electronics

Article

Is Initial Performance in a Course Informative? Machine Learning Algorithms as Aids for the Early Detection of At-Risk Students

Maura A. E. Pilotti [1,*], Emaan Nazeeruddin [2], Mohammad Nazeeruddin [2], Ibtisam Daqqa [1], Hanadi Abdelsalam [1] and Maryam Abdullah [1]

[1] Department of Science and Human Studies, Prince Mohammad Bin Fahd University, Al Khobar 31952, Saudi Arabia; idaqqa@pmu.edu.sa (I.D.); habdelsalam@pmu.edu.sa (H.A.); mabdullah@pmu.edu.sa (M.A.)
[2] Department of Engineering and Computer Science, Prince Mohammad Bin Fahd University, Al Khobar 31952, Saudi Arabia; 202000080@pmu.edu.sa (E.N.); nmohammad@pmu.edu.sa (M.N.)
* Correspondence: maura.pilotti@gmail.com

Abstract: The extent to which grades in the first few weeks of a course can predict overall performance can be quite valuable in identifying at-risk students, informing interventions for such students, and offering valuable feedback to educators on the impact of instruction on learning. Yet, research on the validity of such predictions that are made by machine learning algorithms is scarce at best. The present research examined two interrelated questions: To what extent can educators rely on early performance to predict students' poor course grades at the end of the semester? Are predictions sensitive to the mode of instruction adopted (online versus face-to-face) and the course taught by the educator? In our research, we selected a sample of courses that were representative of the general education curriculum to ensure the inclusion of students from a variety of academic majors. The grades on the first test and assignment (early formative assessment measures) were used to identify students whose course performance at the end of the semester would be considered poor. Overall, the predictive validity of the early assessment measures was found to be meager, particularly so for online courses. However, exceptions were uncovered, each reflecting a particular combination of instructional mode and course. These findings suggest that changes to some of the currently used formative assessment measures are warranted to enhance their sensitivity to course demands and thus their usefulness to both students and instructors as feedback tools. The feasibility of a grade prediction application in general education courses, which critically depends on the accuracy of such tools, is discussed, including the challenges and potential benefits.

Keywords: predictive validity; general education; learning algorithms; COVID-19; online learning; face-to-face learning

Citation: Pilotti, M.A.E.; Nazeeruddin, E.; Nazeeruddin, M.; Daqqa, I.; Abdelsalam, H.; Abdullah, M. Is Initial Performance in a Course Informative? Machine Learning Algorithms as Aids for the Early Detection of At-Risk Students. *Electronics* **2022**, *11*, 2057. https://doi.org/10.3390/electronics11132057

Academic Editor: Hung-Yu Chien

Received: 17 June 2022
Accepted: 28 June 2022
Published: 30 June 2022

1. Introduction

General education courses in undergraduate programs are assumed to ensure that all college students acquire the foundational interdisciplinary knowledge, as well as the analytical and communication skills that are necessary to address the demands of their selected majors and those of their chosen professions [1]. Failures in general education courses may ignite a cascade of undesirable effects, which can span from mild (the repetition of a course) to severe (e.g., academic dismissal, delayed degree attainment, and loss of financial aid eligibility). Thus, how students perform in such courses can be considered key in determining their academic success, including retention and graduation [2,3]. It is an accepted fact that the timing of the identification of at-risk students is a critical aspect of the effectiveness of the implementation of remedial interventions [4–6]. However, educators usually have very little information about students' performance during the first half of the semester, which can make the identification of at-risk students both challenging and broadly consequential if erroneous conclusions are reached. To wit, unrecognized difficulties (an event classified

in signal detection theory as a miss) are likely to lead to course failure. Notwithstanding the need for valid early predictions of students' academic performance, which rely on limited information, most of the research on algorithms that are intended to assist educators' performance forecasts has relied upon much greater amounts of information collected within a much larger timeframe and has often involved discipline-specific subject matters [7–9]. Examples are predictions of final course grades in a particular subject matter based on students' grade point average (GPA), as well as grades in pre-requisite courses [10], or more simply, on students' academic history, as exemplified by their performance in past courses [11]. Yet, the algorithms that yield optimal results tend to vary considerably [12,13], along with a myriad of innovative stand-alone or hybrid solutions that appear as a regular stream in the extant literature [14]. As a result, the selection and subsequent use of a suitable technique for predicting at-risk students may become so challenging and overwhelming for an educator whose expertise is other than computer science that ignoring potentially viable technical solutions is the most likely course of action [15]. For such educators, continued reliance on personal intuition and conscious reasoning may seem preferable to the ordeal of understanding the technically dense machine learning literature. However, this comes at the cost of an increased likelihood of biases affecting the processing of students' information for assessment and decision-making [16,17], including personal preferences for the parameters to take into account and for the type and amount of data that are necessary to generate sensible predictions. Consider that, as the semester progresses, the amount of information about students that is available to an educator accumulates, but its utility decreases as remedial actions become harder to implement and their success becomes more uncertain [18,19]. Earlier predictions are unquestionably more valuable than later predictions, but at the beginning of a course, very little information is available to the educator, making predictions about a student's difficulties even more uncertain (e.g., is initial poor performance symptomatic of a momentary hurdle, perhaps linked to the idiosyncrasies of an assignment, or a reliable indication of serious issues?).

The COVID-19 pandemic has complicated the prediction matter by suddenly relocating students, most of whom were exclusively accustomed to face-to-face instruction, to online instruction. Although the synchronous online mode that is adopted by many institutions of higher learning has replicated many aspects of the face-to-face mode (e.g., real-time interactions in a virtual classroom), physical distance, technical idiosyncrasies, competencies, and other issues (e.g., students' degree and manner of adaptation to environmental changes) may have made learning in online courses different from that of face-to-face courses [20,21]. For instance, it has been proposed that the online mode has fostered the practice of a more continuous engagement in learning activities [20]. As evidence of change, studies have reported higher online performance (as measured by course grades) than pre-pandemic face-to-face performance [21–25]. However, other studies have reported declines or no change at all [4,25–27]. Thus, because uncertainty endures as to whether remote instruction during the pandemic has fostered relevant changes to students' learning, it remains unclear whether performance predictions that are made online and face-to-face can be considered equivalent.

In the present study, we examined whether algorithms that are commonly used for the predictions of final grades could be of assistance to educators in both face-to-face and online courses when the only information available to the educators is students' performance on the first test and assignment. Both assessment measures can be classified as formative assessment tools [28]. These are tools that are used by students to assess their learning in a course and by educators to determine the effectiveness of the instruction they deliver, thereby defining formative assessment as serving both diagnostic and feedback functions. In a course, formative assessment measures can be said to be particularly critical to students' academic success, since the information they provide has the potential to foster change in the way that students approach the curriculum and understand its demands, as well as in the way that educators teach. Thus, in principle, the earlier the assessment, the higher may be its impact on both students and educators. Formative assessment differs

from summative assessment (i.e., final tests), whose primary aim is to measure learning comprehensively across the entire semester as an evaluation of the extent to which it meets pre-set learning outcomes. A summative assessment indicator is the final course grade that is given to each student at the end of the semester. The effectiveness of early formative assessment measures, each of which covers a portion of the curriculum to be acquired in a course, resides in their ability to adequately predict final course grades which reflect the student's learning of the entire course curriculum.

It is customary for institutions of higher education to demand that students meet a minimum performance requirement to gain access and remain enrolled in any degree program. At the institution that was selected for the present research, this requirement entails maintaining a GPA that is better than a C (greater than 79%). Thus, to ensure authenticity, we classified the final course grades into three performance categories: high (H—equal to or greater than 90%); medium (M—80–89.99%), and low (L—79% or below). This stringent classification scheme created categories of comparable size, while it minimized the impact of grade inflation and educators' grading idiosyncrasies, as well as reflected the standards of academic success at the selected institution.

At the outset, we recognized that the predictive validity of a forecast may refer to a variety of key parameters, such as accuracy [(hits + correct rejections)/all responses, including hits, correct rejections, misses, and false alarms)]; precision [hits/(hits + false alarms)]; and sensitivity [hits/(hits + miss)]. In the task of identifying at-risk students, however, correct rejections are not particularly relevant. Furthermore, false alarms are much less costly or even less relevant than misses. Namely, false alarms are likely to reflect cases of temporary difficulties experienced by individual students which are mistakenly identified as enduring and/or severe (a false alarm), thereby creating unnecessary but fleeting stress in such students. Thus, in the present study, we relied on sensitivity as a measure of the predictive validity of forecasts of at-risk students (i.e., learners receiving an L grade at the end of the semester). A sensitivity score for an L classification was conceptualized as a proportion, including the number of correctly classified L grades divided by the number of grades that were either misclassified as H and M or correctly classified as L.

The study involved female undergraduate students of a society that is in transition from a patriarchal order to one that is akin to gender equity in education and employment [29–32]. In such a society, of which a prototypical example is the Kingdom of Saudi Arabia (KSA), female students of college age are the main target of top-down gender-equity interventions. Decrees and massive financial investments aim to re-set the country's social structure to favor meritocracy for both sexes at the expense of tribal and patriarchal favoritism [33]. Thus, the academic success of female college students is a priority for the adequate development of the economic and social engine of KSA, making our research a window into the performance of this highly valued population, as well as into the utility of early performance assessment in the said population.

The current study tested several popular learning algorithm(s) to answer two interrelated questions:

a. Can at-risk students (defined as those with an end-of-the-semester score of L in a general education course) be effectively identified by very early performance indicators (i.e., grades on the first test and first assignment) through one of these algorithms?
b. Do predictions of at-risk students vary between face-to-face instruction and synchronous online instruction, as well as with the specific subject matter taught in a course?

We selected a sample of courses that are representative of the general education curriculum of a Saudi higher education institution that follows a curriculum of U.S. import and a student-centered pedagogy. The courses had been taught by the same instructor both online (during the pandemic) and face-to-face (before the pandemic) for at least three semesters in each mode. The acceptable sensitivity threshold for the selected algorithms

was determined by a sample of educators who taught similar courses. We predicted that if early performance indicators cannot be relied upon to identify at-risk students, early predictions will exhibit a sensitivity score at the identified subjective threshold or below. This outcome is likely to be present if instructors are more lenient at the start of the semester, thereby making the results of the first assignment and test less representative of the demands that are placed upon students in the courses they teach. Alternatively, the higher a sensitivity score is above the threshold, the more the first assignment and test can be said to represent students' overall performance. The description of the specific algorithms that we selected and the rationale behind their selection are included in the Methods section.

2. Methods

2.1. Sample

The data set of the present study included the grades of 5158 female students that were enrolled in a general education course at a university in the Middle East (KSA). In the set, a random number uniquely identified each student and was associated with her grades on the first test and homework assignment, as well as her final course grade, all measured on a scale from 0 to 100. The data set included students who completed one of the following general education courses, which were carefully chosen to ensure an adequate representation of the general education curriculum that was adopted by the selected university and a minimal overlap of students: Arabic Cultural Studies (n = 1314); Introductory Psychology (n = 847); Statistics (n = 612); Wellness Education (n = 1390); and Written Communication (n = 995). The courses had been taught both face-to-face (before the pandemic; n = 2614) and synchronously online (during the pandemic; n = 2544) by the same faculty (n = 10). All faculty had at least 5 years of teaching experience in the sampled courses. If a student appeared in more than one course in our dataset, only the grades of one course would be included. Random selection dictated the course for which grades would be entered into the data set.

Both instructional modes relied on Blackboard for the posting of course materials, submission of assignments, and testing. Online classes also relied on Blackboard Collaborate, a platform that created a virtual classroom. In it, video, audio, and chat functions permitted participants to interact with each other in real-time. To ensure proper conduct during the tests that were administered in the virtual classroom, students were required to activate the video and microphone function of Blackboard Collaborate and rely on a lockdown browser application. In both online and face-to-face classes, anti-plagiarism software (i.e., Turnitin) was also used. Important to note is that the general education curriculum at the selected university followed a U.S. model in content and practice. Namely, the curriculum, whose content had been developed by the Texas International Education Consortium (TIEC) and approved by the Saudi Ministry of Education, required that English be used as the primary mode of communication and that the pedagogy adopted for instruction be student-centered. The university was known to rely on a standard-based grading system, according to which the performance of students in a course was defined by their attainment of the learning objectives that defined the curriculum of the course, regardless of how other students performed [34]. Faculty teaching any of the selected courses were required to comply with syllabi that were developed and approved by TIEC, thereby demanding activities (tests and assignments) that met an identical set of learning outcomes. In each of the selected courses, the first assignment and test generally covered at least 1/3 of the learning objectives that were specified by the entire curriculum. Since preliminary analyses yielded null differences for the variable "faculty" within the same course, this variable was not included in the data analyses described in the result section. The present research was conducted under the purview of the Deanship of Research of the selected institution.

2.2. Materials and Procedure

For convenience, students' grades (range: 0–100) were organized into 3 categories, high (H), medium (M), and low (L) performance. H stood for grades in the 90–100 range, M referred to grades in the 80–89 range, and L included grades between 79 and 0. Six different algorithms that are commonly used in educational research [35,36] were chosen to make predictions of students' final grades. These algorithms are K-Nearest Neighbor (KNN); Linear Regression (LR); Multi-Layer Perceptron (MLP); Naïve Bayes (NB); Random Forest (RF); and Support Vector Machine (SVM). Each algorithm is briefly described below. For each algorithm, an article from the pertinent literature describing its mathematical details is mentioned for the interested reader.

K-Nearest Neighbor (KNN) is a non-parametric classifier [37]. That is, it selects a letter grade as the final grade for a given student. KNN merely classifies a target object (e.g., a student's initial grades) by relying on a majority vote of its K neighbors (e.g., students with similar initial grades). The distance between the target object and each of its neighbors is what matters. In our research, to predict final grades (H, M, and L), the number of nearest neighbors used was 21. This number was chosen after fine-tuning the algorithm. The Euclidian distance between any two points was used to find the nearest neighbors.

Logistic regression [38] is a method used to predict a dependent variable (also called outcome variable) for a given set of independent variables (i.e., predictors). It is specifically used to calculate the probability of a categorical dependent variable (e.g., H, M, and L grades). For this study, multiclass logistic regression was used, as predictions needed to fall into more than two categories.

Multi-Layer Perceptron (MLP) is a type of Artificial Neural Network (ANN), which attempts to mimic how the human brain is organized and functions [39]. Namely, an ANN is a network of neurons (i.e., nodes) with the strength of their connections defined by their weight. The MLP that was used in this study was organized into an input layer, three hidden layers, and an output layer. It was a feed-forward network with a back-propagation algorithm for training. Adam optimizer, a stochastic gradient-based optimizer, was applied to optimize the log-loss function. A RELU activation function was used for the hidden layer.

The algorithm named Naïve Bayes (NB) works on the principles of Bayes' theorem, whose aim is to find the conditional probability of a given event [40]. NB assumes that each feature makes an individual and equal contribution to the outcome. For a dataset, NB first creates a frequency table for each attribute, which is then molded into likelihood tables. Lastly, the Bayesian theorem is used for calculating the posterior probability for each class. The class with the highest posterior probability is the prediction that is made by NB. Although in the grade prediction task, the assumption of independence for the variables that are used to make estimates is likely to be violated (e.g., first assignments and tests are unlikely to be independent measures of early performance), we included NB because of its recurrent presence in grade prediction research (see [35]).

The Random Forest (RF) algorithm relies on a tree-like structure consisting of nodes and leaves [41]. Each node corresponds to an input variable (e.g., a student's initial grades), whereas each leaf represents all the different outcomes that could be achieved (e.g., final grades illustrating H, M, and L performance). RF consists of many decision trees that are determined by the user. Each decision tree gives one prediction and the prediction with the highest number of votes becomes the model's final prediction. The main concept behind this algorithm is that many uncorrelated trees working as a community can outperform the working of a single tree. For the present research, we used 100 trees.

The core of a Support Vector Machine (SVM) is an algorithm where each data point is plotted in an n-dimensional space [42]. In such a space, the number of features is n and the value for each feature is a coordinate. Classification is achieved by finding a hyperplane, which is a line that separates groups of data. Finding the hyperplane with the maximum margin for all groups is an optimization problem. SVM is sensitive to the data points that are close to the border of any two groups. In our study, a multi-class SVM algorithm based

on Gaussian Radial Basis Function (RBF) kernel was used to predict students' H, M, and L performance, as measured by final grades in the courses in which they were enrolled. The algorithm works by solving a single optimization problem that maximizes the margins between all the designed classes simultaneously.

For measurement purposes, the dataset was partitioned into 70% training and 30% testing. The parameter of sensitivity was used to illustrate the predictive validity of the estimates that were generated by each algorithm, since in grade prediction matters it is particularly costly to miss a student at risk. Sensitivity = [hits/(hits + miss)]. In this equation, a hit (true positive) is the number of students who are correctly identified as receiving a particular letter grade (e.g., the algorithm predicts an L for a particular student, which is indeed what the student has received at the end of the semester). A miss (false negative or type II error) is the number of students who are incorrectly identified as not receiving a particular letter grade (e.g., the algorithm predicts a grade other than an L for a particular student, but the student's grade is an L). We excluded other measures of performance as either irrelevant to the task of predicting at-risk students, such as a correct rejection, or minimally costly, such as a false alarm. In the grade prediction task, a false alarm (false positive or type I error) is the number of students who are incorrectly identified as receiving a particular letter grade (e.g., the algorithm predicts an L for a particular student, but the student has received a grade better than an L). A correct rejection (true negative), instead, is the number of students who are correctly identified as not receiving a particular letter grade (e.g., the algorithm predicts a grade other than an L for a particular student, which is indeed what the student has received).

3. Results

The results of the present study are organized into the following sections: a description of students' performance, and a description of the performance of the chosen algorithms in predicting the final course grades.

3.1. Students' Performance

To obtain a sample of grades that adequately reflected students' key performance levels in the courses in which they were enrolled, and bypassed grade inflation and instructors' grading idiosyncrasies, we classified the final course grades into three performance categories: High (equal to or greater than 90%); Moderate (80–89.99%); and Low (79% or below). The latter category included at-risk students. Table 1 displays the percentage of grades that were assigned to H, M, and L performance by course and instructional mode.

Table 1. Percentage of students with H, M, and L performance by course and instructional mode.

Course	Mode	High Performance	Medium Performance	Low Performance
ACS	Face-to-face	37.9%	37.6%	24.4%
	Online	65.6%	15.9%	18.5%
PSY	Face-to-face	25.1%	40.5%	34.4%
	Online	41.7%	36.4%	21.9%
STA	Face-to-face	16.7%	34.5%	48.8%
	Online	33.1%	50.4%	16.5%
WCO	Face-to-face	30.8%	33.6%	35.6%
	Online	41.9%	31.1%	27.0%
WED	Face-to-face	29.7%	31.7%	38.6%
	Online	57.3%	13.5%	29.2%

Note: ACS = Arabic Cultural Studies; PSY = Psychology; STA = Statistics; WCO = Written Communication; and WED = Wellness Education.

Overall, a greater percentage of students yielded L or M performance in face-to-face classes than in online classes, whereas a greater percentage of students yielded H performance online, χ^2 (2, n = 5158) = 285.34, $p < 0.001$. However, when we examined the fre-

quency of grades H, M, and L in individual courses, a more nuanced pattern emerged about the relationship between performance level and instructional modality, χ^2 (2, n = 612–1390) $\geq$ 14.66, $p \leq 0.001$. In online classes, H was the most frequent score. The only exception was online STA for which M was the most frequent score. In face-to-face WCO and WED classes, there was a somewhat even distribution of L, M, and H scores. In face-to-face STAT classes, L was the most frequent score, whereas in face-to-face ACS classes most scores were either H or M. These patterns of frequency distribution were reflected in students' end-of-course feedback surveys for which STA was judged as a difficult course, but less so online. Although the other classes were reported not to be as difficult as STA, they were also seen as easier when they were online. However, at the start of the semester, STA was judged as more difficult when it was online.

3.2. Algorithms' Performance

The final course grades, labeled as H, M, or L, were used for estimation. We applied the selected algorithms to assess their ability to predict at-risk students in each of the selected performance categories. We relied on sensitivity scores as a measure of the quality of the estimation that was made. Table 2 displays the sensitivity scores of the L-performance category, which indexed at-risk students, as a function of instructional mode and type of course.

Table 2. Prediction of at-risk students (sensitivity scores) by instructional mode and course.

Algorithm	Modality	ACS	PSY	STA	WED	WCO	Mean
KNN	FtF	0.83	0.43	0.67	0.74	0.70	*0.70*
	Online	0.41	0.52	0.71	0.68	0.57	*0.50*
LR	FtF	0.86	0.43	0.67	0.69	0.74	*0.70*
	Online	0.41	0.58	1.00	0.63	0.76	*0.63*
MLP	FtF	0.86	0.55	0.71	0.69	0.70	*0.56*
	Online	0.41	0.52	0.90	0.62	0.68	*0.59*
NB	FtF	0.86	0.43	0.64	0.69	0.79	*0.73*
	Online	0.31	0.58	0.76	0.46	0.76	*0.55*
RF	FtF	0.76	0.43	0.76	0.74	0.70	*0.56*
	Online	0.41	0.48	0.71	0.69	0.57	*0.59*
SVM	FtF	0.75	0.40	0.67	0.69	0.63	*0.48*
	Online	0.25	0.48	1.00	0.61	0.61	*0.52*
Mean		*0.59*	*0.49*	*0.77*	*0.66*	*0.68*	

Note: KNN = K-Nearest Neighbor; LR = Logistic Regression; MLP = Multi-Layer Perceptron; NB = Naive Bayes; RF = Random Forest; SVM = Support Vector Machine. ACS = Arabic Cultural Studies; PSY = Psychology; STA = Statistics; WCO = Written Communication; and WED = Wellness Education.

To determine the extent to which actual instructors in actual classrooms would tolerate misclassifications of at-risk students as likely to do well in their courses, we presented 10 faculty who had experience in at least one of the selected courses with the following scenario: "Imagine that after a couple of weeks into a semester, you are asked to take over a class from another instructor who was abruptly granted a leave of absence for health reasons. A colleague offers you an algorithm that can help you identify students who will not do well in this class. Imagine that at present, unbeknownst to you, 10 students will not do well in this class without some sort of early intervention. What is the maximum number of students out of 10 that the algorithm can fail to correctly identify as being at-risk for you to deem the algorithm unlikely to be useful? Alternatively, what is the minimum number of students out of 10 that the algorithm should correctly identify as being at-risk for you to deem the algorithm useful? Keep in mind that the early identification of at-risk students in real life is a complex task, which may lead educators to misclassify some students as likely to do well (misses or false negatives). Thus, provide a realistic number that would apply to you as representative of your teaching experience". The answers included sensitivity rates ranging from 0.9 to 0.6, leading to an average of 0.7 (the value the arrow points to).

Thus, in our study, we considered a sensitivity score of 0.7 or above as acceptable, which was treated as the threshold of subjective effectiveness.

Was there an algorithm that could be described as superior in both face-to-face (FtF) and online instruction? Figure 1, which plots sensitivity as a function of the type of algorithm and mode of instruction, shows that KNN, LR, and NB consistently performed more effectively face-to-face than online. SVM, MLP, and RF yielded poor performance both online and face-to-face. Interestingly, no algorithm performed adequately (i.e., above the threshold of subjective effectiveness) in online classes.

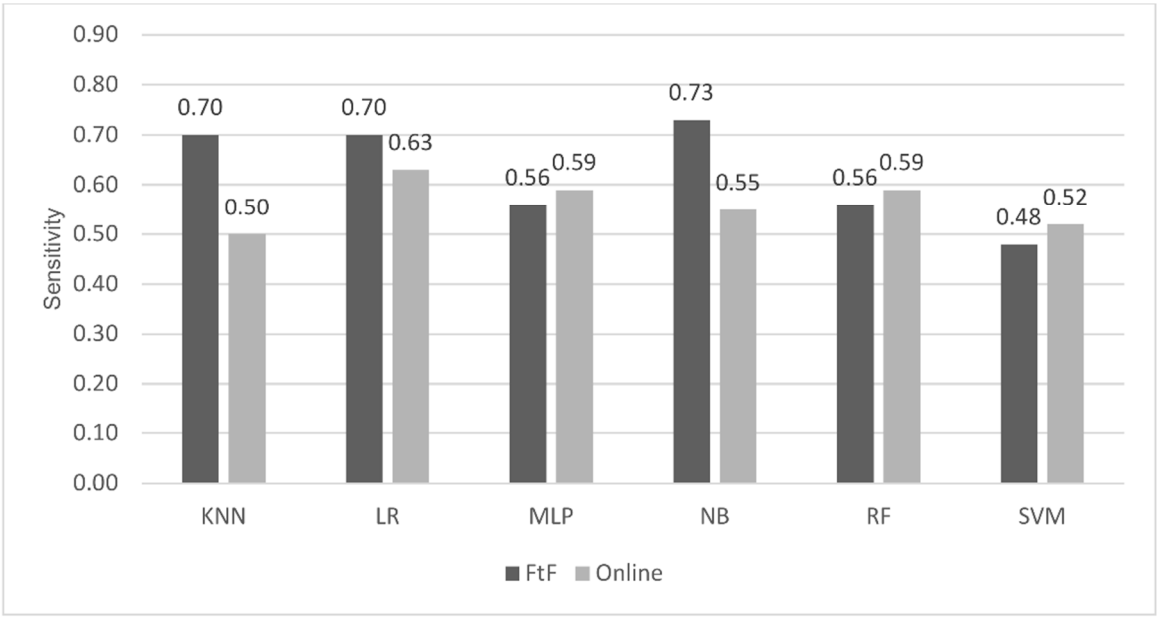

Figure 1. Prediction of at-risk students (sensitivity scores) by instructional modality and algorithm.

Across algorithms, did sensitivity change as a function of the type of course and mode of instruction? Figure 2, which plots sensitivity scores as a function of course and mode of instruction, shows two dissimilar patterns: ACS, WCD, and WED yielded more effective predictions face-to-face than online. Instead, STA yielded more effective predictions online. However, except for ACS and less so for STA, the differences between the modes of instruction were minor.

To better understand the pattern that was yielded by ACS, WCO, WD, STA, and PSY, we examined whether specific algorithms contributed to it. Figures 3–7, which plot the sensitivity scores as a function of the algorithm and mode of instruction in each of these courses, illustrate that the effectiveness of algorithms in making predictions depended on both the type of course and the mode through which the instruction was delivered. Thus, educators would be well advised to consider both variables in selecting algorithms for predicting at-risk students in the classes they teach. For instance, although ACS, WCO, and WED showed greater effectiveness in the prediction task (as measured by the threshold of subjective effectiveness of 0.7) in the face-to-face mode, not all algorithms did so across all courses. To illustrate, ACS was an exception, as all algorithms made effective predictions in face-to-face classes and none in online classes. However, such a clear pattern was not obtained in WCO and WED. Specifically, only KNN and RF made predictions that were above the threshold of effectiveness in face-to-face WCO classes, whereas all algorithms, except SVM, made predictions that were above the threshold of subjective effectiveness in face-to-face WED classes. Furthermore, only LR and NB made effective predictions in online WED classes. In contrast to the checkered pattern of WCO and WED, STAT exhibited a pattern that was largely the opposite of ACS. Namely, all algorithms yielded predictions that were above the threshold of subjective effectiveness online, whereas only MLP and RF made predictions that were above said threshold face-to-face. Irrespective of whether

PSY was delivered face-to-face or online, the predictions of all the algorithms were poor, all well below the threshold of subjective effectiveness.

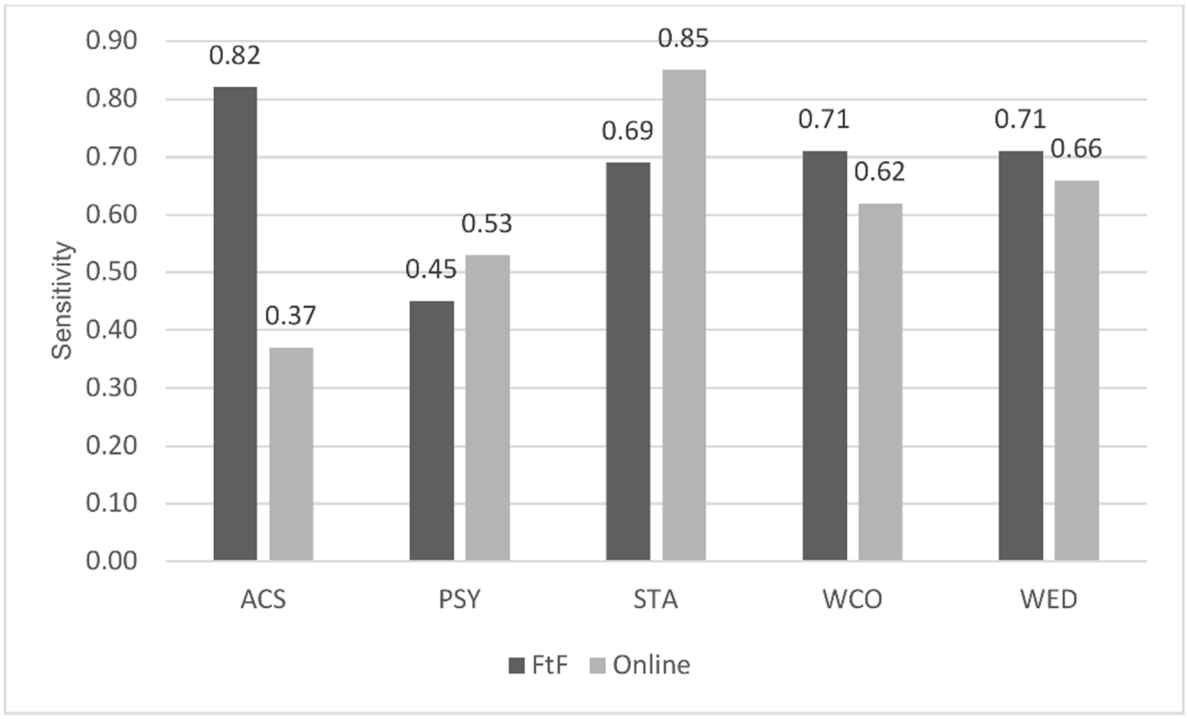

Figure 2. Prediction of at-risk students (sensitivity scores) by instructional modality and course type.

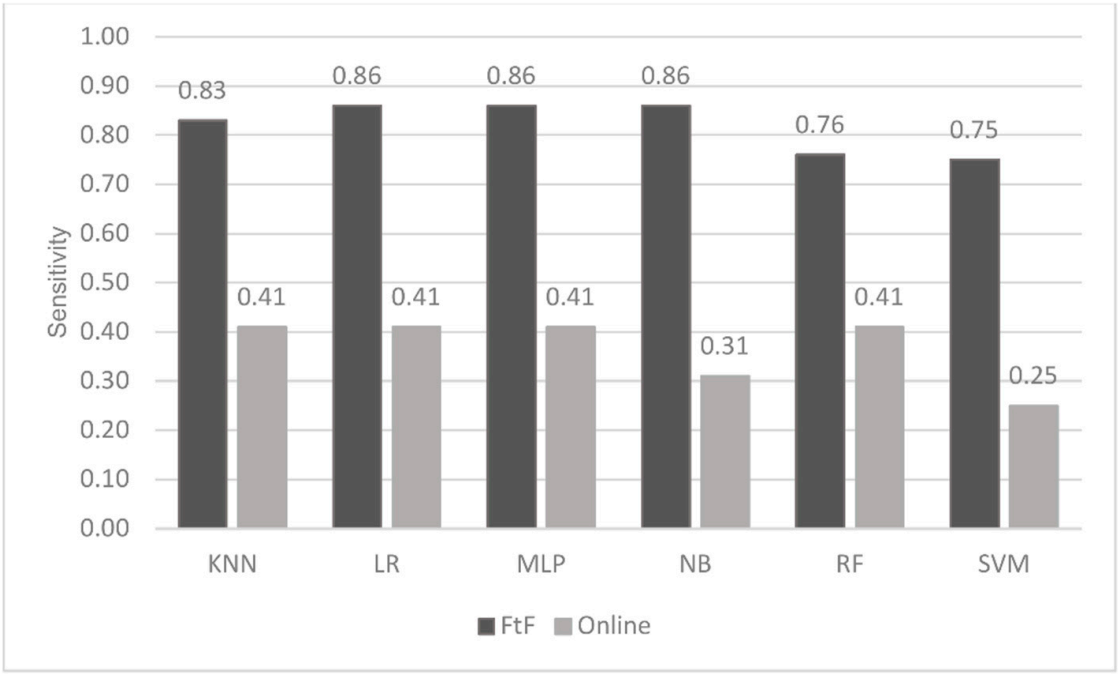

Figure 3. ACS course: prediction of at-risk students (sensitivity scores) by instructional mode.

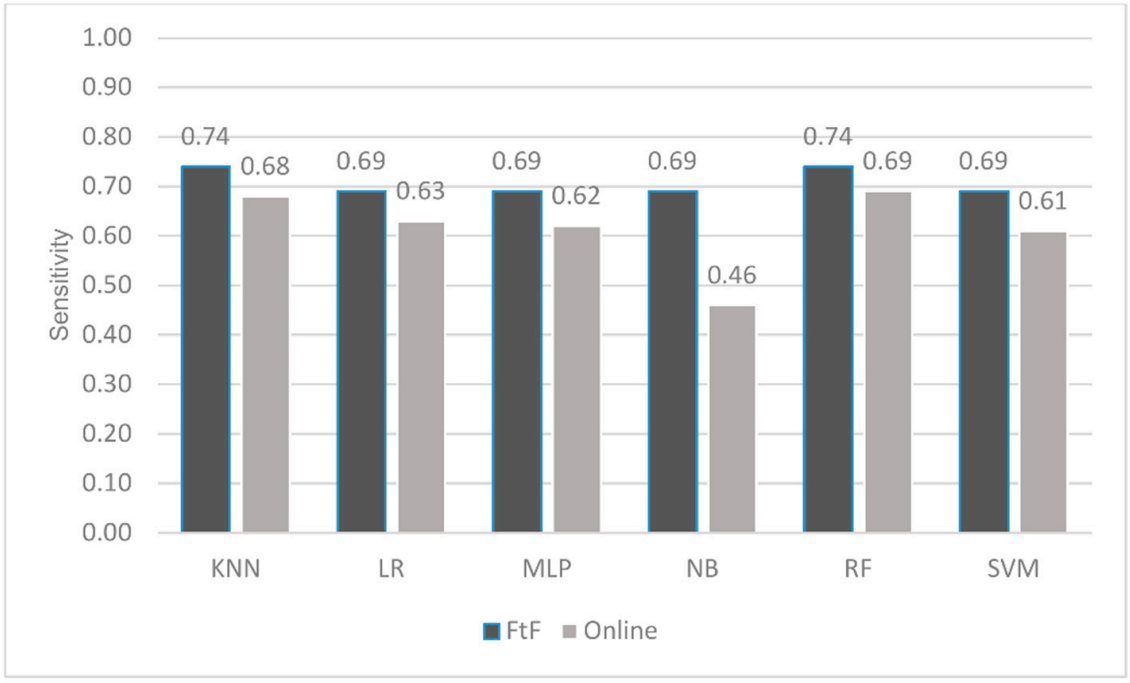

Figure 4. WCO course: prediction of at-risk students (sensitivity scores) by instructional mode.

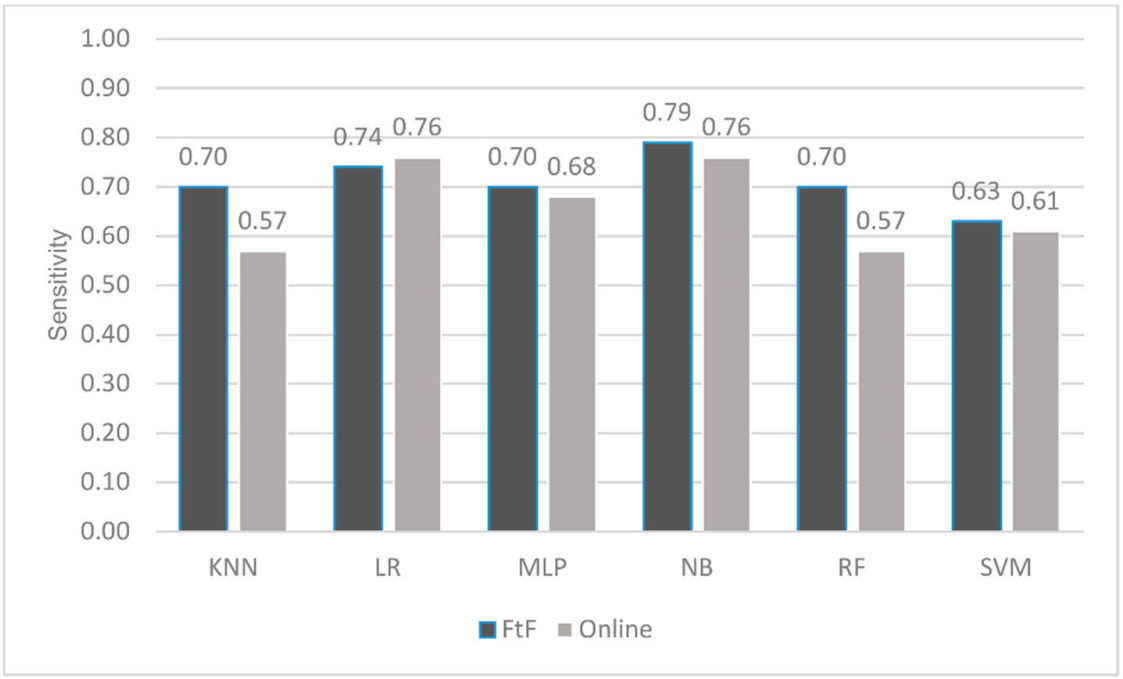

Figure 5. WED course: prediction of at-risk students (sensitivity scores) by instructional mode.

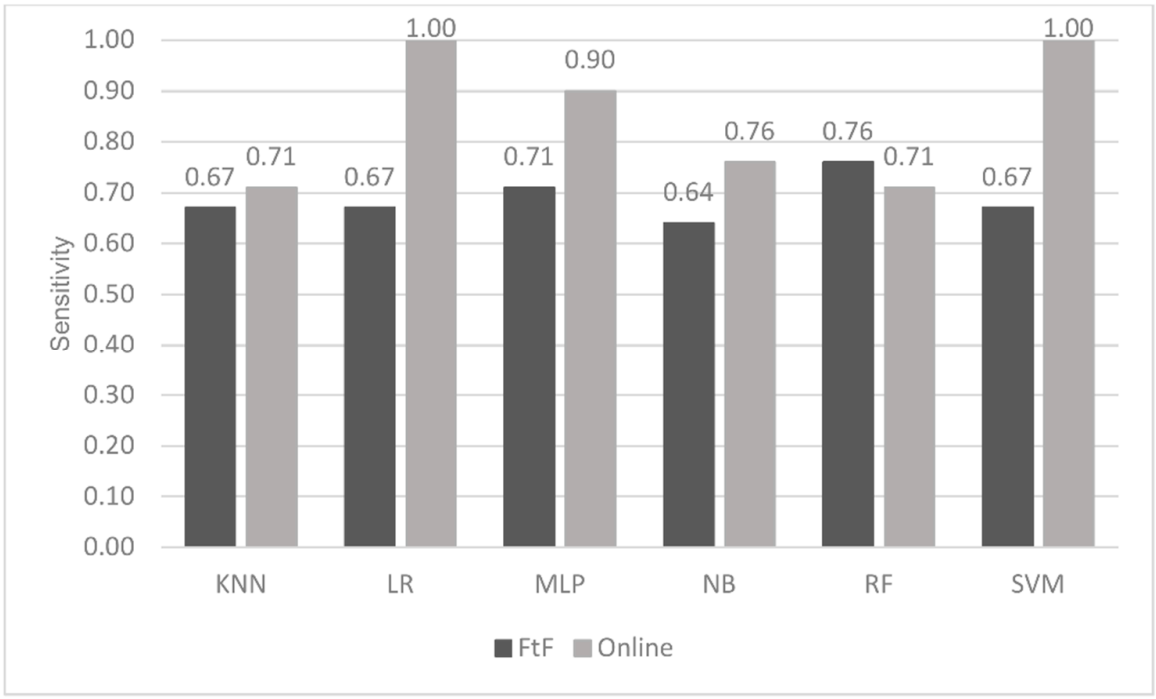

Figure 6. STA course: prediction of at-risk students (sensitivity scores) by instructional mode.

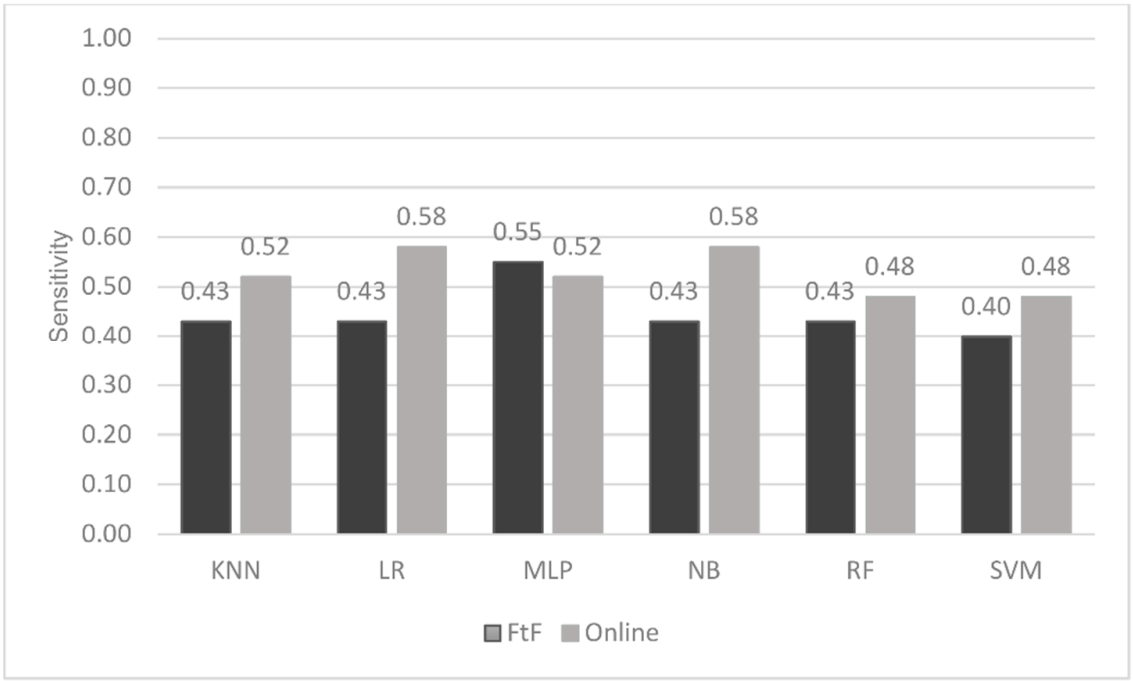

Figure 7. PSY course: prediction of at-risk students (sensitivity scores) by instructional mode.

4. Discussion

The findings of the present research can be summarized in two points. First, if the overall predictive validity of machine learning algorithms is of interest, our evidence suggests that they tend to yield a higher predictive validity (as indexed by sensitivity scores) in face-to-face than in online classes. Is this higher predictive validity due to changes in the way that students approach the curriculum of a course? Is it due to changes in the way that educators deliver content and/or assess learning? We explicitly selected courses that were taught by the same experienced educators and followed the same curriculum

requirements online and face-to-face. Educators' self-reports did not illustrate that the standards of early formative assessment (i.e., the first assignment and test) were changed between face-to-face and online courses. Yet, the lower predictive validity of early formative assessment online indicated that these measures were less useful to both educators and students when embedded in online classes. Inquiries through focus groups and informal exchanges with both students and faculty did not clarify this puzzle, mostly leading to the acknowledgment by both parties that the adaptation to online courses was more challenging for students than the adaptation to face-to-face courses. The following themes were frequently mentioned by students and corroborated by faculty: more time devoted to understanding how to navigate materials posted online (Blackboard) and how to use them; feelings of isolation and perceived distance from the instructor; and fewer opportunities for informal interactions with the instructor and classmates. Instructors reported more initial inquiries regarding course contents and requirements online than face-to-face, often noting that students who were accustomed to on-campus classes required more time to navigate and feel comfortable with the online mode. Thus, qualitative evidence seemed to point to educators who, aware of students' adaptation challenges, might have become tacitly more lenient when assessing performance, even though they purported not to have changed their standards of assessment. However, this pattern may offer a misleading picture since the variables course type and mode of instruction interacted.

Second, algorithms such as KNN and RF were consistently better predictors of at-risk students in face-to-face courses in the humanities and social sciences, such as ACS, WCO, and WED, whereas they were better predictors online when the course covered mathematical knowledge and skills (STA). The flexibility of KNN and RF may be particularly useful if the SARS-CoV-2 virus, which causes the COVID-19 disease, persists in affecting people's lives, thereby forcing university administrators to continue relying on the online mode or adopt hybrid modes for courses that are offered at their institutions.

As for the differences in algorithms' predictive validity between online and face-to-face, the interaction of course type and mode of instruction was not entirely clarified by the self-reports that were produced by students and educators; however, a consistent theme emerged. At the very beginning of the semester, some classes were reported by students as likely to be more difficult online (e.g., STA), whereas others (e.g., ACS and WED) were seen as potentially easier online. These biases could be thought of as capable of shaping students' behavior inside and outside the virtual classroom. For instance, consider that the anxiety that was experienced by female students towards Math courses increased considerably when the courses were delivered online. Increased anxiety might have led students to pay more attention in class and devote more time to class activities across the entire semester, thereby potentially leading to three interrelated outcomes: (a) enhancing their performance across the entire semester; (b) rendering a view, at the end of the semester, of the online STAT course as easier to manage than expected; and (c) making even initial formative assessment measures more likely to reflect overall course performance online than face-to-face. Instead, the initial expectation of easier online courses, coupled with educators' purported leniency that was driven by the opposite expectation, might have had a quite different impact. Namely, expectations might have unnecessarily lessened students' effort towards class activities, and relaxed educators' grading standards to ease students' adaptation to online courses, thereby making initial formative assessment less likely to reflect overall course performance online.

Our research suggests that particular machine learning algorithms can be used to make informed predictions regarding students' performance attainment, but the predictive validity of each algorithm has to be first assessed as a function of two important variables: course type and instructional mode. Our study adds to the growing body of grade prediction studies that rely on machine learning algorithms [8,10,43–45] by pointing to the relevance of such variables to interventions that are intended to foster academic success in an understudied student population. In our research, the latter is represented by young women of college age from a society that has only recently implemented and enforced

gender equity guidelines. Our research also contributes to the extant literature by relying on a subjective criterion of effectiveness that is produced by faculty with direct experience in teaching the courses that are included in our sample. Too often, studies examining the predictive validity of different algorithms have focused on relative comparisons but have failed to give readers an idea of how to conceptualize a desirable outcome for the actual situations/conditions they face.

5. Conclusions

We believe that research in educational settings should be motivated by the intention to improve participants' existing conditions [46–48]. As such, we subscribe to the main tenets of action research according to which the aim of a research project is practical. Namely, it is to identify a problem, condition, or situation; propose and implement a solution that is intended to bring improvement to the very people who participate in the research; assess the effectiveness of the solution; and either (a) start from the beginning if the outcome is unsatisfactory, or (b) broaden the reach of the purported solution if the outcome is within the expected parameters [49]. Thus, our goal is to rely on machine learning algorithms as feedback tools for students to assess their learning, and for faculty to assess their teaching. If improvement is needed, such tools can also inform the nature of the changes to be implemented in a university's curriculum and instruction.

To this end, we recognize that too often, algorithms and related data mining techniques are mainly accessible to educators who possess a background in computer science, and, more precisely, in artificial intelligence [15]. Educators with diverse backgrounds are frequently unable to access data mining techniques, thereby preventing their application to a much wider educational field. Our goal at the selected institution is to offer faculty with backgrounds outside computer science access to such techniques via workshops and mentorship efforts. Specifically, we plan to develop an easy-to-use early-warning system that relies on KNN to identify students at risk in particular courses, depending on the mode of instruction that is used to deliver their content. Currently, we have data supporting the effectiveness of an early-warning system using KNN in face-to-face ACS, WCD, and WED courses, as well as in online STA. The choice of KNN is based on its yielding the best relative performance for different modes of instruction in the courses that are selected for our study. However, the dismal performance of KNN, along with that of all the other algorithms in PSY, suggests that the early formative assessment measures of this course need to be examined closely to determine whether they indeed fit the learning outcomes of the course. A similar examination of the early formative assessment measures of the other courses that were selected for the present examination may also be warranted to improve their predictive validity. Of course, the predictive validity of KNN for at-risk students in general education courses that were not included in the present research, especially those involving natural sciences and math, will also need to be examined.

Existing early warning systems for the identification of at-risk students may be too general and thus become poor predictors of academic difficulties in particular courses. They also often rely on norm-referenced scores, which take into account how other students perform, instead of criterion-referenced scores, which consider how students perform relative to the learning outcomes that are set by the curriculum of different courses (as illustrated by summative assessment measures). Thus, an algorithm, such as KNN, that uses criterion-based scores to predict at-risk students may be seen as particularly helpful by educators. Indeed, it has the ability not only to effortlessly identify students who experience difficulties, but also to inform revisions of the curriculum and assessment protocol to ensure adequate coverage of the learning outcomes of a course. The attainment of such learning outcomes is particularly critical to general education courses, which lay the foundations for academic success in major-specific courses [50–53].

Two important lessons that are learned from our investigation and the pertinent extant literature are reminders of the limitations of our study. First, machine learning solutions for grade prediction, albeit most useful early in a course, may require adaptation

to the particular student population and the academic environment that an educator or administrator has selected for assessment and intervention. Namely, each educational setting may have features that are common to other educational settings (e.g., reliance on the synchronous online mode), thereby allowing a technical solution to be transferred, and features that may be unique to it. Unique features introduce uncertainties by questioning the transfer of the solution to other settings. For instance, the students selected for the present investigation are exposed to undergraduate courses guided by the principles of student-centered instruction intended to promote deep learning at the expense of rote learning. It is unclear whether our findings generalize to students who are exposed to a different type of instructional principles. Thus, uncertainties, defined by their statistical properties (e.g., parametric or non-parametric factors), and origin (e.g., internal to the learner or environmental, etc.), are likely to depend on the student populations that educators select for assessment and intervention, and the specific factors they deem relevant. Second, a lesson that is learned from the extant literature is that, in a vast array of problem domains, computational models are relentlessly evolving and are often so complex that educators without knowledge in computer science are left out. Innovative algorithmic solutions may be applied to the grade prediction needs of an institution and its faculty [54,55] if the unique properties of the grade-prediction conundrum in any given setting (including students and academic environment) are integrated and computing resources for training and inference purposes are made available e.g., [56]. However, such models also need to become more transparent and user-friendly for non-experts to ensure broad and reliable adoption [15]. Our paper is a modest call for action that specifically targets non-expert educators and administrators to approach the field of machine learning due to its potential benefits to the quality of learning and teaching.

Author Contributions: All authors M.A.E.P., E.N., M.N., I.D., H.A. and M.A., contributed equally to the research, including conceptualization, methodology, formal analysis, data curation, writing—original draft preparation, writing—review and editing, and project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are available upon request.

Acknowledgments: We thank the members of the Cognitive Science Cluster for their feedback.

Conflicts of Interest: The authors have no conflict of interest to declare that are relevant to the content of this article.

References

1. Lei, S.A.; Lei, S.Y. General education curricula affecting satisfaction and retention of undergraduate students: A review of the literature. *Education* **2019**, *139*, 197–202.
2. Warner, D.B.; Koeppel, K. General education requirements: A comparative analysis. *J. Gen. Educ.* **2009**, *58*, 241–258. [CrossRef]
3. Millea, M.; Wills, R.; Elder, A.; Molina, D. What matters in college student success? Determinants of college retention and graduation rates. *Education* **2018**, *138*, 309–322.
4. AbdelSalam, H.M.; Pilotti, M.A.E.; El-Moussa, O.J. Sustainable math education of female students during a pandemic: Online versus face-to-face instruction. *Sustainability* **2021**, *13*, 12248. [CrossRef]
5. Daniel, A.M. Identification of skill-appropriate courses to improve retention of at-risk college freshmen. *J. Coll. Stud. Retent. Res. Theory Pract.* **2020**, *24*, 126–143. [CrossRef]
6. Hannafin, K.M. Technology and the support of at-risk students. *J. Gen. Educ.* **1991**, *40*, 163–179. Available online: https://www.jstor.org/stable/27797135 (accessed on 10 December 2021).
7. Alyahyan, E.; Düştegör, D. Predicting academic success in higher education: Literature review and best practices. *Int. J. Educ. Technol. High. Educ.* **2020**, *17*, 3. [CrossRef]
8. Bujang, S.D.A.; Selamat, A.; Ibrahim, R.; Krejcar, O.; Herrera-Viedma, E.; Fujita, H.; Ghani, N.A.M. Multiclass prediction model for student grade prediction using machine learning. *IEEE Access* **2021**, *9*, 95608–95621. [CrossRef]
9. Rastrollo-Guerrero, J.L.; Gomez-Pulido, J.A.; Durán-Domínguez, A. Analyzing and predicting students' performance by means of machine learning: A review. *Appl. Sci.* **2020**, *10*, 1042. [CrossRef]
10. Huang, S.; Fang, N. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Comput. Educ.* **2013**, *61*, 133–145. [CrossRef]

11. Bydžovská, H. Are collaborative filtering methods suitable for student performance prediction? In *Progress in Artificial Intelligence: Lecture Notes in Computer Science*; Pereira, F., Machado, P., Costa, E., Cardoso, A., Eds.; Springer: New York, NY, USA, 2015; Volume 9273, pp. 230–425. [CrossRef]
12. Fahd, K.; Venkatraman, S.; Miah, S.J.; Ahmed, K. Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. *Educ. Inf. Technol.* **2021**, *27*, 3743–3775. [CrossRef]
13. Albreiki, B.; Zaki, N.; Alashwal, H. A systematic literature review of student' performance prediction using machine learning techniques. *Educ. Sci.* **2021**, *11*, 552. [CrossRef]
14. Kanetaki, Z.; Stergiou, C.; Bekas, G.; Jacques, S.; Troussas, C.; Sgouropoulou, C.; Ouahabi, A. Grade Prediction Modeling in Hybrid Learning Environments for Sustainable Engineering Education. *Sustainability* **2022**, *14*, 5205. [CrossRef]
15. Lee, K.M.; Yoo, J.; Kim, S.W.; Lee, J.H.; Hong, J. Autonomic machine learning platform. *Int. J. Inf. Manag.* **2019**, *49*, 491–501. [CrossRef]
16. Gilovich, T.; Griffin, D.; Kahneman, D. *Heuristics and Biases: The Psychology of Intuitive Judgment*; Cambridge University Press: Cambridge, UK, 2002.
17. Meegan, D.V. Zero-sum bias: Perceived competition despite unlimited resources. *Front. Psychol.* **2010**, *1*, 191. [CrossRef]
18. Lu, O.H.T.; Huang, A.Y.Q.; Huang, J.C.; Lin, A.J.Q.; Ogata, H.; Yang, S.J.H. Applying learning analytics for the early prediction of students' academic performance in blended learning. *Educ. Technol. Soc.* **2018**, *21*, 220–232.
19. Quinn, R.J.; Gray, G. Prediction of student academic performance using Moodle data from a Further Education setting. *Ir. J. Technol. Enhanc. Learn.* **2020**, *5*, 1–19. [CrossRef]
20. Gonzalez, T.; de la Rubia, M.A.; Hincz, K.P.; Comas-Lopez, M.; Subirats, L.; Fort, S.; Sacha, G.M. Influence of COVID-19 confinement on students' performance in higher education. *PLoS ONE* **2020**, *15*, e0239490. [CrossRef]
21. Iglesias-Pradas, S.; Hernández-García, Á.; Chaparro-Peláez, J.; Prieto, J.L. Emergency remote teaching and students' academic performance in higher education during the COVID-19 pandemic: A case study. *Comput. Hum. Behav.* **2021**, *119*, 106713. [CrossRef]
22. Engelhardt, B.; Johnson, M.; Meder, M.E. Learning in the time of Covid-19: Some preliminary findings. *Int. Rev. Econ. Educ.* **2021**, *37*, 100215. [CrossRef]
23. El Said, G.R. How Did the COVID-19 Pandemic affect higher education learning experience? An empirical investigation of learners' academic performance at a university in a developing country. *Adv. Hum. Comput. Interact.* **2021**, *2021*, 6649524. [CrossRef]
24. Elzainy, A.; El Sadik, A.; Al Abdulmonem, W. Experience of e-learning and online assessment during the COVID-19 pandemic at the College of Medicine, Qassim University. *J. Taibah Univ. Med. Sci.* **2020**, *15*, 456–462. [CrossRef] [PubMed]
25. Zheng, M.; Bender, D.; Lyon, C. Online learning during COVID-19 produced equivalent or better student course performance as compared with pre-pandemic: Empirical evidence from a school-wide comparative study. *BMC Med. Educ.* **2021**, *21*, 495. [CrossRef] [PubMed]
26. Foo, C.C.; Cheung, B.; Chu, K.M. A comparative study regarding distance learning and the conventional face-to-face approach conducted problem-based learning tutorial during the COVID-19 pandemic. *BMC Med. Educ.* **2021**, *21*, 141. [CrossRef]
27. Hussain, A.; Chau, J.; Bang, H.; Meyer, L.; Islam, M. Readiness, reception, and performance of students in a communications course delivered amid the pandemic. *Am. J. Pharm. Educ.* **2021**, *85*, 8617. [CrossRef]
28. Boston, C. The concept of formative assessment. *Pract. Assess. Res. Eval.* **2002**, *8*, 9. [CrossRef]
29. Ahmed, W. Women empowerment in Saudi Arabia: An analysis from education policy perspective. *Middle East Int. J. Soc. Sci.* **2020**, *2*, 93–98.
30. Barry, A. Gender differences in academic achievement in Saudi Arabia: A wake-up call to educational leaders. *Int. J. Educ. Policy Leadersh.* **2019**, *15*, 17. [CrossRef]
31. Pilotti, M.A.E. What lies beneath sustainable education? Predicting and tackling gender differences in STEM academic success. *Sustainability* **2021**, *13*, 1671. [CrossRef]
32. Syed, J.; Ali, F.; Hennekam, S. Gender equality in employment in Saudi Arabia: A relational perspective. *Career Dev. Int.* **2018**, *23*, 163–177. [CrossRef]
33. Hvidt, M. Economic diversification and job creation in the Arab Gulf countries: Applying a value chain perspective. In *When Can Oil Economies be Deemed Sustainable?* Luciani, G., Moerenhout, T., Eds.; Palgrave Macmillan: London, UK, 2021; pp. 281–300.
34. Sadler, D.R. Interpretations of criteria-based assessment and grading in higher education. *Assess. Eval. High. Educ.* **2005**, *30*, 175–194. [CrossRef]
35. Marbouti, F.; Diefes-Dux, H.A.; Madhavan, K. Models for early prediction of at-risk students in a course using standards-based grading. *Comput. Educ.* **2016**, *103*, 1–15. [CrossRef]
36. Romero, C.; Ventura, S. Educational data mining: A review of the state of the art. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2010**, *40*, 601–618. [CrossRef]
37. Zhang, Z. Introduction to machine learning: K-nearest neighbors. *Ann. Transl. Med.* **2016**, *4*, 136. [CrossRef] [PubMed]
38. Kleinbaum, D.G.; Dietz, K.; Gail, M.; Klein, M.; Klein, M. *Logistic Regression*; Springer: New York, NY, USA, 2002.
39. Kruse, R.; Mostaghim, S.; Borgelt, C.; Braune, C.; Steinbrecher, M. (Eds.) Multi-layer perceptrons. In *Computational Intelligence*; Springer: New York, NY, USA, 2022; pp. 53–124.
40. Xu, S. Bayesian Naïve Bayes classifiers to text classification. *J. Inf. Sci.* **2018**, *44*, 48–59. [CrossRef]

41. Boulesteix, A.L.; Janitza, S.; Kruppa, J.; König, I.R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2012**, *2*, 493–507. [CrossRef]
42. Bhavsar, H.; Panchal, M.H. A review on support vector machine for data classification. *Int. J. Adv. Res. Comput. Eng. Technol.* **2012**, *1*, 185–189.
43. Pereira, F.D.; Oliveira, E.H.; Fernandes, D.; Cristea, A. Early performance prediction for CS1 course students using a combination of machine learning and an evolutionary algorithm. In Proceedings of the 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), Maceio, Brazil, 15–18 July 2019; Volume 2161, pp. 183–184. [CrossRef]
44. Friedman, J.H.; Bentley, J.L.; Finkel, R.A. An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Softw. TOMS* **1977**, *3*, 209e226. [CrossRef]
45. Hung, J.L.; Wang, M.C.; Wang, S.; Abdelrasoul, M.; Li, Y.; He, W. Identifying at-risk students for early interventions—A time-series clustering approach. *IEEE Trans. Emerg. Top. Comput.* **2015**, *5*, 45–55. [CrossRef]
46. Sáez Bondía, M.J.; Cortés Gracia, A.L. Action research in education: A set of case studies? *Educ. Action Res.* 2021; *latest articles*. [CrossRef]
47. Ali, A.D. Implementing action research in EFL/ESL classrooms: A systematic review of literature 2010–2019. *Syst. Pract. Action Res.* **2020**, *33*, 341–362. [CrossRef]
48. Lufungulo, E.S.; Mambwe, R.; Kalinde, B. The meaning and role of action research in education. *Multidiscip. J. Lang. Soc. Sci. Educ.* **2021**, *4*, 115–128.
49. Jacobs, S.D. A history and analysis of the evolution of action and participatory action research. *Can. J. Action Res.* **2018**, *19*, 34–52. [CrossRef]
50. Benander, R.; Lightner, R. Promoting transfer of learning: Connecting general education courses. *J. Gen. Educ.* **2005**, *54*, 199–208. [CrossRef]
51. Aloi, S.L.; Gardner, W.S.; Lusher, A.L. A framework for assessing general education outcomes within the majors. *J. Gen. Educ.* **2003**, *52*, 237–252. [CrossRef]
52. Jeske, J. Nurturing rich general education courses. *J. Gen. Educ.* **2002**, *51*, 103–114. [CrossRef]
53. Landon-Hays, M.; Peterson-Ahmad, M.B.; Frazier, A.D. Learning to Teach: How a Simulated Learning Environment Can Connect Theory to Practice in General and Special Education Educator Preparation Programs. *Educ. Sci.* **2020**, *10*, 184. [CrossRef]
54. Tutsoy, O.; Polat, A.; Çolak, Ş.; Balikci, K. Development of a multi-dimensional parametric model with non-pharmacological policies for predicting the COVID-19 pandemic casualties. *IEEE Access* **2020**, *8*, 225272–225283. [CrossRef]
55. Tutsoy, O. COVID-19 epidemic and opening of the schools: Artificial intelligence-based long-term adaptive policy making to control the pandemic diseases. *IEEE Access* **2021**, *9*, 68461–68471. [CrossRef]
56. Kang, S.J.; Lee, S.Y.; Lee, K.M. Performance comparison of OpenMP, MPI, and MapReduce in practical problems. *Adv. Multimed.* **2015**, *2015*, 7. [CrossRef]