Saqr and López-Pernas
Smart Learning Environments (2024) 11:52
https://doi.org/10.1186/s40561-024-00343-4

Smart Learning Environments

RESEARCH Open Access

Why explainable AI may not be enough: predictions and mispredictions in decision making in education

Mohammed Saqr[1*] and Sonsoles López-Pernas[1]

*Correspondence:
mohammed.saqr@uef.fi

[1] School of Computing, University of Eastern Finland, Joensuu Campus Yliopistokatu 2, FI-80100 Joensuu, Finland

Abstract

In learning analytics and in education at large, AI explanations are always computed from aggregate data of all the students to offer the "average" picture. Whereas the average may work for most students, it does not reflect or capture the individual differences or the variability among students. Therefore, instance-level predictions—where explanations for each particular student are presented according to their own data—may help understand how and why predictions were estimated and how a student or teacher may act or make decisions. This study aims to examine the utility of individualized instance-level AI, its value in informing decision-making, and—more importantly—how they can be used to offer personalized feedback. Furthermore, the study examines mispredictions, their explanations and how they offer explanations or affect decision making. Using data from a full course with 126 students, five ML algorithms were implemented with explanatory mechanisms, compared and the best performing algorithm (Random Forest) was therefore selected. The results show that AI explanations, while useful, cannot achieve their full potential without a nuanced human involvement (i.e., hybrid human AI collaboration). Instance-level explainability may allow us to understand individual algorithmic decisions but may not very helpful for personalization or individualized support. In case of mispredictions, the explanations show that algorithms decide based on the "wrong predictors" which underscores the fact that a full data-driven approach cannot be fully trusted with generating plausible recommendations completely on its own and may require human assistance.

Introduction

The recent advances in artificial intelligence (AI) have been both revolutionary and transformative across several domains, and education was not an exception (Došilović et al., 2018). Several studies have indicated the potential of AI in providing personalized feedback, enhancing students' reflection, and offering individualized recommendations and personalized reports (Khosravi et al., 2022). As we currently stand, AI has been used in numerous applications and extended to all areas of teaching and learning (Ilkka, 2018). For instance, AI has been used to offer support for students' learning in academic writing, language, science and mathematics. AI has also been used to automate grading, provide real-time feedback, and analyze students' performance data to

guide instructional adjustments, making assessments more scalable and less demanding (Nagy & Molontay, 2023; Saqr et al., 2024). AI has also been used to personalize learning and adapt content to students' needs to offer a more inclusive learning environment (Khosravi et al., 2022). AI also has been used to offer support through chatbots, virtual assistants and AI tutors. Indeed, the scale of AI applications is widening to most areas of education technology and research and practice as well as in analyzing students' data, predicting performance and forecasting outcomes (Jang et al., 2022; Khosravi et al., 2022). While initially encouraging, several concerns have been raised regarding fairness, bias and transparency (Bernard & Balog, 2023). As such, an increasing number of studies are emerging using explainable AI techniques that explicitly show which factors are involved in the prediction and their relative importance (Adnan et al., 2022; Nagy & Molontay, 2023; Saqr et al., 2024). In doing so, learners and teachers are offered the opportunity to understand the "behind the scenes" process of predictions and more importantly, use this information as a ground for feedback for students (Khosravi et al., 2022). For instance, in a predictive model designed to predict students' final course grades, the algorithm will show the variables that have contributed to the positive outcome, e.g., regular course work, practicing formative assessment, or reading course materials (Khosravi et al., 2022); teachers can use this information to guide students and make future decisions regarding course design. Nevertheless, AI explanations are always offered from the aggregate data of all the students (i.e., at the dataset level) and therefore, offer the "average" overall picture (Saqr et al., 2024). Whereas the average gives an idea about what works for most students (i.e., feature importance), it does not reflect or capture the individual differences or the variability among students (Bobrowicz et al., 2024; Malmberg et al., 2022). Put another way, aggregate predictions are accurate in general but not accurate in any particular case. Therefore, instance-level predictions—where explanations for each particular student are presented according to their own data—can help understand how and why predictions were estimated and how a student or teacher may act (Jang et al., 2022; Nagy & Molontay, 2023). This study aims to examine the utility of individualized instance-level predictions, their value in informing decision-making, and—more importantly—how they can be used to offer feedback. Furthermore, the study examines mispredictions, their explanations, and how they may—or may not—affect decision-making.

Background

Prediction in education

Predicting students' performance has been a goal for educational researchers for several decades. The premise is that, if we can identify students who face problems or are lagging behind, we can initiate a proactive intervention when it matters (Saqr et al., 2022). At the turn of the twenty-first century, the widespread digitization of education and the emergence of the learning analytics field stimulated a remarkable surge in forecasting students' future performance (McCalla, 2023). The interest was further motivated by the rapid progress in AI—and more specifically machine learning (ML)—and the successful applications in the industry (Ilkka, 2018). The initial application of ML in learning analytics had a focus on optimizing accuracies and model performance metrics and indeed several studies have demonstrated remarkable performances (Adnan et al., 2022;

Jang et al., 2022; Nagy & Molontay, 2023). Nonetheless, accurate predictive models are rather problematic if they are an opaque "black box" with no insights into why or how they have produced such predictions (Khosravi et al., 2022). Further, for stakeholders to entrust a system, they have to understand how the decisions are made and most importantly, these decisions need to be ethical, fair, transparent, and accountable (Bernard & Balog, 2023). Therefore, attention was diverted to the alignment of ML with learning theories and the value of predictions in offering teachers and learners alike with necessary tools for feedback through transparent explainable algorithms.

Misprediction

Misprediction occurs when algorithms inaccurately predict student performance, behavior, needs or misclassify students in the wrong category (Biecek, 2018). Such flawed algorithmic predictions would result in misguided instructional decisions, needless interventions or wrong recommendations to mention a few (Baker & Hawn, 2022; Barredo Arrieta et al., 2020; Kordzadeh & Ghasemaghaei, 2022). Several factors contribute to misprediction that include poor data quality, biased datasets, and limitations in modeling techniques (Baker & Hawn, 2022). Also, some algorithms may not be able to capture the complex nature of educational data leading to inaccuracies (Barredo Arrieta et al., 2020; Biecek, 2018). As the use of algorithms increases in educational technologies, it is only expected that mispredictions will be everywhere. Even when algorithms improve or accuracy increases, they will not be perfect, and some students may pay the prices (Baker & Hawn, 2022).

Therefore, identifying mispredictions, and understanding why mispredictions happen would allow researchers to improve fairness of data-driven decision-making which would eventually results in a more fair and personalized learning process (Khosravi et al., 2022). This would also result in refining instructional strategies to ensure that interventions are more effectively targeted to students' diverse needs and individual differences (Ilkka, 2018). In doing so, we can avoid unnecessary support, missed opportunities for help or applying the wrong type of support or recommendations. Furthermore, students' trust and acceptance of AI would increase, their reliance on technology will also increase resulting in more adoption as well as possible better outcomes (Barredo Arrieta et al., 2020; Kordzadeh & Ghasemaghaei, 2022).

When we reduce misprediction, we essentially reduce bias and inequities by ensuring that predictive tools work well across diverse student populations (Barredo Arrieta et al., 2020). Furthermore, identifying mispredictions promotes a more balanced integration of AI into education, where algorithms are used not as definitive solutions but as tools to complement teachers' expertise. All the more so, identifying the sources of bias, and the students who are likely to be impacted by bias (Barredo Arrieta et al., 2020; Kordzadeh & Ghasemaghaei, 2022).

Explainable AI

A wide consensus in the AI community is that the higher the accuracy of an AI model (e.g., deep learning), the more likely the model is to be opaque and—consequently—the less likely it is to be interpretable, entrusted, or justifiable. Yet, it is no longer an option to trade accuracy for interpretability (Bernard & Balog, 2023). Therefore, researchers

resorted to either using transparent explainable models (the model is understandable by itself) or using mechanisms to augment models with explanations (Biecek & Burzykowski, 2021). White-boxing aims to extract information from the model, such as variable importance, visualizing the contribution of variables to aid the interpretability of the model (Khosravi et al., 2022). Offering interpretability can be done at the whole dataset level, referred to as global explainability (e.g., course, school, or cohort) level; or at the single instance level, referred to as local explainability—or instance-level explainability—(e.g., a student) (Biecek, 2018). Besides offering information about how and why AI decisions were made, instance-level explainability allows us to know which variables may affect the decisions if acted upon. For instance, what happens if the students get more engaged in coursework? And to what extent might that affect their grades? More importantly, an instance-level explanation offers a clue as to why mispredictions happened, be it an optimistic overestimate or a biased underestimate (Biecek & Burzykowski, 2021).

Explainable AI in education

Examples of explainable AI in the literature are emerging with a focus on predictions at the course level using inherently explainable algorithms, such as decision trees, and offering global explanations (Khosravi et al., 2022). Nevertheless, research with a focus on local interpretability is rather rare in the literature despite the important implications and benefits it could offer. A few examples can be seen here in the work by Jang et al. (2022), who studied data from seven Korean courses and visualized two cases where students were classified at-risk using SHapley Additive exPlanation. The visualization showed low engagement with homework was the most important variable in both cases. Similar example cases using the SHapley algorithm were visualized by Nagy and Molontay (2023) to explain student dropout. Some other examples using local explanations used subsets of the dataset, e.g., students who improved vs. not (Lin et al., 2023), or students who failed (Adnan et al., 2022). To that end, it is clear that the potential of single-instance explanation has been barely harnessed and only regarding examples of explainability. Yet, explaining misprediction and more importantly, how algorithms could offer a clue of what to improve to help students attain their desired outcome has not been explored.

The current study

Given the increasing reliance on AI, there is a pressing need for humanizing AI to be fair, transparent, and explainable and, in particular, a need to understand when a prediction or decision made by AI is correctly or incorrectly made and why. If an AI explanation is provided, it is necessary to know how it is justified, or if it is reasonable enough to be acted upon by a human, for instance, in an intervention. In this study, we use the SHapley Additive exPlanation (SHAP) method, a method based on game theory principles, to examine the correctly and incorrectly predicted cases (i.e., to offer local explanations for such cases) and to understand how and why prediction and misprediction happened (Lin et al., 2023; Nagy & Molontay, 2023). SHAP works by averaging the contribution of variables over several possible orderings to compute the "additive contribution" of the given variables and minimize the possible role of interactions (Strumbelj & Kononenko,

2010). We further use *ceteris-paribus* plots (a type of partial dependence plots) to determine the influence of individual variables on the outcome. *Ceteris-paribus* plots estimate what happens when a variable changes while all others are held constant (which is *ceteris paribus* in Latin). We use *ceteris-paribus* plots to examine if and to what extent AI-generated recommendations work in both correctly and incorrectly predicted cases (Biecek, 2018).

Motivation of the study

While explainable AI has been increasingly adopted to understand and predict student performance, there is a significant gap in investigating the limitations thereof and why these limitations take place. Attention has always been directed to what algorithms can achieve, not to when and why they un-achieve (e.g., mispredict). We have limited knowledge of the products of mispredictions, the decisions that may be based on them, and how they may affect educational practice. Even in the few instances in which mispredictions are studied, they are studied on the aggregate levels (across the whole dataset level) overlooking the instance or individualized impact of mispredictions. This study aims to address these gaps by examining the utility of individualized instance-level predictions and exploring how mispredictions may impact decision-making and feedback mechanisms. To that end, this study aims to answer the following research questions:

RQ1: Using explainable AI, to what extent can we predict students' performance? And to what extent can the AI recommendations be used as feedback?

RQ2: In cases where AI mispredicts students' performance, what are the factors behind mispredictions and how useful are AI recommendations in cases of misprediction?

Methods

Context

The dataset in this study comes from a course that teaches the subject of growth and human development to first-year medical students. The course teaches basic medical science subjects (anatomy, histology, physiology, and pathology) related to human growth (childhood, adulthood, reproduction, etc.). To integrate these subjects together, the program uses problem-based learning (PBL) as the pedagogical approach. In PBL, students are given weekly clinical scenarios (referred to as problems) that, for instance, tell a patient's story of a child growing up with references to different aspects of child development and related issues. The problems are ill-structured by design, requiring students to go through a series of steps of reading, brainstorming, discussing solutions, and reflecting on the process. Most of the process occurs online, except for an introductory meeting where the students read the problem together and discuss vocabulary and objectives, and an end-of-week meeting where they reflect on their solutions and the process. Whereas the program is a blended learning program, the Learning Management System (LMS) is the main platform for conducting online PBL, delivery of the lectures, support as well as formative assessment.

Data collection

The data collection was operationalized following the literature on online engagement which entails collection of traces of students' online learning. In particular, the

Interactive, Constructive, Active, and Passive (ICAP) framework offers an intuitive conceptual model for explaining learning outcomes using different engagement levels (Chi & Wylie, 2014). In that, deeper levels of engagement are associated with better learning and so it is expected that learning increases according to engagement from "passive to active to constructive to interactive, their learning will increase" (Chi & Wylie, 2014). Therefore, actions were coded to make the logs more meaningful and interpretable and to combine similar activities (e.g., "open quiz", "navigate quiz", and "attempt quiz" were all coded as "evaluate").

The fully coded activities and their meaning are explained in Table 1. Codes such as "construct" can be thought of as belonging to the interactive constructive end of the spectrum; "formative", "number of sessions", and "active days" belong to the active category, and, lastly, "evaluate", "course view", "read PBL" and "learning resources" belong to the passive category. Further, while most of the engagement indicators are counts of the number of times each action was performed in the LMS, the "session counts", the "active days", and the "duration" of the total time working online have been computed based on the timestamps of the logged actions. The course outcome was operationalized as the final grade of the course exam which tests what students have learned through a well-balanced multiple-choice examination.

Data processing

The course was divided into two equal periods and data up to the mid-course was used for the prediction task given the aim of early prediction where proactive action is possible.

The following variables were used as predictors: Course view, Construct, Read PBL, Formative, Evaluate, Learning resources, Duration, Session count, Active days,

Table 1 Description of the Moodle LMS actions available in the trace data and computed engagement indicators

Frequency of actions	Description
Course view	Number of times a student views the course main page, which contains links to learning resources, discussions, and all other learning activities, as well as announcements and updates.
Evaluate	Number of times a student interacts with the formative quizzes.
Instructions	Number of times a student accesses the course instructions, guides or course booklet
Learning resources	Number of times a student accesses to all learning materials e.g., lectures, learning links, videos, pages, or videos.
Construct	Number of times a student composes a PBL post .
Read PBL	Number of times a student reads the PBL posts by other students.
Socialize	Number of times a student interacts in non-learning forums e.g., talks about life events.
Support	Number of times a student poses queries in the support discussions or reads posts posed by their classmates.
Computed engagement indicators	
Duration	The total time of time spent on online learning activities.
Session count	The number of sessions a student makes while learning (a session is an interrupted time spent on learning activities).
Active days	The number of days a student has accessed using learning activities.
Formative	Grades in formative assessment

Instructions, Support, and Socialize. Table 2 shows the descriptive statistics of the LMS actions and computed engagement indicators. The data was partitioned into training (0.7) and testing sets (0.3).

Data analysis

ML algorithms

Five common ML algorithms were used for the task of predicting student performance:

(1) *Random Forest*: Random Forest (RF) is a type of ensemble algorithm. This means that, instead of using a single model (like one decision tree), it computes a large number of decision trees and averages the prediction of individual trees as the final result. This process makes it less sensitive to noisy data, and less prone to overfitting (Hastie et al., 2009). RF can be used both for classification and for regression (such as in our case). There are many instances of their use in education research for predicting students' grades (Alamri et al., 2020; Nachouki et al., 2023). We implemented RF in our analysis through the R package *randomForest* (Liaw & Wiener, 2002).

(2) *XGBoost*: eXtreme Gradient Boosting—or most commonly XGBoost—is a gradient boosting algorithm that builds several decision trees to form an ensemble where each tree builds on the results of the former trees and the final results are computed by combining all tree predictions (Chen & Guestrin, 2016). The main difference with RF is that XGBoost builds trees sequentially, with each new tree trying to correct the errors made by the previous trees. XGBoost has started to be used in educational research in the last few years, mostly to predict student performance (Asselman et al., 2021; Yan, 2021). We have implemented XGBoost in our analysis through the R package *xgboost* (Chen et al., 2024).

(3) *Neural networks*: Neural networks are a class of models inspired by the structure of the human brain. In its simplest form, a neural network consists of input nodes, which correspond to the features in your data, and output nodes, which produce the final prediction. These nodes are connected through weights that represent the

Table 2 Descriptive statistics of the predictors

Predictor	M	SD
Course view	37.28	28.11
Evaluate	7.69	7.04
Instructions	8.05	5.16
Learning resources	15.07	8.42
Construct	13.15	13.21
Read PBL	87.05	74.72
Socialize	2.98	3.56
Support	1.23	2.55
Duration	21,665.71	21,328.22
Session count	29.26	16.49
Active days	11.75	3.36
Formative	5.18	5.31

strength of connections between them. Neural networks have been operationalized in education research to evaluate team collaboration (Barmaki & Guo, 2020) and, of course, to predict student performance (Thomas & Ali, 2020). We have implemented XGBoost in our analysis through the R package *parsnip* (Kuhn & Vaughan, 2023).

(4) *Linear regression*: Linear regression is one of the most basic techniques for statistical modeling and ML. It works by finding a line of best fit that predicts the relationship between a dependent variable and one or more independent variables (features). Linear regression has been used in educational research in many ways, but mostly for performance prediction, just like the other methods (Saqr et al., 2017). In this analysis, we have used the base R implementation of linear regression through the *lm* function.

(5) *Support Vector Machine***.** Support Vector Machine (SVM), which works similarly linear regression to find a function that represents the association between the predictors and the predicted outcome, while minimizing errors (Biecek & Burzykowski, 2021). The difference with linear regression is that SVM tries to fit the best line within a margin of tolerance around the predicted values, focusing on keeping the prediction errors within a defined range. The use of SVM in education research in general, and in performance prediction specifically, has been widespread (Alamri et al., 2020). In our analysis, we implemented SVM through the *e1071* R package (Meyer et al., 2023).

Explainable AI

The training data was used to fit an ML model using each of the algorithms described in the previous subsection. The performance of the five algorithms was computed using the test data (30% of the original dataset) and compared using the following performance measures: Mean Squared Error (**MSE**) and Mean Absolute Deviation (**MAD**), Root Mean Squared Error (**RMSE**), and the Coefficient of Determination (**R-squared** or R^2). Further, the residuals of the five algorithms were plotted and compared to evaluate the performance of the algorithms and select the best-performing one.

Model explainability was carried out using the *DALEX* package (Biecek, 2018), which provides diagnostic tools to explore and explain the models. Variable importance was estimated using the RMSE loss function which quantifies the magnitude of loss of RMSE if the target variable was removed. When important variables are removed, model performance worsens. The process was repeated thousands of times to quantify the average variable importance over multiple permutations (Biecek, 2018).

Local explainability was estimated using the Shapley additive values. Given that SHAP can be affected by the order, the predictors were ordered according to the ICAP engagement framework (i.e., from constructive to passive). Further, we permuted the values 1000 times to compute the average contribution and eliminate the possible ordering problem (Biecek & Burzykowski, 2021).

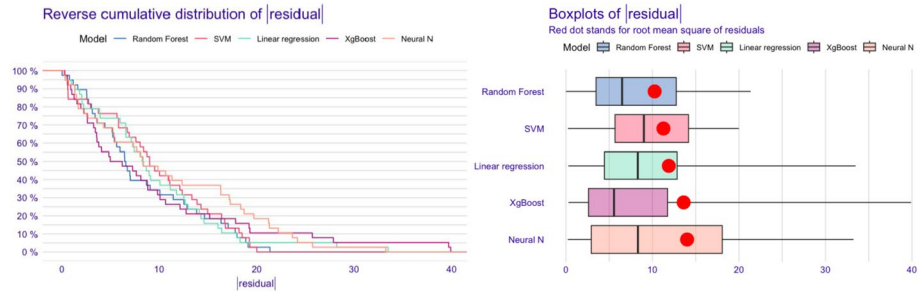

Fig. 1 Residual analysis for the five algorithms

Table 3 Performance comparison of the five algorithms

Metric	Random forest	Linear regression	XGBoost	Neural network	SVM
MSE	**117.54**	155.52	169.54	149.02	127.89
RMSE	**10.84**	12.47	13.02	12.21	11.31
R2	**0.3**	-0.06	-0.16	-0.02	0.13
MAD	**6.65**	8.65	8.06	7.98	10.22

Results

RQ1: general model results

As a first step, the five models were compared based on performance to select the best-performing model. The XGBoost model had the most samples with the lowest levels of residuals, but a small share of samples had very high levels. The Random Forest model had the most consistent low levels of residuals. Figure 1 (right) confirms these findings by showing the box plot of the residual distributions for each model. The Random Forest algorithm shows the lowest mean residual (RMSE), followed by SVM, linear regression, and XGBoost (though XGboost had the lowest median), whereas the single-layer neural network shows the highest mean residual. Figure 1 shows the results of the residual analysis for each of the five algorithms. Figure 1 (left) shows the reverse cumulative distribution plot where it can be seen that a large number of samples in the neural network model had large residuals.

Regarding accuracy measures (Table 3), Random Forest (in bold) was the best performing algorithm, with the lowest MSE, RMSE, and MAD, and the highest—by far—R-squared and therefore was selected for the analysis.

Explaining predictions using RF

Regarding variable importance, examining the loss of RMSE (using 1000 permutations with boxplots), the RF model showed that *formative* grades, the number of posts (i.e., *construct*), and viewing the *instructions* as well as *learning resources* were the most important predictors. Other predictors such as using *support* forums, *socializing*, number of *active days*, *duration* of online work, and using the quiz module (i.e., *evaluate*) were far less important. The Shapley values in Fig. 2 show each of the predictors' contributions to each instance (student in the testing dataset). Obviously,

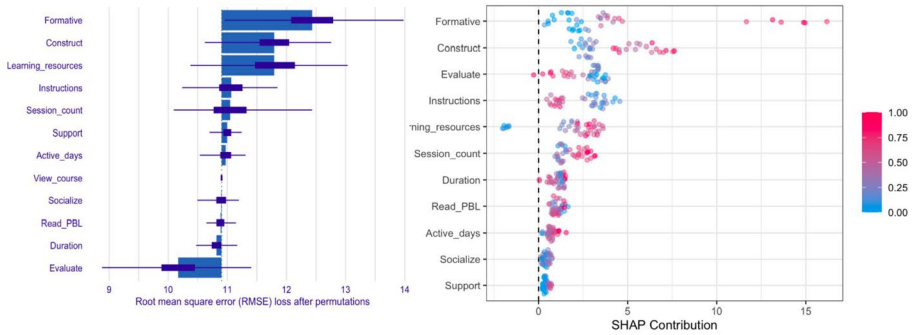

Fig. 2 Distribution of RMSE and of SHAP contribution for each engagement indicator across students

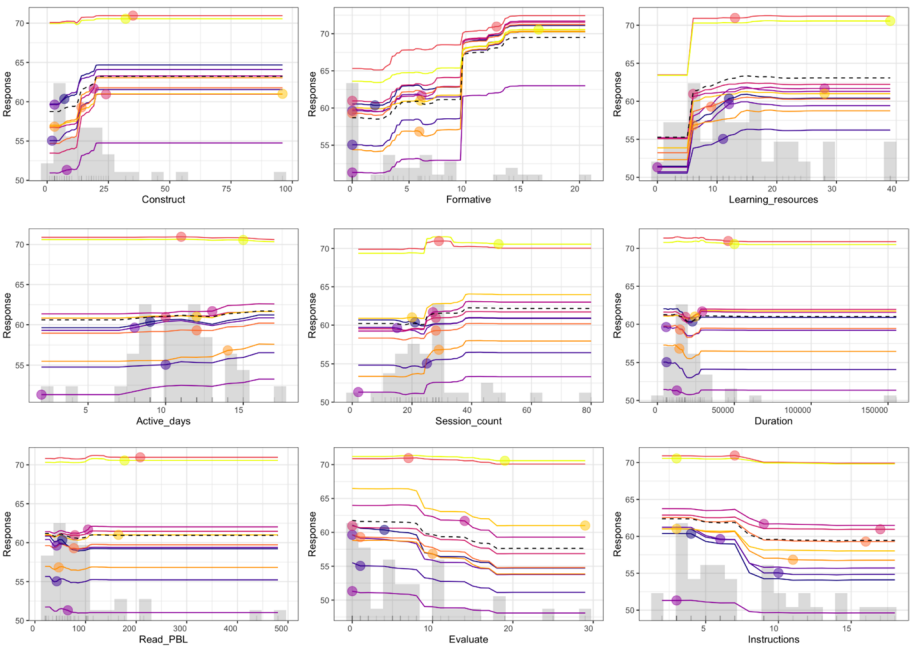

Fig. 3 Partial dependence plots, each line represents percentile, the x-axis represents the number of activities, the y-axis represents the expected grade. The gray histogram on each plot represents the actual distribution of values

there are wide variations between the students in some variables. For example, formative assessment shows a wide variance: on one end, few students scored high in their formative assessment quizzes, and, therefore, *formative* assessment as a predictor contributed significantly to their predictions. Similarly, albeit with lower variance, the *construct* predictor varied moderately between students.

Acting on explanations

More important than explaining the reasons behind a certain prediction is to show which variables could be manipulated (changed) to improve the chance of a positive prediction (success in our case). Partial dependence plots (or *ceteris-paribus* plots) of each variable are plotted in Fig. 3. Partial dependence plots show what happens if a given variable is changed (i.e., increased) while all other variables are held constant. Put another

way, in the hypothetical scenario that a student increases learning activity, would the probability of a better grade improve? If so, to what extent? As Fig. 3 shows, increasing the *construct* variable could potentially increase students' grades: the increase is steep and gradual till around 20 posts and then a plateau is reached. Likewise, an increase in grades is associated with an increase in *formative* assessment and *learning resources*. Similarly, though with lower magnitude, *active days* and *session count* variables show similar very gradual curves. Duration and reading PBL show flat curves, whereas *evaluate*, and *instructions* show increasing these variables may worsen grades, or rather distract students from the more meaningful activities.

RQ2: explaining misprediction

Whereas the average predictions were useful, they—as any other predictive models—are far from perfect. In that, an algorithm always generates accurate and inaccurate predictions, the former is well-studied and the latter is barely studied. All the more so, we are interested in understanding why and how students were mispredicted and if the *ceteris-paribus* plots give useful information about what is recommended for improvement and, in the case where the students were mispredicted, what kind of recommendations are given.

We start here with students ($n=6$) who were mistakenly predicted to have higher grades than they ended up having in reality. We use a threshold of more than 10% difference between the predicted and observed grades. We chose the 10% threshold because in our context, we use a letter grading system, a 10% change means completely changing the student grade one full letter grade from A to B for example, or C to D, or even D to F which results in significant consequences regarding graduation and later training opportunities. The predicted grade *mean* was 60.5, SD = 5.1, and the actual grade *mean* = 44.8, SD = 6.7. The waterfall plot (Fig. 4) on the left side shows that these students typically attained positive prediction based on their engagement indicators that were mostly *passive* according to the ICAP framework (e.g., viewing learning resources, reading PBL, reading instructions, and duration). Typically, these students did worse on *formative* or

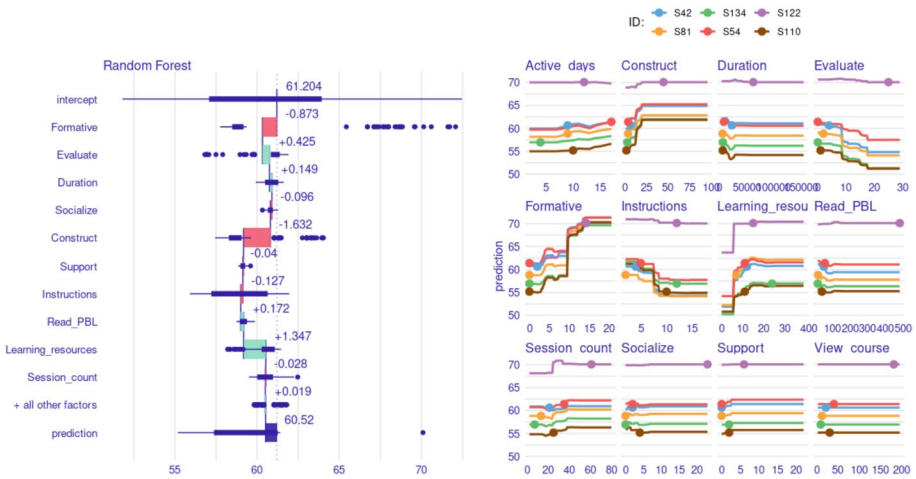

Fig. 4 Waterfall plot (left) and *ceteris-paribus* plot (right) of the students whose grade was mispredicted as higher than it actually was

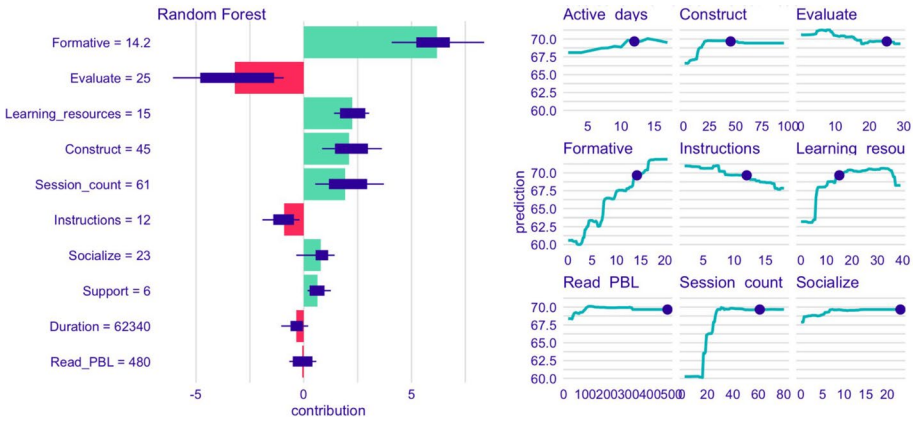

Fig. 5 Waterfall plot (left) and ceteris paribus plot (right) of student S122 who obtained a grade of 57 whereas the prediction by Random Forest was 69.66

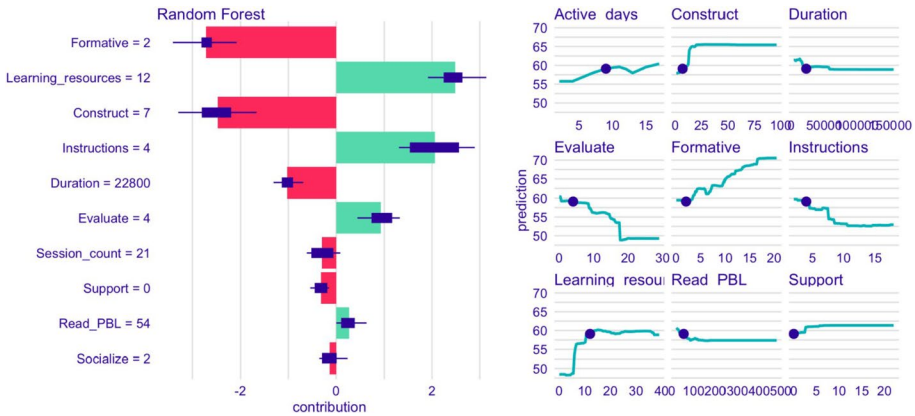

Fig. 6 Waterfall plot (left) and ceteris paribus plot (right) of student S42 who obtained a grade of 39 whereas the prediction by Random Forest was 59.1

construct activities which are closer to the positive side of the spectrum. Interestingly, they were also predicted higher for navigating the quiz (*evaluate*) which does not necessarily imply taking it. The *ceteris-paribus* plot (Fig. 4 - right) shows each student and their expected change in grades when increasing their activities. In general, the trends are similar to the overall model, however, individually, some curves are flat and thus not useful for some students, e.g., S122 (more detailed in Fig. 5) whose predicted grade was 69.7 where the actual grade was 57. In that sense, recommendations based on explainable interpretable AI can be inadequate. In other cases, such as S42—who was predicted 59 where the actual grade was 39—, the suggestions were more aligned with the general model (Fig. 6).

Mispredicted lower

Students who had their grades predicted pessimistically lower than their observed grades (10% difference) were grouped together in this category. The predicted grade mean was 63.75, SD = 4.85 and the actual grade was 79.43, SD = 6.45, i.e., these students scored in the highest range. According to Shapley values, the predicted grades decreased

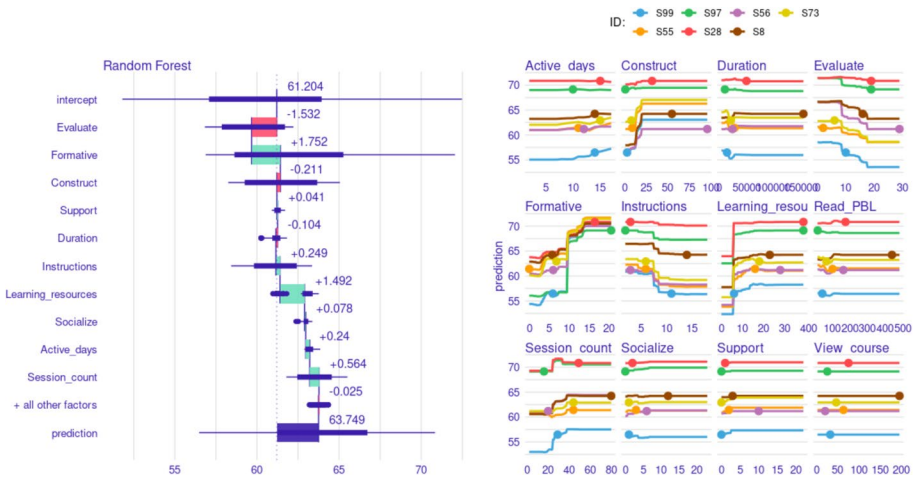

Fig. 7 Waterfall plot (left) and ceteris paribus plot (right) of the students whose grade was mispredicted as lower than it actually was

because of fewer interactions with the quiz modules (i.e., evaluate) although the mispredicted formative assessment grades were the highest positive variable that contributed to the grades. Furthermore, grades were predicted lower partially due to lower posting activity (i.e., construct). Other activities that increased the predicted grades include the learning resources, the instruction, and the session count. In other words, the students in this group did not conform to the general picture and therefore, were offered wrong underestimates of their grades. The *ceteris-paribus* plots show mostly flat lines (notice the points of the students' position). In that sense, one can observe that students who were mispredicted around the average may not benefit at all from AI-based suggestions (Fig. 7).

Discussion

The quest to harness the power of AI has been accelerating over the past two decades. Yet, haste led to a troubled path paved with costly consequences in terms of human suffering, bias, and wasted resources (Kordzadeh & Ghasemaghaei, 2022). Thereupon, the development of interpretable, fair and just ML was born out of the necessity of ethical and transparent AI (Baker & Hawn, 2022; Bernard & Balog, 2023). Our study is a step in the interpretability direction. We aimed to explore if and to what extent explainable AI models can help understand students' performance. In particular, we focus on instance-level predictions, mispredictions, and recommendations. This is because instance-level interpretability could help teachers offer individualized feedback, understand the contribution of variables to the predictive model, and more importantly, offer a look into mispredictions or why models have missed or made the wrong decision about a student or another.

Our findings have shown that the interpretability of AI models offered a transparent view of how the model worked and what variables matter for our course achievement prediction. Interpretability made it clear why students were predicted and what variables were taken into account to produce the results (Došilović et al., 2018; Roscher et al., 2020). Nonetheless, algorithmic predictions are made solely based on a data-driven

approach, and their connections with theory and pedagogy are not guaranteed. Of course, it is understood that, by design, ML learns from the data that reflects the behavior of the majority—and this majority defines what the algorithm considers important and that can result in both plausible and implausible conclusions (Barredo Arrieta et al., 2020). For instance, based on our results, one can tell students that the more they engage in highly cognitive tasks (formative assessment and constructing knowledge), the more they are likely to score higher grades. On the other hand, it does not seem plausible to suggest that reading less the course instructions, quizzes, or reading responses of their colleagues in the discussions will be detrimental to their achievement. Put another way, explaining the predictions is rather insufficient on its own and does not alleviate the need for a human (teacher or the student) who can interpret and make sense of the results (Peeters et al., 2021).

Explaining mispredictions has shown that algorithmic predictions may make the wrong decisions based on the "wrong predictors" (Chi & Wylie, 2014). In our case, students were predicted far higher than their actual grades based on their engagement with the "passive" side of engagement which is less cognitively demanding and in fact, less likely to lead to better outcomes according to existing research or theory (i.e., ICAP) (Chi & Wylie, 2014). By the same token, students who were mispredicted with lower grades than their actual observed ones were predicted so because they did not engage much (i.e., excessively read the quiz) with passive engagement indicators (i.e., navigating the quiz module without actually attempting them). These results, while helpful, cannot be taken as offering an "accurate" picture of what can be considered individualized support.

In comparison with previous research, our results show similarities and differences, which can be justified by the different context, models, and approach. In that, we studied medical education, using LMS data as well as focused on the local explanation and misprediction. The work by (Lin et al., 2023) concluded that using explainable AI may have shown the important variables, but it is yet to be verified if these important variables are actually useful. Indeed, our study shows that transparency, usefulness, and pedagogical value may not converge to the same thing. In other words, transparency and interpretability may be achieved without actual usefulness. Similarly, Jang et al. (2022) showed that individual explainability can be useful for some cases and rather less useful for others as a basis for intervention.

Taken together, our results show that AI explanations, while useful, are far from being practical without a nuanced human involvement (i.e., hybrid human-AI collaboration). This may be explained by the inherent deficiencies of online data that account for only the online part of learning. It can also be explained by the shortcomings and the limits of AI. Instance-level explainability may allow us to understand individual algorithmic decisions, but they are far from perfect for being useful in offering personalization or individualized support. Possibly, for individualizing or personalizing recommendations, an idiographic approach that entails single-subject analysis using adequate data may be a better solution (Bobrowicz et al., 2024; Saqr & López-Pernas, 2021). It is yet to be verified whether the inclusion of more data to capture the multidimensional nature of learning can help AI algorithms deliver a more nuanced understanding of learners' behavior. These results underscore the fact that a fully data-driven approach can't be fully trusted

with generating plausible recommendations completely on its own and may require human assistance.

Conclusions

Online data and AI algorithms have inherent deficiencies and imperfections that may fail in capturing the complexities of human learning or align with educational objectives and can even lead to implausible conclusions. As such, AI alone may not provide the practical recommendations or the individualized support that educators long for. While explainable AI offers insights into important variables affecting course achievement, it has demonstrated an obvious gap between educational theories and practice and algorithmic predictions. This disconnection underscores the need for a hybrid human-AI approach that combines contextual understanding, human expertise and technology. A more comprehensive approach that considers the multidimensional nature of learning and incorporates detailed individual analysis can offer a deeper understanding of learner behavior.

Authors' contributions

Mohammed Saqr: Conceptualization, Methodology, Software, Data Collection, Writing - Original Draft Preparation, Writing - Reviewing and Editing, Funding Acquisition, Supervision. Sonsoles López-Pernas: Methodology, Software, Data Collection, Writing - Original Draft Preparation, Writing - Reviewing and Editing, Supervision.

Funding

The paper is co-funded by the Academy of Finland for the project TOPEILA, Decision Number 350560 which was received by the first author.

Availability of data and material

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare no competing interests.

Received: 20 July 2024 Accepted: 24 October 2024
Published online: 18 November 2024

References

Alamri, L. H., Almuslim, R. S., Alotibi, M. S., Alkadi, D. K., Khan, U., I., & Aslam, N. (2020). Predicting student academic performance using support vector machine and random forest. In *Proceedings of the 2020 3rd international conference on education technology management*. ICETM 2020: 2020 3rd International conference on education technology management, London United Kingdom. https://doi.org/10.1145/3446590.3446607

Adnan, M., Irfan Uddin, M., Khan, E., Alharithi, F. S., Amin, S., & Alzahrani, A. A. (2022). Earliest possible global and local interpretation of students' performance in virtual learning environment by leveraging explainable AI. *IEEE Access : Practical Innovations, Open Solutions*, *10*, 129843–129864.

Asselman, A., Khaldi, M., & Aammou, S. (2021). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments, 31*, 1–20.

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *32*(4), 1052–1092.

Barmaki, R., & Guo, Z. (2020). Deep neural networks for collaborative learning analytics: Evaluating team collaborations using student gaze point prediction: Evaluating team collaborations using students' gaze point prediction. *Australasian Journal of Educational Technology*, *36*(6), 53–71.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *An International Journal on Information Fusion*, *58*, 82–115.

Bernard, N., & Balog, K. (2023). A systematic review of fairness, accountability, transparency and ethics in information retrieval. *ACM Comput Surv*. https://doi.org/10.1145/3637211

Biecek, P. (2018). Dalex: Explainers for complex predictive models in R. *Journal of Machine Learning Research: JMLR*, *19*, 1–5.

Biecek, P., & Burzykowski, T. (2021). *Explanatory model analysis: Explore, explain, and examine predictive models*. CRC.
Bobrowicz, K., López-Pernas, S., Teuber, Z., Saqr, M., & Greiff, S. (2024). Prospects in the field of learning and individual differences: Examining the past to forecast the future using bibliometrics. *Learning and Individual Differences*, *109*, 102399.
Chen, T., & Guestrin, C. (2016). XGBoost. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. KDD '16: The 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco California USA. https://doi.org/10.1145/2939672.2939785
Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., & Yuan, J. (2024). *xgboost: Extreme Gradient Boosting*. https://CRAN.R-project.org/package=xgboost.
Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243.
Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, 0210–0215.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). Random forests. *The elements of statistical learning* (pp. 587–604). Springer.
Ilkka, T. (2018). *The impact of artificial intelligence on learning, teaching, and education*. European Union.
Jang, Y., Choi, S., Jung, H., & Kim, H. (2022). Practical early prediction of students' performance using machine learning and eXplainable AI. *Education and Information Technologies*, *27*(9), 12855–12889.
Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence, 3*, 100074.
Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409.
Kuhn, M., & Vaughan, D. (2023). *parsnip:* A common API to modeling and analysis functions. https://CRAN.R-project.org/package=parsnip
Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.
Lin, J., Dai, W., Lim, L. A., Tsai, Y. S., Mello, R. F., Khosravi, H., Gasevic, D., & Chen, G. (2023). Learner-centred analytics of feedback content in higher education. In *LAK23: 13th international learning analytics and knowledge conference*, 100–110.
Malmberg, J., Saqr, M., Järvenoja, H., Haataja, E., Pijeira-Díaz, H. J., & Järvelä, S. (2022). Modeling the complex interplay between monitoring events for regulated learning with psychological networks. In M. Giannakos, D. Spikol, Di D. Mitri, K. Sharma, X. Ochoa, & R. Hammad (Eds.), *The Multimodal Learning Analytics Handbook* (pp. 79–104). Springer International Publishing.
McCalla, G. (2023). The history of artificial intelligence in education – the first quarter century. *Handbook of Artificial Intelligence in Education* (pp. 10–29). Edward Elgar Publishing.
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2023). e1071: misc functions of the department of statistics, probability theory group (formerly: e1071), tu wien. https://CRAN.R-project.org/package=e1071
Nachouki, M., Mohamed, E. A., Mehdi, R., & Abou Naaj, M. (2023). Student course grade prediction using the random forest algorithm: Analysis of predictors' importance. *Trends in Neuroscience and Education*, *33*, 100214.
Nagy, M., & Molontay, R. (2023). Interpretable dropout prediction: Towards XAI-Based personalized intervention. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-023-00331-8
Peeters, M. M. M., van Diggelen, J., van den Bosch, K., Bronkhorst, A., Neerincx, M. A., Schraagen, J. M., & Raaijmakers, S. (2021). Hybrid collective intelligence in a human–AI society. *AI & Society*, *36*(1), 217–238.
Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *Ieee Access : Practical Innovations, Open Solutions*, *8*, 42200–42216.
Saqr, M., Cheng, R., López-Pernas, S., & Beck, E. D. (2024). Idiographic artificial intelligence to explain students' self-regulation: Toward precision education. *Learning and Individual Differences, 114*, 102499.
Saqr, M., Fors, U., & Tedre, M. (2017). How learning analytics can early predict under-achieving students in a blended medical education course. *Medical Teacher*, *39*(7), 757–767.
Saqr, M., Jovanovic, J., Viberg, O., & Gašević, D. (2022). Is there order in the mess? A single paper meta-analysis approach to identification of predictors of success in learning analytics. *Studies in Higher Education, 47*, 1–22.
Saqr, M., & López-Pernas, S. (2021). Idiographic learning analytics: A definition and a case study. In *2021 International conference on advanced learning technologies (ICALT)*, 163–165.
Strumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. https://www.jmlr.org/papers/volume11/strumbelj10a/strumbelj10a.pdf?ref=https://githubhelp.com
Thomas, J. J., & Ali, A. M. (2020). Dispositional learning analytics structure integrated with recurrent neural networks in predicting students performance. *Advances in Intelligent systems and Computing* (pp. 446–456). Springer International Publishing.
Yan, K. (2021). Student performance prediction using XGBoost method from A macro perspective. In *2021 2nd international conference on computing and data science (CDS)*, pp. 453–459.

Publisher's note