



Academic Behavior Analysis in Virtual Courses Using a Data Mining Approach

Dario Delgado-Quintero¹, Olmer Garcia-Bedoya²(✉) , Diego Aranda-Lozano¹,
Pablo Munevar-Garcia¹ , and Cesar O. Diaz³ 

¹ Universidad Nacional Abierta y a Distancia, Bogota, Colombia

² Universidad de Bogota Jorge Tadeo Lozano, Bogota, Colombia

olmer.garciab@utadeo.edu.co

³ OCOX AI, Bogota, Colombia

cdiaz@ocox-ai.co

Abstract. Virtual education is one of the educational trends of the 21st century; however knowing the perception of students is a new challenge. This article presents a proposal to define the essential components for the construction of a model for the analysis of the records given by the students enrolled in courses in a virtual learning platform (VLE). The article after a review of the use of data analytics in VLE presents a strategy to characterize the data generated by the student according to the frequency and the slice of the day and week that access the material. With these metrics, clustering analysis is performed and visualized through a map of self-organized Neural Networks. The results presented correspond to five courses of a postgraduate career, where was found that students have greater participation in the forums in the daytime than in the nighttime. Also, they participate more during the week than weekends. These results open the possibility to identify possible early behaviors, which let to implement tools to prevent future desertions or possible low academic performance.

Keywords: Learning management systems · Educational data mining · SOM Networks · Virtual education

1 Introduction

The impact of data mining in the industry is increasingly evident, the proper management of organizational data and how to learn to identify the beneficial information for organizations and turn it into beneficial information for commercial purposes is one of the pillars of its success [6]. Like the other sectors of the industry where data mining has been gaining strength, the educational sector has not been the exception, both in terms of data mining for educational environments, as a research topic or as an innovative factor in the research segment. And also, as an investment in administrative terms, because big data and

Supported partially by Colciencias.

© Springer Nature Switzerland AG 2019

H. Florez et al. (Eds.): ICAI 2019, CCIS 1051, pp. 17–31, 2019.

https://doi.org/10.1007/978-3-030-32475-9_2

analytics remain top priority for CIOs, due to the return of the investment (ROI) in this kind of projects [9,22].

Educational data mining (EDM) [7] describes the process of converting raw information from a platform or educational system to be used as a knowledge asset for educational entities. EDM seeks to generate value from the information they can gather from the interaction of their students or teachers with their systems. This paper seeks to address the concept of EDM in e-learning platforms, and how data from the interaction records of students or tutors in these platforms can be used to define tools or policies that directly impact the retention and permanence. In traditional learning environments, teachers can obtain feedback of learning through direct interaction with students, enabling a continuous evaluation by teachers [21]. The interaction and observation of students' behavior in the classrooms, as well as the analysis of the history of the courses, give data to estimate appropriate pedagogical strategy to apply in the classroom. However, for work with students in virtual learning environments, this monitoring is more complicated. Tutors should look for other sources than direct observation for audit the learning process of students in the virtual classroom. One of these options is the Web platforms used by educational institutions, which collect large amounts of information automatically from the interaction with their educational systems. The data that these tools collect, after some EDM, can provide multiple dimensions of student behavior to teachers and the institution.

This paper presents a data analytic exercise with the registers (LOGS) coming from a virtual learning platform for higher education programs in virtual mode as an input for the definition of retention and permanence policies. Article is structured as follows: start with the description of the methodology implemented, followed by the exploration and use of data for the recognition of records from educational platforms, and finally the contextualization and use of the same in the definition of student retention and permanence policies.

2 Previous Work

Virtual learning environment (VLEs) or course management systems (CMS) are part of modern pedagogical approaches. These platforms host information about the interaction of the users with these. According to [8,14,23], this information has the potential to improve pedagogical approaches. From this premise, it is found multiple interrelated approaches that seek to exploit the registers or LOGs. In Table 1 is presented different approaches for the analysis of logs in virtual learning environments.

For this particular work, there is special interest in the analysis of the behavior of students and tutors in the VLEs. Taking into account the works related to this topic, Table 2 presents multiple approaches or algorithms centered in different dimensions of student behavior in virtual platforms.

Table 1. Approaches for the analysis of logs in virtual learning environments

Subject	Description	References
Improve the experience in e-learning	Use the Logs of VLEs to identify best practices or policies that allow improving the user experience in these	[3, 12, 20]
Visualization	Log visualization schemes for monitoring and understanding virtual classrooms interactions	[13, 16, 18]
EDM for performance improvement in VLEs	Data mining approaches that allow students to improve academic performance in VLEs	[4, 5, 8, 23, 25]
Platform behavior analysis	Pattern analysis approaches for the analysis of the behavior of students and tutors in VLEs	[1, 2, 10, 14, 15, 17, 19, 24, 25]

Table 2. Mining techniques for the analysis of logs for behavior analysis.

Mining techniques	Algorithm	Description	References
Association rules	Apriori	Find rules of association between behaviors and patterns of records	[14]
Classification	Decision tree, Support vector machines (SVM), NaiveBayes, NearesNeighbours	Design functions to determine the membership of certain classes from patterns in the registers	[14, 25]
Clustering	K-means, Elbow Algorithm, Fuzzy C-Means Clustering	Determine patterns of association between the records of different students	[1, 2, 14, 24]
Process Mining	Heuristic Miner, Fuzzy Miner	Discover underlying processes in event logs	[19]
Intent mining	Viterbi Algorithm, Baum-Welch Algorithm	Model the processes according to the purpose of the actors	[19]
Mining of sequential patterns	Generalized Sequential Patterns (GSP), Sequential Pattern Mining (SPAM), Sequential Pattern Mining Consider, probabilistic models	Find common patterns among the data samples where the values are delivered in a sequence	[15, 17, 19]
Graph mining	Branch-and-bound, On-line Plan Recognition, Recursive Matrix (R- MAT)	Extract patterns (sub-graphics) of interest from the graphs describing the underlying data	[19]

3 Methodology

In this work we seek to develop an approach to identify student behaviors in virtual mode for an academic program using the registers (LOGS) left by them in the virtual learning platforms. The general approach of this work can be summarized by Fig. 1, where:

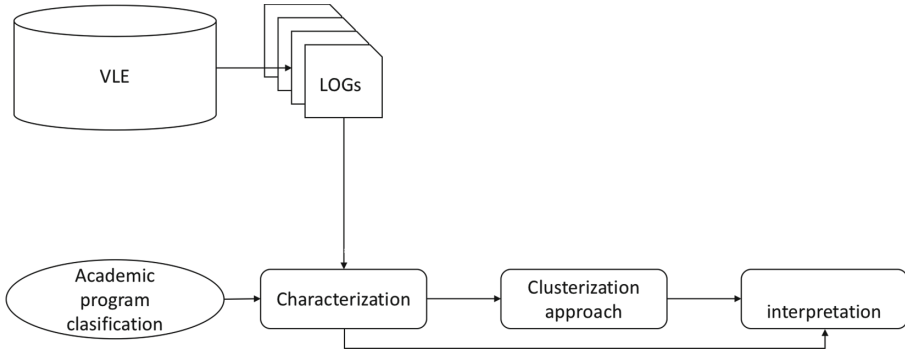


Fig. 1. Academic behavior analysis in academic programs

- **Extraction:** From the VLE databases a representative sample of the students enrolled in the different courses of an academic program is taken. From these students, once they have finished the courses, their LOGs of interaction with the platforms and their notes are extracted.
- **Data Characterization:** Once the logs extracted from a sample of students enrolled in an academic program, these records are transformed by formulating metrics related to their behavior that are easily associated with their academic behavior.
- **Clustering:** Once a database of coded information from the sample drawn from an academic program has been consolidated, an unsupervised clustering algorithm is used to identify the latent behavior patterns in the data sample.
- **Analysis and interpretation:** The information related to the academic performance of the extracted sample is used to label the different groups identified with the clustering algorithms.

3.1 Data Characterization

In a virtual learning platform (Course Management System - CMS), A log is a sequential file with temporal records associated with all events in an academic course product of the student's interactions with the CMS. For finalized courses, we can obtain a set of records of the students and tutors behavior in an specific configuration of the CMS.

For this work, the CMS configuration is defined as following. The set of activities, $A = \{a_1, a_2, \dots, a_n\}$ where n is the number of activities, to performing during a specific course. The set of forums $F = \{f_1, f_2, \dots, f_n\}$ defined for each activity, the agenda (interval time) $T = \{[t_1, t_2], [t_3, t_4], \dots, [t_{(n-k)}, t_n]\}$ defined for each activity, the evaluative weight $P = \{p_1, p_2, \dots, p_n\}$ for each activity where $\sum_{i=1}^n p_i = 500$, the course materials, $M = \{m_1, m_2, \dots, m_w\}$ where w is the number of folders (folders with books, articles, videos, etc.). The students enrolled in the course, $E = \{e_1, e_2, \dots, e_m\}$ where m is the number of students, and the academic ponderation $N = \{n_1, n_2, \dots, n_m\}$ for each student. According to the CMS configuration, a course is defined as the function $C(A, F, T, P, M, E) \rightarrow N$.

The Logs list $L = \{l_1, l_2, \dots, l_m\} \in C$ for each student where $l_i = \{f_i, u_i, ua_i, ec_i, c_i, en_i, o_i, ip_i\} \in e_i$ contains the temporal records of each student, see Table 3.

Table 3. Log structure

Component	Description
f_i	LOG date and time
u_i	User name
ua_i	Interaction with the user
ec_i	Event context
c_i	Course component accessed
en_i	Event name
o_i	Access origin (mobile, web)
ip_i	IP address

Standard logs representation is not enough to contextualize this information in academic terms [3, 13]. To transform the logs into academic relevant information we realize a characterization process in order to measure a set of variables that allows realize an academic interpretation of the behavior of students and tutors in the CMS. We propose a set of variables measured using the logs $l_i \in e_i$ for each student, see Table 4.

In a first approximation, $\forall e_i \in E$ we can characterize each student as $g(l_i) \rightarrow x_i \mid x_i = \{Tp_i, Tt_i, Ef_i, Er_i, Dp_i, De_i, Vm_i, Nl_i, Ls_i, Lf_i, Ld_i, Ln_i, Pa_i, Fa_i\}$. However, this characterization need a special encoding to approach multiple courses particularities.

3.2 Clustering

Let Σ the database of information encoded for an academic program, define the function $F(\Sigma) \rightarrow W$ where $W = [c_1, c_2, \dots, c_h]$ are the different types of behavior that students adopt in the CMS. Assuming that it is not known a priori what

Table 4. Logs characterization for each student

Variable	Description
Tp	Total time spent by student in the CMS
Tt	Average time spent by a student in an activity
Ef	Average posts in forums for each student
Dp	Average time (days) spend by each student to participate in an activity
De	Average time (days) spend by each student to send the activity products
Vm	Number of visits to the course materials
Nl	Number of logs generated by the student
Ls	Number of weekday generated logs
Lf	Number of weekend generated logs
Ld	Number of diurnal generated logs
Ln	Number of nocturnal generated logs
Pa	IP ratio
Fa	Access frequency

types of behavior exist, the problem is to classify or segment the behavior inside the virtual courses. Therefore, the type of function proposed is an unsupervised grouping algorithm [26], specifically for this work, the Self-organizing map (SOM) networks proposed by Kohonen [11] following by a hierarchical clustering are used.

SOM networks are an algorithm based on unsupervised neural networks. The main functionality of the SOM networks is their ability to project nonlinear data with high dimensionality in a regular grid of low dimensionality (usually in 2D). The algorithm look for points that are near each other in the input space to be transformed to nearby map units in the SOM.

For our particular case of study, each information element is the vectors, $X_i \in \Sigma$, where each of its components is a variable with a defined meaning. The grid generated by the SOM network can be used as a basis on which vectors with similar characteristics can be projected using a color-based coding, and based on these generated groupings, explain the possible types of behaviors that can be found in virtual courses.

For our study, this topological mapping consists of projecting a set of vectors X k -dimensional in a two-dimensional discrete mesh (2-D) of M positions, see Fig. 2. Each position in the output is characterized by a node $h_j (j = 1, 2, \dots, M)$. For each h_i node, a position in the output space is associated by \mathbf{w} , which is obtained through an optimization process, which reduce the distance between all the inputs and the output in the new space of M positions.

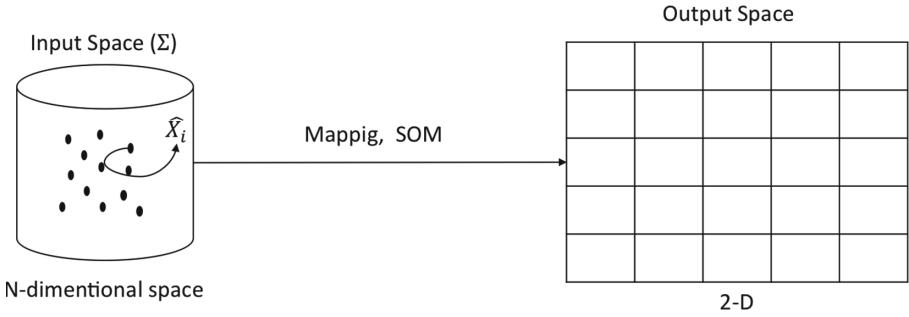


Fig. 2. The SOM mapping pass from a high-dimensional space into a 2-D space.

Starting from the database $\Sigma_{14+(w-1),E}$ (input space), where $14 + (w - 1)$ is the number of characteristics, w the number of classifications within an academic program and E the number of students that have been used to create the database. In Fig. 3 is observed that although the student condition is the same for any sub-classification within an academic program, it is assumed that their behavior may vary depending on the type of subject they are studying.

Group 1	$\{tp_i, Tt_i, \dots Fa_i\}$	0	0	0	0
	$\{tp_i, Tt_i, \dots Fa_i\}$	0	0	0	0

Group 2	0	$\{tp_i, Tt_i, \dots Fa_i\}$	0	0	0
	0	$\{tp_i, Tt_i, \dots Fa_i\}$	0	0	0

Group 3	0	0	$\{tp_i, Tt_i, \dots Fa_i\}$	0	0
	0	0	$\{tp_i, Tt_i, \dots Fa_i\}$	0	0

Group w	0	0	0		$\{tp_i, Tt_i, \dots Fa_i\}$
	0	0	0	0	$\{tp_i, Tt_i, \dots Fa_i\}$

Fig. 3. Input space configuration.

What in terms of the clustering algorithm, will be introduced w primary clusters to which they are looking to perform an additional subcluster, see Fig. 4. To perform this clustering Hierarchical clustering, is used, which is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each one is distinct from the others, and the objects within each cluster are broadly similar.

4 Results

The information of the LOGS and the notes of a semester in five courses of a virtual postgraduate offered by the university was used. The 14 variables already described (Table 4 and N) were obtained, and the students with NaN values in any variable were eliminated. After this preprocessing, the results proposed here are based on a total of 175 students.

Visual Analytics is usually the best way to understand the results of data analytics compared with descriptives techniques, since it facilitates to identify relations between the information. The first exploration, presented in Fig. 5, explore the distribution of the each variable through the diagonal of image table. The distribution of all variables are spread over all the range of data, but some variables can be approximated to some probabilistic distributions. Also, the relation between all the variables of vector X is presented, where it is possible o view some correlation between the variables like is presented in the Fig. 6.

Then, the next step was to obtain a SOM network of dimension five by five. Over this SOM Network the process of grouping, using hierarchical clustering, was done, which allows combining the nodes that are similar and that are side by side in the SOM grid. The results of the grouping with three groups are presented in Fig. 7. To analyze the behavior of each variable, Fig. 9 presents heat maps of the inputs, in which it is possible to analyze the relationship between the clusters and the ranges of the data. Finally, the Fig. 8 presents the histogram of the variable N for each cluster. Both the blue and the green cluster are made up of students with low and medium academic performance, respectively, and are characterized by having little participation in the forums (Ef), low frequency of

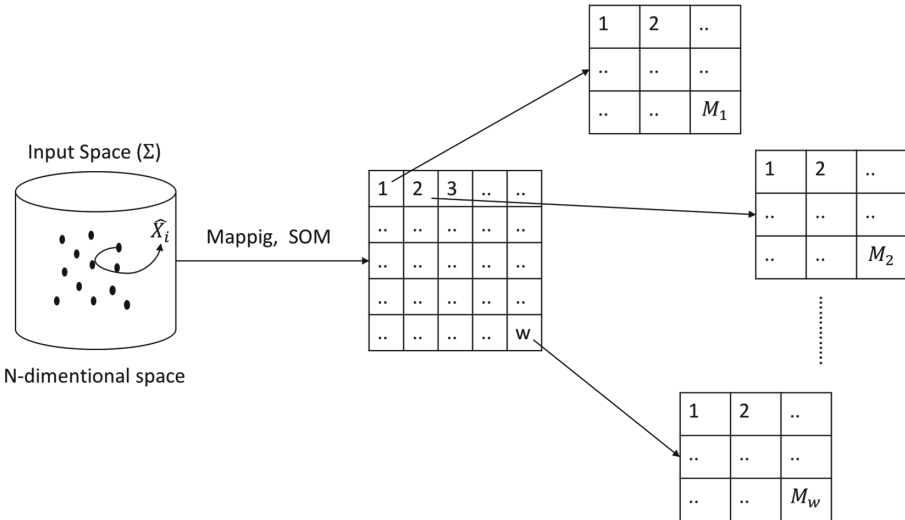


Fig. 4. Clusterization scheme.

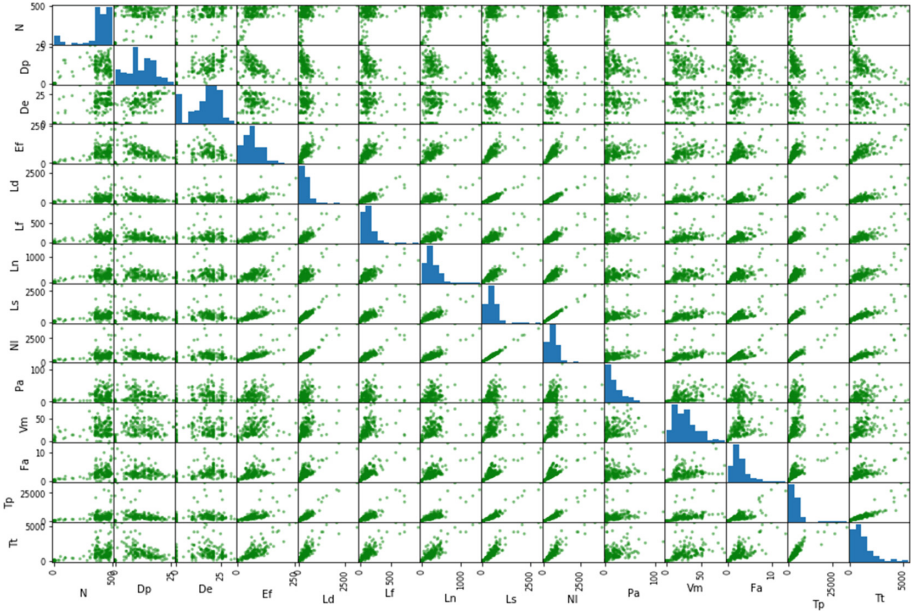


Fig. 5. SOM network 5×5 with cluster using hierarchical clustering of three. (Color figure online)

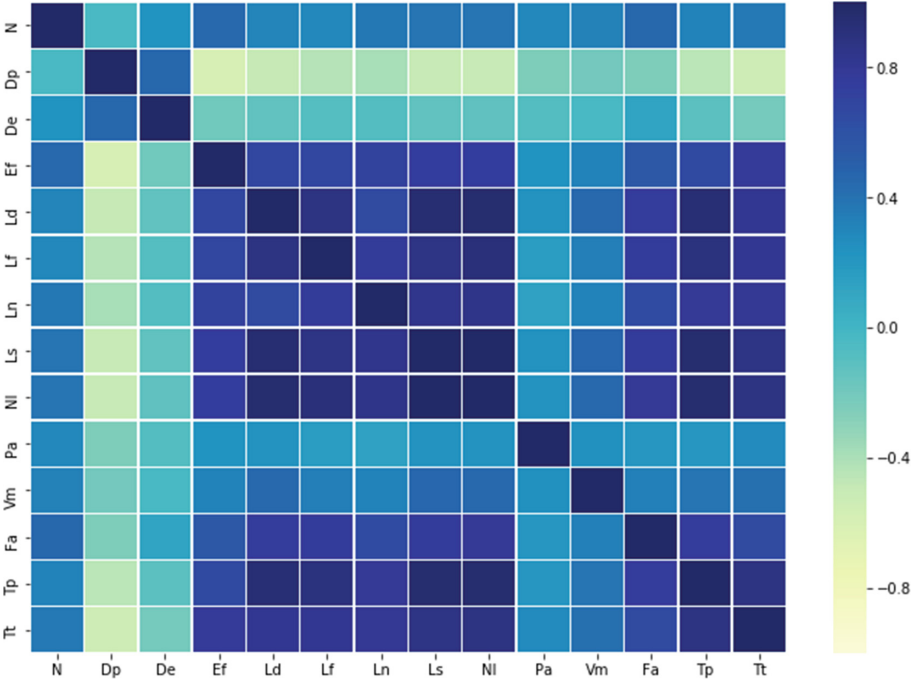


Fig. 6. Correlation of the input data x .

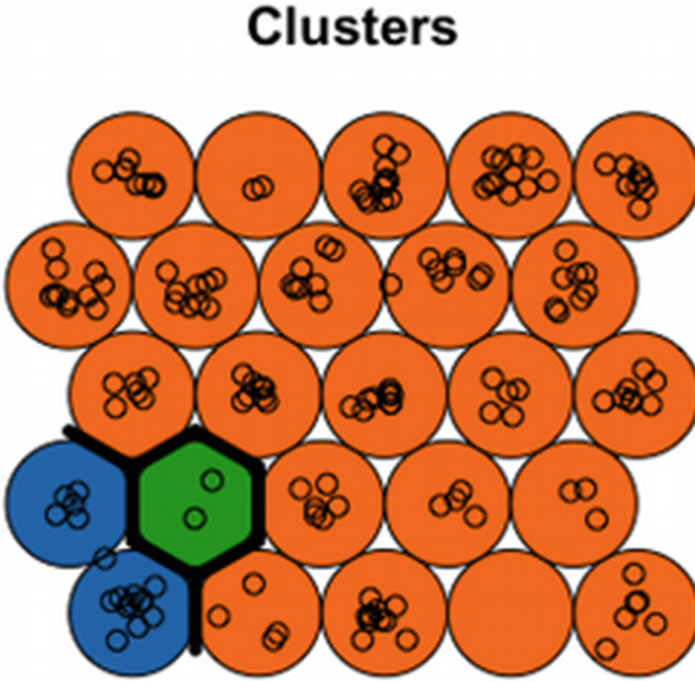


Fig. 7. SOM network 5×5 with cluster using hierarchical clustering of three.

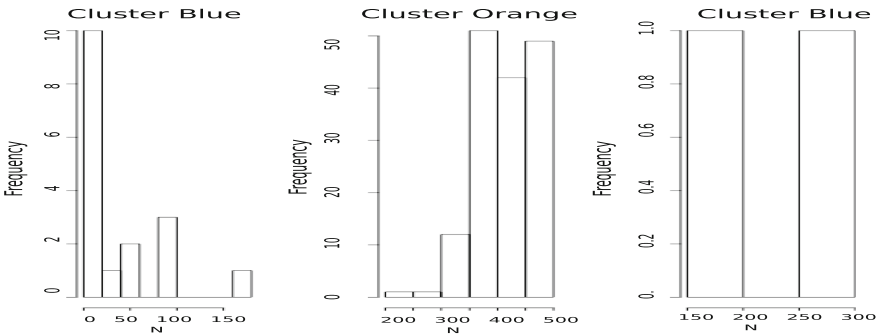


Fig. 8. Histogram of variable Notes (N) for the three clusters.

access (Fa) and little access to the course material (Vm). Otherwise, it happens in the orange cluster, made up of students with the best grades in their majority.

This procedure was repeated for a different number of groups as well as different input variables, which allowed to obtain the following results:

- In the clusters there is a differentiation in a shorter time of participation in the forums (Dp); likewise, there is a more significant number of contributions

- between the groups. Regarding the number of logs generated during the week and the frequency of access, certain groups are more abundant.
- Students with low grades, where the total is less than 250 are mainly those who do not participate as much in the activities as in the deliveries and represent 10% of the sample.

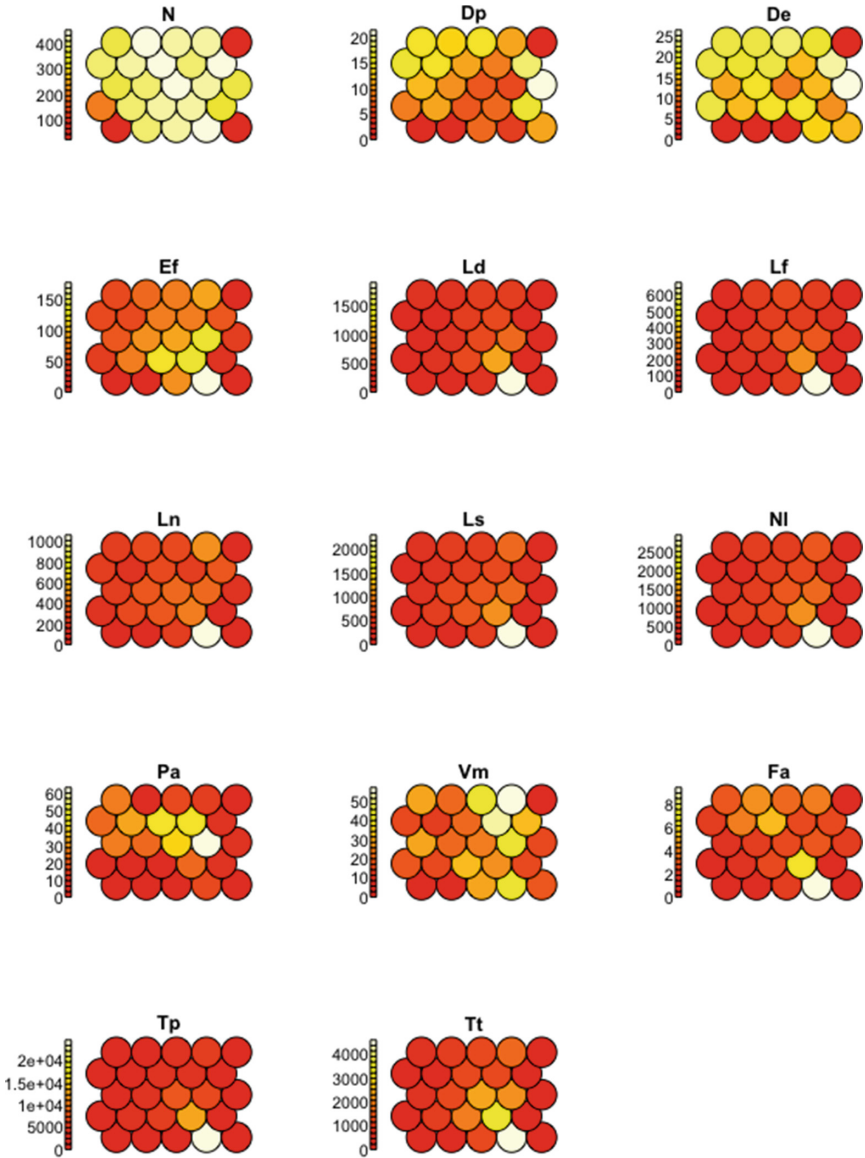


Fig. 9. Heat Maps of each variable over the SOM network architecture.

- The time it takes a student to make the deliveries (Of), as well as the participation in the forums (Dp) does not have an impact on the notes.
- Few contributions in the forums (Ef) and low frequency of access (Fa) to the platform is positively correlated with better grades.
- The variables Ln , Ld , Ls , Lf , Vm , Tt_{min} , Tp_{min} , Nl , and Pa have a weak correlation with the notes.
- Students have greater participation in the forums in the daytime than in the nighttime. Also, they participate more during the week than weekends.
- Students participate about 55% on the day versus 45% at night; however, there is a 76% participation in the platform during the week compared to 24% on the weekend.

5 Discussion

Logs analysis in educational platforms is not a new issue. Even, in terms of characterizing the students' behavior in this kind of technological tools. Research papers as shown in Tables 1 and 2 afford similar approaches like this one. However, the principal contribution of this study is the application of this kind of analysis in a particular population with a socio-economical, cultural, geographical and political environment, in order to afford in future works, use this information to define policies, design new interfaces for technological tools, between others.

This work is framed in the analysis of Colombian virtual students behavior. principally, to understanding how to prevent desertion in virtual high education programs. In this project phase, the characterization approach, and the mining tool was defined, in order to in a future phase use this information to understand the Colombian virtual Student and also improve the technological tools.

6 Conclusion

Teaching or coaching tasks in virtual environments imply new challenges in pedagogy. Virtual Learning environments provide useful tools for the interaction between students, teachers, learning materials, and also with the educational institutions. However, it creates barriers between the players participating in the learning process, such as the understanding of students or teachers behavior beyond the actions in the virtual platforms.

Mathematical approaches that allow continuous monitoring of student behavior according to the logs analysis in VLE, can help to define educational policies in order to take preventive actions in order to reduce the desertion. This also suggests the need to design involving a greater number of variables related to the behavior of the teacher and whose purpose is focused on determining how they affect the performance of their students.

In the development of this work, we will include a representative sample to the analysis of a complete higher educational institution, and also include socioeconomic and geographic information concerning understand particularities

of virtual students. With this information we want to develop a support tool to aid policies makers to plan in order to reduce desertion and increase the quality of service.

References

1. Alias, U.F., Ahmad, N.B., Hasan, S.: Mining of E-learning behavior using SOM clustering. In: 6th ICT International Student Project Conference: Elevating Community Through ICT, ICT-ISPC 2017, pp. 1–4 (2017). <https://doi.org/10.1109/ICT-ISPC.2017.8075350>
2. Bara, M.W., Ahmad, N.B., Modu, M.M., Ali, H.A.: Self-organizing map clustering method for the analysis of e-learning activities. In: 2018 Majan International Conference (MIC), pp. 1–5, March 2018. <https://doi.org/10.1109/MINTC.2018.8363155>
3. Baruque, C.B., Amaral, M.A., Barcellos, A., da Silva Freitas, J.a.C., Longo, C.J.: Analysing users' access logs in Moodle to improve e learning. In: Proceedings of the 2007 Euro American Conference on Telematics and Information Systems, EATIS 2007, pp. 72:1–72:4. ACM, New York (2007). <https://doi.org/10.1145/1352694.1352767>
4. Charitopoulos, A., Rangoussi, M., Koulouriotis, D.: Educational data mining and data analysis for optimal learning content management: applied in Moodle for undergraduate engineering studies. In: 2017 IEEE Global Engineering Education Conference (EDUCON), pp. 990–998, April 2017. <https://doi.org/10.1109/EDUCON.2017.7942969>
5. Conde, M., Garca-Pealvo, F., Fidalgo-Blanco, Sein-Echaluce, M.: Study of the flexibility of a learning analytics tool to evaluate teamwork competence acquisition in different contexts. In: CEUR workshop Proceedings, vol. 1925, pp. 63–77 (2017). ceur-ws.org/Vol-1925/paper07.pdf. cited By 0
6. Dhingra, S., Chaudhry, K.: A study of the impact of data warehousing and data mining implementation on marketing effort. *Int. J. Adv. Stud. Comput. Sci. Eng.* **7**(1), 13–20 (2018)
7. Elaal, S.: E-learning using data mining. *Chin. Egypt. Res. J. Helwan Univ.* (2013)
8. Gamie, E.A., El-Seoud, M.S.A., Salama, M.A., Hussein, W.: Pedagogical and elearning logs analyses to enhance students' performance. In: Proceedings of the 7th International Conference on Software and Information Engineering, ICSIE 2018, pp. 116–120. ACM, New York (2018). <https://doi.org/10.1145/3220267.3220289>. Cited by 0
9. Grover, V., Chiang, R.H., Liang, T.P., Zhang, D.: Creating strategic business value from big data analytics: a research framework. *J. Manag. Inf. Syst.* **35**(2), 388–423 (2018)
10. Hernández-García, Á., Acquila-Natale, E., Iglesias-Pradas, S., Chaparro-Peláez, J.: Design of an extraction, transform and load process for calculation of teamwork indicators in Moodle. In: LASI-SPAIN (2018). ceur-ws.org/Vol-2188/Paper7.pdf
11. Kohonen, T.: The self-organizing map. *Proc. IEEE* **78**(9), 1464–1480 (1990). <https://doi.org/10.1109/5.58325>

12. Kolekar, S.V., Pai, R.M., Manohara Pai, M.M.: Adaptive user interface for Moodle based E-learning system using learning styles. *Procedia Comput. Sci.* **135**, 606–615 (2018). <https://doi.org/10.1016/j.procs.2018.08.226>. The 3rd International Conference on Computer Science and Computational Intelligence (ICCCSCI 2018): Empowering Smart Technology in Digital Era for a Better Life
13. Konstantinidis, A., Grafton, C.: Using Excel Macros to Analyse Moodle Logs. *UK Research.Moodle.Net*, pp. 4–6 (2013). http://research.moodle.net/pluginfile.php/333/mod_data/content/1233/UsingExcelMacrostoAnalyseMoodleLogs.pdf
14. Moreira Félix, I., Ambrósio, A.P., Silva Neves, P., Siqueira, J., Duilio Brancher, J.: Moodle predicta: a data mining tool for student follow up. In: *Proceedings of the 9th International Conference on Computer Supported Education 1 (CSEDU)*, pp. 339–346 (2017). <https://doi.org/10.5220/0006318403390346>
15. Poon, L.K.M., Kong, S.-C., Wong, M.Y.W., Yau, T.S.H.: Mining sequential patterns of students' access on learning management system. In: Tan, Y., Takagi, H., Shi, Y. (eds.) *DMBD 2017. LNCS*, vol. 10387, pp. 191–198. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61845-6_20
16. Poon, L.K.M., Kong, S.-C., Yau, T.S.H., Wong, M., Ling, M.H.: Learning analytics for monitoring students participation online: visualizing navigational patterns on learning management system. In: Cheung, S.K.S., Kwok, L., Ma, W.W.K., Lee, L.-K., Yang, H. (eds.) *ICBL 2017. LNCS*, vol. 10309, pp. 166–176. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59360-9_15
17. Qiao, C., Hu, X.: Discovering student behavior patterns from event logs: Preliminary results on a novel probabilistic latent variable model. In: *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, pp. 207–211, July 2018. <https://doi.org/10.1109/ICALT.2018.00056>
18. Raga, R.C., Raga, J.D.: A comparison of college faculty and student class activity in an online learning environment using course log data. In: *2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCOM/IOP/SCI)*, pp. 1–6, August 2017. <https://doi.org/10.1109/UIC-ATC.2017.8397475>
19. Ros, S., Lázaro, J.C., Robles-Gómez, A., Caminero, A.C., Tobarra, L., Pastor, R.: Analyzing content structure and Moodle milestone to classify student learning behavior in a basic desktop tools course. In: *Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality, TEEM 2017*, pp. 42:1–42:6. ACM, New York (2017). <https://doi.org/10.1145/3144826.3145392>
20. Porras, J.T., Alcántara-Manzanares, J., García, S.R.: Virtual platforms use: a useful monitoring tool. *EDMETIC* **7**(1), 242–255 (2018). <https://doi.org/10.21071/edmetic.v6i2.8696>
21. Sheard, J., Ceddia, J., Hurst, J., Tuovinen, J.: Inferring student learning behaviour from website interactions: a usage analysis. *Educ. Inf. Technol.* **8**(3), 245–266 (2003). <https://doi.org/10.1023/A:1026360026073>
22. Shim, J.P., French, A.M., Guo, C., Jablonski, J.: Big data and analytics: issues, solutions, and ROI. *CAIS* **37**, 39 (2015)
23. Smith, S.M., et al.: How might the development of data mining and log analysis systems for the Moodle virtual learning environment improve computer science students' course engagement and encourage course designers' future engagement with data analysis methods for the evaluation of course resources? Ph.D. thesis, University of Lincoln (2017). <http://eprints.lincoln.ac.uk/30882/>

24. Vega, A.B.: Mejora en el descubrimiento de modelos de minería de procesos en educación mediante agrupación de datos de interacción con la plataforma Moodle. Ph.D. thesis, Universidad de Córdoba (2018)
25. Verma, A., Rathore, S., Vishwakarma, S., Goswami, S.: Multilevel analysis of students feedback using Moodle logs in virtual cloud environment. *Int. J. Comput. Sci. Inf. Technol.* **9**, 15–28 (2017). <https://doi.org/10.5281/zenodo.2558650>
26. Wu, X., et al.: Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **14**(1), 1–37 (2008). <https://doi.org/10.1007/s10115-007-0114-2>