Computers and Education: Artificial Intelligence 8 (2025) 100369

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Evaluation of early student performance prediction given concept drift

Benedikt Sonnleitner [a,b,*], Tom Madou [a], Matthias Deceuninck [a], Filotas Theodosiou [a], Yves R. Sagaert [a]

[a] VIVES University of Applied Sciences, Doorniksesteenweg 145, 8500 Kortrijk, Belgium
[b] Fraunhofer IIS, Nordostpark 84, 90411 Nuremberg, Germany

ARTICLE INFO

Keywords:
Data science applications in education
Distance education and online learning

ABSTRACT

Forecasting student performance can help to identify students at risk and aids in recommending actions to improve their learning outcomes. That often involves elaborate machine learning pipelines. These tend to use large feature sets including behavioral data from learning management systems or demographic information. However, this complexity can lead to inaccurate predictions when concept drift occurs, or when a large number of features are used with a limited sample size. We investigate the performance of different machine learning pipelines on a data set with change in study behavior during the Covid-19 period. We demonstrate that (i) LASSO, a shrinkage estimator that reduces complexity and overfitting, outperforms several machine learning models under these circumstances, (ii) a linear regression relying on only two handcrafted features achieves higher accuracy and substantially less predictive bias than commonly used, more complex models with large feature sets. Due to their simplicity, these models can serve as a benchmark for future studies and a fallback model when substantial concept or covariate drift is encountered.

1. Introduction

The use of educational technology (EdTech) such as Learning Management systems (LMS) in higher education has increased substantially over the past decades (Dziuban et al., 2018). This increase provides educational institutions with extensive data related to their learners. In addition to student background data, (e.g. course enrollment, prior degrees) learning behavior such as online interactions with EdTech is now being recorded along learning outcomes (i.e. course grades). Educational Data Mining (EDM) develops methods to explore such data and use it to better understand students and the educational settings they learn in (Baker & Yacef, 2009; Dutt et al., 2017). One of the main goals of EDM is to assist teachers and students in education (Dutt et al., 2017). Despite a consensus that EDM has the potential to improve teaching and learning (Ifenthaler & Yau, 2020), there is little evidence on a widely deployed use of educational data for this purpose (Viberg et al., 2018; Tsai et al., 2020).

Predicting student success, a key focus of EDM, has received large attention in the last decade (Balaji et al., 2021; Namoun & Alshanqiti, 2020; Xu et al., 2017). Typically student profiles (e.g. parents' education, prior degrees, etc.) and trace data from EdTech (e.g. online learning environments) are used to predict students' learning results (e.g. pass/fail or course grades on a continuous scale) (Gašević et al., 2016; Namoun & Alshanqiti, 2020). One application is to identify at-risk students early for timely intervention (Xu et al., 2017). This is important as the growing use of EdTech can lead to larger student groups per teacher, potentially diminishing the ability to provide personalized attention to each student (Laurillard, 2013). Meaningful student interventions require early, accurate and explainable predictions (Meier et al., 2016; Namoun & Alshanqiti, 2020) which is challenging due to limited data available, variation in academic assignments and exams, and diverse student backgrounds (Meier et al., 2016).

All these factors change over time, and accordingly can lead to covariate drift where the distribution of the features (or covariates) that are used to predict student performance changes, and concept drift where the relation between forecast target and the predictive features changes (Tiukhova et al., 2024; Conijn et al., 2017).

This again can lead to poor performance of predictive models and potentially systematically biased predictions (Bayram et al., 2022). Nevertheless student performance prediction often entails large behavioral feature sets and hard to explain machine learning models, that are specifically vulnerable to concept and covariate drift, due to the sheer amount of potentially drifting feature-target relations. We show that un-

* Corresponding author at: VIVES University of Applied Sciences, Doorniksesteenweg 145, 8500 Kortrijk, Belgium.
E-mail address: benedikt.sonnleitner@posteo.de (B. Sonnleitner).

https://doi.org/10.1016/j.caeai.2025.100369
Received 15 July 2024; Received in revised form 10 January 2025; Accepted 15 January 2025
Available online 23 January 2025

der data and concept drift, simple and sparse prediction methods can be superior to these elaborate prediction methods.

Our research has four main contributions: (i) we provide a descriptive analysis of collected data on freshmen students in nursing and find that their learning behavior shifted substantially between academic years. The changing teaching environment during the Covid 19 pandemic may have contributed to this shift. We show that this negatively biases the predictions of different models. (ii) Using LASSO feature selection and a linear regression analysis, we assess the predictive power of various features proposed in literature and crafted based on our own experience to predict student success. These features include background data on students (e.g. prior degrees), the results of the Motivated Strategies for Learning Questionnaire (MSLQ) (i.e. a validated instrument to assess learning strategies and motivation in students) and higher level features generated from trace data on the learning platform that was used. We find that only a small subset of these relate significantly to student success. (iii) We demonstrate that a simple linear benchmark model using only the midterm trial exam and whether the respective student has a general education degree performs better than several elaborate machine learning pipelines. (iv) We discuss the actionability of our results to improve education.

Section 2 reviews related works. Section 3 describes the methodology and framework. Section 4 presents the results and discussion. We conclude in Section 5.

2. Background

One of the most important applications of EDM and Learning Analytics is to develop predictive models for students' success. These models can be used in early warning systems (EWS), to provide at-risk students with recommendations to improve their learning strategies or to gain insights into the positive or negative effects of certain study behaviors (Bernacki et al., 2020; Jovanović et al., 2021). Typically, these predictive models are built using large feature sets and more or less transparent and complex prediction methods (for example Howard et al., 2018; Tiukhova et al., 2024; Riestra-González et al., 2021). The features and models used should thereby reflect the targeted use-case: if the goal is simply to identify students at risk in an EWS, the prediction itself triggers an action, and it foremost needs to be early and accurate, so that students can still act on it. However, if causal insights are the motivation, the actionability of a specific feature depends on the specific educational aspect that one wants to improve. For example, age or prior degrees might be strong predictors of course success, and accordingly improve predictive accuracy for EWS, however, they do not allow for recommendations on better study behavior. This would require features that directly relate to study behavior, enabling actionable insights that students can use to improve their learning behavior (Bernacki et al., 2020).

While few Learning Analytics Dashboards currently integrate predictive analytics, the trend is growing (Ramaswami et al., 2023). Positive effects are often attributed to interventions triggered by early warnings from predictive models, which can lead to self-reflection and positive behavioral changes in students. For example Bernacki et al. (2020) report that students predicted to perform poorly that received and accessed an early warning outperformed their peers on a similar trajectory that did not access or receive the early warning. However, many dashboards lack detailed insight and transparency into the predictions, hindering the ability to offer tailored interventions (Ramaswami et al., 2023).

2.1. Predictors for student success

Among the most common used features to predict students' success is behavioral data retrieved from logged variables from LMS (Ifenthaler & Yau, 2020). These can for example comprise the consistency and frequency of the student's learning behavior based on the number of log ins. Whether these hold actually predictive value is however unclear: Bernacki et al. (2020) found that an early warning system, built solely on behavioral data such as access to the different study resources helped at-risk students to improve their outcome, and Jovanović et al. (2021) found several aggregates of activity logs from an LMS being significant in a regression analysis. Contrary Conijn et al. (2017) and Tempelaar et al. (2016) found that these hold only limited predictive value and that they are not stable across different courses in different disciplines. Log data is often used to assess student self-regulation, which is related to student success (You, 2016). In our study, we construct several behavioral features from log data on which we elaborate in Section 3.2. Additionally, we assess the motivation and learning strategies of students, by the validated Motivated Strategies for Learning Questionnaire (MSLQ). This allows us to capture factors in the study behavior of students that might influence student performance but are not reflected in the log data. For example Wu (2021) find that self-reported help seeking tendency correlates positively with student success and self-reported procrastination tendency correlates negatively with student success.

Additionally, the usage of discussion forums has been reported as a valuable predictor in several other studies (Morris et al., 2005; Chen & Cui, 2020; López-Zambrano et al., 2020; Romero et al., 2013) while others, such as Lauría et al. (2012) have reported only weak correlations.

In addition, the assessments that occur during the course are frequently reported to be strong predictors (Macfadyen & Dawson, 2010; Tempelaar et al., 2016; Howard et al., 2018; Riestra-González et al., 2021). Depending on the course setup, these can exhibit predictive power, simply because they count into the overall course results, as for example in Riestra-González et al. (2021). However, by nature in these cases this already limits the opportunities to act upon a prediction, as part of the bad grade is already fixed at the time of prediction (Bernacki et al., 2020). We include voluntary trial exam data in our modeling as it is conducted early enough (halfway the semester), and is not included in the overall grade at the end of the semester.

Including all the aforementioned features leads to relatively big feature sets that are often evaluated on a limited number of observations (for example Howard et al., 2018; Romero et al., 2013; Costa et al., 2017). Although this lies to some extent in the nature of the use case, since student consent might be required for the analysis, and researchers often only have access to data from a limited amount of courses, it poses specific challenges in modeling. Features might be found significant in regression analysis by chance: the calculation of p-values is based on the quality of the estimation of model coefficients, which is dependent on sample size. Exploratory models that have a low ratio of observations to parameters carry the risk of weak estimation that can lead to spurious p-values (Hastie et al., 2008; Wasserstein & Lazar, 2016).

Additionally, this set up can lead to overfitting on the training data, as the chance of spurious correlations is high when many features on few observations are present (Lever et al., 2016). In the education analytics literature this is for example reflected in the findings of Howard et al. (2018), where principal component regression (PCR) outperforms all benchmarks. PCR is a method from statistics, that first compresses the initial large feature set into few features (components) by projecting the initial features onto fewer new ones such that most of the variance of the original features is kept. Then a linear regression is fit onto these new artificial features. This potentially looses a part of the information present in the initial feature set, but makes the prediction model more robust due to fewer used features (Stock & Watson, 2004). Additionally, Costa et al. (2017) finds that a data preparation pipeline that includes a non-specified feature selection method improves predictions over modeling on the raw data.

2.2. Predictive models

Predictive models used to predict student performance include Adaptive Boosting, Bayesian Additive Regressive Trees, Bayesian Network, Decision Trees (J48), Gradient Boosting Trees, K-Nearest-Neighbour, Linear Regression, Logistic Regression, Naïve Bayes, Neural Networks

(MLP), Random Forests, Support Vector Machines, and XGBoost (Namoun & Alshanqiti, 2020; López-Zambrano et al., 2021; Alwarthan et al., 2022). These prediction methods are typical for tabular data and are applied across various domains and discussed in textbooks on machine learning and statistics (for example Hastie et al., 2008). XGBoost and Gradient Boosting Trees tend to outperform neural networks on medium-sized data sets with roughly 10.000 observations, (Grinsztajn et al., 2022). The use of plain decision trees is uncommon in other domains because these are weak learners and generally achieve less performance compared to their more advanced counterparts, such as Random Forest, XGBoost, and Gradient Boosting Trees (Breiman, 2001; Chen & Guestrin, 2016; Grinsztajn et al., 2022). Given the often limited available data set size for studies in education analytics, we suspect that the performance of predictive algorithms lies more in finding simple, robust patterns rather than complex ones, as the latter also increases the risk of overfitting. For linear models this can be addressed by shrinkage-estimators like LASSO and RIDGE regression, where the sum of the magnitudes or sum of the squares of the coefficients are penalized (p.68 ff Hastie et al., 2008), and accordingly the model focuses only on very prevalent patterns and is more robust to noise.

2.3. Concept and covariate drift

Change in context - and accordingly in data - between training phase and prediction phase can lead to deteriorating performance of predictive models (Bayram et al., 2022). This change can be split in covariate drift, where the distribution of the features (also refereed to as covariates) changes, and concept drift where the relation between features and target change. Covariance drift is typically detected by some test statistic that measures the difference between the distribution of the features at training and prediction phase (Gama et al., 2014). Concept drift is typically detected by a significant change of the prediction errors.

In student performance prediction the context changes from course to course and year to year (Tiukhova et al., 2024; Conijn et al., 2017). Accordingly covariance and concept drift is likely to occur. Covariance drift can be detected by assessing the features of the students in the courses that was learned on, versus the courses that need to be predicted. However, the detection of concept drift using the prediction errors is of limited help in the discussed use case since we observe the prediction errors of all students in a course only at the end of this course, and so no adaptation to the specific course is possible. This is in contrast to systems with continuous data streams, where the model can be gradually adapted to the drift (Gama et al., 2014). Instead we hypothesize that simple models with few features are less prone to concept drift, simply because they rely on less feature-target relations that might drift. Illustratively: if behavioral (LMS) data is not included in the predictive model, the concept drift between student behavior in the LMS data and student performance is not relevant to the model.

3. Methodology

We analyze the causal effect of behavioral features derived from interactions with the learning management system through regression analysis, controlling for previous education of the students. Additionally, we focus on student success predictions for EWS, and accordingly evaluate accuracy of predictions halfway through the semester, early enough for students to respond to the warnings.

Our methodology is outlined below. In Subsection 3.1, we describe the investigated course and the collected data, then in Subsection 3.2, we describe the constructed features. In Section 3.3 we outline the predictive models used.

3.1. Participants and setting

We collected data from an introductory physiology course designed for freshmen students in nursing over a period of three consecutive academic years (AY): comprising a training set spanning AY 2020-2021 and AY 2021-2022, and a test set from AY 2022-2023. We refer to these periods as AY 20-21, AY 21-22 and AY 22-23 respectively. The synchronous classes during AY 20-21 and AY 21-22 primarily occurred online due to public health restrictions during the Covid-19 pandemic. This means we train on data collected during the Covid-Period to predict post Covid-Period data. We point out the implications of this in Section 4. The course used a blended learning method, integrating a majority (> 90%) of asynchronous online activities with limited synchronous (i.e. 'traditional') face-to-face teaching. The online learning content was organized within a LMS (i.e. Moodle) and consisted of instructional video clips, quizzes and moderated discussion forums. The course implemented a student centered asynchronous learning approach (Bossant et al., 2022), providing students with the flexibility to progress through the material at their own pace. Students had access to all learning materials and were able to engage with course content according to their individual schedules and their preferred sequence. In addition to the quizzes per instructional video clip, the students had the opportunity to partake in a mid-semester trial exam. Throughout the semester a limited number of synchronous face-to-face sessions were organized allowing students to ask questions. No new content was introduced during these sessions. Data of students that dropped out of the course or did not give their consent was excluded from the study, ensuring the integrity and ethical compliance. Table 1 provides key descriptive statistics pertaining to the final grades for each academic semester.

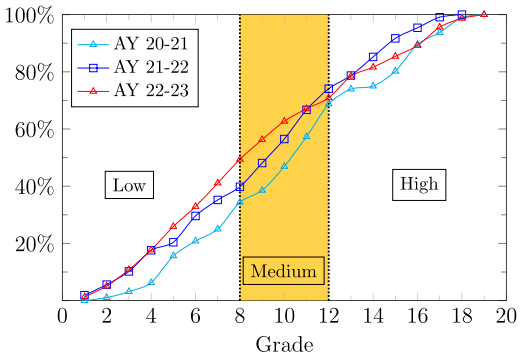

Fig. 1. Cumulative distribution of students' grades. A student is considered 'Medium' when their grade is between 8 and 12.

In Fig. 1 the cumulative distribution of the students' final grades is shown. Despite passing the course being the most important result, the grades are split into three levels: 'Low' (0-7), 'Medium' (8-12) and 'High' (13-20). We chose the thresholds of 8 and 13 because they represent the tolerable fail and distinction mark in Belgium, where the data was collected. A tolerable fail means that students can progress in the curriculum despite failing the course according to government regulations. As can be seen on the graph, the three grade categories -low, medium, and high- are relatively balanced with similar numbers of students in each. However, performance varies substantially over the years. For instance, in AY 22-23 (post Covid-period), there was a notable increase of students who failed the course. Additionally, the proportion of students scoring between 8 and 12 was lower, accounting for only 22%, compared to 35% during the Covid-period (AY 20-21 and AY 21-22).

3.2. Predictive features

At the start of the semester, we conducted a comprehensive survey to gather information about students' demographic and educational background, including factors such as their highest attained degree and the type of Secondary Education they completed.

In addition, participants completed the Motivated Strategies for Learning Questionnaire (MSLQ Pintrich, 1991). This questionnaire is a self-report instrument using a 7-point Likert scale that provides a framework for assessing students' motivations and learning strategies. In line

Table 1
Overview of results. Only results from students that gave consent were included in the study.

Period	# students	Passed exam	Average grade	Std. Dev. (grade)	Median grade
AY 20-21	96	61.5%	10.71	4.3	11
AY 21-22	108	51.9%	9.44	4.3	10
AY 22-23	158	43.7%	9.30	4.8	9

with Bossant et al. (2022) the following constructs were assessed using 50 questions from a validated translation of the MSLQ:

(a) Rehearsal
(b) Elaboration
(c) Organization
(d) Critical Thinking
(e) Self Regulation
(f) Peer Learning
(g) Effort
(h) Environment
(i) Help Seeking

Students responses were converted into numerical features by computing the mean for each category (Bossant et al., 2022). We merged these survey-derived features (i.e. student demographics and MSLQ) with data extracted from the LMS. For this purpose we converted the raw LMS event data into meaningful features on which we elaborate in the following. The cleaning process for the LMS log data addressed several potential issues (e.g. instances where system logs generated by specific actions could result in duplicate counts or data). We omitted 12 observations with missing data (e.g. if students did not take the final exam).

As emphasized earlier, this work focuses on providing timely support to students, which requires early predictions. Therefore, our features only include data up to week 8, which is halfway the semester and immediately following the trial exam. For example, the quiz results considered in our analysis are limited to those recorded up to week 8.

We adopted a methodology similar to Swamy et al. (forthcoming 2022) to develop a framework for feature categorization and engineering. Our dataset comprises a rich collection of diverse features, providing a holistic view of students' academic profiles. At the same time this introduces the mentioned modeling challenges in Section 2. We categorize the considered features into six groups based on their characteristics and relevance to our prediction task:

- **Engagement** features (ENG) quantify the time spent on online content.
- **Performance** features (PRF) encompass the students' scores on various assessments, including quizzes and trial exam.
- **Participation** features (PAR) indicate whether the student participated in a specific learning task or activity
- **Regularity** features (REG) capture the consistency and frequency of the student's learning behavior.
- **Repetition** features (REP) quantify the extent to which the student repeats certain learning tasks.
- **Learning profile** features (LPR) encompass personal educational background and learning-related attributes captured by MSLQ responses.

In Table 2 an overview of the key features is given for each group. The correlation ρ^2 indicates the strength of the linear relationship between each feature and the final grade. Participation features and performance on trial and quizzes show the highest correlation with the final grade.

3.3. Model building

We approach the problem as a classification and regression task, the two most common set ups in student performance prediction (Abu Saa et al., 2019). First, we classify each student into one of the three performance groups (High, Medium, Low). Second, we directly predict the final exam grade for each student. In application these predictions can be used to identify students at risk (when the model predicts *Low*) as it is common in EWS systems, or offer tailored advice to all groups of students. Accordingly, our goal is twofold: first, we want to provide accurate predictions about the students' success. For that, we evaluate different Machine Learning models with varying complexity on their predictive performance. Second, we evaluate the effect of the different suggested features onto student success, using LASSO-feature selection and a linear model. We first provide a brief introduction to the prediction models used, then outline LASSO feature selection.

3.3.1. Prediction models

We evaluate different prediction models from very simple white-box approaches to black-box machine learning models. We evaluate simple models with from our own teaching experience handpicked feature set: (i) using the trial exam scores directly as predictions for the final grade, replacing any missing values with a zero score (*TrialExamPure*). (ii) Fitting a linear/ logistic regression solely on the trial exam grade (*TrialExamLm*). (iii) Fitting a linear/ logistic regression solely on the trial exam grade and the GeneralEducation feature (*TrialExamGenEduLm*).

Additionally, we evaluate pipelines that include the overall feature sets: (iv) A linear model with prior feature selection by LASSO (*LmLASSO*). (v) two tree based machine learning models that are among the top performing in several benchmarks and common in the learning analytics literature across the whole feature set: xgboost (*Xgb*) and random forest (*Rf*) (Breiman, 2001; Chen & Guestrin, 2016; Grinsztajn et al., 2022), and support vector machines (*Svm*). Additionally, we assess these models with prior LASSO-feature selection (*XgbLasso*, *RfLasso*, *SvmLasso*). For an introduction into the evaluated models, see for example Hastie et al. (2008).

The simpler models cannot account for various complex patterns that might be present in the data, but are also less demanding in training/ estimation. Moreover they are by design fully explainable and relatively easy to explain to practitioners in their fitting procedure, as well as their predictions. Besides various neural network architectures, the Machine Learning ones are also the most frequently used models in forecasting educational performance (López-Zambrano et al., 2021; Balaji et al., 2021). While the predictions of these models are in principle also explainable, for example by Shapley values (Lundberg & Lee, 2017; Tiukhova et al., 2024), this adds additional complexity in the modeling pipeline. Table 3 summarizes the used software packages. We use the default hyper-parameters for all models.

3.3.2. LASSO feature selection

Given the relatively large number of features (and depending on the model, large number of free parameters) in relation to the relatively few number of observations, there is a high risk to overfit on the training data, and have poor generalization, and thus bad predictive performance in application (Lever et al., 2016). In a linear model setting this can be addressed by LASSO, where in fitting, we not only aim to minimize the in sample error, but also penalize the sum of the squared β-coefficients. This reduces overfitting and can thereby lead to more accurate out of sample predictions. Additionally this can be used for feature selection, as in fitting the β-parameters of the non-relevant features are shrinked to zero. We optimize the regularization-parameter (λ), that determines how aggressively the parameters are shrinked, via a in sample cross validation. For an introduction into LASSO see (Hastie et al., 2008, p. 68).

Table 2
Overview of key features used in model.

Set	Feature	ρ^2	Description
ENG	TotalActions	0.39	Total actions performed by student.
	MaterialViews	0.30	Number of times student visited the course material.
	DiscussionViews	0.38	Number of times student visited forum.
	TrialExamDuration	-0.08	Total time spent on trial exam.
	DiscussionsCreated	0.23	Number of new forum discussions started.
	ActiveDays	0.39	Total number of days with at least one action.
	PostsCreated	0.24	Number of forum contributions.
	QuizAttempts	0.23	Total number of quiz attempts.
PRF	TrialExamScore	0.39	Student's grade on trial exam.
	TrialExamCorrect	0.39	Number of correct answers on trial exam.
	TrialExamWrong	-0.19	Number of wrong answers on trial exam.
	TrialUnanswered	-0.23	Number of questions left unanswered on trial exam.
	AvgQuizGrade	0.31	Average score on quizzes.
	MinQuizGrade	0.28	Lowest score on a quiz.
	StdQuizGrade	0.17	Std. deviation of quiz scores.
	AvgQuizPercentile	0.39	Student's average quiz score relative to other students.
	FailedQuizzes	-0.13	Number of quiz attempts with a failing score.
	Q_X	≤0.36	Score on question X of trial exam.
PAR	DiscussionsViewed	**0.43**	% of forum posts viewed relative to total posts available.
	UniqueQuizzes	0.35	Number of unique quizzes completed.
	MaterialAccessed	0.39	% of material accessed relative to total material.
	ChaptersAccessed	0.35	% of chapters accessed.
REG	FirstActionDay	-0.12	First day student performed certain action.
	WeekVariability	0.35	Std. deviation of actions per week.
	MaxActivityGap	-0.28	Max number of days between subsequent actions.
	StdDaysBetween	-0.21	Std. deviation of days between actions
	RatioWeekend	0.25	% of actions in weekend
REP	AttemptsPerQuiz	0.18	Average number of attempts per quiz.
	ViewsPerMaterial	0.33	Average number of views per course material.
	ViewsPerDiscussion	0.38	Average number of views per forum discussion.
	QuizRetakes	0.18	Number of quizzes completed more than once.
LPF	GeneralEducation	0.39	Student completed general Secondary Education.
	TechnicalEducation	-0.31	Student completed technical Secondary Education.
	MSLQ_X	-	Average response to MSLQ construct X.
	HighestDegree	-	Highest degree of the student.

Table 3
Used prediction models.

Method	Package	Reference
linear regression	`glmnet` R package, version 4.1-8	Friedman et al. (2010)
LASSO regression	`glmnet` R package, version 4.1-8	Friedman et al. (2010)
random forest	`randomForest` R package, version 4.7-1.1	Breiman (2001), Liaw and Wiener (2002)
xgboost	`xgboost` R package, version 1.7.7.1	Chen and Guestrin (2016)
support vector machine	`e1071` R package, version 1.7-14	Vapnik (1997), Chang and Lin (2011)

3.3.3. Evaluation

For the regression set up, we evaluate model performance by the root mean squared error (RMSE), mean absolute error (MAE), and mean error (ME) to evaluate the bias in predictions. For the classification set up, we report accuracy (Acc), precision (Prec), and recall (Rec). We denote the i-th actual observation as y_i, and its prediction (in regression) as $\hat{y}_i$ for each of the n students. For classification we denote true positive predictions by TP, true negative predictions by TN, false positive predictions by FP, and false negative predictions by FN. The metrics are then defined as:

$$\text{RMSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad \text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|, \quad \text{ME} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i),$$

$$\text{Acc} = \frac{TP+TN}{TP+TN+FP+FN}, \quad \text{Prec} = \frac{TP}{TP+FP}, \quad \text{Rec} = \frac{TP}{TP+FN}.$$

4. Results and discussion

We provide an analysis of the collected data and a concept drift analysis from the Covid-period to the non-Covid period in Section 4.2. Next, in Section 4.3, we summarize the selected and significant features, before presenting the results of the accuracy benchmark in Section 4.4.

4.1. Data collection

Data was collected on participants' demographic, educational background and their interactions with the learning management system. Additionally, all participants completed a validated questionnaire assessing their motivations and learning strategies (i.e. MSLQ). Research shows that students' learning in online environments is linked to their self-regulatory skills, which might impact their learning outcomes (Van Laer & Elen, 2017). For example, online or blended learning appears to be particularly challenging for learners with lower self-regulatory abilities; however, the reverse is also true: students who can effectively regulate their own learning tend to perform well in these environments (Van Laer & Elen, 2017). Various theories on self-regulated learning have been developed over recent decades (Van Laer & Elen, 2017), which are rooted in a social cognitive perspective (Bandura, 1977), and

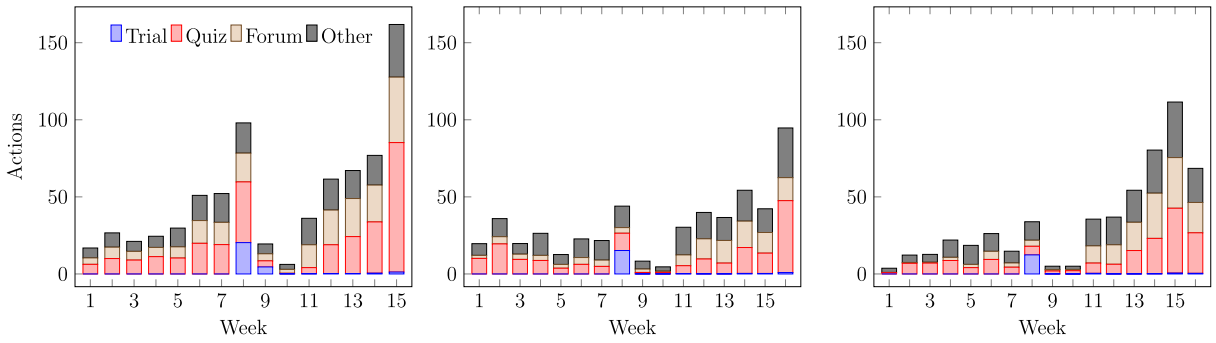

Fig. 2. Average weekly number of actions per student per activity group. See colored figure online.

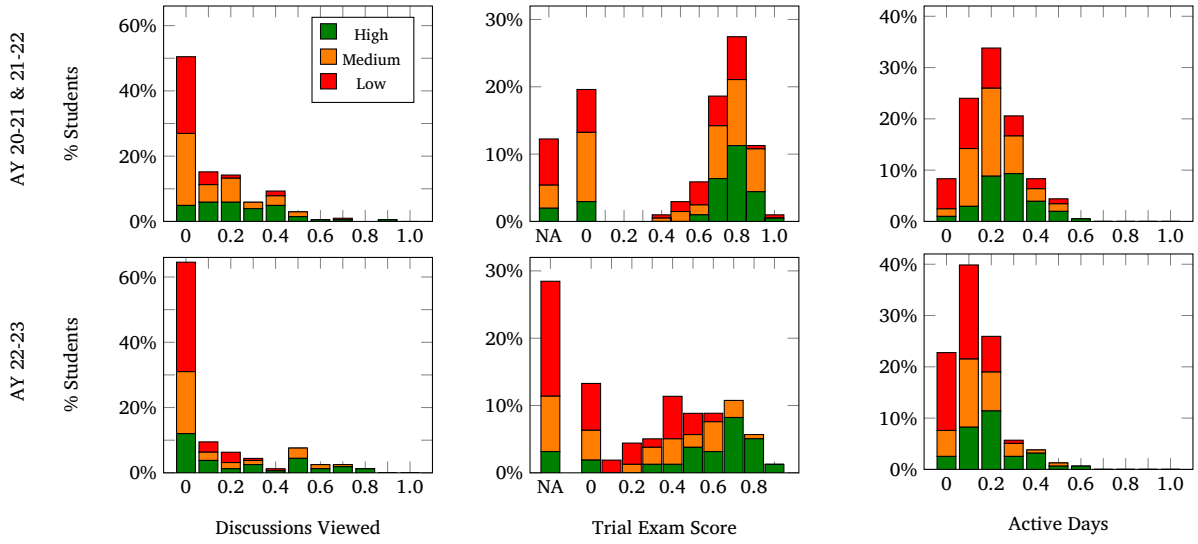

Fig. 3. Student distribution of LMS variables with the highest ρ^2 for AY 20-21 and AY 21-22 (training set) and their respective distributions in AY 22-23 (test set). See colored figure online.

view learning as an activity actively undertaken by students themselves rather than simply a result of instruction (Zimmerman, 1986).

4.2. Descriptive analysis and concept and covariate drift analysis

We first focus on frequency of student access to the online resources to better understand student engagement with the course material. Fig. 2 depicts the weekly distribution of student activity for the three academic years, categorized into the groups Trial Exam, Quiz, Forum, or Other. In all years, the activity increases towards the end of the semester, averaging approximately 35 actions per student per week. However, in the post-Covid period (AY 22-23), the activity during the first eight weeks was approximately 29% lower compared to the same weeks during Covid-19 (AY 20-21, AY 22-23). Using the Kolmogorov–Smirnov test we find a significant covariate drift for most engagement features, all performance features, all participation features, most regularity features, and most repetition features. In contrary we do not find covariate drift for the learning profile features. Table A.7 provides an overview of the p-values per feature.

Next we analyze the concept drift between AY20-21/AY21-22 (training data, Covid period) and AY22-23 (test data, post-Covid period) for some key features: Fig. 3 illustrates the proportion of students falling into specific grade categories for different features. For all these we find considerable concept drift: In the Covid period, only 9.7% of the students that viewed less than 10% of the forum discussions during the first half of the semester achieved a *High* grade. In the post-Covid period the share rose up to 18.6%. When students viewed at least 10% of the forum discussions, the proportions substantially rose to 47.5% and 48.2% respectively. Similarly, regarding the days active, in the Covid-period only 12.7% of students who were active less than 20% of days achieved a High-grade, but 17.2% of these did in the post-Covid period.

Regarding the trial exam, there was a decrease in the number of students participating in the post-Covid period (AY 22-23). In the Covid period, 42.7% of students who obtained a High score on the trial exam also achieved a High score on the final exam. Conversely, in the post-Covid period, this proportion increased to 72.2%. 48.5% and 56.5% Of students who failed or did not participate in the trial exam, achieved a Low score on the final exam. The results of the trial and final exams, correlate with 49% for AY 22-23.

Fig. 4 shows the distributions of the AY 21-22 (Covid period) and AY 22-23 (post-Covid period) for participation features. As mentioned above, the students in AY 22-23 were overall less active during the first half of the semester.

4.3. Feature relevance

We use LASSO for a first step feature selection and then evaluate the relevance of each selected feature for predicting the grade of each student, with a linear model. For the regression analysis here we do so on train and test data jointly. The feature selection in the accuracy benchmark (Section 4.4) is done only on the training data to prevent data leakage. We report the selected features, along with the fitted coeffi-

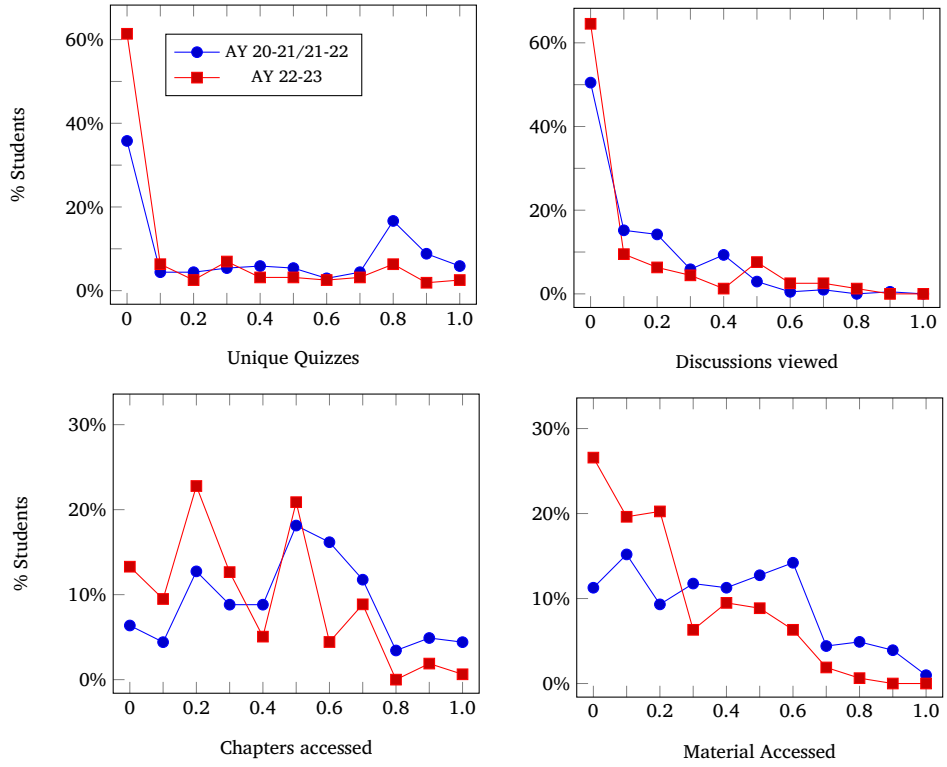

Fig. 4. Distribution of participation features for AY20-21/21-22 (training set), and AY22-23 (test set). Upper left: ratio of quizzes attempted, upper right: forum discussions viewed, lower left: chapters accessed, and lower right: % of total material accessed. See colored figure online.

Table 4
Regression: fitted coefficients and their significance.

Set	Feature	coef (± std. error)
ENG	TrialExamDuration	$'$-0.31 (± 0.18)
PRF	TrialExamWrong	***-1.07 (± 0.24)
	Q15	*0.82 (± 0.41)
	Q7	*0.68 (± 0.34)
	TrialExamScore	0.39 (± 0.36)
	TrialExamCorrect	0 (± 0.52)
	AvgQuizPercentile	0.21 (± 0.27)
PAR	DiscussionsViewed	0.4 (± 0.47)
	ChaptersAccessed	0.33 (± 0.31)
	MaterialAccessed	0.15 (± 0.28)
REG	ViewsPerDiscussion	0.52 (± 0.46)
	StdDaysBetween	-0.24 (± 0.22)
LPR	MSLQ Elaboration	*0.46 (± 0.18)
	GeneralEducation	***1.24 (± 0.19)
	TechnicalEducation	**-0.51 (± 0.19)

Significance in the linear model: $^{***}: p < 0.001$, $^{**}: p < 0.01$, $^{*}: p < 0.05$, $': p < 0.1$.

cients and their significance in the regression setting in Table 4, and for classification in Table 5. Features that were not selected by LASSO are omitted from the table. The feature selection beforehand allows for a more robust regression analysis, since substantially fewer independent variables are considered and as outlined in Section 2, evaluating a large set of features on relatively few observations can lead to spurious findings. Table 4 shows that LASSO selects 15 features, and of these the OLS regression identifies six variables as significant for student success: the trial exam results (TrialExamWrong and answers to two specific questions), the answers to the MSLQ Elaboration questionaire, and the previous highest degree of the student (GeneralEducation, OtherEducation). In the classification setting, LASSO selects 19 features of which 8 are significant in the OLS regression over all classes (High - Medium - Low). The main difference to the regression setting is the share of forum discussions viewed (DiscussionsViewed) and TrialExamScore. All other features, many of them used to predict student success by previous studies are insignificant in our analysis or omitted by LASSO.

The duration of the trial exam and the number of wrong answers on the trial exam have a negative effect on the student success. In contrast, discussions viewed and materials accessed have a positive impact on the student success. The general education level has a positive impact on the student success. The same can be seen in Table 5 as a positive coefficient for the category 'High' and negative coefficient for the category 'Low'. Technical education, the alternative to general education

Table 5
Classification: fitted coefficients and their significance.

Set	Feature	High	Medium	Low
ENG	TrialExamDuration	*-0.53 (± 0.22)	-	-
	DiscussionsCreated	′0.42 (± 0.25)	-	-
	MaterialViews	0.12 (± 0.27)	-	-
PRF	TrialExamScore	* 0.71 (± 0.35)	-	′-0.44 (± 0.26)
	TrialExamCorrect	-0.61 (± 0.5)	-	-0.05 (± 0.36)
	TrialExamWrong	**-0.71 (± 0.23)	-	-
	AvgQuizPercentile	-0.03 (± 0.23)	-	-
	Q15	-	-	-0.07 (± 0.25)
	Q16	*0.64 (± 0.32)	-	-
	Q7	*0.61 (± 0.27)	-	-
PAR	DiscussionsViewed	0.24 (± 0.41)	-	*-0.52 (± 0.22)
	MaterialAccessed	-0.15 (± 0.33)	-	-
REG	StdDailyActions	-	-	-0.1 (± 0.27)
	StdDaysBetween	-	-	0.19 (± 0.16)
REP	ViewsPerChapter	0.2 (± 0.25)	-	-0.1 (± 0.25)
	ViewsPerDiscussion	0.12 (± 0.45)	-	-
LPF	MSLQ Elaboration	* 0.35 (± 0.16)	-	-
	MSLQ_Submit_Date	-0.15 (± 0.18)	-	′0.27 (± 0.15)
	GeneralEducation	***0.82 (± 0.15)	-	***-0.79 (± 0.15)

Significance in the linear model: *** : $p < 0.001$, ** : $p < 0.01$, * : $p < 0.05$, ′ : $p < 0.1$.

Table 6
Accuracy benchmark results. Accuracy (Acc.) relates to the classification setting, and RMSE to the regression setting.

Model[a]	RMSE	MAE	ME	Acc.	Precision[a]			Recall[a]		
					H	M	L	H	M	L
TrialExamPure	5.91	4.68	2.76	0.54	0.73	0.33	0.54	0.59	0.26	0.73
TrialExamLm	4.05	3.41	0.46	0.50	0.94	0.35	0.55	0.37	0.55	0.56
TrialExamGenEduLm	3.75	3.12	**0.27**	**0.55**	0.81	0.39	0.72	0.37	0.70	0.58
LASSO	3.79	3.22	-0.74	0.44	0.69	0.34	-	0.48	0.83	0.09
LmLASSO	**3.65**	**3.02**	-0.74	0.48	0.70	0.34	0.65	0.50	0.67	0.31
RfLASSO	3.91	3.16	1.40	0.49	0.98	0.35	0.58	0.11	0.52	0.75
Rf	4.51	3.64	2.11	0.40	-	0.12	0.44	0.00	0.06	0.99
SvmLASSO	4.76	4.00	0.26	0.31	-	0.31	-	0.00	1.00	0.00
Svm	4.92	4.24	-1.22	0.31	-	0.31	-	0.00	1.00	0.00
XgbLASSO	5.70	4.58	3.66	0.46	0.56	0.35	0.59	0.19	0.53	0.61
Xgb	6.73	5.54	5.36	0.53	0.59	0.36	0.58	0.28	0.30	0.90

[a] H: High mark, M: Medium mark, L: Low mark; Precision for zero predictions in the respective class cannot be calculated.

hinders the students' success. A high score on the MSLQ Elaboration feature improves student performance, and is particularly helpful to obtain a 'High' grade. MSLQ Elaboration is a self report on a specific cognitive strategy. This includes paraphrasing, summarizing, creating analogies and generative note-taking and helps students storing information into long-term memory by connecting it to prior knowledge. The finding that self-reported elaboration is linked with student performance can be valuable information for the teacher of this specific course. Based on Pintrich's original work the teacher might suggest to paraphrase and summarize important information (Pintrich, 1991). Also, encouraging students to imagine themselves as the teacher explaining the content to others could improve their understanding and performance (Pintrich, 1991). As with all predictive features we believe that MSLQ constructs are subject to instructional conditions meaning that different courses might find value in different constructs (Gašević et al., 2016). Overall these findings are in line with our expectations from our own teaching experience.

4.4. Predictive accuracy benchmark

We report forecast accuracy for the described 12 different predictive modeling set ups with varying complexity and explainability. We do so for the classification set up and the regression set up. For the pipelines that include LASSO feature selection, this was done only on the training set (in contrast to our regression analysis). Table 6 provides the accuracy results on the test set.

Besides our very simple models being (among) the most accurate, we also find that the predictions of the Random Forest and XGBoost are substantially biased (ME). Fig. 5 illustrates the predicted versus the actual grades for a single seed of *XgbLasso*. The predictions systematically underestimate the true grades of high-performing students. This bias can probably be attributed to the concept drift from AY 21-22 to AY 22-23. Throughout the models, the bias is lower when feature selection is done beforehand, and the simplest approaches achieve a substantially lower bias. *Svm* and *SvmLASSO* predict "Medium" for each student and are such also fairly unbiased but also has no predictive power.

5. Conclusions

We collected data from freshmen students in nursing aiming at early, accurate and explainable prediction of student success in a physiology course. For this purpose we extracted and analyzed student interaction variables from a LMS and combined them with demographic information on students and the results from a questionnaire on self regulation.

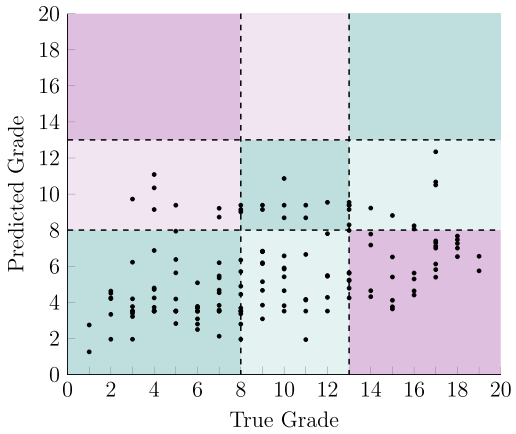

Fig. 5. Predicted grades compared to true grades for a single seed of XgbLasso model. See colored figure online.

We find notable shifts in online student engagement over successive academic years. Despite maintaining a comparable overall level of activity throughout the entire semester, students in the AY 22-23 were less active during the first half of the semester compared to the previous year. Additionally, a smaller proportion of students participated in the trial exam and the average score of those that did was lower.

The transition from full remote learning modalities during the Covid-19 pandemic back to limited face-to-face teaching in AY 22-23 may have contributed to the observed change in student behavior. We find that students had a more stable online learning pattern when face-to-face education was prohibited. This may explain (in part) a covariate and concept drift in the data that makes learning from the pandemic-years challenging. However, based on previous work we believe that similar data shifts are common in educational settings, making the potential use of predictive algorithms in learning analytics challenging (Conijn et al., 2017; Tiukhova et al., 2024). So even though Covid-19 was a highly unusual and maybe unique study period, drifts in study behavior are not uncommon - similar drifts might for example occur due to changes in teachers, curriculums or student profile. Additionally, also other crisis such as natural disasters, or armed conflict likely cause substantial change in student behavior. Accordingly, our findings might transfer to these circumstances as well, and are not unique to the special circumstances under investigation.

Using the collected data, we evaluated a number of features on their significance and relation to student performance. For this purpose we combined student background data (e.g. the MSLQ questionnaire or prior degrees) with online learning behavior such as the performance on a voluntary trial exam halfway the semester and trace data from the learning platform. We found the following features to be significant: (i) the student's prior education degree (ii) the MSLQ Elaboration feature, (iii) the trial exam performance, (iv) to what extent the online forum discussions were used. The finding that student background data holds predictive value confirms previous research (Aggarwal et al., 2021) although this is not a stable finding (Tomasevic et al., 2020). The MSLQ Elaboration feature is up to now not commonly used in models for student performance prediction, but it can guide teachers in the choice of pedagogic methods, and is such valuable beyond accuracy improvements. It will be interesting to see whether our result holds in other courses, similar to the nursing course in our study. That predictive accuracy benefits from early assessments like the trial exam is in line with previous research (Namoun & Alshanqiti, 2020). Finally, although the predictive value of interacting with discussion forums has been reported before, this is also not a stable finding in published literature. Other features that have previously been linked with predictive power (e.g. regularity features) were not significantly linked to student performance in this work. Overall, these results confirm that predictive modeling in education is specific to instructional context and should include student background data.

In our predictive accuracy evaluation, we find that a linear model that takes into account only performance on the trial exam and the GeneralEducation feature (i.e. prior degree) identified 72% percentage of students who failed the course already halfway the semester, performed best in the classification setting, and ranked second in the regression setting even when compared to several elaborate machine learning pipelines. Next to its good performance, this model is also explainable in fitting and predictions. The more elaborate models that used large behavioral feature sets provided substantially biased predictions that systematically underestimated student performance, likely due to the mentioned concept drift. This effect was also mitigated by the simple sparse model. Accordingly, for scientific studies we suggest to use linear models with similar and few features at least as a benchmark to predict student performance in a blended learning environment, especially if there is only few data available or data shift has to be expected, as it was the case in our study. Even though this model does not directly allow for behavioral insights, practitioners and EdTech software suppliers can use it for use-cases where the prediction itself triggers an action, such as in EWS.

We acknowledge the following limitations. Firstly, the collected data was on a single course, which limits the generalizability of our findings to other subjects or academic contexts. Second, this also leads to a fairly small data set, where we would not have expected machine learning algorithms to blossom. On substantially larger data sets, these might perform better than our simple benchmarks. However, given the substantial concept drift in the data, we suspect that any method that learns complex patterns in the data will have similar difficulties, irrespective of data set size. Third, although nearly all learning activities were organized online, we do not report on teacher behaviors that might influence student' learning behavior or results. While data collection in a real-life setting can enhance ecological validity, it complicates the standardization of teacher behavior. For instance, variations in the extent of teacher encouragement to engage with online study materials are not reported and differed across academic years (e.g. in the AY 22-23 the teacher sent personalized messages encouraging interaction with the learning material following the trial exam). This might have impacted some of the data drift that we report. We believe our accuracy and regression study results remain nevertheless valid since these do not directly depend on the cause of the data drift. Additionally, we did not capture offline learning activities, such as studying from textbooks, notes or participating in face-to-face classes. These activities impact student performance beyond online learning, and might be addressed in future research.

CRediT authorship contribution statement

Benedikt Sonnleitner: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Tom Madou:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Matthias Deceuninck:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation, Conceptualization. **Filotas Theodosiou:** Writing – review & editing, Writing – original draft, Software, Methodology, Conceptualization. **Yves R. Sagaert:** Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization.

Open data and ethics

The study was ethically approved by the leading authors university, reference [PWOAIinEducation]. Informed consent was obtained from all participants, and their privacy rights were strictly observed. The data can be obtained by sending request e-mails to the corresponding author.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Elena Tiukhova for friendly review. This research was funded by the Vives University of Applied Sciences, grant [PWOAIin-Education].

Appendix A. Kolmogorov–Smirnov test result table

Table A.7
Kolmogorov-Smirnov test results.

Set	Feature	p-Value
ENG	TotalActions	< 1e-06
	MaterialViews	0.009
	DiscussionsViews	< 1e-06
	TrialExamDuration	0.002
	DiscussionsCreated	0.7
	ActiveDays	< 1e-06
	PostsCreated	0.389
	QuizAttempts	< 1e-05
PRF	TrialExamScore	< 1e-06
	TrialExamCorrect	< 1e-06
	TrialExamWrong	< 1e-05
	TrialUnanswered	< 1e-06
	AvgQuizGrade	< 1e-03
	StdQuizGrade	< 1e-05
	MinQuizGrade	< 1e-03
	MaxQuizGrade	< 1e-04
	AvgQuizPercentile	< 1e-04
	FailedQuizzes	< 1e-03
	Q_X	$[< 1e-03, < 1e-06]$
PAR	DiscussionsViewed	< 1e-04
	UniqueQuizzes	< 1e-05
	MaterialAccessed	< 1e-06
	ChaptersAccessed	< 1e-05
REG	FirstActionDay	< 1e-06
	WeekVariability	< 1e-05
	MaxActivityGap	0.001
	StdDaysBetween	< 1e-05
	RatioWeekend	0.006
REP	AttemptsPerQuiz	< 1e-03
	ViewsPerMaterial	< 1e-05
	ViewsPerDiscussion	< 1e-03
	QuizRetakes	0.161
LPF	GeneralEducation	1
	TechnicalEducation	0.993
	MSLQ_Rehearsal	0.408
	MSLQ_Elaboration	0.203
	MSLQ_Organization	0.999
	MSLQ_Critical_Think	0.877
	MSLQ_Self_Regulation	0.612
	MSLQ_Environment	< 1e-06
	MSLQ_Effort	< 1e-05
	MSLQ_Peer_Learning	< 1e-03
	MSLQ_Help_Seeking	0.003
	MSLQ_WeightedAVG	0.78
	HighestDegree	1

References

Abu Saa, A., Al-Emran, M., & Shaalan, K. (2019). Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques. *Technology, Knowledge and Learning*, *24*(4), 567–598. https://doi.org/10.1007/s10758-019-09408-7. ISSN 2211-1662.

Aggarwal , D., Mittal, S., & Bali, V. (2021). Significance of non-academic parameters for predicting student performance using ensemble learning techniques. *International Journal of System Dynamics Applications*, *10*(3), 38–49. https://doi.org/10.4018/ijsda.2021070103. ISSN 2160-9772.

Alwarthan, S. A., Aslam, N., & Khan, I. U. (2022). Predicting student academic performance at higher education using data mining: A systematic review. *Applied Computational Intelligence and Soft Computing*, *2022*, 1. https://doi.org/10.1155/2022/8924028. ISSN 1687-9732.

Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–17. https://doi.org/10.5281/zenodo.3554657. Retrieved from https://jedm.educationaldatamining.org/index.php/JEDM/article/view/8.

Balaji, P., Alelyani, S., Qahmash, A., & Mohana, M. (2021). Contributions of machine learning models towards student academic performance prediction: A systematic review. *Applied Sciences (Switzerland)*, *11*(21), 11. https://doi.org/10.3390/app112110007. ISSN 2076-3417.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191. ISSN 1939-1471.

Bayram, F., Ahmed, B. S., & Kassler, A. (2022). From concept drift to model degradation: An overview on performance-aware drift detectors. *Knowledge-Based Systems*, *245*, Article 108632. https://doi.org/10.1016/j.knosys.2022.108632. ISSN 0950-7051.

Bernacki, M. L., Chavez, M. M., & Uesbeck, P. M. (2020). Predicting achievement and providing support before STEM majors begin to fail. *Computers and Education*, *158*, 12. https://doi.org/10.1016/j.compedu.2020.103999. ISSN 0360-1315.

Bossant, L., Madou, T., Theodosiou, F., & Sagaert, Y. R. (2022). The effects of self-regulation and instructional control on learning behaviour and performance of undergraduate nursing students. In *In ACM international conference proceeding series* (pp. 39–43). Association for Computing Machinery. ISBN 9781450397766.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324. ISSN 1573-0565.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–27.

Chen, F., & Cui, Y. (2020). Utilizing student time series behaviour in learning management systems for early prediction of course performance. *Journal of Learning Analytics*, *7*(2), 1–17. https://doi.org/10.18608/JLA.2020.72.1. ISSN 1929-7750.

Chen, T., & Guestrin, C. (2016). XGBoost. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). ACM. ISBN 9781450342322.

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. ISSN 1939-1382.

Costa, E. B., Fonseca, B., Almeida Santana, M., Ferreira de Araújo, F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, *73*, 247–256. https://doi.org/10.1016/j.chb.2017.01.047. ISSN 0747-5632.

Dutt, A., Ismail, M. A., & Herawan, T. (2017). *A Systematic Review on Educational Data Mining. IEEE Access*, *5*, 15991–16005. https://doi.org/10.1109/ACCESS.2017.2654247. ISSN 2169-3536.

Dziuban, C., Graham, C. R., Moskal, P. D., Norberg, A., & Sicilia, N. (2018). Blended learning: The new normal and emerging technologies. *International Journal of Educational Technology in Higher Education*, *15*(1), 12. https://doi.org/10.1186/s41239-017-0087-5. ISSN 2365-9440.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01. Retrieved from https://www.jstatsoft.org/index.php/jss/article/view/v033i01.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, *46*(4), 1–37. https://doi.org/10.1145/2523813. ISSN 0360-0300.

Gašević , D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, *28*, 68. https://doi.org/10.1016/j.iheduc.2015.10.002. ISSN 1096-7516.

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In S. Koyejo,

S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems: Vol. 35* (pp. 507–520). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2022/file/0378c7692da36807bdec87ab043cdadc-Paper-Datasets_and_Benchmarks.pdf.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning: Data mining, inference, and prediction* (2 edition). Springer.

Howard, E., Meehan, M., & Parnell, A. (2018). Contrasting prediction methods for early warning systems at undergraduate level. *The Internet and Higher Education*, *37*, 66. https://doi.org/10.1016/j.iheduc.2018.02.001. ISSN 1096-7516.

Ifenthaler, D., & Yau, J. Y. K. (2020). Utilising learning analytics to support study success in higher education: A systematic review. *Educational Technology Research and Development*, *68*(4), 1961–1990. https://doi.org/10.1007/s11423-020-09788-z. ISSN 1556-6501.

Jovanović, J., Saqr, M., Joksimović, S., & Gašević, D. (2021). Students matter the most in learning analytics: The effects of internal and instructional conditions in predicting academic success. *Computers and Education*, *172*, 10. https://doi.org/10.1016/j.compedu.2021.104251. ISSN 0360-1315.

Lauría, E. J. M., Baron, J. D., Devireddy, M., Sundararaju, V., & Jayaprakash, S. M. (2012). Mining academic data to improve college student retention. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 139–142). ACM. ISBN 9781450311113. https://doi.org/10.1145/2330601.2330637.

Laurillard, D. (2013). *Teaching as a design science: Building pedagogical patterns for learning and technology*. Routledge.

Lever, J., Krzywinski, M., & Altman, N. (2016). Model selection and overfitting. *Nature Methods*, *13*(9), 703–704. https://doi.org/10.1038/nmeth.3968. ISSN 1548-7091.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*(3), 18–22. Retrieved from https://CRAN.R-project.org/doc/Rnews/.

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 (Switzerland)*, *10*(1), 1. https://doi.org/10.3390/app10010354. ISSN 2076-3417.

López-Zambrano, J., Torralbo, J. A. L., & Romero, C. (2021). Early prediction of student learning performance through data mining: A systematic review. *Psicothema*, *33*(3), 456–465. https://doi.org/10.7334/psicothema2021.62. ISSN 1886-144X.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems: Vol. 30*. Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, *54*(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008. ISSN 0360-1315.

Meier, Y., Xu, J., Atan, O., & Van Der Schaar, M. (2016). *Personalized grade prediction: A data mining approach*In *Proceedings - IEEE international conference on data mining, ICDM, volume 2016-January* (pp. 907–912). Institute of Electrical and Electronics Engineers Inc. ISBN 9781467395038.

Morris, L. V., Finnegan, C., & Wu, S.-S. (2005). Tracking student behavior, persistence, and achievement in online courses. *The Internet and Higher Education*, *8*(3), 221–231. https://doi.org/10.1016/j.iheduc.2005.06.009. ISSN 1096-7516.

Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, *11*(1), Article 237. https://doi.org/10.3390/app11010237. ISSN 2076-3417.

Pintrich (1991). A manual for the use of the motivated strategies for learning questionnaire (MSLQ). Technical report. National Center for Research to Improve Postsecondary Teaching and Learning.

Ramaswami, G., Susnjak, T., Mathrani, A., & Umer, R. (2023). Use of predictive analytics within learning analytics dashboards: A review of case studies. *Technology, Knowledge and Learning*, *28*(3), 959–980. https://doi.org/10.1007/s10758-022-09613-x. ISSN 2211-1662.

Riestra-González, M., del Puerto Paule-Ruíz, M., & Ortin, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Computers and Education*, *163*, 4. https://doi.org/10.1016/j.compedu.2020.104108. ISSN 0360-1315.

Romero, C., López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers and Education*, *68*, 458–472. https://doi.org/10.1016/j.compedu.2013.06.009. ISSN 0360-1315.

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*(6), 405–430. https://doi.org/10.1002/for.928. ISSN 0277-6693.

Swamy, V., Radmehr, B., Krco, N., Marras, M., & Käser, T. (2022). Evaluating the explainers: Black-Box explainable machine learning for student success prediction in MOOCs. Retrieved from arXiv:2207.00551.

Tempelaar, D. T., Rienties, B., & Giesbers, B. (2016). Verifying the stability and sensitivity of learning analytics based prediction models: An extended case study. In S. Zvacek, M. T. Restivo, J. Uhomoibhi, & M. Helfert (Eds.), *Computer supported education* (pp. 256–273). Cham: Springer International Publishing, ISBN 978-3-319-29585-5.

Tiukhova, E., Vemuri, P., Lopez Flores, N., Islind, A. S., Oskarsdottir, M., Poelmans, S., Baesens, B., & Snoeck, M. (2024). Explainable learning analytics: Assessing the stability of student success prediction models by means of explainable AI. *Decision Support Systems*, Article 114229. https://doi.org/10.1016/j.dss.2024.114229. ISSN 0167-9236.

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers and Education*, *143*, 1. https://doi.org/10.1016/j.compedu.2019.103676. ISSN 0360-1315.

Tsai, Y. S., Rates, D., Moreno-Marcos, P. M., Muñoz-Merino, P. J., Jivet, I., Scheffel, M., Drachsler, H., Delgado Kloos, C., & Gašević, D. (2020). Learning analytics in European higher education—trends and barriers. *Computers and Education*, *155*, 10. https://doi.org/10.1016/j.compedu.2020.103933. ISSN 0360-1315.

Van Laer, S., & Elen, J. (2017). In search of attributes that support self-regulation in blended learning environments. *Education and Information Technologies*, *22*(4), 1395–1454. https://doi.org/10.1007/s10639-016-9505-x. ISSN 1360-2357.

Vapnik, V. N. (1997). The support vector method. In *Artificial neural networks - ICANN'97* (pp. 261–271). Berlin, Heidelberg: Springer Berlin Heidelberg, ISBN 978-3-540-69620-9.

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, *89*, 98. https://doi.org/10.1016/j.chb.2018.07.027. ISSN 0747-5632.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108. ISSN 0003-1305.

Wu, J.-Y. (2021). Learning analytics on structured and unstructured heterogeneous data sources: Perspectives from procrastination, help-seeking, and machine-learning defined cognitive engagement. *Computers & Education*, *163*, Article 104066. https://doi.org/10.1016/j.compedu.2020.104066. ISSN 0360-1315.

Xu, J., Moon, K. H., & Van Der Schaar, M. (2017). A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal on Selected Topics in Signal Processing*, *11*(5), 742–753. https://doi.org/10.1109/JSTSP.2017.2692560. ISSN 1932-4553.

You, J. W. (2016). Identifying significant indicators using LMS data to predict course achievement in online learning. *Internet and Higher Education*, *29*, 23. https://doi.org/10.1016/j.iheduc.2015.11.003. ISSN 1096-7516.

Zimmerman, B. J. (1986). Becoming a self-regulated learner: Which are the key subprocesses? *Contemporary Educational Psychology*, *11*(4), 307–313. https://doi.org/10.1016/0361-476X(86)90027-5. ISSN 0361-476X.