Technology in Society 76 (2024) 102474

Contents lists available at ScienceDirect

Technology in Society

journal homepage: www.elsevier.com/locate/techsoc

Predicting student dropouts with machine learning: An empirical study in Finnish higher education

Matti Vaarma [a,*], Hongxiu Li [b]

[a] Department of Information and Knowledge Management, Faculty of Management and Business, Tampere University, Finland
[b] Department of Information and Knowledge Management, Faculty of Management and Business, Tampere University, Finland

ARTICLE INFO

Keywords:
Student dropout prediction
Learning management systems
Machine learning
Higher education
Data analysis

ABSTRACT

This study uses three machine learning models to predict student dropouts based on students' transcript, demographic, and learning management system (LMS) data from a Finnish university. The contribution of this research lies in 1) comparing the relative importance of LMS (Moodle) data with transcript and demographic data in degree program dropout prediction, 2) examining the predictive importance of different data features monthly as a function of time from enrollment, hence extending the prior end-of-semester research to a mid-semester analysis, and 3) measuring the prediction performance of the models monthly. The results identify "accumulated credits" (transcript) the "number of failed courses" (transcript), and "Moodle activity count" (LMS) as the most important features, suggesting LMS has significant predictive power and should be considered alongside transcript and demographic data when predicting degree program dropouts. Moreover, we visualize how these factors' importance and prediction performance vary over time, revealing general longitudinal trends and fluctuations within semesters. Finally, we elaborate upon this study's contributions before highlighting its limitations.

1. Introduction

Student dropout is a key challenge for higher education institutions because of its negative impact on both students and society. Higher education institutions have taken action to improve student retention, and research on this subject has been conducted since the 1970s [1]. However, this topic remains important since substantial improvements have been difficult to achieve [2]. For example, according to a recent study conducted in the United States, about 40% of students fail to obtain bachelor's degrees within six years of entering college [3].

In the last two decades, the availability of digitalized student data has opened new opportunities to identify at-risk students and analyze the factors that lead to their dropping out. Institutional transcript and students' demographic data have been fed into machine learning algorithms to predict at-risk degree program students and analyze which features are most important for student retention [4–10]. Moreover, the data collected by various learning management systems (LMS), such as Moodle, provide a rich source for educational data mining [11]. However, although previous research has applied transcript and student demographic data to degree program dropout prediction in higher education, LMS data have mainly been applied to dropout prediction at the course level [12–17]. Few studies have linked LMS data to degree dropouts in higher education, but they have not compared the relative importance of LMS data with that of transcript or demographic data in student dropout prediction [18–20].

Furthermore, although previous research has applied machine learning algorithms to degree program dropout prediction in higher education, these studies have utilized almost exclusively data collected at the end of a semester [21–26]. In other words, the prior literature has not focused on investigating different machine learning techniques' prediction power within a semester. Additionally, although prior studies have investigated different factors in student dropout predictions, they have rarely examined how the important factors that contribute to such predictions behave as a function of time. According to [22], educational researchers should consider the temporal dimension when examining student dropout.

To address these research gaps, in the current study, we analyzed and ranked the relative importance of data features from 8,813 students' transcript, demographic, and LMS (Moodle) data at a Finnish university of applied sciences. Moreover, we examined the dropout prediction performance of three different machine learning models—CatBoost (CAT), neural networks (NN), and logistic regression (LR)—as a function

* Corresponding author.
E-mail address: matti.vaarma@tuni.fi (M. Vaarma).

https://doi.org/10.1016/j.techsoc.2024.102474
Received 18 September 2023; Received in revised form 20 December 2023; Accepted 1 February 2024
Available online 3 February 2024

of time from enrollment. We also conducted a time-dependency analysis of the most important features to see how their importance varied over time.

Thus, in the current study, we endeavored to offer the following contributions to the student dropout literature. First, we analyzed LMS data's importance in degree program dropout prediction compared to transcript and demographic data in higher education. Second, we examined different data features' monthly predictive importance as a function of time from enrollment to identify general longitudinal trends, as well as fluctuations within semesters, thus extending the prior end-of-semester research to a midsemester analysis. Third, we analyzed the three machine learning models' monthly prediction performance. Thus, the present study provides an in-depth understanding of the predictive power and role of the most important data features of different machine learning techniques for degree program dropout prediction in higher education. In the current study, we also aimed to offer practical contributions by helping the representatives of higher education institutions understand how and when to apply machine learning in student dropout prediction and supporting them to make decisions on prevention strategies.

2. Literature review

The prior literature on student dropout prediction in higher education can be divided into at least three different research streams according to their settings. The first stream is predicting dropouts at the course level [12–14,27–31]. The second is predicting dropouts within a degree program [21,24,32]. Finally, the third stream is predicting dropouts within a university or multiple degree programs [4–6,22,26]. In this study, we focused on the third stream.

Prior research has examined dropout prediction performance metrics and feature importance analysis. [33] used the demographic, transcript, academic, and macro-economic data of 4,433 students in higher education to predict whether students would drop out of a degree program after their enrollment year. The synthetic minority over-sampling technique (SMOTE), support vector machine (SVM), random forest (RF), RUSBoost (RB), and easy ensemble (EE) machine learning models were employed in predictions with f1-scores from 58% to 66%. In that study [33] found "accumulated credits" to be an important feature in dropout prediction. [7] predicted student dropouts using the demographic, first-term transcript, and behavioral engagement data of 50, 095 students from four US universities. Using the elastic net (EN) and RF methods, they obtained area under curve (AUC) scores of 0.65–0.79, depending on the data set. Moreover, they found "grade point average" (GPA) and "behavioral metrics" (e.g., campus event attendance) to be important features. [8] also applied several machine learning classifiers to predict student dropouts using the demographic and academic history data of 6,398 students from a Hungarian university. In their study, Catboost (CAT)—the best-performing classifier—achieved an AUC score of 0.774 when predicting dropouts in a cohort of students who had enrolled in 2017. The authors found out that "high school GPA", "high school math scores", and "the number of years elapsed between high school graduation and university enrollment" were the most important features in predicting student dropouts. [34] employed six different machine learning classifiers to predict student dropouts using the demographic, transcript, and attendance data of more than 36,000 students from a South Korean university. "The number of scholarships", "tuition fees", and "access year" were identified as the most important features. These machine learning classifiers achieved a precision of 72%–83%.

[35] used generalized linear model (GLM), RF, and decision trees (DT) to predict early (no later than the third semester) and late (after the first three semesters) dropouts using the demographic, academic history, and transcript data of 31,071 students from an Italian university with AUC scores of 0.871–0.961, depending on the data set. In their study "accumulated credits during the first semester" was found to be the most important feature. Berka and Merek [4] applied DT, RF, and LR to mine the demographic and transcript data of 3,339 Czech bachelor students at the end of four semesters. Prediction accuracies of around 80% were obtained using varying models and data sets. Those authors found the models' most important feature to be "the percentage of lost credit vouchers in the last semester," a variable expressing the effort a student must exert to finish a semester. [36] used the demographic, financial, and transcript data of 93,457 students with gradient-boosted trees (GBT) and LR to predict whether the students would return after their first year of study. The residential students' dropout prediction precision ranged from 83.6% to 83.9%, and the recall ranged from 53.2% to 54.1%, depending on the model and the data set. Those authors found that some protected data features—such as "gender", "first-generation college student status", "underrepresented minority status", and "high financial need"—were less important in dropout prediction. [37] used GBT, extreme gradient boosting (XGB), and NN to predict student dropouts using the demographic, academic history, and transcript data of 10,196 Hungarian university students. The applied machine learning classifiers achieved prediction precision of 67%–86% and recall of 74%–81%, depending on the data set and the classifier. They also conducted feature importance analysis, finding that "the credit index", "credits earned", and "students' age at the time of their enrollment" were the most important features.

[25] employed RF, NN, SVM, and LR to predict student dropouts based on the demographic, transcript, and academic history data of 15, 720 students who had enrolled at the Instituto Tecnológico de Costa Rica. The prediction algorithms were applied at the ends of 12 different semesters, and the prediction precisions averaged 82%, while the sensitivity averaged 71%. They did not identify the most important factors contributing to the prediction performance. [38] used the demographic, transcript, academic history, and financial data of 7,800 students collected from a US university to predict dropouts with DT, naïve Bayes (NB), NN, and rule induction at the end of the students' first three semesters. The dropout prediction precision progressed over time from 66% to 74%, with recall ranging from 24% to 52%. In their study, the most important features that contributed to the prediction were related to first-year academic performance. [5] applied DT, NN, SVM, and LR to predict whether 16,066 US students would return to university after their freshmen year. The students' demographic, transcript, academic history, and financial data were fed into the algorithms, resulting in prediction precision of 75%–87%, depending on the algorithm. Three variables—"earned hours divided by registered hours," "student loan at spring," and "fall GPA"—were identified as the most important features contributing to the prediction performance.

[39] applied event history analysis to the data of 6,730 US students that had been collected through a longitudinal survey to identify how financial and racial factors were related to college students' risk of dropping out. They found that minority and lower-income students faced a higher risk of dropping out than their counterparts. [40] employed a multinomial logit model to study how academic, demographic, psychosocial, situational, and institutional factors could predict enrollment and degree outcomes, including dropping out. Based on the data of 4,500 US college students from 2003 to 2008, they found that high GPA, high motivation, and full-time enrollment increased students' probability of obtaining college degrees. [23] investigated sociodemographic, academic history, and financial pre-college factors' effects on college outcomes (e.g., timely graduation or dropping out) based on the data of 4,427 US college students using employing event history analysis. They found that the risk of dropping out among first-generation students was significantly higher than that of their counterparts. [22] applied a time-inclusive event history model to study the most important features of timely college graduation and student dropout based on the demographic, academic, financial, and attitudinal data of 2,373 US students. They found GPA to be positively related to timely graduation, and students with high GPA were less likely to drop out than students with lower GPA.

A vast amount of data created within learning management systems (LMS) has been applied, for example, to analyze specific features that influence students' academic achievements [41], fed to through various machine learning methods to predict dropouts in e-learning courses and massive open online courses (MOOCs) [12–14,42] and to design interventions that increase students' engagement in e-learning courses [43]. The majority of LMS-related dropout research, however, has focused on predicting dropouts within a course [15–17,27,44] or within multiple courses [45]. For example, [46] used various types of LMS data to predict course dropouts on a weekly basis so that course instructors could intervene effectively and promptly, while Baneres [60] studied course dropout prediction performance using the LMS data of 889 online courses at an all-online-university. When the focus shifts to predicting dropouts at the higher education degree program or university levels, however, the number of publications that have used LMS data falls drastically.

The few studies that have linked LMS data to dropout predictions at the degree program level include the following. [18] studied the LMS data of 362 Israeli students and found that 66% of the students who were labeled as at-risk did not finish their course or graduate. [20] analyzed the Moodle activity data of 36 laboratory engineering and 47 ICT engineering students collected during their first year at a Finnish university of applied sciences. Using LMS data, the results successfully flagged 17 of the 32 students who actually dropped out of their degree programs. [19] utilized the LMS, demographic, and transcript data of 98,685 online university students and predicted dropouts with RF, SVM, and NN. They achieved 95% prediction precision and 76% recall. However, they did not perform any kind of feature importance analysis. In fact, none of these LMS-related dropout prediction studies compared the relative importance of the LMS data to that of other data sources.

Table 1 summarizes the most relevant research included in the current literature review. As this table shows, though previous studies have analyzed the most important dropout-related features, they have not used LMS data. Moreover, the only paper to have used demographic, transcript, and LMS data [19] did not include any kind of feature importance analysis.

To identify the longitudinal aspects of students' dropouts, some studies have performed analyses at the end of multiple semesters [21–26,38]. However, very little attention has been paid to studying how data features behave as a function of time, such as mid-semester. The only study which has analyzed mid-semester data at three different time points in a semester, but they all occurred within the first semester [33]. Additionally, no studies have included LMS data to make student dropouts prediction at different time points of studies.

Finally, in recent years, scholars in Finland have also researched student dropouts in Finnish universities. Some of this research has assessed dropping out from other angles to help explain this phenomenon, such as the reasons behind students' dropping out [47,48], student tutoring at Finnish universities [49], student burnout [50], and students' transition from their first year to their second year [51]. Some scholars have applied machine learning to predict student dropouts. [52] used LR to predict dropouts using the data of 20,000 students from four Finnish universities, achieving a classification rate of 87.4%. They identified "gender" and "starting age" as the most important variables.

3. Methodology

3.1. Background on Finnish higher education

As in other countries, student dropout is a major challenge for Finnish universities. For example, in the Finnish universities of applied sciences from 2001 to 2007, the average dropout percentage was 9.0%; an average of 9,622 students dropped out of their programs every year [47]. Higher education is tuition-free for Finnish citizens and residents of Finland. Finnish universities' current financing model is based exclusively on their number of graduations. For example, if a student drops out after three years of his or her studies, though the university has already allocated considerable resources for that student's education, the university gains no financial support from the Finnish government based on that student. Thus, student dropouts cause serious financial losses for Finnish universities. In Finland, degree program students at universities of applied sciences aim to graduate after three and a half, four, or four and a half years for 210, 240, or 270 ETCS degree programs, respectively. Students must apply to extend their study times when these periods end.

Additionally, in the Finnish higher education system, students can earn credits at any time of year—not just at the end of a semester. For example, accumulated credits, GPAs, and numbers of failed courses are updated daily, providing transcript data throughout a semester.

Following the Finnish education system, in the current research, we defined *a dropout* as a student who either actively discontinued his or her studies or whose study time ended before his or her graduation without

Table 1
Summary of the previous research analyzing the most important features of dropout predictions at the degree program level.

Author	Amount of data	Data types	Algorithms	Performance	Feature importance
[4]	3,339	demographic, transcript	DT, LR, RF	accuracies around 80%	the percentage of lost credit vouchers in the last semester
[35]	31,071	demographic, transcript, academic history	DT, GLM, RF	AUC scores from 0.87 to 0.96	accumulated credits in the first year
[18]	362	LMS	–	precision 66%	in-semester clickstream (Moodle activity count)
[5]	16,066	demographic, transcript, academic history, financial	DT, LR, NN, SVM	precision from 75% to 87%	earned hours divided by registered hours, student loan at spring, fall GPA
[38]	7,800	demographic, transcript, academic history, financial	DT, NB, NN, rule induction	precision from 66% to 74%, recall from 24% to 52%	academic performance
[37]	10,196	demographic, transcript academic history	GBT, NN XGB	precision from 67% to 86%, recall from 74% to 81%	credits index, credits earned (accumulated credits), age at the enrollment
[33]	4,433	demographic, transcript, social-economic	EE, RB, RF, SMOTE	f1-scores from 58% to 66%	accumulated credits, but varies with time at 3 points within the first semester
[7]	50,095	demographic, transcript, behavioral engagement	EN, RF	AUC scores from 0.65 to 0.79	GPA, campus engagement
[8]	6,398	demographic, academic history	CAT	AUC score 0.774	high school GPA, math score, years elapsed between high school graduation and university enrollment
[19]	98,685	demographic, transcript, LMS	NN, RF, SVM	precision 95%, recall 76%	–
[34]	36,000	demographic, transcript, attendance	DT, LightGBM, LR, RF, SVM, XGB	precisions from 72% to 83%	number of scholarships, tuition fee, access year
[20]	83	LMS	–	recall 100%	number of Moodle log events (Moodle activity count)
[36]	93,457	demographic, transcript, financial	GBT, LR	precision 84%, recall 54%	features gender, first-generation college student, underrepresented minority and high financial need are not important

an application for or the granting of an extension. In the context of the Finnish universities of applied sciences, most dropout cases are the latter type, occurring on the final day of a semester. Their study time concludes on the semester's final day before their planned graduation, and they have not applied for or been granted additional study time. For example, according to the data set used in this study, 82% of dropout cases occurred on either July 31 (60%) or December 31 (22%), the dates when semesters end at Finnish universities. Thus, if student dropouts are predicted only at the end of a semester, many at-risk students actually drop out, and universities may miss the chance to intervene by helping or motivating these students to continue their education. If, however, predictions can be made during a semester, universities can intervene, retaining students before the semester ends.

3.2. Data description

This study's data included the demographics and transcripts of degree program students who had enrolled at the university in 2010–2020. Unfortunately, the LMS (Moodle) data were only available from January 1, 2015. To compare the relative importance of the different data sets, data completeness was required. Therefore, the data were filtered to include only students who had either graduated or dropped out in 2015–2020. In total, we obtained data from 8,813 students, of whom 6,664 (76%) had graduated and 2,149 (24%) had dropped out. Fig. 1 visualizes the dropouts' cumulative distribution versus the time from enrollment. As this figure shows, a considerable number of dropout cases occurred during later program stages; this trend has also been evident at universities outside of Finland [34]. Moreover, only 71% of the students represented by the current study's data set had graduated at the planned times.

Data features were selected based on a review of the previous literature and the data's availability. Many demographic and transcript data features—such as "gender", "degree program", "age at enrollment", "accumulated credits", and "GPA"—have been used extensively in previous research (see Table 1). Recent studies have also argued that LMS data are important in predicting dropouts [46]. Especially, the literature has shown that a student's Moodle activity count—a simple count of every action the student has taken in Moodle, including logging in, opening links, returning assignments, commenting, and discussing—is very significant to dropout prediction since it depicts the student's overall learning activity on Moodle [18,20]. Following the literature, in this study, we used students' demographic, transcript, and LMS (Moodle) data to predict dropouts. These data were available from the case university. The data were collected with approval from the case university's data protection office and in compliance with data protection guidelines. Table 2 presents more details on the data features, including the categorical feature distributions.

Table 2
Short description of the data features.

	Feature	Values (and distributions in categorical variables)
Time-invariant	Gender	Categorical: male (51%), female (49%)
	Former education level	Categorical: high school (55%), vocational education (31%), other (14%)
	Application priority[a]	Categorical: top priority (69%), not top priority (31%)
	Degree program	Categorical: engineering (33%), nursing (17%), business (15%), hospitality (5%), data processing (5%), other 210 ETCS programs (14%), other 240 ETCS programs (12%)
	Mother language	Categorical: Finnish (94%), not Finnish (6%)
	Start semester	Categorical: autumn (89%), spring (11%)
	Age at enrollment	Numerical, decimal: years
	Output	Categorical: graduation (76%), dropout (24%)
Time-varying	Amount of months present[b]	Numerical, integer: in months (1–60)
	Amount of months absent	Numerical, integer: in months (1–60)
	Accumulated credits (ETCS)	Numerical, integer
	GPA (grade point average)[c]	Numerical, decimal: 1–5
	Number of failed courses	Numerical, integer
	Number of exchange days	Numerical, integer
	Moodle activity count[d]	Numerical, integer
	Moodle activity trend[e]	Categorical: decreasing (43%), increasing (37%), stagnating (20%)

[a] In the Finnish higher education enrollment process, the students prioritize their enrollment options.
[b] To standardize the results, all the students who started their studies in autumn were labeled as starting at the first of August, and all the students who started at spring, first of January. This feature was only used in preprocessing.
[c] GPA uses the data of all previous passed courses, not just the passed courses of the current semester
[d] Moodle activity: log-ins, link openings, completed tasks, etc., also referred to as click stream
[e] Calculated from Moodle activity count by comparing the count of the current month to the previous month, with $\pm 10\%$ threshold

3.3. Data pre-processing, feature engineering, and classifier models

The data set was pre-processed before the data were analyzed. The original time series data were reindexed so that every student had exactly one entry of data features for every month of their study time (the time from enrollment), dated with the final day of the month.

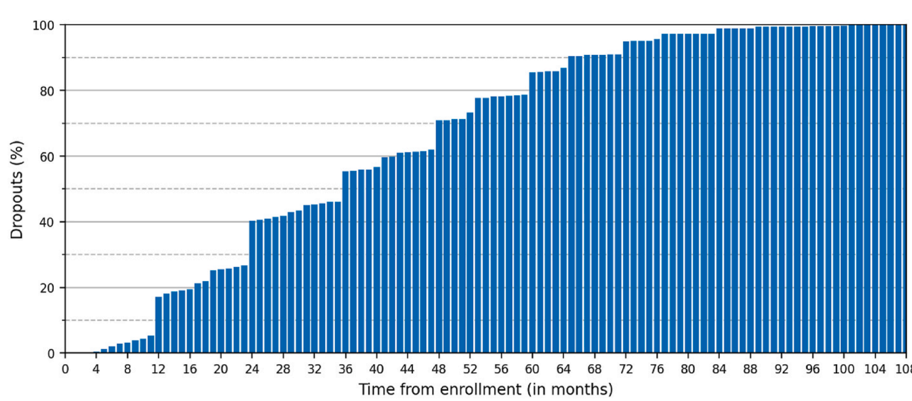

Fig. 1. The cumulative distribution of dropout versus time.

"Moodle activity trend" was calculated by comparing the "Moodle activity count" of the previous month to the current month. If the percentage changes exceeded ±10%, the "Moodle activity trend" was labeled "increasing" or "decreasing," and if the changes were less than or equal to ±10%, they were labeled as "stagnating." Moreover, some variables were normalized. "Amount of months absent" was normalized by dividing the amount of months absent by the time from enrollment. "Accumulated credits" was normalized by dividing the number of accumulated credits by semester number multiplied by 30; that is, the accumulated credits of the first semester were divided by 30, the accumulated credits of the second semester were divided by 60, and so on. "Number of failed courses" was similarly divided by the number of semesters. Finally, the input features' Spearman correlations (not assuming a normal data distribution) were calculated. The results are displayed as a heatmap in Fig. 2.

For every given month present—or the time from enrollment, t, in months where months of absence did not count—between 1 and 60, all students' corresponding monthly entries formed a data subset, D_t. That is, every D_t data set included the time-invariant variables and the time-variant variables at time t. If a student had graduated or dropped out before a given month, t, the data were not included in D_t. Moreover, if a student had a missing value in any of the data features, the entry of the student was removed from the data subset, D_t; that is, the data were complete, and no values were imputed. Every data set, D_t, was then split for the training and test data sets, following the idea of the cohort test set as described in [8,36]. That is, the test set was chosen so that it included only students who had enrolled in 2015, and the training set included all the other students. In total, the test set comprised 1,834 students, of whom 22% dropped out. The training set comprised 6,989 students, of whom 25% dropped out.

After the split, the input categorical features were one-hot-encoded. That is, for each category of the feature, a binary column of ones and zeros was created, with ones denoting that the entry belonged in the category. To avoid collinearity, the first binary column of every feature category was removed. After one-hot encoding, the numerical features were standardized by removing the mean and scaling to the unit variance. That is, the samples, x, of a numerical feature in both the training and test datasets were transformed via $x = \frac{x-u}{s}$, where u and s represent the mean and variance of the numerical feature in the training set, respectively. Standardizing only with the training data set mean and variance ensured that no information leaked from the test set to the training set.

Finally, the classifier models were trained with the training data set and evaluated with the test data set for every data subset, D_t. That is, every classifier model was separately trained for every data set, D_t, totaling 60 trainings for every classifier. Ten different classifier models were employed: Logistic regression, Random Forest, Support vector machine, Linear discriminant analysis, K-nearest neighbors, Naïve Bayes, Neural networks, xgboost, CatBoost and lightgbm. Of these models, the three best-performing models were selected for a feature importance comparison. The classifier algorithms were provided by scikit-learn [53]. The default values of the models' hyper-parameters were used, and no threshold optimization was performed for the following reasons. First, this study aimed not to achieve the highest prediction performance but to analyze the relative importance of LMS, transcript, and demographic data. Second, we sought to analyze how the models and feature importance varied over time. Moreover, since the analysis involved 60 times 10 different trained models, no hyper-parameter tuning was performed to ensure the method's simplicity and repeatability.

3.3.1. Model performance evaluation

Model performance was evaluated using the following metrics: precision, recall (also called sensitivity), and their harmonic mean, f1-score. Precision is defined by $\frac{TP}{TP+FP}$, recall is defined by $\frac{TP}{TP+FN}$, and f1-score is defined by $2 \cdot \frac{precision \cdot recall}{precision \cdot recall}$, where TP denotes for true positives (in this study: correctly predicted dropouts), FP for false positives (incorrectly predicted dropouts), and FN false negatives (incorrectly predicted graduations). Thus, prediction is a metric that evaluates how accurate the dropout predictions of the model are, and recall is a metric that evaluates the amount of dropout predictions the model can return.

We selected precision and recall as the model's performance metrics for the following reasons. First, they are easy to interpret. Second, the model's purpose is to identify potential dropouts—not potential graduates. Precision and recall provide information about the model's

Fig. 2. Spearman correlations of the entire data set's input variables.

performance in a class (dropouts), whereas accuracy would evaluate the model's performance in both classes (dropouts and graduates). Moreover, the data set was imbalanced (24% dropouts and 76% graduates); therefore, an overall performance metric could provide misleading information about the model's performance with a class. For example, a dummy model predicting every student's graduation would achieve an accuracy of 76%.

The 10 different machine learning models' prediction performance was initially compared using the area under the receiver operating characteristic curve (AUC) score. This score was first calculated for every month, t, and then averaged. The higher the AUC score, the better the models' predictive performance. Moreover, a closely related metric, average precision (AP), was similarly computed for all the models at every time, t.

To represent the trade-off between the model's precision and recall, precision–recall curves were computed for each trained model, M_t. Precision–recall curves are useful when classes are imbalanced because they focus only on one class [54]. A high AUC would represent high precision and high recall; that is, it would suggest that the model's results were accurate and that most of the positive results (dropouts) were returned. High precision but low recall would accurately predict a low number of dropouts accurately. High recall and low precision would suggest the opposite: a high number of results but low accuracy.

We include this note for readers who are unfamiliar with the threshold concept in classification problems. Many classifier models return the prediction probability, a decimal between 0 and 1, for each instance of the test set. This instance is further classified as *positive* (a dropout in our case) if the model returns a decimal greater than the chosen threshold value. Similarly, if the decimal is smaller than the threshold value, the instance is classified as negative (a graduate). The default threshold is usually 0.5, but it can be changed. Changing the threshold can achieve different precision and recall values for the same model. This trade-off is visualized with a precision–recall curve. In the current study, we used a classification threshold of 0.6 for all models to slightly favor precision over recall. In practice, this would result in favoring precise interventions to the number of at-risk students identified.

3.3.1.1. Feature importance. To identify the most important features that contribute to dropout predictions, we used the permutation importance method. Permutation importance is defined as a decrease of a trained model's performance score when a feature's values are randomly shuffled and used to make predictions [53,55]. This shuffling makes the chosen feature's data meaningless, thus decreasing the performance score. The more the score has decreased, the greater the feature's importance to the model.

In this study, the permutation importance score was chosen to be the f1-score because it summarizes both the precision and recall of the dropout class. Therefore, the permutation importance score, p, of a feature corresponds to a drop of p in the f1-score. Accordingly, a feature's high permutation importance indicates that the feature is important to the model's dropout prediction, and vice versa.

To make the permutation importance scores more robust, the data of the feature were randomly shuffled five times with different permutations. Then, the average of the permutation importance scores was calculated. This process was repeated for every feature of every data set, D_t. In the current study, the categorial features were shuffled before one-hot encoding. Therefore, the categorical features' permutation importance scores yield insights into the features themselves, rather than their individual categories.

Features' multicollinearity can greatly affect permutation importance analysis [53]. If the features are collinear, the model's performance can remain the same—even if a feature's data are permutated—because the model can obtain the same information from a correlated feature. In the worst case, permutation importance would show that none of the features were important. Fig. 2 presents the features' Spearman correlations.

Finally, permutation importance does not indicate a feature's importance in itself; rather, it indicates the feature's importance to the particular model. Therefore, a feature of low importance to a certain model might be highly important to a different model, and vice versa [53]. Nevertheless, the permutation importance method is model-independent, and it can be applied even to nonopaque models. This, in turn, enables a comparison of different models' permutation importance scores. Therefore, comparing several models' feature importance can provide insights into features' importance in themselves.

4. Results

The prediction performance scores—AUC, AP, f1-score, precision, and recall, with dropouts as the positive class—were calculated for each classifier and for each individual time, t, in the 60-month interval. The scores' averages were then calculated for every model, and Table 3 presents the results. As Table 3 shows, the three best-performing models by AUC score were CAT, NN, and LR. These models were chosen for further analysis. The analysis results are presented visually in the following three figures (Figs. 3–5).

The three classifier models' performance metrics are visually summarized in Fig. 3, where the metrics are plotted against the time from enrollment. As Fig. 3 shows, the general trend is that the prediction performance is increasing with time. In the first four months after students' enrollment, the models' prediction performance was low. This result could be explained by the low amount of data that the students had accumulated. In Months 5–11, the precision and recall both are rapidly improving. In the time interval of 12–53, the precision ascends roughly from 70% to 90% and recall from 40% to 60%. From Month 54 onwards, the recall descends rapidly. The f1-scores, which summarize both the precision and recall of the three models, showed a similar trend; low values were observed for the first four months, versus a rapid increase from Months 5–11, and a slightly rising trend up to 53 months, after which, the scores descent rapidly. A comparison of the f1-score curves indicated a similar performance among the three models. Additionally, the prediction performance of all three models increased within every year or semester. Thus, the f1-score was higher toward the end of the year in the intervals of 1–12, 13 to 24, 25 to 36, and—perhaps to a lesser extent—37 to 48 months. The only year during which the prediction performance decreased was the fifth year, from Months 49–60.

To identify the most important features, the mean of the permutation importance scores over the time interval of 1–60 months was computed for all the features and all three models. Table 4 presents these results. As the table shows, three of the features clearly stood out: "accumulated credits", "number of failed courses" and "Moodle activity count".

To visualize the most important features' time series behavior, the permutation importance scores (f1-scores, with dropouts being the positive class) were plotted versus the time from enrollment in Fig. 4. To make this figure easier to read, only the three most important features

Table 3
Average prediction performance scores over the 60-month interval for all ten machine learning classifiers.

Classifier	AUC	AP	f1 (%)	Precision (%)	Recall (%)
CAT	0.853	0.721	59	81	47
NN	0.844	0.698	58	76	47
LR	0.842	0.696	56	82	42
LDA	0.841	0.682	53	82	39
RF	0.839	0.700	57	82	43
LGBM	0.832	0.696	59	78	48
XGB	0.823	0.681	59	75	49
SVM	0.823	0.693	56	81	44
NB	0.780	0.558	52	52	58
KNN	0.777	0.566	55	72	45

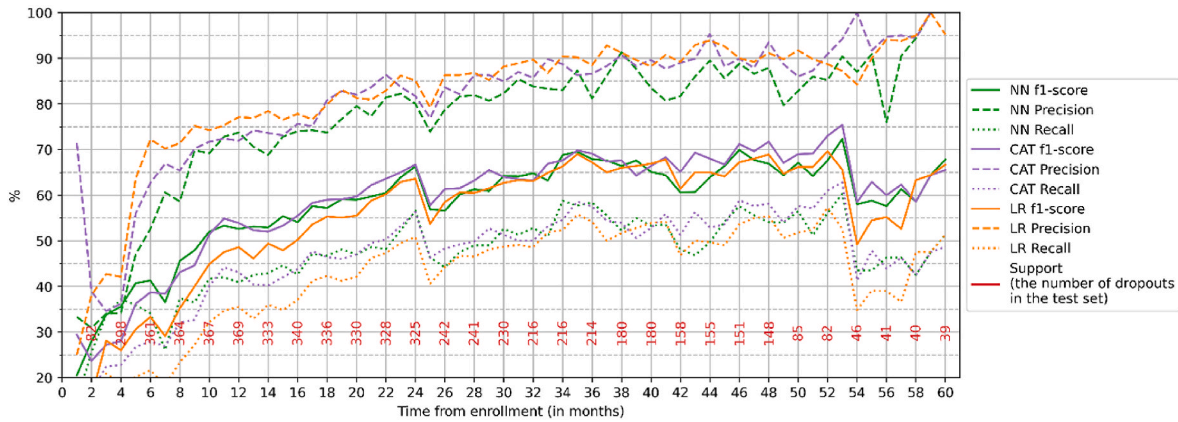

Fig. 3. The prediction performance scores (precision, recall, and f1-score with dropout being the positive class) of the three classification models versus time. An example of how to read the curves: at $t = 12$, the CAT model classified 369 dropouts with the precision of 72%, recall 43%, and f1-score 54%.

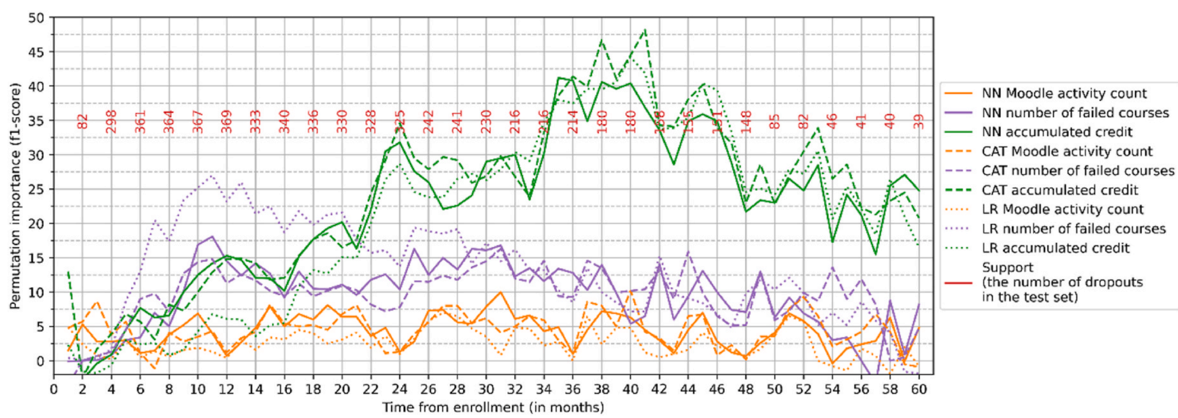

Fig. 4. The permutation importance scores (f1-score with dropout being the positive class) of the 3 most important features of the three classifier models versus study time. An example of how to read the graph: at $t = 24$, shuffling the test set data (of 325 dropouts) of the feature "accumulated credits" would result in a loss of 35% f1-score of the CAT model.

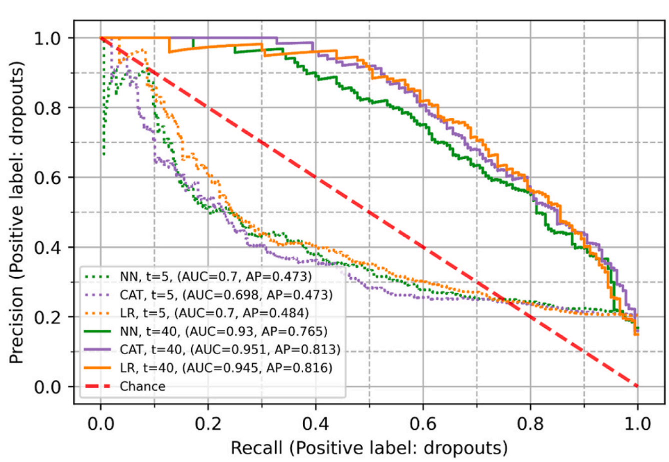

Fig. 5. Precision-recall curves of the three classifier models at two different times. At $t = 5$, the model performances are at their worst and at $t = 40$ are at their best (excluding months 1–4). As an example of how to read the curves, at $t = 40$, changing the threshold value of the CAT model to raise precision from 80% to 90% would result in a fall of recall from 60% to 52% in dropout prediction.

are shown. As Fig. 4 shows, "accumulated credits" was the most important feature for all three models from Month 21 onward, especially during the fourth year. This finding suggests that this feature's

Table 4
Means of all analysis features' permutation importance scores over the time interval 1–60 month.

Feature	permutation importance CAT	permutation importance NN	permutation importance LR
Accumulated credits (ETCS)	24.0	22.2	21.2
Number of failed courses	9.4	9.4	12.8
Moodle activity count	4.5	4.4	2.5
Gender	0.9	1.2	1.3
Degree Program	0.7	1.3	1.2
Moodle activity trend	0.4	0.9	0.6
Number of exchange days	0.3	0.3	0.1
Application priority	0.2	−0.1	0.3
Age at enrollment	0.2	0.1	1.5
Mother language	0	−0.1	0
Amount of months absent	−0.1	0.2	0.1
Former education level	−0.2	−0.1	0.4
Start semester	−0.2	−0.2	−0.1
GPA	−0.4	0	0.5

importance to dropout predictions is high, regardless of the model. Moreover, as the figure shows, the relative importance of "accumulated credits" significantly fluctuated over time. It increased up to Month 41 and then decreased from that month onwards. LR most emphasized "the number of failed courses" for the first 20 months, after which "accumulated credits" became the most important feature. This trend, however, was not observed with CAT or NN, for which "accumulated credits" was the most important feature from as early as 12 months, and "number of failed courses" was the second most important feature. Furthermore, all three models showed that the importance of "number of failed courses" slowly decreased over time. For all three models, the "Moodle activity count" was the third most important feature over a wide time range, and the most stable of the three features in terms of their fluctuations over time.

To display the trade-off between precision and recall, we plotted the three models' precision–recall curves at two different times, t, in Fig. 5. Because showing the curves for every month would not have been practical, $t = 5$ and $t = 40$ were chosen since, at $t = 5$, the models performed at their worst, and at $t = 40$, they performed at their best (excluding the first four months' performances because, as Fig. 3 shows, the performance scores were low during this interval). As Fig. 5 shows, at $t = 5$, the models' performance remained low.

5. Discussion and conclusions

In this study, we examined the dropout prediction data features' importance in the three machine learning models (CAT, NN, and LR). We used the model-independent permutation importance method as a function of time from enrollment with the demographic, transcript, and LMS data of 8813 students from the Finnish case university. The most important features of all three models were "accumulated credits," "the number of failed courses," and "Moodle activity count".

The current study merged LMS (Moodle) data with transcript and demographic data and compared their relative importance in degree program dropout predictions. Our findings suggest that LMS data are indeed important and should be considered in degree program dropout predictions alongside course dropout predictions. Our finding further validates the similar findings of [18,20] although their works did not include transcript or demographic data, and their data sets were small. Moreover, our finding that "accumulated credits" and "the number of failed courses" were the most important features is consistent with prior research findings [33,35,37]. However, although GPA has been identified as a major predictor, especially in the US context [5,7,33,40,56], this importance was not observed in the current study. We found that GPA had little importance to the predictive power of all the three models. Finally, in alignment with our findings concerning the Finnish system, the research of [52] has shown that "gender" is one of the most important features in dropout predictions. In our study, however, "gender" was not included among the three most important features.

Concerning feature importance's time-dependency, generally, we found that some features' importance changed significantly over time. Particularly during the interval from 21 months to 42 months, the permutation importance of "accumulated credits" drastically increased, from about 15% to 45%, for all the models. Moreover, the features' relative importance or rank also changed over time. For example, the "number of failed courses" was the most important feature for LR in the first 20 months, but from 21 to 60 months, "accumulated credits". was the most important feature. Moreover, although "number of failed courses" and "accumulated credits" were fairly equally important for CAT and NN up to 16 months, "accumulated credits" then became by far the most important feature.

Another of this study's contributions is its extension of prediction performance and feature importance analysis to multiple time points within semesters. Our findings suggest that some features' importance indeed fluctuates significantly over time intervals shorter than a semester. For example, the permutation importance of "accumulated credits" increased from about 17% to 30% within just four months (from 20 to 24 months) for all three models. Moreover, this feature's importance steeply fell—from 45% to 35%—in just one month at 41 months. Prediction performance also fluctuates in time intervals shorter than semesters. For example, the performance of all the models showed a clear increasing trend within the first three years or semesters. That is, f1-score increased toward the end of the year during the first three years. This finding aligns with the results of previous research showing that prediction precision increases toward the end of the enrollment year [33]. Moreover, generally over longer periods, our results confirmed that, as time progresses, machine learning models' prediction performance fluctuates [5,25,38].

From a theoretical perspective, the current study's findings contribute to the literature on student dropout predictions in the following ways. First, this study enriches the literature on student dropout predictions by having compared the relative importance of demographic, transcript, and LMS data. Our findings suggest that LMS data features have relatively high predictive power and should be considered as input variables in dropout predictions. Second, we analyzed the predictive features' importance as a function of students' study durations monthly, within semesters. These findings confirm that the data features' importance—and even their relative importance or rank—change over time, and they may steeply increase or decrease in time intervals shorter than a semester. Third, the classification models' prediction performance was, likewise, analyzed as a function of time. The results provide insights into how reliably dropouts can be predicted at any given time, rather than just at the end of a semester.

From a practical perspective, the current study offers suggestions for higher education institutions. As the results show, machine learning models can precisely predict a large portion of dropouts using the demographic, transcript, and LMS data available to universities—at least in the interval of 10–53 months from students' enrollment. These findings provide some practical insights into the intervention time and strategies needed to prevent students from dropping out of Finnish universities. While some retention strategies aim to identify at-risk students and intervene as early as possible, other retention strategies are also available. Interventions can occur as late as the fourth or fifth year in Finnish universities for at least two reasons. First, late-stage at-risk students have already shown their study motivation for years and probably accumulated the majority of credits they need to graduate. Intuitively, therefore, such students seem easier to motivate and help toward graduation than early-stage at-risk students who must still accumulate over 200 credits. The second reason is financial. Finnish universities are financially supported by the Finnish government, based exclusively on their number of student graduations, and most students at Finnish universities do not pay tuition fees. These universities have allocated considerable resources to educate at-risk late-stage students. Therefore, these universities are highly motivated to help such students graduate so that these institutions can receive financial support from the Finnish government based on these students' graduation. Additionally, since student retention campaigns must be implemented mid-semester, this study's findings can provide insights into when the predictions for such campaigns would be the most precise and which features to consider the most when planning retention strategies.

6. Limitations and future research directions

The current study faced some limitations. First, we only considered demographic, transcript-related, and LMS-related features as inputs in our prediction models. However, evidence suggests that data features from different sources—such as surveys—might also be important to dropout predictions [21,26,57–59]. Accordingly, in future studies, researchers could consider applying other features to the prediction models in order to determine whether their inclusion could improve their predictions. Moreover, this study used only two LMS data features, "Moodle activity count" and "Moodle activity trend", —although evidence suggests that other LMS data features might be important as well, at least in course dropout predictions [46]. Finally, the current study's data were collected from a single Finnish higher education institution, so the results might not be generalizable to other countries or institutions.

CRediT authorship contribution statement

Matti Vaarma: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Hongxiu Li:** Supervision, Writing – review & editing.

Data availability

The data that has been used is confidential.

References

[1] V. Tinto, Dropout from higher education: a theoretical synthesis of recent research, Rev. Educ. Res. 45 (1) (1975) 89–125, https://doi.org/10.3102/00346543045001.

[2] V. Tinto, Research and practice of student retention: what next? J. Coll. Stud. Retent: Res. Theory Pract. 8 (1) (2006) 1–19, https://doi.org/10.2190/4YNU-4TMB-22DJ-AN.

[3] J. McFarland, B. Hussar, J. Zhang, X. Wang, K. Wang, S. Hein, M. Diliberti, E. F. Cataldi, F.B. Mann, A. Barmer, The Condition of Education 2019. 144, National Center for Education Statistics (US), 2019.

[4] P. Berka, L. Marek, Bachelor's degree student dropouts: who tend to stay and who tend to leave? Stud. Educ. Eval. 70 (2021) 100999 https://doi.org/10.1016/j.stueduc.2021.100999.

[5] D. Delen, A comparative analysis of machine learning techniques for student retention management, Decis. Support Syst. 49 (4) (2010) 498–506, https://doi.org/10.1016/j.dss.2010.06.003.

[6] S. Maldonado, J. Miranda, D. Olaya, J. Vásquez, W. Verbeke, Redefining profit metrics for boosting student retention in higher education, Decis. Support Syst. 143 (2021) 113493, https://doi.org/10.1016/j.dss.2021.113493.

[7] S.C. Matz, C.S. Bukow, H. Peters, C. Deacons, A. Dinu, C. Stachl, Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics, Sci. Rep. 13 (2023) 5705, https://doi.org/10.1038/s41598-023-32484-w.

[8] M. Nagy, R. Molontay, Interpretable dropout prediction: towards XAI-based personalized intervention, Int. J. Artif. Intell. Educ. (2023), https://doi.org/10.1007/s40593-023-00331-8 available online).

[9] D. Olaya, J. Vásquez, S. Maldonado, J. Miranda, W. Verbeke, Uplift modeling for preventing student dropout in higher education, Decis. Support Syst. 134 (2020) 113320, https://doi.org/10.1016/j.dss.2020.113320.

[10] J.L. Rastrollo-Guerrero, J.A. Gómez-Pulido, A. Durán-Domínguez, Analyzing and predicting students' performance by means of machine learning: a review, Appl. Sci. 10 (3) (2020) 1042, https://doi.org/10.3390/app10031042.

[11] C. Romero, S. Ventura, E. García, Data mining in course management systems: Moodle case study and tutorial, Comput. Educ. 51 (2008) 368–384, https://doi.org/10.1016/j.compedu.2007.05.016.

[12] R.B. Basnet, C. Johnson, T. Doleck, Dropout prediction in Moocs using deep learning and machine learning, Educat. Inform Technology 27 (2022) 11499–11513, https://doi.org/10.1007/s10639-022-11068-7.

[13] C. Jin, MOOC student dropout prediction model based on learning behavior features and parameter optimization, Interact. Learn. Environ. 31 (2) (2023) 714–732, https://doi.org/10.1080/10494820.2020.1802300.

[14] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, V. Loumos, Dropout prediction in e-learning courses through the combination of machine learning techniques, Comput. Educ. 53 (2009) 950–965, https://doi.org/10.1016/j.compedu.2009.05.010.

[15] L.P. Macfadyen, S. Dawson, Mining LMS data to develop an "early warning system" for educators: a proof of concept, Comput. Educ. 54 (2010) 588–599, https://doi.org/10.1016/j.compedu.2009.09.008.

[16] N. Nistor, K. Neubauer, From participation to dropout: quantitative participation patterns in online university courses, Comput. Educ. 55 (2010) 663–672, https://doi.org/10.1016/j.compedu.2010.02.026.

[17] M.M. Tamada, R. Giusti, J.F. de M. Netto, Predicting students at risk of dropout in technical course using LMS logs, Electronics 11 (3) (2022) 468, https://doi.org/10.3390/electronics11030468.

[18] A. Cohen, Analysis of student activity in web-supported courses as a tool for predicting dropout, Educ. Technol. Res. Dev. 65 (2017) 1285–1304, https://doi.org/10.1007/s11423-017-9524-3.

[19] H.S. Park, S.J. Yoo, Early dropout prediction in online learning of university using machine learning, Int. J. Inform. Visualiz. 5 (4) (2021) 347, https://doi.org/10.30630/joiv.5.4.732.

[20] S. Suhonen, H. Kinnari-Korpela, Using Moodle data for early warning of dropping out, in: 47th Annual Conference Proceedings of European Society for Engineering Education, 2019. September 2019, Budapest, Hungary.

[21] J. Bayer, H. Bydžovská, J. Géryk, T. Obšívač, L. Popelínský, Predicting drop-out from social behaviour of students, in: Proceedings of the 5th International Conference on Educational Data Mining, EDM 2012, Berlin, 2012.

[22] S.L. DesJardins, D.A. Ahlburg, B.P. McCall, A temporal investigation of factors related to timely degree completion, J. High Educ. 73 (5) (2002) 555–581, https://doi.org/10.1080/00221546.2002.11777168.

[23] T.T. Ishitani, Studying attrition and degree completion behavior among first-generation college students in the United States, J. High Educ. 77 (5) (2006) 861–885, https://doi.org/10.1080/00221546.2006.11778947.

[24] L. Kemper, G. Vorhoff, B.U. Wigger, Predicting student dropout: a machine learning approach, Eur. J. High Educ. 10 (1) (2020) 28–47, https://doi.org/10.1080/21568235.2020.1718520.

[25] M. Solis, T. Moreira, R. Gonzalez, T. Fernandez, M. Hernandez, Perspectives to Predict Dropout in University Students with Machine Learning, 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), 1–6, San Carlos, 2018, https://doi.org/10.1109/IWOBI.2018.8464191. IEEE.

[26] S. Wild, L. Schulze Heuling, Student dropout and retention: an event history analysis among students in cooperative higher education, Int. J. Educ. Res. 104 (2020) 101687, https://doi.org/10.1016/j.ijer.2020.101687.

[27] R. Cerezo, M. Esteban, M. Sánchez-Santillán, J.C. Núñez, Procrastinating behavior in computer-based learning environments to predict performance: a case study in Moodle, Front. Psychol. 8 (2017) 1403, https://doi.org/10.3389/fpsyg.2017.01403.

[28] F. Dalipi, A.S. Imran, Z. Kastrati, MOOC dropout prediction using machine learning techniques: review and research challenges, in: IEEE Global Engineering Education Conference, vols. 1007–1014, 2018, https://doi.org/10.1109/EDUCON.2018.8363340 anta Cruz de Tenerife, Spain.

[29] B.J. Evans, R.B. Baker, T.S. Dee, Persistence patterns in massive open online courses (MOOCs), J. High Educ. 87 (2) (2016) 206–242, https://doi.org/10.1080/00221546.2016.11777400.

[30] M. Fei, D.-Y. Yeung, Temporal Models for Predicting Student Dropout in Massive Open Online Courses, 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 256–263, Atlantic City, NJ, US, 2015.

[31] Y. Liu, S. Fan, S. Xu, A. Sajjanhar, S. Yeom, Y. Wei, Predicting student Performance using clickstream data and machine learning, Educ. Sci. 13 (1) (2022) 17, https://doi.org/10.3390/educsci13010017.

[32] G.W. Dekker, M. Pechenizkiy, J.M. Vleeshouwers, Predicting students drop out: a case study, in: The 2nd International Conference on Educational Data Mining, vols. 41–50, 2009. Cordoba, Spain.

[33] M.V. Martins, L. Baptista, J. Machado, V. Realinho, Multi-class phased prediction of academic performance and dropout in higher education, Appl. Sci. 13 (8) (2023) 4702, https://doi.org/10.3390/app13084702.

[34] Z. Song, S.-H. Sung, D.-M. Park, B.-K. Park, All-year dropout prediction modeling and analysis for university students, Appl. Sci. 13 (2) (2023) 1143, https://doi.org/10.3390/app13021143.

[35] M. Cannistrà, C. Masci, F. Ieva, T. Agasisti, A.M. Paganoni, Early-predicting dropout of university students: an application of innovative multilevel machine learning and statistical techniques, Stud. High Educ. 47 (9) (2022) 1935–1956, https://doi.org/10.1080/03075079.2021.2018415.

[36] R. Yu, H. Lee, R.F. Kizilcec, Should college dropout prediction models include protected attributes?, in: Proceedings of the Eighth ACM Conference on Learning @ Scale ACM, Virtual Event Germany, 2021, pp. 91–100, https://doi.org/10.1145/3430895.3460139.

[37] B. Kiss, M. Nagy, R. Molontay, B. Csabay, Predicting dropout using high school and first-semester academic achievement measures, in: 2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA), vols. 383–389, IEEE, 2019, https://doi.org/10.1109/ICETA48886.2019.9040158. Starý Smokovec, Slovakia.

[38] A. Djulovic, D. Li, Towards freshman retention prediction: a comparative study, Int. J. Inform. Educat. Technol. 3 (5) (2013) 494–500.

[39] R. Chen, S.L. DesJardins, Investigating the impact of financial aid on student dropout risks: racial and ethnic differences, J. High Educ. 81 (2) (2010) 179–208, https://doi.org/10.1080/00221546.2010.11779048.

[40] S.F. Porchea, J. Allen, S. Robbins, R.P. Phelps, Predictors of long-term enrollment and degree outcomes for community college students: integrating academic, psychosocial, socio-demographic, and situational factors, J. High Educ. 81 (6) (2010) 680–708, https://doi.org/10.1080/00221546.2010.11779077.

[41] R. Cerezo, M. Sánchez-Santillán, M.P. Paule-Ruiz, J.C. Núñez, Students' LMS interaction patterns and their relationship with achievement: a case study in higher education, Comput. Educ. 96 (2016) 42–54, https://doi.org/10.1016/j.compedu.2016.02.006.

[42] P.M. Moreno-Marcos, P.J. Muñoz-Merino, J. Maldonado-Mahauad, M. Pérez-Sanagustín, C. Alario-Hoyos, C. Delgado Kloos, Temporal analysis for dropout prediction using self-regulated learning strategies in self-paced MOOCs, Comput. Educ. 145 (2020) 103728, https://doi.org/10.1016/j.compedu.2019.103728.

[43] I. Borrella, S. Caballero-Caballero, E. Ponce-Cueto, Taking action to reduce dropout in MOOCs: tested interventions, Comput. Educ. 179 (2022) 104412, https://doi.org/10.1016/j.compedu.2021.104412.

[44] C. Romero, M.I. López, J.M. Luna, S. Ventura, Predicting students' final performance from participation in on-line discussion forums, Comput. Educ. 68 (2013) 458–472, https://doi.org/10.1016/j.compedu.2013.06.009.

[45] L.-V. Poellhuber, B. Poellhuber, M. Desmarais, C. Leger, N. Roy, M. Manh-Chien Vu, Cluster-based performance of student dropout prediction as a solution for large scale models in a Moodle LMS, in: 13th International Learning Analytics and Knowledge Conference, vols. 592–598, ACM, Arlington TX USA, 2023, https://doi.org/10.1145/3576050.3576146.

[46] A.A. Mubarak, H. Cao, W. Zhang, Prediction of students' early dropout based on their interaction logs in online learning environment, Interact. Learn. Environ. 30 (8) (2022) 1414–1433, https://doi.org/10.1080/10494820.2020.1727529.
[47] R. Kalima, Opintojen Pitkittyminen Ja Keskeyttäminen Ammattikorkeakoulussa, University of Tampere, 2011.
[48] J. Mäkinen, E. Olkinuora, K. Lonka, Students at risk: students' general study orientations and abandoning/prolonging the course of studies, High Educ. 48 (2004) 173–188, https://doi.org/10.1023/B:HIGH.0000034312.79289.ab.
[49] E. Kuurila, Uraohjaus Ja Urasuunnittelu Ammattikorkeakouluissa (Career Planning and Counselling at a University of Applied Sciences), University of Turku, 2014.
[50] K. Salmela-Aro, S. Read, Study engagement and burnout profiles among Finnish higher education students, Burnout Res. 7 (2017) 21–28, https://doi.org/10.1016/j.burn.2017.11.001.
[51] V. Korhonen, M. Inkinen, M. Mattsson, A. Toom, Student engagement and the transition from the first to second year in higher education, in: E. Kyndt, V. Donche, K. Trigwell, Lindblom-Ylänne, S. edited (Eds.), Book Higher Education Transitions: Theory and Research, vols. 113–134, 2019, https://doi.org/10.4324/9781315617367-10.
[52] V. Korhonen, J. Rautopuro, Identifying problematic study progression and "at-risk" students in higher education in Finland, Scand. J. Educ. Res. 63 (7) (2019) 1056–1069, https://doi.org/10.1080/00313831.2018.1476407.
[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[54] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, PLoS One 10 (3) (2015) e0118432, https://doi.org/10.1371/journal.pone.0118432.
[55] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[56] T.T. Ishitani, S.L. DesJardins, A longitudinal investigation of dropout from college in the United States, J. Coll. Stud. Retent.: Res. Theory Pract. 4 (2) (2002) 173–201, https://doi.org/10.2190/V4EN-NW42-742Q-2NTL.
[57] S. Gilardi, C. Guglielmetti, University life of non-traditional students: engagement styles and impact on attrition, J. High Educ. 82 (1) (2011) 33–53, https://doi.org/10.1080/00221546.2011.11779084.
[58] E.J. Lizarte Simón, J. Gijón Puerta, Prediction of early dropout in higher education using the SCPQ, Cogent Psychology 9 (1) (2022) 2123588, https://doi.org/10.1080/23311908.2022.2123588.
[59] A. Scheunemann, T. Schnettler, J. Bobe, S. Fries, C. Grunschel, A longitudinal analysis of the reciprocal relationship between academic procrastination, study satisfaction, and dropout intentions in higher education, Eur. J. Psychol. Educ. 37 (2022) 1141–1164, https://doi.org/10.1007/s10212-021-00571-z.
[60] D. Baneres, M.E. Rodriguez, M. Serra, An early feedback prediction system for earners at-risk within a first-year higher education course, IEEE Trans. Learn. Technol. 12 (2) (2019) 249–263, https://doi.org/10.1109/TLT.2019.2912167.