Learning Analytics in Mathematics Teacher Education at the Ceará State University

João Batista Carvalho Nunes[1][0000-0002-1270-0026], Viviani Maria Barbosa Sales[2][0000-0002-7927-4711] and João Bosco Chaves[1][0000-0001-7994-0772]

[1] Ceará State University (UECE), Fortaleza, CE, 60714-903, Brazil
[2] Fortaleza City Council, Fortaleza, CE, Brazil
joao.nunes@uece.br

Abstract. Enrollment in higher education is growing in Brazil. Teacher education, particularly, faces several problems pointed out by studies in the country. Low completion rates of undergraduate teacher education programs requires urgent measures to change this situation with quality. Learning analytics emerges as an option to improve students' learning, reducing the risk of academic failure. This research, as a part of a larger project, sought to develop a statistical model that helps to predict students at risk of not being successful in the subjects of the Undergraduate Mathematics Teacher Education Program, in distance education modality, at the Open University of Brazil / Ceará State University (UAB/UECE). For that, a statistical method was used. It was decided to use data of the students who entered the university in 2009, in poles of two cities in the countryside of Ceará: Mauriti and Piquet Carneiro. The students' access and interaction data in the Moodle environment and final performance grades in the subjects were collected. In data analysis, binary logistic regression was used. The predictive model developed has the academic situation of the student (approved / failed) as a dependent variable and, as independent variables, the student's action categories in Moodle. This model has six independent variables, related to the activities: choice, forum, questionnaire and task, and to the file resource. It has a high overall efficiency, with statistically significant predictor variables and high accuracy percentages to predict failure and non-failure.

Keywords: Learning Analytics, Distance Education, Mathematics Teacher Education.

1 Introduction

1.1 Initial Teacher Education in Brazil

Enrollment in higher education is growing in Brazil. Taking only undergraduate programs into account, the object of interest of this research, from 2005 to 2014, enrollments in that stage increased by 71.4%, from 4,567,798 in 2005 to 7,828,013 in 2014 [1].

It is necessary to highlight the great evolution of undergraduate distance programs in relation to the on-site ones in that period. While enrollment in face-to-face undergraduate programs grew by 45.7%, based on 2005, enrollment in distance education programs increased by 1,070.5%, reaching 17.1% (1,341,842) of total enrollments in 2014. Regarding undergraduate teacher education programs, the basis for the initial teacher education, 36.9% (540,693) of the enrollments in 2014 occurred in the distance education modality [1].

Particularly, teacher education faces several problems pointed out by studies in Brazil [2-4]. On the one hand, the low completion rates of undergraduate teacher education programs, not exceeding 25% [5], requires urgent measures to increase, with quality, this percentage. On the other hand, 41.2% of the enrollments in 2014 in such programs occurred in public institutions, corresponding to 604,623 people [1]. This is a high investment of the public sector in teacher education for the various areas of knowledge, facing a national framework that, on the one hand, needs teachers with a specific degree in their areas of activity and, on the other, still cannot guarantee that the participants will complete those education programs.

Regarding the education of mathematics teachers, 85,402 students were enrolled in 2016, 62.2% (53,134) of them in face-to-face undergraduate teacher education programs and 37.8% (32,268) in distance learning programs. However, in that same year only 10,919 students completed undergraduate mathematics teacher education programs, maintaining the approximate percentage of 62.2% (6,789) from face-to-face programs and 37.8% (4,130) from distance programs [6].

1.2 Learning Analytics in Teacher Education

Research upon learning analytics has intensified since 2010, after the 1st International Conference on Learning Analytics and Knowledge (LAK) held in Banff (Canada) in 2011, followed by the constitution of the Society for Learning Analytics Research (SoLAR) in the same year [7-8]. Since then, learning analytics has been used to improve students' learning and the educational process.

In Latin America, the Latin American Workshop on Learning Analytics (LALA) was held in 2015, within the framework of the IV Congresso Brasileiro de Informática na Educação (CBIE) and the X Latin American Conference on Learning Technologies (LACLO). From the six papers presented and published in LALA 2015, three of them have Brazilian participation.[1]

Nunes [9] elaborated the state of the art in scientific journals on the use of the learning analytics in Latin America. He used bibliographic research in two abstract databases that covered journals of this region's countries: Scielo and Dialnet. Only two articles were found in Scielo and eight were found in Dialnet. From those ten, only two met the research criterion, indicating that there is much to be done in the field of learning analytics research in Latin America.

A recent work has sought to map the initiatives of learning analytics in Latin America [10]. For that, a systematic mapping of papers from 2011 to May 2016 was carried out, as well as an open questionnaire to obtain data on research groups in learning analytics. At the end, 30 texts and 28 research groups were found. According to [10],

[1] Available in http://www.br-ie.org/pub/index.php/wcbie/issue/view/139.

"although the number of articles can be considered small, it is possible to see a clear trend of increasing".

In Brazil, researches currently developed by the Laboratory of Analytics, Educational Technology and Free Software (LATES) of the Ceará State University (UECE) seeks to use the learning analytics in the education of basic education teachers. Two master's dissertations [11-12] and two Ph.D. theses [13-14] have already been defended for that purpose.

In his study in the UAB[2]/UECE Undergraduate Mathematics Teacher Education Program, with students who entered the university in 2009, Chaves [11], using Spearman's rank correlation coefficient, found out that among the set of five activity types offered in Moodle[3] (questionnaires, tasks, forums, chats and choices), "only two have a positive and strong influence on performance – questionnaires and tasks. Forums have a positive influence, however moderate to weak".

Aguiar [12], using the same groups of students of [11] and the binary logistic regression technique, found a statistically significant model that associates dropout (dependent variable) with the sum of all interactions (independent variable). According to the research, when there is an increase in a student's interaction in the program environment, the chance of avoiding decreases by 0.001351%, that is, the more interactions in the VLE, the greater the probability that the student will continue until the end of the program.

Sales's [13] research was developed in the UAB/UECE Undergraduate Pedagogy Program. Using the statistical technique of binary logistic regression, she made use of the actions carried out in the Moodle by students from Mauriti and Missão Velha who entered the university in 2009. She was able to prove the thesis that "[...] through analysis of student data available in the VLE, it is possible to assess student performance, predict and identify future problems, and propose models capable of identifying students at risk of academic failure".

Based on the learning analytics, Gonçalves [14] analyzed the education offered to students of the UAB/UECE Undergraduate Pedagogy Program in the field of school management and their performance in the program. Using a mixed approach and statistical and documentary analysis methods, she identified that the "[...] actions of uploading a file to a task, clicking on a discussion topic in a forum, viewing a resource, and viewing all tasks on the same screen directly interferes with students' performance".

2 Methodology

This research is part of a larger project, whose research problem was: how to improve the academic performance of UAB/UECE undergraduate teacher education students? This text is limited to developing a statistical model that helps to predict students at risk of not being successful in the subjects of the Undergraduate Mathematics Teacher Education Program, in distance education modality, at the UAB/UECE. For that, a

[2] Open University System of Brazil. More information available in http://www.capes.gov.br/uab/o-que-e-uab.

[3] The UAB/UECE programs use Moodle virtual learning environment (VLE).

statistical method was used. It was decided to use the data from the students who entered the university in 2009, in poles of two cities in the countryside of Ceará: Mauriti and Piquet Carneiro.

According to [15], the learning analytics process generally consists of three main steps: data collection and pre-processing, analytics and action, and post-processing. This research comprised only the data collection and pre-processing stage and the "analytics" sub-stage of the second stage. The "action" sub-step and the post-processing stage will be carried out in a research that will give continuity to the larger project.

For the data collection and pre-processing stage, an exact copy of the Moodle environment used in the UAB/UECE for the 2009-2012 period was generated. With this copy, operational variables were defined that would be used, based on an analysis of the environmental structure adopted by the institution. Next, the VLE reporting tool was used to extract the required access and interaction data. The final performance marks in the subjects were those officially registered in the academic system of the university. The student results (performance and frequency) were converted into two categories: "approved" (reference category – received value "0") and "failed" (event of interest – received value "1").

The data were grouped into a single spreadsheet in Excel format. Each line corresponds to a student's (coded) id in a given semester of the program subject of a polo of the analyzed program (example: MAT_MT_E01_S1D1). There are also 41 variables, non-sequentially coded from A04 to A99, corresponding to each of the action categories performed by the students. These action categories comprise a set of same-kind action logs executed by students in Moodle VLE. In addition, there are the variables program (1 - Mathematics), pole (6 - Mauriti, 8 - Piquet Carneiro), and result (0 - approved, and 1 - failed).

In the "analytics" sub-step, the data in Excel format were entered in the Stata software [16-17]. Since the obtention of a predictive model was expected, the initial application of multiple linear regression was attempted. Due to non-compliance with the assumptions that deal with the relations between the dependent and independent variables in the multiple linear regression, the statistical technique of binary logistic regression was used [18-20].

3 Results

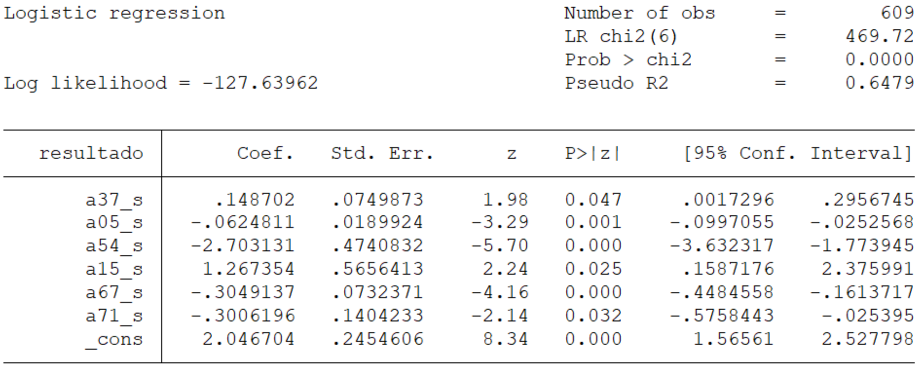
The analyzed program has 660 records, 331 (50.2%) referring to the polo in Mauriti City and 329 (49.8%) to the one in Piquet Carneiro City. There are 488 (73.9%) records of approval in the program subjects and 172 (26.1%) of failure.

Using the result as a dependent variable and, as independent variables or predictors, all 41 variables related to action categories (A04, A05, ..., A99), the Stata logit command was applied to obtain the complete binary logistic regression model estimated by maximum likelihood, at the 95% confidence level for the estimated parameter ranges.

For this calculation, the variables A10, A20, A40, A41, A64, A68 and A72 were eliminated, due to predicting the fault perfectly, as well as variables A42 and A77 as a result of multicollinearity. Consequently, 51 observations were deleted.

The maximum values of the log-likelihood function for the complete model and for the null model are, respectively, -115.60113 and -362.49747. The χ2 test with 32 degrees of freedom was 493.79, with p-value equal to 0.000 (p <0.001). Therefore, at the significance level of 0.05, it is possible to reject the null hypothesis that all parameters βj (j = 1, 2, ..., 41) are statistically equal to zero. At least one independent variable is statistically significant to explain the probability of a student failing.

Since the null hypothesis cannot be rejected for several parameters (for example, β1, β3, β4, etc.), at the significance level of 0.05, when the column P> | z | is observed, the binary logistic regression using the *stepwise* procedure in Stata was estimated. The result is shown in Fig. 1. For this model, the χ2 test with 6 degrees of freedom was 469.72, with p-value equal to 0.000 (p <0.001). Hence, at the significance level of 0.05, it is possible to reject the null hypothesis that all parameters βj (j = 1, 2, ..., 41) are statistically equal to zero.



```
Logistic regression                             Number of obs     =        609
                                                LR chi2(6)        =     469.72
                                                Prob > chi2       =     0.0000
Log likelihood = -127.63962                     Pseudo R2         =     0.6479

------------------------------------------------------------------------------
   resultado |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       a37_s |    .148702   .0749873     1.98   0.047     .0017296    .2956745
       a05_s |  -.0624811   .0189924    -3.29   0.001    -.0997055   -.0252568
       a54_s |  -2.703131   .4740832    -5.70   0.000    -3.632317   -1.773945
       a15_s |   1.267354   .5656413     2.24   0.025     .1587176    2.375991
       a67_s |  -.3049137   .0732371    -4.16   0.000    -.4484558   -.1613717
       a71_s |  -.3006196   .1404233    -2.14   0.032    -.5758443    -.025395
       _cons |   2.046704   .2454606     8.34   0.000      1.56561    2.527798
------------------------------------------------------------------------------
```

Fig. 1. Result of binary logistic regression in Stata by means of the *stepwise* procedure.

The likelihood-ratio test was applied to verify the fit quality of the complete model compared to the fit of the estimated model using the *stepwise* procedure. The results show that the estimation of the final model (Fig. 1), keeping only six variables (A05, A15, A37, A54, A67 and A71) and discarding the others, did not alter the quality of the adjustment, at the significance level of 0.05. As p = 0.5716 (p> 0.05), one cannot reject the null hypothesis that the fit quality of both models is statistically the same. Thus, the final model is preferable to the complete model.

In the final model, all the parameters βj, as well as the constant, have a z of Wald test with an associated p-value less than 0.05. The final expression of the estimated probability of a student's failure in the Undergraduate Mathematics Teacher Education Program at the UAB/UECE, class of 2009.1 (the group of students who entered the university in the first semester of 2009) is (1).

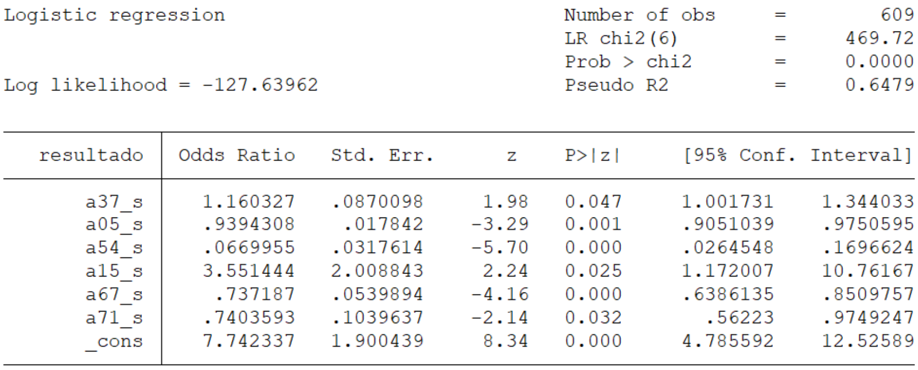
$$p_i = \frac{1}{1+e^{-(2{,}046704-0{,}0624811\times A05_i+1{,}267354\times A15_i+0{,}148702\times A37_i-2{,}703131\times A54_i-0{,}3049137\times A67_i-0{,}3006196\times A71_i)}} \quad (1)$$

The action categories that influence the probability of a student's failure in a subject of the Undergraduate Mathematics Teacher Education Program, class of 2009.1, are listed in Table 1.

Table 1. Action categories that influence the likelihood of a student's failure in a subject.

Identifier	Action Category Description	Functionality
A05	View the link of a task	Task
A15	Select a choice	Choice
A37	Message error sent to e-mail in forum	Forum
A54	Click the button: *Try to answer the questionnaire now*	Questionnaire
A67	View a resource (a PDF file, for example)	File
A71	This action is logged whenever the student uploads a file	File

To identify the chances of occurrence of the interest event (to fail) when the corresponding predictive variable was changed in one unit, while remaining the other constant conditions, the *logistic* command of Stata was applied. The result can be seen in Fig. 2.



```
Logistic regression                               Number of obs     =        609
                                                  LR chi2(6)        =     469.72
                                                  Prob > chi2       =     0.0000
Log likelihood = -127.63962                       Pseudo R2         =     0.6479

------------------------------------------------------------------------------
   resultado | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       a37_s |   1.160327   .0870098     1.98   0.047     1.001731    1.344033
       a05_s |   .9394308    .017842    -3.29   0.001     .9051039    .9750595
       a54_s |   .0669955   .0317614    -5.70   0.000     .0264548    .1696624
       a15_s |   3.551444   2.008843     2.24   0.025     1.172007    10.76167
       a67_s |    .737187   .0539894    -4.16   0.000     .6386135    .8509757
       a71_s |   .7403593   .1039637    -2.14   0.032       .56223    .9749247
       _cons |   7.742337   1.900439     8.34   0.000     4.785592    12.52589
------------------------------------------------------------------------------
```

Fig. 2. Result of binary logistic regression in the Stata by *stepwise* procedure and *logistic* command.

Fig. 2, instead of bringing the coefficients indicated in Fig. 1, presents the odds ratios. Based on the values obtained, it is possible to indicate that the probability of a student's failure falls, on average, while remaining the other constant conditions: 6.06% for each addition of a link visualization of a task (A05); 93.30% for each new click on the button: *Try to answer the questionnaire now* (A54); 26.28% when you increase by one unit the view of a resource (a PDF file, for example) (A67); and 25.96% when you upload a new file (A71). This result indicates that the accomplishment of the questionnaire activity by the students is the one that contributes most to decrease the chance of a student's failure and, consequently, to approve, remaining the other constant conditions.

On the other hand, it can be stated that the chance of a student's failure is, on average, maintained the other conditions: 255.14% higher when selecting a choice (A15); and 16.03% higher when a new e-mail message error is made in the forum (A37).

Sensitivity analysis with a cutoff of 0.5 for the estimated final model shows that the overall efficiency of the model, that is, the "total percentage of correctness in the classification" [16], is 92.28 %. The sensitivity (percentage of correctness considering only the observations that were about failure) is 88.95%; while the specificity (percentage of correctness considering only the observations that were about approval) is 93.59%.

The curve of sensitivity and specificity versus probability cutoff for the final model (Fig. 3) shows that the approximate value of 0.36 for cutoff equals sensitivity to specificity. It reflects that, for cutoff values above 0.05, the specificity curve shows high hit percentages (75% or higher). In addition, only for cutoff values above 0.80, the sensitivity curve shows percentages of accuracy inferior to 75%. So, it can be concluded that the model is good for predicting either failed (event) and approved (non-event) students.

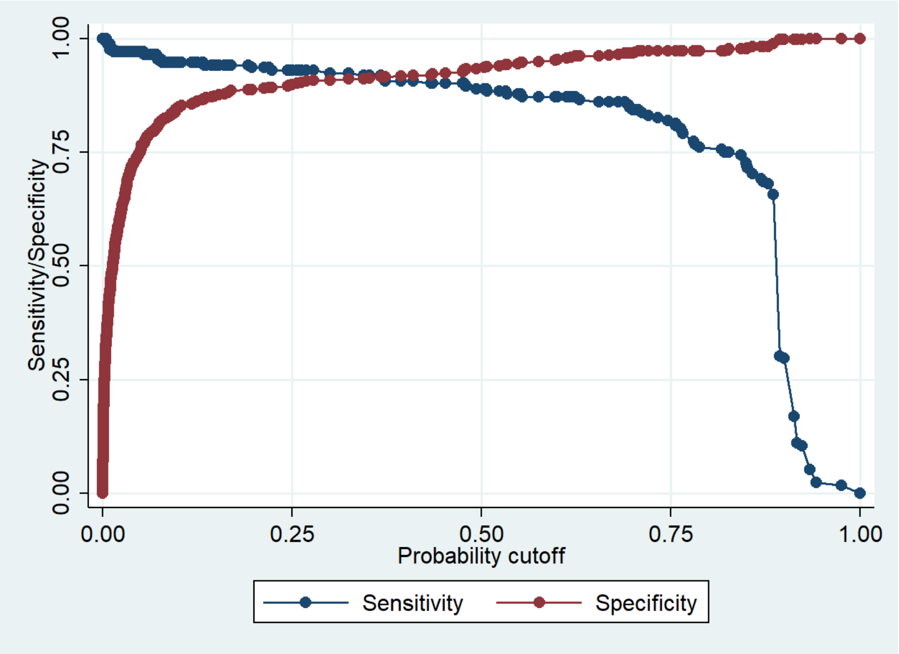


Fig. 3. Sensitivity curve for the final model.

Finally, the final model estimated is shown with high overall efficiency, with statistically significant predictive variables and high percentage of correctness for the event and non-event, ratified by an area under the ROC (AUC) curve of 0.9640 (Fig. 4), denoting an excellent value for forecasting purposes.

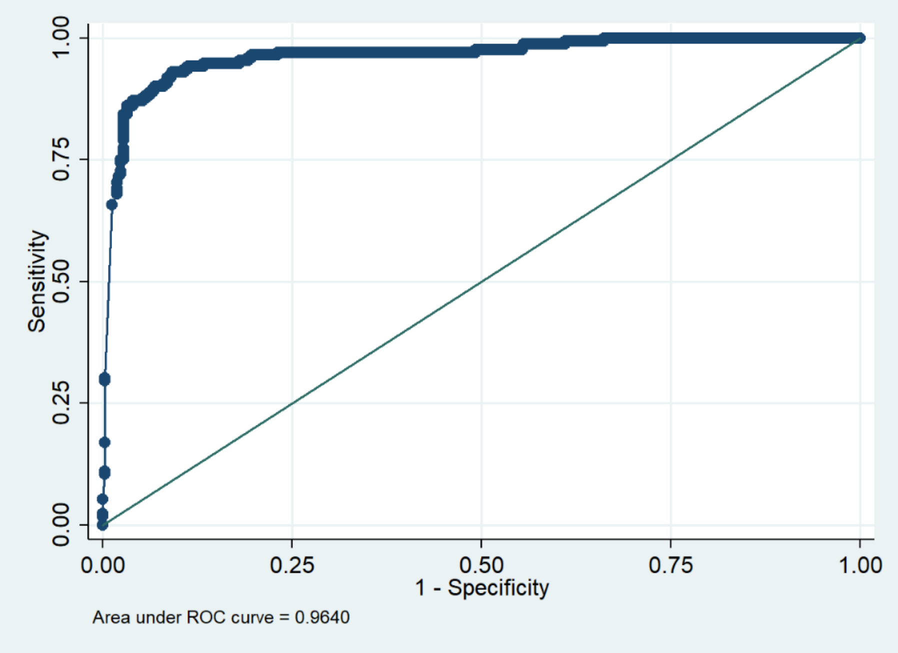


Fig. 4. ROC curve for the final model.

4 Conclusion

Learning analytics is a new, but very promising field of research. This research is one of the pioneers on this topic in Brazil, carried out by researchers in the area of education. It sought to develop a statistical model that helps to predict students at risk of not being successful in the subjects of the Undergraduate Mathematics Teacher Education Program, taught at distance by the Ceará State University, in partnership with the Open University System of Brazil. It was centered in two classes that started in 2009, in the poles located in the cities Mauriti and Piquet Carneiro.

The predictive model developed has as a dependent variable the academic situation of the student (approved / failed) and, as independent variables, the students' action categories in the Moodle: A05 (view the link of a task), A15 (select a choice), A37 (message error sent to e-mail in forum), A54 (click the button: *Try to answer the questionnaire now*), A67 (view a resource – a PDF file, for example), A71 (this action is logged whenever the student uploads a file). Accordingly, the functionalities that influence the probability of a student being failed are the activities: choice, forum, questionnaire and task, besides the file resource. The model shows high overall efficiency, with statistically significant predictive variables and high percentage of correctness to predict failure and non-failure.

Based on the results obtained, it is possible, in addition to other actions, to delineate pedagogical actions in order to strengthen the achievement of action categories that impact on the increase of the chance of approval, and to weaken the action categories that generate growth in the chance of disapproval, as well as to stimulate the use of Moodle activities and resources which are not yet present in the program structure under analysis. In addition, the results will serve as a basis for creating tools for the real-time application of learning analytics, favoring the observation of the predictive model performance at the beginning, middle and end of the program, and its subsequent improvement.

References

1. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP): Resumo técnico: censo da educação superior 2014. Inep, Brasília (2016).
2. Gatti, B. A., Barretto, E. S. S.: Professores do Brasil: impasses e desafios. Unesco, Brasília (2009).
3. Gatti, B. A., Barretto, E. S. S., André, M. E. D. A.: Políticas docentes no Brasil: um estado da arte. Unesco, Brasília (2011).
4. Saviani, D.: Formação de professores no Brasil: dilemas e perspectivas. Poíesis Pedagógica. 9(1), 7-19 (2011). doi: 10.5216/rpp.v9i1.15667
5. Gatti, B. A.: Formação de professores no Brasil: características e problemas. Educação e Sociedade. 31(113), 1355-1379 (2010). doi: 10.1590/S0101-73302010000400016
6. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP): Sinopse Estatística da Educação Superior 2016. Inep, Brasília (2017).
7. Cooper, A.: A brief history of analytics. CETIS Analytics Series. 1(9), 1-21 (2012).
8. Ferguson, R.: Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning. 4(5-6), 304-317 (2012). doi: 10.1504/IJTEL.2012.051816
9. Nunes, J. B. C.: Estado da arte sobre analítica da aprendizagem na América Latina. In: Anais dos Workshops do IV Congresso Brasileiro de Informática na Educação 2015. pp. 1024-1033. SBC, Maceió (2015). doi: 10.5753/cbie.wcbie.2015.1024
10. Santos, H. L., Cechinel, C., Nunes, J. B. C., Ochoa, X.: An initial review of learning analytics in Latin America. In: 2017 Twelfth Latin American Conference on Learning Technologies (LACLO). pp. 1-9. IEEE, La Plata (2017). doi: 10.1109/LACLO.2017.8120913
11. Chaves, J. B.: Formação a distância de professores em Matemática pela UAB/UECE: relação entre interação e desempenho à luz da analítica da aprendizagem. Dissertation, Universidade Estadual do Ceará (2015).
12. Aguiar, A. N.: Evasão no curso de Licenciatura em Matemática a distância da UECE sob a perspectiva da analítica da aprendizagem. Dissertation, Universidade Estadual do Ceará (2016).
13. Sales, V. M. B.: Analítica da aprendizagem como estratégia de previsão de desempenho de estudantes de curso de Licenciatura em Pedagogia a distância. Ph.D. thesis, Universidade Estadual do Ceará (2017).
14. Gonçalves, M. T. L.: Formação do pedagogo para a gestão escolar na UAB/UECE: a analítica da aprendizagem na educação a distância. Ph.D. thesis, Universidade Estadual do Ceará (2018).
15. Chatti, M.A., Dyckhoff, A.L., Schroeder, U., Thüs, H.: A reference model for learning analytics. International Journal of Technology Enhanced Learning. 4(5-6), 318-331 (2012). doi: 10.1504/IJTEL.2012.051815
16. Fávero, L. P.: Análise de dados. Elsevier, Rio de Janeiro (2015).

17. Fávero, L. P., Belfiori, P., Takamatsu, R. T., Suzart, J.: Métodos quantitativos com Stata: procedimentos, rotinas e análises de resultados. Elsevier, Rio de Janeiro (2014).
18. Agresti, A., Finlay, B.: Métodos estatísticos para as Ciências Sociais. Penso, Porto Alegre (2012).
19. Field, A.: Descobrindo a Estatística usando o SPSS. Artmed, Porto Alegre (2009).
20. Hair JR, J. F, Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L.: Análise multivariada de dados. Bookman, Porto Alegre (2009).