Contributions to real-time monitoring and analysis of heterogeneous learning environments

Lucia Uguina-Gadella[1,2,*]

[1] IMDEA Networks, Leganés, Spain
[2] Universidad Carlos III de Madrid, Leganés, Spain
lucia.uguina@imdea.org

Abstract. The ubiquity and flexibility of heterogeneous learning environments allows gathering a huge amount of data from students' interactions. Applying learning analytics and data mining to these data, as well as a self-regulated learning criterion, is a well-accepted method to learn students' behavior and ultimately predict their learning outcomes. In order to enrich the learning experience, the prediction should be done before the failure occurs. Thus, the thesis proposed in this paper aims to contribute with several prediction algorithms based in students' interactions gathered through events in a real-time basis. This could be used to early detect students at risk and help them to succeed.

Keywords: Learning analytics, heterogeneous learning environments, prediction, self-regulated learning, real-time

1 Introduction

Over the past twenty years, technological innovation has changed the educational environment [1]. Information technologies such as personal computers, mobile phones and the Internet, have altered the education experience. Thanks to these technologies, students are now connected to teachers and could access educational resources every day and everywhere [2].

The main objective of these technologically enhanced learning environments is to help students in their learning process [3]. However, these environments are the perfect situation for students to generate a large and meaningful amount of data while working on their assignments. "The measurement, collection, analysis and reporting of these data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" is called learning analytics [4]. There are different challenges in the learning analytics field, such as analysis and visualization of data, predicting student's performance, providing feedback for instructors, student modeling, detecting undesirable students' behavior, recommendation for students…

[3]. Nowadays, one of these challenges, predicting student's performance, is generating lots of research publications. Nevertheless, most of them focus on only distance educational environments such as MOOCs (Massive Online Open Courses) or online university degrees [5], [6] or rely only on past courses grades to predict students' performance [7]. These research publications neglect the flexibility and ubiquity of blended learning environments, those which combine online and face-to-face learnings [8].

The blended learning environments not only provide advantages to students but also creates great challenges in terms of engagement and autonomy [9]. The courses with a blended or online environment expect from their students a high level of self-regulation [10]. Literature suggests that self-regulated learning and autonomy could contribute to students' success [11], [12].

The predictive approach, enhanced with a self-regulated learning consideration, could be used for solving one of the other challenges in learning analytics, which is to provide students the appropriate guidance at the right time [13]. As there exists a necessity from both teachers and students to obtain feedback of their behavior in a real-time basis [14] the early prediction of students' grades could lead to an appropriate intervention performed by the instructor in order to help students at risk of failing the subject.

The aim of this thesis is to develop and test a predictive alert system, based on the Lostrego system infrastructure [15], that could detect students at risk in a real-time basis adapting its prediction to the students' behavior. In order to achieve this, several research questions have to be addressed:

RQ1. *Is it possible to obtain an accurate prediction of the students' grades using actions on their working environments and, if possible, midterm exams scores in order to detect student at-risk at course early stages?*

RQ2. *If RQ1 seems possible, could instructor's interventions to students at-risk improve their performance in the subject?*

RQ3. *Are students self-aware of their learning behavior? Could self-regulation be related with grades?*

2 Related work

Forecasting future outcomes in education is a popular research topic with numerous contributions [16], [17], [18], [19]. Due to this, a systematic review on this topic has been done in order to understand present state of the problem addressed as well as possible existing solutions.

Online learning environment is, nowadays, the most popular environment for predictive analysis as MOOCs platforms offer internal records of students' interactions [20]. For instance, several contributions done over MOOCs proofs that video viewing time correlates with grades [21], Pérez-Lemonche et al. [6] predicted students' performance in a MOOC with a Median Absolute Difference (MAD) of around 10% and Moreno-Marcos et al. [5] used exercises attempts and forum participation as predictive variables for forecasting students' performance in a MOOC.

Other approximations to predictive analysis are done over blended environments but taking only online resources as data sources for the predictive variables. Jokhan et al. [22] used the Moodle resource to identify students at risk with an accuracy of 60.8%. Nguyen et al. [14] created a model to predict students' results in online learning with an accuracy of only 50%. In true blended environments the issue is that the analysis is usually done after the course is finished, some examples are [23] or [24].

Although several researchers only considered past grades or CGPA (Cumulative Grade Point Average) [7], [18], [25], [26], [27], [28], [29], Aguiar et al. [16] concluded that, on the majority of the situations, CGPA was not enough to predict consistently students' performance or dropout rates.

Moreover, some other researchers have developed intelligent tutoring systems (ITSs) in order to support students [30] or [31]. Nevertheless, the problem with ITSs is that it is required for the student to use the tutoring system purposely in order to gather data or to provide feedback. In order to address this problem, the Ztreamy [32] and Lostrego [15] systems are the infrastructure for data gathering and event processing. The students generate data and events while working on a virtual machine provided at the very beginning of each course. This data generation could also be gathered in any Linux system (as Linux architecture is required for this specific subject). Thus, the events can be analyzed without the issue of the students' interaction with a specific tool.

3 Research methodology

In order to achieve the goal of this thesis, several steps should be considered for the research methodology.

First of all, the data gathering. This step is divided in three different data sources:

- Real-time data from students of the Systems Architecture subject (Bachelor in Telecommunication Technologies Engineering, 2nd course). These are gathered through Ztreamy [32] and Lostrego [15] systems. These infrastructures provide a real-time gathering of events generated by students through both their personal computers and their access (from any device) to the course webpage.
- Self-learning regulation tests and pre-knowledge tests done at the very beginning of the course.
- Students monitoring through mid-term exams and weekly polls.

Secondly, method test and validation:

- Compilation and analysis of different clustering methods in order to identify students at-risk.
- Compilation and analysis of different machine learning algorithms for students' grade prediction through gathered data.

- Method validation, both clustering and prediction, with data from past and present years.

Finally, development of the tool and prediction methods:

- Real-time prediction of grades and of the classification of the students (at risk or safe) through a web-based graphical tool that could be used both by teacher and student.
- Help and tutoring of students thanks to the prediction and classification following a design-based research. These interventions will be analyzed with experimental and control groups in order to validate whether the interventions are useful for the students.

4 Expected contribution

The expected contribution of this thesis could be divided in three different approaches:

- Students' final grades prediction with the development of machine learning algorithms in heterogeneous environments through the gathering of real-time events.
- Analysis of the students' self-regulated learning and the research of a possible relationship between self-regulation and final grade.
- Improvement of the learning experience through specific instructor's interventions motivated by the analysis of the students' behavior during the course.

These proposed solutions address not only the challenge of providing useful information at the right time, but they also create the need to analyze whether interventions are ultimately useful for the students' success. Moreover, as self-regulated learning is taken into consideration, it could provide a greater insight of the students' self-awareness of their learning behavior.

This contribution, if compared to the existing research publications mentioned in Related work section, gives a more detailed and real-time basis approach than the other publications as the events are gathered throughout the course without any additional interaction from the student. Moreover, this contribution could also be extended to different course types as it only needs an event gathering type of environment.

5 Current state

The current state of this thesis could be described as a middle state. Several advances have been done in some of the stages of the research methodology.

Two consecutive years of full data (web events, system events and working sessions) have been recorded and stored. One year of partial web data and full events and session is also recorded and stored as well as one year of only events and session. Self-learning regulation tests as well as pre-tests have been provided at the very beginning of two consecutive years and their results have been partly analyzed as well as stored. Students have been monitored with mid-term exams without variations in terms of course structure during two consecutive years and with slight variations in a third one. Polls have been provided weekly during two consecutive years.

The compilation and analysis of different clustering methods was done at the very first stage of the thesis work as well as the analysis of machine learning algorithms. These analyses were performed in two consecutive years with the statistics shown in **Table 1**.

The identification of failing students as well as the analysis of several algorithms in order to predict the final grade of those students is an ongoing work near to the publication phase.

Table 1. Course statistics

	2017		2018	
	Per student	Total	Per student	Total
Total events	4,984.3	618,053	5,189.41	871,821
Web events	3,741.54	463,951	3,624.18	608,863
Working sessions	324.31	40,215	382.28	64,223

Moreover, an analysis of the students' self-awareness of their learning behavior as well as the analysis of the relationship between self-regulated learning and students' performance was carried out in an accepted EDUCON 2020 paper [33]. In this work, a group of students answered a self-regulated learning questionnaire with a question about how they spaced their working time sessions. The answers were the ones shown in **Table 2**.

Table 2. Questionnaire responses

	Spaced their study sessions	Studied in one session before test
Followed subject	60	11
Dropouts	51	9

The events generated by students while working allowed to determine that, in the majority of the cases, students were self-aware of their working patterns. In addition to this, several correlation and ANOVA analyses were done in order to study the relationship of self-regulated learning and final marks. These analyses showed that variance (how students regulated their working sessions through time) was a promising predicting variable with a remarkable relationship with the students' final grade. Thus, this led to a further research in how self-regulated learning and time management could affect to students' performance. The ANCOVA analysis results can be shown in **Table 3**.

Table 3. ANCOVA analysis

	SS	DF	F	p-unc
Questionnaire answer	2.024	1	0.229	0.633986
Variance	72.253	1	8.169	0.005744

Total time spent	4.007	1	0.453	0.503339

Finally, another paper was published [34] that presented a case study of a flipped classroom analyzed through learning analytics and data-driven learning design. Some of the conclusions extracted in the publication were that students were highly mark and deadline oriented and they delayed their work, transforming the flipped approach into a conventional classroom. This research gave a preliminary introduction to how students work and what motivates them in order to adapt future interventions to their work and study patterns.

Future research plans for this thesis work are the ones listed below.

- Analysis of learning behavior through process mining. With this analysis, several work patterns or processes could be labeled as desired paths and students with dangerous or more risky processes could be alerted.
- Analysis of self-regulated learning throughout the entire course.
- Continuation of interventions in sample groups and comparing the results with control groups during several years.

Acknowledgement

This PhD proposal is done under the advisory of Iria Estévez Ayres and Jesús Arias Fisteus. It is partially funded by: FEDER/Ministerio de Ciencia, Innovación y Universidades-Agencia Estatal de Investigación, through the project Smartlet (TIN2017-85179-C3-1-R); the Community of Madrid through its regional project "eMadrid" (S2018/TCS-4307); the Erasmus+ programme of the European Union through the InnovaT project (598758-EPP-1-2018-1-AT-EPPKA2-CBHE-JP).

References

1. M. Escueta, V. Quan, A. J. Nickow and P. Oreopoulos, «Education technology: An evidence-based review,» *National Bureau of Economic Research*, nº w23744, 2017.
2. G. J. Hwang and H. Chang, «A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students,» *Computers & Education*, vol. 56, nº 4, pp. 1023-1031, 2011.
3. C. Gonzalez-Nespereira, A. Fernandez-Vilas and R. P. Diaz-Redondo, «Am I failing this course? Risk prediction using e-learning data,» de *Proceedings of the 3rd International Conference on Technological Ecosystems for Enhancing Multiculturality*, 2015.
4. «Proceedings of the 1st International Conference on Learning Analytics and Knowledge,» Association for Computing Machinery, New York, 2011.
5. P. Moreno-Marcos, P. J. Muñoz-Merino, C. Alario-Hoyos, I. Estévez-Ayres and C. Delgado-Kloos, «Analysing the predictive power for anticipating assignment grades in a massive open online course,» *Behaviour & Information Technology*, vol. 37, nº 10-11, pp. 1021-1036, 2018.

6. A. Pérez-Lemonche, G. Martínez-Muñoz and E. Pulido-Cañabate, «Analysing Event Transitions to Discover Student Roles and Predict Grades in MOOCs,» *Artificial Neural Networks and Machine Learning – ICANN,* pp. 224-232, 2017.
7. J. Xu, K. H. Moon and M. van der Schaar, «A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs,» *IEEE Journal of Selected Topics in Signal Processing,* vol. 11, nº 5, pp. 742-753, 2017.
8. R. T. Osguthorpe and C. R. Graham, «Blended learning environments: Definitions and directions,» *Quarterly review of distance education,* vol. 4, nº 3, pp. 227-233, 2003.
9. N. Gedik, E. Kiraz and M. Y. Ozden, «The Optimum Blend: Affordances and Challenges of Blended Learning For Students,» *Online Submission,* vol. 3, nº 3, pp. 102-117, 2012.
10. Y. Zhu, W. Au and G. Yates, «University students' self-control and self-regulated learning in a blended course,» *The Internet and Higher Education,* vol. 30, pp. 54-62, 2016.
11. J. Maldonado-Mahauad, M. Pérez-Sanagustín, P. Moreno-Marcos, C. Alario-Hoyos, P. Muñoz-Merino and C. Delgado-Kloos, «Predicting Learners' Success in a Self-paced MOOC Through Sequence Patterns of Self-regulated Learning,» *Lifelong Technology-Enhanced Learning.*, pp. 355-369, 2018.
12. B. J. Zimmerman, «Self-Regulated Learning and Academic Achievement: An Overview,» *Educational Psychologist,* vol. 25, nº 1, pp. 3-17, 1990.
13. G.-J. Hwang, «Definition, framework and research issues of smart learning environments - a context-aware ubiquitous learning perspective,» *Smart Learning Environments,* vol. 1, nº 1, 2014.
14. V. A. Nguyen, Q. Nguyen and V. T. Nguyen, «A Model to Forecast Learning Outcomes for Students in Blended Learning Courses Based On Learning Analytics,» de *Proceedings of the 2nd International Conference on E-Society, E-Education and E-Technology - ICSET,* 2018.
15. I. Estévez-Ayres, J. A. Fisteus and C. Delgado-Kloos, «Lostrego: A distributed stream-based infrastructure for the real-time gathering and analysis of heterogeneous educational data,» *Journal of Network and Computer Applications,* vol. 100, pp. 56-68, 2017.
16. E. Aguiar, G. C. N. Alex Ambrose, V. Goodrich and J. Brockman, «Engagement vs Performance: Using Electronic Portfolios to Predict First Semester Engineering Student Persistence,» *Journal of Learning Analytics,* vol. 1, nº 3, pp. 7-33, 2014.
17. R. S. K. McKenzie, «Who Succeeds at University? Factors predicting academic performance in first year Australian university students,» *Higher Education Research & Development,* vol. 20, nº 1, pp. 21-33, 2001.
18. A. Olani, «Predicting First Year University Students' Academic Success,» *Electronic Journal of Research in Education Psychology,* vol. 7, nº 19, 2017.
19. S. Abdulkadir, «Predicting Students' First-Year Academic Performance Using Entry Requirements for Faculty of Science in Kaduna State University, Kaduna – Nigeria,» *American Journal of Computer Science and Technology,* vol. 2, nº 1, p. 9, 2019.
20. A. del Blanco, A. del Blanco, A. Serrano, M. Freire, I. Martinez-Ortiz and B. Fernandez-Manjon, «E-Learning standards and learning analytics. Can data collection be improved by using standard data models?,» de *IEEE Global Engineering Education Conference (EDUCON)*, 2013.
21. D. Vu, P. Pattison and G. Robins, «Relational event models for social learning in MOOCs,» *Social Networks,* vol. 43, pp. 121-135, 2015.

22. A. Jokhan, B. Sharma and S. Singh, «Early warning system as a predictor for student performance in higher education blended courses,» *Studies in Higher Education*, pp. 1-12, 2018.
23. O. H. T. Lu, A. Y. Q. Huang, J. C. Huang, A. J. Q. Lin, H. Ogata and S. J. H. Yang, «Applying Learning Analytics for the Early Prediction of Students' Academic Performance in Blended Learning,» *Journal of Educational Tehcnology & Society*, vol. 21, nº 2, pp. 220-232, 2018.
24. F. Okubo, T. Yamashita, A. Shimada and H. Ogata, «A neural network approach for students' performance prediction,» de *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK*, 2017.
25. D. M. D. Angeline, «Association rule generation for student performance analysis using apriori algorithm,» *The SIJ Transactions on Computer Science Engineering & its Applications (CSEA)*, vol. 1, nº 1, pp. 12-16, 2013.
26. M. Mayilvaganan and D. Kalpanadevi, «Comparison of classification techniques for predicting the performance of students academic environment,» de *International Conference on Communication and Network Technologies*, 2014.
27. S. Natek and M. Zwilling, «Student data mining solution–knowledge management system related to higher education institutions,» *Expert Systems with Applications*, vol. 41, nº 14, pp. 6400-6407, 2014.
28. T. M. Christian and M. Ayub, «Exploration of classification using NBTree for predicting students' performance,» de *International Conference on Data and Software Engineering (ICODSE)*, 2014.
29. A. M. Shahiri, W. Husain and N. A. Rashid, «A Review on Predicting Student's Performance Using Data Mining Techniques,» *Procedia Computer Science*, vol. 72, pp. 414-422, 2015.
30. S. Graf, Kinshuk and C. Ives, «A Flexible Mechanism for Providing Adaptivity Based on Learning Styles in Learning Management Systems,» de *10th IEEE International Conference on Advanced Learning Technologies*, 2010.
31. M. Feng, N. T. Heffernan and K. R. Koedinger, «Predicting State Test Scores Better with Intelligent Tutoring Systems: Developing Metrics to Measure Assistance Required,» *Intelligent Tutoring Systems*, pp. 31-40, 2006.
32. J. Fisteus, N. F. Garcia, L. Fernandez and D. Fuentes-Lorenzo, «Ztreamy: A Middleware for Publishing Semantic Streams on the Web,» *SSRN Electronic Journal.*
33. L. Uguina-Gadella, I. Estévez-Ayres, J. Arias Fisteus, C. Delgado-Kloos, «Application of learning analytics to study the accuracy of self-reported working patterns in self-regulated learning questionnaires,» *EDUCON – IEEE Global Engineering Education Conference*, 2020.
34. I. Estévez-Ayres, J. Arias Fisteus, L. Uguina-Gadella, C. Alario-Hoyos and C. Delgado-Kloos, «Uncovering Flipped-Classroom Problems at an Engineering Course on Systems Architecture Through Data-Driven Learning Design,» *International Journal of Engineering Education*, vol. 34, nº 3, pp. 865-878, 2018.