Integrating an at-risk student model into learning management systems[*]

Júlia Marques Carvalho da Silva

Instituto Federal do Rio Grande do Sul
Av. Osvaldo Aranha, 540
Bento Gonçalves, RS, Brasil
julia.silva@bento.ifrs.edu.br

Darin Hobbs

Athabasca University
10011 109 Street Northwest,
Edmonton, AB, Canada
darinh@shaw.ca

Sabine Graf

Athabasca University
10011 109 Street Northwest,
Edmonton, AB, Canada
sabineg@athabascau.ca

ABSTRACT

Every year, many students are at risk of failing a course. Teachers can detect whether students are having problems through their behavior and performance during the class. However, in online courses, it can be difficult to get information about students' behaviour and how they are doing in the course since such information is not always available in learning management systems. This study proposes a model to identify at-risk students based on four factors: percentage of activities delivered, average grades, percentage of resources viewed, and attendance. The model has been evaluated with 88 students of nine undergraduate courses in the middle and in the end of the term. The results show an accuracy of 85% in at-risk student identification.

K.3.1 [**Computers and Education**]: Computers Uses in Education – *Computer-managed instruction (CMI), Distance learning*

General Terms

Algorithms, Experimentation, Human Factors.

Keywords

Students' retention, at-risk factors, Moodle.

1. INTRODUCTION

According to the latest study from the Organisation for Economic Co-operation and Development (OECD), only 68% of students complete university programs and only 79% are able to get a secondary diploma [11]. It means that many people drop out of programs, after failing or withdrawing in one or more courses. The retention rate could rise if teachers and course managers put more effort in diagnosing these students and helping them before they drop out. However, the task of identifying these at-risk students is not easy, because it requires observing students' behaviors, such as low yield, retention, behavioral problems, low attendance, low socioeconomic status, etc. [15]. In addition, at-risk students may be found in all educational levels (from kindergarten to postgraduate) and modalities (face-to-face, distance or blended). In face-to-face classes, teachers are more aware to diagnosis learning difficulties through students' behaviors, and by the results on assessments and daily activities. However, online or blended courses require more teacher effort to identify at-risk students.

Currently, more technology has enhanced learning. One example are learning management systems (LMSs), which require students' and teachers' registration to access courses and learning objects. In such LMSs, students can read texts, contribute with a post, and submit an assessment. On the other hand, teachers can monitor students' actions, for example, if they had delivered an assignment, or communicated with other participants. Typically, all actions of students are captured and stored by the LMS that maintain high quantity and quality of data [4, 9]. However, while this data can provide rich feedback about students' achievement, the analysis, comprehension, and interpretation of what the data is indicating is not always easy to do. The information is dispersed, so that teachers need to bring all information together to get an overall status of a student in the course and sometimes, some data are not accessible for teachers at all.

Our research addresses this shortcoming and proposes a model to identify at-risk students based on students' performance and behavior, considering the respective course design. A course design shows how a teacher planned and organized learning objects according to the learning goals. Some courses use many readings, while others focus on assignments and/or forums. Behavior of a student can be defined as how a student uses the learning objects. It includes, for example, students' content visits and attendance in LMSs. The proposed model is designed to work with an LMSs, gathering data, identifying students who are at-risk, and visualizing this information to teachers. Moreover, the proposed model forecasts at-risk levels of students in real-time. The model was evaluated through a well-known LMS namely, Moodle, with real data from 88 students in nine courses. Because the aim of this research is to inform teachers in advance about the students who might not complete the course, teachers can modify their teaching strategies or focus on those identified as at-risk, which can help in increasing the course retention rate.

The remainder of the paper is structured as follows. Section 2 describes a literature review on factors used to identify at-risk students. Section 3 addresses details about the proposed model, including the factors selected, the algorithm proposed, and the report developed in an LMS. Section 4 reports the validation process of the model, i.e., the sample description, the study design, results and discussions. Finally, conclusions are presented in section 5.

---

2. AT-RISK STUDENTS

At-risk students are students who have a high probability of failing a course or withdrawing. It means that a student enrols in a course but he/she starts to present poor performance or does not appear to the classes. Some studies try to explain this behavior; addressing personal reasons (health, family or financial) while others describe issues related with the engagement to the course (lack of study-time, learning difficulties, low self-confidence). However, more than just understanding the reasons, it is important to identify such situation while students are still in class, so that teachers may intervene in order to increase the retention rate.

Previous studies report about the importance of students' background. According to Levy [8], from this data it is possible to compare one learner with others and predict what the profile of a non-completer student is. However, such findings might be only useful for the local neighborhood studied, i.e., as demographical data is being analysed, local context must be known and considered in generating the student's profile.

Early studies escribed the inefficiency of using only demographic data [8, 10, 19]. Billings [2] and Powell, Conway and Ross [12] used students' behavior to understand the difference between completers and non-completers. Moreover, students' awareness and attitudes may indicate their commitment to the course.

Another feature cited by many studies is the capacity of how accurately learners can estimate their required study time [6, 14, 18, 20]. This factor affects students' learning outcome directly, as a bad estimation of time does not allow them to read all course content and do all activities.

Lately, with the introduction of LMSs, more data has been used to understand the students' behavior better; and, consequently, it has made it possible to investigate the factors to identify at-risk students more precisely. According to Morris *et al*. [10], attendance and number of learning objects viewed are good references to distinguish students from being at-risk. Xenos *et al*. [20], Macfadyen and Dawson [9], and Er [3] investigated additional factors, such as average grade and number of assessments delivered.

More than only establishing that specific factors can identify at-risk students, some studies propose models to predict the risk based on demographics and behavior. Kotsiantis *et al*. [7] used seven demographic and four performance factors through a data experiment to generate a dropout-preventing model. Their results show an accuracy of 63% in early stages of the course using only demographic data and 83% in the middle of the course with all factors. In another study, Smith and Sweely [16] developed a model based on the frequency of logins and interaction with course materials, which resulted with a 70% accuracy in predicting unsuccessful students.

According to these studies, we noticed a gap between factors and models proposed with current LMSs. Few researchers have been presenting practical implementations that can be easily integrated in existing systems. Therefore, our study aims to build a model that can be applied to any LMS.

3. PROPOSED MODEL

This section describes the proposed model to integrate the functionality of identifying at-risk students in LMSs. The model is designed in such a way that it can be integrated easily into different LMSs with minimum required changes. Three steps were conducted for proposing this model. First, based on the literature reviewed as well as our requirement to make the model easy to use in different LMSs, the factors that can help in identifying at-risk students were selected. Second, depending on the identified factors, a decision tree based algorithm was designed. In the following subsections, these steps are described in more detail.

3.1. Factors Selected

According to the literature review, there are two types of factors that can help in identifying at-risk students namely, demographic and performance/behavior factors [1, 2, 8, 12, 19]. Demographic factors describe the students' profile and background to identify the students' probability to complete a course or not. It is an easy solution for programs offered in the same place, like blended courses. However, as our model aims to reach online and blended courses, demographic data does not work particular well in our case since students can be from anywhere in the world [19].

On the other hand, performance/behavior factors depend on the students' effort and course design, i.e., what learning objects are offered by the course and how students use them. Arnold and Pistilli [1] also suggested the consideration of course design, if courses are structured differently and place singular demands on learners.

Performance/behavior factors consider students' actions in a course (e.g., what he/she viewed, interacted or submitted) as well as their performance on activities/assignments. Some learning objects can only be viewed, such as text, image, video, forum of news, or assignment description. For others, students can also interact, like to submit an assignment, to send a new entry to database or glossary, and to post a message in a forum discussion. Those actions can then be evaluated using grades.

Based on student performance and behavior, we selected four factors to help in at-risk students' identification: percentage of activities done, average grades, percentage of resources viewed and attendance. The following subsections provide more details about each of these four factors.

3.1.1. Percentage of Activities Conducted

Activities are considered as any learning object that allows some interaction between students and the LMS, such as submitting assignments, posting a forum message, solving quizzes, talking in chatrooms, contributing in wikis, adding a term in glossaries, inserting a record in databases, etc. Billings [2], Xenos *et al*. [20], Macfadyen and Dawson [9], and Morris, Finnegan & Wu [10] used this factor in their works, which indicates that it is a known factor to identify at-risk students. It helps to check if students are committed with the course by showing his/her interest through active participation.

This factor only determines whether or not a certain activity has been performed, rather than counting how often a student engaged in this activity. For example, if a student has submitted two attempts in one quiz; we consider the student to be engaged in the quiz but do not count the attempts.

3.1.2. Average Grades

Grades refer to the value assigned by teachers to activities requested and delivered by students. Indeed, the value depends on the scale chosen by teachers (e.g., from 0 to 100 or using labels). Moreover, the teachers can indicate which activities will receive a grade, so that not all assignments have to be graded.

In the learning context, average grades are considered reliable information about students' performance [5]. It provides the teacher with an overview about the students' achievement in the course. Consequently, it is not just about whether a student has done the activity, but the quality of his/her work and the content comprehension. This factor has been frequently used, for example, by Macfadyen and Dawson [9] and Arnold and Pistilli [1] for designing an early at-risk students' warning systems.

The proposed model calculates the average grade, i.e., all grades are added and divided by the total of graded activities available, considering the respective weights of each activity. However, some observations are considered for this model. If a student did not deliver an activity before its deadline, he/she receives a grade of zero; and if a teacher has not provided the grade yet, this activity is not added to the total.

3.1.3. Percentage of Resources Viewed

Resources are any learning objects available in the LMS that do not require direct interaction between the student and the LMS, i.e., any text, image, video, etc. The number of resources viewed demonstrates the students' interest in the course. If he/she accesses the course's resources frequently, it indicates that he/she wants to learn about it. Previous studies have described that the total number of visits of resources may indicate students' interest [9], but others affirm that the time spent in one resource may diagnose the quality of learning [10]. However, the new generation of students seem to have new behaviors in the learning process. They are used to keep more than one browser's tab open at a time, accessing many contents at the same time [17], or sometimes they prefer to download the content once and never access the LMS again [19]. Consequently, how many times a student has accessed or how much time he/she has spent might be a misleading measurement.

Therefore, we only consider whether or not a resource has been visited. In other words, we assume that a student should access every resource at least once and it does not matter if he/she has accessed a resource more than once.

3.1.4. Attendance

Since many years, the importance of students having a regular time to study has been discussed [12], and how the lack of self-regulation of the time needed to their study affects course completion directly [14, 18, 20]. More recently, studies describe how weekly attendance helps in students' commitment with the course [3, 21] and how the changes in students' actions in LMSs can be a reliable indicator of failure [19].

In the proposed model, the attendance factor is using three sub-factors: weekly attendance, last access, and action flow. The sub-factor Weekly Attendance checks whether students accessed the LMS at least once a week. The sub-factor Last Access looks at whether a student has done an action in the last 20% of the course duration. For example, if the course takes five months and the student has not visited the LMS for more than one month, and the course is still ongoing, it shows that his/her attendance is very poor and he/she is at risk of failing the course. Finally, the sub-factor Action Flow measures how many actions the student did in the last 20% of the course duration, and compares this number to the previous period (e.g., assuming again a course with a duration of five months, we compare the number of actions in the last month with the number of actions in the month before).

Sometimes a student keeps visiting the LMS, but does fewer actions than he/she was used to do, which may indicate a decrease in commitment and again can indicate that this student is at risk. For this factor, we only consider actions related to learning (e.g., visiting a resource, submitting an assignment, posting a message, etc.). We do not consider actions that are not related to learning such as accessing courses' homepage and students' profile pages.

3.2. Algorithm

In order to identify the at-risk status of students based on the four factors (percentage of activities conducted, average grades, percentage of resources viewed and attendance), an algorithm was designed based on decision tree model. Wolff [19] affirmed that this technique is suitable for predicting at-risk conditions in early stages of the course while it also works until the end, once the model has more information to diagnose. The decision tree checks each factor in the sequence of its importance in the prediction.

As shown in Figure 1, the decision tree starts by checking if a course has activities and if so, what percentage of activities a student has done. Second, if the course includes graded activities and the teacher has provided the student's grade, it is checked. If a student has received low grades then he/she is at risk, but if he/she has high grades it means there is no risk. However, in the case of values in the medium range, the remaining two factors, percentage of resources viewed and attendance are checked.
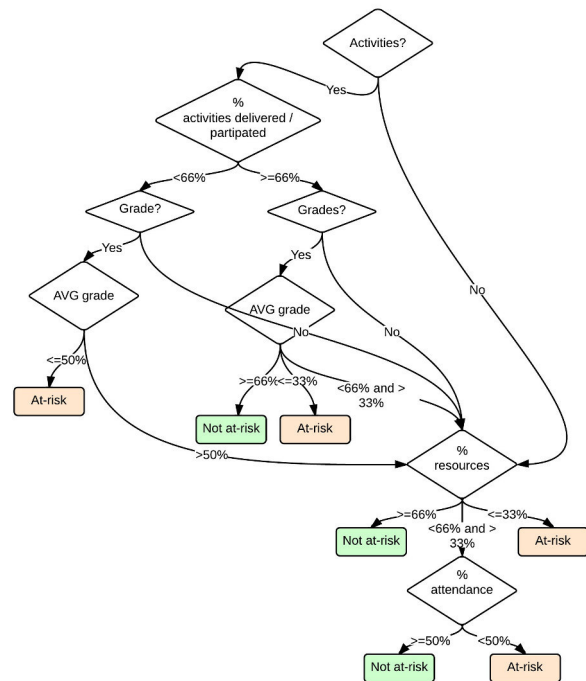

Figure 1. Proposed model

This decision tree was designed to be applicable in any LMS since the factors used are present in most LMSs.

3.3. Reporting tool

Based on the decision tree explained in the previous section, a reporting tool has been implemented. While the decision tree was general designed to work for different LMSs, the reporting tool has been implemented for the LMS Moodle. For this

implementation, the Moodle database was studied, and queries were developed to get discretized results that were needed for each of the four factors. The decision tree was converted to rules in order to be implemented in the reporting tool. Each time the reporting tool is executed, it uses current data about the students and suggests their at-risk status at the time the tool is executed. In addition, the reporting tool shows the students' start date, the percentage of activities done, the average grade, the percentage of resources viewed, and the attendance (Figure 2).

User	Course start date	Activities delivered	AVG grade	Resources viewed	Attendance	At-risk status
Ana Paula	2013-04-01	86%	85%	38%	84%	Not At-risk
Anderson	2013-04-01	4%	71%	62%	100%	Not At-risk
Andriel	2013-04-01	0%	-	2%	7%	At-risk
Betania	2013-04-08	64%	95%	52%	100%	Not At-risk
Carol	2013-04-01	21%	0%	2%	4%	At-risk
Caroline	2013-04-01	86%	100%	75%	74%	Not At-risk
Cristina	2013-04-01	25%	0%	12%	6%	At-risk
Custódia	2013-04-01	68%	85%	30%	70%	Not At-risk
Daiane	2013-04-01	86%	100%	58%	100%	Not At-risk
Daniela	2013-04-01	18%	0%	4%	3%	At-risk
Deisi	2013-04-03	86%	80%	25%	95%	Not At-risk
Delaine	2013-04-01	36%	80%	6%	11%	At-risk

Figure 2. Interface of reporting tool.

4. EVALUATION

The methodology followed to validate the proposed model is the post-fact technique, which uses data from past events to comprehend a phenomenon. In this case, the data from finished courses was analyzed including: (1) database copies from two different moments of the course (the middle and the end) and (2) reports containing students' outcomes in the course.

The sample of this study contains 88 students enrolled in nine undergraduate courses from one program. Students have 20 face-to-face classes offered weekly, but teachers may use up to three classes to deliver classes online using Moodle LMS. Moreover, the LMS is also used during the face-to-face classes to support the learning process (e.g., through exercises or readings), or the teacher may request some extra activities to be done at home.

4.1. Study design

The validation process occurred through three steps. First, we installed two Moodle instances, one with a data snapshot at the middle of the courses and one with a data snapshot of the completed courses. The Moodle version used was 2.2 with a MySQL database.

The second step was to match students by their names from the outcome reports to their profile in the Moodle database. Students are typically asked to create a Moodle account using their real name, but sometimes they omit part of their surname. Therefore, manual verification was required so that it was possible to determine whether students had completed each course successfully or failed.

The third step was to execute the reporting tool for each course in both Moodle instances. The reporting tool's results were compared to real students' outcomes, and were classified in one of three statuses:

- correct (a student successfully completed the course and he/she was identified as not at-risk; or the student failed the course and he/she was identified as at-risk);

- overestimated (a student successfully completed the course, but he/she was identified as at-risk);

- incorrect (a student failed the course, but he/she was identified as not at-risk).

4.2. Results and Discussions

According to Table 1, the overall results indicate that 85% of students were correctly identified, 13% were overestimated and 2% were incorrectly identified in both analyses. In addition, most of the courses got good individual results, except for course B that presented fewer accuracy results at the end of the courses. We did a deeper investigation into course B and found that the teacher used many group activities, i.e., when a group of students does the assignment but usually only one student submits the assignment in the LMS. Consequently, only one student delivers the task and the LMS records only points for this one student.

Table 1. Results in the middle and at the end of the course.

Courses	Middle			End		
	Correct	Overestimated	Incorrect	Correct	Overestimated	Incorrect
A	14	1	0	14	1	0
B	15	5	0	11	9	0
C	19	5	0	22	1	1
D	6	0	0	6	0	0
E	4	0	0	4	0	0
F	5	0	0	5	0	0
G	6	0	0	6	0	0
H	6	0	2	7	0	1
Total	**75**	**11**	**2**	**75**	**11**	**2**

Given that the model's goal is to identify at-risk students during the course rather than based on the full data (at the end of the course), we can consider that our proposed model presented good results in predicting at-risk conditions, with 85% of students being successfully identified, 13% of students were overestimated and only 2% were identified incorrectly in the middle of the courses. It should be mentioned here that identifying 100% of students correctly is very unlikely since a few students may change their behavior during the course, sometimes starting to put more effort into the course or dropout due to unforeseen events.

From this study, we found that the four factors used in the model to predict at-risk students seem well suited for the at-risk identification. The main one was the percentage of activities conducted. This factor is a good predictor, not only based on our results, but also according to other studies. For example, Macfadyen and Dawson [9] described similar findings with respect to this factor. On the other hand, grades contain accurate information, but it depends on having graded activities and on the teacher's effort to enter these grades in the LMS. Sometimes, courses offer assessments; however, a teacher can request a group activity, where not all students have their grades entered in the LMS. In addition, there are cases where a teacher creates many activities, but they are not mandatory. The other two factors are also essential, but they should be combined with the previous ones. Although, attendance is one of the most cited factor in

previous studies [3, 9], we found it important to check a student's grades and actions before.

5. CONCLUSIONS

This paper discussed how to integrate an at-risk student identification model to existing LMSs. In order to do this, first, the factors that can help in identifying at-risk students were studied from literature, and then based on this literature review, four factors (namely percentage of activities conducted, average grades, percentage of resources viewed and attendance) were selected. According to the literature review and in order to build a generic model, performance and behavior factors were selected over demographic factors that were used in past studies [1, 2, 8, 12, 19], because demographic profiles can depend on particular locations and cultures. Second, based on the identified factors, an algorithm based on a decision tree was designed and then implemented in Moodle as a reporting tool to visualize at risk levels of students to teachers. While the design of the algorithm is generic and can be applied to any LMS, the implementation is specific to Moodle.

This model was evaluated with a dataset of 88 students in nine undergraduate courses with a duration of 20 weeks. The results showed an accuracy of 85% in at-risk students' identification, which is higher than other similar works such as the studies conducted by Kotsiantis *et al*. [7] and Smith and Sweeley [16].

Despite this promising result, some limitations can affect the model's outcome, such as group assessments, lack of grades entered by teachers, learning objects that are not mandatory to access, different course designs that either display all the content or hide it in the begin of the course, etc. As future work, the last two can be solved by improvements in the model and can further enhance the accuracy of the model.

REFERENCES

[1] Arnold, K. E. & Pistilli, M. D. (2012) Course signals at Purdue: Using learning analytics to increase student success. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. ACM.

[2] Billings, D. M. (1987) Factors related to progress towards completion of correspondence courses in a baccalaureate nursing programme." Journal of Advanced Nursing 12.6, pp. 743-750.

[3] Er, E. (2012) Identifying At-Risk Students Using Machine Learning Techniques: A Case Study with IS 100. International Journal of Machine Learning and Computing. 2(4)

[4] Graf, S; Ives, C.; Rahman, N. & Ferri, A. (2011) AAT: A tool for accessing and analysing students' behaviour data in learning systems. Proceedings of the 1st International Conference on Learning Analytics and Knowledge, pp. 174-179

[5] Haig, T.; Falkner, K. & Falkner, N. (2013). Visualisation of Learning Management System Usage for Detecting Student Behaviour Patterns.

[6] Jensen, L. F. & Almeida, and O. C. S. (2009) A correlação entre falta de interatividade e evasão em cursos a distância. XV Congresso Internacional de Educação a Distância.

[7] Kotsiantis, S. B.; Pierrakeas, C. J. & Pintelas, P. E. (2003) Preventing student dropout in distance learning using

machine learning techniques. Knowledge-Based Intelligent Information and Engineering Systems. Springer Berlin Heidelberg.

[8] Levy, Y. (2007) Comparing Dropouts and Persistence in e-Learning Courses. Computers & education, v. 48, n. 2, pp. 185-204.

[9] Macfadyen, L. P. & Dawson, S. (2010) Mining LMS data to develop an "early warning system" for educators: A proof of concept. Computers & Education 54.2, pp. 588-599.

[10] Morris, L. V.; Finnegan, C. & Wu, S. (2005) Tracking student behavior, persistence, and achievement in online courses. The Internet and Higher Education 8.3, pp. 221-231.

[11] OECD (2013), Education at a Glance 2013: Highlights, OECD Publishing. ISBN 9789264201057

[12] Powell, R; Conway, C. & Ross, L. (1990) Effects of student predisposing characteristics on student success. The Journal of Distance Education/Revue de l'Éducation à Distance 5.1, pp. 5-19.

[13] Prinsloo, P.; Slade, S. & Galpin, F. (2012) Learning Analytics: challenges, paradoxes and opportunities for mega open distance learning institutions. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. ACM, pp. 130-133.

[14] Shin, N. & Kim, J. (1999) An exploration of learner progress and drop-out in Korea National Open University." Distance education 20.1, pp. 81-95.

[15] Slavin, R. E. & Madden, N. A. (1989). What works for students at risk: A research synthesis. Educational leadership, 46(5), 4-13.

[16] Smith, V. C. & Sweeley, D. (2010) Predictive Analytics: Increasing Online Student Engagement and Success. Retrieved from http://www.uregina.ca/orp/PapersPresentations/SCUP_Webnar_Sept29_2010.pdf

[17] Thompson, M. M. (1998) Distance learners in higher education. In C.C. Gibson (Ed.), Distance learners in higher education, Madison, WI: Atwood Publishing, pp. 9-24.

[18] Vergidis, D. & Panagiotakopoulos, C. (2002) Student Dropout at the Hellenic Open University: Evaluation of the graduate program, Studies in Education. The International Review of Research in Open and Distance Learning. v.3. n.2.

[19] Wolff, A.; Zdrahal, Z.; Nikolov, A.; & Pantucek, M. (2013). Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment. In Proceedings of the Third International Conference on Learning Analytics and Knowledge, pp. 145-149.

[20] Xenos, M.; Pierrakeas, C.; & Pintelas P. (2002) A survey on student dropout rates and dropout causes concerning the students in the Course of Informatics of the Hellenic Open University. Computers & Education 39.4, pp. 361-377.

[21] Zhang, H.; Almeroth, K.; Knight, A.; Bulger, M. & Mayer R. (2007) Moodog: tracking students online learning activities in Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, pp. 4415-4422.