A Brief Review about Educational Data Mining applied to Predict Student's Dropout

G. A. S. Santos[1], A.L. Bordignon[2], S.L.G. Oliveira[3], D.B.Haddad[1], D.N.Brandão[1], K.T.Belloze[1]

[1] PPCIC - Programa de Pós-graduação em Ciências da Computação (CEFET/RJ)
Caixa Postal 20271-204 – 455, Maracanã, Rio de Janeiro - RJ

[2]Instituto de Matemática
Universidade Federal Fluminense
Niterói - RJ

{kele.belloze, diego.brandao}@cefet-rj.br

[3]Universidade Federal de Lavras
Lavras - MG

sgonzaga@dcc.ufla.br

Abstract. *Educational Data Mining (EDM) may be a very useful technique as much to understand student behavior as to plan and manage government investments in education. EDM helps to analyzes and to expose the hidden information of educational data. Particularly, an important application of EDM is to predict or analyze the students' dropout. This problem affects several educational institutions in Brazil and the world, and identify its origin has been a relevant research motivator. This paper presents a brief introduction about EDM applied to predict students' dropout and analyzes some important articles during the period from 2013 to 2018.*

1. Introduction

Brazilian society has suffered annually financial damages when the students occupy the vacancies and dissociate themselves from the universities without completing the course in which they registered, causing what we call school dropout. The vacancies that were formerly occupied by these students now become idle and will hardly be fulfilled. Data from the census of Higher Education in Brazil for the year 2016 show more than 10.6 million vacancies in undergraduate courses were offered, being 73.8% new vacancies and 26.0% remaining vacancies. About the remaining vacancies, only 12.0% were occupied [INEP 2017]. According to Organization for Economic Co-operation and Development (OCDE), in 2013 the average annual cost per undergraduate student in Brazilian public education was $15,771.67 per student [Nascimento and Verhine 2017].

School dropout is a topic that reaches the most diverse Higher Education Institution (HEIs) in the world, both public and private. The identification of the origin of this problem has been a subject of study for researchers in education. According to the Special Committee on Evasion Studies of the MEC, the dropout concept can be characterized in the following ways [de Evasão 1996]:

- Dropout of course: the student disconnects from the higher course in situations such as transfer of course, abandonment (course or discipline) or exclusion by institutional norm;

- Dropout of the institution: the student gives up the institution in which is enrolled;
- Dropout of the system: the student permanently or temporarily abandons higher education.

Brazil has sought to stimulate higher education through various mechanisms that favor access to a Higher Education Institution. However, the amount of wasted investments due to dropout is much higher than desired [Silva Filho et al. 2007]. The educational institutions in Brazil and in the world have devoted a lot of attention in order to understand the causes of this disorder [Rodriguez 2014].

The reasons for a student's decision whether or not to pursue a course or institution are diverse and vary at levels of personal, social and institutional. These problems can be situated from the student's aptitude, personal context and vocational guidance, even to the inadequate infrastructure of the HEI or faculty still in training [Baggi and Lopes 2011].

The work of [Silva Filho et al. 2007] also mentions the question of losing students who start but do not finish their courses; which is considered a social, academic and economic waste. More explicitly, the authors detail that in the public sector, the investment is due to tax collection and with evasion, there is no return to society. In the private sector, avoidance leads to loss of revenue. Therefore, dropout becomes the source of the idleness of teachers, employees, equipment, and physical space.

Thus, it becomes understandable to question that, given a student of an HEI and its average cost, what efficient actions should be applied to reduce the impact of dropout? And, given the length of stay of this student in HEI, when these actions should be taken?

In last years, educational institutions have acquired a huge data about students that might be used to assist in the process of understanding some of these questions. Educational data mining (EDM) is an area of computer science that has been very promising in the analysis of these data. EDM research has mainly been done to analyze students retention and to predict students dropout [DeBerard et al. 2004, Yukselturk et al. 2014, Dutt et al. 2017].

In order to understand the domain of scientific studies that address this central issue and, mainly, to identify the most significant contributions regarding the problem of school dropout, this paper proposes a brief review of the literature, highlighting the recent work on the topic that make use of EDM techniques.

This work consists of three more sections. Section 2 presents the growth of published scientific works on the EDM area. Section 3 describes several studies on EDM applied to students' dropouts. Section 4 describes the final remarks on this work.

2. Educational Data Mining

Considered as an interdisciplinary area, Educational Data Mining (EDM) uses data mining and machine learning techniques associated with pedagogical concepts to educational data sets aiming solving academic and educational issues [Baker and Yacef 2009, Romero and Ventura 2010, Dutt et al. 2017].

The research in EDM increased in the last two decades. A search realized in Scopus Database using a query string "educational data mining" demonstrate it (Figure 2). In 2012 there were 60 articles and this number goes up to 235 in 2017. This raise rate demonstrates the importance EDM has received in the scientific area.

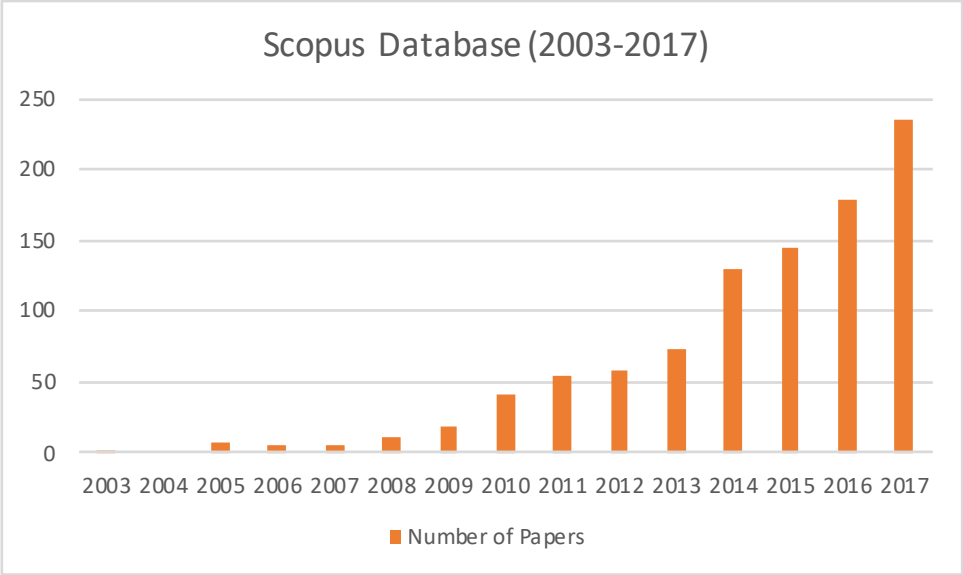

Figure 1. Number of Papers about EDM in Scopus Database in the period from 2003 to 2017.

According to [Baker et al. 2010], the works in EDM may be categorized in five class: 1-prediction; 2-clustering; 3-relationship mining; 4-distillation of data for human judgment, and; 5-discovery with models. The first three classes are very common and traditional in data mining research. The class number four consists of visualization and statistics analysis. The last class is the most recent in EDM and it consists of the development of a model to describe the problem and this model is associated with another technique as a new component.

In the context of student's dropout we observed that the main articles have been in prediction class. We obtained 35 articles in Scopus combining the terms "educational data mining" and "dropout". From those, 22 were about prediction techniques, 6 were about clustering and 7 combining both. A specific search for the Brazilian context was also performed using two terms, "brazil" and "brazilian", which were added to the above query string. Considering the first term, no articles were returned and for the second one four articles were returned, which are describe in the next section.

3. EDM Applied to Students' Dropout

Students' dropout is an extremely important element for enrollment management as it affects not only the ranking of universities, but also school reputation and financial support [Delen 2010]. This problem has been the focus of some works in the area of computing, which use the EDM techniques for the analysis of data from educational environments. In the international context, we can highlight recent works such as those by [Sarra et al. 2018], [Ahuja and Kankane 2017] and [Sultana et al. 2017], who analyze data from undergraduate students with the objective of identifying students with higher chances of dropout. Sarra *et al.* (2018) used a Bayesian Profile Regression method to analyze the questionnaire responses of more than 500 students who answered on specific questions aimed at retrieving information about academic competencies, motivations, and resilience. The authors verified the relevant factors to identify the groups with the high-

est chance of dropout, including items from the Academic Resilience Scale, scores of motivational items, scores on difficulties and satisfaction during academic life.

Ahuja and Kankane (2017) used several algorithms such as Naïve Bayes, Logistic regression, K-nearest neighbors, Ctree, Rpart, Random Forest and C5.0 to predict the probability of completion of the undergraduate course. Data used included student grades and social and demographic data. The authors also compared the methods. The experiments showed that the Random Forest and Ctree algorithms were the best for classification and prediction of the results. The authors point out that non-academic data contribute to improved outcomes.

The work of Sultana *et al.* (2017) analyzed data from Electrical Engineering students, provided through different questionnaires. The work made use of different methods such as Decision Tree, Logistic Regression, Naïve Bayes, and Neural Networks to explore cognitive and non-cognitive characteristics of students to predict dropout results. The authors describe that cognitive aspects improve predictive accuracy in decision tree methods, but not in other methods. In a complementary way, the work of [Burgos et al. 2017] also makes use of Logistic Regression for the grades analysis of more than 100 students of several distance learning courses. The aim of the work is the proposal of a method for dropout detection that allows producing, in a timely manner, a tutoring plan of action to prevent it. The results showed that the prediction combined with the plan of action helped reduce dropout during the 2014/15 school year compared to the other years when such an approach had not been implemented. Another important work presents a compilation of several works in the area [Chaturvedi 2017]. The author does not analyze a data set, however, presents some tools that can be used to analyze the data sets and summarizes the algorithms used in the area describing their basic characteristics.

In Brazil, we highlight the works of [Couto and Santana 2017], [da Cunha et al. 2016], [Manhães et al. 2014a] and [Manhães et al. 2014b]. Do Couto and de Santana (2017) present a paper that aims to create subsidies to assist managers of higher education institutions in identifying students prone to dropout or retention in their courses. For this, the authors used classification algorithms applied to data of several undergraduate students. The Random Forest and the Bayesian Network methods were the most satisfactory to analyze the data and thus to assist the managers. The work of da Cunha *et al.* (2016) analyzed data from courses at different education levels to detect which attributes on the students most influenced school dropouts and disapprovals in order to draw a profile of dropout and disapproval situations. To do this, a Decision Tree method and the Analysis Services tool were applied. From the obtained results, the authors raise some preventive measures for the institution managers to minimize the rates of dropout and disapproval.

The work of Manhaes *et al.* (2014a) also aims to assist the institution managers. The authors identified attributes that help detect students who are poorly performing or at dropout risk. Data from undergraduate students of a large educational institution were analyzed using the following classification algorithms: Decision Tree, Support Vector Machine, Naïve Bayes, and Multilayer Perceptron Neural Network. The Naïve Bayes model was used to present a quantitative approach. The authors estimate that the use of EDM methods helps to identify the students with greater chances of dropout and still enrolled, as well as to offer the institution an additional analysis form for the problem of

high dropout rate. In a subsequent work, Manhaes *et al.* (2014b) presents an architecture named WAVE that uses EDM techniques to predict and identify students who are at dropout risk. The architecture uses only student data stored in the academic management system, not requiring social or economic data.

4. Final Remarks

Research in Educational Data Mining has increased significantly over the last years. Particularly, EDM's application to educational dropout prediction has been the main interest for many researchers. In this context, this paper presented a brief introduction about EDM applied to predict students' dropout and analyzed some important articles during the period from 2013 to 2018.

Designing a prediction model to describe students' dropout proved to be a hard task. Different algorithms have been used, as Decision Tree, Regressions, Neural Networks and others, but the results did not reach a satisfactory precision yet. Although many approaches to predict dropout have focused mainly on academic data, some research indicate that it is important to combine strategies from approaches based on academic and non-academic data. New studies should be conducted to evaluate different methods of machine learning as kernel methods, in addition to evaluations involving the combination of different types of data.

References

Ahuja, R. and Kankane, Y. (2017). Predicting the probability of student's degree completion by using different data mining techniques. In *Image Information Processing (ICIIP), 2017 Fourth International Conference on*, pages 1–4. IEEE.

Baggi, C. A. d. S. and Lopes, D. A. (2011). Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. *Avaliação: Revista da Avaliação da Educação Superior*, 16(2).

Baker, R. et al. (2010). Data mining for education. *International encyclopedia of education*, 7(3):112–118.

Baker, R. S. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM— Journal of Educational Data Mining*, 1(1):3–17.

Burgos, C., Campanario, M. L., de la Peña, D., Lara, J. A., Lizcano, D., and Martínez, M. A. (2017). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*.

Chaturvedi, M. (2017). Data mining and it's application in edm domain. In *Intelligent Computing and Control Systems (ICICCS), 2017 International Conference on*, pages 829–834. IEEE.

Couto, D. and Santana, A. (2017). Educational data mining applied to the identification of variables associated with evasion and retention [mineração de dados educacionais aplicada à identificação de variáveis associadas à evasão e retenção]. volume 1877, pages 333–344.

da Cunha, J. A., Moura, E., and Analide, C. (2016). Data mining in academic databases to detect behaviors of students related to school dropout and disapproval. In *New Advances in Information Systems and Technologies*, pages 189–198. Springer.

de Evasão, C. E. d. E. (1996). Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas. *Avaliação, Campinas*, 1(2):55–65.

DeBerard, M. S., Spielmans, G., and Julka, D. (2004). Predictors of academic achievement and retention among college freshmen: A longitudinal study. *College student journal*, 38(1):66–80.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4):498–506.

Dutt, A., Ismail, M., and Herawan, T. (2017). A systematic review on education data mining. *IEEE Access*, (1):15991–16005.

INEP (2017). Instituto nacional de estudos e pesquisas educacionais anísio teixeira. mec e inep divulgam dados do censo da educação superior 2016. `http://download.inep.gov.br/educacao_superior/censo_superior/documentos/2016/censo_superior_tabelas.pdf`. Accessed: 2018-07-10.

Manhães, L. M. B., da Cruz, S. M. S., and Zimbrão, G. (2014a). The impact of high dropout rates in a large public brazilian university. In *Proceedings of the 6th International Conference on Computer Supported Education-Volume 3*, pages 124–129. SCITEPRESS-Science and Technology Publications, Lda.

Manhães, L. M. B., da Cruz, S. M. S., and Zimbrão, G. (2014b). Wave: an architecture for predicting dropout in undergraduate courses using edm. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, pages 243–247. ACM.

Nascimento, P. and Verhine, R. (2017). Considerações sobre o investimento público em educação superior no brasil. `http://repositorio.ipea.gov.br/bitstream/11058/7648/1/Radar_n49_considera%C3%A7%C3%B5es.pdf`. Accessed: 2018-07-10.

Rodriguez, A. (2014). Fatores de permanência e evasão de estudantes do ensino superior privado brasileiro um estudo de caso. *Caleidoscópio*, 1(3):31–43.

Romero, C. and Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618.

Sarra, A., Fontanella, L., and Di Zio, S. (2018). Identifying students at risk of academic failure within the educational data mining framework. *Social Indicators Research*, pages 1–20.

Silva Filho, R. L. L., Motejunas, P. R., Hipólito, O., and Lobo, M. B. C. M. (2007). A evasão no ensino superior brasileiro. *Cadernos de pesquisa*, 37(132):641–659.

Sultana, S., Khan, S., and Abbas, M. A. (2017). Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts. *International Journal of Electrical Engineering Education*, 54(2):105–118.

Yukselturk, E., Ozekes, S., and Türel, Y. K. (2014). Predicting dropout student: an application of data mining methods in an online education program. *European Journal of Open, Distance and E-learning*, 17(1):118–133.