Decision Support Systems 182 (2024) 114229

Contents lists available at ScienceDirect

Decision Support Systems

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

Explainable Learning Analytics: Assessing the stability of student success prediction models by means of explainable AI

Elena Tiukhova [a],[*], Pavani Vemuri [a], Nidia López Flores [b], Anna Sigridur Islind [b], María Óskarsdóttir [b], Stephan Poelmans [a], Bart Baesens [a,c], Monique Snoeck [a]

[a] KU Leuven, LIRIS, Naamsestraat 69, Leuven, 3000, Belgium
[b] Reykjavik University, Department of Computer Science, Menntavegi 1, Reykjavík, 102, Iceland
[c] Department of Decision Analytics and Risk, University of Southampton, University Road, Southampton SO17 1BJ, UK

ARTICLE INFO

Keywords:
Learning analytics
Self-regulated learning
Explainable AI
Model stability

ABSTRACT

Beyond managing student dropout, higher education stakeholders need decision support to consistently influence the student learning process to keep students motivated, engaged, and successful. At the course level, the combination of predictive analytics and self-regulation theory can help instructors determine the best study advice and allow learners to better self-regulate and determine how they want to learn. The best performing techniques are often black-box models that favor performance over interpretability and are heavily influenced by course contexts. In this study, we argue that explainable AI has the potential not only to uncover the reasons behind model decisions, but also to reveal their stability across contexts, effectively bridging the gap between predictive and explanatory learning analytics (LA). In contributing to decision support systems research, this study (1) leverages traditional techniques, such as concept drift and performance drift, to investigate the stability of student success prediction models over time; (2) uses Shapley Additive explanations in a novel way to explore the stability of extracted feature importance rankings generated for these models; (3) generates new insights that emerge from stable features across cohorts, enabling teachers to determine study advice. We believe this study makes a strong contribution to education research at large and expands the field of LA by augmenting the interpretability and explainability of prediction algorithms and ensuring their applicability in changing contexts.

1. Introduction

Understanding student success, what factors contribute to it, and how, remains an open question in learning analytics (LA) and educational data mining (EDM) research. Predictive LA is a rapidly growing area of research that focuses on modeling future learner outcomes. This growth is facilitated by borrowing various techniques from several disciplines such as machine learning (ML) to solve various LA tasks [1,2], with predicting student success being among them [3]. Modern Learning Management Systems (LMS) allow to trace the history of learners' activities. According to [4], "self-regulated learning (SRL) is a behavioral expression of meta-cognitively guided motivation". In line with the SRL theory, studying trace data can be considered a representation of a motivated learning choice and further used to engineer higher-level indicators that represent different aspects of learning [5].

The most popular ML models used to learn from the study data are black-box algorithms that offer high predictive power [3,6], but at the cost of interpretability [7]. One solution is eXplainable Artificial Intelligence (XAI), which focuses on generating understandable explanations for a model's decisions, which is essential for delivering responsible analytics [8]. The use of XAI for LA is a relatively new but rapidly growing area of research, with most applications aiming to improve interpretability through visualization, e.g., to provide personalized instruction [9] or data-driven feedback through learning dashboards [10]. A popular class of XAI techniques are additive feature attribution methods, which provide feature importance rankings and effects [11]. Using these techniques for LA can provide stakeholders with the success factors [12] that allow the promotion of desirable learning behaviors.

However, due to the changing contexts of learning, it is important to ensure not only the explanatory power but also the stability of models across years [13]. Technological advances and changes in learning contexts can lead to a mismatch between historical patterns of learning and current data, known as concept drift [14]. For example, the outbreak of COVID-19 drastically affected the way education

[*] Corresponding author.
E-mail addresses: elena.tiukhova@kuleuven.be (E. Tiukhova), bart.baesens@kuleuven.be (B. Baesens).

https://doi.org/10.1016/j.dss.2024.114229

Available online 26 April 2024

is delivered, thereby affecting learning patterns and outcomes [15]. Model stability is essential when advising students based on learning indicators that have been found to be significant for academic success based on historical data. Traditionally, this has been assessed either by the stability of performance scores and linear model coefficients or by statistical analyses such as correlation studies [13,16].

This study introduces a novel approach of applying XAI to investigate the stability of student success prediction models across years. To this end, we make the following contributions to the field of LA and XAI. First, we map trace data from three academic years onto learning-related features based on best practices in the SRL literature and enrich the feature set with the financial consumption indicators adapted for LA. Next, we use concept drift detection techniques to assess the extent to which the distributions of (1) the study data and (2) model predictions change across academic years. Next, we use the engineered features as inputs to eight ML algorithms widely used in predictive LA [3,6,17] and use statistical tests to evaluate model performance degradation across years. We then use SHapley Additive eXplanations (SHAP) [11] to generate feature importance rankings used in the model stability assessment using the feature agreement, rank agreement and rank correlation metrics widely adopted in the XAI domain [18]. By examining the stability of feature importance across years, we can gain valuable insights into the stability of learning patterns and the applicability of models in changing study contexts. Finally, we show how the generated explanations for the stable models can illustrate desired SRL behavior. To test the generalizability of our findings, we perform an external validation using data from a different course.

In our study, we highlight the importance of ensuring model stability in light of current European legislation that prevents interventions without explicit student consent [19]. Therefore, past models based on anonymized historical data become crucial as they serve as the most reliable means to generate study advice, provided that the models remain stable. Thus, we demonstrate the dual utility of SHAP, which not only enables the study of stability but also serves as a tool for identifying stable success factors that facilitate student guidance.

The paper is organized as follows. Section 2 outlines the related work and identifies the research gap addressed. Section 3 describes the methodology. Section 4 presents the results and discusses them. Section 5 covers theoretical and practical implications, and Section 6 concludes by summarizing main findings, limitations, and future research opportunities.

2. Related work

2.1. Concept drift

In most real-world ML applications, the quality of the deployed predictive models degrades over time, a phenomenon known as model degradation [20]. The first type of model degradation comes from changes in the hidden context not captured by existing data collection methods, hence unknown to the model [20]. Since these changes cannot manifest themselves in the data, ML research focuses primarily on concept drift, formally defined as a change of the distribution underlying the data [21]. The problem of concept drift is exacerbated in ML, where concepts are learned from past data and applied to current data, assuming they still hold. Thus, continuous tracking and evaluation of predictive systems is essential for the usefulness of ML models for decision making [20]. In the LA domain, changes in both the internal and external context of learning over time are reflected in the collected data. It is vital to understand these changes to make well-informed decisions [22].

Concept drift detection methods include statistical tests evaluating significant changes in raw data distributions, model performance analysis, and parameter changes assessment [20,21]. Distribution change is primarily detected using nonparametric statistical tests such as Kolmogorov–Smirnov [23,24] and Wilcoxon [25] tests. Model performance can also be used to detect concept drift, with increasing error rates being a sign of change. Concept drift is also reflected in changes in the optimal parameters of the model under the uncertainty it causes [26]. The issue of concept drift is well recognized in areas where learning from evolving data is essential, e.g., anomaly detection, credit scoring and churn prediction [27,28].

Another method of quantifying distributional shifts is based on credit risk research, where the Population Stability Index (PSI) is an industry benchmark for assessing differences between statistical distributions [29]. The PSI helps to identify variations in model predictions, by comparing their distributions on a current test cohort with those on a train set when the model was developed. Higher PSI values indicate deteriorating model stability [30].

2.2. Predictive LA and model stability

Predictive modeling in LA has shifted from finding the right features and building highly accurate models to stronger theoretically supported approaches [5]. The current study is based on Winne and Hadwin's (1998) constructivist model of SRL [31], which views learners as active agents who process information to create learning artifacts that support their learning goals. Learners engage in SRL by evaluating learning materials, tools, and tactics through metacognitive monitoring, which is influenced by internal conditions such as motivation, prior knowledge, and affective states, and external conditions such as teacher roles and course requirements. The relevance of the SRL model in face-to-face, blended and online learning settings is well established [6,32], and the influence of both internal [33] and external conditions [34,35] has already been investigated [5].

LA is no exception when it comes to the problem of concept drift, as technological developments and changes in the learning contexts lead to shifts in student learning patterns [14]. This has implications for the use of historical educational data as, under concept drift, the models trained on past data are no longer valid for decision making on current student cohorts [14], making concept drift detection essential for producing generalizable models.

Predictive models' stability has been primarily explored by comparing the algorithms' performance and the features' predictive power. Early research on the stability and sensitivity of models across time cohorts suggests the strongest predictive patterns remain apparent in both cohorts [13]. However, instructional design changes affect the patterns connected with them, influencing the predictive power and stability of the features measuring them [13]. A recent study by [16] conducts a meta-analysis to estimate the predictive performance of study indicators across homogeneous courses. The suggested statistical method included the Pearson correlation between grades and study indicators separately, a meta-analysis, and a random effects model [16]. Some indicators were found to be consistent and others to show variability with moderate heterogeneity. When heterogeneity exists, variability occurs, and more homogeneous courses have higher stability in their indicators' predictive power [16].

Predictive LA studies mostly focus on early prediction to orchestrate interventions for at-risk students [36]. However, studies that examine portability/stability of indicators across contexts (e.g., [5,16]) analyze the relationship between SRL indicators and grades using trace data of a complete course. Full course data can offer insights into study profiles, recommended behavior and advise for future students, rather than intervening with individual at-risk students. This allows to solve the ethical conundrum of targeting only at-risk students [37] and not the entire student body. The patterns that lead to better academic success can also benefit students who are not at risk. In addition, because trace data is complex and not always available, educators often have to rely on historical data.

2.3. eXplainable AI for LA

The increasing use of ML algorithms for LA has raised concerns about transparency and fairness. Recent advances in XAI have improved the interpretability and transparency of ML models. Although XAI for education is similar to its broader application, it has unique requirements as summarized in the XAI in EDucation (XAI-ED) framework [38], which emphasizes the benefits of improved student–teacher interactions, enhanced AI literacy, and increased trust. The approaches to achieving explainability in education align with those of general XAI. However, a misinterpretation of an explanation can cause AI systems to malfunction, so it is essential to carefully assess the experience, needs, and capabilities of target users, involving not only ML experts but also key stakeholders of education: instructors and students [38,39]. Pedagogical experts may need more complex explanations, but learners can benefit from even simple explanations to understand AI decisions [39].

The popularity of XAI in education has led to diverse XAI applications for various LA tasks. The study by Mu et al. (2020) [40] uses the TreeSHAP algorithm to provide explanations for models that predict wheel-spinning, in which students attempt an educational task repeatedly without learning a specific skill. By leveraging actionable features and feature contributions returned by TreeSHAP, the most effective interventions can be identified to assist struggling students [40]. In another study by [12], different XAI approaches, including LIME [41], Kernel SHAP [11], PermutationSHAP [11], Contrastive Explanation Method [42], and Diverse Counterfactual Explanations [43], were evaluated for explaining a Bidirectional LSTM model designed to predict student success in Massive Open Online Courses (MOOCs). Although the models align on the top-contributing features, significant variations exist in the importance scores across different explainability methods [12]. Moreover, the choice of a technique impacts the feature importance more than the data or model itself, emphasizing the need to carefully select an appropriate explainability method based on the desired properties [12]. XAI also finds its applications in dropout assessment tasks [44] where the effectiveness of Shapley values, SHAP, and LIME frameworks is evaluated for the task of explaining predictions made by a black-box Multi-Layer Perceptron (MLP) model. Notably, the SHAP framework yields the highest explainability index among the evaluated methods [44]. SHAP has also been adopted for explaining AI-based student success prediction models [45] trained on social media usage and demographics data. The study revealed that certain sensitive demographic factors had a substantial impact on the model's predictions, thereby highlighting concerns about the fairness and trustworthiness of the deployed model [45].

2.4. Research gap and novelty

In this paper, we introduce a novel perspective on stability that differs from previously established definitions. We situate our research in the domain of student success prediction, where the models' stability is evaluated in the context of using the model's insights to provide study advice. For black-box models, these insights are revealed using XAI techniques, allowing educators to understand the model output and use it for advising. Due to ethical and privacy concerns, this advice is primarily based on data from past course runs. In this context, stability can be defined as the consistency of the "model's reasoning", specifically the consistency of the global feature importance revealed by the explainability technique. For student success prediction, these features represent students' self-regulatory capabilities, and the importance of a particular aspect of self-regulation is crucial for advising students. The extent to which a feature's importance remains stable from one course run to another is at the core of the definition of stability in this research.

Existing studies focus on assessing the stability and portability of linear models using various statistical techniques. In contrast, our study investigates the stability of more complex and often black-box ML algorithms for predicting student success. To this end, we draw inspiration from the broad field of XAI, and emphasize the importance of model interpretability, focusing on the analysis of global feature importance. This aspect is critical in our quest to identify consistent student success indicators and to understand their variation across courses and course sequences. We use SHAP, a widely accepted feature attribution technique, to examine the stability of key student success indicators. Furthermore, for those indicators that show stability, we examine their feature importance rankings and feature effects, which are critical for advising students and promoting desirable SRL behaviors. To the best of our knowledge, this study is the first of its kind to investigate the stability of student success prediction models through a comprehensive analysis of global feature importance.

3. Methodology

Following the general ML pipeline, we first prepare the data as described in Section 3.1 and check for concept drift in the data (Section 3.2). Next, we perform the predictive modeling task with the SHAP explanations generated on top of the trained models (Section 3.3). Finally, the model predictions are used to analyze prediction drift (Section 3.2), while the performance metric values and the generated SHAP estimates (Section 3.4) are used to check performance stability and compute agreement metrics as described in Section 3.5.

3.1. Data

We use the data from the mandatory first-year bachelor courses: first, the Accountancy course taught in the fall (first) semester (used for the main analysis) and, second, the Global Economics course taught in the spring (second) semester (used for the external validation) in four programs offered on one campus by the Faculty of Economics and Business in a higher education institution in Belgium. LMS trace data was collected from these courses for three academic years (AY): AY 2018–2019 (AY 18–19 for short), AY 2019–2020 (AY 19–20 for short) and AY 2020–2021 (AY 20–21 for short). The choice of academic years allows us to consider the effect of COVID-19. For the Global Economics course taught in the spring semester, the comparison of AY 18–19 with AY 19–20 shows the transition from non-COVID mode to partial COVID mode (first lockdown), while the comparison of AY 19–20 with AY 20–21 shows the transition from partial COVID mode (first lockdown) to full COVID mode (subsequent lockdowns). For the Accountancy course taught in the fall semester, the comparison of AY 18–19 to AY 19–20 is unaffected by the pandemic while the comparison of AY 19–20 to AY 20–21 shows a transition from non-COVID mode to full COVID mode.

3.1.1. Feature engineering

We base our feature engineering on the SRL theory using the widely adopted study indicators as summarized in [46], which grounds feature engineering in the study of [5] and enriches the feature set with features inspired by financial consumption patterns that measure exploration, exploitation, and plasticity of human consumption [47]. These features have been shown to be informative in previous research [46]. A detailed feature engineering is described below. To measure study success, we collect the summative first exam attempt grades (on a scale of 0–20 with passing grades starting at 10) and transform them into a binary pass/fail feature.

The original events captured by an LMS are used to generate higher-level learning actions as they are proven to be more meaningful [5,48]. We refer to these actions as features or indicators hereafter. The study of [5] proposes to capture study behavior on the levels of activity and study regularity. In addition, for both levels, the overall and learning action specific patterns are captured resulting in four types of features: overall level of activity (OLA), learning action specific level of activity (LALA), overall regularity of study (ORS), and learning action specific regularity of study (LARS) [5]. Regarding the specific level of activity, four different types of learning actions are further delineated: forum

Table 1
Engineered features. Those remaining after preprocessing are displayed in bold (Section 3.1.2).

Type	Feature	Description
OLA	**Total number of sessions**	Total number of sessions of non-zero duration
	Total sessions duration	Sum over all the sessions duration during the course, seconds
	Median session duration	Median calculated over all the sessions' duration during the course, seconds
	Median number of actions per session	Median calculated over all the sessions' total learning actions counts
	Proportion of active days	Total number of active days relative to the duration of the course in days
	Median number of active days per week	Median calculated over all the weeks with active days during the course
	Median difference between active days	Median calculated over the time distances between consecutive active days during the course.
	Proportion of active weeks	Total number of active weeks relative to the duration of the course in weeks
LALA	Proportion of active days: course material/main page	Total number of active days with course materials/main page views relative to the duration of the course in days
	Proportion of active weeks: course material/main page	Total number of active weeks with course materials/main page views relative to the duration of the course in weeks
	Proportion of posts read - forum consumption	Total number of posts read on the forum relative to the total number of posts available on a discussion forum
	Total number of created posts - forum contribution	Total number of posts written on discussion forum during the course's duration
ORS	**Constancy of clicks**	Entropy calculated with the probabilities estimated as the proportion of the number of learning actions per session relative to the total number of learning actions across all sessions
	Constancy of session length	Entropy calculated based on the probabilities estimated as the proportion of a session's length relative to the total sessions length across all sessions
	Proportion of weeks with first-day activity	Total number of weeks with activity on the first day relative to the duration of the course in weeks
	Proportion of first-day-of-week activity	Median calculated over all the weeks for the proportion of the learning actions performed on the first day of the week relative to the total number of actions performed in this week
LARS	Constancy of clicks: course material/main page daily	Entropy calculated based on the probabilities estimated as the proportion of the number of course material/main page views per day relative to the total number of course material/main page views
	Constancy of clicks: course material/main page weekly	Entropy calculated based on the probabilities estimated as the proportion of the number of course material/main page views per week relative to the total number of course material/main page views

contribution, forum consumption, access to learning materials and access to the main course page.

The engineered features are displayed in Table 1. A session is defined as a sequence of consecutive learning actions performed within a time frame of maximum eight hours. An event during the session is captured as the most granular activity per timestamp. Due to the logging idiosyncrasies of the LMS, it is possible to have several events per timestamp as opening one directory also logs opening all its subdirectories. These cases were preprocessed by replacing multiple events with one event as it better represents an action performed by a student. We use the terms view, click and access interchangeably to describe these student actions. We define active days and active weeks [5]: an active day is a day with at least one learning action while an active week is a week with a number of active days equal or above the course's average number of active days per week. The motivation for using the weekly indicators is similar to [5]: the course has a weekly periodicity.

Due to the unavailability of granular event data for forum consumption and contribution, we cannot measure them on a timely basis. Hence, we express forum consumption as a proportion of total posts read relative to the total number of posts available on the forum. Forum contribution is expressed as the total number of posts written on the forum.

To compute the ORS and LARS features, we follow the approach of [5] and use entropy. We give the features more intuitive names by replacing the entropy term with the term 'Constancy'. This is motivated by the fact that a higher entropy corresponds to a more uniform distribution of learning over time, resulting in a higher regularity, i.e. constancy of learning. Constancy is calculated using Shannon's entropy H formula as $H = -\sum_{i=1}^{M} P_i \, log_2 \, P_i$ with the probabilities P_i calculated as explained in Table 1.

The aforementioned indicators do not fully capture the temporal aspects of studying which is crucial as learning unfolds over time. To this end, the features of [5] are complemented with the features of [47] to incorporate more sophisticated ways of capturing self-regulation behavior [46]. Originally, [47] studies human consumption patterns across time and space. In particular, these patterns are captured by exploration, exploitation, and plasticity. These traits can be proxied by diversity, loyalty, and regularity metrics, respectively. These metrics have been shown to play an important role in predicting the financial outcomes of individuals [47,49]. In this study, we adapt these metrics for the SRL context.

Firstly, diversity represents the extent to which users spread their transactions over time (temporal diversity) or space (spatial diversity) with a higher diversity corresponding to spreading the transactions almost equitably over the bins [47], i.e., $D_i = \frac{-\sum_{i=j}^{N} P_{ij} \, log_2 \, P_{ij}}{log_2 \, M}$ where P_{ij} is the fraction of transactions in bin j for user i, N is the total number of bins with M of them being non-empty. In the context of students studying for a course, this translates to temporal diversity with weekly bins. This concept of diversity is closely tied to the concept of entropy but at the scale of sessions. Sessions are chosen as transactions as the constancy of individual learning actions is already measured by the features displayed in Table 1. The diversity metric is in the range $[0, 1]$, with larger values corresponding to spreading studying more evenly over the semester. Hence, we change the name from 'Diversity' to 'Uniformity' as it better represents uniformly spreading study effort throughout the semester.

Second, the loyalty metric from [47] represents how transactions are distributed across different bins (spatial or temporal), with higher values corresponding to having most transactions within the top N of the bins, i.e., $L_i = \frac{f_i}{\sum_{j=1}^{N} P_{ij}}$ where f_i is the combined fraction of all transactions that occur in the top three most-frequented bins, P_{ij} is the fraction of transactions in bin j for user i. We rename 'Loyalty' into 'Bingeing' as it better expresses the concentration of study effort

Table 2
Outcome feature distribution.

	Accountancy			Global Economics		
	AY 2018–2019	AY 2019–2020	AY 2020–2021	AY 2018–2019	AY 2019–2020	AY 2020–2021
Pass	383 (53%)	391 (56%)	457 (64%)	333 (46%)	420 (62%)	305 (46%)
Fail	335 (47%)	309 (44%)	261 (36%)	389 (54%)	256 (38%)	357 (54%)

in a small number of time slots. We calculate temporal bingeing with weekly bins, with sessions as transactions. We purport that the action of bingeing on learning materials to catch up in a course or to study for an exam is a self-regulation action.

Lastly, regularity measures the differences between behavioral patterns over shorter and longer study periods, i.e., $R_i = 1 - \frac{\sqrt{(D_i^1 - D_i^T)^2 + (L_i^1 - L_i^T)^2}}{\sqrt{2}}$ where D_i^1 and L_i^1 are the diversity and the loyalty in a shorter period (e.g., one month), D_i^T and L_i^T is the diversity and the loyalty in the entire period. Since the label ‘Regularity’ is quite intuitive even in study contexts, we keep it the same. We compute temporal Regularity with D_i^1 and L_i^1 set to the Uniformity and the Bingeing in the semester weeks, respectively, and D_i^T and L_i^T set to the Uniformity and the Bingeing in the entire course, respectively. The purpose of Regularity is to compare how students disperse and self-regulate their periods of Uniformity and Bingeing.

3.1.2. Data preprocessing

We deal with missing values by dropping the students who did not take the exam or have no corresponding activity on the LMS platform. Table 2 shows the final distribution of the outcome variable after data preprocessing. Missing values for the entropy features are imputed as $V = log_2 B$, where the value of B varies according to the granularity of the entropy. In particular, for the daily and weekly features, B represents the course duration in days and weeks, respectively. For the session features, B is set to the average number of sessions calculated over all the students. This design choice is based on the fact that the absence of study represents a “missing not at random” situation, where missing data is systematically related to unobserved data. We assume that when there is no studying at all, the regularity of studying is uniform, which can be represented by the maximal possible entropy value of V. Missing values for the uniformity, bingeing and regularity metrics are set to the highest possible value of 1.

Correlation analysis revealed that some of the features are highly correlated and represent essentially the same information (see online Appendices A and B[1]). Including such features may obscure important interactions. In addition, we follow Occam’s razor and prefer a less complex model. Therefore, we drop some of the LALA and LARS features and end up with 13 features marked in bold in Table 1 and the Bingeing, Uniformity and Regularity features.

3.2. Concept drift detection

As argued in Section 2.1, the Kolmogorov–Smirnov (KS) nonparametric test is widely used in concept drift detection [21,22] especially for smaller sample sizes. The null hypothesis of the two-sample KS test is that two samples come from the same distribution, with the test statistic calculated as $D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}|$ where n, m represent the number of observations in the reference and current samples, respectively, $F_{1,n}$ and $F_{2,m}$ are the cumulative distribution functions of these samples, and $\sup$ is the supremum function. The null hypothesis is rejected at α if $D_{n,m} > \sqrt{-\ln(\frac{\alpha}{2}) \cdot \frac{1+\frac{m}{n}}{2m}}$. To control alpha inflation, we apply Bonferroni correction, adjusting α to $\frac{\alpha}{m}$, where m is the number of tests.

The stability of the distribution of the model predictions can be assessed with PSI scores (as outlined in Section 2.1). The PSI calculation includes binning the model predictions into a number of equal-sized bins, and then calculating the difference between the proportion of predictions in each bin on the current dataset and the proportion of predictions in the same bin on the reference dataset with the final PSI score calculated as the sum of these differences, i.e., $PSI = \sum_{k=1}^{K} \left((P_{t2}^k - P_{t1}^k) \cdot \ln \left(\frac{P_{t2}^k}{P_{t1}^k} \right) \right)$ where P_{t1}^k represents the number of observations that fall within bin k in the reference dataset while P_{t2}^k represents the number of observations that fall within bin k in the current dataset. According to [29], a PSI value above 0.25 indicates a significant shift in the score distribution.

The detected distribution and prediction drifts can be seen as a first signal of model instability, since the model is no longer used in the same context in which it was trained. We can also use the results of this step to benchmark our stability assessment approach to see if it confirms the findings seen from the concept drift alone.

3.3. Predictive modeling

A blueprint ML pipeline starts with splitting the data. As we want both to optimize the model hyperparameters and evaluate the model’s generalization capabilities, the nested cross-validation strategy is used (Figs. 1 and 2). It consists of outer and inner loops of splitting the data and running the models that are used for both the data of the past academic year (hereafter, *AY T-1 data*) and the data of the current academic year (hereafter, *AY T data*) (Fig. 1). The outer loop is represented by 10 stratified folds each being split into 10% test data and 90% train data. The *AY T-1 train data* is further used in the inner loop where it is split into train and validation data following a stratified 5 fold cross-validation approach used to search for the best model configuration. Within each outer fold, we use min–max normalization to transform the data for the models that require it (Table 3).

The optimal model is retrained using all the *AY T-1 train data* defined in the outer loop (hereafter, *Model AY T-1*) and is applied on the *AY T-1 test data* to obtain the performance metrics. Also, the global feature importance scores are calculated using the *Model AY T-1* and the *AY T-1 test data* (step 1 in Fig. 2). As using a sample of data as a background dataset is advised in the official SHAP implementation [11], we follow their guidance and use a random sample of 100 observations as a background dataset to speed up the calculations.

The next step is to apply the *Model AY T-1* on the *AY T test data* and obtain the performance metrics and the global feature importance scores (step 2 in Fig. 2). This step comes in line with the performance change evaluation, one of the concept drift detection methods [21]. As retraining the model is the simplest way to respond to concept drift [21], the *Model AY T-1* is retrained on the *AY T train data* (hereafter *Model AY T-1 retrained*) and is used to make predictions on the *AY T test data* to obtain the performance metrics and global feature importance scores (step 3 in Fig. 2). However, retraining the model alone may not be sufficient to fully account for concept drift. Therefore, the optimization procedure is performed using the inner loop and the *AY T* data. The resulting model (hereafter *Model AY T*) is applied to the *AY T test data* to obtain the performance metrics and the global feature importance scores (step 4 in Fig. 2).

The above four steps are repeated 10 times based on the data split in the outer loop in order to account for the different ways of splitting

[1] Online appendices and source code can be found at https://github.com/tiu-elena/LA-stability

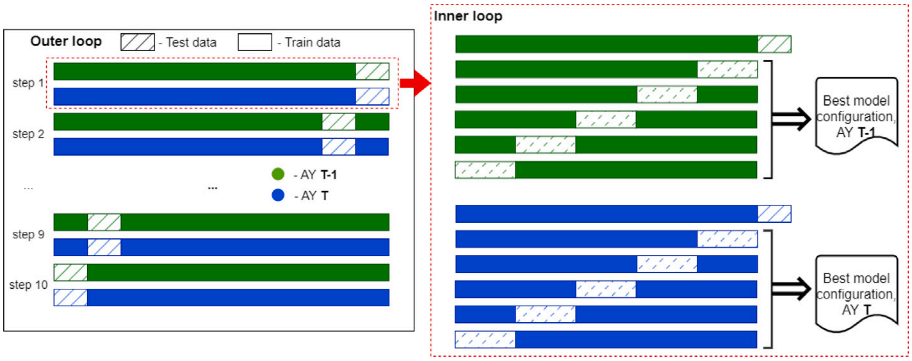


Fig. 1. Methodology: outer loop.

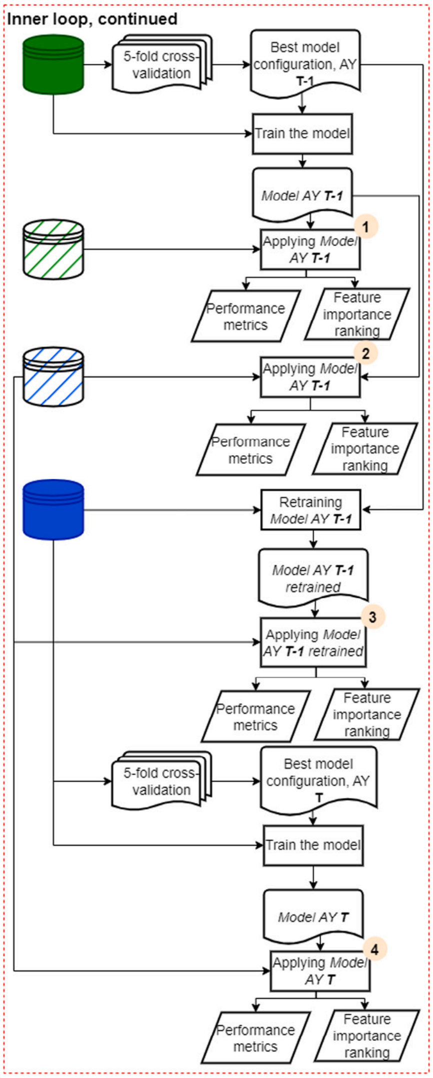


Fig. 2. Methodology: inner loop.

the data and reducing the impact of splitting the data only once. The pipeline is run on eight ML algorithms tailored to solve the binary classification task that gained immense popularity in the field of the LA literature [3,6,17]. These algorithms and their hyperparameter search spaces are shown in Table 3.

3.4. eXplainable AI

In this study, we chose to use the Kernel SHAP algorithm, motivated by its wide adoption in the field and its ability to provide model-agnostic explanations through feature contributions [11]. Inspired by the original Shapley values [61] and LIME [41], Kernel SHAP is an additive feature attribution method that approximates the original model $f(x)$ as a sum of the feature effects ϕ_i as $g(z') = \phi_0 + \Sigma_{i=1}^{M} \phi_i z'_i$ where $z' \in \{0,1\}^M$, M is the number of simplified input features, and $\phi_i \in \mathbb{R}$. The feature effects are calculated as the marginal contribution of a feature value across all possible coalitions as $\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!}[f_x(z') - f_x(x' \backslash i)]$ where $|z'|$ is the number of non-zero entries in z', $z' \subseteq x'$, $f_x(z') = f(h_x(z')) = E[f(z)|z_S]$ and S is the set of non-zero indexes in z' [11].

Kernel SHAP simplifies the input mapping $h_x(z') = z_S$ and approximates $f(z_S)$ with $E[f(z)|z_S]$, where z_S has missing values for features not in the set S. The latter conditional expectation is in turn approximated with $E_{z_{\bar{S}}}[f(z)]$, assuming feature independence. Initially seen as a limitation of Kernel SHAP [11], the feature independence has been shown to be a correct way of looking at the feature contributions from a causal perspective, as the observational conditions are flawed and lead to the failure of sensitivity [62]. This problem does not arise if we consider the feature attribution problem from a causal perspective treating inputs as causes of the output. In this case, the marginal expectation $E_{z_{\bar{S}}}[f(z)]$ corresponds to a correct view of the feature contribution effect. Adopting a causal perspective involves considering a hypothetical intervention, leaving the actual causal relation in the real world aside [62]. In other words, the interventional distribution might create observations that are far from the data distribution. This raises a problem of a trade-off between being “true to the model” and being “true to the data” [63]. As this research investigates model stability, being “true to the model” is preferable, justifying using the Kernel SHAP algorithm.

3.5. Model agreement evaluation

To measure the agreement between models, we (1) compare the values of the predictive performance metrics and their statistical significance and (2) adapt the metrics from the research on the general disagreement problem between explainability techniques [18,64] to be applied to the rankings generated from the SHAP estimates. Below, we describe the metrics and the aspects of disagreement they measure. Note that these metrics are computed for each fold and subsequently averaged to obtain a final estimate of the metric value, as SHAP estimates are calculated per each outer fold, resulting in ten sets (Fig. 2). Note also that for the metrics that require the specification of the top-k features, we set $k = 8$ (half of the total number of features), as this provides a fair trade-off between variability in feature importances for calculating agreement metrics and visualization capabilities for using the explanations for decision making.

Table 3
Students success prediction models. The models displayed in bold require normalized data.

Model	Grid search space	Used in
Naïve Bayes (NB) [50]	Smoothing parameter: 100 values spaced evenly on a log scale from $[10^{-9}, 1]$	[3,6,17]

Support vector machine (SVM) [51]	Regularization parameter C: {0.1, 1, 10, 100}
	Kernel coefficient: {1, 0.1, 0.01, 0.001}
	Kernel function: {RBF, Polynomial, Sigmoid}
		[3,6,17]

Artificial Neural Network (Multi-Layer Perceptron - MLP) [50]	Hidden layer sizes: {(#features/2,), (#features/2, #features/4,)}
	Activation function: {Identity, Logistic, Hyperbolic tangent, ReLu}
	L2 regularization term α: {10^{-6}, 10^{-5}, 10^{-4}}
	Solver: {L-BFGS, SGD, ADAM}
	Learning rate: {10^{-4}, 10^{-3}}
		[3,6,17]

Logistic regression (LR) [52]	Penalty: L1, L2 Inverse of regularization strength C: {10^{-3}, 10^{-2}, 10^{-1}, 10^{1}, 10^{2}, 10^{3}} Solver: {Liblinear, Newton-CG, L-BFGS}	[3,17]
K-Nearest Neighbors (KNN) [53]	#neighbors: [1, 30] with a step of 5 Weight function: {Uniform, Distance}	[3]

Random Forest (RF) [54]	Maximal depth: {–, 3, 5, 10}
	Min. #samples to split: {2, 5, 10}
	Criterion: {Gini, Entropy, Log. Loss}
	#estimators: {50, 100, 200}
	#samples to train: {#features, #features/2}
	#features for best split: {–, $\sqrt{\#features}$, $log_2\#features$}
		[6,55]

XGBoost (XGB) [56]	Maximal depth: {–, 3, 5, 10}
	Min. #samples to split: {2, 5, 10}
	Learning rate: {0.1, 0.01}
	#estimators: {50, 100, 200}
	#features for best split: {–, $\sqrt{\#features}$, $log_2\#features$}
		[55,57]

TabNet (TAB) [58]	Decision prediction layer width: {8, 16, 24, 32} Number of steps in the architecture: {1, 2, 3} Feature reusage in the masks coef.: {1, 1.2, 1.4} #shared Gated Linear Units: {1, 3} Sparsity loss coefficient: {1e−6, 1e−3}	[59,60]

Performance metrics. We use balanced accuracy as the metric for evaluating predictive performance because accuracy is one of the most widely used metrics for predicting student success [6] as well as in concept drift detection analysis [20], and we use its balanced version to account for a slight imbalance in the data. The balanced accuracy is calculated as $\frac{sensitivity+specificity}{2} = \frac{\frac{TP}{TP+FN}+\frac{TN}{TN+FP}}{2}$ where TP (FP) is the number of true (false) positives and TN (FN) is the number of true (false) negatives. To test whether the differences between the four different sets of performance measures are statistically significant, we perform a nonparametric Friedman test with the null hypothesis that the means of three or more groups are equal. The test statistic is calculated as $T_1 = \frac{12}{bk(k+1)} \sum_{i=1}^{k} (R_i - b(k+1)/2)^2$ where $R_j = \sum_{i=1}^{b} R(X_{ij})$ is the sum of the ranks $R(X_{ij})$; if there are ties, then $T_1 = \frac{(k-1)\sum_{i=1}^{k} (R_i - \frac{b(k+1)}{2})^2}{A_1 - C_1}$ where $A_1 = \sum_{i=1}^{b} \sum_{j=1}^{k} (R(X_{ij}))^2$ and $C_1 = \frac{bk(k+1)^2}{4}$.

To test the statistical significance of the individual differences, we use a nonparametric Wilcoxon signed-rank test [25], which tests the null hypothesis that two related paired samples come from the same distribution. The test statistic is calculated as $W = min(W-, W+)$ where $W+$ and $W-$ are the sums of the positive and negative ranks, respectively. To avoid alpha inflation due to multiple testing, we use Bonferroni correction of p-value as $\frac{\alpha}{m}$ where α is the original significance level and m is the number of tests.

Feature Agreement. In the general disagreement problem in the XAI domain, the main notion of disagreement between models lies in the differences in the top-k important features returned by the explainability technique [18]. Consequently, a feature agreement metric has been introduced [18] as $FeatureAgreement(E_a, E_b, k) = \frac{|top\_feat(E_a,k) \cap top\_feat(E_b,k)|}{k}$ where $top\_feat(E, k)$ is the set of top-k features returned by the XAI technique E.

Rank Agreement. Not only the agreement of the top features is vital but also their order. Accordingly, a stricter metric of rank agreement is introduced by considering both the commonality of top-k features and their position in the respective rankings [18] as $RankAgreement = \frac{|\cup_{s\in S}\{s|s\in top\_feat(E_a,k) \wedge s\in top\_feat(E_b,k) \wedge rank(E_a,s)=rank(E_b,s)\}|}{k}$ where S is a feature set, $rank(E, s)$ is the rank of feature s returned by the technique E.

Rank correlation. The differences between explanations also manifest themselves in the changes in the relative feature importance rankings returned by the explainability technique [18]. In this respect, the agreement between them can be measured by the rank correlation (RC) coefficients [65], i.e, $RC(E_a, E_b, F) = r_s(Ranking(E_a, F), Ranking(E_b, F))$ where r_s is the Kendall rank correlation, F is the top eight features returned by the Kernel SHAP algorithm of the *AY T-1 model* applied on the AY T-1 test data, $Ranking(E, F)$ is a rank assignment to the features in F returned by E. We use the Kendall rank correlation because of its suitability for small sample sizes with ties present in data, i.e., $r_s = \frac{P-Q}{\sqrt{(P+Q+T)\cdot(P+Q+U)}}$ where P is the number of concordant pairs in $Ranking(E_a, F)$ and $Ranking(E_b, F)$, Q is the number of discordant pairs in $Ranking(E_a, F)$ and $Ranking(E_b, F)$, T is the number of ties only in $Ranking(E_a, F)$, and U is the number of ties only in $Ranking(E_b, F)$.

4. Results and discussion

4.1. Concept drift detection

Table 4 shows the results of our concept drift analysis on the independent features (visualizations of distributions can be found in online appendix D, Fig. 13–16). Since we are using the same data

Table 4
KS test *p*-values. Significant values ($\alpha = 0.025$) are displayed in bold = drifted features.

	Accountancy		Global Economics	
	AY 18-19 vs. AY 19–20	AY 19-20 vs. AY 20–21	AY 18-19 vs. AY 19–20	AY 19-20 vs. AY 20-21
Bingeing of sessions	**0.01**	**0**	**0**	**0**
Constancy of clicks	0.79	**0**	**0**	0.10
Constancy of session length	**0**	**0**	**0**	0.03
Median diff. between active days	**0**	**0**	**0**	1
Median number of actions per sessions	**0**	**0**	**0**	**0**
Median number of active days per week	**0**	**0**	**0**	0.84
Median session duration	**0**	**0**	**0**	**0**
Proportion of active days	0.11	**0**	**0**	0.29
Proportion of first-day-of-week activity	0.13	**0**	**0.01**	**0**
Proportion of posts read	**0**	0.10	**0.02**	0.14
Proportion of weeks with first-day activity	0.52	**0**	**0**	**0.01**
Regularity of sessions	0.07	**0**	0.09	**0**
Total number of created posts	**0.002**	0.75	0.99	0.95
Total number of sessions	**0.02**	**0**	**0**	**0.01**
Total sessions duration	0.10	**0**	**0**	**0**
Uniformity of sessions	**0**	**0**	0.05	**0**

of AY 19–20 for two comparisons for each course, we correct the significance level α as $\frac{0.05}{2} = 0.025$. For the Accountancy course, comparing the distributions of AY 18–19 and AY 19–20 reveals that 10 out of 16 (62.5%) features exhibit concept drift, as we can reject the null hypothesis at a significance level of $\alpha = 0.025$. Comparing AY 19–20 to AY 20–21, we observe an even more pronounced shift, with 14 out of 16 (87.5%) features showing evidence of concept drift. This finding underscores the profound impact of COVID-19 on study behavior and is consistent with the research findings of [15], which found significant differences in study patterns depending on the year of study. Note that since this is a first-semester course, and COVID-19 affected only the second semester, the drift from AY 18–19 to AY 19–20 is much smaller than the drift from AY 19–20 to AY 20–21.

For the Global Economics course, our analysis confirms the above reasoning. Comparing the distributions from AY 18–19 and AY 19–20, we see a significant shift: 13 out of 16 (81.3%) features experience concept drift, marking a transition from the pre-COVID mode to a partial COVID setting, specifically during the first lockdown. This shift was a direct response to the unanticipated impact of COVID, resulting in a profoundly disruptive influence on the educational experience. The AY 19–20 vs. AY 20–21 comparison shows 9/16 (56.3%) drifting features, which is a smaller change than observed in the previous comparison. This contrast highlights the diminishing effects of COVID as the educational system gradually adapted to the new circumstances, moving from a partial COVID mode to a full COVID mode.

The identified concept drift in the distribution of independent features serves as an early indicator of potential model instability as the historic data used to train the model does not come from the same distribution that the model is being applied to in the future.

4.2. Prediction drift

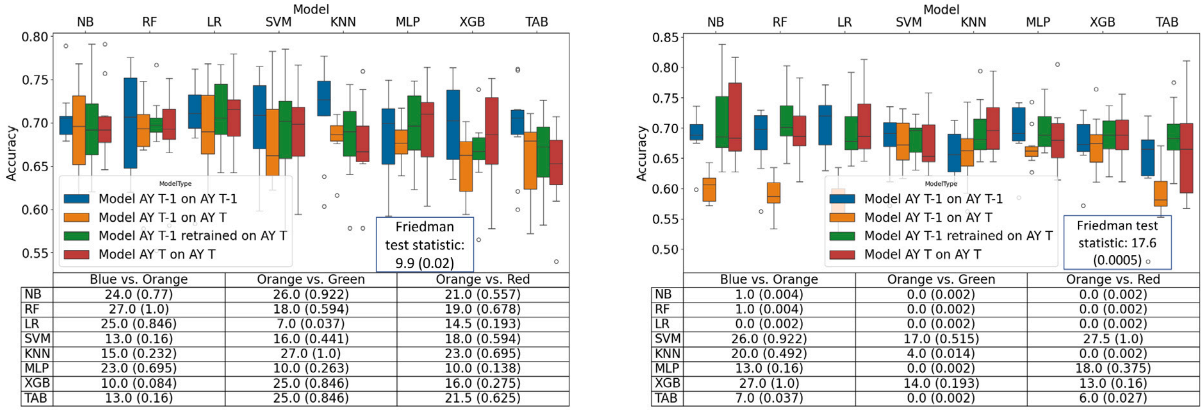
In addition to the model performance stability analysis, we perform a prediction drift analysis, i.e., we use the PSI index to measure the extent to which the distributions of model predictions are stable across years. In particular, we calculated PSI scores for the comparison between the predictions of the *Model AT T-1* applied to AY T-1 test data and the predictions of the *Model AT T-1* applied to AY *T* data (online appendix C, Table 1). With the PSI we examine if the prediction distribution changes when used on data from different academic years while keeping the model constant. We see that the threshold of 0.25 is exceeded in almost all comparisons, indicating a prediction drift of the models. For the Accountancy course, this drift is higher for the comparison happening within the COVID context, i.e., AY 19–20 vs. AY 20–21, while for the Global Economics course the values are higher for the shift from the non-COVID mode to the partial-COVID mode, i.e., the AY 18–19 vs. AY 19–20 comparison.

4.3. Model performance stability

Figs. 3 and 4 show the Friedman test results and the box plots of the balanced accuracy values obtained from the outer folds of the nested cross-validation. The tables below the box plots show the Wilcoxon test statistics and their *p*-values. For the individual comparisons of the model applications, we are interested in three performance comparisons: (1) *Model T-1* applied to test data from AY T-1 vs. *Model T-1* applied to test data of AY *T* to understand how well the model can perform as-is without modification; (2) *Model T-1* applied to test data from AY *T* vs. *Model T-1 retrained* applied to test data from AY *T* to understand the performance gains (if any) from retraining the model; (3) *Model T-1* applied to test data from AY *T* vs. *Model T* applied to test data from AY *T* to further evaluate the added value of hyperparameter optimization. In line with these comparisons, the significance level α for the Wilcoxon test is corrected to $\frac{0.05}{3} = 0.016$. Note that an analysis of the comparative performance between the algorithms is beyond the scope of this research.

For the Accountancy course, the *p*-values of the Friedman test statistic show that the null hypothesis can be rejected for both the comparisons, i.e., the differences in model performance are statistically significant. However, looking at the pairwise differences for the comparison of AY 18–19 and AY 19–20 (Fig. 3(a)), both the box plots and the Wilcoxon test statistics show that the differences in the performance of most of the models in the above comparisons are not statistically significant. From the box plots, it can be seen that for all the algorithms, using the model trained on the data of past AY results in a lower mean performance than using it on the data of the same AY on which it was trained, but the margin is rather small and not statistically significant. The same is true for retraining the model: for 6/8 of the models (i.e., except for NB and TAB), the average balanced accuracy is higher than when using the original model, but the difference is not statistically significant.

In contrast, when looking at the results for the Accountancy course and comparing AY 19–20 with AY 20–21 (Fig. 3(b)), retraining the model has an added value for the performance of most of the models, as we can see that the differences between just applying the model trained on the past data and the model retrained on the new data are statistically significant for 5/8 of the models. The visual observation of the box plots also confirms these findings, as we can clearly see that the mean values of the updated models are higher than the original ones, although there is no clear added value of optimizing the hyper-



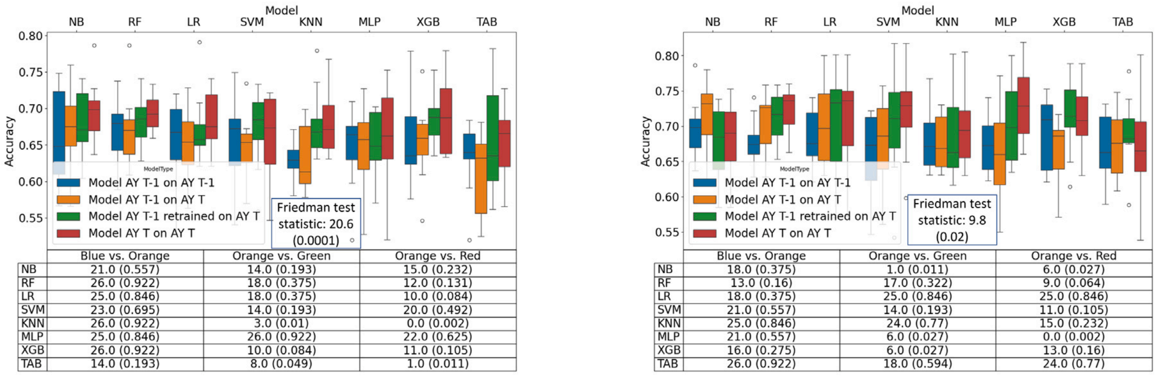
	Blue vs. Orange	Orange vs. Green	Orange vs. Red
NB	24.0 (0.77)	26.0 (0.922)	21.0 (0.557)
RF	27.0 (1.0)	18.0 (0.594)	19.0 (0.678)
LR	25.0 (0.846)	7.0 (0.037)	14.5 (0.193)
SVM	13.0 (0.16)	16.0 (0.441)	18.0 (0.594)
KNN	15.0 (0.232)	27.0 (1.0)	23.0 (0.695)
MLP	23.0 (0.695)	10.0 (0.263)	10.0 (0.138)
XGB	10.0 (0.084)	25.0 (0.846)	16.0 (0.275)
TAB	13.0 (0.16)	25.0 (0.846)	21.5 (0.625)

(a) AY 18-19 vs. AY 19-20

	Blue vs. Orange	Orange vs. Green	Orange vs. Red
NB	1.0 (0.004)	0.0 (0.002)	0.0 (0.002)
RF	1.0 (0.004)	0.0 (0.002)	0.0 (0.002)
LR	0.0 (0.002)	0.0 (0.002)	0.0 (0.002)
SVM	26.0 (0.922)	17.0 (0.515)	27.5 (1.0)
KNN	20.0 (0.492)	4.0 (0.014)	0.0 (0.002)
MLP	13.0 (0.16)	0.0 (0.002)	18.0 (0.375)
XGB	27.0 (1.0)	14.0 (0.193)	13.0 (0.16)
TAB	7.0 (0.037)	0.0 (0.002)	6.0 (0.027)

(b) AY 19-20 vs. AY 20-21

Fig. 3. Model performance: Accountancy - accuracy and statistical tests results (*p*-values are in brackets).



	Blue vs. Orange	Orange vs. Green	Orange vs. Red
NB	21.0 (0.557)	14.0 (0.193)	15.0 (0.232)
RF	26.0 (0.922)	18.0 (0.375)	12.0 (0.131)
LR	25.0 (0.846)	18.0 (0.375)	10.0 (0.084)
SVM	23.0 (0.695)	14.0 (0.193)	20.0 (0.492)
KNN	26.0 (0.922)	3.0 (0.01)	0.0 (0.002)
MLP	25.0 (0.846)	26.0 (0.922)	22.0 (0.625)
XGB	26.0 (0.922)	10.0 (0.084)	11.0 (0.105)
TAB	14.0 (0.193)	8.0 (0.049)	1.0 (0.011)

(a) AY 18-19 vs. AY 19-20

	Blue vs. Orange	Orange vs. Green	Orange vs. Red
NB	18.0 (0.375)	1.0 (0.011)	6.0 (0.027)
RF	13.0 (0.16)	17.0 (0.322)	9.0 (0.064)
LR	18.0 (0.375)	25.0 (0.846)	25.0 (0.846)
SVM	21.0 (0.557)	14.0 (0.193)	11.0 (0.105)
KNN	25.0 (0.846)	24.0 (0.77)	15.0 (0.232)
MLP	21.0 (0.557)	6.0 (0.027)	0.0 (0.002)
XGB	16.0 (0.275)	6.0 (0.027)	13.0 (0.16)
TAB	26.0 (0.922)	18.0 (0.594)	24.0 (0.77)

(b) AY 19-20 vs. AY 20-21

Fig. 4. Model performance: Global Economics - accuracy and t-test results (*p*-values are in brackets).

parameters as retraining alone improves the performance. These results are consistent with the data drift discussed in Section 4.1, where for the Accountancy course we observed a higher concept drift in the data distribution of AY 20–21 compared to AY 19–20, reinforcing the need to update the model with the latest data. These results also illustrate the impact of COVID-19, which had a tremendous impact on education worldwide and changed the nature of studying [66].

To validate the results, we present the same analysis in Fig. 4 for the Global Economics course. For both the comparisons, we can see that the results of the Friedman tests show the statistical significance of the differences between the models. However, for most of the algorithms, the pairwise tests show no significant differences in performance for the pairs of the models we are interested in. Comparing AY 18–19 and AY 19–20, visual observation of the box plots shows that for most of the algorithms (6/8), the average performance decreases when the historical model is applied as-is to the current data, while optimizing and retraining the model improves the performance for all the algorithms. However, these differences are not statistically significant. On the other hand, the comparison between AY 19–20 and AY 20–21 (Fig. 4(b)) does not show the decrease in performance when the historical model is applied to the current data, and the differences are not statistically significant. The effect of (optimizing and) retraining the model is less obvious than in the previous comparison with 5/8 algorithms benefiting from the retraining and 6/8 algorithms benefiting from the optimization and retraining. However, these differences are not statistically significant.

In the context of the Global Economics course, both comparisons are affected by COVID-19 effects. The first comparison concerns the transition from a non-COVID to a partial-COVID mode, while the second the transition from a partial-COVID to a full-COVID mode. For both the comparisons, we observed no statistically significant changes in the historical model performance when applied to new data and no significant effects of optimizing and retraining the model. This comes in line with the fact that, for both comparisons, the contexts of the academic years share similarities, i.e., either a (partial) non-COVID context for the first comparison and a (partial) COVID context for the second comparison. This common study context, when manifested in the data, allows the model to learn the patterns that remain stable across academic years, leading to performance stability. Nevertheless, the insights gained from a visual observation of the box plots are consistent with the detected concept drift (Section 4.1): the performance gains from optimizing and retraining the model are more apparent for the case of a larger concept drift (AY 18–19 vs. AY 19–20).

4.4. Model agreement evaluation

Table 5 shows the agreement metrics for three comparisons of the rankings: (1) *Model T-1* applied to the AY T-1 test data vs. *Model T-1*

Table 5
Agreement metrics. Darker colors represent higher agreement.

Model	AY 18-19 vs. AY 19-20									AY 19-20 vs. AY 20-21								
	Feature agreement			*Rank correlation*			*Rank agreement*			*Feature agreement*			*Rank correlation*			*Rank agreement*		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
	Accountancy																	
NB	1,00	0,91	0,91	0,84	0,69	0,66	0,58	0,40	0,38	0,93	0,89	0,85	0,74	0,69	0,69	0,28	0,25	0,28
RF	0,93	0,86	0,88	0,89	0,63	0,65	0,56	0,20	0,23	0,93	0,65	0,69	0,81	0,11	0,14	0,60	0,06	0,09
LR	0,95	0,78	0,68	0,85	0,65	0,22	0,59	0,24	0,11	0,94	0,70	0,68	0,76	0,39	0,31	0,40	0,18	0,13
SVM	0,95	0,93	0,78	0,73	0,35	0,28	0,31	0,18	0,15	0,93	0,59	0,63	0,79	0,31	0,25	0,45	0,15	0,10
KNN	0,94	0,86	0,84	0,74	0,49	0,49	0,46	0,19	0,20	0,90	0,81	0,81	0,79	0,29	0,29	0,41	0,13	0,15
MLP	0,86	0,80	0,69	0,73	0,66	0,36	0,40	0,29	0,19	0,93	0,81	0,60	0,82	0,68	0,54	0,46	0,35	0,20
XGB	0,91	0,79	0,80	0,67	0,32	0,36	0,33	0,15	0,16	0,90	0,54	0,56	0,81	0,18	0,21	0,44	0,10	0,11
TAB	0,90	0,61	0,61	0,80	0,36	0,34	0,46	0,05	0,10	0,85	0,59	0,58	0,61	0,11	0,22	0,24	0,16	0,10
	Global Economics																	
NB	0,95	0,84	0,70	0,66	0,50	0,33	0,51	0,36	0,24	0,88	0,85	0,73	0,82	0,77	0,66	0,73	0,70	0,26
RF	0,93	0,81	0,78	0,89	0,47	0,35	0,64	0,26	0,23	0,91	0,81	0,76	0,83	0,66	0,66	0,60	0,38	0,38
LR	0,89	0,75	0,68	0,65	0,57	0,29	0,39	0,33	0,23	0,96	0,65	0,58	0,89	0,47	0,31	0,66	0,26	0,18
SVM	0,93	0,68	0,69	0,77	0,36	0,41	0,50	0,23	0,13	0,88	0,74	0,83	0,88	0,40	0,35	0,56	0,21	0,14
KNN	0,89	0,81	0,79	0,86	0,66	0,61	0,59	0,29	0,33	0,88	0,78	0,76	0,71	0,46	0,49	0,34	0,28	0,24
MLP	0,89	0,81	0,68	0,76	0,67	0,49	0,44	0,38	0,26	0,94	0,89	0,61	0,84	0,78	0,55	0,48	0,46	0,30
XGB	0,88	0,76	0,74	0,66	0,39	0,35	0,36	0,19	0,21	0,95	0,76	0,76	0,85	0,58	0,58	0,59	0,36	0,36
TAB	0,84	0,59	0,69	0,65	0,24	0,32	0,35	0,16	0,16	0,86	0,66	0,65	0,62	0,35	0,25	0,29	0,11	0,18

applied to the AY T test data; (2) *Model T-1* applied to the AY $T-1$ test data vs. *Model T-1 retrained* applied to the AY T test data; (3) *Model T-1* applied to the AY T-1 test data vs. *Model T* applied to the AY T test data. The choice of the comparisons is based on the fact that we are interested in the changes in the rankings of the historical model application with its current applications (including applying the model as-is, retraining it, and optimizing along with retraining).

In the context of both Accountancy and Global Economics (Table 5), and for both AY comparisons, we can see that the highest agreement values are obtained for the feature agreement metric while the rank agreement metric shows the lowest agreement. This is consistent with the logic of the metrics calculation: rank agreement is the strictest agreement measure among the metrics used in this paper, as it requires the feature not only to appear in the top important features but also to hold the same ranking. Another clear pattern is that the agreement decreases as the models are updated: the agreement is the highest when the model is applied unchanged to new data and is the lowest when the model is optimized and retrained on new data. Also, for both courses and comparisons, the Naive Bayes algorithm shows the highest agreement. This stability can be explained by the way the algorithm works: it is the only one of the algorithms used in this study that belongs to the family of generative learning algorithms, while the others are discriminative. It uses the probabilistic structure of the data to learn the patterns behind the different classes. Therefore, if this structure remains stable, the model itself remains stable despite the distributional changes.

In the context of the Accountancy course (Table 5), we can observe that higher agreement is obtained for all model comparisons when comparing AY 18–19 and AY 19–20. This is also the case for the smaller distributional shift as described in Section 4.1 due to the same instructional context in both academic years (no COVID effects). In contrast, when comparing AY 19–20 and AY 20–21, we can see a lower agreement, especially for the rank agreement metric. Together with the observed concept drift and the significant performance gains of the (optimized) retrained model, this finding illustrates the need to keep the model updated as soon as the learning context changes. The analysis of the data from the Global Economics course confirms these findings: we observe a lower agreement when comparing AY 18–19 and AY 19–20, which corresponds to a switch to the (partial) COVID mode and is also consistent with a higher concept drift (Table 5).

In Table 6, we display the indicators that consistently appear in the top-eight important features across all four different model applications (Fig. 2). In particular, we first ensure that the feature appears in the top-eight in at least 8 out of 10 folds. Then, for the features that meet this requirement, we calculate the number of models for which it appears in the top-eight important features for all model applications. We can see that the two most stable features represent the overall level of activity (Total number of sessions and Total session duration). The features representing the regularity of study (i.e., Constancy of clicks and Constancy of session length) also appear to be portable across years; however, they appear in less of the algorithms. The findings of overall activity and regularity indicators being stable across contexts are in line with the findings of Saqr et al. (2022) [16].

Fig. 5 shows SHAP summary plots summarizing both the feature importance ranking and the feature effects for the top eight features of the Naive Bayes model that was found to be the most stable according to the agreement metrics. Positive SHAP values contribute positively to the prediction, and vice versa. We can see that both the feature effects and the top important features remain stable with slight changes in ranking. The plots reveal how study indicators relate to academic success, emphasizing that higher overall activity (total number of sessions and their total duration) positively influences academic success. SHAP plots can be used as a visual aid by teachers to improve their decision-making when providing study guidance and advice to students, similar to their adoption in studies on early success prediction [9] and formative feedback generation [10].

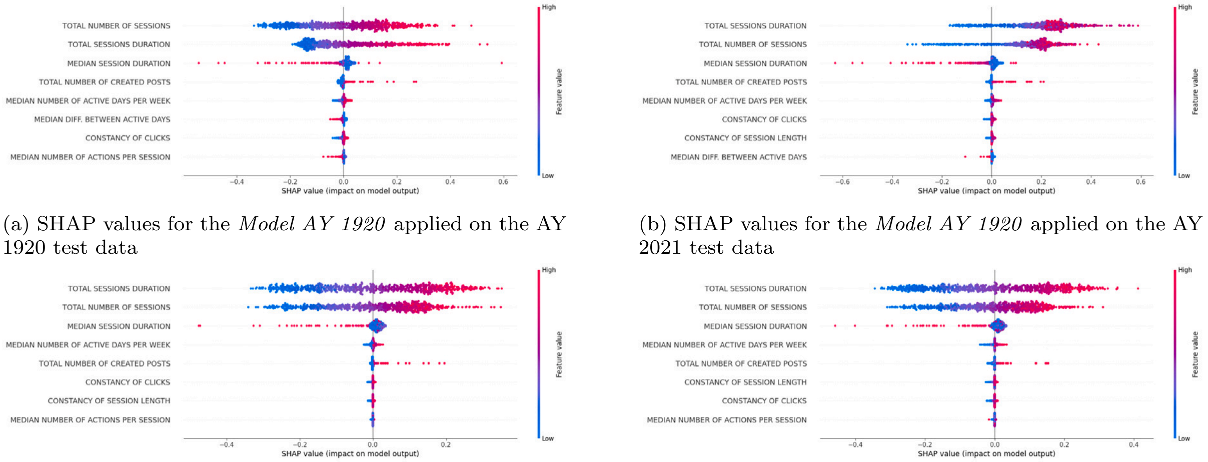
5. Theoretical and practical implications

Comparing the results of the concept drift detection, the performance stability analysis, and the feature agreement evaluation, we can see some common patterns between the techniques as well as the differences between them. On the one hand, we could see a general agreement between the techniques in terms of model stability measured with the data alone (distribution drift), the model itself (prediction drift and performance stability) and the feature agreements, which is also consistent with a general contextual change due to COVID-19. On the other hand, we could observe that even when the predictive performance is unstable (as in the case of the Accountancy course when comparing AY 19–20 and AY 20–21), some of the indicators remain stable, making the historical model still useful for decision making. In contrast, even when the differences in model performance are not significant (as in the case of the Global Economics course when comparing AY 18–19 and AY 19–20), we can still see the decreasing agreement between the models once they are updated with the new data. The above results highlight the added value of each technique for stability analysis and the overall advantage of using all techniques in combination for a comprehensive stability assessment.

The usefulness of each technique is highly dependent on the teaching and learning context in which predictive models are to be used. In

Table 6
The indicators portable across the models under investigation. Values in brackets represent the number of algorithms the feature is found to be stable for.

Accountancy		Global Economics	
AY 18-19 vs. AY 19-20	*AY 19-20 vs. AY 20-21*	*AY 18-19 vs. AY 19-20*	*AY 19-20 vs. AY 20-21*
Total sessions duration (8/8)	Total sessions duration (8/8)	Total sessions duration (8/8)	Total sessions duration (8/8)
Total number of sessions (6/8)	Constancy of clicks (5/8)	Total number of sessions (6/8)	Proportion of posts read (7/8)
Proportion of weeks with first-day activity (6/8)	Median number of active days per week (4/8)	Constancy of session length (3/8)	Total number of sessions (5/8)
Constancy of clicks (5/8)	Total number of sessions (3/8)	Proportion of posts read (3/8)	Proportion of active days (4/8)
Median number of active days per week (4/8)	Proportion of active days (3/8)	Median difference between active days (1/8)	Constancy of session length (3/8)



(a) SHAP values for the *Model AY 1920* applied on the AY 1920 test data

(b) SHAP values for the *Model AY 1920* applied on the AY 2021 test data

(c) SHAP values for the *Model AY 1920 retrained* applied on the AY 2021 test data

(d) SHAP values for the *Model AY 2021* applied on the AY 2021 test data

Fig. 5. SHAP summary plots for Naive Bayes.

particular, in places where GDPR guidelines are strictly enforced, access to real-time data may not be easy, and relying on the insights from past models is the only information available to the instructors. In this scenario, predictive performance is not the most important criterion for assessing stability, as the predictions are made post hoc and are not actionable. On the contrary, the feature importances returned by this model are actionable as they can be used for study advice. In this scenario, the stability of these importances is at the core of the stability evaluation, as it is critical to understand whether the insights from a previous year are still valid for the current year. Since our analysis includes eight widely used ML algorithms, it also aids in decision-making about which algorithms could be further used for decision making in cases where model stability is crucial. In this study, in line with its generative nature, the Naive Bayes algorithm was found to be the most stable according to the agreement metrics for both courses and for both academic years. However, the stability of its performance scores remains course dependent, as its optimization and retraining results in performance gains for the Accountancy course when comparing AY 19–20 with AY 20–21.

In addition, we believe that not only the stability results presented in this paper, but the stability assessment process itself is valuable to LA practitioners. The fact that we do not use raw log data, but rather engineer high-level study indicators, makes the framework generalizable to other LA datasets. Moreover, the engineered SRL indicators are widely used in the LA literature [5], highlighting the general availability of data for these indicators.

6. Conclusion

The quality of real-world ML applications depends heavily on the stability of the models when they are applied to new, unseen data. This stability is traditionally evaluated by detecting distributional changes in data and model predictions, shifts in model performance and changes in model parameters [20,21]. Particularly in the LA domain, the stability of predictive models has been based on performance scores and model coefficients (mostly using linear models, which severely limits the insights that can be extracted). The stability problem in LA is exacerbated due to the data protection regulations and ethical concerns that make historical models the best proxy for data-driven decision making.

To this end, in addition to investigating concept drift and performance stability of the academic success prediction models, this study demonstrated how feature importance rankings obtained from the SHAP framework can be used to evaluate their stability. Trace data was mapped to long-standing, well-established SRL indicators and newer indicators derived from customer consumption studies by drawing parallels between student study sessions and customer transactions. Specifically, uniformity, regularity, and bingeing of/in study sessions were included as SRL characteristics. These indicators were then used to predict success using eight ML algorithms. Concept drift was detected in both the SRL indicators and the model predictions using the KS statistical test and the PSI index. Performance metrics along with SHAP explanations were generated for different applications of these models and then used to further investigate the stability of both predictive power and feature importance. In addition, SHAP explanations were used to provide study advice.

First, with the concept drift analysis, we demonstrated that external context changes manifest themselves in the collected LMS study data traces and subsequently in the SRL indicators engineered based on this data. These changes come in line with the severity of the contextual changes, particularly, COVID effects researched in this study.

Second, we found that model performance also changes from one year to another, with the change influenced by the contextual changes similar to the concept drift detection analysis. However, the results for another course show that when the study contexts are similar, the performance differences are not significant and the predictive power shows stability.

Third, we introduced a novel way to measure model stability using the global feature importance rankings obtained from the SHAP framework. We showed that the indicators representing the overall level of activity remain stable even in the presence of concept drift and performance instability. We also illustrated how the agreement between indicators decreases when the model is changed, with the simultaneous hyperparameter optimization and retraining resulting in the lowest agreement. Additionally, we showed how the SHAP plots generated for the stable models can be used by instructors in determining study advice.

We find SHAP to transcend several limitations of linear methods as generated SHAP explanations can display both feature importance ranking and feature effects that can be subsequently used for decision-making. This bi-fold usefulness of SHAP summary plots enables (1) stability evaluation and (2) the summarization of information in an interpretable way to support instructors and researchers in recognizing how the study indicators relate to academic success. It goes without saying that these findings when shared with students could also equip them to better self-regulate toward success.

We acknowledge several limitations of our study. First, local explanation methods may be more relevant to decision making than the global methods we examined, suggesting a promising area for future investigation. Second, while our paper delves into data drift literature, there is room for expansion in sample selection methods, validation strategies, and modeling approaches. Third, we focused on feature importance ranking for model stability, leaving the analysis of feature effects stability and algorithm performance comparison for future studies. Moreover, exploring alternative XAI techniques and validating our findings across different educational contexts are areas for future research. Finally, extending the application of our proposed stability assessment beyond LA remains an avenue for exploration.

CRediT authorship contribution statement

Elena Tiukhova: Conceptualization, Data curation, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. **Pavani Vemuri:** Conceptualization, Data curation, Methodology, Writing – review & editing. **Nidia López Flores:** Conceptualization, Methodology, Writing – review & editing. **Anna Sigridur Islind:** Conceptualization, Methodology, Supervision, Writing – review & editing. **María Óskarsdóttir:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Stephan Poelmans:** Supervision, Writing – review & editing. **Bart Baesens:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Monique Snoeck:** Conceptualization, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

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.

Data availability

The data that has been used is confidential.

Acknowledgments

The third author acknowledges the support of the Icelandic Research Fund (IRF) [grant number 239408-051].

References

[1] K. Coussement, M. Phan, A. De Caigny, D.F. Benoit, A. Raes, Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model, Decis. Support Syst. 135 (2020) 113325.

[2] M. Phan, A. De Caigny, K. Coussement, A decision support framework to incorporate textual data for early student dropout prediction in higher education, Decis. Support Syst. 168 (2023) 113940.

[3] A. Abu Saa, M. Al-Emran, K. Shaalan, Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques, Tech. Knowl. Learn. 24 (2019) 567–598.

[4] P.H. Winne, R.S. Baker, et al., The potentials of educational data mining for researching metacognition, motivation and self-regulated learning, J. Educ. Data Min. 5 (1) (2013) 1–8.

[5] J. Jovanović, M. Saqr, S. Joksimović, D. Gašević, Students matter the most in learning analytics: The effects of internal and instructional conditions in predicting academic success, Comput. Educ. 172 (2021) 104251.

[6] A. Khan, S.K. Ghosh, Student performance analysis and prediction in classroom learning: A review of educational data mining studies, Educ. Inf. Technol. 26 (2021) 205–240.

[7] N. Sghir, A. Adadi, M. Lahmer, Recent advances in predictive learning analytics: A decade systematic review (2012–2022), Educ. Inf. Technol. (2022) 1–35.

[8] K.W. De Bock, K. Coussement, A. De Caigny, R. Slowiński, B. Baesens, R.N. Boute, T.-M. Choi, D. Delen, M. Kraus, S. Lessmann, et al., Explainable AI for operational research: A defining framework, methods, applications, and a research agenda, European J. Oper. Res. (2023).

[9] Y. Jang, S. Choi, H. Jung, H. Kim, Practical early prediction of students' performance using machine learning and explainable AI, Educ. Inf. Technol. (2022) 1–35.

[10] M. Afzaal, J. Nouri, A. Zia, P. Papapetrou, U. Fors, Y. Wu, X. Li, R. Weegar, Explainable AI for data-driven feedback and intelligent action recommendations to support students self-regulation, Front. Artif. Intell. 4 (2021) 723447.

[11] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, in: NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 4768–4777.

[12] V. Swamy, B. Radmehr, N. Krco, M. Marras, T. Käser, Evaluating the explainers: Black-box explainable machine learning for student success prediction in MOOCs, in: Proc. of the 15th International Conference on EDM, 2022, p. 98.

[13] D.T. Tempelaar, B. Rienties, B. Giesbers, Verifying the stability and sensitivity of learning analytics based prediction models: An extended case study, in: Computer Supported Education, Springer International Publishing, Cham, 2016, pp. 256–273.

[14] A. Mathrani, T. Susnjak, G. Ramaswami, A. Barczak, Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics, Comput. Educ. Open 2 (2021) 100060.

[15] N.G. López Flores, A.S. Islind, M. Óskarsdóttir, Effects of the COVID-19 pandemic on learning and teaching: A case study from higher education, 2021, arXiv preprint arXiv:2105.01432.

[16] M. Saqr, J. Jovanovic, O. Viberg, D. Gašević, Is there order in the mess? A single paper meta-analysis approach to identification of predictors of success in learning analytics, Stud. High. Educ. 47 (12) (2022) 2370–2391.

[17] A. Peña-Ayala, Educational data mining: A survey and a data mining-based analysis of recent works, Expert Syst. Appl. 41 (4) (2014) 1432–1462.

[18] S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, H. Lakkaraju, The disagreement problem in explainable machine learning: A practitioner's perspective, 2022, arXiv preprint arXiv:2202.01602.

[19] T. Karunaratne, For learning analytics to be sustainable under GDPR—Consequences and way forward, Sustainability 13 (20) (2021) 11524.

[20] F. Bayram, B.S. Ahmed, A. Kassler, From concept drift to model degradation: An overview on performance-aware drift detectors, Knowl.-Based Syst. 245 (2022) 108632.

[21] N. Lu, G. Zhang, J. Lu, Concept drift detection via competence models, Artificial Intelligence 209 (2014) 11–28.

[22] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, G. Zhang, Learning under concept drift: A review, IEEE Trans. Knowl. Data Eng. 31 (12) (2018) 2346–2363.

[23] A. Kolmogoroff, Confidence limits for an unknown distribution function, Ann. Math. Stat. 12 (4) (1941) 461–463.

[24] N.V. Smirnov, Approximate laws of distribution of random variables from empirical data, Uspekhi Mat. Nauk (10) (1944) 179–206.

[25] F. Wilcoxon, Individual comparisons by ranking methods, Biometr. Bull. 1 (6) (1945) 80–83.

[26] B. Su, Y.-D. Shen, W. Xu, Modeling concept drift from the perspective of classifiers, in: 2008 IEEE Conference on Cybernetics and Intelligent Systems, 2008, pp. 1055–1060.

[27] E. Lima, C. Mues, B. Baesens, Monitoring and backtesting churn models, Expert Syst. Appl. 38 (1) (2011) 975–982.

[28] I. Žliobaitė, M. Pechenizkiy, J. Gama, An overview of concept drift applications, in: Big Data Analysis: New Algorithms for a New Society, Springer, 2016, pp. 91–114.

[29] B. Baesens, D. Roesch, H. Scheule, Credit Risk Analytics: measurement Techniques, Applications, and Examples in SAS, John Wiley & Sons, 2016.

[30] B. Yurdakul, J. Naranjo, Statistical properties of the population stability index, J. Risk Model Valid. 14 (4) (2019).

[31] P.H. Winne, A.F. Hadwin, Studying as self-regulated learning, in: Metacognition in Educational Theory and Practice, Lawrence Erlbaum Associates Publishers, 1998, pp. 277–304.

[32] R.A. Rasheed, A. Kamsin, N.A. Abdullah, Challenges in the online component of blended learning: A systematic review, Comput. Educ. 144 (2020) 103701.

[33] J. Jovanovic, N. Mirriahi, D. Gašević, S. Dawson, A. Pardo, Predictive power of regularity of pre-class activities in a flipped classroom, Comput. Educ. 134 (2019) 156–168.

[34] D. Gašević, S. Dawson, T. Rogers, D. Gasevic, Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success, Internet High. Educ. 28 (2016) 68–84.

[35] G. Lust, N.A.J. Collazo, J. Elen, G. Clarebout, Content management systems: Enriched learning opportunities for all? Comput. Hum. Behav. 28 (3) (2012) 795–808.

[36] S. Van Goidsenhoven, D. Bogdanova, G. Deeva, S.v. Broucke, J. De Weerdt, M. Snoeck, Predicting student success in a blended learning environment, in: Proceedings of the tenth international conference on learning analytics & knowledge, 2020, pp. 17–25.

[37] D. Tzimas, S. Demetriadis, Ethical issues in learning analytics: A review of the field, Educ. Technol. Res. Dev. 69 (2021) 1101–1133.

[38] H. Khosravi, S.B. Shum, G. Chen, C. Conati, Y.-S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, D. Gašević, Explainable artificial intelligence in education, Comput. Educ.: Artif. Intell. 3 (2022) 100074.

[39] R. Farrow, The possibilities and limits of XAI in education: A socio-technical perspective, Learn. Media Technol. (2023) 1–14.

[40] T. Mu, A. Jetten, E. Brunskill, Towards suggesting actionable interventions for wheel spinning students, in: Proceedings of the 13th International Conference on EDM, 2020, International Educational Data Mining Society, 2020.

[41] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" explaining the predictions of any classifier, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 1135–1144.

[42] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, P. Das, Explanations based on the missing: Towards contrastive explanations with pertinent negatives, in: Advances in Neural Information Processing Systems, vol. 31, 2018.

[43] R.K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 607–617.

[44] E. Melo, I. Silva, D.G. Costa, C.M. Viegas, T.M. Barros, On the use of explainable artificial intelligence to evaluate school dropout, Educ. Sci. 12 (12) (2022) 845.

[45] F. Afrin, M. Hamilton, C. Thevathyan, On the explanation of AI-based student success prediction, in: International Conference on Computational Science, Springer, 2022, pp. 252–258.

[46] E. Tiukhova, P. Vemuri, M. Óskarsdóttir, S. Poelmans, B. Baesens, M. Snoeck, Discovering unusual study patterns using anomaly detection and XAI, in: Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS), 2024.

[47] V.K. Singh, B. Bozkaya, A. Pentland, Money walks: Implicit mobility behavior and financial well-being, PLoS One 10 (8) (2015) e0136628.

[48] N.L. Flores, A.S. Islind, M. Óskarsdóttir, A learning analytics-driven intervention to support students' learning activity and experiences, in: Digitalization and Digital Competence in Educational Contexts, Routledge, 2023, pp. 81–102.

[49] M. Óskarsdóttir, C. Bravo, C. Sarraute, J. Vanthienen, B. Baesens, The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics, Appl. Soft Comput. 74 (2019) 26–39.

[50] C. Sammut, G.I. Webb, Encyclopedia of Machine Learning, Springer Science & Business Media, 2011.

[51] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intell. Syst. Appl. 13 (4) (1998) 18–28.

[52] J.A. Nelder, R.W.M. Wedderburn, Generalized linear models, J. R. Stat. Soc. 135 (3) (1972) 370–384.

[53] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inform. Theory 13 (1) (1967) 21–27.

[54] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[55] W. Chen, C.G. Brinton, D. Cao, A. Mason-Singh, C. Lu, M. Chiang, Early detection prediction of learning outcomes in online short-courses via learning behaviors, IEEE Trans. Learn. Technol. 12 (1) (2018) 44–58.

[56] J.H. Friedman, Greedy function approximation: A gradient boosting machine, Ann. Statist. (2001) 1189–1232.

[57] A. Asselman, M. Khaldi, S. Aammou, Enhancing the prediction of student performance based on the machine learning XGBoost algorithm, Interact. Learn. Environ. 31 (6) (2023) 3360–3379.

[58] S.O. Arik, T. Pfister, TabNet: Attentive interpretable tabular learning, in: Proc. AAAI Conference on Artificial Intelligence, vol. 35, (no. 8) 2021, pp. 6679–6687.

[59] M. Baranyi, M. Nagy, R. Molontay, Interpretable deep learning for university dropout prediction, in: Proc. SIGITE '20, 2020, pp. 13–19.

[60] B.N. Anh, N.H. Giang, N.Q. Hai, T.N. Minh, N.T. Son, B.D. Chien, An university student dropout detector based on academic data, in: ISIEA, IEEE, 2023, pp. 1–8.

[61] L.S. Shapley, Notes on the n-Person Game—II: The Value of an n-Person Game.(1951), Lloyd S Shapley, 1951.

[62] D. Janzing, L. Minorics, P. Blöbaum, Feature relevance quantification in explainable AI: A causal problem, in: AISTATS, PMLR, 2020, pp. 2907–2916.

[63] H. Chen, J.D. Janizek, S. Lundberg, S.-I. Lee, True to the model or true to the data? 2020, arXiv preprint arXiv:2006.16234.

[64] S. Müller, V. Toborek, K. Beckh, M. Jakobs, C. Bauckhage, P. Welke, An empirical evaluation of the rashomon effect in explainable machine learning, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2023, pp. 462–478.

[65] M. Neely, S.F. Schouten, M.J. Bleeker, A. Lucic, Order in the court: Explainable ai methods prone to disagreement, 2021, arXiv preprint arXiv:2105.03287.

[66] Z. Lei, H. Zhou, W. Hu, G.-P. Liu, Impact of COVID-19 pandemic on engineering education: Case study with the online laboratory ncslab, Int. J. Eng. Educ. (2022) 1505–1512.

Elena Tiukhova is a Ph.D. student at the Research Center for Information Systems Engineering, KU Leuven. She holds a Master of Science in Business and Information Systems Engineering with a focus in Computer Science. Her research is about using machine learning and deep learning techniques for anomaly detection and XAI with applications in marketing, credit risk and learning analytics.

Pavani Vemuri is a Ph.D. student in the LIRIS research group at the KU Leuven with research interests in blended learning and adaptive blending learning environments that facilitate personalized learning. Her primary research is on how insights from Learning analytics can be put to use in helping teachers in orchestration or to redesign.

Nidia Guadalupe López Flores is a Ph.D. candidate at the Department of Computer Science at Reykjavik University, Iceland. Her research interests focus on data and network science with applications to learning analytics and educational data mining. Her Ph.D. research focuses on learning analytics and data science in education to investigate learner profiles and improve teaching and learning practices. Before starting her Ph.D., Nidia worked in industry for several years and obtained an M.Sc. in Business Analytics and Management Sciences from the University of Southampton, United Kingdom.

Anna Sigridur Islind has a Ph.D. in informatics from University West in Sweden. She is an Associate Professor at the Department of Computer Science at Reykjavik University and is the co-director of the research centre CISDAS: Centre for Information Systems and Data Science Research. Moreover, she is the leader of the digital innovation in the Sleep Revolution project, a large-scale Horizon 2020 project with 39 partners across Europe funded by the European Union, and she is the program manager for the M.Sc. in Digital Health, a novel multi-disciplinary study line at Reykjavik University. She has published 70 peer-reviewed papers most of which focus on information systems and doing good through being socially aware and ethically driven when designing, developing, and using digital platforms, data and artificial intelligence in general.

María Óskarsdóttir is an Associate Professor at the Department of Computer Science at Reykjavík University and an Adjunct Professor at Western University. She holds a Ph.D. in Business Analytics from the Faculty of Economics and Business at KU Leuven in Belgium. Her research is focussed on the intersection of network science and machine learning, looking at practical applications of data science and analytics whereby she leverages advanced machine learning techniques, network science, and various sources of data with the goal of increasing the impact of the analytics process and facilitating better usage of data science for decision making in various domains, such as finance, learning, marketing, health care and sustainability. She has over 50 publications in high-impact journals and conferences in the domains of operations research, network science and information systems. She serves as editor at Machine Learning.

Stephan Poelmans is a professor of the Research Centre for Information Systems Engineering (LIRIS) at the faculty of Economics and Business of the KU Leuven (campus Brussels). His research interests include process and conceptual modeling, technology acceptance and usability, and technology-enhanced teaching and learning. Stephan teaches courses in data management, data mining, and ICT management in business engineering and business administration programs.

Professor Bart Baesens is a professor of Data Science at KU Leuven (Belgium), and a lecturer at the University of Southampton (UK). He co-authored more than 250 scientific papers and 10 books. Bart received the OR Society's Goodeve medal for best JORS paper in 2016 and the EURO 2014 and EURO 2017 award for best EJOR paper. His research is summarized at www.dataminingapps.com. Bart is listed in the top of Stanford University's new Database of Top Scientists in the World. He was also named one of the World's top educators in Data Science by CDO magazine in 2021 and has educated tens of thousands of data scientists across the globe. Bart also has his own ON-LINE learning BlueCourses platform: www.bluecourses.com.

Monique Snoeck is full professor at the KU Leuven, Research Center for Management Informatics (LIRIS), Belgium, and visiting professor at the UNamur, Belgium. Her research focuses on conceptual modeling, enterprise modeling, requirements engineering, model-driven engineering, and business process management and the teaching and learning of these topics. Previous research has resulted in the Enterprise Information Systems Engineering approach MERODE, and its companion e-learning and prototyping tools MERLIN and MERLIN Prototyper. In the domains of Smart Learning environments and Technology enhanced learning, she focuses on intelligent feedback provisioning and learning analytics with the aim of defining features predictive for learner success and learner engagement. She has (co-)authored over 130 peer-reviewed papers, half of which peer-reviewed journal papers. She is involved in numerous conferences in the domains of Information Systems such as CAiSE, PoEM, ER, and EMMSAD.