Trustworthy Academic Risk Prediction with Explainable Boosting Machines

Vegenshanti Dsilva, Johannes Schleiss(✉), and Sebastian Stober

Otto-von-Guericke University, Magdeburg, Germany
{vegenshanti.dsilva,johannes.schleiss,stober}@ovgu.de

Abstract. The use of predictive models in education promises individual support and personalization for students. To develop trustworthy models, we need to understand what factors and causes contribute to a prediction. Thus, it is necessary to develop models that are not only accurate but also explainable. Moreover, we need to conduct holistic model evaluations that also quantify explainability or other metrics next to established performance metrics. This paper explores the use of Explainable Boosting Machines (EBMs) for the task of academic risk prediction. EBMs are an extension of Generative Additive Models and promise a state-of-the-art performance on tabular datasets while being inherently interpretable. We demonstrate the benefits of using EBMs in the context of academic risk prediction trained on online learning behavior data and show the explainability of the model. Our study shows that EBMs are equally accurate as other state-of-the-art approaches while being competitive on relevant metrics for trustworthy academic risk prediction such as earliness, stability, fairness, and faithfulness of explanations. The results encourage the broader use of EBMs for other Artificial Intelligence in education tasks.

Keywords: Explainable AI in Education · Responsible AI · Trustworthy Machine Learning · Academic Risk Prediction · Virtual Learning

1 Introduction

Predictive models in education are used for a range of tasks, such as predicting enrollment numbers or student performance, engagement and satisfaction of classroom activities, or identifying at-risk students [22,25,29]. In this context, academic risk prediction focuses on predicting if a student fails or drops out of a course or class based on either past academic data or learning behaviors [1,3,22,25]. Thus, the use of predictive models in education should support learners in their learning journey and provide enriched information for educators or staff to give additional support in the classroom or on a system level [37].

Like in other AI domains, we face the trade-off between interpretable models and accurate models [9,17]. Interpretability in this context refers to a situation where humans can understand or even predict a model's output [21].

N. Wang et al. (Eds.): AIED 2023, LNAI 13916, pp. 463–475, 2023.
https://doi.org/10.1007/978-3-031-36272-9_38

To bring AI to the high-stake environment of education, it is core to address the need for stakeholders to understand the models' predictions from a user and a regulatory perspective [14]. The European General Data Protection Regulation (EU GDPR) [8], for example, states that "the existence of automated decision-making should carry meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject." Thus, users require additional information on the underlying logic of automated decision-making. Similarly, trustworthy AI [37], ethics of AI [14], and the importance of explainability [9,17] are discussed in the context of AI in education. In this context, Fiok et al. [9] mention transparency, causality, privacy, fairness, trust, usability, and reliability as goals of explainable AI (XAI) that build the basis for the responsible and trustworthy use of AI in education. This is especially relevant when aiming to understand the most important causes and underlying factors in academic risk prediction [3,12,25,36].

When interpreting model results, explanations can be categorized as either *ante-hoc* or *post-hoc*. *Post-hoc* explanations, such as LIME [26] or SHAP [20], attempt to explain the predictions of a black-box model after it has been trained. However, these methods are limited by the assumption that a model can be approximated and require a lot of computing time. Additionally, Swamy et al. [35,36] found that *post-hoc* XAI methods often disagree and that human experts are not good validators of explainability. Therefore, using inherently interpretable models, or *ante-hoc* explanations, is recommended.

Recently, Explainable Boosting Machines (EBMs) [24] were introduced as an extension to Generative Additive Models [13]. Through its additive nature, the model is inherently interpretable while achieving state-of-the-art performance on tabular data on a range of tasks [4,23,24,38]. This paper explores the use of EBMs as an interpretable model for the task of academic risk prediction. The main contribution of the paper is to demonstrate the benefits of EBMs trained with online learning behaviors. We show how the interpretability of EBMs aligns with the causality a teacher might use to provide support to students in real life. Moreover, we compare EBMs to other state-of-the-art models on performance and relevant metrics in trustworthy academic risk prediction, such as earliness, stability, fairness, and faithfulness of explanations.

The paper is structured as follows. Section 2 gives the background on academic risk prediction, explainable AI in education, and EBMs. Next, Sect. 3 introduces the analysis approach including the data, training, and the proposed model assessment. Section 4 presents the experimental results of performance and other metrics and discusses the implications of the results. Last, Sect. 5 concludes the study and gives an outlook on future research.

2 Background

2.1 Academic Risk Prediction

Academic risk prediction is a part of predictive learning analytics and describes the task to identify students that might fail or drop a class or a course as a classification task [22,25,29]. Before the Covid-19 pandemic, demographic data and prior academic data were typically used as features for academic risk prediction [25,27]. Recently, more online learning behavioral data, such as interactions with Learning Management Systems, is available and used as input data for predictive models [25,29]. Common approaches include Artificial Neural Networks, Naive Bayes, K-Nearest-Neighbor as well as tree-based models like Random Forest, Decision Tree, or Gradient Boosting [25,29]. Typically, classification models are evaluated using performance measures such as the accuracy, F-measure, precision, or area under the curve [22,25,29]. In this context, most papers for academic risk prediction only assess a one-dimensional performance rubric (e.g. accuracy or precision), missing out on key considerations of trustworthy and explainable AI, such as earliness, stability, or quantitative explainability metrics [2,31].

2.2 Explainable AI in Education

Explanations are needed to gain the user's trust, improve the design, support the user's understanding of the recommendation and prediction, set the context of the prediction, and justify and rationalize an action, as means to communicate results [33]. Thus, explanations should focus on stakeholders' needs, and improve the perception, trust, and acceptance of users. Shin et al. [30] show that users evaluate explanations based on existing beliefs and partly on their understanding of the model. Moreover, they find that fairness, accountability, and transparency influence causability and trust in AI systems.

Similar to XAI in general, XAI in education focuses on global and local approaches for explanation [9,17,36]. However, state-of-the-art explainable post-hoc methods are not yet applied to most models of student performance predictions [2,6]. Moreover, the review of Almari et al. [2] shows that none of the analyzed studies has quantified the explainability of the proposed models, making it difficult to compare them on these metrics. They conclude that explainability metrics should be included next to the accuracy metric in the analysis and development of models for AI in education.

2.3 Explainable Boosting Machines

EBM [24] is an improved additive machine learning model based on Generalized Additive Models [13]. It can compete with state-of-the-art machine learning models on tabular data while being inherently interpretable [4,24,38]. EBMs address the need for local explanations through their additive model design, where each feature contributes individually to the prediction [24]. This means it has a feature importance build-in, which can simplify the understanding of the model's predictions. We demonstrate the explainability of EBMs in the context of academic risk prediction in Sect. 4.1. Moreover, Nori et al. [23] also have added Differential Privacy to EBM, which addresses the need to protect privacy, for example in educational data [28].

An EBM consists of feature functions $f_j(x_j)$ that are learned individually for each feature x_j, and pairwise feature interactions $f_{ij}(x_i, x_j)$ for the most important feature pairs. In the training phase, individual trees are trained on each feature separately in a round-robin fashion with a low learning rate (to diminish the effects of feature sequence dependence) for several iterations using boosting and bagging methods. The learned trees are averaged to a feature function that provides the feature score for different feature value bins. Based on this, the pairwise feature interactions are trained using the FAST algorithm [19]. The final prediction $E[y]$ can be expressed as,

$$g(E[y]) = \beta_0 + \sum_j f_j(x_j) + \sum f_{ij}(x_i, x_j) \tag{1}$$

with x_i and x_j expressing features, f_j the feature function, and f_{ij} the pairwise interaction function for the most important feature pairs. $g(.)$ refers to the link function for different learning tasks and β_0 describes the intercept.

Related to AI in education, Jayasundara et al. [16] conducted performance prediction of students with an EBM using socio-economic features and pre-course academic performance and compared the model accuracy to other approaches. In terms of interpretability, they analyzed the correlation between features and labels and compared them with global and local feature importance. Building up on this, we propose training EBMs on data of online learning behaviors. Furthermore, we explore the use of EBMs as trustworthy models for academic risk prediction by demonstrating their explainability and evaluating them on various metrics relevant for trustworthy academic risk prediction.

3 Methods

3.1 Study Area and Data

Task. In our investigation, we evaluate EBM for two academic risk prediction tasks: failure prediction and dropout prediction.[1] We implement the task as binary classification so that the feature functions generated by EBM are easy to interpret.

Dataset. The study uses the Open University Learning Analytics Dataset (OULAD) [18]. It contains data from 32,593 students for seven course modules offered in two terms, each course lasting for approximately nine months. The dataset includes demographic information (e.g. gender and age), pre-course information (e.g. studied credits and whether a course was previously attempted by the student), and clickstream data from the virtual learning environment (VLE) for 20 types of VLE resources. Moreover, it contains four target classes: Distinction, Pass, Fail, and Withdrawn.

[1] The code is available under https://gitlab.com/vegeedsilva/trustworthy-academic-risk-prediction-with-explainable-boosting-machine.git.

Online Learning Behaviors as Features. The clickstream data available through the interaction with VLE can reveal learning behaviors of students. Existing research explores how learning behaviors like engagement, regularity, and curiosity influence students' academic performance. Using learning behaviors as features can allow predictions to be actionable when provided with a relevant explanation [11,32]. Here, we created learning behaviors as features based on the interactions of a learner with the VLE. Table 1 describes the learning behaviors used as features for this dataset.

Table 1. Online learning behaviors used as features for academic risk prediction with the count indicating the number of a type of feature

Learning Behavior	Description	Count
Engagement	Total number of interactions as sum of clicks that a learner has with each VLE resource	20
Session Count	Total number of study sessions spent by a learner for each VLE resource	20
Regularity	Count of blocks (continuous weeks of learning interactions with the VLE) as maximum, minimum, average, and variance	5
Lateness	Percentage of assessment deadlines missed by a learner	1
Curiosity	Percentage of VLE resource covered by the learner	20
Assessment Coverage	Percentage of assessments completed by a learner	1

After generating all features, we have 67 features in total. The target labels are encoded as 1 for *at-risk* and 0 for *non-risk* learners. We perform feature scaling and select 30 features using the MinMaxScaler and SelectK functions from the scikit-learn package. After filtering student records with multiple attempts for a course and only retaining the latest records in such cases, we have 31,284 student records. The dataset is split randomly with a 70-10-20 split for training, validation, and test set.

3.2 Training

To evaluate EBMs against other state-of-the-art models, we train a Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Feed-forward Neural Network (FNN) for failure prediction and dropout prediction. The length of the longest course in OULAD is 39 weeks. Hence, we compute the appropriate feature values based on the course week and train the models with the final result as the target. The models have been optimized for their respective hyperparameters. As a separate model was trained for each week for the two tasks, we have only mentioned important parameter details of the selected models. The selected EBM models were trained for 5000 rounds with 10 feature pairs and a quantile method for binning. The LR models were trained for 200 iterations. The DT and RF models used Gini impurity as their splitting criteria and the selected FNN models were composed of three layers, with ReLU activation, Adam optimizer, and binary cross-entropy loss.

3.3 Model Assessment

Accuracy. As the OULAD is a balanced dataset, we use accuracy as a performance measure and compare the accuracy development through the 39 weeks of the course period for different models.

Earliness and Stability. Especially in the context of education, early prediction of risk can facilitate scaffolding and intervention at the right time and enable learners to improve on their learning path [1,5]. Hence, Soussia et al. [31] introduced earliness and stability measures to evaluate academic risk prediction systems. *Earliness* is computed as the average time period when the first correct predictions for each record are made. One drawback here is that earliness does not reflect the accuracy of the model. Hence, the authors propose using the harmonic mean (HM) of earliness and accuracy to have a well-rounded assessment of prediction systems [31]. We use HM to evaluate the earliness of different models as follows,

$$HM = \frac{2 * (1 - earliness) * accuracy}{(1 - earliness) + accuracy} \tag{2}$$

In this context, a higher HM value is desirable for early and accurate predictions.

(Temporal) stability helps to gain user trust in the model's ability to make correct predictions over time. It describes the average of the highest number of weeks the model can deliver successive correct predictions over the course period. A higher stability score indicates a stable predictor. Thus, earliness and stability are significant parameters for evaluation when investigating the suitability of academic risk predictors.

Fairness. Bias can slide into a machine learning pipeline at several stages, including imbalanced datasets or missing data, bias during model training, and user perception during data collection. Here, we assess the algorithmic bias of the dataset and the mentioned models. For this purpose, we compute group fairness metrics such as the *statistical parity difference (SPD)* [7] and *equal opportunity difference (EOD)* [10]. *SPD* is defined as the difference between the probabilities of the protected and the majority groups in obtaining a favorable decision. It is described as

$$SPD = P(Y_{pred} = 1 \mid F = min) - P(Y_{pred} = 1 \mid F = maj) \tag{3}$$

where Y_{pred} are the model predictions and F is the group of the sensitive attribute. A SPD value of zero indicates fairness.

EOD also takes into account the class label and measures the deviation from equal opportunity. In this context, equal opportunity implies that both the privileged and unprivileged group have the same probability of obtaining a favorable class. Thus, an EOD value of zero indicates fairness. EOD can be expressed as

$$EOD = P(Y_{pred} = 1 \mid F = min, Y = 1) - P(Y_{pred} = 1 \mid F = maj, Y = 1) \tag{4}$$

where Y_{pred} are the model predictions, F is the sensitive attribute and Y are the true labels.

Table 2. Mean feature importance score assigned by EBM for predicting dropouts in week 20 for the top five features.

Feature	Mean Importance Score
Lateness	0.667
Curiosity: quiz	0.422
Assessment Coverage	0.414
Regularity: Block count	0.366
Curiosity: subpage	0.234

Faithfulness of Local Explanations. A faithful explanation should provide insight into the rationale of the model to arrive at its prediction. Although there is no standard method to quantitatively evaluate the faithfulness of explanations, different measures have been introduced in the field of XAI [15,34]. Here, we measure the faithfulness of a model by computing the average of recall obtained after re-training the model over the top important features that contribute at least 50% towards the final outcome in the local explanations of the test set. This measure is called *recall on important features (ROIF)* [26]. It allows evaluating the consistency between the feature importance in the local explanations and the true importance given by the model.

4 Experimental Results

4.1 Explainability of EBMs

Before providing the results of the model assessments, we want to demonstrate the interpretability of EBMs in the context of educational data. The first step is to understand how EBMs work on inference time. To provide a prediction outcome, EBMs use feature functions and pairwise feature interaction heatmaps as lookup tables. Thus, EBMs provide the possibility to access the feature importance by default.

For example, Table 2 shows the overall mean importance of the features in EBM for dropout prediction in week 20 for the five most important features. Furthermore, we can also plot the feature shape function to understand the influence of the feature, as demonstrated in Fig. 1 for the feature "lateness". We can observe that when the lateness is very low (below 0.2), the model gives a score of -1, pushing toward the *non-risk* class (i.e. label 0). Correspondingly, when lateness is high, its score is positive, pushing the overall prediction towards *at-risk* (i.e. label 1). In terms of model understanding, this corresponds to our didactic understanding. For example, assume a teacher observes that a student has been submitting most assignments late, we can expect that the teacher will give special attention to the student, as there might be a risk of dropout or failure. Thus, using EBMs with online learning behavior data allows interpretable and reasonable predictions that are easy to access and provide support for students and teachers.

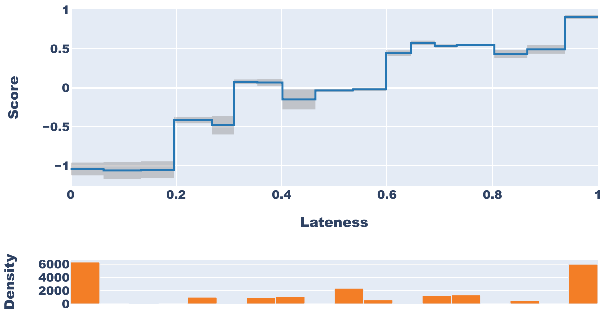

Fig. 1. Feature function for the feature "lateness" generated by EBM for predicting learners at risk of dropout in week 20 with the gray boxes representing variance around a score. The lower graph represents the density in the different feature value bins.

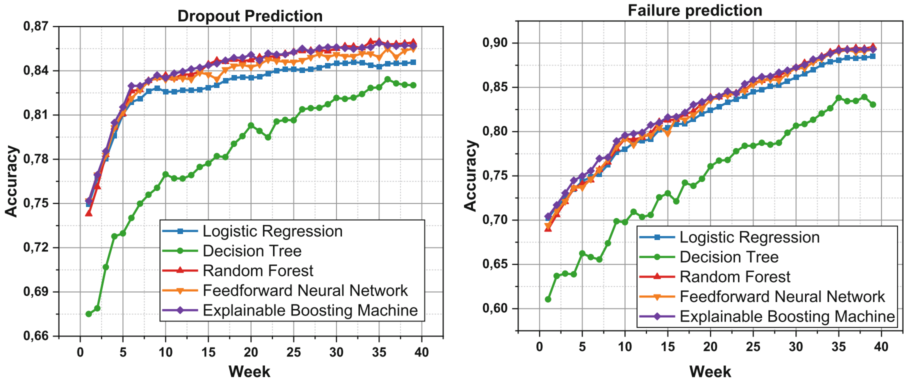

Fig. 2. Comparison of model accuracy evolution through course period for dropout prediction (left) and failure prediction (right).

4.2 Accuracy

The accuracy of the models per course week for the dropout and failure prediction is shown in Fig. 2. We can observe that the accuracy to predict the events increases over time, with all models except DT performing in a similar accuracy range. This demonstrates that EBMs perform equally well compared to established methods in the context of academic risk prediction with online learning behaviors.

4.3 Earliness and Stability

The earliness and stability results are presented in Table 3. When evaluating models for earliness the group of learners *at-risk* is of special focus to identify such learners as early as possible. For the task of predicting dropouts, we observe similar HM earliness of around 85% for the *at-risk* group for EBM along with DT, RF, and FNN. For the failure prediction, DT has the highest HM earliness score. Here, the earliness of EBM, although not the best, is at par with the remaining models (except DT). Concerning the stability of models, we can observe that LR provides stable correct predictions for the longest period. At the same time, the EBM is comparably stable in both tasks.

Table 3. Harmonic mean earliness and stability measures for task of dropout prediction and failure prediction for *at-risk* (AR) and *non-risk* (NR) students.

Model	HM Earliness (Percentage)				Stability (No. of weeks)			
	Dropout		Failure		Dropout		Failure	
	(AR)	(NR)	(AR)	(NR)	(AR)	(NR)	(AR)	(NR)
EBM	85.53	**74.64**	78.79	70.02	22.86	33.97	25.67	**36.47**
LR	84.15	73.94	77.40	**71.42**	**23.94**	**34.35**	**26.24**	36.28
DT	85.19	60.64	**82.77**	54.62	22.54	24.59	20.51	26.94
RF	85.61	71.55	79.26	65.95	**23.93**	33.12	24.31	36.10
FNN	**85.88**	74.61	79.01	67.46	22.98	33.57	24.33	36.09

Table 4. Algorithmic fairness for gender and disability expressed through statistical parity difference (SPD) and equal opportunity difference (EOD) for both tasks of dropout and failure prediction.

Model	Gender				Disability			
	Dropout		Failure		Dropout		Failure	
	SPD	EOD	SPD	EOD	SPD	EOD	SPD	EOD
EBM	–0.031	–0.024	–0.024	**0.003**	0.053	0.028	0.054	0.012
LR	**–0.023**	**–0.010**	**–0.012**	0.011	0.048	0.028	0.056	0.018
DT	–0.028	–0.014	–0.023	–0.006	**0.036**	0.026	**0.050**	**–0.009**
RF	–0.039	–0.025	–0.018	0.008	0.050	**0.017**	0.063	0.028
FNN	–0.032	–0.017	–0.022	0.006	0.056	0.034	**0.050**	0.015

4.4 Fairness

Before looking at the model behavior, we need to assess the balance of the dataset for the gender and disability attributes. The SPD of the OULAD for the attribute gender is 0.002 for dropout and −0.019 for failure prediction with females as the privileged group for both cases. Respectively, the SPD for the attribute disability is 0.076 for dropout and 0.065 for failure prediction with non-disabled learners as the privileged group. Thus, the dataset is balanced for both attributes. Table 4 illustrates the results for fairness metrics SPD and EOD in week 39 for all trained models. With this balanced dataset in place, we can observe that all models perform mostly similarly concerning the evaluated fairness metrics.

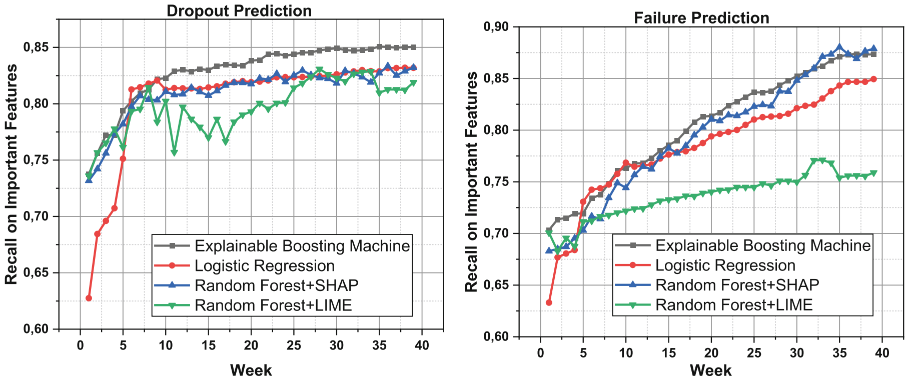

Fig. 3. Average ROIF (for top five features) for dropout prediction (left) and failure prediction (right).

4.5 Faithfulness of Explanations

We compare the faithfulness of local explanations given by EBM to LR, another inherently interpretable model, and to the application of post-hoc explainability methods such as LIME [26] and SHAP [20] on a RF model. Figure 3 shows the ROIF score of the methods through the course period. In the dataset, the initial period is critical for retaining students as 60% of the total dropouts occur in the first ten weeks of the course period. From our results, we can observe that for dropout and failure prediction the ROIF score for EBM is generally higher than all other explainability methods.

4.6 Discussion

The experiments evaluated EBM in comparison to other state-of-the-art models for academic risk prediction on online learning behavior data. Furthermore, we showed how EBMs are inherently interpretable and how this can benefit the trustworthiness of their use in an academic risk prediction task. The experiments on one dataset indicate that EBMs are equally accurate as other models and perform very stable. All models perform in a similar range on other metrics such as earliness and fairness. Concerning the faithfulness of local explanations, EBMs show slightly better ROIF scores compared to LR as another interpretable model or the use of model-agnostic post-hoc explainability methods, such as SHAP and LIME used in combination with RF. At the same time, more analysis is needed to generalize these results.

Using EBMs in production is well supported through the InterpretML package [24]. They are especially useful for datasets that are well understood and features that have no complex interdependence. However, EBMs currently only support tree structures as base learners. Furthermore, the application of EBMs for multi-class prediction settings, for example identifying the risk of failure and dropouts in the same setting, is still under ongoing research [24]. This also supports the argument for improving and using model-agnostic post-hoc interpretability methods. In any case, to develop trustworthy academic risk prediction, we need to make it a standard to include quantified explainability metrics. Regarding EBMs, we can conclude that it is a useful model for tabular datasets in education in tasks that require interpretability of the model.

5 Conclusion and Outlook

To address the need for trustworthiness, we demonstrated the use of EBMs for the task of academic risk prediction using data from online learning behaviors. Through their additive nature, EBMs are inherently interpretable as the contribution of each feature is visible. In our experiments, we demonstrated the interpretability of EBMs and quantified the trustworthiness of the model in comparison to other state-of-the-art models using metrics such as faithfulness of explanation, fairness, earliness, and stability. We observe that EBMs are able to capture the influence of different learning behaviors on the risk of dropout and failure in accordance with the causality a teacher might use to provide support to students in real life. Moreover, we find that EBMs perform similarly to other models for the metrics discussed in the academic risk context. Thus, we can conclude that its explainability could be a significant advantage that can give insights and create trust for all stakeholders involved [35]. The results encourage broader use of EBM for other tasks in AI in education that use tabular data.

Future work will test the model on other open access datasets, compare it to further developments of GAMs, and explore the suggested application of Differential Privacy on EBM [23] with educational data. Moreover, a user study is planned to analyze the impact of explainability in a real scenario.

References

1. Adnan, M., et al.: Predicting at-risk students at different percentages of course length for early intervention using machine learning models. IEEE Access **9**, 7519–7539 (2021)
2. Alamri, R., Alharbi, B.: Explainable student performance prediction models: a systematic review. IEEE Access **9**, 33132–33143 (2021)
3. Baranyi, M., Nagy, M., Molontay, R.: Interpretable deep learning for university dropout prediction. In: Proceedings of the 21st Annual Conference on Information Technology Education, pp. 13–19 (2020)
4. Bussmann, N., Giudici, P., Marinelli, D., Papenbrock, J.: Explainable machine learning in credit risk management. Comput. Econ. **57**(1), 203–216 (2021)
5. Chen, F., Cui, Y.: Utilizing student time series behaviour in learning management systems for early prediction of course performance. J. Learn. Anal. **7**(2), 1–17 (2020)
6. Cohausz, L.: Towards real interpretability of student success prediction combining methods of XAI and social science. In: Proceedings of the 15th International Conference on Educational Data Mining, pp. 361–367 (2022)

7. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the Innovations in Theoretical CS Conference, pp. 214–226 (2012)
8. EU: Regulation EU 2016/679 of the European Parliament and of the Council of 27 April 2016. Official Journal of the European Union (2016)
9. Fiok, K., Farahani, F.V., Karwowski, W., Ahram, T.: Explainable artificial intelligence for education and training. J. Defense Model. Simul. **19**(2), 133–144 (2022)
10. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Adv. Neural Inf. Process. Syst. **29**, 3315–3323 (2016)
11. Hasan, R., Fritz, M.: Understanding utility and privacy of demographic data in education technology by causal analysis and adversarial-censoring. Proc. Priv. Enhanc. Technol. **2022**(2), 245–262 (2022)
12. Hasib, K.M., Rahman, F., Hasnat, R., Alam, M.G.R.: A machine learning and explainable AI approach for predicting secondary school student performance. In: IEEE 12th Annual Computing and Communication Workshop and Conference, pp. 0399–0405. IEEE (2022)
13. Hastie, T., Tibshirani, R.: Generalized additive models: some applications. J. Am. Stat. Assoc. **82**(398), 371–386 (1987)
14. Holmes, W., et al.: Ethics of AI in education: towards a community-wide framework. Int. J. Artif. Intell. Educ. **32**(3), 504–526 (2022)
15. Hooker, S., Erhan, D., Kindermans, P.J., Kim, B.: A benchmark for interpretability methods in deep neural networks. Adv. Neural Inf. Process. Syst. **32**, 9737–9748 (2019)
16. Jayasundara, S., Indika, A., Herath, D.: Interpretable student performance prediction using explainable boosting machine for multi-class classification. In: 2022 2nd International Conference on Advanced Research in Computing (ICARC), pp. 391–396. IEEE (2022)
17. Khosravi, H., et al.: Explainable artificial intelligence in education. Comput. Educ. Artif. Intell. **3**, 100074 (2022)
18. Kuzilek, J., Hlosta, M., Zdrahal, Z.: Open university learning analytics dataset. Sci. Data **4**(1), 1–8 (2017)
19. Lou, Y., Caruana, R., Gehrke, J., Hooker, G.: Accurate intelligible models with pairwise interactions. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 623–631 (2013)
20. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. **30**, 4765–4774 (2017)
21. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
22. Namoun, A., Alshanqiti, A.: Predicting student performance using data mining and learning analytics techniques: a systematic literature review. Appl. Sci. **11**(1), 237 (2020)
23. Nori, H., Caruana, R., Bu, Z., Shen, J.H., Kulkarni, J.: Accuracy, interpretability, and differential privacy via explainable boosting. In: International Conference on Machine Learning, pp. 8227–8237 (2021)
24. Nori, H., Jenkins, S., Koch, P., Caruana, R.: Interpretml: a unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223 (2019)
25. de Oliveira, C.F., Sobral, S.R., Ferreira, M.J., Moreira, F.: How does learning analytics contribute to prevent students' dropout in higher education: a systematic literature review. Big Data Cogn. Comput. **5**(4), 64 (2021)

26. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
27. Rubiano, S.M.M., Garcia, J.A.D.: Formulation of a predictive model for academic performance based on students' academic and demographic data. In: 2015 IEEE Frontiers in Education Conference (FIE), pp. 1–7. IEEE (2015)
28. Schleiss, J., Günther, K., Stober, S.: Protecting student data in ML pipelines: an overview of privacy-preserving ML. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds.) Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium. AIED 2022. LNCS, vol. 13356, pp. 532–536. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-11647-6_109
29. Sghir, N., Adadi, A., Lahmer, M.: Recent advances in predictive learning analytics: a decade systematic review (2012–2022). Educ. Inf. Technol. 1–35 (2022)
30. Shin, D.: The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. Int. J. Hum. Comput. Stud. **146**, 102551 (2021)
31. Soussia, A.B., Labba, C., Roussanaly, A., Boyer, A.: Assess performance prediction systems: Beyond precision indicators. In: Proceedings of the 14th International Conference on Computer Supported Education, pp. 489–496 (2022)
32. Soussia, A.B., Treuillier, C., Roussanaly, A., Boyer, A.: Learning profiles to assess educational prediction systems. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds.) Artificial Intelligence in Education. AIED 2022. LNCS, vol. 13355, pp. 41–52. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-11644-5_4
33. Srinivasan, R., Chander, A.: Explanation perspectives from the cognitive sciences-a survey. In: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, pp. 4812–4818 (2021)
34. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning, pp. 3319–3328 (2017)
35. Swamy, V., Du, S., Marras, M., Kaser, T.: Trusting the explainers: teacher validation of explainable artificial intelligence for course design. In: LAK23: 13th International Learning Analytics and Knowledge Conference, pp. 345–356 (2023)
36. Swamy, V., Radmehr, B., Krco, N., Marras, M., Käser, T.: Evaluating the explainers: black-box explainable machine learning for student success prediction in MOOCS. In: Proceedings of the International Conference on Educational Data Mining (2022)
37. Vincent-Lancrin, S., van der Vlies, R.: Trustworthy artificial intelligence (AI) in education. OECD Educ. Work. Pap. **218** (2020)
38. Wang, C., Han, B., Patel, B., Rudin, C.: In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. J. Quant. Criminol. **39**, 519–581 (2023)