Journal of Formative Design in Learning (2025) 9:101–113
https://doi.org/10.1007/s41686-025-00104-7

Early Detection of At-Risk Students in Physics Through Remediated Online Formative Assessments and Machine Learning

Charlie V. Sarmiento[1] · **Germano Maioli Penello**[1,2] · **Lucas Sigaud**[3]

Accepted: 30 April 2025 / Published online: 2 June 2025

Abstract
Traditional distance learning evaluations often fail to diagnose students' deficiencies in the early stages. This research tackles this challenge by implementing online formative assessments and analyzing the records of an introductory physics course, aiming to predict at-risk students. These online assessments go beyond mere evaluation, offering valuable benefits for students, such as immediate feedback, the opportunity to retake assessments, and learn from mistakes, all fostering deeper understanding. Three machine learning algorithms were used to predict students' final situation. All algorithms demonstrated good classification performance. However, the support vector machine (SVM) algorithm surpassed the others. This result allows us to predict potential failure during the very first formative assessment. This method empowers instructors to intervene early and improve student success, potentially leading to higher retention rates. These findings pave the way for personalized learning interventions in distance learning education. Potentially transforming students' outcomes and fostering a more engaging learning experience.

Keywords Teaching physics · Higher education · Online learning · Data mining · Learning strategies · Formative assessments

Introduction

Educational data mining constitutes a subset of education research that uses extensive data systems and machine learning algorithms to forecast student academic success and/or their progression towards graduation (Alam, 2023; Khan & Ghosh, 2021; Sghir et al., 2023). The origins of this field within education research date back to the early 1990s (Nandeshwar et al., 2011) and investigations within the domain of physics education research (PER) show the importance of the use of data mining techniques (Aiken et al., 2019; Hansen, 2023; Richards & Kelly, 2023; Yang et al., 2020; Zabriskie et al., 2019).

Physics courses are central in STEM education (science, technology, engineering, and mathematics). Recognizing the critical need for a more robust STEM workforce, the President's Council of Advisors on Science and Technology (PCAST) outlined strategies in 2012 to improve student retention in STEM fields (Olson & Riordan, 2012). This initiative stemmed from the goal of adding one million more STEM professionals to the United States workforce within a decade. However, there is a significant challenge: less than 40% of students entering STEM fields actually graduate with a STEM degree. Increasing retention rates to just above 50% would be instrumental in achieving the objective. The issue of STEM student retention is not new, and numerous studies have explored this persistent challenge (Hall et al., 2015; Sithole et al., 2017; Winberg et al., 2019).

An integral component of course design is the evaluation method (Richlin, 2023; Sudakova et al., 2022), which constitutes a cornerstone in enhancing and optimizing retention rates. Assessments are usually distinguished as formative and summative, which are better understood as assessments for learning and of learning, respectively. Formative assessments (FA) can provide educational outcomes by providing

Charlie V. Sarmiento is now participating of the Programa de Capacitação Institucional (PCI) at the Centro Brasileiro de Pesquisas Físicas, Rio de Janeiro, Brazil.

✉ Charlie V. Sarmiento
charliesarmiento@cbpf.br

1 Universidade Federal do Rio de Janeiro, Instituto de Física, Rio de Janeiro, Brazil

2 Universidade de São Paulo, Instituto de Física, São Paulo, Brazil

3 Universidade Federal Fluminense, Instituto de Física, Niterói, Rio de Janeiro, Brazil

evidence on student needs (Wiliam, 2011). Several investigations show the importance of well-structured or creatively thought about formative assessments to engage both teachers and learners in a better learning process (Bennett, 2011; Gikandi et al., 2011; Villarroel et al., 2018). In 2016, Wylie and Lion showed that formative assessments enable the adjustment of instructional strategies and the achievement of academic objectives through feedback from peers and teachers (Wylie & Lyon, 2016). However, the proposed method requires training for peers and teachers to ensure its effectiveness, which complicates its implementation. With technological advancements, these formative assessments can be designed, implemented, and corrected online, providing students with immediate feedback (Bulut et al., 2019, 2023). Moreover, technology-supported formative assessment generates valuable data that can be leveraged for student monitoring and/or to assess the mechanistic reasoning using machine learning algorithms (Martin & Graulich, 2023).

This study aims to identify and provide guidance to students who are at risk of academic failure. To achieve this goal, we implemented a formative assessment within the Moodle platform as part of an undergraduate introductory physics course, and the data collected was utilized to train machine learning algorithms for predicting and classifying students' final situations. Initially, this assessment was offered as an optional test for students in 2019, but has now become an integral part of the assessment methodology. The purpose of this FA is to empower students to evaluate their own learning process and identify areas where they may need improvement before taking proctored exams (PE). In doing so, we can closely monitor students' academic performance, identify those who require additional support, and facilitate direct communication with tutors.

In a nutshell, this study addresses the following research questions:

RQ-1. How can we support the learning of students within our logistic limitations?

RQ-2. How do institutional data and students' performance in the new FA predict the final situation in an introductory course of physics?

Background—Related Literature

In this section, we introduce the theoretical concepts underlying the learning-by-doing method and the resubmission process, both of which are essential for categorizing the implemented formative assessment. Furthermore, we review relevant literature to highlight their significance. Lastly, we provide a brief discussion of existing research on the application of machine learning algorithms to predict the final performance of students in physics courses.

Theoretical Background

This study used the learning-by-doing (LBD) educational approach. John Dewey (1859–1952), an American philosopher, is the most influential figure in experiential education (Williams, 2017). Unlike traditional memorization, LBD emphasizes student involvement in projects and tasks to solidify understanding. As Dewey believed, true learning is based on experience.

This approach encompasses various learning processes based on practical activities and experimentation, both mental and physical. The benefits of implementing LBD in STEM education have been documented in several studies (Das Dores et al., 2023; Nantsou et al., 2021, 2024a, b; Nantsou & Tombras, 2022; Niiranen & Rissanen, 2017) and other educational fields (Effie Steriopoulos & Harkison, 2022; Molly George et al., 2015; Sangpikul, 2020). These advantages include deeper understanding, improved problem-solving skills, increased motivation, and better information retention.

In the specific case of distance learning education, online learners often lack face-to-face interaction that can be beneficial for learning. For that reason, the implementation of LBD activities in online learning might be particularly crucial. Several studies in online education highlight the importance of Learning-by-Doing in developing problem-solving skills (Hettiarachchi, 2013; Rossano et al., 2020; Sim & Lau, 2018).

Resubmission as a Remediation Process

The goal of remediation is to support the learning process of students by identifying gaps in their knowledge and helping them reach the expected level (Boylan & Saxon, 1999). Resubmission provides an opportunity to submit a revised version of an activity for evaluation, aligning with the learning-by-doing approach. Several studies emphasize the importance of resubmissions in the online evaluation process. Ardid et al. (2015) used the idea of resubmission as a form of continuous assessment in their online course evaluations. They observed that these new assessments had the potential to distinguish between students who might fail or pass the course.

Pinchbeck and Heaney (2017) permitted only one resubmission in their evaluations, along with two interventions: (1) an online synchronous tutorial session and (2) a supporting asynchronous forum. They found that these intervention measures improved the quality and quantity of resubmissions. Howard et al. (2019) allowed students to resubmit their quizzes along with explanations for errors, with the poten-

tial to earn an additional mark, as a remediation proposal. They observed that this proposed remediation had a positive impact on the final grades of the students, particularly among those who had initially achieved higher grades.

Use of Machines Learning Methods to Predict Outcomes in Physics

Machine learning algorithms are increasingly used to predict the final performance of students in physics. Zabriskie et al. (2019) developed various models for forecasting using logistic regression and random forest algorithms. These models were constructed for two courses: Introduction Calculus-based to Mechanics and Electricity and Magnetism. They trained the algorithms using institutional data and grades from homework assignments and exams. Notably, they found the logistic regression model can accurately predict students who are likely to earn grades lower than B, with accuracy rates exceeding 70%, as early as the first week of the course. Among the used data, homework grades were identified as the most crucial factors in model development, while demographic variables, gender and race were deemed less significant for algorithm training.

Yang et al. (2020) extended the earlier research by building predictive models. In their study, they used the random forest-supervised machine learning algorithm. They incorporated data from the Introductory Mechanics course offered at two institutions with distinct demographic profiles. Their focus was on developing models capable of predicting students earning grades or withdrawing, a task that had not been achieved in previous research. They recognized the importance of selecting appropriate metrics to address unbalanced data to achieve higher accuracy percentages. They also emphasized that certain data, such as underrepresented minorities, first-generation college status, and low socioeconomic status, did not significantly contribute to the construction of predictive models.

Richards and Kelly (2023) examined how academic coursework, performance, and demographic characteristics of community college students enrolled in astronomy courses influence their performance. They employed logistic regression to identify factors that predict students' grades in astronomy. Bahr's deconstructive approach served as their theoretical framework (Bahr, 2013). This framework emphasizes understanding the educational pathways of students within institutions and how these pathways impact their progress towards academic goals. Richards and Kelly (2023) found that difficulties in mathematics, potentially indicated by failing or retaking mathematical courses, were associated with graduation delays, lack of interest in STEM disciplines, and STEM attrition. Notably, they observed that demographic variables such as ethnicity, socioeconomic status, gender, and age were not significant predictors.

Methodology

Institutional Context

Introduction to Physical Sciences I (ICF1), offered by CEDERJ (Distance Education Center of the State of Rio de Janeiro), is the title of the course in which this work was developed. The course is coordinated by the Physics Institute of the Federal University of Rio de Janeiro and has the particularity of being present in six different degrees (Physics, Chemistry, Mathematics, Biological Sciences, and both Meteorological and Production Engineering), with students present in 23 different cities from the Rio de Janeiro State. This course is typically supported by two professors, 7 distance tutors (DT), and more than 40 face-to-face tutors (FT). Throughout the semester, only FTs have direct contact with students, providing assistance with physics experiments that are done on pre-schedule days. Professors and DTs have indirect contact and primarily respond to inquiries regarding course material.

In ICF1, exams generally adhere to the CEDERJ standard assessment process: home exams (HE) and proctored exams in person (PE) account for 20% and 80% of the final grade, respectively. Before COVID-19 pandemic, these home exams consisted of two types of assessment: two optional assessments (OA) and two distance assessments (DA). Both types were considered formative assessments (FA) and, in 2018, we recognized the need for improvement. Consequently, in 2019, we improved the OA, which received positive feedback from students. As a result, we decided to implement mandatory FAs for every topic of the course, including introduction,[1] optics, vectors, kinematics, forces, and astronomy. The primary objective of these FAs is to encourage practice, through exercises, focusing on fundamental content of each topic without penalizing mistakes.

Formative Assessments (FAs)

The new assessment templates were implemented on the CEDERJ online platform (which uses Moodle language), as follows:

- Students perform the FA on the online platform by logging in to their profiles.
- The FA consists of an average of twenty conceptual questions or practical exercises that directly involve some fundamental concepts of the subject.
- All responses must be provided in a written format inside an input text box. The number of significant figures or

[1] The topic "introduction" here refers to the expected prior content of basic mathematics and physics concepts that the students should already have dominated before the course.

decimal places required in the answer text box is specified. Some examples of the questions are presented in the Appendix.

- The FA can be taken as many times as students want without penalties and they are informed that each attempt lasts an average of 40 to 60 min. The FA can be submitted in a predefined time frame of up to three weeks.
- Each time the FA is started, the questions are shuffled in a random order, and the numerical values used in each question are also generated randomly. Students have at most three hours to complete the assessment. Thus:
  - In each new attempt, students will receive a shuffled sequence of questions, making it difficult to sequentially memorize how to solve each question.
  - The new attempt will only be available to the student after one hour to avoid memorization of the questionnaire.
  - The student must have to recalculate all the questions in a new attempt since their numerical values are always "renewed" by random selection within a pre-established values domain. This feature forces the student to solidify the concepts involved in the FA.
  - Sharing exact solutions and copies among students becomes less possible. Even if two students decide to do the exercises simultaneously, the order of each question will be different for each student and the numerical values will also be different for similar questions.
- The platform is programmed to automatically correct the questions and immediately report the student of his final score providing the student with the correct answer to each question.
- In case students want to improve their scores, they must perform a new attempt knowing that the new attempt overwrites the last score regardless of its new value. This feature gives students the possibility to achieve the maximum score.
- The system records each attempt by each student, allowing for subsequent monitoring by the course coordinators not only of the effectiveness of the practice but also of the performance of those students who sought to make the FA.

It is important to mention that to address student inquiries: we offer a comprehensive support system with both synchronous and asynchronous options. Recognizing the importance of personalized attention, we provide live video tutoring from 8:00 AM to 7:00 PM, Monday through Friday. In addition, students can submit questions and receive a response from a qualified tutor within 24 h.

Collected Data

The ICF1 course is usually attended by more than 1000 students every semester, and with the implemented assessments, a great amount of data is generated and can be used to make predictions about the final situation of students. In this context, machine learning algorithms offer a good alternative to predict the final situation of students, outperforming simple statistical methods. To make these predictions, the data to be used should be properly selected. It is important to note that all data were collected from a Brazilian public institution, making the evaluation results also public by Brazilian law. Furthermore, all data were anonymized, ensuring that no individual student could be identified in our results and this article, thus exempting the need for student consent, as outlined in the General Data Protection Regulation (GDPR) and the Brazilian Data Protection Law (Lei Geral de Proteção de Dados, LGPD).

The data used to make the predictions ranged from the second semester of 2020 (2020-2) to the second semester of 2022 (2022-2). The first four semesters were used to train and evaluate machine learning algorithms, while the last semester was employed to simulate how the algorithm would have predicted the final situation of students.

These semesters were chosen because they have an equal number of assessments. In Table 1, all predictors used in this study are presented together with their respective descriptions. The total number of students across the 5 semesters is 5476, and it is important to note that some students took the course more than once, but this was not taken into account during the training of the algorithms.

Table 1 Full list of predictors

Predictor	Type	Description
Course	Nominal	Students' course
Campus	Nominal	City where students are enrolled
Semester	Numeric	Semester of the year (1 or 2)
Introduction	Numeric	FA of basic concepts expected prior to the course
Practice1	Numeric	Optics experiments and analysis
Optics	Numeric	FA of geometric optics
Vectors	Numeric	FA of vectors
PE1	Numeric	PE of geometric optics and vectors
Kinematics	Numeric	FA of kinematics
Practice2	Numeric	Kinematics/Force experiments and analysis
Forces	Numeric	FA of forces
Astronomy	Numeric	FA of astronomy
PE2	Numeric	PE of kinematics, forces and astronomy
Score trial	Numeric	Pre-grades for all FAs
Frequency trial	Numeric	Total number of trials for each FA

Data Preprocessing

The data generated and stored on the Moodle platform were pre-analyzed using some Python 3.7.11 libraries such as Numpy 1.21.5 (Harris et al., 2020), Pandas 1.3.4 (Team, 2020; Wes et al., 2010), Matplotlib 3.5.3 (Hunter, 2007) and Seaborn 0.12.2 (Waskom, 2021). From the FAs results, three features were extracted from the Moodle platform, related to time, frequency, and scoring. Regarding time, we considered the total time allocated to students to obtain the highest possible grade, as this time limit was varied across the five semesters under study. In particular, it was observed that some students participated up to 17 attempts on certain FAs, while others attempted them only once. For this reason, the number of attempts by each student was counted for each FA and was considered as a frequency feature.

This type of data often presents certain particularities that need to be addressed. The first was related to data cleaning, since different datasets exhibited different formats and/or missing values. The missing data were filled with a value of negative one to make it clear for the algorithm that the student did not take that assessment. Since the course and campus are nominal data, a data transformation was performed from nominal to binary, so all the data could be interpreted as numbers by the algorithms.

After analyzing all predictors, it was observed that differences in the time window allotted to students to complete each FA did not show a change in the prediction of their final situation. Therefore, this feature was excluded from the final results.

Machine Learning Methods

In order to classify the final situation of the students, three supervised machine learning algorithms were chosen. These algorithms are trained with data that already have a classification or label, allowing the algorithm to be trained with a smaller amount of data. The chosen algorithms are described below:

1. Logistic regression (LR) is used for a binary or multinomial classification (Bertsimas & King, 2017). The classification is performed by expressing the variables or attributes (numerical or categorical) in a linear equation as follows:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots = \theta^T \boldsymbol{X}, \tag{1}$$

where $\boldsymbol{X}$ and θ^T are vectors of the variables (x_i) and the regression parameters (θ_i), respectively. This equation is then transformed into a probability using the sigmoid function

$$p(\boldsymbol{X}) = \exp(\theta^T \boldsymbol{X})/(1 + \exp \theta^T \boldsymbol{X}). \tag{2}$$

In order to optimize the θ_i parameters, the newton-cg solver from scikit-learn was employed, which maximizes the likelihood (Fan et al., 2008; Pedregosa et al., 2011; Peng et al., 2002).

2. Support vector machine (SVM) was initially designed for binary classification and has been extended for multiclass classification and regression (Chen et al., 2005). The underlying concept of this algorithm involves identifying a hyperplane that can effectively separates two classes, expressed mathematically as follows:

$$\boldsymbol{W} \cdot \boldsymbol{X} + b = 0. \tag{3}$$

In this equation, $\boldsymbol{W}$ denotes the hyperplane or hypervector, $\boldsymbol{X}$ represents all the variables or attributes, of the classes, and b is the bias which determines the gaps between the hyperplane and the nearest data point (variables). The algorithm's objective is to determine $\boldsymbol{W}$ while maximizing the value of b (Smola & Schölkopf, 2004). In many cases, the classes are not linearly separable. To address this, the notion of kernel trick was introduced. This technique involves mapping the original variables into a higher-dimensional space where linearly separability of classes can be achieved. Within Scikit-Learn, four kernels have been incorporated: linear, multinomial, radial base function, and sigmoid kernels. In situations where classes are not linearly separable, the algorithm employs the gradient descent method to optimize the hyperplane and bias. Scikit-Learn employs the Sequential Minimal Optimization (SMO) method for this optimization process (Chang & Lin, 2011; Fan et al., 2008; Platt, 1998).

3. Stochastic gradient boosting method (GBM) combines the power of decision trees with optimization techniques (Friedman, 2002). It offers the flexibility to adjust both the depth of the user-defined trees and the total number of trees in the ensemble, although it is recommended to limit the depth to 6 (Hastie et al., 2001). Once the depth and number of tree parameters are configured, the process continues using a stochastic technique. This entails that, at each step, a random subset of variables is selected and distributed among the trees. Each tree is constructed and trained to perform a classification, often referred to as a weak classifier.
Classification within each tree is accomplished by formulating the variables in a linear equation and applying a sigmoid function to assign to one of the two possible classes in a binary classification scenario. By utilizing these weak classifiers, a comprehensive classifier is constructed. Subsequently, the quality of the classification is evaluated using the cross-entropy function (or log-loss). Further enhancements of this classification are made using a gradient descent method to minimize the clas-

sification error. At the end, a function F(x) is found to map variables x to their output values y (Hastie et al., 2001).

Metrics

Metrics in machine learning are quantitative values used to evaluate task performance. An important metric commonly employed to evaluate binary classification tasks is the area under the receiver operating characteristic curve (AUC) (Hanley & McNeil, 1982). Unlike the accuracy metric, AUC is a reliable predictor when dealing with imbalanced data, which is often encountered in real-world datasets and is also addressed in this work.

The AUC value is determined by calculating the area under the curve of the sensitivity versus (1 - specificity). The sensitivity and specificity represent the probabilities of correctly predicting the students who will pass and fail the course, respectively. To achieve a balance between sensitivity and specificity, a threshold τ must be determined (Brown & Davis, 2006). An optimal cut-off point for τ can be obtained by maximizing the Youden index (Perkins & Schisterman, 2006), a metric commonly used in other machine learning classifications (Bertolini et al., 2021; Fluss et al., 2005).

Data Analysis

Some considerations should be expressed regarding the algorithm training. As explained in the previous section, to strike a balance between specificity and sensitivity, the Youden index was maximized to determine the cut-off point threshold. Cross-validation was employed for training the algorithms, which is a resampling method. Kohavi et al. (1995) determined the number of folds for cross-validation for real datasets to be equal to 10 and this was used to train the algorithms.

By analyzing the data, it was observed that the number of students participating in each attempt decreased by half in each FA. Therefore, based on this and some predicting tests, it was considered better to use only the first two attempts of each FA.

Results and Discussions

The online learning environment, a Moodle-based platform, used in CEDERJ courses allows course coordinators to have a large amount of useful data at hand to monitor the progress of students throughout the semester. In order to respond to the first question (RQ-1), online FAs were implemented in Moodle. The following section will present an analysis of student grades from one of the FA, illustrating how their progress can be effectively tracked.

Mean Grades by Attempt

As a case test, the evolution of the students' average scores for the Introductory FA of the second semester of 2020 (for a maximum score of 10.0) is presented in Fig. 1. The darker vertical gray bars indicate the region between the end of the first quartile and the beginning of the last quartile, showing that 50% of the students have grades in this region. The small horizontal lines at each attempt indicate the average grade of that attempt (for example, at the first attempt the average score was slightly over 5.3). Students are divided into groups with red triangles (labeled "still trying to improve") and blue hexagons (labeled "last attempt") that represent the average grade of students who made and did not make the next attempt, respectively. This separation shows the improvement of students who make the second attempt. The gray area of the plot for each attempt (referred to from now on as 'violin plots' due to its resemblance) helps to visualize the distribution of students for each score.

For example, when analyzing the first attempt in Fig. 1, the red triangle on the first attempt indicates the average score of the set of students who also made the second attempt. In addition, the horizontal line of the second attempt is calculated using all the students in this second attempt. By separating the students in this way, the horizontal line of the second

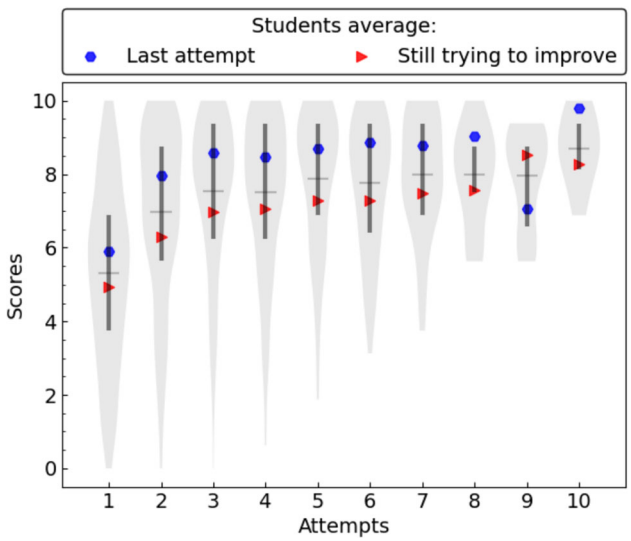

Fig. 1 Evolution of the students' average scores for the introductory assessment of 2020-2, distinguishing between groups that proceeded to the next attempt and those that did not

attempt represents the same set of students as the red triangle of the first attempt. This shows that the students who make the second attempt obtain a higher average than the students who only made the first attempt (identified by the blue hexagon of the first attempt), thus clearly indicating an average improvement of the students that performed the next attempt.

This analysis can be done for every pair of subsequent attempts to identify the score's improvement. It is observed that from the third attempt until the ninth attempt, the student's average fluctuates from 7.5 to 8, showing that students are having difficulties to improve. Figure 1 is limited to the tenth attempt because the subsequent attempts were performed by too few students to allow us to discuss an "average" behavior.

The fact that the numerical values of the questions are randomly selected among a predefined interval makes each attempt always different from each other and drives the student to study to obtain an improvement in his performance. This should be reflected in their understanding of the underlying concepts, and consequently, better performance in the PE is expected for those students who make more attempts.

Table 2 shows the average scores of the students and the standard deviation, in parentheses, for the first three attempts of each FA determined for the five semesters of study. Generally, students begin the first attempt of each FA with an average grade of around 5. Vector's topic appears to be the most challenging FA, with an average grade that ranges from 4.25 (in the first semester) to 2.75 (in the last semester). In contrast, Astronomy seems to be the easiest, with an average starting grade of 7.05, which decreases with each subsequent semester up to 6.40. In general, the average grades consistently increase with the number of attempts, suggesting that some students struggle initially but improve with newer attempts. It is noteworthy to mention the high standard deviations, showing significant variability in students' performance.

The analysis presented in Fig. 1 could be performed for each student individually. However, given the large number of students in the discipline and the high distribution of grades (presented in Table 2), it is challenging to identify those who require more attention. Therefore, we chose to train machine learning algorithms to determine if accurate predictions of the final situation of the students could be made. The results will be discussed in the next section.

Machine Learning Results

In this section, the results of the AUC metric obtained for all machine learning algorithms used will be discussed. In Fig. 2, the metric AUC is presented as a function of the assessment in chronology order. T1_'FA' and T2_'FA' represent the first and second trials for each FA assignment - Introductory (I), Optics (O), Vectors (V), Kinematics (K), Forces (F) and Astronomy (A) -, N_Trials represents the number of trials for that assignment (additional data from FAs).

There are three graphs which represent the results obtained for a specific algorithm (from up to down, LR, SVM, and GBM). The red triangles represent the AUC values when the

Table 2 Students' average (standard deviation) grades of the first three trials in each FA, by semester and year

Sem.-Year	Trial	Introduction	Optics	Vectors	Kinematics	Forces	Astronomy
II-2020	1	5.31(2.42)	4.70(2.93)	4.25(3.28)	4.11(3.44)	5.93(3.56)	7.05(2.86)
	2	6.97(2.34)	6.70(2.52)	6.40(2.91)	6.32(3.23)	7.90(2.65)	7.88(2.17)
	3	7.55(2.04)	7.01(2.42)	6.94(2.64)	6.74(3.12)	8.24(2.26)	8.92(2.19)
I-2021	1	4.83(2.71)	4.75(2.96)	4.08(3.54)	4.54(3.76)	6.26(3.74)	7.63(2.76)
	2	6.44(2.47)	6.85(2.47)	6.45(3.16)	6.62(3.16)	7.75(2.57)	8.08(2.47)
	3	7.41(2.13)	7.26(2.36)	7.33(2.86)	7.21(2.70)	8.54(2.16)	8.51(2.43)
II-2021	1	4.25(2.53)	3.78(2.94)	3.46(3.17)	4.37(3.89)	5.58(3.47)	7.13(3.04)
	2	5.86(2.39)	5.95(2.59)	5.36(2.94)	6.12(3.37)	7.36(2.52)	7.77(2.52)
	3	6.64(2.19)	6.67(2.33)	6.26(2.39)	6.85(3.18)	8.00(2.14)	8.38(2.63)
I-2022	1	4.20(2.62)	3.67(3.14)	3.11(3.11)	4.13(3.74)	5.28(3.53)	6.93(2.92)
	2	5.36(2.43)	6.14(2.60)	4.99(3.08)	5.68(3.47)	7.27(2.74)	7.80(2.33)
	3	6.13(2.19)	6.87(2.33)	5.83(2.81)	6.15(3.22)	7.50(2.60)	8.49(2.18)
II-2022	1	4.19(2.53)	3.05(2.96)	2.75(3.05)	2.80(3.42)	3.75(3.48)	6.40(2.93)
	2	5.32(2.30)	5.10(3.05)	4.92(2.92)	4.56(3.71)	6.59(2.74)	7.70(2.18)
	3	6.07(2.55)	5.24(3.22)	5.17(3.29)	4.88(3.16)	7.14(2.52)	8.40(1.73)

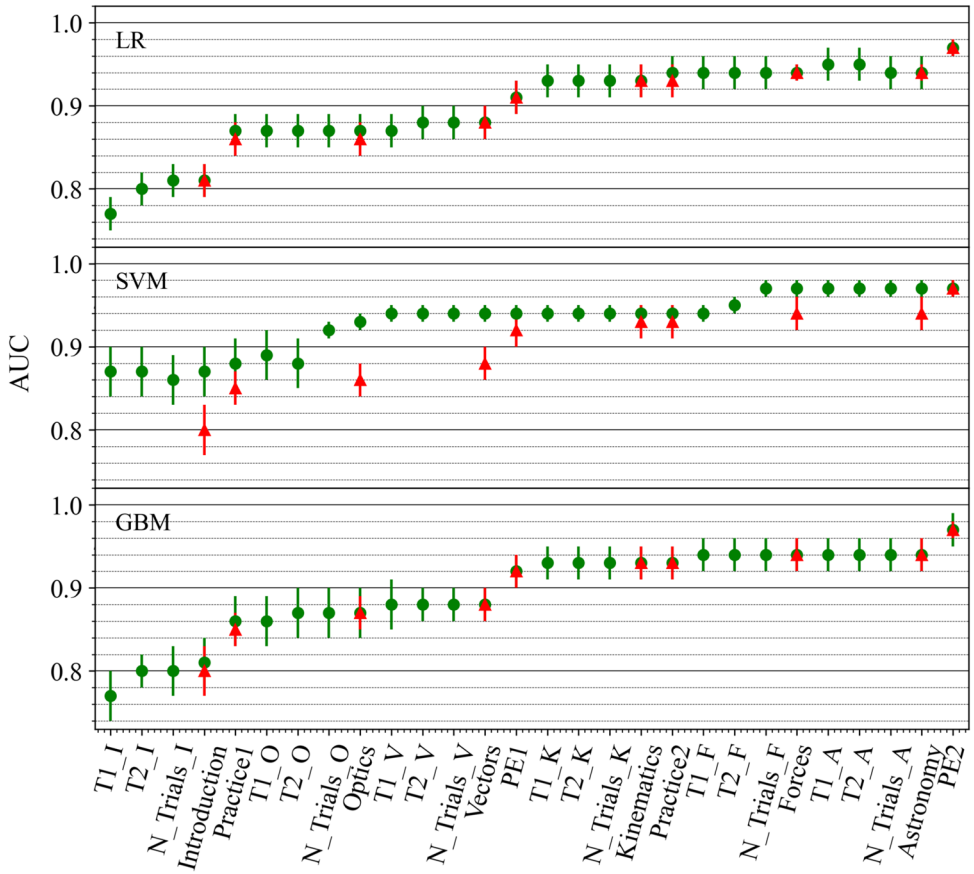

Fig. 2 AUC results as a function of the respective evaluation for the LR, SVM, and GBM algorithms. Red triangles represent results using final grades for each assessment, while green circles incorporate grades from the first two trials (T1_'FA' and T2_'FA'), the number of trials and final grades. Vertical lines represent one standard deviation derived from cross-validation

algorithm was trained considering only the final grades of all assessments, and the green circles show the AUC values considering also the grades from the first two trials of every FA and the total number of trials. For all results, a standard deviation was added as vertical lines as an error bar. It can be observed that the AUC results obtained show an increasing trend over each new assessment.

Taking into account the final grade of the introductory FA for the LR, SVM, and GBM algorithms, the AUC value obtained is 0.80 ± 0.03, 0.80 ± 0.03, and 0.81 ± 0.02, respectively. With the additional data from the FA these values are 0.81 ± 0.03, 0.87 ± 0.03, and 0.81 ± 0.02 also for the algorithms LR, SVM, and GBM, respectively. As all values are equal to or greater than 0.8, it could be considered that all algorithms are good classifiers, as suggested by Hosmer Jr et al. (2013).

It is interesting to note that considering the additional data from the FA leads to an increase in the AUC value when using the SVM algorithm. This increase ranges from 0.07 in the first FA (as the highest increment) to 0.01 in the kinematics FA (as the lowest value). In the case of LR, an increase in the AUC value is observed in the first two FAs, which does not exceed 0.01. No improvement is observed when using the GBM algorithm. Based on these results, the SVM algorithm makes a better prediction of the final situation of the students. Recently, the SVM algorithm showed acceptable performance accuracy with unbalanced data (TK & Midhunchakkravarthy, 2023).

It is important to note that the additional data from the FAs do not have a specific temporal point as students have a time window to complete their attempts. Consequently, this suggests that interventions could be carried out while the FA time frame period is still open. This underscores the importance of using formative assessment data and aligns with the findings of Bulut et al. (2023), who advocate the inclusion of formative assessment data in the development of learning analytic models, instead of complex data such as event logs or clickstream data, for example.

An important inquiry is about the ability of algorithms to predict the upcoming semester results. Figure 3 helps to

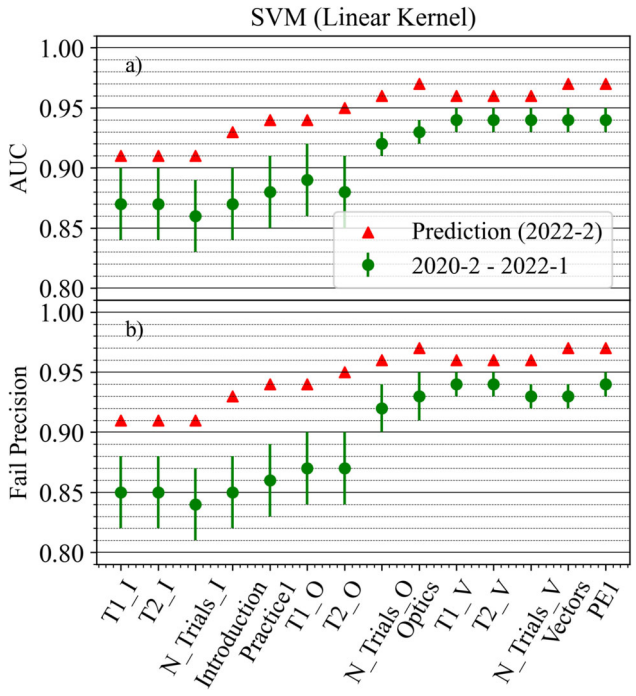

Fig. 3 Performance metrics of the SVM algorithm with a Linear Kernel: (a) AUC and (b) Fail Precision as a function of the assessments. Red triangles indicate the prediction evaluation for the 2022-2 semester, while green circles represent the training evaluation using data from semesters 2020-2 to 2022-1

answer this inquiry. Figure 3(a) and (b) depict the result derived from the AUC values and the Fail Precision (or the precision of the students who failed), respectively, as a function of the respective assessment during the semester. The results obtained from the trained SVM algorithm (semesters data from 2020-2 to 2022-1) are represented as green dots, while the "prediction[2]" data for 2022-2 is shown as red triangles.

In order to identify an earlier intervention time, we have limited our presentation to the results up to PE1 evaluation. From these results, it is evident that the projected AUC and Fail Precision values for the 2022-2 semester surpass the results from the previous semesters. This trend holds across all evaluations. The Fail Precision results offer more specific insight as they reveal the proportion of students correctly identified as failing the discipline from the total number of students predicted to fail. The Fail Precision value for the initial attempt of the introductory assessment is 0.91, and this value steadily increases over time. This outcome suggests that at the beginning of the semester, only 9% of the students would not be correctly identified as at risk of failing.

[2] The term "prediction" here is used as a benchmark of the pre-trained algorithm's ability to make predictions.

Final Discussion

With the increasing advancement of technology and the development of machine learning and artificial intelligence algorithms, their use in the field of education has grown strongly in recent years (Sghir et al., 2023). However, the most published works are in the computing area, which is somewhat expected because professionals in this field are the ones who predominantly create algorithms (Sghir et al., 2023).

The main objective of this work is to provide better support to students in the introductory physics course. Based on our logistical limitations and the large number of students, it was deemed appropriate to improve the formative assessments of the course. This implementation was carried out on a Moodle-based platform, which provides different tools to simplify the implementation. One way of learning is through the identification of gaps in our knowledge, and this should not be penalized. These formative assessments created here allow students to identify these gaps because they do not penalize students for the mistakes they make in their learning progress. This approach is a type of remediation, as discussed in the background section, and allows us to give students the opportunity to engage in self-regulated study.

Assessment questions were designed as key elements or building blocks of a more comprehensive question with the aim of incrementally building knowledge. The current design of the synchronous and asynchronous tutoring system limits our ability to analyze student-tutor interaction. This is particularly concerning because a significant portion of students do not ask questions. Many students, citing work commitments and time constraints, reportedly do not utilize the tutoring system.

Limitations

The presented results are strongly dependent on the data used, making replication potentially challenging due to diversity and the number of students, the topics covered in the course, and the number of degrees for which the ICF1 course is offered. In addition, this research does not focus on demographic factors because previous studies have not observed relevant contributions from these variables to prediction accuracy. However, in this study, we recognize that demographic factors have a minor positive influence on prediction outcomes. A more detailed analysis of these factors is an important avenue for future research.

The data used to train the algorithms was collected mainly during pandemics. Although the algorithms accurately predicted student performance for the 2022-2 semester (perhaps after the post-pandemic), it is likely that ongoing adjustments will be necessary to the algorithm training data. This neces-

sity arises from the disparities in teaching and assessment methods (e.g., PE) across different semesters. Furthermore, course coordination continuously implements modifications, which also require data updates. Another contributing factor will be interventions to support students at risk of failure. It is important to devise a method for incorporating these data to enhance the performance of the algorithm.

Conclusions

In this study, we implemented formative assessments for each topic of the course not only to quantify students' learning progress, but also to serve as a way to identify a specific group of students needing special attention during the semester. The primary benefit is that students actively participate in improving their own knowledge. They receive their grades as feedback as soon as they submit their assessment and have the opportunity to review the content and make further attempts to get better grades. A secondary benefit is that this implementation generates huge amounts of data that could be used for the benefit of the students.

To analyze these data, we employ three machine learning algorithms to classify the final situation of the students, namely: logistic regression, supporting vector machine, and stochastic gradient boosting machine. When considering only the final grades of the assessments, all algorithms demonstrated excellent classification performance, achieving an AUC value of 0.8 or higher even for the initial assessment. In particular, incorporating data from the first two attempts of the first formative assessment further boosted the SVM's AUC value by 7%. This result shows that one should incorporate formative assessment data when building learning analytic models and establishes an advantage for the prediction of students who could fail because we can address some support as soon as they take the first attempt of the first formative assessment. The beauty of this method lies in its simplicity. It requires minimal resources and can be easily implemented, making it an attractive option.

Appendix

A few questions of the formative assessments are presented. The questions were translated into English for presentation in this paper. In order to evaluate a simple and direct concept, the questions are short and straightforward.

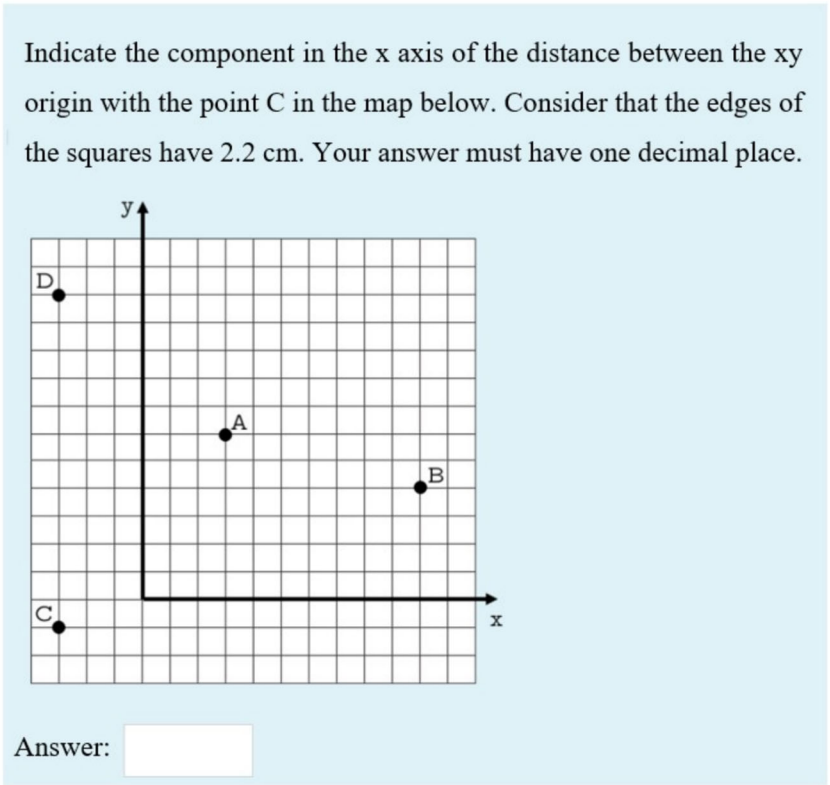

Fig. 4 A sample of a basic question belonging to the vector FA, where the edge length of the squares is randomly selected between 1 and 10, with one decimal place

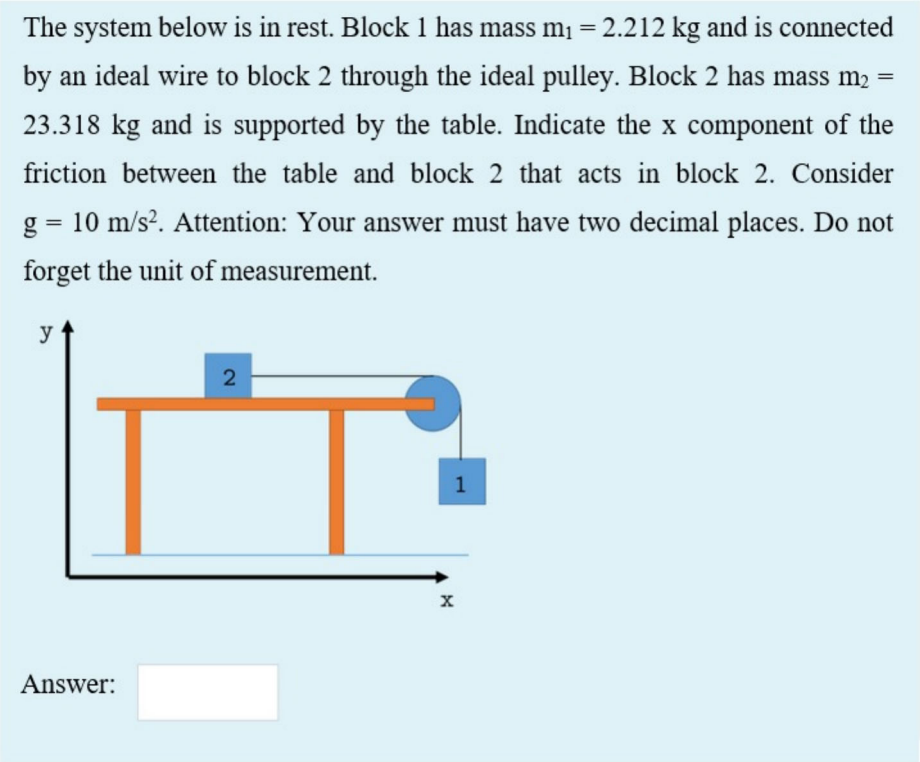

Fig. 5 A sample question of forces, where the values of mass m_1 and m_2 are randomly selected. The values are limited to a range between 0.5 kg and 4.5 kg for m_1 and 15 kg and 25 kg for m_2

Acknowledgements The authors would like to thank "Centro de Ciências e Educação Superior a Distância do Estado do Rio de Janeiro" foundation (CECIERJ) and "Universidade Aberta do Brasil" (UAB).

Author Contributions Charlie V. Sarmiento: conceptualization, methodology, investigation, data analysis, draft writing. Germano Maioli Penello: conceptualization, editing and reviewing, supervision. Lucas Sigaud: conceptualization, editing and reviewing.

Data Availability The data that support the study's findings are accessible upon request to the corresponding author.

Declarations

Ethics Approval Not applicable

Consent to Participate Not applicable

Conflict of Interest The authors declare no competing interests.

Conflict of Interest of Generative AI in Scientific Writing During the preparation of this work, the author (Charlie V. Sarmiento) used ChatGPT in order to make an English grammar correction. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Aiken, J. M., Henderson, R., & Caballero, M. D. (2019). Modeling student pathways in a physics bachelor's degree program. *Phys. Rev. Phys. Educ. Res., 15*, Article 010128. https://doi.org/10.1103/PhysRevPhysEducRes.15.010128

Alam, A. (2023). Improving learning outcomes through predictive analytics: Enhancing teaching and learning with educational data mining. In 2023 7th international conference on intelligent computing and control systems (iciccs) (p. 249-257). https://doi.org/10.1109/ICICCS56967.2023.10142392

Ardid, M., Gómez-Tejedor, J. A., Meseguer-Dueñas, J. M., Riera, J., & Vidaurre, A. (2015). Online exams for blended assessment. Study of different application methodologies. Computers and Education, 81 , 296-303. Retrieved from https://www.sciencedirect.com/science/article/pii/S0360131514002309. https://doi.org/10.1016/j.compedu.2014.10.010

Bahr, P. R. (2013). The deconstructive approach to understanding community college students' pathways and outcomes. *Community College Review, 41*(2), 137–153. https://doi.org/10.1177/0091552113486341

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy and Practice, 18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Bertolini, R., Finch, S. J., & Nehm, R. H. (2021). Testing the impact of novel assessment sources and machine learning methods on

predictive outcome modeling in undergraduate biology. *Journal of Science Education and Technology, 30*, 193–209. https://doi.org/10.1007/s10956-020-09888-8

Bertsimas, D., & King, A. (2017). Logistic regression: From art to science. *Statistical Science*, *32*(3), 367–384. Retrieved 2024-07-26, from http://www.jstor.org/stable/26408297

Boylan, H. R., & Saxon, D. P. (1999). What works in remediation: Lessons from 30 years of research. Unpublished report. Retrieved October , 14 , 2006.

Brown, C. D., & Davis, H. T. (2006). Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems, 80*(1), 24–38. https://doi.org/10.1016/j.chemolab.2005.05.004

Bulut, O., Cutumisu, M., Aquilina, A. M., & Singh, D. (2019). Effects of digital score reporting and feedback on students' learning in higher education. Frontiers in Education, 4 . Retrieved from https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2019.00065. https://doi.org/10.3389/feduc.2019.00065

Bulut, O., Gorgun, G., Yildirim-Erbasli, S. N., Wongvorachan, T., Daniels, L. M., Gao, Y., & Shin, J. (2023). Standing on the shoulders of giants: Online formative assessments as the foundation for predictive learning analytics models. *British Journal of Educational Technology, 54*(1), 19–39. https://doi.org/10.1111/bjet.13276

Chang, C.-C., & Lin, C.-J. (2011). Libsvm: A library for support vector machines. , 2 (3). Retrieved from https://doi.org/10.1145/1961189.1961199. https://doi.org/10.1145/1961189.1961199

Chen, P.-H., Lin, C.-J., & Schölkopf, B. (2005). A tutorial on v-support vector machines. *Applied Stochastic Models in Business and Industry, 21*(2), 111–136. https://doi.org/10.1002/asmb.537

Das Dores, A. J. L., Burbano, L. F. C., Aristizábal Henao, I. D., Galvez, L. F. P., Zabalaga, M. A. C., Mario Weyerstall, W., & Garzón, J. A. C. (2023). Learning by doing strategy for electronic engineer students at Unidad Central del Valle del Cauca. In 2023 ieee global humanitarian technology conference (ghtc) (p. 263-269). https://doi.org/10.1109/GHTC56179.2023.10354915

Effie Steriopoulos, E. G., & Harkison, T. (2022). Practical teaching tips on designing authentic assessments in tourism, hospitality and events (the) higher education. *Journal of Teaching in Travel and Tourism, 22*(4), 425–433. https://doi.org/10.1080/15313220.2022.2096181

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). Liblinear: A library for large linear classification. *The Journal of machine Learning research, 9*, 1871–1874.

Fluss, R., Faraggi, D., & Reiser, B. (2005). Estimation of the Youden index and its associated cutoff point. *Biometrical Journal, 47*(4), 458–472. https://doi.org/10.1002/bimj.200410135

Friedman, J. H. (2002). Stochastic gradient boosting. Computational statistics and data analysis, 38 (4), 367-378. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167947301000652 (Nonlinear Methods and Data Mining) https://doi.org/10.1016/S0167-9473(01)00065-2

Gikandi, J., Morrow, D., & Davis, N. (2011). Online formative assessment in higher education: A review of the literature. Computers and Education, 57 (4), 2333-2351. Retrieved from https://www.sciencedirect.com/science/article/pii/S0360131511001333. https://doi.org/10.1016/j.compedu.2011.06.004

Hall, C. W., Kauffmann, P. J., Wuensch, K. L., Swart, W. E., DeUrquidi, K. A., Griffin, O. H., & Duncan, C. S. (2015). Aptitude and personality traits in retention of engineering students. *Journal of Engineering Education, 104*(2), 167–188. https://doi.org/10.1002/jee.20072

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

Hansen, J. D. (2023). Methods for analyzing physics student retention and physics curricula. https://doi.org/10.33915/etd.12201

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., & Cournapeau, D.,., others,. (2020). Array programming with NumPy. *Nature, 585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hastie, T., Friedman, J., & Tibshirani, R. (2001). Kernel methods. In the elements of statistical learning: Data mining, inference, and prediction (pp. 165–192). New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-21606-5_6

Hettiarachchi, K. (2013). Technology-enhanced assessment for skill and knowledge acquisition in online education (Ph.D. thesis, Universitat Oberta de Catalunya. Internet Interdisciplinary Institute (IN3), Catalunya, Spain). Retrieved from http://hdl.handle.net/10609/31041

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. Hoboken: John Wiley and Sons.

Howard, E., Meehan, M., & Parnell, A. (2019). Quantifying participation in, and the effectiveness of, remediating assessment in a university mathematics module. *Assessment and Evaluation in Higher Education, 44*(1), 97–110. https://doi.org/10.1080/02602938.2018.1476670

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science and engineering, 9*(03), 90–95. https://doi.org/10.1109/MCSE.2007.55

Khan, A., & Ghosh, S. K. (2021). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and information technologies, 26*(1), 205–240. https://doi.org/10.1007/s10639-020-10230-3

Kohavi, R., et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, pp. 1137–1145). Retrieved from https://www.researchgate.net/profile/Ron-Kohavi/publication/2352264_A_Study_of_Cross-Validation_and_Bootstrap_for_Accuracy_Estimation_and_Model_Selection/links/02e7e51bcc14c5e91c000000/A-Study-of-Cross-Validation-and-Bootstrap-for-Accuracy-Estimation-and-Model-Selection.pdf

Martin, P. P., & Graulich, N. (2023). When a machine detects student reasoning: A review of machine learning-based formative assessment of mechanistic reasoning. *Chem. Educ. Res. Pract., 24*, 407–427. https://doi.org/10.1039/D2RP00287F

Molly George, S. L., Lim, Helen, & Meadows, R. (2015). Learning by doing: Experiential learning in criminal justice. *Journal of Criminal Justice Education, 26*(4), 471–492. https://doi.org/10.1080/10511253.2015.1052001

Nandeshwar, A., Menzies, T., & Nelson, A. (2011). Learning patterns of university student retention. Expert Systems with Applications, 38 (12), 14984-14996. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417411008323. https://doi.org/10.1016/j.eswa.2011.05.048

Nantsou, T. P., Kapotis, E., & Tombras, G. S. (2024). Learning-by-doing as a method for teaching the fundamentals of light to physics educators and students online. In M. E. Auer, U. R. Cukierman, E. Vendrell Vidal, & E. Tovar Caro (Eds.), Towards a hybrid, flexible and socially engaged higher education (pp. 53–64). Cham: Springer Nature Switzerland.

Nantsou, T. P., Kapotis, E., Tsirou, A., Nistazakis, H. E., & Tombras, G. S. (2024). Hands on experiments on atomic structure and particle physics for primary teachers at CERN. In M. E. Auer, U. R. Cukierman, E. Vendrell Vidal, & E. Tovar Caro (Eds.), Towards a hybrid, flexible and socially engaged higher education (pp. 300–311). Cham: Springer Nature Switzerland.

Nantsou, T. P., Kapotis, E. C., & Tombras, G. S. (2021). A lab of hands-on stem experiments for primary teachers at CERN. In 2021 IEEE Global Engineering Education Conference (EDUCON) (p. 582-590). https://doi.org/10.1109/EDUCON46332.2021.9453915

Nantsou, T. P., & Tombras, G. (2022). Hands-on physics experiments for k-6 teachers at CERN. In 2022 IEEE Global Engineering Education Conference (EDUCON) (p. 180-188). https://doi.org/10.1109/EDUCON52537.2022.9766515

Niiranen, S., & Rissanen, T. (2017). Learning by doing and creating things with hands: Supporting students in craft and technology education. In Patt: proceedings. Retrieved from https://jyx.jyu.fi/bitstream/handle/123456789/56354/1/niiranenrissanen.pdf

Olson, S., & Riordan, D. G. (2012). Engage to excel: producing one million additional college graduates with degrees in science, technology, engineering, and mathematics. report to the president. Executive office of the president. Retrieved from https://eric.ed.gov/?id=ed541511

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ..., others (2011). Scikit-learn: Machine learning in python. the Journal of machine Learning research, 12 , 2825–2830. Retrieved from https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https:/

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96*(1), 3–14. https://doi.org/10.1080/00220670209598786

Perkins, N. J., & Schisterman, E. F. (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology, 163*(7), 670–675. https://doi.org/10.1093/aje/kwj063

Pinchbeck, J., & Heaney, C. (2017). Case report: The impact of a resubmission intervention on level 1 distance learning students. *Open Learning: The Journal of Open, Distance and e-Learning, 32*(3), 236–242. https://doi.org/10.1080/02680513.2017.1348290

Platt, J. C. (1998). Fast training of support vector machines using sequential minimal optimization. In Advances in Kernel Methods: Support Vector Learning. The MIT Press. Retrieved from https://doi.org/10.7551/mitpress/1130.003.0016

Richards, Z., & Kelly, A. M. (2023). Predicting community college astronomy performance through logistic regression. *Phys. Rev. Phys. Educ. Res., 19*, Article 010119. https://doi.org/10.1103/PhysRevPhysEducRes.19.010119

Richlin, L. (2023). *Blueprint for learning: Constructing college courses to facilitate, assess, and document learning*. Taylor and Francis.

Rossano, V., Roselli, T., & Quercia, G. (2020). Coding and computational thinking: Using arduino to acquire problem-solving skills. In P. Isaias, D. G. Sampson, & D. Ifenthaler (Eds.), Technology supported innovations in school education (pp. 91–114). Cham: Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-030-48194-0_6

Sangpikul, A. (2020). Challenging graduate students through experiential learning projects: the case of a marketing course in Thailand. *Journal of Teaching in Travel and Tourism, 20*(1), 59–73. https://doi.org/10.1080/15313220.2019.1623150

Sghir, N., Adadi, A., & Lahmer, M. (2023). Recent advances in predictive learning analytics: A decade systematic review (2012–2022). *Education and information technologies, 28*(7), 8299–8333. https://doi.org/10.1007/s10639-022-11536-0

Sim, T. Y., & Lau, S. L. (2018). Online tools to support novice programming: A systematic review. In 2018 IEEE conference on e-learning, e-management and e-services (ic3e) (p. 91-96). https://doi.org/10.1109/IC3e.2018.8632649

Sithole, A., Chiyaka, E. T., McCarthy, P., Mupinga, D. M., Bucklein, B. K., & Kibirige, J. (2017). Student attraction, persistence and retention in stem programs: Successes and continuing challenges. *Higher Education Studies*, *7*(1), 46–59. https://eric.ed.gov/?id=EJ1126801

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing, 14*, 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Sudakova, N. E., Savina, T. N., Masalimova, A. R., Mikhaylovsky, M. N., Karandeeva, L. G., & Zhdanov, S. P. (2022). Online formative assessment in higher education: Bibliometric analysis. Education Sciences, 12 (3). Retrieved from https://www.mdpi.com/2227-7102/12/3/209. https://doi.org/10.3390/educsci12030209

Team, T. P. D. (2020). pandas-dev/pandas: Pandas. Zenodo, February.

TK, S., & Midhunchakkravarthy. (2023). Academic performance prediction of at-risk students using machine learning techniques. In 2023 3rd international conference on advance computing and innovative technologies in engineering (icacite) (p. 1222-1227). https://doi.org/10.1109/ICACITE57410.2023.10183199

Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. *Assessment and Evaluation in Higher Education, 43*(5), 840–854. https://doi.org/10.1080/02602938.2017.1412396

Waskom, M. L. (2021). Seaborn: Statistical data visualization. Journal of Open Source Software, 6 (60), 3021. 10.21105/joss.03021

Wes, M., van der Walt, S., & Millman, J. (2010). Proceedings of the 9th python in science conference. SciPy Austin, Texas.

Wiliam, D. (2011). What is assessment for learning? Studies in Educational Evaluation, 37 (1), 3-14. Retrieved from https://www.sciencedirect.com/science/article/pii/S0191491X11000149 (Assessment for Learning) https://doi.org/10.1016/j.stueduc.2011.03.001

Williams, M. K. (2017). John Dewey in the 21st century. *Journal of Inquiry and Action in Education*, *9*(1), 7. Retrieved from https://digitalcommons.buffalostate.edu/jiae/vol9/iss1/7

Winberg, C., Adendorff, H., Bozalek, V., Conana, H., Pallitt, N., Wolff, K., ..., & Roxå, T. (2019). Learning to teach stem disciplines in higher education: A critical review of the literature. *Teaching in Higher Education, 24*(8), 930–947. https://doi.org/10.1080/13562517.2018.1517735

Wylie, C., & Lyon, C. (2016). Using the formative assessment rubrics, reflection and observation tools to support professional reflection on practice. Formative Assessment for Teachers and Students (FAST) State Collaborative on Assessment and Student Standards (SCASS) of the Council of Chief State School Officers (CCSSO). Retrieved from https://www.commonsense.org/sites/default/files/pdf/2018-05/document-formative-assessment-rubrics-and-observation-tools-document.pdf

Yang, J., DeVore, S., Hewagallage, D., Miller, P., Ryan, Q. X., & Stewart, J. (2020). Using machine learning to identify the most at-risk students in physics classes. *Phys. Rev. Phys. Educ. Res., 16*, Article 020130. https://doi.org/10.1103/PhysRevPhysEducRes.16.020130

Zabriskie, C., Yang, J., DeVore, S., & Stewart, J. (2019). Using machine learning to predict physics course outcomes. *Phys. Rev. Phys. Educ. Res., 15*, Article 020120. https://doi.org/10.1103/PhysRevPhysEducRes.15.020120