Predicting early dropout students is a matter of checking completed quizzes: the case of an online statistics module

Josep Figueroa-Cañas [0000-0002-6790-9142] and Teresa Sancho-Vinuesa [0000-0002-0642-2912]

Universitat Oberta de Catalunya, Barcelona, Spain
[jfigueroa, tsancho]@uoc.edu

Abstract. Higher education students who either do not complete the subjects they enrolled in or interrupt indefinitely their studies without certification, the so-called college dropout problem, still continues to be a major concern for practitioners and researchers. Within the subjects, an early prediction of dropout students has aided teachers to focus their intervention in order to reduce dropout rates. Several machine-learning techniques have been used to classify/predict dropout students, including the tree-based methods which are not the best performers, but in their favour, are easily interpretable. This study presents a procedure to identify dropout-prone students at an early stage in an online statistics module, based on decision tree models. Although the attributes initially considered in the creation of the trees were mainly related to quiz completion, participation in the forum and access to the bulletin board, the final models show that the former is the only attribute with significant discriminatory power. We have evaluated the classification performance by means of a validation set. The performance measure of accuracy shows values above 90%, whereas that of recall and precision slightly under 90%.

Keywords: Dropout prediction, decision trees, quiz completion, online education.

1 Introduction

Among education practitioners and reseachers, students who do not complete a single module/subject or indefinitely interrupt their studies without having achieved the certificate have been a matter of considerable concern for a long time. These students are usually called dropout students. In online courses, the high dropout rates of students justify the abundant research on this particular topic, as shown in the extensive review of [1], where 159 studies published between 1999 and 2009 were analysed. More recently, in the European framework, reducing the dropout student rate in higher education is considered a key strategy to attain the ambitious objective of not less than 40% of people in their thirties who have completed higher education studies by 2020 [2]. Concerned as teachers and guided by European strategy, the authors have decided to carry out research on dropout students in the statistics module at the Universitat Oberta de Catalunya.

In a higher education context, two levels of dropout can be differentiated: (a) the micro-level dropout, and (b) the macro-level one. In the former, the fact of dropout takes place inside the module or subject [3], where teachers can intervene in case they have convenient information at an early stage in order to reduce it. In line with that, Burgos [4] shows a reduction of 14% in dropout rates of students by means of a tutoring plan action after the dropout-prone students have been identified early. In macro-level dropout, withdrawal from studies occurs, in general, outside the subjects so that the interventions are the responsibility of other staff different from the teachers of the subject.

The main purpose of the present study is to design a procedure to identify as many dropout-prone students as possible in an online statistics module, as soon as possible. This procedure is based on the prediction/classification provided by binary conditional decision trees generated in several instants of time throughout the module duration, from the data related, mainly to test completion and participation in both the online forum and the bulletin board.

2 Literature review

According to [1], there is an absence of consensus on the definition of both the micro-level dropout and the macro-level one. With regard to the latter, even online and face-to-face universities do not share the dropout definition [5]. Grau-Valldosera [5] claims the time accepted without any enrolled subjects in an online university has to be extended compared with that in a face-to-face university because of the students' characteristics.

As illustrations of the micro-level dropout definitions, we have chosen the three that follow. First, Liu [6] straightforwardly associates subject dropout with subject failure. Dropout students are those who do not attain A, B, or C, that is, those who fail the subject. Second, Levy [7] defines dropout students as those who do not complete the subject and their tuition fees have not been refunded. And third, Dupin [8] considers dropout students as those who are non-completers, understood in a broad sense.

The studies about dropout students by Cohen [3], Burgos [4], Costa [9], Santana [10], Lykourentzou [11], Lara [12] and Kotsiantis [13] are focused on the micro level (university subjects), all in an online but [3] blended environment. In addition, all of them are concerned with early prediction and show considerable high values of several evaluation measures of classification performance, such as accuracy, recall, precision or F1-measure. Cohen [3] reports a maximum precision of 80%, Burgos [4] a recall of 96.73%, Costa [9] a maximum F1-measure of 82%, Santana [10] a maximum accuracy of 86%, Lykourentzou [11] a maximum recall of 95% and Kotsiantis [13] a maximum accuracy of 83.89%. Lara [12] found an accuracy above 90%, a figure that is "a very acceptable percentage for the problem domain" [12, pp. 31]. In the following four paragraphs, we present a comparative review between [3-4, 9-14] regarding dropout definition, single/multiple predicting instants of time, attributes selected as predictors and classification method to carry out the prediction.

The dropout definition from the failure perspective [6] is the one used in the studies of Cohen [3], Costa [9] and Santana [10]. The definition of Levy [7] is explicitly mentioned in Lykourentzou [11], who adds another requirement: that the dropout student has to access the e-learning platform at least once throughout the subject duration. That means the student has to leave a trace in the information system before leaving the subject in order to be considered a dropout student. For Burgos [4] and Lara [12] students who do not sit the final exam are those defined as dropout students. And finally, Kotsiantis [13] does not precisely define the non-completer students.

Predicting in a single instant of time is the option chosen by Santana [10] and Kotsiantis [13]. The latter argues that prediction has to be released before the subject is half over because otherwise it would not be useful for the teachers to intervene in time. Santana [10] predicts dropouts after the first exam, which also coincides with half of the subject duration. In contrast, multiple instants of time, albeit not the same ones, are contained in the proposals of [3-4, 9, 11-12]. Lykourentzou [11] released predictions into each of the 7 sections that the subject is divided into. Similarly, Burgos [4] predicts in each of the 12 assessment activities. The proposals of [3,9,12], based mainly on regular time intervals, are slightly different: Cohen [3] predicts dropouts monthly, in a one semester course, Lara [9] weekly in 15-20 week courses, and finally Costa [9] also weekly in a 10-week course and after releasing the mid-course exam marks.

All the attributes employed in [3-4, 9-13] can be grouped into three main categories: demographics, usage of educational tools, and assessment activities or exam performance. The first category is formed by time-invariant data available at the beginning of the course, whereas the other two categories include time-varying data which are incrementally collected throughout the course. Demographic attributes such as gender and professional information are used by [9-11, 13] alike. Some studies also consider other specific demographic attributes, like English language literacy [13]. The usage of educational tools in general, and particularly participation in the forum is included in the set of attributes that form the models of Cohen [3], Costa [9], Santana [10], Lykourentzou [11] and Lara [12]. Finally, the marks attained in assessment activities or exams are analysed in the studies of Burgos [4], Costa [9], Santana [10], Lykourentzou [11] and Kotsiantis [13].

Regarding classification methods, apart from Cohen [3] who uses a unique method based on comparing changes in attribute values of a student with respect to the mean of attribute values of the whole group of students, the studies of [4, 9-13] use a great variety of machine-learning techniques. Algorithms based on neural networks and support vector machines are common to [4, 9-13], whereas naive Bayes and decision tree classifiers are only employed by Costa [9], Santana [10] and Kotsiantis [13]. Finally, logistic regression is also included in the set of classifiers of Burgos [4], Lara [12] and Kotsiantis [13].

Although the study of Romero [14] does not explicitly mention the dropout problem, as it aims to predict the final performance of students by classing them as passed or failed, it could be deemed as a dropout problem according to Liu's definition [6]. Moreover, like some of the references previously reviewed, an early prediction is released, and the usage of the forum is the source of information to feed the attributes. The study stands out for the comparative performance of 14 classification algorithms and reaches the conclusion that the sequential minimal optimization (SMO) algorithm, related with support vector machines, is the better performer. It is worth recalling that the studies of [3-4, 9-13] all included that machine-learning technique.

The high dropout rates are also a major source of concern in Massive Open Online Courses [15] and, in order to reduce them, several studies have dealt with their early prediction [15-17]. These studies differ both in the type of dependent variables and the machine learning methods used in their models. First, whereas the studies by Ruiperez-Valiente [15] and Sharma [16] include the scores awarded after assignment submission, Yang [17]'s only takes into account the behaviour in the discussion forum. And second, prediction algorithms based on artificial neural networks are the ones chosen by Sharma [16], while Ruiperez-Valiente [15] implemented random forests, generalised boosted regression modelling, K-nearest neighbours and a logistic regression, and Yang [17] used a survival model. Sharma [16] finds a relationship between students failing in assignments and dropping out of the course.

3 Methodology

3.1 Participants and learning context

The participants in this study were the 197 students enrolled in the first semester of the 2018/19 fully asynchronous online one-semester statistics module, which formed part of the Computer Engineering degree at the Universitat Oberta de Catalunya.

The teaching plan for this statistics module allowed students to complete optional quizzes (Quizzes) and constructed-response questions (R.Questions) that had to be solved by using the statistical program R. Six different pairs (Quiz, Rquestion), named continuous assessment tests, were scheduled throughout the semester. Quizzes were corrected and marked immediately, providing automated feedback. R.Questions required manual teacher correction and feedback was delayed. The scores attained, which formed part of the continuous assessment mark, could be included in the final mark. The module included two assessment instruments: (a) a compulsory in-person final exam, and (b) non-compulsory online continuous assessment throughout the semester. The final mark for the module was mainly based on the final exam mark, which could be modified slightly by the continuous assessment mark. In addition, during the first week teachers assigned an initial test to ascertain students' prior knowledge of secondary-education statistics. In order to encourage participation, students who voluntarily completed and submitted the test obtained a bonus, which also formed part of the continuous assessment mark.

An e-learning platform provides students enrolled in the statistics module of the Universitat Oberta de Catalunya with a communication tool: a forum, and an information tool: a bulletin board. The latter was used by teachers to upload course information which was mostly only accessible by students via that bulletin board. The former allowed students and teachers to interact with each other, in general, asynchronously. The e-learning platform also included direct access to view the teaching plan, which contained precise information about the assessment system. All reading access to the bulletin board, forum and teaching plan, as well as writing access to the forum were recorded by the information system of the Universitat Oberta de Catalunya.

3.2 Measure and data collection

The data has been collected in four instants of time, which coincide with the first four continuous assessment test submission deadlines, the only ones in the first half of the course. The separation between submission deadlines is variable, ranging from 1 to 3 weeks. We define four periods of time (Period.1, ..., Period.4) from the previous submission deadline as follows: Period.1 is the interval of time between the first day of the semester and the first submission deadline, Period.2 is the interval of time between the first and second submission deadlines, and so on for Period.3 and Period.4.

During the first period (Period.1), we gathered students' register data such as the number of courses enrolled on in the semester and whether they were repeater students or not . This data, contained in the information system of the Universitat Oberta de Catalunya and anonymously delivered to us, filled the instances of the attributes Repeating and Enrolled_Courses (see Table.1). The Moodle activity log was the source of information to determine whether the student had submitted the initial test or not, and likewise the first continuous assessment test. With that data, the instances of the attributes Initial_Test, Quiz_Till_Period.1 and R.Question_Till_Period.1 were filled (see Table.1). The e-learning platform activity log provided the date and time of all access to the platform which, after being pre-processed, filled the instances of the attributes BBoard_Till_Period.1, Forum_Wr_Till_Period.1, Forum_Re_Till_Period.1 and Teaching_Plan_Viewed_Till_Period.1 (see Table.1). All the previous data, transferred to the second period (Period.2) and incremented with the specific information collected in Period.2, filled the attributes ending in _Till_Period.2. This procedure was repeated for Period.3 and Period.4 (see Table.1)

Table 1. Attributes for the Period.i, with i=1,..., 4

Name	Description	Types and Vaules
Repeating	Indicates whether the student is repeating the subject or not	Type: Boolean. Values: I.RP, N.RP
Enrolled _Courses	Indicates the total of courses enrolled on in the semester.	Type: Integer Values: {1, ...}
Initial_Test	Indicates whether the student has or has not completed and submitted the initial test.	Type: Boolean. Values: H.IT, N.IT
Teaching_Plan_Viewed	Indicates whether the student has or has not viewed the teaching plan until the last day of the Period.i	Type: Boolean. Values: H.TPV, N.TPV
Quiz_Till_Period.i	Indicates the number of quizzes completed and submitted until the last day of the Period.i	Type: Integer Values: {0, 1, ..., i}

R.Question_Till_Period.i	Indicates the number of quizzes completed and submitted until the last day of the Period.i	Type: Integer Values: {0, 1, ..., i}
BBoard_Till_Period.i	Indicates the number of periods in which the student has accessed the board until the last day of the Period.i	Type: Integer Values: {0, 1, ..., i}
Forum_Wr_Till_Period.i	Indicates the number of periods in which the student has written messages on the forum until the last day of the Period.i	Type: Integer Values: {0, 1, ..., i}
Forum_Re_Till_Period.i	Indicates the number of periods in which the student has read messages on the forum until the last day of the Period.i	Type: Integer Values: {0, 1, ..., i}

The attribute selection of our study is based on the references in section Literature review [4,9]. Nevertheless, we have not considered the scores of assessment activities as attributes as [4] does. Instead, we have opted for the completion or non-completion of Quizzes and R.Questions. There are two main reasons for this decision. The first reason is Quizzes and R.Questions submission data are available faster than definite marks since both R.Questions are marked manually (as mentioned in section 3.1), and students may apply for marking reviews. The second reason is the likely high correlation between completion of assessment activity and its mark, as is shown in a calculus module of the same degree and in a very similar educational context [18].

In the present study, we have defined, based on [7], a dropout student as the student who attains a final mark of “Not Completed”, which means the student has not taken the compulsory final exam. That approach is in line with that of the [4, 12]. The boolean response variable Y=Dropout indicates whether the student complies (I.Dropout) or not (I.Completer) with the previous definition, that is, whether they belong to the dropout student or to the completer student class. To fill the instances of that variable, the information system of the Universitat Oberta de Catalunya has anonymously delivered the final marks to us. By combining attributes, like predictors, and response variable Y, four sets of data are available, and each of them contains the instances of attributes of each period and the instances of the Dropout variable.

3.3 Classification method

We pose a classification problem, the result of which will be a binary classification model or binary classifier in order to predict whether a student will be classed as a dropout student or completer student at the end of the semester. In addition, we require the classifier to be easily interpretable, although at the expense of it not being the best performer in terms of the usual evaluation measures of classification performance like accuracy, precision or recall. Due to "tree-based methods being simple and useful for interpretation " [19, pp. 303], we have decided to use those methods of classification in our study. Basically, a binary decision tree is an oriented graph that starts in a node called root, follows through arcs called branches, and ends in the terminal nodes, called leaves. Each nonterminal node, including the root, represents an attribute (a test on the attribute), and each leaf represents one of the two classes (dropout student or completer student) or the proportion of students that belong to each class. The branches that come out of a node represent the values of the attribute associated with the node (the answer to the test on the attribute) [20].

Given our attribute selection (see Table.1), we observe that not all the attributes have the same number of possible values. A widely identified issue detected in studies using decision tree models is the bias, in creating the nodes, to attributes with a large number of possible values [21]. Conditional tree models mitigate that bias [21], and for that reason those models are the classification methods we have chosen. To grow our conditional trees, we have used the ctree() function provided by the statistical program R.

For each of the four data sets a classification model has been built (Model.1, Model.2, Model.3, Model.4). In order to evaluate the performance of the models, firstly the whole data set can be split into two mutually exclusive sets: the training set and the validation one. Secondly, with the training set the classification model is fitted. And finally, the evaluation of the performance is carried out using the validation set [19]. In our study, we have conducted a random stratified split into a training set (80% of the whole set) and a validation set (20%), keeping the same class distribution of the whole set in each subset.

Taking into account that our main purpose was to identify dropout-prone students, we have considered students predicted as dropouts, that is, those whose predicted class is I.Dropout, as “Positive” cases, and the others, those whose predicted class is I.Completer, as “Negative” cases. Moreover, as usual, we differentiate between “True” or “False” depending on whether the predicted class coincides with the observed class or not, respectively. Table.2 depicts the four possible pairs when applying the validation set to the model fitted with the training set.

Table 2. Possible pairs in terms of predicted and observed classes

	Predicted class I.Dropout	**Predicted class I.Completer**
Observed class I.Dropout	True Positive (TP)	False Negative (FN)
Observed class I.Completer	False Positive (FP)	True Negative (TN)

In our study we have decided to use three evaluation measures of the classification performance: Accuracy (1), Precision (2) and Recall (3), according to the following definitions [19] :

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

4 Results

The four classifiers created from the training set are extremely simple, each of them contains one single node. Table.3 depicts each model as a decision rule. The only attribute shown in the models, a result that reveals that it is the one with the strongest association with the response Dropout [20], is the completion of Quizzes (Quiz_Till_Period.i). Using the first three models (Model.1, Model.2, Model.3), students are classified/predicted as dropout students (class I.Dropout) if they have not completed all the Quizzes scheduled until the end of the period associated with the model, in other words, if they have not completed one or more of those Quizzes. As an example, at the end of Period.3, students that has not completed all three Quizzes corresponding to the first three continuous assessment tests are classified/predicted as a dropout students. Only students who have completed all three Quizzes are classified/predicted as completer students (class I.Completer). In Model.4, the last condition is softened, so that students are classified/predicted as completer students even if they have not completed all four Quizzes. They can have decided to skip one Quiz, at the most.

Table 3. Decision rules for the four models: Model.1, Model.2, Model.3, Model.4

Model.1*:	**Model.2***:

IF Quiz_Till_Period.1=1
THEN I.Completer
ELSE I.Dropout
	IF Quiz_Till_Period.2=2
	THEN I.Completer
	ELSE I.Dropout

Model.3*:	**Model.4***:

IF Quiz_Till_Period.3=3
THEN I.Completer
ELSE I.Dropout
	IF Quiz_Till_Period.4 >2
	THEN I.Completer
	ELSE I.Dropout

* p-value < 0.001. H0: D(Dropout | Quiz_Till_Period.i) = D(Dropout), that is, H0: The response Dropout is independent of the predictor Quiz_Till_Period.i

Using the validation test, three results stand out in the evaluation measures of the classification performance (see Table.4). First, Accuracy shows a gradual increase from the first model and, in Model.3 passes the figure of 90%, which is considered acceptable by [12]. Second, Precision also grows from the second model and reaches the value 100% in Model.4. And third, Recall also rises from the first model and reaches its highest value in Model.3, just when Accuracy attains the level of “acceptable”.

Table 4. Performance measures

	Recall	Precision	Accuracy
Model.1	53.8%	87.5%	82.9%
Model.2	61.5%	80.0%	82.9%
Model.3	84.6%	84.6%	90.2%
Model.4	84.6%	100%	95.1%

5 Discussion

From the very beginning we aimed to find a classification model that was easily interpretable, even at the expense of not finding the best performer classifier. The four classification models (see Table.3) entirely comply with the previous requirement. In the rest of the section, we discuss the following three statements: (a) completing evaluative quizzes is the only attribute that determines the classification process, (b) the simplicity of the models eases the creation of an overall classification procedure that includes all the models, and (c) applying the models separately, Model.3 is the best.

Above all, it is worth noticing that only one attribute, the Quiz_Till_Period.i, intervenes in the classification process as the four models show (see Table.3). The Quiz_Till_Period.i attribute, directly related with the completion of Quizzes, has basically an evaluative character, which sets it apart from the attributes related to the usage of the e-learning platform, such as the forum. The dominance of evaluative attributes is in line with the study of Costa [9], who found that the most important attribute was the midterm marks. On the other hand, completion of R.Questions likewise has evaluative character, but nonetheless does not intervene in the final models. The main difference between Quizzes and R.Questions lies in that the latter require students to apply higher level skills than the former. Consequently, it seems reasonable to argue that students who do not even complete the least-demanding assessment assignments, such as Quizzes, are the most prone to becoming dropout students. And last but not least, the three first models separate dropout and completer students depending on whether they have or have not completed all the Quizzes scheduled until the moment the model is applied. Therefore, we can interpret that the continued “doing” of Quizzes is the relevant aspect in differentiating those who complete the course from those who do not.

The simplicity of the model reduces the volume of information actually being used to only that related to completion of Quizzes, which in turn entails two beneficial consequences: (a) the obvious elimination of time spent gathering and processing the rest of attributes, (b) the teacher himself/herself can collect the required data directly from the Moodle activity log. Using the three first models in cascade, at the end of the first continuous assessment test submission deadline, the teacher can create a list of dropout-prone students by selecting those who have not completed the first Quiz. After the second submission deadline, the teacher can add new dropout-prone students to the previous list by selecting those who have not completed the second Quiz, and likewise regarding those who have not complete the third one. So, by following that simple procedure the teacher step by step adds to the list of dropout-prone students, which can be useful when deciding possible measures in order to change the unsuccessful predicted result.

The performance measures (see Table.4) indicate that, for Model.1, Precision is quite high, but Recall is not, which can be interpreted as follows: that a limited amount of students classed as a I.Dropout will eventually become completers, whereas a significant number of students classed as I.Completer will finally become dropouts. As a result, a limited number of students can be the target of unnecessary teacher intervention, but what is worse, a significant number of students will be outside the scope of teacher intervention, which would have been useful if they had been correctly classified. Due to the fact that our purpose is to identify as many dropout-prone students as possible, Recall prevails over Precision. As a consequence, Model.1 turns out to have a low degree of satisfaction. Model.2 is slightly more satisfactory than the Model.1 because of its higher Recall, but Model.3 is the best option owing to its reasonably high values of Precision, Recall, and also Accuracy (90.2%, which is therefore acceptable according to [12]). Moreover, Model.3 can be applied after the seventh week of the course, some way before the halfway point of the semester. And finally, because Model.4's Recall does not improve that of the Model.3, and given that our purpose included identification "as soon as possible", we can state that Model.3 is better than Model.4.

6 Conclusion and further research

The main contribution of the present study is to provide a simple and easy-to-use procedure, by means of several classification conditional tree-based models, to identify dropout-prone students before the halfway point of the semester. Firstly, it is simple since there is only a single attribute that contributes to classifying students. That attribute is related to students' behaviour with respect to the completion of low-stake assessment assignments such as quizzes posed by teachers and not related to the usage of the e-learning platform, like forum participation. And secondly, it is easy to use because simply by knowing every time a student has not completed one of the first three posed quizzes is enough to identify him/her directly as a dropout-prone student. Furthermore, because the information required is not only easily accessible by the teacher, but also does not need to be processed, teachers can control the procedure by themselves and implement it once the first quiz is submitted. If the performance measures entail a serious concern for the teacher, the previous procedure has to be modified in some way, although it remains simple and easy-to-use. The procedure consists of checking whether students have completed all of the first three quizzes. If the answer is no, the student is identified as a dropout student.

According to the methodology selected, the students that belong to the training set, with whom the classification models have been fitted, and the students of the validation set, whose performance has been evaluated, are enrolled all together in the same academic year. This limitation could lead to further research. The studies of Lykourentzou [11], Lara [12] and Kotsiantis [13], which create the training set in one academic period and the test set in a different one, are references that it would be useful to bear in mind. A second aspect that could be included in further research is the extension of the identification procedure to the fail-prone students [6], so that a richer approach to the dropout prediction problem could be achieved.

Acknowledgements

This paper has been partially supported by a Fundació IBADA grant. We would like to thank Dr Laura Calvet and Mr Paul Garbutt for their valuable contributions in helping to improve this study.

References

1. Lee, Y., Choi, J.: A review of online course dropout research: Implications for practice and future research. Educ. Technol. Res. Dev. 59, 593–618 (2011).
2. Vossensteyn, H., Kottmann, A., Jongbloed, B., Kaiser, F., Cremonini, L., Stensaker, B., Hovdhaugen, E., Wollscheid, S.: Drop-Out and Completion in Higher Education in Europe - Literature Review. (2015).
3. Cohen, A.: Analysis of student activity in web-supported courses as a tool for predicting dropout. Educ. Technol. Res. Dev. 65, 1285–1304 (2017).
4. Burgos, C., Campanario, M.L., Peña, D. de la, Lara, J.A., Lizcano, D., Martínez, M.A.: Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. Comput. Electr. Eng. 66, 541–556 (2018).
5. Grau-Valldosera, J., Minguillón, J.: Rethinking dropout in online higher education: The case of the universitat oberta de catalunya. Int. Rev. Res. Open Distance Learn. 15, 290–308 (2014).
6. Liu, S., Gomez, J., Yen, C.-J.: Community College Online Course Retention and Final Grade: Predictability of Social Presence. J. Interact. Online Learn. 8, 165–182 (2009).
7. Levy, Y.: Comparing dropouts and persistence in e-learning courses. Comput. Educ. 48, 185–204 (2007).
8. Dupin-bryant, P.A.: Pre-Entry Variables Related to Retention in Online Distance Education. Am. J. Distance Educ. 18, 199–206 (2011).
9. Costa, E.B., Fonseca, B., Santana, M.A., de Araújo, F.F., Rego, J.: Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Comput. Human Behav. 73, 247–256 (2017).
10. Santana, M.A., Costa, E.B., Neto, B.F.S., Silva, I.C.L., Rego, J.B.A.: A predictive model for identifying students with dropout profiles in online courses. In: Workshop Proceedings of the EDM 2015 International Conference on Educational Data Mining Vol 1446.
11. Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., Loumos, V.: Dropout prediction in e-learning courses through the combination of machine learning techniques. Comput. Educ. 53, 950–965 (2009).
12. Lara, J.A., Lizcano, D., Martínez, M.A., Pazos, J., Riera, T.: A system for knowledge discovery in e-learning environments within the European Higher Education Area - Application to student data from Open University of Madrid, UDIMA. Comput. Educ. 72, 23–36 (2014).
13. Kotsiantis, S.B., Pierrakeas, C.J., Pintelas, P.E.: Preventing Student Dropout in Distance Learning Using Machine Learning Techniques. In: Proceeding of the 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, KES 2003, pp. 267–274. , Oxford, UK (2003).

14. Romero, C., López, M.I., Luna, J.M., Ventura, S.: Predicting students' final performance from participation in on-line discussion forums. Comput. Educ. 68, 458–472 (2013).
15. Ruiperez-Valiente JA, Muñoz-Merino PJ, Andújar A, Delgado-Kloos C.: Early Prediction and Variable Importance of Certificate Accomplishment in a MOOC. Proceedings of the European Conference on Massive Open Online Courses, 263-272 (2017).
16. Sharma K, Kidzinski L, Jermann P, Dillenbourg P.: Towards Predicting Success in MOOCs: Programming Assignments. Proceedings of the European Stakehold SUMMIT on Experiences and Best Practices Around MOOCs (EMOOCS), 135–148 (2016).
17. Yang D, Sinha T, Adamson D, Penstein Rose C.: "Turn on, Tune in, Drop out": Anticipating Student Dropouts in Massive Open Online Courses. Proceedings of the 2013 NIPS Data-driven education workshop, 1-8 (2013).
18. Figueroa-Cañas J, Sancho-Vinuesa T.: Investigating the relationship between optional quizzes and final exam performance in a fully asynchronous online calculus module. Interact Learn Environ. (2018).
19. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning. (2013).
20. Kotsiantis SB, Zaharakis ID, Pintelas PE (2006) Machine learning: A review of classification and combining techniques. Artif Intell Rev 26:159–190.
21. Hothorn, T., Hornik, K., Zeileis, A.: Unbiased recursive partitioning: A conditional inference framework. Research Report Series 8, Department of Statistics and Mathematics, WU Wien, 2004. J. Comput. Graph. Stat. 15, 651–674 (2006).