“Earlier” Warning Systems: Making the Most out of the First Signs of Student Underperformance

Martín Liz-Domínguez
AtlanTTic Research Center
University of Vigo
Vigo, Spain
mliz@det.uvigo.es

Martín Llamas-Nistal
AtlanTTic Research Center
University of Vigo
Vigo, Spain
martin@det.uvigo.es

Manuel Caeiro-Rodríguez
AtlanTTic Research Center
University of Vigo
Vigo, Spain
mcaeiro@det.uvigo.es

Fernando Mikic-Fonte
AtlanTTic Research Center
University of Vigo
Vigo, Spain
mikic@det.uvigo.es

***Abstract*—This Research-to-Practice Work in Progress addresses the need for predictors in educational environments to make reliable predictions of student performance as early as possible. This is the main foundation of early warning systems, tools that aim to detect students at risk of failing or dropping out of a course, and do so at a point in time at which it is still possible to execute effective interventions to help these students. This study, carried out in the context of a first-year university course, presents a classifier of students depending on whether they are expecting to pass or fail the course, using exclusively data available before any assessment activities are performed. This classifier is based on the Random Forest algorithm. The obtained results, while limited by the scarcity of data, show promise that it is indeed possible to detect students at risk of failing with acceptable reliability in the studied educational context.**

***Index Terms*—blended learning, early warning systems, learning analytics, learning management systems, machine learning**

I. Introduction

Among the multiple sub-disciplines that exist within the field of learning analytics, the application of predictive algorithms has been a particularly interesting topic for researchers in recent years. In educational contexts, these algorithms are popular for their potential to predict the success rate of students in the courses or educational programs that they are enrolled in [1].

An important challenge that predictors in educational contexts need to face is the need for reliable predictions of student performance to be available as early as possible, so that teachers or academic advisors are able to perform effective interventions to help struggling students. To this end, researchers began to experiment with early warning systems (EWS) in the educational field. EWS were typically used in knowledge fields such as medicine or natural disaster prevention. Nowadays, many authors have developed their own EWS applied to education, each of them targeting specific academic contexts [2].

This paper has been partially funded by Xunta de Galicia (Galicia’s Singular Research Center, accreditation 2019-2022), the European Union (European Regional Development Fund - ERDF), and the Galician Regional Government under project ED431B 2020/33.
This work is partially financed by public funds granted by the Galician regional government, with the purpose of supporting research activities carried out by PhD students. (“Programa de axudas á etapa predoutoral da Xunta de Galicia — Consellería de Educación, Universidade e Formación Profesional”).

The present study focuses on the context of a first-year engineering course at University of Vigo, and aims to explore if it is possible to reliably predict students’ final success or failure based on data generated at the beginning of the course. Specifically, we are interested in working with data available before students need to perform any exam or task that would affect their course score. If the obtained predictions are trustworthy, instructors or student advisors could use them to offer specific help to students who need it early in the course, before they start struggling with tasks and exams, or experiencing disengagement from the course itself.

The rest of this paper is structured as follows: after this introduction, some applications of EWS in education that exist in the literature are introduced. Afterwards, the academic context which this study targets is described. Next, the experiment conditions are summarized, including the available data and the applied analysis procedure. Following this, analysis results are presented and discussed. Finally, some closing thoughts are provided, along with insights regarding future work.

II. Related work

The following are some notable examples of EWS applied in educational contexts:

- Arnold and Pistilli [3] are the authors of Course Signals, one of the earliest examples of an EWS applied in a higher education context. It is used at Purdue university to assess students’ risk of failure, defining three levels of increasing risk: green, yellow and red. Course Signals bases its estimations on data obtained from Purdue’s LMS, Blackboard Vista, as well as students’ prior academic history. Course Signals also provides features to communicate directly with students and provide feedback.
- Krumm *et al.* [4] created the Student Explorer EWS, which is similar to Course Signals in concept: it classifies students into a three-level risk scale by assessing LMS data. Specifically, Student Explorer relies on grades and student log-in events from the LMS.
- Plak *et al.* [5] performed a large-scale experiment at a Dutch university, including data from 758 first-year students belonging to twelve different bachelor programs. The developed EWS estimated a percentage chance for each student to drop out of the program, at regular

intervals within the academic year. Data used included student demographic characteristics, previous academic history and current program results, as these were made available. Analysis results were shown to student advisors via a dashboard.
- Bañeres *et al.* [6] presented a unique approach to an EWS targeting courses following a continuous assessment system. Using data from previous academic years, the EWS estimates the grade that each student should obtain in an upcoming continuous assessment task in order to have a favorable chance to pass the course. For subsequent tasks, the system calculates the necessary grade for each student as a function of the results of previous assignments. These estimations are shown to students as a form of feedback.
- Lee and Chung [7] developed an EWS to predict student dropout in the context of South Korean high schools. This study involves very large amounts of data, as the authors had access to information regarding about 165000 high schools and 2 million students. This study identified the most common reasons why South Korean students may drop out of high school and used attendance, behavior and course performance data to identify students at risk. The main challenge in this experiment was dealing with an imbalanced data set, as only around 2% of students dropped out each academic year.

III. Academic Context

This experiment was carried out using data from a first-year Telecommunications Engineering course at University of Vigo. This course is divided in two parts: theory and lab assignments. However, at the start of the semester, course workload focus almost exclusively on the theory part. As aforementioned, this study focuses on analyzing student activity at the very beginning of the course, and as such, data related to the theory contents corresponding to the first weeks will be considered in this paper.

In the theory part of this course, students evaluated via a continuous assessment model, performing short summative exams every two to three weeks covering course contents up to that point. Additionally, in terms of teaching methodology, this course follows the flipped classroom system: lectures are made available in the form of videos for students to watch at home, and weekly in-classroom sessions are used for solving practical problems, answer student questions and performing the summative exams [8]. The course uses the institutional LMS, based on Moodle, to manage educational resources.

Data used in this study corresponds to the 2020/2021 academic year, during which 177 students took part in the theory part of the course.

The first summative exam was scheduled to be performed at the start of the fourth week of the course. Thus, data from the first three weeks will be considered in this experiment. During this period of time, the following resources and activities were made available on the LMS for students to access:

- 18 video lectures covering the first contents of the course. Each one covered a specific sub-topic or practical problem, and their duration ranged between about two minutes for the shortest ones, and ten minutes for the longest.
- Two optional self-assessment questionnaires including exam-like problems for students to practice on.

The compiled data set, which is described in detail in Section IV-A, takes into account the use that students made of these resources during the studied time period.

IV. Experiment Description

A. Available Data

The composed data set has one row per enrolled student, with a series of features associated to each one of them. Table I summarizes the available features and their meanings.

TABLE I
PERMUTATION IMPORTANCE OF FEATURES IN THE DATA SET

Feature	Description	Value range
Videos	Number of available videos which the student accessed at least once.	$[0, 18]$
Self-assessment 1	Whether the student completed the first self-assessment test.	[true, false]
Self-assessment 2	Whether the student completed the second self-assessment test.	[true, false]
Enrollments	Number of times that the student has been enrolled in the subject in previous academic years.	$[0, 5]$
Sessions	Number of learning sessions that the student carried out on the institutional LMS.	$[0, 123]$

The *Videos* and *Self-assessment* features are directly related to students' use of course resources at their disposal during the first three weeks of the course, as explained in Section III.

Additionally, the number of enrollments in the course during previous academic years is also considered, as it was observed that the fact that students may need to retake the same course multiple times can have an important impact over the interpretation of LMS traces [9].

Finally, the *Sessions* parameter represents the number of learning sessions registered by each student on the LMS. The concept of "learning session" is borrowed from Jovanovic *et al.* [10], and represents a chain of actions that a student performs consecutively in the LMS. For example, a student may log into the platform, watch some of the videos and perform a self-assessment test in the same session. In this study, consecutive events by a student are considered part of the same session if the difference between their observed timestamps is no bigger than 50 minutes. This time window was carefully chosen: if it is too big, unrelated events will be considered as part of the same session, but if it is too small, events that clearly belong to the same learning sequence may be split up into different sessions. The 50 minute cutoff time was observed to work well in the analyzed scenario, as it minimized instances of the two aforementioned problems [11].

B. Analysis Procedure

The goal of this experiment is classifying students regarding their course outcome. The following two scenarios will be tested:

- Classification of students into three categories: *Pass*, *Fail* or *No-show*.
- Binary classification of students into *Pass* or *Fail* categories. Here, the *Fail* category would also include students who were labeled as *No-show* in the previous experiment.

Naturally, the *Pass* category represents students who passed the course, and the *Fail* category those who did not. A student belongs to the *No-show* category if they did not attend any of the exams throughout the course.

The actual labels observed for the 177 students in the data set were distributed as follows:

- 42 students belonged to the *Pass* category.
- 83 students belonged to the *Fail* category.
- 52 students belonged to the *No-show* category.

Regarding the predictive algorithm that was used, we decided to apply the Random Forest (RF) algorithm. This choice was based on the fact that RF is able to appropriately and reliable handle data sets with multiple types of features, in this case, a combination of integers and booleans [12].

Before applying the predictive algorithm, the data set was divided into training and test splits, assigning 75% and 25% of samples to them respectively. The RF model was fitted using the training set, additionally running five-fold cross-validation, which was preferred over the definition of an additional validation split due to the limited size of the data set.

V. Results

A. Three-label classification

In the first defined scenario, where three different labels are considered, a cross-validation accuracy value of 0.613 was achieved. On the other hand, the proportion of correct label assignments on the test set was 0.578. Figure 1 displays the confusion matrix for predictions performed over the test set in this scenario.

Observing the confusion matrix, it can be observed that 71% of students in the *Fail* category are classified correctly, compared to just 50% of those in the *Pass* category. Another interesting aspect is that, out of the incorrectly classified instances of *Fail* or *Pass* students, almost none of them were labeled as *No-show*.

On the other hand, only 7 out of the 15 instances of *No-show* students in the test set were classified correctly, with most mislabeled instances having been assigned to the *Fail* category.

It is worth pointing out that the quality of this classified might be hindered by the fact that the data is fairly imbalanced, with instances of the *Fail* category significantly outnumbering any of the other two groups.

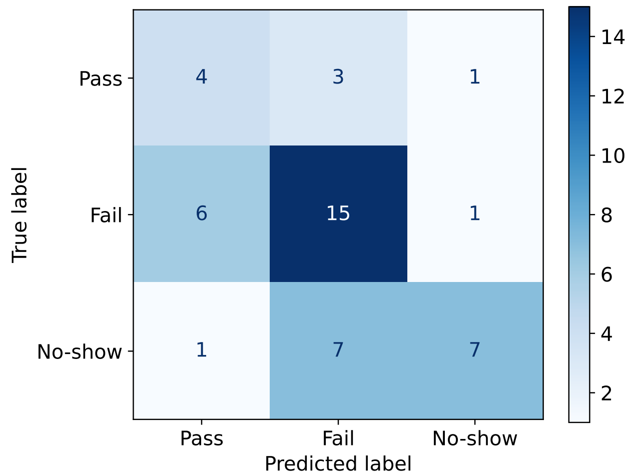

Fig. 1. Confusion matrix corresponding to the three-label classification scenario.

B. Binary classification

In the second scenario, instances of *No-show* students were merged into the *Fail* category to convert this problem into one of binary classification. This new RF classifier achieved a cross-validation accuracy value of 0.758 and a test set classification accuracy of 0.778, a significant improvement over what was observed in the previous experiment. Figure 2 represents the confusion matrix for test set labeling in this scenario.

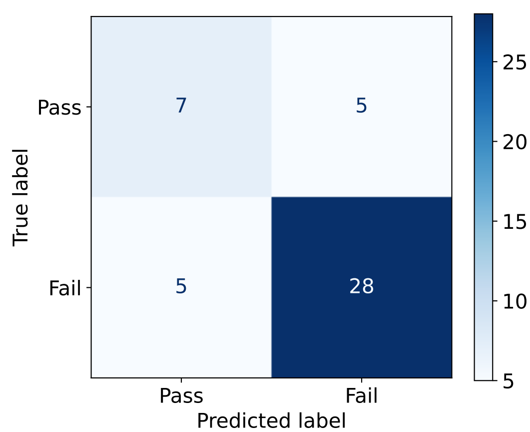

Fig. 2. Confusion matrix corresponding to the binary classification scenario.

As in the previous case, imbalanced data may hurt the performance of this classifier. In the binary classification scenario, more than 75% of students in the data set belong to the *Fail* category. As a consequence of this, the classifier struggles at correctly labeling students in the *Pass* category, resulting in almost half the instances in this group to be misclassified.

C. Feature relevance

In addition to assessing the performance of RF classifiers, we were interested in knowing which of the features were the most important for predicting students' course outcomes. In order to do this, the relevance of each feature was computed using the permutation importance algorithm. This metric was chosen over the decrease in Gini index (also known as decrease in impurity) because, even though the latter is very commonly used for assessing RF classifiers, it tends to be biased towards numeric features in data sets that include variables of different types, like the one that is used in this experiment [13].

Table II summarizes the permutation important obtained for each feature in both analyzed scenarios: classifier 1 being three-label classification, and classifier 2 being binary classification.

TABLE II
PERMUTATION IMPORTANCE OF FEATURES IN THE DATA SET

Feature	Classifier 1	Classifier 2
Videos	0.1876	0.1754
Self-assessment 1	0.0599	0.0095
Self-assessment 2	0.0859	0.0628
Enrollments	0.0734	0.1442
Sessions	0.3305	0.1930

In the three-label classification scenario, the number of sessions is by far the most relevant feature, followed by the number of videos watched. The other three parameters are comparatively much less relevant.

On the other hand, in the binary classification scenario, the number of enrollments is much more relevant than in the previous case, being almost on par with the number of videos and sessions. The relevance of the number of sessions is significantly lower than in the previous scenario, although it is still the most relevant feature out of the five present in the data set. Completion of self-assessment tests is still relatively irrelevant compared to the rest of the features.

The big changes in significant values for the number of enrollments and number of sessions between both scenarios can be explained by the removal of the *No-show* label as an independent category. Participating in a low number of LMS sessions, or in some cases not participating in any at all, is the most important telltale sign that a student may drop out of the course, regardless of whether they were enrolled in the subject for the first time or they had been enrolled in a previous academic year. Alternatively, if only *Pass* and *Fail* categories are considered, the number of enrollments becomes a more important feature. It was observed that students who had already been enrolled in the past often have a head start compared to those enrolled for the first time, since the former may retain some knowledge of the subject from previous attempts, even if they failed the course [9].

VI. CONCLUSION AND FUTURE WORK

The study presented in this paper aimed to prove the feasibility of student classification regarding whether they are expected to pass or fail a course, in a higher education scenario applying the flipped classroom and continuous assessment methodologies. While exclusively using data available before any assessment activities were carried out in the course, two different Random Forest classifiers were built: one that considered no-show students as an independent category, and another one that merged them together with the failing students. The obtained accuracy values for test sets in both classifiers were 0.578 and 0.778, which we consider acceptable considering the restrictions of the study and the existing potential for improvement.

The most important limitation of this experiment was the lack of abundant data. Only 177 students from the 2020/2021 academic year were considered. Future iterations of this study will feature data from students belonging to other academic years in order to increase the number of records available to the classification algorithm.

An additional objective would be adding more features to the data set. For example, as part of a parallel study, students were asked to periodically answer short questionnaires regarding different types of self-regulated learning habits [14]. Answers from questionnaires performed at the beginning of the course could be processed and incorporated as new features in this data set.

Finally, further testing of different machine learning algorithms would be beneficial for comparing the performance of different classifiers. The RF classifier was chosen due to its ability to handle features of different types, as well as having acceptable performance when working with imbalanced data sets. Nevertheless, it would be interesting to repeat the experiment using some of the other existing classification algorithms, such as SVM or kNN.

REFERENCES

[1] R. Jindal and M. D. Borah, "A Survey on Educational Data Mining and Research Trends," *International Journal of Database Management Systems*, vol. 5, no. 3, pp. 53–73, jun 2013. [Online]. Available: http://www.airccse.org/journal/ijdms/papers/5313ijdms04.pdf

[2] M. Liz-Domínguez, M. Caeiro-Rodríguez, M. Llamas-Nistal, and F. A. Mikic-Fonte, "Systematic Literature Review of Predictive Analysis Tools in Higher Education," *Applied Sciences*, vol. 9, no. 24, p. 5569, dec 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/24/5569

[3] K. E. Arnold and M. D. Pistilli, "Course signals at Purdue," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, ser. LAK '12. New York, NY, USA: ACM, apr 2012, pp. 267–270. [Online]. Available: https://dl.acm.org/doi/10.1145/2330601.2330666

[4] A. E. Krumm, R. J. Waddington, S. D. Teasley, and S. Lonn, "A Learning Management System-Based Early Warning System for Academic Advising in Undergraduate Engineering," in *Learning Analytics*. New York, NY: Springer New York, 2014, pp. 103–119. [Online]. Available: http://link.springer.com/10.1007/978-1-4614-3305-7_6

[5] S. Plak, I. Cornelisz, M. Meeter, and C. Klaveren, "Early warning systems for more effective student counselling in higher education: Evidence from a Dutch field experiment," *Higher Education Quarterly*, vol. 76, no. 1, pp. 131–152, jan 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/hequ.12298

[6] D. Bañeres, M. E. Rodríguez, A. E. Guerrero-Roldán, and A. Karadeniz, "An Early Warning System to Detect At-Risk Students in Online Higher Education," *Applied Sciences*, vol. 10, no. 13, p. 4427, jun 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/13/4427

[7] S. Lee and J. Y. Chung, "The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction," *Applied Sciences*, vol. 9, no. 15, p. 3093, jul 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/15/3093

[8] M. Llamas-Nistal, F. A. Mikic-Fonte, M. Caeiro-Rodriguez, and M. Liz-Dominguez, "Supporting Intensive Continuous Assessment With BeA in a Flipped Classroom Experience," *IEEE Access*, vol. 7, pp. 150 022–150 036, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8865067/

[9] M. Liz-Dominguez, M. Llamas-Nistal, M. Caeiro-Rodriguez, and F. Mikic-Fonte, "Lms logs and student performance: The influence of retaking a course," in *2022 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, March 2022, pp. 1970–1974. [Online]. Available: https://ieeexplore.ieee.org/document/9766691/

[10] J. Jovanović, D. Gašević, S. Dawson, A. Pardo, and N. Mirriahi, "Learning analytics to unveil learning strategies in a flipped classroom," *The Internet and Higher Education*, vol. 33, pp. 74–85, apr 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1096751617300684

[11] M. Liz-Dominguez, M. Llamas-Nistal, M. Caeiro-Rodriguez, and F. A. Mikic-Fonte, "Profiling students' self-regulation with learning analytics: A proof of concept," *IEEE Access*, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9812587/

[12] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, oct 2001. [Online]. Available: https://link.springer.com/article/10.1023/A:1010933404324

[13] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, may 2010. [Online]. Available: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq134

[14] M. Liz-Domínguez, M. Caeiro-Rodríguez, M. Llamas-Nistal, and F. Mikic-Fonte, "Monitoring students' self-regulation as a basis for an early warning system," in *Learning Analytics Summer Institute Spain 2021: Learning analytics in times of COVID-19: Opportunity from crisis*, vol. 3029. Barcelona, Spain: CEUR-WS, 2021, pp. 38–51. [Online]. Available: http://ceur-ws.org/Vol-3029/paper04.pdf