An Analysis of Self-Regulated Learning Behavioral Diversity in Different Scenarios in Distance Learning Courses

Aldo A. Cavalcanti[1], Raphael A. Dourado[2], Rodrigo L. Rodrigues[3], Nathan Alves de L. Silva[1], João Sedraz[4], Jorge L. C. Ramos[4]

[1]Computer Department - Federal Rural University of Pernambuco - Brazil

[2]Computer Center - Federal University of Pernambuco - Brazil

[3]Department of Education - Federal Rural University of Pernambuco - Brazil

[4]Federal University of Vale do São Francisco - Brazil

aldoacf@gmail.com, raphaaugusto@gmail.com, rodrigo.linsrodrigues@ufrpe.br, nathan.alves.r2n@gmail.com, jsedraz@gmail.com, jorgeluiscavalcanti@gmail.com

Abstract. *The increasing volume of student behavioral data within virtual learning environments (VLE) provides many opportunities for knowledge discovery. Thus, techniques that make it possible to predict the academic performance of students become essential tools to assist distance learning instructors. This article shows the results of the development of a student performance predictive model, based on behavioral indicators of self-regulated learning in a database extracted from the Moodle VLE. In addition, we attempted to develop specialized predictive models for three distinct scenarios (general, divided by course and divided by semester). The results showed that the variation in the student behavior through the different semesters has a strong influence on the model's predictive power.*

1. Introduction

The availability of Distance Education courses in Brazil has increased in recent years [Censo 2016] and became a viable form of knowledge acquisition, especially for underprivileged people. As a result of this growth, the use of Virtual Learning Environments (VLEs) is gaining popularity, since it is the main medium for interaction between students and instructors in distance education. These interactions generate an increasing amount of data, which can provide instructors with information related to their students' degree of self-regulatory behavior — an important skill in this context [Barnard et al. 2009; Romero and Ventura, 2013].

Methods from an emerging research area known as Educational Data Mining (EDM) can help us understand these self-regulatory behaviors through the student's interactions on VLEs. EDM's main objective as a discipline is to develop methods to explore data sets collected in VLEs in order to understand and improve learning [Romero e Ventura, 2013]. Currently, this area has established itself as a strong and consolidated line of research with great potential for discovering new knowledge and helping to improve the quality of teaching [Baker, Isotani, and Carvalho, 2011].

DOI: 10.5753/cbie.sbie.2018.1493

Through the use of EDM techniques, it is possible to find indicators of self-regulatory behavior among the students' interactions and understand how they manage their cognitive process. This is important because, according to Pintrich and DeGroot (1990), there is a strong correlation between an individual's self-regulatory skills and his performance, making it convenient to use EDM for predicting academic success.

Following this line of inquiry, this work proposes a predictive model of student performance based on his/her indicators of self-regulated learning (SRL) in a VLE. For this purpose, the dataset used in the study was stratified into three distinct scenarios: (1) General, containing all courses; (2) per Period (which in this case is given per semester), containing 8 subsets, one for each course semester; and (3) per Course, with 4 subsets referring to the 4 courses (Administration, Biology, Literature, and Pedagogy). By analyzing these subsets, we hoped to find which scenarios the prediction would be more accurate, thus answering the following research questions:

- *Are there differences between the development of performance classifiers, in relation to the course to which the student belongs and the period in which he/she is enrolled?*
- *What are the variables that most influence students' performance prediction?*

To answer these research questions, we organized this paper's structure as follows: Section 2 deals with self-regulation, its impact on student performance, and also mentions some related work; Section 3 describes the development of classifiers using the Logistic Regression Model, which is the prediction technique adopted in the present work; Section 4 presents the purpose of this research, the details of the used dataset through some descriptive analysis of the variables, the development process of the proposed predictive models, and the discussion of the results; Section 5 presents the educational implications of the experiment results; and finally, Section 6 reports the conclusions and perspectives for future work.

2. Self-Regulated Learning and its Relationship with Academic Performance

Self-regulated learning describes the individual's ability to manage his cognitive processes, regulating his actions to achieve better learning. Students who manage to self-regulate their learning are more active, effective, efficient, and demonstrate high levels of motivation [Gaitero, Román, and Real García, 2016]. In the context of distance learning courses, EDM techniques can help to uncover SRL patterns, providing the necessary variables for the identification of self-regulation standards. For example, among these variables, a VLE can provide indicators of student's time management strategies, as seen in [Cho and Shen, 2013], which can be used to seek for correlations between the amount of time the student spends in the platform and his academic performance.

Thus, students' academic performance in distance learning courses can be directly related to their behavior within the VLE [Cicchinelli et al., 2018]. Several authors believe that these interactions between students, instructors, and the virtual environment constitute valuable information to be explore by EDM [Peña-Ayala, 2014; Costa, 2013; Ko and Leu, 2016].

Many studies expose the importance of EDM in helping to understand the learning processes in distance learning courses and its potential to become a powerful ally of education. These findings point to the relationship between VLE access [Murray et al., 2013], clicks within the environment [Dickson et al., 2005], and the student's performance. Also regarding clicks within the VLE, Cicchinelli [Cicchinelli et al., 2018] was able to find patterns that represent students' self-regulating skills.

The prediction of student evasion in face-to-face courses is also possible, as shown by Manhães et al. (2011) and Cechinel et al. (2015). Finally, [Rodrigues, 2017] shows that it is possible to understand through EDM techniques how students self-regulate their learning and how this impacts their academic performance. These works are closely related to the theme proposed in this paper and provided the background for the development of the present study.

3. Development of the Logistic Regression Classifiers

To develop a predictive model, it is necessary to apply algorithms to the chosen dataset. According to the studies developed by Rodrigues (2017), Logistic Regression is the best algorithm for SRL studies. The Logistic Regression method is a statistical technique capable of estimating the probability of occurrence of a certain event, described in a binary and categorical manner, from the exploration of previously known variables — also called *independent variables* — that can be categorical or non-categorical [Peng, 2002].

In regression analysis, the response variable Y_j, also known as *dependent variable*, can take two values, $Y_j = 0$ and $Y_j = 1$, which in the context of this study means "passed" and "failed", respectively. The event of interest in this paper is to predict when a student will fail, according to his/her self-regulatory behavior within the VLE.

Logistic Regression can provide such predictions of binary events by generating a model with all the useful predictor variables, capable of inferring the probability of occurrence for each case. If the probability calculated by the algorithm is greater than 0.5, then the expected event result is 1, meaning that the student will fail. Conversely, if it is less or equal to 0.5, the expected event result is 0, which means that the student will pass.

A logistic regression model is represented by Equation (1), which depicts the *logit* transformation of p_j, where p_j points to the probability of occurring the event of interest, $x_1 \dots x_n$ is the vector of independent variables and β indicates the model coefficients, or how the model can explain the observed values.

$$logit(p_j) = ln\left(\frac{p_j}{1-p_j}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \tag{1}$$

4. The Experiment

The experiments described in this paper used a dataset containing students' interactions with educational artifacts of four undergraduate courses in an online learning platform used by University of Pernambuco. The main goal was to find an ideal data model for predicting student failure through Logistic Regression classifiers.

For this experiment, we stratified the original dataset by course and semester, arriving at three different scenarios, namely: (1) General Dataset, representing the original data, (2) t per Period, where 8 subsets were generated for each semester of the courses; and (3) per Course, divided in 4 subsets referring to the 4 existing courses in the institution (Literature, Biology, Administration and Pedagogy). The purpose of these different scenarios was to find out in which of them the predictions are more accurate, testing whether there is difference in performance for each generated model and finding the variables that have the most influence on their performance.

Finally, to evaluate these differences, the performance score (accuracies) of each prediction model was compared to each other by Analysis of Variance (*ANOVA*), as well as a visual comparison through box-and-whisker diagrams (or boxplot) generated for each scenario.

4.1. Descriptive Analysis

To build the predictive models used in this study, we used a dataset that represents behaviors of students within the Moodle learning platform. This dataset comprises 7 years of interactions between students and the VLE.

The dataset contains 30,217 rows, representing students distributed among the courses of Administration, Biology, Literature, and Pedagogy, along 34 variables that describe indicators of their self-regulatory behaviors, grouped in 6 SRL constructs: 1) Environment structuring; 2) Search for help; 3) Strategies for completing tasks; 4) Time management; 5) Goal setting; and 6) Self-evaluation. This allocation of variables based on self-regulation constructs followed the definitions established in [Rodrigues, 2017]. All these variables are described in detail in Table 1. In addition to these, we created the variable BINARY_PERFORMANCE, which is the target variable to be predicted in this study and represents the final situation of the student in the course.

The original variable that represents the student performance, used as the target variable, was composed by weighing the grades of each task as follows: *exam = 5.5*, *interaction in forums = 2.0* and *homework = 2.5*. We have taken into consideration replacement exams in case the student has missed one of the evaluations. Equation (2) summarizes the formula used to calculate the performance of each student:

$$Performance = (MEAN_{EXAMS} * 5{,}5) + (MEAN_{FORUM} * 2{,}0) + (MEAN_{WEBQUEST} * 2{,}5) \quad (2)$$

After composing the performance variable as a continuous value, we transformed it into a binary value using the following rule: values lower than 0.5 were converted to 0 and values higher or equal to 0.5 were converted to 1.

Table 1. Description of variables

Variable	Description of the variables	Construct
VAR01	Number of different locations (IPs) from which the student accessed the environment	**Environment structuring**
VAR02	Number of messages sent by the student to the instructor within the environment	**Search for Help**
VAR03	Number of messages sent per student to the instructor within the environment	
VAR04	The general amount of messages sent by the student within the environment	
VAR05	The general quantity of messages received by the student within the environment	
VAR06	Number of topics created by the student in the "strip-doubt" forum	
VAR07	Number of posts in the "strip-doubt" forum	
VAR08	Number of posts in forums that have been answered by other students	
VAR09	Number of posts in forums that have been answered by the instructor	

VAR10	Number of different colleagues to whom the student sent messages within the environment	
VAR12	Number of views in the "Content" tab (syllabus)	**Strategies for completing tasks**
VAR13	The time that the student has conducted more activities	
VAR14	Day-shift in which the student performed more activities	
VAR16	Number of out-of-term activities delivered by the student, by discipline	
VAR17	The average time between the opening of the activity and its submission	
VAR18	Number of readings made to the forum (pageviews)	
VAR20	Number of responses to the main topic (remake opinion in the forum)	**Self-evaluation**
VAR21	Number of pageviews to the chart of notes	
VAR22	Number of times the student viewed the "Activities Checklist"	
VAR23	Number of notes views per activity	
VAR24	The weekly average of the number of students accesses to the environment	**Time Management**
VAR25	The average time between the creation of a topic and the first post of the student	
VAR28	Number of Time Out	
VAR31	Number of student's access to the environment	
VAR31b	Number of different days that the student has accessed the course	
VAR31c	Number of different days that the student accessed the platform	
VAR32a	Number of student's access to the environment per shift (morning)	
VAR32b	Number of student's access to the environment per shift (afternoon)	
VAR32c	Number of student's access to the environment per shift (night)	
VAR32d	Number of student's access to the environment per shift (dawn)	
VAR33	Number of activities delivered by the student on time, by discipline	**Goal Setting**
VAR34	The overall amount of student posts in forums	
VAR35	Number of instructor responses to student questions on forums	

After the stratification, the three proposed scenarios showed the following characteristics: in the **General Dataset** scenario, which includes the entire dataset, we had a total of 30,217 rows containing information regarding the interactions of students from different courses in different semesters. In the **"per Course" scenario**, we had the following four subsets: "Administration", with 2,892 rows, "Biology", with 6,526 rows, "Literature" with 6297 rows, and "Pedagogy", with 14,502 rows. Each of these datasets contains information regarding students' interactions for each course in its various semesters, with the exception of the courses "Administration" — which lacked information for the 6th and 7th semesters — and "Pedagogy" — which lacked information for 6th semester. Finally, in the **"per Period" scenario**, we had a total of eight subsets, one for each period of the course (all courses are divided into eight periods) resulting in the following distribution: 5,815 rows for the 1st period, 1,700 for the 2nd, 6,321 for the 3rd, 8,865 for the 4th, 2,365 lines for the 5th, 441 for the 6th, 1,309 for the 7th, and 3,401 for the 8th.

4.2. Model Development

In this phase, we aimed not only to develop accurate models for student performance prediction in distance education courses, but also to understand under what scenario these predictions are more accurate. Thus, after the data understanding and pre-processing steps described in Section 4.1, we used the Logistic Regression method in each of the scenarios in order to predict student failure. The choice of the Logistic Regression algorithm was based on the results presented in [Rodrigues, 2017], which achieved satisfactory results in a similar context using this algorithm, compared to three other ones: *SVM*, *Random Forest,* and *Decision Tree*.

To measure the developed models' precision, we chose the *accuracy* metric, which attained, in our experiments, values that ranged from 83.9% to 93.3% for the "per

Period" scenario, 84.5% to 89 % for the "by Course" Scenario, and 85.9% to 86.4% for the "General" Scenario, as shown in Figure 1.

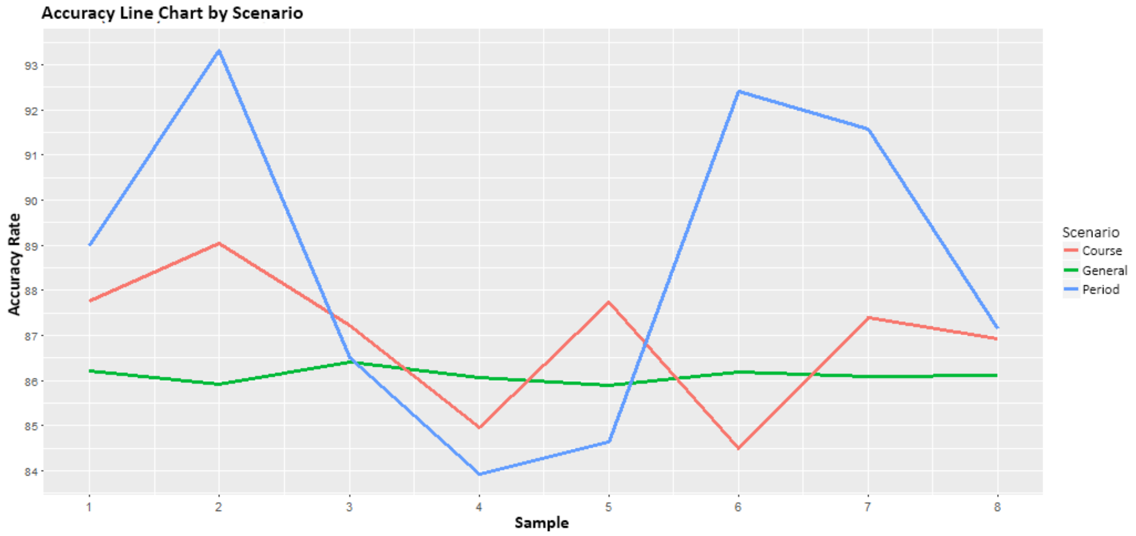

Figure 1. Line chart representing the accuracy variation for each scenario

Table 2 shows three vectors depicting the accuracy levels for each scenario and divided by period. Since the "per Period" scenario has the largest number of subsets, we had to adapt the accuracy vectors so that all the visualizations were generated uniformly, facilitating the analysis of the results. Then, in this scenario, only 1 accuracy value was generated for each period, totaling 8. For the "by Course" scenario, 2 accuracy values were generated for each of the 4 courses in order to total 8 samples and fit to the visualizations. As for the General Scenario, 8 values of accuracy were generated for the base.

In the "by Course" scenario we reorganized the vector so that it represents 1 course to every 2 indexes, in this case the indexes 1 and 2 show values referring to the Administration course, the indexes 3 and 4 concerns Biology course, the indexes 5 and 6 refer to Literature, and the indexes 7 and 8 present the values of accuracy for the Pedagogy. In the case of the General and Period scenarios, the values appear in the order in which they were obtained.

Table 2. Vectors with accuracy values for the scenarios "General", "by Course" and "by Period".

General Scenario							
1st	2nd	3rd	4th	5th	6th	7th	8th
86.2	85.9	86.4	86.1	85.9	86.2	86.1	86.1
Scenario by Course							
1st	2nd	3rd	4th	5th	6th	7th	8th
87.8	89	87.2	85	87.8	84.5	87.4	86.9
Scenario by Period							
1st	2nd	3rd	4th	5th	6th	7th	8th
89	93.3	86.5	83.9	84.6	92.4	91.6	87.2

Table 2 showed the models' accuracy values for the three distinct scenarios. For the "General" scenario, we repeated the experiment eight times and for the "by Couse" scenario two times. There were no repetitions for the "by Period" scenario. We can see that in most cases, in accordance with Figure 1 and Table 2, the "by Period" scenario showed the best accuracy values, which can be seen as an indication that it is the most suitable setting for generating classifiers based on self-regulation behavioral data. Beyond this inference based on descriptive statistics, the next subsection shows a deeper analysis of the results through boxplot charts and variance analysis.

4.3. Evaluation of the Models for the Different Scenarios

According to the boxplot charts shown in Figure 2, the "by Period" scenario is the ideal one for generating classifiers, since it has the highest median value.

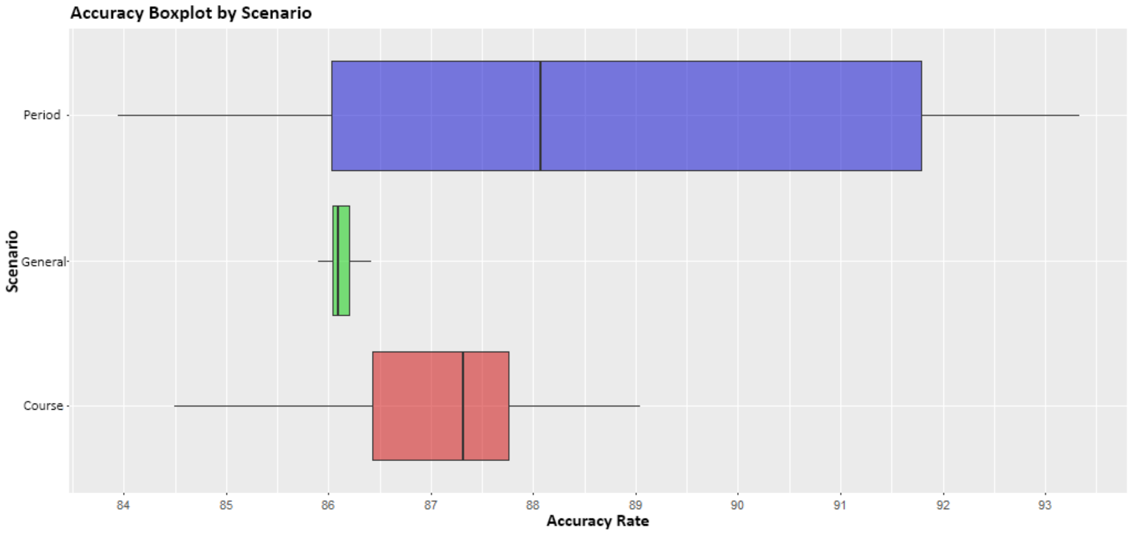

Figure 2. Boxplot charts depicting the distribution of accuracy values for each scenario

Although the boxplot charts graphically show the **"by Period"** scenario as the best one, even with its high variability, a variance test (ANOVA) is necessary to compare the accuracy values between the three scenarios and assert that there is statistically significant difference between them. As shown in Table 3, the ANOVA test did not found significant differences that explain choosing one scenario over another.

Table 3. Multiple comparisons of Tukey averages with confidence of 95%

	Difference	Lower	Upper	Adjusted p-value
General - Course	-1.00750	-3.7658158	1.750816	0.6335700
Period - Course	1.37375	-1.3845658	4.132066	0.4351312
Period - General	2.38125	-0.3770658	5.139566	0.0988566

Table 3 describes the Tukey Test, used in conjunction with ANOVA, which aims to find significantly different means between, in this case, the pairs of scenarios. The obtained results pointed out that none of the comparisons is statistically significant since, as shown in the "Adjusted p-value" column, all values are higher than the adopted significance level of 0.05.

Given the obtained results, we focused the study in the "per Period" scenario and sought to understand which variables most influenced the prediction of students' performance for this scenario according to each of the eight semesters. This analysis is shown in Figure 3, where the five most significant variables of each model for each period is shown.

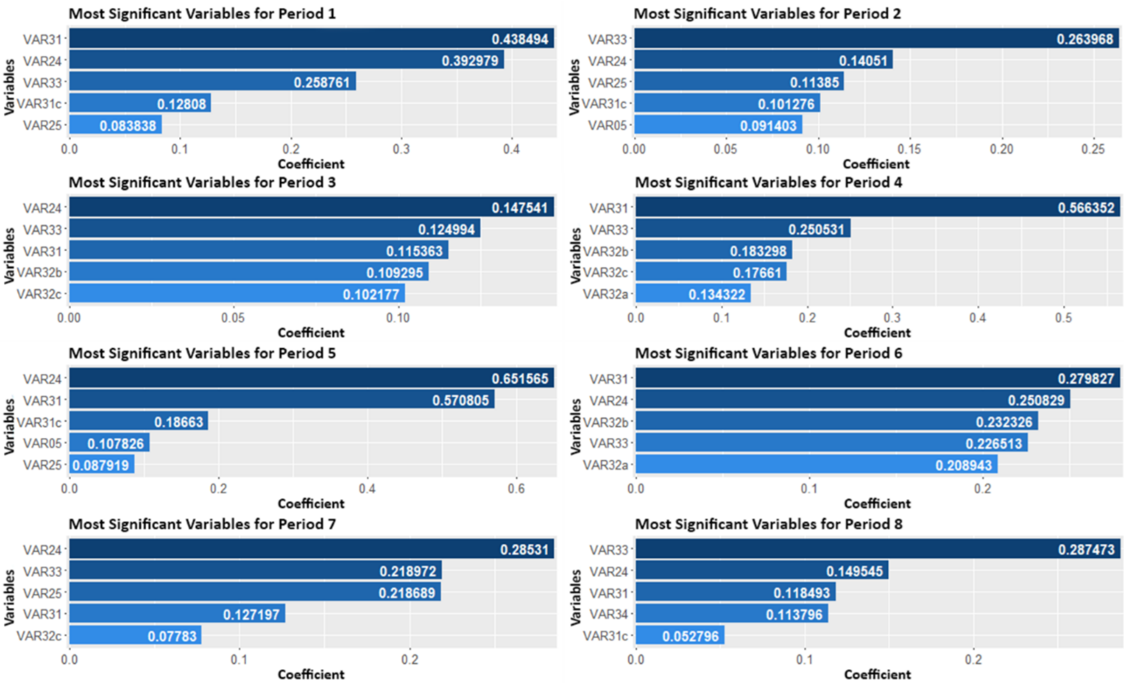

Figure 3. Bar charts showing the five most significant variables on each model

In Figure 3, we can observe that variables VAR24, VAR31 and VAR33 — which represent, respectively, the "weekly average of the number of student accesses to the environment", the "number of student accesses to the environment" and the "amount of activities delivered by the student per discipline" — are the most significant ones.

No single variable appears as one of the five most significant in all eight semesters, but VAR31 only does not appear in semester 2, VAR24 only does not appear in semester 4 and VAR33 only does not appear in semester 5. But VAR33 appears in all the semesters of the first half of the courses and the variables VAR24 and VAR31 appear in all semesters of the second half of the courses.

As a result, we can conclude, on a preliminary basis, that the “Time Management” construct is the most influential one in predicting the students' performance followed by the “Goal Setting” construct as the second most influential.

4.4. Educational Implications

The first step to increase student academic success is identifying the students with a high risk of failure. Nowadays, almost a third of distance education students in Brazil drop out after the first year [Censo, 2016]. These students’ low academic performance is seen as one of the most influential causes of this dropout rate [Essa and Ayad, 2012]. We hope that the results of this work will contribute to minimize retention rates by identifying in advance the factors that affect performance, allowing managers and instructors to make strategic educational decisions regarding students at risk and provide them the means to improve their performance.

5. Conclusion

Baker et al (2011) point out the importance of using automatic models to identify students with high risk of failure. The increasing volume of educational data makes this a timely situation for the development of these models and gain even more knowledge about how students are managing their own learning process, helping to understand the patterns behind their performances.

In this paper, we initially sought to apply the Logistic Regression technique in the dataset to obtain predictions of student failure. To obtain even more precise results, three distinct scenarios were generated that separated the dataset in General, by Course and by Period. Through variance and graphical analysis of performance rates (accuracy) of each of the three scenarios, it was determined that the scenario "per Period" yields the best results.

In this way, it was possible to find the variables that best describe the behaviors of the students within the VLE and how much they influence their performances. It was also found that the student's performance is directly related to the number of interactions within the environment, being the ability to manage the time in which he engages in academic activities and the ability to establish his goals the most critical activities for his learning.

Thus, it becomes clear the need for further research on these two constructs in order to develop new solutions that best exploit these behaviors.

Acknowledgements

The authors thank PRPPG/UFRPE and CNPq for the financial support through the Scientific Initiation grants. They also thank NEAD/UPE for the databases used in this study.

References

Baker, R., Isotani, S., & Carvalho, A. (2011). Mineraçao de dados educacionais: Oportunidades para o Brasil. Brazilian Journal of Computers in Education, 19(02), 03.

Barnard, L., Lan, W. Y., To, Y. M., Paton, V. O., & Lai, S. L. (2009). Measuring self-regulation in online and blended learning environments. The internet and higher education, 12(1), 1-6.

Cechinel, C., Araujo, R. M., & Detoni, D. (2015). Modelling and Prediction of Distance Learning Students Failure by using the Count of Interactions. Brazilian Journal of Computers in Education, 23(03), 1.

Censo, E. A. D. (2016). Relatório Analítico da Aprendizagem a Distância no Brasil 2016 - Censo EAD.

Cho, M. H., & Shen, D. (2013). Self-regulation in online learning. Distance education, 34(3), 290-301.

Cicchinelli, A., Veas, E., Pardo, A., Pammer-Schindler, V., Fessl, A., Barreiros, C., & Lindstädt, S. (2018). Finding traces of self-regulated learning in activity streams.

Costa, E., Baker, R. S., Amorim, L., Magalhães, J., & Marinho, T. (2013). Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. Jornada de Atualização em Informática na Educação, 1(1), 1-29.

Dickson, W. P. (2005). Toward a deeper understanding of student performance in virtual high school courses: Using quantitative analyses and data visualization to inform decision making. A synthesis of new research in K–12 online learning, 21-23.

Essa, A., & Ayad, H. (2012, April). Student success system: risk analytics and data visualization using ensembles of predictive models. In Proceedings of the 2nd international conference on learning analytics and knowledge (pp. 158-161). ACM.

Gaitero, Ó. G., Román, Ó. C., & García, J. J. R. (2016). A New Theoretical Construct in the Concept of Self-Regulated Learning. Educational Excellence, 2(1), 49-61.

Ko, C. Y., & Leu, F. Y. (2016, March). Applying Data Mining to Explore Students' Self-Regulation in Learning Contexts. In Advanced Information Networking and Applications (AINA), 2016 IEEE 30th International Conference on (pp. 74-78). IEEE.

Manhães, L. M. B., Da Cruz, S. M. S., Costa, R. J. M., Zavaleta, J., & Zimbrão, G. (2011). Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. In Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE) (Vol. 1, No. 1).

Murray, M., Pérez, J., Geist, D., & Hedrick, A. (2013, July). Student interaction with content in online and hybrid courses: Leading horses to the proverbial water. In Proceedings of the Informing Science and Information Technology Education Conference (pp. 99-115). Informing Science Institute.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. Expert systems with applications, 41(4), 1432-1462.

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. The journal of educational research, 96(1), 3-14.

Pintrich, P. R., & DeGroot, E. (1990). Quantitative and qualitative perspectives on student motivational beliefs and self-regulated learning. In Annual Meeting of the American Educational Research Association, Boston, MA (Vol. 128).

Rodrigues, R., Gomes, A. S., & Adeodato, P. (2017). Uma abordagem de Mineração de Dados Educacionais para previsão de desempenho a partir de padrões comportamentais de Autorregulação da Aprendizagem. In Anais dos Workshops do Congresso Brasileiro de Informática na Educação (Vol. 6, No. 1, p. 13).

Romero, C., & Ventura, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.