PAELLA:

Personalized student **A**ctivation in **E**ngineering-education: **L**everaging **L**earning **A**nalytics for an engaging blended learning course design

PerActiLA:
Personalized Student Activation through Learning Analytics-based insights about students' learning processes

Progress Report R3
The Canvas data modeling

Version 1: course 1

24.06.2022

Eindhoven University of Technology

Authors: Dr Rianne Conijn, Dr. Uwe Matzat, Dr. Ad Kleingeld, Prof. dr. Chris Snijders, and Sonja Kleter

1. Introduction

In report R3 we describe the models that we develop to identify students at risk (nonengaged students who are likely to profit more from a mindset intervention). A major theme in the field of learning analytics is the prediction of student performance. These models often include learner interaction within an online course environment, such as a learner management system, where - generally - more activity is related with a higher student performance. These models are generated based on the full student interaction in the course. However, it could be argued that this is not useful for predicting student at risk, or students in need of support (Romero & Ventura, 2019). Being able to predict student performance at the end of the course would be too late to meaningfully intervene. Therefore, in the current project we will examine how accurately we can predict student performance within a course, already after two weeks in the course. Below, we specifically describe the early prediction of student performance within course 1, a first year bachelor course in the major Psychology and Technology. The model will be created on the academic year 2020-2021, and then used for the intervention in academic year 2021-2022 (see report R5). For the prediction, both statistical and machine learning models will be used.

2. Description of the data

The data used in this study was collected from two data sources: Osiris and Canvas. The Osiris data included in-between assessment data, final exam grades, and final course grades. The Canvas data included all student interactions (clicks) within the learning management system Canvas. Based on these clickstream data a wide range of indicators were extracted that could provide insight into students who are doing well in the course, or students who might benefit of an intervention (students 'at risk'). Here, we specifically opted for relatively general indicators, which would be present in most courses. These indicators are generated over the full-course time period, as well as per week, to allow for timely prediction of student performance. The indicators can be divided into five categories:

- **Students’ behavior at the course start**. This can provide an initial measure of time management. Namely, students who require longer to access the course for the first time after its publication, and those who access the schedule or study guide later, are potentially less motivated or have less time management skills.
- **Study sessions**. Here, a session is defined as a series of online activities, with a new session beginning when a student was inactive for at least 30 minutes. Such descriptive indicators of online activity are useful especially in the first phase of the course when no assignments or quizzes have been submitted yet and too few data is present to infer learning patterns. The indicators include, the number of sessions, the mean time of a session, the mean time between sessions, and the number of clicks. In addition, the variance of session time and time between sessions is included. Especially the latter are considered to be more informative measures of time

management, as they indicate learning patterns rather than frequencies (Li et al., 2018).

- **Assignment submissions.** The third category of LMS indicators inform on contextual information around quizzes and assignments submissions. Although the in-between assessment data carries much information about performance on the assignments and quizzes, this provides no information on the learning behavior leading to these grades. Examples of quiz indicators include the number of practice quizzes performed, or the number of quizzes attempted more than once. For the assignment indicators, an important contextual indicator is the number of late submissions which can signal a lack of time management.
- **File access**. This indicates how often students access the course resources, including the number of clicks and downloads of course files.
- **Discussion forum usage**. Several previous studies have incorporated indicators concerning the usage of discussion forums and announcement forums (Conijn et al., 2017; López-Zambrano et al., 2020; Sandoval et al., 2018). Students who more actively use these forums are expected to be more concerned with the course material. It should be noted, however, that there is often small variation in these indicators, given that only few students post on a forum. More students will solely read the forum posts, looking for information or an answer to their question. The number of clicks on posts is expected to represent this behavior accurately.

It should be noted that previous studies also often describe the use of video data to predict student performance, or students who are 'at risk'. However, these video data are not directly accessible via the Canvas data. Video data would need to be collected from separate sources, such as TU/e Mediasite, Panopto, or YouTube. These sources were not available for the current analysis, and hence not included.

Descriptive statistics

The course offering in 2020-2021 consisted of 208 students. Of these, 28 people did not make the final exam (13%). The average exam grade was 5.1 (SD = 1.9). An overview of the final exam grades can be found in Figure 1.

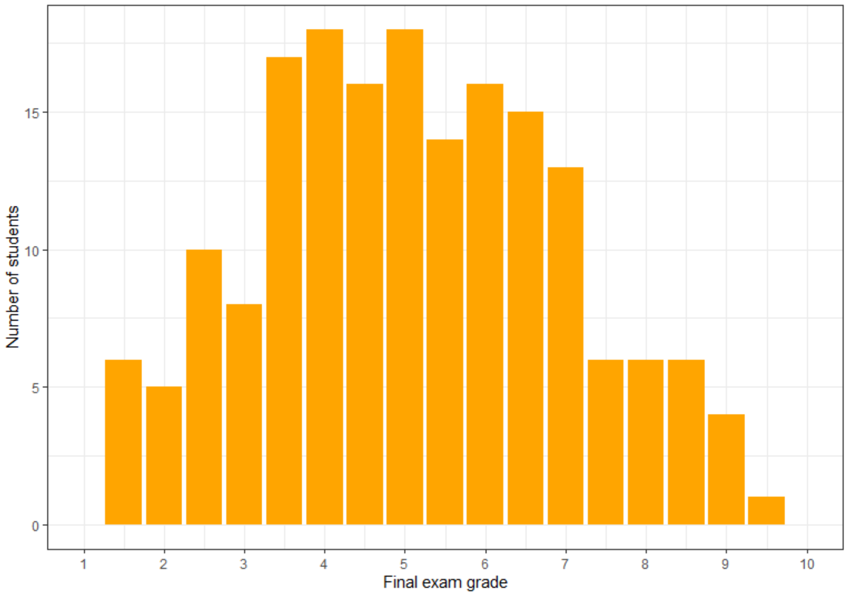

Figure 1. Distribution of final exam grade across the course.

For the Canvas data, it was found that 18 (9%) people did not access the course page on Canvas at all. In addition, there were no quizzes in the course, so there were no quiz activities found in the data. An overview of the descriptive statistics per category can be found in Table 1. As can be seen, it took students on average more than 7 days before they accessed the Canvas course page, after the course was published. The default course schedule and course information were only accessed by a handful of students, which could indicate that this information was also provided elsewhere (on Canvas, or via the lectures). On average, students had a total of 27 study sessions in the course, which lasted for 20 minutes. There were only three assignments in the course, which results in relatively low activity for the assignments. Also, the file access and especially file downloads were relatively low. Finally, the discussion forum was rarely used, and if it was used, the students posted a new topic, but rarely replied on an existing topic. This already shows us that some of the indicators show very little variance within this course. Therefore, these variables will not be used for the prediction.

Table 1. Canvas Indicators for the course Brain, Body and Behavior (2020-2021).

Category	Indicator	N	Mean	SD	Min	Max
Course start	Time to first login (hours)	190	174	263	0	1853
	Time to first schedule (hours)	16	145	142	8	368
	Time to first course info (hours)	3	39	44	8	89
Study sessions	Number of clicks	190	221	257	2	2821
	Number of sessions*	190	27	15	1	94
	Total session time (min)*	190	649	1037	0	10172
	Mean session time (min)*	190	20	18	0	147
	SD session time (min)*	183	38	43	1	364
	Mean time between sessions (hours)*	186	67	60	11	505
	SD time between sessions (hours)*	183	90	57	24	519
	Mean start time (hour of the day)*	190	13:06	02:12	05:30	23:40
Assignment submissions	Number of assignment clicks*	184	14.3	11.3	0	106
	Number of assignments**	184	2.8	0.5	1	3
	Number of late assignments**	184	0.1	0.4	0	2
File access	Number of file clicks*	184	20.7	25.2	1	155
	Number of file access*	184	4.7	2.6	1	21
	Number of file downloads	184	0.3	0.6	0	3
Discussion forum usage	Number of forum clicks*	167	2.3	4.5	0	24
	Number of announcement clicks*	167	8.0	7.5	0	45
	Number of forum topic posts**	167	0.2	0.6	0	4
	Number of forum reply posts	167	0.0	0.2	0	2

Note. * Indicator is used in the final prediction model. ** Indicator is in the final prediction model, but as categorical predictor.

3. Prediction of final exam grade (over full course)

We now use linear regression to predict the final exam grade, based on the data over the full course. It is shown that mean interval time between sessions, the mean start time on the day, and the number of file downloads can be used to predict the final grade. All relations are roughly equal in size (standardized effect sizes are about 0.2). The correlation between the predicted values of the model and the true scores equals 0.53, which is a reasonable result. The direction of the effects of the three predictors with significant contributions is straightforward for the mean time between sessions (longer absence results in lower grades), and the number of file downloads (more file downloads results in higher grades). Perhaps somewhat surprising is the effect of the mean start time on the day. We defined a day as starting from 6 am, and it turns out that the students who start their working day relatively late, score lower grades on average. When controlling for these three predictors, all other variables show negligible effects.

Combined, the variables explain 31% of the variance in final exam grade. The mean absolute error is 1.27, so the model is 1.27 points off on average, with the grades ranging from 1-10. This sounds large, but if we consider all students who scored lower than a 5.0 (that is, students who could be considered 'at risk'), the model predicts a score of 6.0 or more only in 6 cases (7%). Next, we will determine whether we can still accurately predict final exam grade after only two weeks in the course.

4. Prediction of final exam grade (after two weeks)

First, we ran a linear regression on all information available over the first two weeks. It should be noted however that 27 of the 208 students did not access the course at all in the first two weeks. For these students, there is no Canvas data to predict the final grade. Of these students, 16 (59%) did not go to the final exam, and the other 11 students scored on average a 5.1. Hence, it could be argued that, on average, students who don't access the course in the first two weeks, would be in need of support.

For the students who did access the course, we ran a linear regression. As the main goal is now prediction (rather than describing the data), we also need to control for overfitting. Therefore, 10-fold cross-validation was used. The only significant predictors are the number of announcement clicks and the mean time between sessions. Here, more clicks leads and a shorter time between sessions results in a higher final exam grade. Combined, the model can only explain 6% of the variance in final grade. Moreover, the model is on average 1.67 points away from the actual grade. It should be noted here that if we would just predict the average grade for each student (baseline model), the prediction would only be 1.52 points away from the actual grade. Hence, it could be argued that linear regression is not accurate enough for the early prediction.

To try to further improve the model, several machine learning algorithms were used, including random forest and support vector machines (radial kernel). Both the random forest model and the support vector machine were slightly better than the linear regression, however, the mean average error (1.57 and 1.58, respectively) were still higher than the baseline model. So, the models did not outperform the baseline. That being said, just predicting the average grade will not be meaningful to distinguish which students are at risk, and who are not (as everyone will receive the same grade). Therefore, we (cautiously) prefer the random forest model.

5. Conclusion

To conclude, it has been shown that leaving less time between sessions, downloading more files, and starting early on results in a higher final exam grade. With these predictors, we can relatively accurately predict student's course performance. However, early on in the course when interventions are still meaningful (e.g., after two weeks), it is hard to predict student performance. Future research should improve these early models, for example by looking into more complex indicators. For example, indicators that are related to the specific course design (e.g., lecture times), or temporal patterns.

For the PerACtiLA project, the results imply that we cannot accurately make an early prediction after two weeks with the models. We therefore need to identify students at risk in our data analyses post-hoc. These post-hoc analyses can make use of the Canvas data of the new and re-designed course until the mid of week 4 when the first application of the intervention took place. Compared with the course of the last year, the re-designed course offers more information about the use of videos and about participation in quizzes during the first weeks which may allow an early enough identification of students at risk.

We (the team of researchers of the project) will present the results of the post-hoc analyses in report R5 (application and evaluation of the intervention). Furthermore, we will continue with the Canvas data modeling for the course 2 and course 3. We will update this report (Report R3) in line with the schedule announced in our proposal.

References

Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2017). Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS. IEEE Transactions on Learning Technologies, 10(1), 17–29. https://doi.org/10.1109/TLT.2016.2616312

Li, H., Flanagan, B., Konomi, S., & Ogata, H. (2018). Measuring Behaviors and Identifying Indicators of Self-Regulation in Computer-Assisted Language Learning Courses. Research and Practice in Technology Enhanced Learning, 13(1), 19. https://doi.org/10.1186/S41039-018-0087-7

López-Zambrano, J., Lara, J. A., & Romero, C. (2020). Towards Portability of Models for Predicting Students' Final Performance in University Courses Starting from Moodle Logs. Applied Sciences 2020, Vol. 10, Page 354, 10(1), 354. https://doi.org/10.3390/APP10010354

Romero, C., & Ventura, S. (2019). Guest editorial: Special issue on early prediction and supporting of learning performance. *IEEE Transactions on Learning Technologies*, *12*(2), 145–147. https://doi.org/10.1109/TLT.2019.2908106

Sandoval, A., Gonzalez, C., Alarcon, R., Pichara, K., & Montenegro, M. (2018). Centralized student performance prediction in large courses based on low-cost variables in an institutional context. The Internet and Higher Education, 37, 76–89. https://doi.org/10.1016/J.IHEDUC.2018.02.002