2017 16th IEEE International Conference on Machine Learning and Applications

Early prediction of college attrition using data mining

Luiz Carlos B. Martins
University of Brasilia (UnB)
Brasilia, DF, Brazil
Email: luizmartins@unb.br

Rommel N. Carvalho
University of Brasilia (UnB)
Brasilia, DF, Brazil
Email: rommelnc@unb.br

Ricardo S. Carvalho
Ministry of Transparency, Monitoring and Control (MTFC)
Brasilia, DF, Brazil
Email: ricardo.carvalho@cgu.gov.br

Márcio C. Victorino
University of Brasilia (UnB)
Brasilia, DF, Brazil
Email: mcvictorino@uol.com.br

Maristela Holanda
University of Brasilia (UnB)
Brasilia, DF, Brazil
Email: maristela.holanda@gmail.com

***Abstract*—College attrition is a chronic problem for institutions of higher education. In Brazilian public universities, attrition also accounts for the significant waste of public resources desperately needed in other sectors of society. Thus, given the severity and persistence of this problem, several studies have been conducted in an attempt to mitigate undergraduate dropout rates. Using H2O software as a data mining tool, our study employed parameter tuning to train 321 of three classification algorithms, and with Deep Learning, it was possible to predict 71.1% of the cases of dropout given these characteristics. With this result, it will be possible to identify the attrition profiles of students and implement corrective measures on initiating their studies.**

***Index Terms*—Data Mining, Classification, University Evasion, Deep Learning**

I. Introduction

One of the great challenges of higher education is to ensure that the majority of the students who enter complete their desired degree programs. However, university attrition is a persistent problem among Brazilian universities [1]. According to [1], based on data obtained from the National Institute for Educational Studies and Research Anísio Teixeira (INEP, *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira*) university dropout rates in Brazil reached an average of 22% with growth trends. Among the various factors that influence university attrition, [2] cites that the main factors are: ability to study, adaptation to the course, academic training prior to joining the course and vocation-profession.

In Brazil, to enter an institute of higher education, each candidate must complete a selection process, which evaluates the students' knowledge, acquired during the levels prior to higher education, especially high school. Public institutions use three types of instruments to select potential candidates:

- **Vestibular**: Test applied by the institution itself to select students.
- **Serial Evaluation Program (PAS, *Programa de Avaliação Seriada*)**: The candidate undergoes tests during high school, where the level of his or her performance is used as the criterion to classify for a vacancy.
- **Unified Selection System (SISU, *Sistema de Seleção Unificado*)**: In this system, the institutions offer a number of vacancies in certain Majors and from the performance they obtained in the National High School Examination (ENEM, *Exame Nacional do Ensino Médio*), a ranking is established, such that vacancies are filled by the highest scoring candidates.

University attrition is a problem that institutions face on a daily basis. However, early identification of students who have a profile that indicates the tendency to drop out of their Major can be a solution, since corrective measures may be applied before the student considers abandoning the course. In this effort, the use of machine learning can be an effective instrument. According to [3] machine learning is the technique of knowledge discovery using computational resources to detect hidden patterns in data through an implicit analysis. Machine learning is also linked to other areas of science such as artificial intelligence and statistics. Data mining is the application of algorithms in data sets for knowledge discovery. In this context, this paper proposes data mining for the identification of students with an attrition profile, so that corrective measures may be adopted early on and these students may complete their university courses.

The article is organized as follows: In Section II we present related works. In Section III we show the steps that were used to carry out the experiment, as well as the resources used, the technologies adopted, how the evaluation of the model was carried out and the result of the experiment. Finally, Section IV discusses the conclusion as well as possible future work.

II. Related work

Based on official data published by INEP, [1] studies the factors that influence university attrition in Brazil. An explicit data analysis was performed to identify a negative correlation between student dropout rates and demand, concluding that the higher the demand, the lower the dropout rate. The study did not aim to propose solutions to the proble

0-7695-6321-X/17/31.00 ©2017 IEEE
DOI 10.1109/ICMLA.2017.000-6

but rather attempted to explain the phenomenon observed in universities. One conclusion of the work is that the Brazilian situation resembles the reality of other countries.

The study of [4] investigated attrition at the Federal University of Minas Gerais (UFMG) from data involving five Majors reporting high dropout rates. Data were collected from three sources: The Department of Registration and Academic Control (DRCA), the analysis of documents regarding institutional rules that regulate attrition, and qualitative research through interviews with students. The three main reasons that influence dropout identified in this study were: performance during the course, exclusive dedication to studies and lack of appropriate qualification for the Major.

The research carried out in [1] and [4] observed the phenomenon through manual statistical analysis and the use of some automated technique for the discovery of knowledge.

There are also several other studies that use classification to identify students' performance. [5] uses the classification algorithms C4.5 and Random Tree from the data of the students of the group Guru Gobind Singh Indraprastha University. [6] also seeks to classify Vikram University students at risk for attrition using the algorithms C4.5, CART and ID3.

In a study carried out with data from the University of Brasília (UnB), it was possible to identify students in the group of risks of abandoning graduation through Machine Learning techniques. Data were collected that identify the profile of the students as sex, age or types of high school as well as data of educational performance during the course. Data were divided into 4 groups by age and course criteria. Students from courses related to information technology and ANN algorithms, Linear Regression, Naive Bayes, Random Forest, SVR and ZeroR were selected. The metric used to evaluate the performance of the models was the F-measure, and in three achieved the value of the median F-measure reached 0.8 or more [7].

The work sought the early identification of students at risk for attrition using data mining. To achieve higher precocity of identification, this study restricts that the analyzed variables were generated before the student entered the Major, while other studies analyzed a classification using student variables during the course. We identified that this activity falls under the classification technique. We chose to use the platform H2O[1]. According to [8] H2O is a platform to aid in the discovery of knowledge in data, has a set of algorithms for data mining as resources, allows parallel processing, the use of distributed clusters and also the creation of grid, where several models are created, and are generated through the combination of parameters. In order to facilitate the work, we selected three algorithms of those already implemented in H2O that have classification characteristics: Deep Learning, Distributed Random Forest and Gradient Boosting Machine.

III. Methodology

The CRISP-DM Methodology (Cross Industry Standard Process Data Mining Model) was used to develop the work. CRISP-DM divides the work into six steps: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. It has the characteristic of being a cyclic process and the various stages communicate with each other, where the process can return to the previous steps to optimize its results [9].

A. Business Understanding

In 1998 the federal government created the ENEM, which aims to evaluate students who have completed or will complete high school and assess the quality of schools [10]. In 2010 the SISU was created, in which public universities offer their available places and the candidates use the scores that they obtained in the last ENEM to apply for the course they wish [11]. This process standardized the selection of students in the federal institutes of higher education so that the comparison between institutions became more feasible.

[12] defines that university attrition is the interruption of the study cycle, being considered a complex social phenomenon that has negative social, academic and economic impacts. As seen previously in [1], the Brazilian rate of university dropout reaches 22%, which makes it necessary to take preventive and corrective measures to mitigate this situation and to help improve the educational level of the population. This policy is referred to as "affirmative action".

B. Data Understanding

This article aims to address how the evaluation of candidates entering undergraduate programs can predict if the student will complete the degree. As explained, the selection is made through tests that classify the candidates considered most suitable to the Major. Each institution has autonomy to apply the vestibular or the PAS in the way that it deems convenient, which makes this process unaddressed. SISU, by using scores obtained by the ENEM candidate to classify him or her, can standardize the selection, making the data more homogeneous.

The ENEM is held over a period of two days, in which the candidate answers questions in five test areas: Human Sciences and their Technologies (CHT, Ciências Humanas e suas Tecnologias), Natural Sciences and their Technologies (CNT, Ciências da Natureza e suas Tecnologias), Languages, Codes and their Technologies (LCT, Linguagens, Códigos e suas Tecnologias), Mathematics and their Technologies (MT, Matemática e suas Tecnologias) and Writing. In each test, the candidate receives a score ranging from 0 to 1,000.

We also used data from the individual (age and sex), institutional (course, course city, university) and whether the candidate was selected for any social quota or ample competition. Finally, we have synchronic data of the individual's current situation: if he/she has already completed the course, if he/she is still a student of the institution or if he/she dropped out.

C. Data Preparation

In the preparation of the data, a process Extraction was carried out, as well as, Transformation and Loading (ETL).

[1]H2O is an open source machine learning platform https://www.h2o.ai/

The Extraction phase is where the data is collected from databases of any origin. For this, we used the Electronic System of the Information Service (e-SIC, *Sistema Eletrônico do Serviço de Informação ao Cidadão*) of the Brazilian government, where any citizen can request data of the federal government for any public agency. We specify that the data should come in a structured format, preferably in Comma-separated values (CSV), containing student data of their personal characteristics (gender and age), academic (Major, city where they study, if they dropped out and reason for dropping out) and the way they were admitted into to the institution (vestibular, PAS, SISU, selection notes and affirmative action). The Federal Fluminense University (UFF) answered the request according to the specifications presented and sent its data.

Transformation is the stage where cleaning and organization is performed. Various data sources, in addition to cleaning, need to be standardized to be comparable. In the cleanup, elements that were corrected were: accentuations and the space of compound words, standardization of nomenclatures of affirmative actions, numerical data converted to international standards, such as 701,3 to 701.3, and the elimination of unnecessary fields in numerical record identifiers.

At the Load step stage, where data is entered into a repository to be queried by the applications, we created a Data Warehouse (DW) using database management systems MYSQL[2]. Figure 1 denotes the steps that were used to prepare the data.

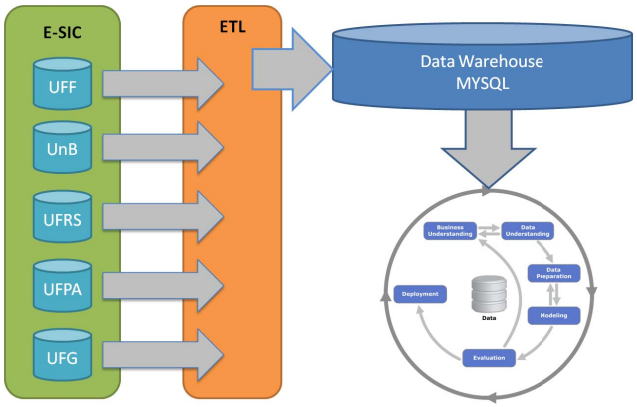

Fig. 1. Data Preparation

In the last step of the preparation, the data is exported to CSV files that are to be used in the modeling process.

Each Major represents a specific area of knowledge, and the ENEM scores will influence dropout rates that will vary according to the Major to which they are applied. Therefore, we chose to use data from only one Math Major, to delimit the problem. In total, we obtained 1,369 records to carry out the study. Table I looks at the fields used for data mining.

[2]https://www.mysql.com/

TABLE I
DATAFRAME FIELDS

Field	Type
City	Character
Age	Numeric (0-99)
Gender	Character[M/F]
Affirmative action	Character
CHT	Numeric (0-1000)
CNT	Numeric (0-1000)
LCT	Numeric (0-1000)
MT	Numeric (0-1000)
Writing	Numeric (0-1000)
Situation	Character [No Attrition/Attrition]

The last field addressed is whether the student has left the course or not, so we created the field "Situation" to include this information.

D. Modeling

A machine with 16Gb of RAM and 8 processing cores was used. We combine R and H2O software as tools for modeling.

The records were split in 80% for training and 20% for the testing. Cross-validation [13] is a statistical technique for the evaluation and comparison of prediction models that divide them into folds that are used to train and validate the cross-model. We used cross-validation with 10-folds to validate the model.

To optimize the training, we used H2O's parallel processing option, which allows tasks to be divided between processors and executed simultaneously, thus reducing the time spent. Another H2O feature we used was the construction of model grids, where we defined hyper parameters that are trained to generate a model for each combination. Following the description of the parameters according to H2O documentation [8]:

- Learn_rate: Specify the learning rate.
- max_depth: Specify the maximum tree depth.
- ntrees: Specify the number of trees to build.
- L1: Specify the L1 regularization to add stability and improve generalization; sets the value of many weights to 0.
- L2: Specify the L2 regularization to add stability and improve generalization; sets the value of many weights to smaller values.
- sample_rate: Determines the sampling rate of the X axis

Table II shows which were the hyper parameters used in the construction of the grids of models of each algorithm.

As the objective is to correctly identify cases of attrition, the models were evaluated by minimizing the "False Negatives". For example, cases in which the record is classified as "No Attrition", but actually "Attrition", because we identified that the error in the identification of a student at risk for dropping out has a much higher cost for the institution than to classify a student as a risk of attrition when this risk does not exist.

In each of the chosen algorithms, the *h2o.grid* function of the H2O package was used to generate a grid of models from the combination of several parameters. At this stage,

TABLE II
HIPER PARAMS

Algorithms	Hiper Params
DP	l1: 0.001, 0.01, 0.1 l2: 0.001, 0.01, 0.1 hidden:(10,10), (20,20), (50,50)
DRF	max_depth: 2, 10, 50 ntrees:10, 50, 100 sample_rate: 0.1, 0.2 ... 1.0
GBM	learn_rate: 0.1, 0.5, 1.0 max_depth: 10, 20, 30 ntrees: 10, 50, 100 sample_rate: 0.001, 0.01, 0.1

we used the separate data for training. After this, from the information of the confusion matrix, we identified which had the capacity to minimize false positives, while the specificity did not have a value considered too low. We selected the models with these characteristics and ran the test data to assess which of them would perform best. The following Table III shows the performance in the identification of false positives by algorithm.

E. Evaluation and Deployment

In total a grid of 321 models were trained: 81 using the algorithm Gradient Boosting Machine, 108 using Deep Learning and 132 Distributed Random Forest. From each algorithm, we selected the model that presented the highest false-negative rate and with not very low specificities, to avoid models obtaining a high false-positive rate by classifying the records in only one class.

In the selected models, the test data was applied to verify the performance and to verify which model could identify the highest percentage of university attrition cases. Table III shows the values of the results of the algorithms.

TABLE III
COMPARING MODELS

Algorithm	True Positive Rate	True Negative Rate
GBM	63.2%	43.7%
DL	71.1%	39.4%
DRF	55.4%	39.4%

From Table III we identified that DRF was able to generate a rate of 55.4% correctly cases of attrition, while GBM of hit cases of 63.24%. Finally, we have that DL hit 71.1% of the cases. On the other hand, we have the rate of true negatives – when a student is predicted to drop out, but he or she does not – which ranged from 42.2% in GBM, to 43.7% in DL and 39.4% in DRF, respectively.

According to our analysis, DL has shown to be more efficient in forecasting our interest class despite having a true negative value slightly higher than the others, this model being the most suitable for predicting attrition.

IV. CONCLUSION AND FUTURE WORKS

Given that the objective of this work is to prevent, as much as possible, the cases of university attrition, we used data that was generated before the student started their University programs, for the training of the models. In this context, the identification of the largest number of possible cases of attrition was our goal. In this work, the use of the Deep Learning algorithm was able to correct 71.1% of the cases of attrition. Considering these preliminary conditions provided by this data, universities are given the the chance to implement corrective actions as soon as the student starts attending classes. This result was possible through a very restricted environment only one university Major, at a single university. Thus, further research, in a broader range of contexts, is necessary to map patterns, which may corroborate or provide a variance to these results.

Future work will include the application of the methodology in data collection, but with regard to different Majors, verifying performance. In addition, other strategies can be performed to try to increase the performance among them. Similarly, other measures include collecting data at a greater number of institutions in order to increase the training data and to identify and collect other variables that may help in the model as the competition rate of the chosen Major. Other variables to consider is the candidate's second choice of a Major, and the distance from the student's home city and the university where he or she will study. This research may also be advanced by including data pertaining to student academic performance in the first semesters of study.

REFERENCES

[1] R. L. L. Silva Filho, P. R. Motejunas, O. Hipólito, and M. Lobo, "A evasão no ensino superior brasileiro," *SciELO Brasil*, vol. 37, pp. 641–659, Sep. 2007.

[2] ANDIFES, ABRUEM, SESu, and MEC, "Diplomação, Retenção e Evasão nos Cursos de Graduação em Instituições de Ensino Superior Públicas," Comissão Especial Sobre a Evasão nas Universidades Públicas Brasileiras, Tech. Rep., 1996.

[3] F. Amaral, *Introdução à Ciencia de Dados: mineração de dados e big data*. Alta Books Editora, 2016.

[4] A. A. C. T. Adachi, "EVASÃO E EVADIDOS NOS CURSOS DE GRADUAÇÃO DA UNIVERSIDADE FEDERAL DE MINAS GERAIS," Ph.D. dissertation, UNIVERSIDADE FEDERAL DE MINAS GERAIS, Dissertação (Mestrado em Educação). Faculdade de Educação–Programa de Pós-Graduação em Educação. Universidade Federal de Minas Gerais. Belo Horizonte, 2009.

[5] M. Tripti, K. Dharminder, and S. Gupta, "Mining Students' Data for Prediction Performance," Feb. 2014.

[6] K. Bunkar, U. K. Singh, B. Pandya, and R. Bunkar, "Data mining: Prediction for performance improvement of graduate students using classification." IEEE, Sep. 2012.

[7] G. Silva and M. Ladeira, "A Machine Learning Predictive System to Identify Students in Risk of Dropping Out of College," vol. I, 2017, pp. 62–68.

[8] T. H. team, "R Interface for H2o," Jun. 2017. [Online]. Available: https://github.com/h2oai/h2o-3

[9] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0 Step-by-step data mining guide," 2000.

[10] E. A. Teixeira, "Enem: documento básico," 1998.

[11] "Portaria Normativa MEC nº 2, de 26 de janeiro de 2010," Jan. 2010. [Online]. Available: http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=2704-sisuportarianormativa2&Itemid=30192

[12] N. P. d. L. Gaioso, "O fenômeno da evasão escolar na educação superior no Brasil," Dissetação, Universidade Católica de Brasília, Brasília, 2005.

[13] L. Liu and M. T. Özsu, *Encyclopedia of database systems*. Springer Berlin, Heidelberg, Germany, 2009, vol. 6.