit.ly/3F99TEk (accessed Jul. 13, 2021).

[9] M. Marta Ferreyra, C. Avitabile, J. Botero Álvarez, F. Haimovich Paz Sergio Urzúa, and D. Humano, "Momento decisivo La educación superior en América Latina y el Caribe," 2017.

[10] J. Hurwitz and D. Kirsch, "Machine Learning for Dummies: Understand machine learning fundamentals," 2018.

[11] D. Sarkar, R. Bali, and T. Sharma, "Practical Machine Learning with Python," *Pract. Mach. Learn. with Python*, 2018, doi: 10.1007/978-1-4842-3207-1.

[12] D. Kučak, V. Juričić, and G. Đambić, "MACHINE LEARNING IN EDUCATION-A SURVEY OF CURRENT RESEARCH TRENDS," pp. 406–0410, 2018, doi: 10.2507/29th.daaam.proceedings.059.

[13] B. Quinto, "Next-generation machine learning with spark: Covers XGBoost, LightGBM, Spark NLP, distributed deep learning with keras, and more," *Next-Generation Mach. Learn. with Spark Cover. XGBoost, Light. Spark NLP, Distrib. Deep Learn. with Keras, More*, pp. 1–355, Jan. 2020, doi: 10.1007/978-1-4842-5669-5.

[14] M. Feurer and F. Hutter, "Hyperparameter Optimization," pp. 3–33, 2019, doi: 10.1007/978-3-030-05318-5_1.

[15] M. F. Musso, C. F. R. Hernández, and E. C. Cascallar, "Predicting key educational outcomes in academic trajectories: a machine-learning approach," *High. Educ. 2020 805*, vol. 80, no. 5, pp. 875–894, Mar. 2020, doi: 10.1007/S10734-020-00520-7.

[16] Bendangnuksung and D. Prabu, "Students Performance Prediction Using Deep Neural Network," 2018.

[17] L. D. F. Zea, Y. F. P. Reina, and J. I. R. Molano, "Machine Learning for the Identification of Students at Risk of Academic Desertion," *Commun. Comput. Inf. Sci.*, vol. 1011, pp. 462–473, 2019, doi: 10.1007/978-3-030-20798-4_40.

[18] H. Chanlekha and J. Niramitranon, "Student performance prediction model for early-identification of at-risk students in traditional classroom settings," *MEDES 2018 - 10th Int. Conf. Manag. Digit. Ecosyst.*, pp. 239–245, Sep. 2018, doi: 10.1145/3281375.3281403.

[19] M. A. Ruiz Palacios, "Factores que influyen en la deserción de los alumnos del primer ciclo de educación a distancia en la Escuela de Administración de la Universidad Señor de Sipán. Períodos académicos 2011-1 al 2013-1: lineamientos para disminuir la deserción," 2018. http://revistas.pucp.edu.pe/index.php/educacion/article/view/19924/19953 (accessed Jul. 13, 2021).