Predicting student performance over time. A case study for a blended-learning engineering course

Juan Antonio Martínez[1] [0000-0002-7696-6050], Joaquim Campuzano[1] [0000-0002-4055-8346], Teresa Sancho-Vinuesa[2] [0000-0002-0642-2912] and Elena Valderrama[1] [0000-0001-7673-2310]

[1] Universitat Autonoma de Barcelona, Edifici D, Campus UAB, 08193 Bellaterra, Spain
[2] Universitat Oberta de Catalunya, Rambla del Poblenou, 156, 08018 Barcelona, Spain
juanan.martinez@uab.cat

Abstract. In recent years, different studies have focused in analyzing whether it is possible to explain and predict performance of students based on information we know about them, and in particular, on that obtained from Learning Management Systems (LMSs). A review of existing literature shows we can still raise no conclusion, and in particular when dealing with face to face (F2F) studies.

In this article, we analyze the performance of a first-year engineering course, offered in a higher education institution (a public university). The course under analysis lasts for 12 weeks and is offered with flipped classroom methodology. Activities that students should follow out of class are scheduled in advance, and communicated to students during the learning period. In addition, there has been a previous effort to align learning activities and learning outcomes.

The goal is to determine if prediction models fed with data gathered during the learning process can provide an accurate estimator of students at risk. This risk evaluation will be done considering as core data those reflecting activity, being of particular relevance, traces stored in LMS as part of the learning process.

Our study demonstrates performance can be estimated based on this data, with increasing accuracy over time. Activity performed by the student is linked to academic result, and this relation is verified even when not taking into account any graded results obtained during the learning process.

Keywords: learning analytics, performance prediction, student modelling.

1 Introduction

General adoption of LMSs has motivated a growing interest for Educational Data Mining and Learning Analytics in general and academic performance prediction in particular. Prediction is one of the most explored areas in both fields, increasing its relative weight over time. Research papers related to prediction were around 30% in 1995-2005 [1], while this number increases to over 40% in more recent studies [2].

This rising interest for prediction is more than justified, due to its potential impact at all educational levels. At the macro level, it can help institutional managers to implement educational policies addressed to reduce failure and increase overall quality. Predicting and understanding the reasons behind prediction results can be a powerful tool to increase global academic performance.

While this macro level is undoubtedly of great interest, we will focus on the teaching impact. We consider prediction a potential lever with different potential applications. If we can provide early prediction, we can redress risk behaviors. This change in behavior can be implemented in a passive way – i.e. just informing students of the potential risk – or in a more active fashion – i.e. implementing specific teaching measures -. If we consider late predictions, they can help to understand causes of failure and redesign pedagogical approaches for forthcoming editions of the same course.

Our study will be carried out in a first course engineering subject in a public on-campus university. Being an on-campus university is relevant, as attending on-campus classes mixes with on-line activities. The subject under analysis has been selected due to its blended-learning methodology, where class attendance mixes with on-line lectures and activities. We aim to estimate probability of success, focusing on behavior of the student regarding the different activities. In particular, considering behavior of the students both in on-campus and on-line activities.

The goal of this research is to evaluate to what extent risk of failure in a flipped face to face (F2F) course can be predicted based on the analysis of the student's behavior. The underlying hypothesis is that student performance (in terms of pass-fail) can be predicted, in good measure, by his/her activity during the course even not taking into account grades gathered in evaluative activities performed during the course. In order to validate the hypothesis we raise two research questions (RQ):

- RQ1: Can student's final performance in terms of "pass-fail" in a course be anticipated by analyzing his/her behavior regarding the fulfillment of the programmed learning activities without taking into account grades obtained in evaluative assessments?
- RQ2: How is this prediction influenced when limiting data to those gathered in the early stages of the course?

Affirmative answer to the first question would suggest activity performance is linked to academic result. This would open a path to better understand the learning process for this particular subject and to suggest potential improvements in pedagogical design. Regarding the second question, early prediction can be useful to redress individual student behavior and reduce overall failure. Answers to both questions would help to improve teaching quality.

2 Theoretical framework

Learning analytics (LA), defined formally by Siemens [3] and Ferguson [4], cover a full set of studies dealing with the extraction of meaningful information from data retrieved in the learning process. Ferguson [4] focuses on "measurement, collection, analysis and reporting of data”. Closely related, but with different goals, we find Educational Data Mining (EDM), more commonly accepted definition by Baker and Yacef [2] which tends to focus on techniques.

A core concept in both fields is data, being of particular interest those gathered during the learning process. While the process of data gathering related to the learning process is intrinsic in online universities, most on-campus universities were not born with this idea in mind. In recent years, and with the general adoption of LMS systems, there has been a shift towards data gathering and analysis.

The extraction of information from LMS data is a topic on its own. Agudo-Peregrina, Iglesias-Pradas, Conde-González, and Hernández-García [5] suggest to begin by classifying information around two main axes: interactions based on agent and interactions based on frequency of use. Each of these axes will include a number of specific variables, which depend on the study and the expected output [6].

Conijn, Snijders, Kleingeld, and Matzat [6] compiled pre-existing work and summarized variables considered of potential interest in the literature. The use of variables which are linked to the learning process is common. In particular, we can find number of resources viewed, quizzes started, sessions or total clicks. It is likely to remark that the different studies analyzed provide different impact and influence of variables depending on the course under consideration.

According to [5] there is also no consensus on the influence of a given variable. This fact is also reflected in [6], concluding that in order to get better results “we need to get a better insight into what the LMS data represents”.

Both compilations ([5, 6]) show the huge number of different variables that are present in different studies. This is also present in LMS related web sites who focus on information gathered from LMS systems [7]. Whichever the initial variable set is, a selection process will be mandatory, in particular if the number of variables is high in relation to the number of samples.

Before entering the prediction process itself, the nature of the problem must be focused. Failure analysis can be approached as a regression problem (i.e. estimating final graded performance of student) or as a classification problem (i.e. analyzing whether the student will pass or fail). The classification approach is common in the literature, with studies suggesting better performance and potential detection of meaningful patterns [8].

Once variables are selected, and considering we face a classification problem, different studies use different methods for prediction. To have some examples, the range goes from simple decision trees [9], to behavioral clustering [10]. Different research compilations regarding techniques ([11, 12]) show there is no universal method that provides suitable results for all situations. In our case, and considering our study is not focused on the techniques themselves, we will evaluate results with most common methods, without being tied to a particular one, putting the focus on the interpretation of results.

Whichever the technique, evaluating model goodness is the next required step. Due to the classification nature of the problem, area under curve (AUC) is a potential indicator of model goodness. AUC “is a one-number measure of a model’s discrimination performance, i.e., the extent to which a model successfully separates the positive and the negative observations” [13]. AUC can also help in cases where we are not dealing with large datasets [14].

Some studies raise concerns about use of AUC as performance indicator [15]. In particular, AUC makes no difference between errors, although when thinking of failure/success classification this can be of potential interest. Depending on the potential application of the prediction, false negative and false positive errors could have different impact, and this information is not contained in AUC.

LA and EDM reseearch does not have unique indicators for evaluating model performance. Due to this fact, comparison of published research works is not straightforward ([16]). Different studies apply different metrics regarding model validation. Bowers, Sprott and Taff ([16]) suggest a framework for comparison. This framework considers global accuracy, but at the same time, includes a graphical view with information related to sensitivity and specificity.

Considering the potential drawback of AUC as a unique indicator, and also the need to compare to previous research, keeping accuracy, sensitivity and specificity besides AUC can help to effectively compare with pre-existing works. AUC can be a general indicator of the overall quality of the classification, while rest of parameters provide additional information and allow to compare with previous research.

This same compilation includes articles with different time scenarios. Best performing models are fed with long-time data (math achievement trajectories from grades 7-12, non-cumulative GPA (Grade Point Average) from grades 9-12, and student engagement trajectories from grades 8-12). All of them include evaluative data as input variables to the model.

The impact of graded activities gathered during the course is common in the literature. At the same time, some studies compiled look for pre-existing variables that could be also of potential interest to performance models. This variables can be external to the learning process, and can include social or economic aspects.

As a final consideration, [16] also concludes that “the predictive utility of many variables is dependent upon course site design and pedagogical goals”. It seems clear that while there has been a technical approach to data mining, there has not been such an evolution on seeking the interpretation or generation of relevant information from the data stored in information systems in general and LMS in particular.

3 Methodology

We face a classification task, without being restricted to a particular data mining technique. Models will be fed with activity data. This activity data will also include a time scope in order to evaluate three kind of models according to the time when the data are gathered: early (4 weeks), medium (8 weeks) and late (12 weeks).

3.1 Suitable techniques

While it is not the goal of the paper to discuss about data mining algorithms, we did not want to restrict our study to a particular technique. We selected those present in literature compilation regarding student performance prediction [12]. Selected techniques were naive Bayes, neural networks, decision trees (including both gradient boosted trees – GBT- and random forest –RF-) and Support Vector Machines (SVM). For all of these techniques, we will keep classification error, sensitivity and specificity as parameters to compare with existing literature, and AUC as an additional check.

Regarding models, we define the true positive class as those students likely to fail who actually fail. Those students marked as failing who really pass will be considered False Positives (i.e. Type I errors), while students marked as passing who really fail will be the false negative class (i.e. Type II errors). This approach will permit direct comparison with results in [16].

3.2 Variables

Due to the different and high number of variables present in the literature, and to the fact that they normally include graded activities, we decided to begin from scratch, but keeping in mind lessons learnt from previous compilations ([5, 6]). In particular, we look for meaningful variables linked to the learning process.

We reviewed our course design, and looked for knowledge derived from our teaching experience. We considered three core concepts as fundamental to explain academic results: class attendance, continuous working and flipped behavior. Not all this piece of information was kept as structured data before performing this study.

In particular, we had no information regarding class attendance. Class attendance is not mandatory, and there is no specific control, as students can decide – without academic impact – whether they attend classes or not. The introduction of the flipped classroom methodology made us think about potential non-intrusive techniques to estimate it.

This estimation was performed through the use of a learning engagement tool (Socrative). This tool was introduced as part of the course design to help the detection of areas that need reinforcement. Questions are performed to students during class to evaluate contents that are clear and those that need reinforcement. Questions have no impact in grades. They help instructors to focus on specific areas depending on the answers students provide. Information in the logs allow us to provide an estimation of student attendance to class. We summarize attendance in each of the periods (early, medium and late attendance). It is an estimation – and not an exact value – as the tool is not used in every class.

Continuous working is complex to evaluate and measure. In order to keep simple and at the same time meaningful variables, we opted to keep the volume of information collected in the LMS log file per user and week. We kept one variable for each week of the course that reflects the amount of log lines the LMS. For each of the periods (early, medium, late) we consolidate work in the whole period into a single variable.

We raised concerns regarding activities performed offline. To capture this offline activity, we reinforced the need to use the LMS as part of the pedagogical design. Users can obviously work offline, but video lessons and problem solving require access to the platform. In this way, we can assume users with greater activity levels are those with greater number of log entries.

The above group of variables can reflect continuous work but does not directly link to flipped behavior. The flipped methodology would make advisable to review certain topics before attending class. The list of required activities and due dates is part of the course design. These activities and dates are communicated to students in advance on a per-week basis. So, we included a new set of variables, reflecting for each week the amount of work that was assigned to that week and was effectively performed on time.

The need to get this information requires that all instructors share a common set of activities instructed to students. Each of the activities will have a unique indicator. Once this indicator is located in the log files for a given user, date can be compared to due date for that activity. This approach makes it possible to compute on-time performance of activities for every student, provided that all instructors set the same dates for activity performance.

So far, we have variables reflecting class attendance for each of the periods. We also have a per-week estimation of workload performed based on the log data, and finally the amount of work assigned to each of the weeks performed on time. For this last two datasets we also keep the total work performed in the period.

Due to the high number of variables, a forward selection process will be necessary. This is done to keep the recommended ratio between number of variables and number of samples avoiding potential overfit [17]. This operation will be done for each of the time scopes (early, medium, late) under analysis.

3.3 A word on pre-existing data

Different studies have analyzed pre-existing variables which can condition students' outcomes [16]. We discarded variables without direct link to the learning process. In our case, and after discussion, we kept the grade you get when entering the university, and the fact of being new or repeating student. Variables such as city of residence or family income were not considered due to our focus on activity.

Regarding the grade the student enters the university with, we thought that under the same conditions, students with higher entering grades should be more likely to pass. Regarding the fact of being new or repeating student, our experience shows that repeating students show different behavior than those being enrolled in the subject for the first time.

3.4 Summary

Table 1 summarizes the information we have provided both for methods and variables. Remember the goal will be to classify students based on probability of passing or not for different moments along the course.

Table 1. Summary of classification methods and variables.

Classificiation methods	Variables (common to all methods)
Naive Bayes Neural Networks (NN)	Class attendance (summarized for early, medium and late period)
Decision trees (DT) – incl. GBT and RF-	Work performed on a per-week basis, estimated through the LMS log file.
Support Vector Machines (SVM)	Total work in early, medium and late period
	Work corresponding to the contents covered in class in each specific week (on-time work)
	Aggregated on-time work for early, medium and late period
	Number of times the student was enrolled in the subject
	University access mark

4 Results

Table 2 shows the results for the different methods and time scopes. For clarity, only Random Forest is shown among decision tree techniques, as gradient boosted and simple decision trees provided no better results. Cross-validation has been performed through k-fold cross-validation.

Table 2. Results for different methods and timelines

		Bayes	NN	RF	SVM
Early	Classif error	36.6	37.1	35.1	33.7
(1st block)	(+/-STD)	(+/-12.3)	(+/-3.7)	(+/-5)	(+/-9.3)
	AUC	0.75	0.734	0.724	0.671
	(+/-STD)	(+/-0.105)	(+/-0.11)	(+/-0.09)	(+/-0.13)
	Sensitivity	53.6	92.4	51	74.8
	(+/-STD)	(+/-31.7)	(+/-6.6)	(+/-10.5)	(+/-9.4)
	Specificity	82.6	33.0	79	51.1
	(+/-STD)	(+/-11.1)	(+/-8.5)	(+/-10.9)	(+/-18.9)
Medium	Classif. error	29.9	25.4	28.1	29.5
(2nd block)	(+/-STD)	(+/-3.7)	(+/-6.7)	(+/-2.2)	(+/-11)
	AUC	0.817	0.831	0.8	0.715
	(+/-STD)	(+/-0.081)	(+/-0.05)	(+/-0.08)	(+/-0.1)
	Sensitivity	64.6	66.9	66.7	50.1
	(+/-STD)	(+/-4.9)	(+/-11.7)	(+/-14.4)	(+/-18)
	Specificity	79.4	80.6	77.4	89.8
	(+/-STD)	(+/-11.6)	(+/-7)	(+/-5.1)	(+/-8.5)
Late	Classif. error	24.4	24.7	35.6	29.9
(3rd block)	(+/-STD)	(+/-5.7)	(+/-3.4)	(+/-5.9)	(+/-5.1)
	AUC	0.86	0.827	0.718	0.826
	(+/-STD)	(+/-0.062)	(+/-0.066)	(+/-0.09)	(+/-0.06)
	Sensitivity	77	73	51.9	85.8
	(+/-STD)	(+/-5.3)	(+/-6)	(+/-16.4)	(+/-11.9)
	Specificity	76.4	78.2	78.4	61.2
	(+/-STD)	(+/-12.6)	(+/-3.9)	(+/-11.8)	(+/-8.7)

We have compared results with the compilation in [16], where impact of different variables in published models is shown. For clarity, we have added just our Bayesian models results, as they offer best performance in terms of AUC for early and late activity, and for mid-term is close to maximum. Results are shown in figure 1, marked as "Early prediction", "Mid-term prediction" and "Late prediction". The compilation includes 110 indicators (depicted as numbers in Figure 1) from 36 different prediction works:

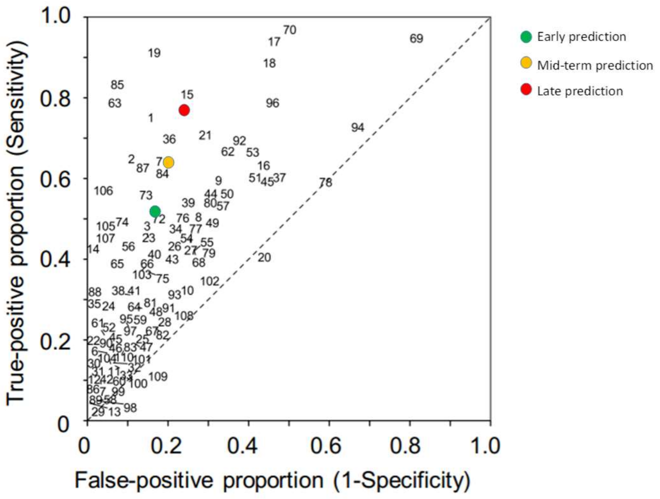

Fig. 1. Comparison of current work and previous research. Adapted from [16]

We also analyzed which were the variables that show relevant in the different time scenarios for the Bayesian method analyzed. Table 3 summarizes the results for the different time scopes:

Table 3. Relevant variables for different periods under analysis

Period under analysis	Selected variables
Early	Attendance to class Work in weeks 1,4 Flipped behavior in weeks 2,3 Access grade to university
Medium	Attendance to class (both during early and medium periods) Work in weeks 1,4,5,7 Overall work in Medium period Flipped behavior in week 3
Late	Attendance to class (all periods) Work in weeks 1,4,7,11 Overall work in Late period Flipped behavior in week 3

5 Discussion

Before answering RQ1 and RQ2, we need to validate to what extent the models outlined in this paper are of potential interest. This question can be answered analyzing Figure 1 and results in Table 2.

Figure 1 shows that final performance of student can be predicted based on the activity data analyzed. We can create models that anticipate success based on this activity data, being those models more accurate the longer the period under consideration. AUC values confirm also models are better the longer the period we analyze.

Prior to comparing with other LA/EDM works, we want to remark our focus has been set on activity. We could obtain potential better models by analyzing partial grades obtained by students, but this could mean losing the focus on the impact of student behavior regarding the subject. This is particularly relevant for us, even more considering we are facing changes in methodology, such as the flipped behavior.

Limiting analysis to early stages provides average results. While we believe a longer time period would be advisable, in particular if the goal is to take actions which can derive costs, performance is similar to other published results. In particular, if we look into the works compiled in [16], we find studies with similar performance, such as [18, 19] – with indicators depicted as 3,72 and 73 in Figure 1–. In the first case, indicators included in the study are socio-economic, while in the second they are related to extra-curricular activities. As a noticeable point, none of them includes grades as predictor variables.

Within our constraints, to get better results it is necessary to broaden the time scope. Doing so – medium and late models – we get results similar to [20, 21] (whose variables are depicted as points 1,15 and 36 in Figure 1). Those studies have also broader time scopes (minimum 1 year) and do not constrain to limit grading data. Best predicting scenarios in [16] include always graded data [22, 23].

Being able to predict results considering whether activities have been performed or not is relevant from a pedagogical point of view and opens potential future lines. While it is not the main goal of this research we also have analyzed results in Table 3 regarding individual variable impact. If we deep into variable details, attending to class, or performing homework makes a difference. Looking this fact from another angle, we can tell students that coming to class and doing what they are instructed to will help them to pass the subject.

Class attendance is present in all cases, independently of time scope. Regarding homework, during the first period, it is relevant your attitude in the first weeks, and just before first partial test. When the period is longer, it becomes more relevant the amount of work performed in the whole period. We believe there is even room for improvement with this same dataset trying to look for other pieces of information that can remain unnoticed inside the huge data volume.

We have compared our findings with results in [5] . In that case, authors consider "there is a relation between some type of interactions and academic performance in online courses, whereas this relation is non-significant in the case of VLE-supported F2F courses". We believe this relation can be found also in F2F or VLE supported studies as long as they include a pedagogical design that requires the use of VLEs. If this is done so, evidences will be gathered in the LMSs and can show differences in behavior.

Consistently with findings by Bowers et al. [16] for dropout flags, and in particular for early prediction, models still lack accuracy (i.e. classification error is high). The potential impact of this error will depend on the purpose of the prediction. If we use them to just raise early alarms, it would not be critical. If deeper pedagogical actions are taken to redress behaviors that can anticipate failure, there would be non-optimal use of resources

Although the models exposed can set the basics for targeted actions, the design of specific policies or pedagogical interventions to reduce failure should take into account not only global accuracy but the impact of false positives and false negatives. For instance, we could consider small group tutoring actions for students classified as likely to fail. The specific design of the action should be made taking into account the false positive rate – i.e. it will affect students who would potentially pass without the action – and at the same time the false negative rate – i.e. there is a group of students marked as passing who will not potentially pass -. We believe this analysis opens a really interesting future line in the field of pedagogical design.

The longer the period, the higher the values for accuracy and sensitivity. In other words, with longer periods we are more certain about final results regarding true positive class. In our case, that means we are more certain behavior of the user could lead to failure.

While this has been a model for an individual subject, we would like to make a reflection about robustness and portability. Computation of data in Table 2 has been done through cross-folding validation, and shows high values of variance in some cases (i.e. sensitivity in early models). To solve this issue it would be advisable to have a higher number of samples (i.e. more students to analyze).

Regarding portability, the process we followed to extract information shows there is a great dependence on course design. Portability of the resulting model itself is not straightforward, but we believe the methodological approach is. The analysis followed can help to obtain models for any flipped classroom course. A pedagogical design that helps gathering evidences from LMS, combined with meaningful variables and, for F2F universities, class attendance should generate models that anticipate potential success based on pure activity data. We believe it will be difficult to generate portable results among different subjects even in same university unless they share a common course design.

Although it was not the goal to establish a comparison among algorithms, Bayesian models have shown good performance related to computational cost. SVM performs also well, but at a higher computational cost. Decision-tree family algorithms can be of interest but would need a higher number on samples to avoid deviations. Finally, deep learning techniques have not provided considerable gain and have higher computation requirements.

To sum up, and going back to the research questions introduced in this paper, final performance of individual students can be anticipated considering only activity data. Relevant aspects for success, considering the course design in our study, include class attendance and different aspects related to homework. Regarding the influence of time, early periods lack accuracy, and would not be optimal if the goal is to set-up actions which involve high costs. As we consider longer periods – medium and late – the models get better. Results for these medium and late models can be of potential help both to redress behavior – in the case of the medium prediction – or – once course is finished – to analyze results and improve course design for future course sessions.

6 Open lines

This paper wants to set the basics for defining specific actions to reduce failure in engineering studies in higher education. Lines of activity include:

- Improve models, in particular in early periods, potentially including new data.
- Deepen into the meaning of the variables selected as more relevant.
- Apply same methodology to other subjects in order to validate and compare results.
- Define actions to reduce failure based on early and medium prediction analysis and to improve pedagogical design based on early, medium and late predictions.

Authors are open to collaboration in previous lines – or to carry out similar research in other environments –. For those interested in carrying out similar research on their own, data processing was done through Python scripts, using Scikit-learn libraries (https://scikit-learn.org/) for modelling algorithms. In particular, sklearn.naive_bayes, sklearn.neural_network, sklearn.tree, sklearn.ensemble (for GBT) and sklearn.svm implementations were relevant among those used [24]. The method does not rely in any particular LMS, but our study was carried out on Moodle platform (https://moodle.org/). Final models were also tested on RapidMiner software (https://rapidminer.com/) to validate results.

References

1. C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Systems with Applications,* vol. 33, no. 1, pp. 135-146, 7 2007.
2. R. S. J. D. Baker and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions".
3. G. Siemens, "What are Learning Analytics?," 2010. [Online].
4. R. Ferguson, "Learning analytics: drivers, developments and challenges Journal Item Learning analytics: drivers, developments and challenges," *International Journal of Technology Enhanced Learning,* vol. 4, no. 5, pp. 304-317.
5. Á. F. Agudo-Peregrina, S. Iglesias-Pradas, M. Á. Conde-González and Á. Hernández-García, "Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning," *Computers in Human Behavior,* vol. 31, pp. 542-550, 2 2014.

6. R. Conijn, C. Snijders, A. Kleingeld and U. Matzat, "Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS," *IEEE Transactions on Learning Technologies,* 2017.
7. "The Indicators Project – Dabbling in analytics," [Online]. Available: https://indicatorsproject.wordpress.com/.
8. P. Strecht, L. Cruz, C. Soares, J. Mendes-Moreira and R. Abreu, "A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance".
9. J. Yoo, S. Yoo, C. Lance, J. Hankins, J. Yoo, S. Yoo, C. Lance and J. Hankins, "Student progress monitoring tool using treeview," in *Proceedings of the 37th SIGCSE technical symposium on Computer science education - SIGCSE '06*, New York, New York, USA, 2006.
10. P. J. H. Y. I.-H. J. Yeonjeong, "Clustering blended learning courses by online behavior data: A case study in a Korean higher education institute," *The Internet and Higher Education,* pp. 1-11, 2016.
11. F. Castro, A. Vellido, À. Nebot and F. Mugica, "Applying Data Mining Techniques to e-Learning Problems".
12. A. M. Shahiri, W. Husain, N. '. Aini and A. Rashid, "ScienceDirect The Third Information Systems International Conference A Review on Predicting Student's Performance using Data Mining Techniques," *Procedia - Procedia Computer Science,* vol. 72, pp. 414-422, 2015.
13. S. Rosset, "Model selection via the AUC.," in *Machine Learning, Proceedings of the 21st International Conference*, 2004.
14. C. Ferri, J. Hernández-Orallo and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognition Letters,* vol. 30, pp. 27-38, 2008.
15. J. M. Lobo, A. Jiménez-Valverde and R. Real, "AUC: a misleading measure of the performance of predictive distribution models," *Global Ecology and Biogeography,* vol. 17, no. 2, pp. 145-151, 1 3 2008.
16. A. J. Bowers, R. Sprott and S. Taff, "Do We Know Who Will Drop Out? A Review of the Predictors of Dropping out of High School: Precision, Sensitivity and Specificity," vol. 96, no. 2, pp. 77-100, 2013.
17. F. E. J. Harrell, K. L. Lee and D. B. Mark, " "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Stat Med,* pp. 361-387, 1996.
18. A. I. ""Mother got tired of taking care of my baby." A study of dropouts," *Austin, Texas: Austin Independent School District.,* 1982.
19. J. L. Mahoney and R. B. Cairns, "Do extracurricular activities protect against early school dropout?," *Developmental psychology,* vol. 33, no. 2, pp. 241-253, 1997.
20. E. Allensworth and J. Easton, "What matters for staying on-track and graduating in chicago public high schools: A close look at course grades, failures, and attendance in the freshman year," 2007.
21. D. Doss, "Ninth grade course enrollment and dropping out," in *Annual Meeting of the American Educational Research Association*, San Francisco, 1986.

22. A. J. Bowers and R. Sprott, "Examining the multiple trajectories associated with dropping out of high school: A growth mixture model analysis," *Journal of Educational Research,* vol. 105, no. 3, pp. 176-195, 1 4 2012.
23. B. Muthén, "Latent Variable Analysis: Growth Mixture Modeling and Related Techniques for Longitudinal Data," in *Handbook of quantitative methodology for the social sciences*, D. Kaplan, Ed., Sage Publications, 2004, pp. 345-368.
24. «API Reference — scikit-learn 0.21.2 documentation» [Online]. Available: https://scikit-learn.org/stable/modules/classes.html.
25. L. P. Macfadyen and S. Dawson, "Mining LMS data to develop an "early warning system" for educators: A proof of concept," *Computers & Education,* vol. 54, no. 2, pp. 588-599, 2 2010.