Influence of course characteristics, student characteristics, and behavior in learning management systems on student performance

Rianne Conijn
Department of Communication and Information Sciences
Tilburg University
NL-5000 LE Tilburg, Netherlands
m.a.conijn@uvt.nl

Ad Kleingeld
Department of Industrial Engineering and Innovation Sciences
Eindhoven University of Technology
NL-5600 MB Eindhoven, Netherlands
p.a.m.kleingeld@tue.nl

Uwe Matzat
Department of Industrial Engineering and Innovation Sciences
Eindhoven University of Technology
NL-5600 MB Eindhoven, Netherlands
u.matzat@tue.nl

Chris Snijders
Department of Industrial Engineering and Innovation Sciences
Eindhoven University of Technology
NL-5600 MB Eindhoven, Netherlands
c.c.p.snijders@tue.nl

Menno van Zaanen
Department of Communication and Information Sciences
Tilburg University
NL-5000 LE Tilburg, Netherlands
mvzaanen@uvt.nl

Abstract

The use of learning management systems (LMS) in education make it possible to track students' online behavior. This data can be used for educational data mining and learning analytics, for example, by predicting student performance. Although LMS data might contain useful predictors, course characteristics and student characteristics have shown to influence student performance as well. However, these different sets of features are rarely combined or compared. Therefore, in the current study we classify student performance using information from course characteristics, student characteristics, past performance, and LMS data. Three classifiers (decision tree, rule-based, and SVM) are trained and compared with the majority class baseline. Overall, SVM is the best classifier to identify pass/fail for a student in a course. However, for more interpretable results, the decision tree or the rule-based algorithm with course characteristics, student characteristics, and midterm data are good second bests. Additionally, it is shown that the different feature sets all have a positive influence on predicting pass/fail. In particular, student characteristics and the midterm grade have a large influence. Compared to these feature sets, LMS data seems less important. Yet, a more fine-grained analysis of the specific metrics found in the learning management system may still yield useful information.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

1 Introduction

Improving learning and teaching is a key topic in educational context. Formerly, this research was mostly done using observational studies and validated questionnaires. With the advancement of computers and the Internet this field entered a whole new era. For example, nowadays the vast majority of educational institutions use learning management systems (LMSs) (Retalis et al., 2006). LMSs are used to provide course content online in a structured way, often in combination with other learning materials, such as presentations, quizzes, assignments, and forums (Piña, 2012). Since every action in an LMS is recorded and stored, a large amount of data is available about students' online behavior. Improvements in data mining techniques in other fields make it possible to deal with those large amounts of data and to conduct more advanced analyses (Clow, 2013).

The fields of educational data mining and learning analytics focus on the use of these educational data to gain insight in learning processes and to improve learning and teaching. A major topic in these fields is the use of data to predict student performance (Romero and Ventura, 2010), where student performance is often quantified by a final exam grade or whether the student passed or failed the course. Predictive modeling of student performance is an important step in learning analytics and educational data mining, as it informs the implementation of intervention, such as personalized feedback. Therefore, in the current study we use several data mining techniques to predict student performance.

Typically LMS data are used for the prediction of student performance (c.f. Hu et al., 2014; Macfadyen and Dawson, 2010; Minaei-Bidgoli and Punch, 2003; Morris et al., 2005; Romero et al., 2013; Zacharis, 2015), while other data such as student characteristics are often not taken into account. However, some student characteristics, such as past performance, personality, and motivation have shown to be generalizable predictors of student performance in various studies in different courses and contexts (e.g. Britton and Tesser, 1991; Conard, 2006; Dollinger et al., 2008; O'Connor and Paunonen, 2007). Nonetheless, these data sources are rarely combined, except for Tempelaar et al. (2015). Tempelaar et al. (2015) even found that LMS data are only of limited value compared to student characteristics and prior performance data. Hence, in the current study we will use LMS data as well as student characteristics (skills and motivation), combined with course characteristics and prior performance data for the prediction of student performance, and determine whether LMS data has additional value next to student characteristics, course characteristics, and in-between assessment grade.

2 Background

2.1 Educational data mining and learning analytics

Research in the area of learning analytics is defined as "the measurement, collection, analysis and reporting of data about learners and their context, for purposes of understanding and optimizing learning and the environments in which it occurs" (Siemens and Baker, 2012). The goal of educational data mining is "to better understand how students learn and identify the settings in which they learn, to improve educational outcomes and gain insight into and explain educational phenomena" (Romero and Ventura, 2013). In educational data mining studies more advanced data mining techniques are used for automated discovery of learner models and automated adaptation of the learning environment (Romero and Ventura, 2013; Siemens, 2011). In contrast, learning analytics studies often use statistical analyses resulting in models that primarily inform teachers about improvements in their teaching (Siemens, 2011). Thus, both fields focus on the improvement of learning and teaching, yet use somewhat different methods to achieve them.

2.2 Predicting student performance

Predicting student performance is one of the major topics of educational data mining and learning analytics (Romero and Ventura, 2010). Already several algorithms have been used for binary classification (pass/fail) and multi-class classification of student performance. For example, using binary logistic regression an overall classification accuracy of 74% (Macfadyen and Dawson, 2010) and 81% (Zacharis, 2015) was found. Minaei-Bidgoli and Punch (2003) compared six classification techniques, and found that k-NN resulted in the highest accuracy (82%) for 2-class prediction of the final grade. CART analysis resulted in the highest accuracy for 3-class (60%) and 9-class (43%)

prediction. A combination of the classification techniques yielded an even better prediction accuracy (87%, 71%, and 51%, for 2, 3 and 9 classes, respectively). Romero et al. (2013) compared 21 classification techniques for 4-class prediction (fail, pass, good, or excellent) and found that CART, GAP, GGP, and NNEP resulted in the best classification (>65% accuracy). When the features were transformed to categorical features (low, medium, or high value) CART and C4.5 resulted in the best classification (>65% accuracy). Hu et al. (2014) found even an accuracy of 93% using the classification techniques CART and C4.5 on pass/fail prediction.

Others used student characteristics and prior performance instead of LMS data to predict whether a student would pass or fail a course. Superby et al. (2006) classified student performance in three classes (high, medium, low) based on student characteristics, such as age, class attendance, and confidence in abilities. They found that linear discriminant analysis resulted in the highest accuracy (57% correctly classified). Cortez and Silva (2008) combined health and family related student characteristics with midterm grades to predict student performance in two courses (Mathematics and Portuguese). Using five classification techniques, they found that the baseline (naive predictor) and a decision tree resulted in the highest accuracy for 2-class classification when midterm grades where included (92 and 93% for both courses, respectively). SVM and random forest resulted in the best accuracy (71% and 85%) when midterms were excluded from the feature set. For 5-class classification the baseline and a decision tree resulted in the highest accuracy when all features where included (79% and 76%). Random forest had the highest accuracy when midterms were excluded (34% and 37%). Kotsiantis and Pintelas (2005) used student characteristics, such as gender, class attendance, and number of children and found that M5rules was the most accurate technique for pass/fail prediction, resulting in an accuracy of 64% and even more than 80% after the middle of the course.

Thus, the variety of classification techniques used result in different accuracies for predicting student performance, where the technique with the highest accuracy varies across studies. Next to the classification techniques, the different feature sets could also result in different accuracies. Although these studies showed that both LMS data and student characteristics can be used to classify student performance with a relatively high accuracy, these feature sets are rarely combined or compared. However, the combination of these feature sets provide more information and hence could even lead to higher accuracies. Additionally, the comparison of these feature sets could give insight in which types of information have most influence on the prediction accuracy. One exception who did combine these feature sets is a study by Tempelaar et al. (2015). They found that student characteristics and performance data have a higher predictive value than LMS data. Accordingly, in the current study we will use data mining techniques with different feature sets: LMS data, prior performance data, course characteristics, and student characteristics to predict student performance to determine (1) which sources lead to the highest accuracy and (2) whether these sources combined result in a higher accuracy.

2.3 Classification algorithms

In addition to analyzing different information sources on the prediction of final grades, we consider a few machine learning classifier approaches. In particular, for teachers it would be useful if they can interpret the reason why a classifier has decided on specific final grade predictions. Classifiers that allow for an explanation based on information that is available early on in a course may help the teacher to improve the learning performance of students. Therefore, it is argued that classifiers such as decision trees and rule based algorithms are preferable (Romero et al., 2013). In this research, we investigate the performance of two classifiers that lead to (at least partially) interpretable results. Additionally, a function based classifier is used, which is expected to lead to somewhat better classification performance at the expense of interpretability.

As the first classifier that may lead to interpretable results, we use a decision tree classifier (J48, which is based on the C4.5 algorithm). This classifier creates a decision tree by recursively identifying the feature with the highest information gain. Splitting the dataset according to the values of the selected feature results in new (sub)datasets, which can then be further subdivided according to other features with high information gain. This results in a tree structure with highly informative features at the top of the tree. The second classifier that leads to interpretable results is a rule inducer algorithm (JRip, which implements the RIPPER algorithm). This algorithm aims to learn rules (boolean conditions

Table 1: Course characteristics.

Course	Calculus A	Calculus B	Appl. Physics formal	Appl. Physics conceptual	Intro to P & T
Quarter	fall	fall	winter	winter	fall
Total number of students	438	1121	836	822	154
Face-to-face hours per week	4.5	5.3	6.0	6.0	4.5
Online hours per week	2.4	2.4	1.0	1.1	0.4
- clicks in content (%)	2.9%	0.6%	1.4%	1.1%	13.2%
- clicks on forum (%)	0.4%	0.5%	0.1%	0.1%	0.2%
- clicks in quizzes (%)	80.2%	85.4%	79.2%	80.5%	46.6%
- clicks in assignments (%)	0.0%	0.0%	0.0%	0.0%	6.1%
N	122	297	45	350	74

based on the features in the dataset) that perfectly classify the training samples in the dataset. Too complex rules are pruned, which essentially makes sure that the remaining rules stay interpretable.

Additionally, we apply a classifier that does not lead to interpretable results, but is expected to lead to high classification accuracy. For this purpose, we selected a support vector based classifier (SMO). The SMO classifier is a binary classifier that aims to learn a boundary in multi-dimensional space to distinguish between the two classes (e.g. pass versus fail).

3 Method

3.1 Participants and course context

For this study, data were used from blended courses taught at Eindhoven University of Technology (The Netherlands) in the first two quarters (fall and winter) of cohort 2014–2015. The data consisted of course characteristics, behavioral data from Moodle LMS, and performance data, which were also used in a previous study (Conijn et al., 2017). In this study, we supplemented this data with student characteristics. Student characteristics came from a survey among prospective students of Eindhoven University of Technology. In total 426 students both participated in the survey and completed at least one course that employed Moodle LMS. Only courses where at least 45 students had taken the test were included, which resulted in a sample of 5 courses with 426 unique students. As some students followed multiple courses (32 students followed 1 course, 326 followed 2, and 68 followed 3), this resulted in a total of 888 students in five courses. As survey data was not available for all students in these five courses, the 888 cases were a sub-sample of all 3,371 cases in these courses (26.3%).

The five courses included were: Calculus A, Calculus B, Applied Physical Sciences formal, Applied Physical Sciences conceptual, and Introduction to Psychology and Technology. These courses were all first-year courses. Calculus and Applied Physical Sciences are compulsory courses for every student, but the type of the courses (A or B, formal or conceptual) depends on the major. Introduction to Psychology and Technology was a compulsory course for students of the major Psychology and Technology. All courses were blended courses, as the course consisted of four to six hours of face-to-face lectures per week and part of the course was presented online in Moodle LMS. According to the classification of blended learning made by Park et al. (2016), all courses could be classified as sharing and submission courses, as they provided content, assignments, and quizzes. Thus, the courses can be considered quite similar. Most activity in the LMS could be found in the quizzes (47%–85% of the clicks). In Introduction to Psychology and Technology a significant amount of activity could be found in the assignment and content modules. Although a discussion forum was provided in all courses, the usage of the forums was low. An overview of the courses and course characteristics can be found in Table 1. Courseid, quarter (fall or winter), total number of students, and number of face-to-face hours per week were included as course characteristics for the prediction of student performance.

3.2 Moodle LMS data

Data from Moodle LMS came from a previous study (Conijn et al., 2017). Data from the courses in the fall quarter were collected from August 25th 2014 (1 week before the lectures started) until

Table 2: Descriptive statistics LMS data, performance data, and student characteristics.

Variable	N	Min	Max	*M*	*S.D.*
LMS data					
Total number of clicks	888	6	5435	799	759
Number of online sessions	888	1	120	36.4	19.2
Total time online (min)	888	7.1	477	979	644
Number of course page views	888	6	1016	241	123
Irregularity of study time	888	0	5805	2051	824
Irregularity of study interval	888	0	1374418	248459	160224
Largest period of inactivity (min)	888	0	71290	18029	10493
Time until first activity (min)	888	786	45617	14992	4642
Average time per session (min)	888	1.0	68.2	26.9	11.0
Performance data					
In-between assessment grade	888	0	9.8	6.9	1.3
Final exam grade	888	0	10	5.3	2.1
Student characteristics					
GPA prior education	394	5.49	8.70	6.87	0.52
Conscientiousness	426	2.33	5.00	3.77	0.50
Time management	426	1.50	5.00	3.75	0.65
Study strategy (lack of)	426	1.00	5.67	2.14	0.92
Self-efficacy	426	2.78	6.89	4.94	0.66
Connection with study program	426	3.00	7.00	5.55	0.64
Confidence with study choice	426	2.75	7.00	5.57	0.89
Amotivation study choice	426	1.00	4.25	1.49	0.63
External regulation study choice	426	1.00	5.50	2.03	0.94

November 9th 2014 (end of the exam week) and grouped per week, which resulted in 11 weeks of data. Data of the courses in the winter quarter were collected likewise from November 3rd 2014 (1 week before lectures started) until February 1st 2015 (end of the exam week). As the two-week Christmas break fell into the winter quarter, this resulted in a total of 13 weeks of LMS data.

Since the LMS provides raw log data, the data needed to be pre-processed first. The pre-processing was done in R, based on the pre-processing steps as discussed in Romero and Ventura (2007). Four aggregated measures were extracted, as these are often used in the literature: total number of clicks, number of online sessions, the total time online, and the total number of views. The aggregated measures were grouped per week, to indicate the amount of activity up to a specific week in the course. Also, five variables related to study patterns were included: the irregularity of study time (*S.D.* of time per session), the irregularity of study interval (*S.D.* of time between sessions), the largest period of inactivity, the time until the first activity, and the average time per session. A detailed description of these variables can be found in Conijn et al. (2017). These LMS variables were combined with course characteristics, student characteristics, and prior performance data. An overview of the descriptive values of all variables can be found in Table 2. For brevity, the aggregated measures over the whole course are listed, not the aggregated measures for every week.

3.3 Student characteristics

The student characteristics were extracted from an online questionnaire, which was part of the university's Study Choice Check for prospective students of bachelor programs at the university. Data used in the current study came from a pilot of the online questionnaire, distributed in the first half of 2014. This pilot only included prospective bachelor students of the departments of Industrial Engineering & Innovation Sciences and Built Environment, resulting in a strong selectivity of students. An invitation to complete the questionnaires was sent three weeks before the prospective students took part in the on-site orientation activity. When students did not complete the questionnaire before the orientation activity, extra time was provided to complete the questionnaire during the activity. This resulted in a response rate of nearly 100% of the students who participated in the orientation

activity. Based on the online questionnaire, an advice concerning the study choice was given to the prospective students, categorized in 'abilities & skills' and 'motivation for study choice'.

The questionnaire measured demographics and a total of nine factors related to abilities & skills (5) and motivation for the study choice (4). Most of the ability/skills and motivation factors were adapted from validated questionnaires. The factors were found significant predictors in a previous longitudinal study on student performance and study continuation at the department of Industrial Engineering & Innovations Sciences (Bipp and Schinkel, 2013).

Skills and capacities consisted of: GPA of prior education, conscientiousness, time management, lack of study strategy, and self-efficacy. GPA was calculated using the average final grade for all courses in prior education, with a higher weight for the courses that are required to enter the study program (Mathematics for all four Bachelor programs, and in addition Physics for Built Environment). Conscientiousness was measured using the validated Dutch translation of the nine-item conscientiousness scale of the Big Five Inventory (Denissen et al., 2008). A sample item is 'Perseveres until the task is finished'. Time management was measured using four items from Kleijn et al. (1994). A sample item is 'I start on time to prepare for an exam'. Lack of study strategy was measured using the lack of strategy scale developed by Harackiewicz et al. (2000). This scale consists of three questions (e.g. 'I often find that I don't know what to study or where to start') and was translated into Dutch. Self-efficacy was measured using a slightly adapted version of the self-efficacy scale of the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich and De Groot, 1990). This scale consisted of nine questions related to students' perceived competence and confidence in their performance in the program (e.g. 'Compared to other students in this class I expect to do well') and was translated to Dutch. Conscientiousness and time management were measured using a five-point scale, ranging from 1 (almost never) to 5 (almost always). Lack of study strategy and self-efficacy were measured using a seven-point scale, ranging from 1 (completely disagree) to 7 (completely agree).

Motivation for study choice consisted of: connection with study program, confidence study choice, amotivation study choice, and external regulation. The connection with the study program was measured via six questions that were selected and adapted from the Dutch 'Startmonitor', a national annual survey among students who start with their higher education (e.g. 'This program fits well with my interests') (Warps et al., 2009). The confidence with study choice items were developed specifically for the Study Choice Check. This scale consists of four questions (e.g. 'I hesitate between this university and other universities'). The lack of motivation (amotivation) for the study choice was measured using the amotivation items from the Situational Motivation Scale (SIMS) (Guay et al., 2000). From this scale, three questions were adapted and translated into Dutch (e.g. 'There may be good reasons to do this program, but personally I don't see any'). External regulation occurs when a student chooses a study program because of a felt obligation. This was measured using the external regulation items from the Situational Motivation Scale (SIMS) (Guay et al., 2000). This scale consists of four questions which were translated to Dutch (e.g. 'I choose this program because I'm supposed to do it'). All motivational factors were measured using a seven-point scale, ranging from 1 (completely disagree) to 7 (completely agree).

3.4 Performance data

The performance data collected for all 888 cases consisted of final exam grade and in-between assessment grade. All grades range from 0 to 10, where grades ≥ 5.5 indicate that a student passed the specific assignment or course and grades < 5.5 represent a fail. The final exam grades where quite low ($M = 5.31$, $S.D. = 2.10$): the average student failed the course. The in-between assessment grades were substantially higher ($M = 6.93$, $S.D. = 1.33$). In-between assessment grade consisted of the grades for the graded assessments during the course (i.e., entry test, assignments, online homework, offline homework, and midterm exam). These assessments could be completed either online in Moodle LMS or offline and handed-in on paper or via other systems. As the weights and types of in-between assessments differed across courses, the (unweighted) average of these grades were used to calculate the in-between assessment grade. As in-between assessment grades were part of the final course grade in all five courses, we used final exam grade as outcome variable.

3.5 Data analyses

After data pre-processing in R, four classification algorithms were run with WEKA 3.6.13: Baseline (predicting the majority class using the ZeroR classifier), decision tree J48, rule inducer JRip, and an SMO support vector machine. All classifiers were run with their default settings. We realize that parameter optimization may lead to higher performance, but here we are mostly interested in the informativeness of the features. For the classifications, four feature sets were used: course characteristics (CC), in-between assessment grade (Midterm), LMS data (LMS), and student characteristics (SC). All combinations of these feature sets were tested, resulting in a total of 15 combinations of feature sets. Classifications were run on three binary target variables: grade ≥ 5.5 (pass), grade ≤ 3 (particularly bad performance), and grade ≥ 8 (particularly good performance). This resulted in a total of 4 (classification techniques) * 15 (data sets) * 3 (target variables) = 240 models. All models were evaluated using 10-fold cross-validation. Afterwards, a series of two-way ANOVAs on the classification method and feature set with Tukey post-hoc tests were conducted using R to determine whether the accuracy of the classification models significantly differed across the classification techniques and data sets.

4 Results

In this section we investigate the performance of the machine learning classifiers and the influence of the different feature sets in three different settings. Firstly, we consider the prediction of pass versus fail for each student. Secondly, we look at the prediction of particularly weak students (those scoring below or equal to 3) and, thirdly, we concentrate on the prediction of good students (those scoring above or equal to 8).

4.1 Predicting pass/fail

The results of the machine learning classifiers on predicting pass or fail can be found in Table 3. Here we see that approximately half of the students fail the courses (the majority class baseline is approximately 50% divided over two classes).

First of all, we see that all classification methods lead to a higher accuracy. Indeed, a two-way ANOVA showed that the accuracies significantly differ across the methods $F(3,540) = 345.6$, $p < .001$. Moreover, a Tukey post-hoc test showed that J48, JRip, and SMO lead to a significantly higher accuracy than the majority class baseline. Additionally, SMO significantly outperformed J48 and JRip. No significant differences were found between J48 and JRip.

Secondly, looking at the impact of the individual feature sets, we see that all feature sets have a positive impact on the classification results (with the exception of the course characteristics when using the JRip classifier). A two-way ANOVA showed that there is a significant difference in accuracy between the combination of feature sets $F(14,540) = 20.7$, $p < .001$. A post-hoc Tukey test on the single feature sets showed that LMS data, student characteristics, and midterm result in a significant higher accuracy than the course characteristics. Moreover, student characteristics and midterm resulted in a significant higher accuracy than LMS data. Student characteristics and midterm did not lead to significant differences in the accuracy. Considering the combination of feature sets, we see that the combination of student characteristics and midterm grade does not lead to a higher accuracy compared to the individual feature sets. It seems that these feature sets provide similar information.

The best results are found when using the SMO classifier and all feature sets combined. This shows that all feature sets do contain useful information that can improve results. However, comparing the impact of the different feature sets, we see that adding the LMS features in general has little (or sometimes even a negative) effect.

4.2 Predicting particularly bad performance

Next to predicting whether a student will pass or fail a course, it is interesting to predict which students perform particularly bad in a course, as these students will need specific attention to improve their performance to a large extent. Therefore, we determined how well the machine learning classifiers could identify students with particularly bad performance, i.e. students with a final grade equal to or below 3 (out of 10).

Table 3: Accuracy of the machine learning systems on predicting pass/fail with the different feature sets: course characteristics (CC), student characteristics (SC), LMS data, and midterm grade.

				Baseline	J48	JRip	SMO
				M% (*S.D.*)	*M*% (*S.D.*)	*M*% (*S.D.*)	*M*% (*S.D.*)
CC				50.7 (0.24)	54.1 (1.96)	50.5 (3.85)	50.9 (4.60)
	SC			50.7 (0.24)	65.3 (4.56)	63.1 (4.78)	62.5 (6.01)
		LMS		50.7 (0.24)	58.7 (4.11)	57.3 (4.07)	59.4 (5.27)
			Midterm	50.7 (0.24)	64.0 (5.01)	62.7 (4.99)	62.8 (5.15)
CC	SC			50.7 (0.24)	61.9 (2.73)	63.1 (4.54)	64.0 (2.87)
CC		LMS		50.7 (0.24)	57.5 (3.89)	57.3 (4.82)	60.7 (3.18)
CC			Midterm	50.7 (0.24)	67.4 (5.17)	65.8 (4.66)	66.8 (5.86)
	SC	LMS		50.7 (0.24)	61.5 (5.61)	60.9 (2.90)	65.1 (4.49)
	SC		Midterm	50.7 (0.24)	61.9 (3.77)	64.6 (3.90)	66.3 (3.63)
		LMS	Midterm	50.7 (0.24)	63.5 (5.74)	63.6 (4.72)	65.3 (4.97)
CC	SC	LMS		50.7 (0.24)	60.6 (3.30)	62.2 (4.75)	65.4 (2.74)
CC	SC		Midterm	50.7 (0.24)	67.1 (5.94)	66.9 (3.69)	68.5 (4.78)
CC		LMS	Midterm	50.7 (0.24)	61.0 (5.60)	64.4 (5.92)	66.7 (5.96)
	SC	LMS	Midterm	50.7 (0.24)	60.2 (5.48)	63.1 (5.54)	67.5 (3.97)
CC	SC	LMS	Midterm	50.7 (0.24)	61.5 (3.94)	61.7 (3.97)	68.7 (3.60)

Table 4: Accuracy of the machine learning systems on particularly bad performance with the different feature sets: course characteristics (CC), student characteristics (SC), LMS data, and midterm grade.

				Baseline	J48	JRip	SMO
				M% (*S.D.*)	*M*% (*S.D.*)	*M*% (*S.D.*)	*M*% (*S.D.*)
CC				86.6 (0.32)	86.6 (0.32)	86.6 (0.32)	86.6 (0.32)
	SC			86.6 (0.32)	86.6 (0.32)	86.5 (0.06)	86.6 (0.32)
		LMS		86.6 (0.32)	84.6 (2.69)	86.0 (1.98)	86.6 (0.32)
			Midterm	86.6 (0.32)	87.9 (3.05)	87.9 (3.05)	86.6 (0.32)
CC	SC			86.6 (0.32)	86.6 (0.32)	86.6 (1.51)	86.6 (0.32)
CC		LMS		86.6 (0.32)	84.0 (3.24)	86.8 (2.68)	86.6 (0.32)
CC			Midterm	86.6 (0.32)	87.1 (2.19)	87.6 (2.25)	86.6 (0.32)
	SC	LMS		86.6 (0.32)	81.9 (2.91)	86.7 (3.16)	86.6 (0.32)
	SC		Midterm	86.6 (0.32)	87.1 (2.04)	86.7 (3.08)	86.6 (0.32)
		LMS	Midterm	86.6 (0.32)	83.8 (2.28)	86.7 (3.46)	86.8 (0.73)
CC	SC	LMS		86.6 (0.32)	83.8 (2.14)	86.9 (2.91)	86.6 (0.32)
CC	SC		Midterm	86.6 (0.32)	87.2 (1.30)	86.6 (3.29)	86.6 (0.32)
CC		LMS	Midterm	86.6 (0.32)	85.1 (3.24)	86.5 (3.04)	87.3 (1.18)
	SC	LMS	Midterm	86.6 (0.32)	84.6 (1.89)	86.4 (3.27)	86.6 (1.51)
CC	SC	LMS	Midterm	86.6 (0.32)	85.0 (2.27)	87.5 (3.18)	87.3 (1.64)

The results (see Table 4) show that most classifiers do not exceed the accuracy of the majority class baseline. This is probably due to the unbalanced classes as only 14.4% of the students got a grade ≤ 3. J48 and JRip with a feature set including midterm grade occasionally show a better accuracy. SMO with course characteristics, LMS data, and midterm or all features also leads to a better result. A two-way ANOVA showed that the accuracy indeed significantly differed across the classifiers and feature sets, $F(42,540) = 2.37$, $p < .001$. However, a post-hoc Tukey test showed that none of the classifiers had a significantly higher accuracy than the majority class baseline. Thus, there seems to be too little information in the feature sets to better identify the weak students.

4.3 Predicting particularly good performance

A small number of students show particularly good performance, the so-called 'excellent' students. These students may be treated differently by a lecturer, for example by providing additional (advanced) materials. As such, it would be interesting to be able to identify these students relatively early in the course. Therefore, we also determine how well the machine learning classifiers are at identifying students with grades equal to or higher than 8 (out of 10). The results of these classifications can be found in Table 5.

Table 5: Accuracy of the machine learning systems on particularly good performance with the different feature sets: course characteristics (CC), student characteristics (SC), LMS data, and midterm grade.

				Baseline	J48	JRip	SMO
				M% (*S.D.*)	*M*% (*S.D.*)	*M*% (*S.D.*)	*M*% (*S.D.*)
CC				90.7 (0.53)	90.7 (0.53)	90.7 (0.53)	90.7 (0.53)
	SC			90.7 (0.53)	90.7 (0.53)	90.9 (1.10)	90.7 (0.53)
		LMS		90.7 (0.53)	90.3 (1.07)	90.2 (1.45)	90.7 (0.53)
			Midterm	90.7 (0.53)	90.3 (1.07)	90.1 (1.01)	90.7 (0.53)
CC	SC			90.7 (0.53)	90.7 (1.06)	89.8 (3.19)	90.7 (0.53)
CC		LMS		90.7 (0.53)	89.3 (1.60)	90.7 (0.53)	90.7 (0.53)
CC			Midterm	90.7 (0.53)	90.7 (0.53)	89.9 (1.67)	90.7 (0.53)
	SC	LMS		90.7 (0.53)	87.8 (2.46)	90.3 (2.37)	90.7 (0.53)
	SC		Midterm	90.7 (0.53)	90.2 (2.04)	90.8 (2.09)	90.7 (0.53)
		LMS	Midterm	90.7 (0.53)	88.7 (2.09)	89.6 (1.65)	90.7 (0.53)
CC	SC	LMS		90.7 (0.53)	88.4 (1.73)	90.1 (1.13)	90.7 (0.53)
CC	SC		Midterm	90.7 (0.53)	87.8 (1.94)	90.1 (1.87)	90.7 (0.53)
CC		LMS	Midterm	90.7 (0.53)	89.0 (1.87)	90.5 (1.20)	90.7 (0.53)
	SC	LMS	Midterm	90.7 (0.53)	89.1 (2.36)	89.8 (1.86)	90.7 (0.53)
CC	SC	LMS	Midterm	90.7 (0.53)	88.3 (1.90)	90.3 (1.07)	90.7 (0.53)

The majority class baseline system shows that the two classes are quite unbalanced. Only a small number of students (83 out of 888 students) received a grade higher than 8. Based on the majority class already high performance is achieved, so the feature sets have to be highly informative to lead to better results. Only the JRip system sporadically leads to a higher accuracy (with student characteristics or student characteristics and midterm grade). Indeed, a two-way ANOVA showed that the accuracy significantly differed across the classifiers and feature sets, $F(42,540) = 2.11$, $p < .001$. However, a post-hoc Tukey test showed that JRip with student characteristics and midterm, or student characteristics only did not lead to a significant higher accuracy than the baseline. Overall, the feature sets do not seem to contain enough information to better identify the good students.

4.4 Interpretability

Although the SMO resulted in the best classifier for prediction pass/fail, it is not an interpretable technique. The decision tree and rule-based classifiers result in better interpretable results. Hence, especially for teachers, these classifiers would be preferable, as these can provide more insight in which variables are especially useful in the prediction of student performance. As the J48 decision tree resulted in a tree with a high depth, this is still hard too read. Therefore, we only report the gain ratios here. Additionally, we provide one of the rules made by JRip.

The gain ratios for the pass/fail classification with all feature sets showed that the number of course page views, and especially the total amount of course page views (over the whole course or up to the last two weeks before the exam) had the highest gain ratios (0.0781 for views up to week 7, 0.0882 for views up to week 8, and 0.0802 for the total course page views). These were followed by the number of clicks and the number of sessions over the whole course or up to the last two weeks before the exam). Additionally, GPA prior performance, midterm grade, and total the number of students in the course were also found important for the pass/fail prediction (gain ratios 0.06322, 0.0587, and 0.0516 respectively). Thus, the aggregated LMS metrics and prior performance result in the highest gain ratios, while study patterns in the LMS, course characteristics (except for the total amount of students) and student characteristics (except for prior GPA) seem to have almost no value in the prediction.

The JRip rule for pass/fail classification (Table 6) mainly consists of midterm grade, the number of clicks and sessions (especially for the whole course, or up to the last weeks before the exam), and prior GPA. This shows that these variables are especially important for the pass/fail prediction. Specifically, high midterm grade (> 7.3) and high prior GPA (> 7.0175) result in a higher probability of passing the course. Course characteristics do not seem to play a role in the rules. Only the number of students in the course is used once in the rules. This indicates that the course characteristics in our feature set are less important for pass/fail prediction than the other variables.

Table 6: JRip rule for pass/fail classification using all feature sets

1 *(midterm ≤ 7) and (clicks_upto_week7 ≥ 298) and (sessions_upto_week7 ≤ 38) and (sessions_total ≤ 34) → pass = false (49.0 instances / 2.0 incorrect)*

2 *(prior_gpa ≤ 6.925) and (midterm ≤ 7.25) and (midterm ≤ 4.966667) → pass = false (39.0 instances / 3.0 incorrect)*

3 *(midterm ≤ 7.25) and (clicks_total ≥ 329) and (midterm ≤ 6.3) → pass = false (61.0 instances / 13.0 incorrect)*

4 *(prior_gpa ≤ 7.0175) and (sessions_upto_week3 ≤ 12) and (courseviews_upto_week8 ≥ 174) and (timeonline_upto_week8 ≤ 43323) → pass = false (57.0 instances / 14.0 incorrect)*

5 *(midterm ≤ 7.3) and (major = Industrial Engineering) and (time_unil_first_act ≤ 825009) and (courseviews_upto_week3 ≤ 88) → pass = false (28.0 instances / 2.0 incorrect)*

6 *(prior_gpa ≤ 7.0175) and (courseviews_upto_week1 ≤ 20) and (connection_program ≥ 6.166667) and (external_regulation ≥ 1.25) → pass = false (25.0 instances / 2.0 incorrect)*

7 *(prior_gpa ≤ 7.0175) and (prior_gpa ≤ 6.275) and (gender = female) → pass = false (50.0 instances / 16.0 incorrect)*

8 *(midterm ≤ 7.3) and (nofstudents ≥ 836) and (irregularity_studytime ≤ 1723.780407) and (amotivation ≤ 2) → pass = false (19.0 instances / 1.0 incorrect)*

9 *(midterm ≤ 7.3) and (timeonline_upto_week1 ≥ 2784) and (confidence_choice ≥ 6.25) → pass = false (40.0 instances / 14.0 incorrect)*

10 *(prior_gpa ≤ 7.0175) and (clicks_upto_week8 ≥ 1082) and (sessions_upto_week2 ≤ 9) → pass = false (36.0 instances / 11.0 incorrect)*

11 *else → pass = true (484.0 instances / 112.0 incorrect)*

The first part of the rule shows that when you have a midterm grade below 7, more than 298 clicks and less than 38 sessions up to two weeks before the exam, and not more than 34 sessions in total, you will have a high probability of failing the course. This may indicate that it is better to have multiple sessions with less clicks instead of fewer sessions with more clicks. Parts 2 and 5 to 9 show that for part of the instances an accurate prediction could be made with the information available up to the midterm, instead of information over the whole course, which makes early intervention more feasible. Interestingly, when you have a midterm grade below 7.3, spent at least 46 minutes online in week 1, and when you have a high confidence in the study choice, you have a higher change to fail the course. This shows that high confidence in the study choice and starting early on in the course, which might be related to high motivation, does not necessarily result in passing the course.

5 Discussion

The aim of this study was to determine the influence of course characteristics, student characteristics, past performance, and behavioral data from learning management systems on the prediction of student performance. Three classifiers (decision tree, rule-based, and SVM) were trained and compared to the majority class baseline on predicting pass/fail, particularly high grades, and particularly low grades.

For the pass/fail predictions it was found that all feature sets have a positive influence on predicting whether a student will pass or fail the final exam. For the individual feature sets, especially student characteristics and in-between (midterm) assessment grade have a large influence. For the combination of feature sets, it seems that LMS data has little added value next to student characteristics and in-between assessment grade. These findings corroborate with the results of Tempelaar et al. (2015) who also found that LMS data are of limited value compared to student characteristics (learner dispositions) and prior performance data. This can be due to the fact that student characteristics and prior performance data are specific measures of (theoretical) concepts, while the LMS metrics currently used (aggregated measures of raw log data) are not.

Across all combinations of feature sets, it was found that all three classifiers significantly exceeded the majority baseline accuracy. No significant differences were found in performance between the decision tree and rule-based algorithm. The SVM algorithm resulted in the highest accuracy across

all classifiers, as expected. SMO with all features resulted in the highest accuracy (68.7%). These results differ from Kotsiantis and Pintelas (2005) who found that a rule-based algorithm (M5rules) outperformed SVM (SMO) in the prediction of student performance using student characteristics. The decision tree and rule-based algorithm only outperformed SMO when only midterm grades were included. These findings corroborate with Cortez and Silva (2008) who found that SVM only outperformed the decision tree when midterm grades were not included.

As SVM does not result in interpretable results, the decision tree and rule-based algorithm can be considered more useful (c.f. Romero et al., 2013). Hence, for more interpretative results, J48 and JRip with student characteristics, course characteristics, and midterm grade are a good second best, with accuracies of 67.1% and 66.9%, respectively.

Additional analysis of the gain ratios provided by the decision tree and one rule made by JRip resulted in more interpretable results about the importance of the individual features for the pass/fail prediction. It appeared that the in-between assessment grade had a high value in the prediction. This in line with previous findings that past performance is an important and robust predictor for student performance (Dollinger et al., 2008; Hattie, 2008). Although student characteristics were one of the most useful feature sets for pass/fail prediction, the additional analyses showed this was mostly due to prior GPA. Other student characteristics, course characteristics, and the study pattern metrics from the LMS data were of little importance for the prediction. The JRip rule showed that some of the students could already be classified early in the course, after the midterm grade. This is especially useful, as early classification can result in personalized interventions at a point in time where it is still effective for the current course (Campbell et al., 2007).

For the prediction of particularly good and particularly bad performance, it was found that the classifiers did not result in a significant higher accuracy compared to the majority baseline. This could be expected, as the classes were highly unbalanced and the majority class already resulted in high performance. Therefore, the feature sets needed to be highly informative to lead to better results. Thus, the current feature sets seem not informative enough to identify these specific small groups of students who perform particularly bad or good.

5.1 Limitations and future work

In the current study, the used feature sets had some limitations. Student characteristics were only available for a sub-sample of students within the five courses. Hence, the findings cannot be generalized to whole course. Additionally, the sample consisted of only five courses, which resulted in a low variety of course characteristics. Hence, it is hard to draw conclusions about the influence of the course characteristics on student performance. Accordingly, future work should include more courses with a higher variety in course characteristics to determine the effect of course characteristics for predicting student performance.

Although the current study found that LMS data were not really useful next to prior performance data and student characteristics, this might be due to fact that the LMS metrics were not specific measurements of concepts. To improve the pass/fail prediction, educational theory needs to be included to convert LMS data in concrete measurements of concepts (Conijn et al., 2017, c.f). Moreover, the current LMS metrics (e.g. total number of clicks) are too raw to provide insight in how students learn, which is one of the aims of educational data mining (Romero et al., 2013). Addition of theory in the extraction of metrics from LMS data could also provide more insight in the learning processes.

Additionally, the classification methods resulted in some limitations as well. We only used three classifiers, while other classifiers, parameter optimization, or ensemble methods could have resulted in higher accuracy. Therefore, future work should also include other methods, to determine whether these methods indeed result in a higher accuracy, and how the accuracies differ across the combinations of feature sets. Lastly, we only performed binary classifications, while it might be more useful to include multi-class classification, to differentiate between more classes of students. For example, students with grades below 4.5, between 4.5 and 5.5 (barely fail), between 5.5 and 6.5 (barely pass) and above 6.5 might need different feedback. Therefore, future work should focus on multiple class prediction as well.

6 Conclusion

To conclude, this study showed that course characteristics, student characteristics, LMS data, and past performance are all useful for the prediction of student performance. However, compared to information found in student characteristics and past performance, LMS data has less value. Hence, it is important to include student characteristics and past performance for the prediction of student performance. Yet, a more fine-grained analysis of the metrics found in the data from learning management systems may still lead to useful information.

References

Bipp, T., K. A. and S. Schinkel
2013. Bachelor entrance study (BEST), onderzoek naar studiesucces en drop-out binnen de bachelor opleidingen aan de faculteit industrial engineering & innovation sciences aan de TU/e. Technical report, Eindhoven University of Technology.

Britton, B. K. and A. Tesser
1991. Effects of time-management practices on college grades. *Journal of educational psychology*, 83(3):405.

Campbell, J. P., D. G. Oblinger, et al.
2007. Academic analytics. *Educause Quarterly*, Pp. 1–20.

Clow, D.
2013. An overview of learning analytics. *Teaching in Higher Education*, 18(6):683–695.

Conard, M. A.
2006. Aptitude is not enough: How personality and behavior predict academic performance. *Journal of Research in Personality*, 40(3):339–346.

Conijn, R., C. Snijders, A. Kleingeld, and U. Matzat
2017. Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. (in press).

Cortez, P. and A. M. G. Silva
2008. Using data mining to predict secondary school student performance. In *Proceedings of 5th FUture BUsiness TEChnology Conference*, Pp. 5–12. EUROSIS.

Denissen, J. J., R. Geenen, M. A. Van Aken, S. D. Gosling, and J. Potter
2008. Development and validation of a dutch translation of the big five inventory (bfi). *Journal of personality assessment*, 90(2):152–157.

Dollinger, S. J., A. M. Matyja, and J. L. Huber
2008. Which factors best account for academic success: Those which college students can control or those they cannot? *Journal of research in Personality*, 42(4):872–885.

Guay, F., R. J. Vallerand, and C. Blanchard
2000. On the assessment of situational intrinsic and extrinsic motivation: The situational motivation scale (sims). *Motivation and emotion*, 24(3):175–213.

Harackiewicz, J. M., K. E. Barron, J. M. Tauer, S. M. Carter, and A. J. Elliot
2000. Short-term and long-term consequences of achievement goals: Predicting interest and performance over time. *Journal of educational psychology*, 92(2):316.

Hattie, J.
2008. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hu, Y.-H., C.-L. Lo, and S.-P. Shih
2014. Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*, 36:469–478.

Kleijn, W., R. Topman, and H. Ploeg
1994. Cognities, studiegewoonten en academische prestaties: de ontwikkeling van de studie management en academische resultaten test (smart). *Ned Tijdschr Psychol*, 49:233–234.

Kotsiantis, S. B. and P. E. Pintelas
2005. Predicting students marks in hellenic open university. In *Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05)*, Pp. 664–668. IEEE.

Macfadyen, L. P. and S. Dawson
2010. Mining lms data to develop an "early warning system" for educators: A proof of concept. *Computers & education*, 54(2):588–599.

Minaei-Bidgoli, B. and W. F. Punch
2003. Using genetic algorithms for data mining optimization in an educational web-based system. In *Genetic and evolutionary computation conference*, Pp. 2252–2263. Springer.

Morris, L. V., C. Finnegan, and S.-S. Wu
2005. Tracking student behavior, persistence, and achievement in online courses. *The Internet and Higher Education*, 8(3):221–231.

O'Connor, M. C. and S. V. Paunonen
2007. Big five personality predictors of post-secondary academic performance. *Personality and Individual differences*, 43(5):971–990.

Park, Y., J. H. Yu, and I.-H. Jo
2016. Clustering blended learning courses by online behavior data: A case study in a korean higher education institute. *The Internet and Higher Education*, 29:1–11.

Piña, A. A.
2012. An overview of learning management systems. *Virtual Learning Environments: Concepts, Methodologies, Tools and Applications. USA: IGI Global*, Pp. 33–51.

Pintrich, P. R. and E. V. De Groot
1990. Motivational and self-regulated learning components of classroom academic performance. *Journal of educational psychology*, 82(1):33.

Retalis, S., A. Papasalouros, Y. Psaromiligkos, S. Siscos, and T. Kargidis
2006. Towards networked learning analytics—a concept and a tool. In *Proceedings of the fifth international conference on networked learning*.

Romero, C., P. G. Espejo, A. Zafra, J. R. Romero, and S. Ventura
2013. Web usage mining for predicting final marks of students that use moodle courses. *Computer Applications in Engineering Education*, 21(1):135–146.

Romero, C. and S. Ventura
2007. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135–146.

Romero, C. and S. Ventura
2010. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618.

Romero, C. and S. Ventura
2013. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27.

Siemens, G.
2011. Learning analytics and educational data announcing open course: Learning and knowledge analytics. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*.

Siemens, G. and R. S. Baker
2012. Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, Pp. 252–254. ACM.

Superby, J.-F., J. Vandamme, and N. Meskens
2006. Determination of factors influencing the achievement of the first-year university students using data mining methods. In *Workshop on Educational Data Mining*, volume 32, P. 234. Citeseer.

Tempelaar, D. T., B. Rienties, and B. Giesbers
2015. In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior*, 47:157–167.

Warps, J., L. Hogeling, J. Pass, and D. Brukx
2009. Studiekeuze en studiesucces. *Een selectie van gegevens uit de Startmonitor over studiekeuze, studieuitval en studiesucces in het hoger onderwijs*.

Zacharis, N. Z.
2015. A multivariate approach to predicting student outcomes in web-enabled blended learning courses. *The Internet and Higher Education*, 27:44–53.