Rakibul Hasan* and Mario Fritz

Understanding Utility and Privacy of Demographic Data in Education Technology by Causal Analysis and Adversarial-Censoring

Abstract: Education technologies (EdTech) are becoming pervasive due to their cost-effectiveness, accessibility, and scalability. They also experienced accelerated market growth during the recent pandemic. EdTech collects massive amounts of students' behavioral and (sensitive) demographic data, often justified by the potential to help students by personalizing education. Researchers voiced concerns regarding privacy and data abuses (e.g., targeted advertising) in the absence of clearly defined data collection and sharing policies. However, technical contributions to alleviating students' privacy risks have been scarce. In this paper, we argue against collecting demographic data by showing that gender—a widely used demographic feature—does not *causally* affect students' course performance: arguably the most popular target of predictive models. Then, we show that gender can be inferred from behavioral data; thus, simply leaving them out does not protect students' privacy. Combining a feature selection mechanism with an adversarial censoring technique, we propose a novel approach to create a 'private' version of a dataset comprising of fewer features that predict the target without revealing the gender, and are interpretive. We conduct comprehensive experiments on a public dataset to demonstrate the robustness and generalizability of our mechanism.

Keywords: Privacy, Educational data mining, Learning analytics, Educational Technology

DOI Editor to enter DOI
Received ..; revised ..; accepted ...

***Corresponding Author: Rakibul Hasan:** CISPA Helmholtz Center for Information Security, E-mail: rakibul.hasan@cispa.de
Mario Fritz: CISPA Helmholtz Center for Information Security, E-mail: fritz@csipa.de

1 Introduction

Many forms of educational technologies (EdTech for short)—ranging from simple online portals to sophisticated AI-enabled applications for smart learning, remote tutoring, and proctoring—are now becoming ubiquitous in educational institutes across the world. The recent pandemic has necessitated inventing new technologies as well as adapting existing ones to fulfill educational purposes. Although many of these technologies were invented as a response to an emergency, they will remain in use in the post-pandemic world, inspire other novel technologies in this domain, and further accelerate the growth of EdTech's worldwide market [53, 72]. According to one estimate, EdTech growth will nearly double by 2025, with an estimated expenditure of $404 billion [47].

The potential for a huge market has attracted many technology development companies to develop web-based interfaces as well as applications for personal computers and mobile platforms [1, 13, 16]. Simultaneously, tech giants like Google and Facebook are creating customized versions of their products and services targeting educational institutes, teachers, and learners at all levels. Google Classroom [35] has doubled its number of users since the pandemic began [34, 58]. Facebook has recently released *Facebook for Education* [27]—a collection of courses, tools, and resources for learners in both K-12 and higher levels of education. EdTech provided by Google and Facebook are hailed for their usability and are lucrative to institutional administrations for being cheaper than other alternatives [4, 58]; consequently, they are being increasingly integrated into educational processes and playing a much larger role in learning and teaching practices than simply being a tool to deliver education [37].

EdTech does not only provide interfaces to facilitate educational processes but also continuously collects multimodal data about students' interactions with them. For example, learning management systems (e.g.,

Canvas[1]) log students' actions when they access course materials using the portal, and remote proctoring systems (e.g., Proctorio[2]) continuously collect audio and video data during remote exams. The availability of such big data has inspired new research directions, such as Educational Data Mining (EDM) and Learning Analytics (LA), and researchers have spent enormous efforts to harness the power of this data using machine learning and data mining techniques. Some of the most popular prediction tasks using educational data include predicting students' course performance (e.g., [91]), identifying students at risk of dropout (e.g., [28], clustering students in terms of what strategic behaviors they adopt (e.g, [3]), and detecting cheating behaviors during remote exams (e.g., [6]). See recent survey papers (e.g., [21, 51, 71, 78]) for details on the current state of the art machine learning models for each prediction task and what types of data are combined (e.g., demographics, background, and log data) to train those models.

Not surprisingly, concerns regarding risking students' privacy by collecting and storing such vast amount of data are growing, particularly fueled by past data breaches [57], as well as EdTech companies' data collection and sharing practices [58, 63, 94]. For example, Klose et al. lists several data breaches and hacks involving companies providing EdTech, leading to millions of students having their identifying and other sensitive data exposed [57]. The data were sold in the black market, used for tax fraud or unlawfully extorting money in exchange for financial aid, and in one case, the breach led to a student's death [57]. Besides data breaches, EdTech companies' practices of data collection, storage, and sharing with 'business partners' were called into question by numerous researchers and privacy activists [63, 94]. Google faced lawsuits for using students' data for advertising purposes, even though their terms of use did not allow such usages [43]. Google classroom also provides APIs (application programming interfaces) for third-party developers, who can integrate their applications to Google classroom and may have access to students' data [37]. Facebook for education is integrated with other products owned by Facebook, such as Instagram, and Facebook ad for education [26]. Triangulating cross-platform data facilitates identification of students even if the original data were deidentified, as well as inferring students' demographic information and behavioral patterns, which might be used for profiling based on demographic characteristics, threatening students' privacy and autonomy.

While public sharing of students' (anonymized) data for research purposes can accelerate scientific progress, it may also increase privacy risks [15, 95]. This heightened fear of violating students' privacy, coupled with the instantiating of stricter privacy laws (such as the General Data Protection Regulations, GDPR, in Europe), discouraged publicizing learning analytics datasets [46, 55]. Indeed, in a recent survey of public MOOC datasets, Lohse, McManus, and Joyner noted that most research papers on learning analytics experiment on proprietary datasets, and no dataset has been made public since 2016 [62]. But concerns regarding data collection and their (improper) uses by EdTech companies remain. Harvesting as much data as possible, including demographic information such as gender and age, is usually justified by the promise of better learning analytics that may improve the learners' experience. Students' demographic attributes are privacy sensitive as they can be used to profile students and target them for surveillance (section 2.2). While such aggressive mining of students' data has been heavily criticized, technical contributions from the research community to reduce the amount of data collected have been inadequate.

In this paper, we make a case against collecting students' demographic attributes that are privacy sensitive (e.g., gender), and using them to train predictive models. We also propose technical means to prevent inferring such attributes from students' behavioral data while allowing to build learning analytical models. Concretely, we make the following three contributions:

- Using causal inference methodologies, we demonstrate that students' gender—a widely used feature to train models to predict, e.g., students' course performance—does not have any *causal* effect on students' course performance. Our analyses suggest that gender's predictive value may be due to spurious correlations with the outcome(s) and thus not causally relevant to predictive modeling of students' performance in educational courses.
- We demonstrate that students' gender can be inferred based on how they interacted with course materials. Thus, simply not collecting such attributes is not enough to prevent profiling students based on demographic factors. We evaluate an adversarial training-based censoring technique to remove gender information from students' behavioral data. This technique transforms input features (i.e., behavioral data) into an intermediate representation. The adversarial training procedure removes infor-

1 https://www.instructure.com/canvas

2 https://proctorio.com

mation about gender from the transformed features, but preserves information about the target (e.g., course performance) outcome. Thus, these features can be used to train new predictive models without risking students privacy.

– We make methodological contributions in identifying a subset of *interpretable* features that are enough to predict students' course performance with high accuracy while preventing gender inference. We formulate the feature selection problem as a combinatorial optimization problem. The exact solution of that problem is intractable. We propose a novel approach to obtain an approximate solution where we combine the adversarial censoring mechanism with a feature selection technique that was implemented as the input layer of a deep neural network-based predictive model. We devised a custom penalty function to apply in the input layer; the function fulfills dual purposes: i) it constrains some model parameters in the input layer to take the on value of zero so that associated features will be left out, and ii) it constrains the remaining features to take on the value of one so that those features combine themselves in interpretable ways. Our method retained a few features, which can predict performance with high accuracy and do not reveal gender information. These features are also interpretable: it is easy to review them and how they combine to gain an intuitive understanding of what features (or combinations of features) are important predictors for a given target task. Additionally, educational experts may audit the selected features and examine if they match their expectations regarding the association between students' behaviors and performance. Finally, since the size of the final set of features is small (with high predictive power), our approach permits building simpler models (e.g., logistic regression), which are easier to understand and explain. We make our methods' implementation public.[3]

We employ our methods on a learning analytics dataset containing students' interactions with course materials along with some demographic attributes (see section 3.1 for the details on the dataset). Our findings strengthen the case for data minimization in this emerging field of building AI-enabled educational technologies. In particular, educational institutes may re-evaluate the necessity to use demographic information (since they may not have a causal effect on the outcome) and may prohibit their collection by EdTech companies while initiating a service contract. Our technical contributions further help to minimize the amount of data collected for predictive purposes. Finally, our approach results in an intuitive feature set and simpler models, enhancing the explainability of predictive models in education technologies.

2 Literature review

2.1 Educational data mining and learning analytics

As web-based technologies such as learning management systems (LMS) were integrated to facilitate educational processes, they opened the door to huge scale learning analytics (LA)—collecting and analyzing students' data to better understand and optimize the learning process and the environment. Education technologies are increasingly getting AI-enabled, where predictive models that are trained on students' data are integrated with these tools. Some of the most common predictive tasks include course performance prediction (see recent research [50, 91] and survey papers [42, 54, 71], predicting the probability of dropping out from a course (see [28, 50, 87] and survey papers [65, 76], students' attention and behavior prediction [54, 73, 97], as well as detecting students' cheating behaviors during remote exams [17, 88].

2.2 Concerns for students' privacy

Concerns related to students' privacy and autonomy have been at the center of the debate on whether EdTech should be deployed in educational institutions [5, 8, 48, 52, 82]. Students' data may be shared among multiple stakeholders including educational institutes and companies providing EdTech [82]. Additionally, third-party applications are integrated with LMS (such as Canvas) for "frictionless" data collection and sharing [64]. Fiebig et al. reported that higher educational institutes are increasingly relying on third-party vendors to collect and store institutional data [29]. The authors noted that this over-reliance may have adverse consequences that go beyond individuals' privacy. In their recent paper, Marachi and Quill cautioned that educational in-

3 https://github.com/rakib062/EdTech-PETS

stitutions are ill-equipped to protect students from data harvesting and exploitation [64]. In line with the seminal work on a taxonomy of privacy violations by Solove [83], Reidenberg and Schaub identified different short- and long-term harms that may be caused by mining students' data and then connected them to the taxonomy [79]. The harms may be caused by excessive information collection, processing, dissemination, and invasion [83], and include students' identification, profiling, surveillance, and online harassment [79].

Many educational institutes and EdTech providers anonymize data before collecting, processing, or publicizing data to protect students' identities. But anonymized data from multiple sources can be combined to identify students. Yacobson et al. mined temporal patterns from de-identified student log data and identified the physical classes and schools with the help of publicly available school data [95]. Chen et al. identified 42% of the learners in a dataset on social media [15]. Students' demographic data, combined with other data sources, may reveal students' identity [86, 95]. In the T3 project, students' anonymized Facebook profile data identified many individuals as being the only Harvard freshman student from a certain state or county [57]. Gursoy et al. classified gender and age as quasi-identifiable attributes that can be combined with other information to identify an individual, and GPA and enrolled courses as sensitive data that students do not want to share with third parties [39]. Many researchers thus advocate not publishing students' demographic information and limiting the amount of metadata or additional information when publicizing anonymized student data for research [11, 57, 95], i.e., following a data-minimization principle [60].

2.3 Using demographic data in EDM

While demographic attributes increase students' risks to privacy harm, they are frequently used as features in predictive models. Paquette et al. surveyed 385 research papers on educational data mining that were published in the previous five years (2015–2019, inclusive) and found that 15% of these publications used demographic variables in EDM [74]. However, the effects of demographic factors (e.g., students' gender and age) on outcome variables (such as course performance and dropout probability) have been largely minor and sometimes inconsistent. For example, Leal et al. found a negative effect of gender (female) on course completion in one sample of data, but the effect was positive in another sample [36]. Focusing on the students who completed an online course, Chen et al. found no difference based on gender and age in learning behaviors [14]. Demographic factors' predictive power may be attributed to spurious correlations with outcome variables; the correlations may vary based on the data sample, causing instability in trained models. Evidence of only spurious correlation between demographic attributes on outcome variables (or equivalently, the absence of any causal effects of the former on the later) strengthen the case against collecting students' demographic data. Unfortunately, we did not find any published research investigating these phenomena.

2.4 Privacy preserving learning analytics

Concerns regarding privacy harms and fair/ethical use of students' data resulted in several initiatives to create or update local and international regulations [46], which may have attenuated the initial excitement regarding learning analytics [23]. Drachsler and Greller provided an eight-point checklist to follow to facilitate a trusted implementation of Learning Analytics [23]. Reidenberg and Schaub proposed policy recommendations for built-in privacy and accountability in learning technologies [79]. Inspired by methodologies in user-centered design, Ahn et al. engaged with K-12 educators to surface how privacy, transparency, and trust interplay in specific settings and how educational technologies can address these dynamics [2]. Technical contributions to achieve these goals, however, have been scarce; we could identify very few research contributions focusing on privacy-preserved learning analytics. Gursoy et al. provided a proof-of-concept implementation of a privacy-preserving interface to access students' data records [39]. Bosch et al. proposed a method to automatically redact students' private information (such as name, location, and contact number) from forum posts so that the sanitized data might be released for research purposes [10]. Recognizing the benefits of sharing students' data for research, Bautista and Inventado proposed publishing synthetic data created by generative adversarial networks (GANs) that were trained on real data [9]. Guo et al. proposed to use federated machine learning to build learning analytics using data from multiple educational institutes without actually sharing the data [38].

2.5 Censored representation learning and Constrained optimization

Different from the above-mentioned works, we aim to reduce privacy risks in learning analytics by censoring feature representations and removing features that are not essential for the target prediction task. We review prior works related to the techniques—adversarial training and constrained optimization—we used to achieve our goals.

Edwards and Storkey proposed an adversarial training scheme to transform intermediate feature representations in a way to remove associations between a (sensitive) attribute and the target variable (i.e., outcome variable of a model trained on the features) [24]. Later works used this idea for data with various modalities, including text [25], image [40], and sensor data [49]. Ganin and Lempitsky used adversarial training to remove information about the source from where data was generated [33]; the goal was to create a domain-invariant representation of the input features so that the model generalizes over multiple distributions of input data. The authors implemented the training scheme as a standalone neural network layer, named "gradient reversal layer" [33], which we use in our implementation of adversarial training.

We combined a constrained optimization technique with adversarial training to binarize our model's parameters. Forcing parameters to take only binary values (e.g., -1 or 1) was first proposed by Courbariaux et al. [19]; they set parameter values deterministically (e.g., using the sign function) or stochastically (e.g., using a probability function of the actual parameter value). This approach was later extended to also binarize the output of activation functions [18]. These prior works binarized parameters and activations to train models more efficiently [18, 19]. In contrast, we binarize parameters for feature selection. Moreover, unlike previous works, we used a penalty function to binarize parameters instead of setting binary values deterministically or stochastically.

3 Methods

We apply our methods on a learning analytics dataset containing students' demographic information, interactions with course materials, and course performance (section 3.1). We show that gender lacks a *causal* effect on course performance (section 3.3) following a matching-based causal inference procedure [85]. But simply leaving out gender information does not protect students' privacy: we built machine learning models that can infer students' gender from their behavioral patterns as logged by learning management systems (section 3.4). To prevent gender inference, we employed an adversarial training procedure that creates an intermediate representation of behavioral data by censoring gender information (section 3.5). Finally, we propose a penalty function and employ it in a constrained optimization setting to identify an interpretable feature subset that best balances privacy-utility trade-offs (section 3.6).

3.1 Dataset description

We used the Open University Learning Analytics Dataset [59] (OULAD)—one of the most widely used datasets in Educational Data Mining (EDM) and Learning Analytics (LA) research (see the paper by Waheed et al. [91] for a list of earlier works based on this dataset). OULAD was created by Open University, the largest distance learning institution in the United Kingdom, and contains information about 32,593 students who participated in 22 courses in 2013 and 2014. Courses typically lasted for approximately nine months, and students were assessed through several assignments and one final exam. OULAD contains students' demographic data (e.g., gender, age group, and the highest level of education), background information (e.g., whether a student took a course previously), and interactions with course materials through a virtual learning environment (VLE). The interactions were recorded as the number of clicks. Interactions were grouped in 20 categories, such as visiting course home- or sub-pages, completing quizzes, and participating in forum discussions. Students' final performance was grouped into four classes: distinction (3,024), pass (12,361), fail (7,052) and withdrawn(10,156). For a complete description of the dataset, please see the original publication [59].

3.2 Preprocessing and feature selection

Since some students participated in multiple courses (or repeated the same course more than once), we kept their interaction data for only one (randomly selected) course and removed data for all other courses. That step retained data from 28,785 students. Next, we removed students for whom no interaction data was recorded. Our final dataset contained interaction logs for 25,245

students, with the following course performance distribution: distinction (2,645), pass (10,883), fail (6,264) and withdrawn (9,043). We extracted summary statistics from this dataset to use as features to train models; the feature selection was done following prior works (based on OULAD and other datasets) as described below and summarized in Table 1.

Following Waheed et al. [91], we computed the total number of interactions with each type of course material. This step yielded 20 features, corresponding to the 20 types of interactions, that may collectively proxy for students' 'engagement' [36]. OULAD does not contain information about (study) sessions, which was defined by Tough [90] as "a period of time devoted to a cluster or sequence of similar or related activities, which are not interrupted much by other activities," and was identified as an important predictor of learning outcomes [22]. We imitate this idea by treating all interactions with a certain type of material that happened consecutively in a single day as belonging to a single 'session', indicating focused period on a single task. Then, we compute the total number of sessions for each type of material throughout a semester, resulting in another 20 features.

Persistence, which captures the extent to which a student continues an activity for a long period, is another long-studied characteristic of students [66]. Students' persistence has been measured in different ways: Whitehill et al. categorized students as persistent who interacted with the course at least once a week [92], while Crues et al. identified three levels of persistence (low, medium, and high) based on the number of weeks students worked in the course [20]. We took a more nuanced view of persistence by identifying patterns of uninterrupted interactions with course material. We define a 'Block' as the number of consecutive weeks a student had at least one session with study material every week, i.e., there was no 'break' in interactions. We counted the number of such blocks of interactions, the maximum, minimum, and average length (in weeks) of the blocks, and the variance of the block lengths.

Features related to course material coverage, e.g., the number or percentage of quizzes attempted, were found to be useful in predicting students' performance [12, 31, 61, 77]. Prior research also reported that features related to exercises were more predictive than other click-stream data [67]. Thus, for each type of course material, we computed what percentage of items in that category a student interacted with. For example, the percentage of external materials reviewed and the percentage of quizzes submitted. There were 20 such features; collectively they indicate the coverage of course material by a student. Combining all types of features, there were 65 features in total.

3.3 Estimating the *causal* effect of gender on course performance

Machine learning-based models that accompany EdTech are usually evaluated in terms of their predictive accuracy. Individual feature's predictive power is assessed by its strength of association with the outcome(s) of interest, regardless of its ability to change the outcome (i.e., causal effect). This approach is problematic when the predictors include sensitive demographic information, as this approach incentivizes collecting such sensitive data without carefully considering their actual (i.e., causal) influence on the outcome. In this section, we first estimate how strongly gender associates with course performance in the original sample of students. Then, using a matching-based procedure, we identify pairs of students with comparable characteristics. Finally, we investigate whether the previously observed association indicates a causal effect using the matched sample.

Estimating association between gender and performance. We assessed the association between gender and course performance by conducting Pearson's chi-squared test [93] that measures association between two categorical variables. Both gender and course performance were represented as binary variables (i.e., two categories). In our sample, 13,374 (52.9%) and 11,871 students self-identified as males and females, respectively, and 13,471 (53.4%) passed the course while 11,774 failed. We found significant association between gender and course performance: $\chi^2(1) = 79.85$, $p < 0.00001$, indicating that gender might be a useful predictor of course performance. Next, we investigate whether this association was causal.

Estimating gender's causal effect. Under the potential outcome framework of causal effects [81], gender's causal effect is the amount of change in the outcome (i.e., course performance) that will be observed by intervening on (i.e., changing) gender, holding everything else (i.e., covariates) constant. In other words, if pairs of students with different genders were 'similar' in terms of covariates (in this case behavioral patterns and engagement with course materials), but performed differently in the course, then the difference in their performance may be attributed to their gender. If there is no such difference, then gender lacks any causal effect on the course performance. Below, we describe the matching-

Feature type	Description	Example	Count
Engagement	Features indicating the total number of interactions with a certain course material such as quizzes.	Total number of interactions while submitting quizzes	20
Focus	Features indicating the number of sessions a student spent on a certain course material throughout the semester.	Number of sessions spent on the course resource page	20
Persistence	Number of 'blocks' (i.e., period of continued efforts); and the minimum, maximum, average, and variance of the blocks' length.	Visited the course homepage at least once every week for 4 weeks (i.e., block length is 4 weeks)	5
Content coverage	The percentage of course material in each of the 20 categories were covered by a student.	Submitted 80% of the quizzes	20

Table 1. Feature types and their descriptions with examples.

based procedure we followed to find pairs of 'similar' students who differ in gender.

Matching. Balancing covariates by matching pairs of 'similar' data samples is one of the widely used methods to make causal inferences from observational data [85]. Similarity between two data points can be estimated using distance-based or propensity score-based metrics. We measured similarity between pairs of students based on their behavioral patterns (i.e., the features computed from the interaction log). We employed Mahalanobis distance [85] in the feature space as the similarity measure (where smaller distance means more similar). Mahalanobis distance-based matching attempts to approximate fully blocked randomized experiments [44] and is preferred to other approaches (such as propensity score-based matching [56]); but this metric may suffer from the 'curse of dimensionality' if too many covariates (i.e., features) are used [85]. Thus, we reduced the dimensionality of the feature set to 13 as detailed below.

Reducing feature dimension. Using principal component analysis (PCA), we reduced the 65-dimensional features to 10-dimensional features. While conducting PCA, initially, we extracted 20 components, and then plotted the amount of variance explained by each component and the cumulative variance explained (i.e., scree plot [30]). We applied the elbow method [30] on this plot and retained 10 components. These 10 components are the new features, which are linear combinations of the original features and collectively explained 70% of the total variance in the original data. In addition to these 10-dimensional interaction features, three other features related to students' background and history were included as covariates in the matching procedure: the number of credits taken by a student, whether a student repeated a course, and the highest level of education a student had prior to enrolling in an online course.

Results of matching. After determining the 13 features to use for matching, each student was paired to another student (who differed in gender) based on the Mahalanobis distance between them in the 13-dimensional feature space (i.e., nearest neighbor matching). We used R package MatchIt [45] with repeated matching enabled (i.e., one control subject can be paired with multiple treated subjects) to reduce the aggregated distance between matched pairs. This procedure matched 5,027 male students to 11,871 female students. We assess the quality of the matched sample in the following paragraph.

Assessing the quality of the matched sample. Matching procedures aim to find pairs with a similar distribution of covariates, i.e., reduce the difference in covariate distributions of the two groups being matched. Thus, the quality of the matched samples can be assessed by examining how similar the covariate distributions of the two groups become after the matching. Standardized mean differences (SMD) in the covariates between two groups is a widely used metric for that purpose [98]. Figure 1 shows the SMDs in the behavioral patterns between male- and female-identifying students, both before and after matching. For all covariates, the standardized mean differences in features between the matched samples are close to zero and much smaller than the differences between the unmatched samples. Thus, the matching procedure successfully identified pairs of 'similar' students who differed only in gender.

Estimating the causal effect of gender. In the matched sample, students who were paired differed only in their gender. Thus, in this sample, any observed differences in course performance across genders could be attributed to gender (i.e., causal effect of gender). To investigate if the previously observed difference in course performance across genders is present in this matched sample, we again computed the association between

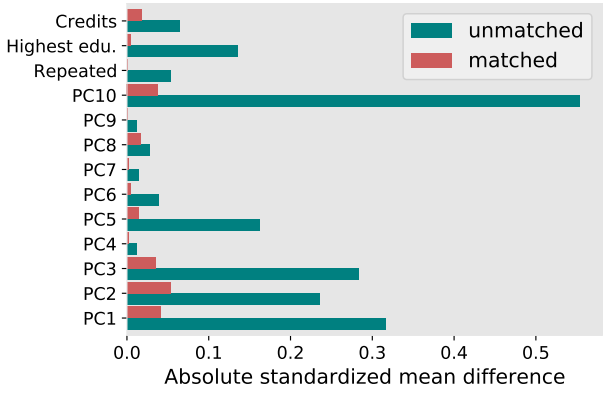

Fig. 1. Standardized mean differences in covariates across genders for the initial (unmatched) and matched samples. Note that, after matching, the standardized differences became much smaller than before and many of them came close to zero, indicating good covariate balance in the matched groups.

those two variables. This time, we used McNemar's test, since the sample is now paired [7]; and found no significant association between gender and course performance ($\chi^2(1) = 0.31$, $p > 0.05$). This result suggests that the previously observed association was not causal.

Conclusions. Based on the above findings, we hypothesize that association without causation is more common in learning analytics than one might realize. This is because causal inference methods such as matching, which are usually employed *to show the existence of a causal effect* (e.g., [41]), assume no unobserved confounders (i.e., the ignorability assumption [85]). This assumption may not hold in practice, but the effects may be accepted as *causal* if they are not very sensitive to unseen confounders (e.g., demonstrated through sensitivity analyses [80]). In our case, we demonstrated the *absence* of any causal effects of gender under the ignorability assumption. When this assumption does not hold or the identified effect is *weak*, it may be too sensitive to confounders, making association disguised as causation commonplace. Thus, we strongly argue against collecting demographic data without specifying their purposes (such as prediction) and verifying their (causal) relevance to those purposes. Additionally, one reason to build predictive models is to help students by interventions [69]. But, there may be no meaningful way to intervene on a variable that does not have a causal effect on the outcome (i.e., varying the variable does not change the outcome), further strengthening the case against demographic data collection.

3.4 Inferring gender from behavioral features

The previous section demonstrated a null causal effect of gender on performance, which could be used to argue against collecting students' demographic information. However, such information might be encoded in behavioral patterns and inferred from interaction data. In this section, we demonstrate how machine learning models can be used to infer students' gender based on how they interacted with course material. We demonstrate gender inference using both simple logistic regression and neural network-based models.

First, we trained a simple logistic regression model (using scikit-learn package [75]) with the 65 features described above, following a 10-fold cross-validation approach with 80%-20% train-test splits. Across the folds, the model had an average prediction accuracy of 73.3% on the test sets.

Next, to investigate if more complex models can predict gender with a higher accuracy, we trained a neural network with three hidden layers (with 30, 20, and 10 nodes, respectively). Hidden layer nodes had Rectified Linear Units (ReLU [70]) as the activation function, and the final layer had the sigmoid activation function. We implemented the model using pytorch[4] framework, all (hyper-)parameters (e.g., parameter initialization strategy) were kept in their default values. We used a 10-fold cross-validation method with 80%-20% train-test splits. The model was trained for 15 epochs using the same 65 features. Across the folds, the model had an average prediction accuracy of 76.2% on the test set, which is comparable to the much simpler logistic regression model.

Conclusions. The above results confirm that students' behavioral data can be used to infer their gender with high accuracy, even by using pretty simple machine learning models. This could bear potentially serious consequences for students' privacy and safety. The next section explores adversarial censoring as one way to prevent gender inference.

3.5 Preventing gender inference

Ideally, we would like to train predictive models that may benefit the students without risking their privacy. One approach to achieve this goal is creating a new rep-

4 http://pytorch.org/

resentation of the input features that contains information about the target outcome but leaves out gender-specific information. These transformed features can then be used to train new models to predict the intended outcome, but not to infer gender. We employ an adversarial training procedure that censors gender information from the input features while preserving information about the outcome. A deep neural network model is used in this step whose architecture is shown in Fig. 2. In this network, the feature encoder (the orange block in the diagram) transforms input features into an intermediate representation. There are two hidden layers in the encoder module, consisting of 20 and 10 nodes, respectively. There are two other modules, 'performance predictor' and 'gender predictor'; each of them consists of a 10-node input layer and a single-node output layer. As before, the output nodes had the Sigmoid function, while nodes in all other layers had the ReLU [70] function as activation. The transformed features from the feature encoder were simultaneously fed to the performance predictor and gender predictor. During training, the goal was to maximize the outcome (i.e., performance in this case) prediction accuracy while minimizing gender prediction accuracy by optimizing the following loss function:

$$L = L_1(X, Y, \theta_I, \theta_1) - \lambda L_2(X, G, \theta_I, \theta_2) \qquad (1)$$

where X represents the feature set, L_1 and L_2 are the loss functions for performance prediction (Y) and gender prediction (G) from X, respectively, θ_I, θ_1, and θ_2 are the parameters in the intermediate layers and the predictive branches, respectively, and λ is a hyperparameter. Consequently, we want to minimize L_1 while maximizing L_2. To achieve these goals, we combined stochastic gradient descent with reverse gradient update [33] to train the model. Concretely, during back-propagation, the parameters were updated along the direction of gradient descent for performance prediction, but gradient ascent for gender prediction. Thus, the learned parameters in the feature encoder layer would transform input features in a representation that can be used to predict students' performance but not their gender. In our implementation of the reverse gradient layer [33], we set λ to 0.3.

One limitation of censoring intermediate representation of raw features is that the transformed representation is difficult to interpreted. The representation is usually high-dimensional, prohibiting their plotting to gain insights. Additionally, if adversaries own auxiliary data that follow the same distribution as the training dataset, in some cases, the transformed features can be de-censored [84]. In the following section, we describe our methodologies aimed to eliminate these two limitations.

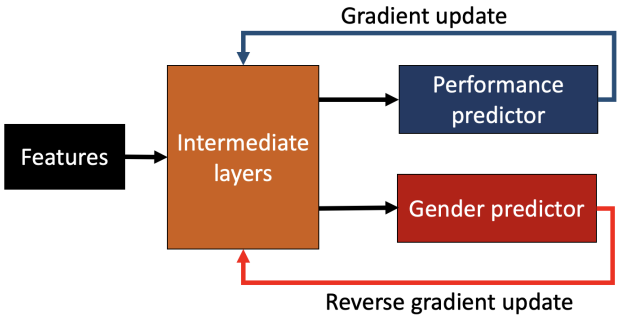

Fig. 2. Schematic diagram of the proposed model. Parameters in the intermediate layers are updated along the gradient sent by the 'performance predictor,' but along the opposite direction of the gradient sent by the 'gender predictor.' Thus, gender information is censored from the intermediate representation of the input features.

3.6 Identifying a *private* and *interpretable* feature subset that preserves the original feature set's predictive power

The initial features were inspired by prior works and based on their relevance to the target prediction task. They are easier to interpret than the transformed (i.e., censored) features, but they also reveal gender. This section describes our methods to identify a subset of the initial features that can predict students' performance with high accuracy without revealing gender. Ideally, we also want the features in the final subset to combine in an interpretable manner when they are propagated through the network. To achieve these goals, we changed the network architecture and training method as follows.

Identifying a smaller subset of features can be achieved by forcing some of the input layer's parameters to be zero. We also want the remaining features to combine among themselves in interpretable ways as they propagate to the following layer. To this end, we only allowed additive combinations of the features by forcing their corresponding (non-zero) parameters to be one. In summary, we want the input layer nodes to have binary parameters (i.e., 0 and 1) and a linear activation function. This feature selection procedure was formulated as a combinatorial optimization problem with additional constraints: from the feature set (X), select a subset

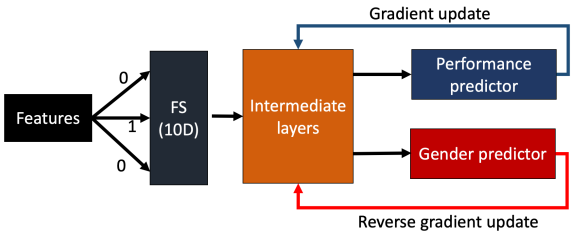

Fig. 3. Schematic diagram of the revised model. Parameters in the feature selection (FS) layer are constrained to be binary (either 0 or 1).

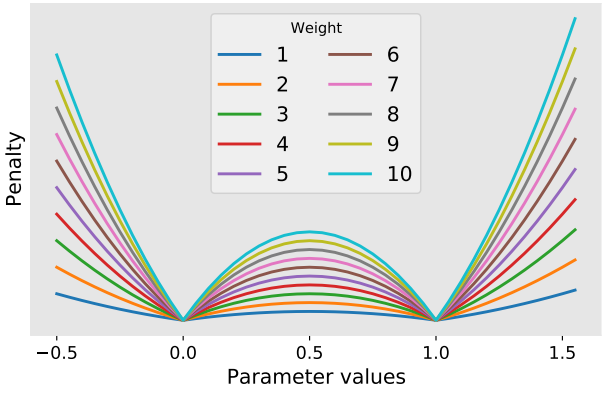

Fig. 4. Plots of the penalty function for different weights (α).

($X^* \in \mathcal{X}$) that optimizes equation 1 (i.e., minimizes performance prediction loss and maximizes gender prediction loss), where $\mathcal{X}$ is the family of all subsets of X. Thus, we arrive at the following constrained optimization problem:

$$\min_{X \in \mathcal{X}} \min_{\theta_I, \theta_1, \theta_2} L_1(X, Y, \theta_I, \theta_1) - \lambda L_2(X, G, \theta_I, \theta_2) \quad (2)$$

With this problem formulation, we want to simultaneously attain three goals: obtain a subset of features that combine among themselves in an interpretable manner, as well as maintain high accuracy for performance prediction and prevent inferring gender using the selected feature subset. Iterating over $\mathcal{X}$ to identify the best feature subset is intractable. Thus, we aim for an approximate solution by introducing a feature selection (FS) layer to the model described in subsection 3.5. Fig.3 shows the modified architecture of the model including the FS layer consisting of 10 nodes; each node's output is a linear combination of its input features and parameters. FS's parameters were constrained to be either 0 or 1 (i.e., a binary constraint). Thus, this layer acts as a feature selector since a zero-valued parameter discards the associated feature. With this formulation, equation 2 becomes

$$\min_{\theta_I, \theta_1, \theta_2, W} L_1(WX, Y, \theta_I, \theta_1) - \lambda L_2(WX, G, \theta_I, \theta_2) \quad \text{with the constraint } w_{i,j} \in \{0, 1\} \quad (3)$$

where W is the parameter matrix in the FS layer with elements $w_{i,j}$. The binary constraint was realized by regularizing FS layer's parameters with a penalty function: $\sum_{i,j} |w_{i,j}^2 - w_{i,j}|$. The terms under the summation factorize as $|(w_{i,j} - 0)(w_{i,j} - 1)|$, and penalizes any value of the parameters other than 0 and 1. Another nice property of this function is that the derivatives around 0 and 1 are symmetric and thus do not introduce any bias. Fig. 4 plots the penalty function with different weights given to it; higher weights increase penalty for the same amount of deviation of the parameters from 0 and 1.

Binarizing parameters serves the two purposes mentioned above. First, if a parameter becomes zero at the end of the training, the associated feature gets discarded, facilitating data minimization through feature selection. Conversely, when a parameter takes the value of one, the associated feature is taken as a whole. Since a linear activation function is used in this (i.e. FS) layer, each node's output is just a sum of all the 'survived' input features. This property of the model may enhance it's interpretability, e.g., by observing how features are combined may hint towards collection of 'interactions' that jointly suggest some meaningful behavior.

To compensate for the FS layer's constraint, later layers' parameters may take arbitrarily large absolute values and render the model unstable. To prevent such instability, all later layers were penalized with *L2* regularization (with weight $\beta = 0.1$). Combining everything, the final objective function becomes

$$\begin{aligned} L = \alpha \sum_{i,j} |w_{i,j}^2 - w_{i,j}| \\ + \beta(||\theta_I||_2{}^2 + ||\theta_1||_2{}^2 + ||\theta_2||_2{}^2) \\ + L_1(WX, Y, \theta_I, \theta_1) - \lambda L_2(WX, G, \theta_I, \theta_2) \end{aligned} \quad (4)$$

The hyperparameter α was initialized to 1 and increased linearly with training iteration. Thus, as training procedure progresses, the misclassification cost decreases while the penalty for non-binary parameters increases, which forces the network to concentrate more on finding binary parameter values during the later iterations.

4 Results

4.1 Preventing gender prediction

As section 3.5 described, we trained a machine learning model following an adversarial censoring procedure to convert input data to an intermediate representation, which can be used to predict students' performance with high accuracy but not to infer gender (i.e., information about gender was censored). Table 2 presents this model's prediction accuracy for performance and gender. The model predicted students' performance with 89.1% average accuracy across 10-folds[5], while predicted gender with only 53.2% accuracy using the same features. For comparison, we trained two baseline models that predict students' performance using the original (i.e., uncensored) feature set: one simple logistic regression model and one two-layer (with 20 and 10 nodes, respectively) neural network model. These two models predicted course performance with 87.4% and 89.2% accuracy, respectively. Thus, the adversarial training procedure yielded a gender-censored feature representation that predicted performance with an accuracy comparable to the original features.

The previous section presented high and low prediction accuracy for performance and gender, respectively, *when they were predicted simultaneously* using the same model that censored gender information. The obtained results match our expectations; however, they could have resulted from the specific model architecture, or the training procedure we adopted, or both. To further validate our results and demonstrate the wider applicability of the censoring mechanism, we decoupled the censoring step from the inference step. That is, we extracted the censored features from the model and used them to train two separate models to *independently* predict performance and gender, respectively. Since the transformed features are only 10-dimensional, we used a logistic regression model in both cases. The two models predicted performance and gender with 89.3% and 52.5% mean accuracy, respectively. These findings suggest that the censored features may be used to train new predictive models without risking students' privacy.

4.2 Obtaining interpretable features

This section reports findings from the model described in section 3.6: a neural network with a feature selection layer and a custom penalty function. The model was trained to identify a feature subset that can predict students' performance but lack gender information. Across 10 train-test splits, the model predicted students' performance and gender with average accuracy of 85.1% and 52.9%, respectively.

Thus, the model achieved the first two goals: high performance prediction accuracy and low gender prediction accuracy. For our third goal, interpretability of the selected features, we examine the selected features and their combinations in the FS layer.

By examining FS layer's parameters, we found that seven out of the 10 nodes in that layer had all 0 values (i.e., all input features were discarded). The remaining three nodes also had 0 values for the parameters corresponding to most of the input features. Following features had associated parameters equal to 1 in those three nodes:

Node 2: number of blocks, number of sessions visiting course homepage, and percentage of submitted quizzes.

Node 5: number of blocks, number of sessions visiting course homepage, and percentage of submitted quizzes.

Node 8: number of blocks, number of sessions visiting course homepage, and percentage of submitted quizzes, percentage of external websites visited, percentage of external quizzes submitted, and percentage of assignments submitted.

All nodes share the first three features, and only six unique features 'survived' the feature selection process. Recall that the parameters corresponding to these features have values equal to 1, and these nodes had a linear activation function. Consequently, in each of the three nodes, the 'survived' features are simply summed and then passed onto the next layer. Each input feature summarize a particular behavior of students. Thus, their combinations, when they are only allowed to be summed together, preserve their interpretive and informative nature. Educational experts may glean insights from these combined features, e.g., how a collection of behaviors impact certain outcomes.

5 we omitted standard deviation because of their small values ($< 0.5\%$ in all cases)

	LR-baseline	NN-baseline	Joint prediction with censoring	Predicting separately with censored features
Performance	**87.4%**	**89.2%**	**89.1%**	**89.3%**
Gender	**73.3%**	**76.2%**	**53.2%**	**52.5%**

Table 2. Average prediction accuracy of course performance and gender by different models (10-fold cross validation). The first two columns provide results from the baseline models described in section 3.4. The third column presents prediction accuracy of performance and gender when both were simultaneously predicted by a single model by censoring input features. The fourth column shows performance and gender prediction accuracy from two different models trained independently using censored features extracted from the previous model. In both cases, the gender the prediction accuracy drops to the chance level after censoring the feature representation.

4.3 Training new models with the interpretable features

As before, to demonstrate our feature selection methods' wider applicability, we decoupled feature selection (with censoring) step and inference step. We trained two separate models with the six features to predict students' performance and gender, respectively. As the number of features is low, we again preferred simple logistic regression models. We conducted 10-fold cross validations with 80%-20% train-test splits in both cases. We found a mean accuracy of 86.5% for performance prediction, and a mean accuracy of 63.5% for gender prediction. Thus, the selected six features can predict performance with the same accuracy as using all 65 features; but unlike the intermediate (censored) features, these six feature can also be used to predict gender, albeit with lower accuracy compared to using all 65 features.

4.4 Training new models with one 'combined' feature

. The FS layer allowed features to only be summed together. We mimic the process by summing the six features to create a 'combined' feature. Then, we again trained two logistic regression models with this 'combined' feature to predict performance and gender, respectively. The first model predicted performance with 84.65% accuracy, while the second model predicted gender with only 52.5% accuracy. Thus, the combination of the six features creates a privacy-preserved version of the original dataset; it predicts students' performance with an accuracy comparable to using all 65 features, while keeping gender prediction accuracy at the chance level.

Conclusions. Our findings show that the adversarial training procedure successfully created a censored representation of the input features. These transformed features may be stored or shared and used to build new predictive models, without risking students' privacy. Further, our feature selection methodology enhances interpretability by identifying a small subset of the features and constraining them to combine linearly, while maintaining high-performance prediction accuracy and preventing gender inference. Again, the selected features or their combination may be stored and shared for commercial or research purposes without harming students' privacy.

4.5 Robustness of our methods

Several sources of randomness (e.g., initializing model parameters) exist in the process of training machine learning models. The obtained solution may be unstable if it depends on the configuration in which it was found. To investigate the stability of our constrained optimization method to identify a small set of features reliably (i.e., robust against randomness), we retrained the model 100 times. We ensured a different random initialization of the parameters in every trial and created different train-test splits of the data. The test set accuracy varied by at most 0.2%, and every time the model identified the same six features. These findings demonstrate that our methodology to obtain a private and interpretive feature subset is stable against random variability.

4.6 Generalizability of our methods

We assessed of our methods' generalizability by applying them to two additional prediction tasks. At first, we kept the target prediction task the same as before (i.e., predicting performance), but focused on a different sensitive attribute: students' age group. Then, we kept the sensitive attribute the same as before (i.e., gender), but focused on a different target prediction task, dropout probability, which is another important task within the

learning analytics community [65].

Inferring age group. OULAD has three age groups: less than 35 years old (N=20145), 35–55 years old (N=8462), and more than 55 years old (N=178). Since the last group is too small, we merged it with the second group, turning the age group inference into a binary classification problem. To balance the class distributions, we randomly down sampled from the majority (first) group. The resulting dataset contained records of 15,432 students equally divided into the two age groups. We repeated the procedure described above to investigate if age group can be predicted from the interaction data, and adversarial training coupled with constrained optimization can provide an interpretable subset of the features that can predict students' performance while censoring age related information.

As before, we first trained a neural network model to predict students' age group; it achieved a mean accuracy of 63.4% (26.8% increased accuracy than chance prediction). Using adversarial censoring with constrained optimization achieved an accuracy of 85.5% for students' performance, while reducing age group prediction to 56.1%. The final subset of features contains the following six features: i) number of blocks, ii) number of sessions visiting course homepage, iii) percentage of external quizzes submitted, iv) percentage of quizzes submitted, v) percentage of pages visited related to site information, and vi) percentage of external websites visited. Note that five of the six features were also identified for gender prediction.

Predicting dropout probability. In the original dataset, 5,995 students dropped out from the courses, while 19,250 students completed the courses. As before, we randomly down sampled the majority class, and the resulting dataset contained 11,990 students, equally divided into the two classes with 54.5% male identifying students. We applied the same set of procedures as before with the following results.

Recall that, the baseline neural network to predict gender achieved an average accuracy of 76.2%. Using adversarial censoring with constrained optimization, our method achieved an accuracy of 80.9% for dropout prediction (60% improvement compared to chance prediction), while reducing gender prediction to 54.7%. The final set of features contains the following seven features: i) number of blocks, ii) number of sessions visiting course homepage, iii) percentage of external quizzes submitted, iv) percentage of quizzes submitted, v) percentage of pages visited related to site information, vi) percentage of external websites visited, and vii) percentage of files (e.g., tutorials and lecture notes) visited. Again, note the overlap with the features identified earlier.

Conclusions. In the above two subsections, we applied our methodology to censor age information while predicting course performance, and to censor gender information while predicting dropout probability. In both cases, our approach successfully identified a private-version of the original dataset consisting of a few (overlapping) features that contain sufficient information about the outcome variable while almost no information about the sensitive attribute. These results demonstrate generalizability of our methods.

5 Discussions, limitations, and future work

Gender does not *causally* affect course performance. No difference in course performance across genders was observed for students who had a 'similar' educational history and engagement with the course content. This result suggests that any observed differences in course performance between male and female students in the original sample may be due to a spurious association between gender and performance. Our results hold for international students who participated in 22 different courses offered under diverse university programs. Thus, we anticipate similar results (i.e., the null causal effect of gender, and perhaps of other demographic attributes) may hold for datasets collected at other institutes, and recommend against sharing such data (with EdTech companies or publicly) without first assessing their effects on target outcomes.

Students' gender can be inferred from behavioral data. Our models predicted gender from activity logs with an accuracy high enough to raise privacy concerns, even though we used basic model architectures for simplicity's shake. EdTech companies' access to much larger and diverse datasets and resources to train more complex models only amplify our privacy concerns. Thus, even if demographic information is not collected directly, EdTech companies may be able to profile students based on inferred demographic factors, and target students, e.g., for advertisements or surveillance, which severely undermines students' privacy and autonomy.

Privacy-preserved versions of a dataset can be obtained using adversarial censoring combined with a feature selection technique. Our novel approach that combines adversarial censoring with feature selection through a penalty function identified a small subset of the original features that contain sufficient information about the target outcome and almost no information about the censored attribute. Educational institutes may employ our approach to identify an interpretable feature set required for their desired prediction task(s), and only allow collecting those features by the EdTechs they deploy. We also demonstrated that new models trained on the selected feature set had high accuracy for target prediction and low accuracy for censored attribute prediction. Thus, the privacy-preserved version obtained through our proposed techniques may be shared publicly to facilitate progress and reproducibility in scientific research related to EDM and LA. Additionally, our methods yielded small feature subsets; thus, after model building, the features' individual and collective (since they combine very intuitively) effects on the outcome variable can be examined easily by educational experts to understand their potential causal relevance, which is one of the key goals of LA [68]. Identifying a small feature subset also facilitates conducting controlled experiments through EdTechs to identify causal effects [68], since the number of variables to manipulate becomes much smaller than the full feature set.

Limitations. Empirical evaluations of our technical contributions are limited to one dataset. Unfortunately, we did not find any other public dataset containing similar features and demographic attributes of students. Nonetheless, we demonstrated the generalizability of our approach by applying it to predict a different target variable (dropout probability), as well as to censor a different demographic attribute (age group). Our approach is also robust against randomness, as evident from the results from repeated trials with different initializations of model parameters and train-test splits.

Future directions. An obvious extension of our work is to apply the proposed methodologies to other datasets (both proprietary and as new datasets become public) to censor different sensitive attributes. In this paper, we used causal inference mechanisms to demonstrate that gender *lack causal effect* on course performance. This same mechanism can be applied to identify features that *causally affect* a target outcome, and build predictive models only using those features. Research in other domains demonstrated that causal models are more transparent and generalizable, as well as less vulnerable to privacy attacks [32, 89, 96]. We hope that our work will pave the way to causality-based model building in the growing fields of EDM and LA to reduce students' privacy risks.

Another interesting future research direction could be dealing with attributes that are privacy sensitive and also causally affects the target outcome(s). Such causally relevant sensitive attributes could be replaced by other correlated, non-causal features; future research could explore methods to find such features. Furthermore, future work could expand our work to jointly censor multiple sensitive attributes. Adversarial censoring-based models, such as ours, do not provide guarantees, as they rely on empirical validations. Conversely, differential privacy (DP) provides formal guarantees of privacy protection for a given privacy budget, but is only applicable to prevent leaking of one's membership to a dataset. Obtaining strong guarantees for issues related to attribute inference remains an open challenge and interesting future research direction.

6 Conclusion

Educational institutes at all levels are deploying EdTech at a rapid pace. In addition to delivering education, such technologies are expanding their roles in other pedagogical and learning practices, such as student assessment. EdTech collects massive amounts of students' behavioral and demographic data, often justified by the need to develop learning analytics to improve the educational process itself. But these data, particularly demographic attributes, pose serious privacy threats to the students, as they can be profiled and, e.g., targeted for advertisement or surveillance based on their demographic characteristics. As EdTech providers aim for 'frictionless' data access by their different services and third-party service providers, inadequate policies and regulations to oversee the storage and sharing of these data have exacerbated the privacy concerns. In this paper, we show that gender—which is often used as a predictor of students' performance and risk of dropout—is not causally relevant to those outcome variables. While we oppose collecting demographic attributes by EdTech, we also show that these attributes can be inferred from students' behavioral data with concerningly high accuracy. We propose a novel method to obtain a privacy-preserving version of the original dataset by combining a feature selection technique with adversarial censoring techniques. The 'private' dataset contains fewer features

that are sufficient to build predictive models for target outcomes but do not reveal information about the sensitive attribute. We demonstrated the efficacy and generalizability of our approach by applying it to censor gender and age information while allowing building models to predict course performance and dropout likelihood. Our methodology can be adopted to restrict educational technologies to collect only essential features for a given task that do not reveal sensitive information about students. Our approach may also facilitate scientific progress and reproducibility by allowing the publication of a privacy-preserved version of institutional datasets.

7 Acknowledgment

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. We thank Hui-Po Wang (CISPA Helmholtz Center for Information Security) for providing valuable suggestions on the code implementation.

References

[1] Adam B. The 101 Hottest EdTech Tools According to Education Experts (Updated For 2020), 6.

[2] June Ahn, Fabio Campos, Ha Nguyen, Maria Hays, and Jan Morrison. Co-Designing for Privacy, Transparency, and Trust in K-12 Learning Analytics. In *LAK21: 11th International Learning Analytics and Knowledge Conference*, pages 55–65. Association for Computing Machinery, New York, NY, USA, 2021.

[3] Nil-Jana Akpinar, Aaditya Ramdas, and Umut Acar. Analyzing student strategies in blended courses using clickstream data. *arXiv preprint arXiv:2006.00421*, 2020.

[4] Reham Mohammad Almohtadi and Intisar Turki Aldarabah. University Students' Attitudes toward the Formal Integration of Facebook in Their Education: Investigation Guided by Rogers' Attributes of Innovation. *World Journal of Education*, 11(1):20–28, 2021.

[5] Kimberly E Arnold and Niall Sclater. Student Perceptions of Their Privacy in Leaning Analytics Applications. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, LAK '17, pages 66–69, New York, NY, USA, 2017. Association for Computing Machinery.

[6] Yousef Atoum, Liping Chen, Alex X Liu, Stephen D H Hsu, and Xiaoming Liu. Automated Online Exam Proctoring. *IEEE Transactions on Multimedia*, 19(7):1609–1624, 2017.

[7] Peter C Austin. Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. *Statistics in medicine*, 30(11):1292–1301, 2011.

[8] Seyyed Kazem Banihashem, Khadijeh Aliabadi, Saeid Pourroostaei Ardakani, Ali Delaver, and Mohammadreza Nili Ahmadabadi. Learning Analytics: A Systematic Literature Review. *Interdisciplinary Journal of Virtual Learning in Medical Sciences*, 9(2):–, 2018.

[9] Peter Bautista and Paul Salvador Inventado. Protecting Student Privacy with Synthetic Data from Generative Adversarial Networks. In *International Conference on Artificial Intelligence in Education*, pages 66–70, 2021.

[10] Nigel Bosch, R Wes Crues, Najmuddin Shaik, and Luc Paquette. "Hello,[REDACTED]": Protecting Student Privacy in Analyses of Online Discussion Forums. In *EDM*, 2020.

[11] Macy A Burchfield, Joshua Rosenberg, Conrad Borchers, Tayla Thomas, Ben Gibbons, and Christian Fischer. Are Violations of Student Privacy "Quick and Easy"? Investigating the Privacy of Students' Images and Names in the Context of K-12 Educational Institution's Posts on Facebook. 2021.

[12] Kursat ; Celik Berkan Cagiltay Nergiz Ercil ; Cagiltay. An Analysis of Course Characteristics, Learner Characteristics, and Certification Rates in MITx MOOCs. *International Review of Research in Open and Distributed Learning*, 21(3):121–139, 2020.

[13] Cassandra Willer. 27 Tech Tools Teachers Can Use to Inspire Classroom Creativity.

[14] Changsheng Chen, Jingyun Long, Junxiao Liu, Zongjun Wang, Minglei Shan, and Yuming Dou. Behavioral Patterns of Completers in Massive Open Online Courses (MOOCs): The Use of Learning Analytics to Reveal Student Categories. In *2020 International Conference on Advanced Education, Management and Social Science (AEMSS2020)*, pages 56–63, 2020.

[15] Guanliang Chen, Dan Davis, Jun Lin, Claudia Hauff, and Geert-Jan Houben. Beyond the MOOC Platform: Gaining Insights about Learners from the Social Web. In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, pages 15–24, New York, NY, USA, 2016. Association for Computing Machinery.

[16] Christopher Pappas. The Best Learning Management Systems (2020 Update), 2020.

[17] Chia Yuan Chuang, Scotty D Craig, and John Femiani. Detecting probable cheating during online assessments based on time delay and head pose. *Higher Education Research & Development*, 36(6):1123–1137, 2017.

[18] Matthieu Courbariaux and Yoshua Bengio. BinaryNet: Training Deep Neural Networks with Weights and Activations }Constrained to +1 or -1. *CoRR*, abs/1602.02830, 2016.

[19] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in neural information processing systems*, pages 3123–3131, 2015.

[20] R Wes Crues, Genevieve M Henricks, Michelle Perry, Suma Bhat, Carolyn J Anderson, Najmuddin Shaik, and Lawrence Angrave. How Do Gender, Learning Goals, and Forum Participation Predict Persistence in a Computer Science MOOC? *ACM Trans. Comput. Educ.*, 18(4), 9 2018.

[21] Ying Cui, Fu Chen, Ali Shiri, and Yaqin Fan. Predictive analytic models of student success in higher education: A review of methodology. *Information and Learning Sciences*, 2019.

[22] Paula G de Barba, Donia Malekian, Eduardo A Oliveira, James Bailey, Tracii Ryan, and Gregor Kennedy. The importance and meaning of session behaviour in a MOOC. *Computers & Education*, 146:103772, 2020.

[23] Hendrik Drachsler and Wolfgang Greller. Privacy and Analytics: It's a DELICATE Issue a Checklist for Trusted Learning Analytics. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, LAK '16, pages 89–98, New York, NY, USA, 2016. Association for Computing Machinery.

[24] Harrison Edwards and Amos J Storkey. Censoring Representations with an Adversary. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[25] Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640*, 2018.

[26] Facebook. Facebook for Education. 2021.

[27] Facebook. https://education.facebook.com/, 2 2021.

[28] Wenzheng Feng, Jie Tang, and Tracy Xiao Liu. Understanding Dropouts in MOOCs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):517–524, 7 2019.

[29] Tobias Fiebig, Seda Gürses, Carlos H Gañán, Erna Kotkamp, Fernando Kuipers, Martina Lindorfer, Menghua Prisse, and Taritha Sari. Heads in the Clouds: Measuring the Implications of Universities Migrating to Public Clouds. *arXiv preprint arXiv:2104.09462*, 2021.

[30] Andy Field, Jeremy Miles, and Zoë Field. *Discovering statistics using R*. Sage publications, 2012.

[31] Alvaro Figueira. Predicting Grades by Principal Component Analysis: A Data Mining Approach to Learning Analytics. In *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*, pages 465–467, 2016.

[32] Sreya Francis, Irene Tenison, and Irina Rish. Towards Causal Federated Learning For Enhanced Robustness and Privacy. *CoRR*, abs/2104.06557, 2021.

[33] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189, 2015.

[34] Gerrit De Vynck and Mark Bergen. Google Classroom Users Doubled as Quarantines Spread , 2020.

[35] Google. Google Classroom, 2021.

[36] Brenda Edith Guajardo Leal, Valenzuela GonzĆ, and others. Student Engagement as a Predictor of xMOOC Completion: An Analysis from Five Courses on Energy Sustainability. *Online Learning*, 23(2):105–123, 2019.

[37] Kalervo Gulson, Carlo Perrotta, Ben Williamson, and Kevin Witzenberger. Should We be Worried about Google Classroom? The Pedagogy of Platforms in Education.

[38] Song Guo, Deze Zeng, and Shifu Dong. Pedagogical Data Analysis Via Federated Learning Toward Education 4.0. *American Journal of Education and Information Technology*, 4(2):56, 2020.

[39] Mehmet Emre Gursoy, Ali Inan, Mehmet Ercan Nergiz, and Yucel Saygin. Privacy-Preserving Learning Analytics: Challenges and Techniques. *IEEE Transactions on Learning Technologies*, 10(1):68–81, 2017.

[40] Jihun Hamm. Minimax filter: Learning to preserve privacy from inference attacks. *The Journal of Machine Learning Research*, 18(1):4704–4734, 2017.

[41] Rakibul Hasan, Cristobal Cheyre Forestier, Yong-Yeol Ahn, Roberto Hoyle, and Apu Kapadia. The Impact of Viral Posts on Visibility and Behavior of Professionals: A Longitudinal Study of Scientists on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 6 2022.

[42] Arto Hellas, Petri Ihantola, Andrew Petersen, Vangel V Ajanovski, Mirela Gutica, Timo Hynninen, Antti Knutas, Juho Leinonen, Chris Messom, and Soohyun Nam Liao. Predicting Academic Performance: A Systematic Literature Review. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, ITiCSE 2018 Companion, pages 175–199, New York, NY, USA, 2018. Association for Computing Machinery.

[43] Alex Hern. Google faces lawsuit over email scanning and student data, 2014.

[44] Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, 15(3):199–236, 2007.

[45] Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, 42(8):1–28, 2011.

[46] Tore Hoel, Dai Griffiths, and Weiqin Chen. The Influence of Data Protection and Privacy Frameworks on the Design of Learning Analytics Systems. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, LAK '17, pages 243–252, New York, NY, USA, 2017. Association for Computing Machinery.

[47] HolonIQ. 10 charts to explain the Global Education Technology Market, 1 2021.

[48] Dirk Ifenthaler. Are higher education institutions prepared for learning analytics? *TechTrends*, 61(4):366–371, 2017.

[49] Yusuke Iwasawa, Kotaro Nakayama, Ikuko Yairi, and Yutaka Matsuo. Privacy Issues Regarding the Application of DNNs to Activity-Recognition using Wearables and Its Countermeasures by Use of Adversarial Training. In *IJCAI*, pages 1930–1936, 2017.

[50] Nikhil Indrashekhar Jha, Ioana Ghergulescu, and Arghir-Nicolae Moldovan. OULAD MOOC Dropout and Result Prediction using Ensemble, Deep Learning and Regression Techniques. In *CSEDU (2)*, pages 154–164, 2019.

[51] Srećko Joksimović, Oleksandra Poquet, Vitomir Kovanović, Nia Dowell, Caitlin Mills, Dragan Gašević, Shane Dawson, Arthur C Graesser, and Christopher Brooks. How do we model learning at scale? A systematic review of research on MOOCs. *Review of Educational Research*, 88(1):43–86, 2018.

[52] Kyle M L Jones, Alan Rubel, and Ellen LeClere. A matter of trust: Higher education institutions as information fiduciaries in an age of educational data mining and learning analytics. *Journal of the Association for Information Science and Technology*, 71(10):1227–1241, 2020.

[53] Katherine Mangan. The Surveilled Student, 2 2021.

[54] Puninder Kaur, Amandeep Kaur, and Rajwinder Kaur. A Systematic Review About Prediction of Academic Behavior Through Data Mining Techniques. *Journal of Computational and Theoretical Nanoscience*, 17(11):5162–5166, 2020.

[55] Mohammad Khalila and Martin Ebner. De-identification in learning analytics. *Journal of Learning Analytics*, 3(1):129–138, 2016.

[56] Gary King and Richard Nielsen. Why Propensity Scores Should Not Be Used for Matching. *Political Analysis*, 27(4):435–454, 2019.

[57] Mark Klose, Vasvi Desai, Yang Song, and Edward Gehringer. EDM and Privacy: Ethics and Legalities of Data Collection, Usage, and Storage. *International Educational Data Mining Society*, 2020.

[58] Daniel G Krutka, Ryan M Smits, and Troy A Willhelm. Don't Be Evil: Should We Use Google in Schools? *TechTrends*, 2021.

[59] Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. Open university learning analytics dataset. *Scientific data*, 4(1):1–8, 2017.

[60] Charles Lang, Charlotte Woo, and Jeanne Sinclair. Quantifying Data Sensitivity: Precise Demonstration of Care When Building Student Prediction Models. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, LAK '20, pages 655–664, New York, NY, USA, 2020. Association for Computing Machinery.

[61] Eitel J M Lauría, Joshua D Baron, Mallika Devireddy, Venniraiselvi Sundararaju, and Sandeep M Jayaprakash. Mining Academic Data to Improve College Student Retention: An Open Source Perspective. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 139–142, New York, NY, USA, 2012. Association for Computing Machinery.

[62] James J Lohse, Christine A McManus, and David A Joyner. Surveying the MOOC Data Set Universe. In *2019 IEEE Learning With MOOCS (LWMOOCS)*, pages 159–164, 2019.

[63] Asim Majeed, Said Baadel, and Anwar Ul Haq. Global triumph or exploitation of security and privacy concerns in e-learning systems. In *International Conference on Global Security, Safety, and Sustainability*, pages 351–363, 2017.

[64] Roxana Marachi and Lawrence Quill. The case of Canvas: Longitudinal datafication through learning management systems. *Teaching in Higher Education*, 25(4):418–434, 2020.

[65] Neema Mduma, Khamisi Kalegele, and Dina Machuve. A survey of machine learning approaches and techniques for student dropout prediction. 2019.

[66] Pedro Manuel Moreno-Marcos, Pedro J Muñoz-Merino, Carlos Alario-Hoyos, and Carlos Delgado Kloos. Re-Defining, Analyzing and Predicting Persistence Using Student Events in Online Learning. *Applied Sciences*, 10(5), 2020.

[67] Pedro Manuel Moreno-Marcos, Ting-Chuen Pong, Pedro J Muñoz-Merino, and Carlos Delgado Kloos. Analysis of the Factors Influencing Learners' Performance Prediction With Learning Analytics. *IEEE Access*, 8:5264–5282, 2020.

[68] Benjamin A Motz, Paulo F Carvalho, Joshua R de Leeuw, and Robert L Goldstone. Embedding Experiments: Staking Causal Inference in Authentic Educational Contexts. *Journal of Learning Analytics*, 5(2):47–59, 8 2018.

[69] Kew Si Na and Zaidatun Tasir. A systematic review of learning analytics intervention contributing to student success in online learning. In *2017 International conference on learning and teaching in computing and engineering (LaTICE)*, pages 62–68, 2017.

[70] Vinod Nair and Geoffrey E Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *ICML*, pages 807–814, 2010.

[71] Abdallah Namoun and Abdullah Alshanqiti. Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. *Applied Sciences*, 11(1), 2021.

[72] Natasha Singer. Learning Apps Have Boomed in the Pandemic. Now Comes the Real Test., 3 2021.

[73] David Nicholas, Hamid R Jamali, Eti Herman, Anthony Watkinson, Abdullah Abrizah, Blanca Rodríguez-Bravo, Cherifa Boukacem-Zeghmouri, Jie Xu, Marzena Świgoń, and Tatiana Polezhaeva. A global questionnaire survey of the scholarly communication attitudes and behaviours of early career researchers. *Learned Publishing*, n/a(n/a).

[74] Luc Paquette, Jaclyn Ocumpaugh, Ziyue Li, Alexandra Andres, and Ryan Baker. Who's Learning? Using Demographics in EDM Research. *Journal of Educational Data Mining*, 12(3):1–30, 2020.

[75] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[76] Bardh Prenkaj, Paola Velardi, Giovanni Stilo, Damiano Distante, and Stefano Faralli. A Survey of Machine Learning Approaches for Student Dropout Prediction in Online Courses. *ACM Comput. Surv.*, 53(3), 5 2020.

[77] Banoor Yousra Rajabalee, Mohammad Issack Santally, and Frank Rennie. A study of the relationship between students' engagement and their academic performances in an eLearning environment. *E-Learning and Digital Media*, 17(1):1–20, 2020.

[78] S Ranjeeth, T P Latchoumi, and P Victer Paul. A Survey on Predictive Models of Learning Analytics. *Procedia Computer Science*, 167:37–46, 2020.

[79] Joel R Reidenberg and Florian Schaub. Achieving big data privacy in education. *Theory and Research in Education*, 16(3):263–279, 2018.

[80] Paul R Rosenbaum. Sensitivity Analysis in Observational Studies. In *Wiley StatsRef: Statistics Reference Online*. American Cancer Society, 2014.

[81] Donald B Rubin. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[82] Jennifer Sabourin, Lucy Kosturko, Clare FitzGerald, and Scott McQuiggan. Student Privacy and Educational Data Mining: Perspectives from Industry. *International Educational Data Mining Society*, 2015.

[83] Daniel J Solove. A taxonomy of privacy. *U. Pa. L. Rev.*, 154:477, 2005.

[84] Congzheng Song and Vitaly Shmatikov. Overlearning reveals sensitive attributes. *arXiv preprint arXiv:1905.11742*, 2019.

[85] Elizabeth A Stuart. Matching Methods for Causal Inference: A Review and a Look Forward. *Statist. Sci.*, 25(1):1–21, 2010.

[86] Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34, 2000.

[87] Mariela Mizota Tamada, José Francisco de Magalhães Netto, and Dhanielly Paulina R de Lima. Predicting and Reducing Dropout in Virtual Learning using Machine Learning Techniques: A Systematic Review. In *2019 IEEE Frontiers in Education Conference (FIE)*, pages 1–9, 2019.

[88] Leslie Ching Ow Tiong and HeeJeong Jasmine Lee. E-cheating Prevention Measures: Detection of Cheating at Online Examinations Using Deep Learning Approach–A Case Study. *arXiv preprint arXiv:2101.09841*, 2021.

[89] Shruti Tople, Amit Sharma, and Aditya Nori. Alleviating Privacy Attacks via Causal Learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9537–9547. PMLR, 8 2020.

[90] Allen Tough. The Adult's Learning Projects. A Fresh Approach to Theory and Practice in Adult Learning. 1979.

[91] Hajra Waheed, Saeed-Ul Hassan, Naif Radi Aljohani, Julie Hardman, Salem Alelyani, and Raheel Nawaz. Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, 104:106189, 2020.

[92] Jacob Whitehill, Kiran Mohan, Daniel Seaton, Yigal Rosen, and Dustin Tingley. Delving deeper into MOOC student dropout prediction. *arXiv preprint arXiv:1702.06404*, 2017.

[93] Wikipedia. Pearson's chi-squared test.

[94] Ben Williamson, Sian Bayne, and Suellen Shay. The datafication of teaching in Higher Education: critical issues and perspectives. *Teaching in Higher Education*, 25(4):351–365, 2020.

[95] Elad Yacobson, Orly Fuhrman, Sara Hershkovitz, and Giora Alexandron. De-identification is not enough to guarantee student privacy: De-anonymizing personal information from basic logs. In *Companion Proceedings 10th International Conference on Learning Analytics & Knowledge (LAK20)*, 2019.

[96] Kui Yu, Xianjie Guo, Lin Liu, Jiuyong Li, Hao Wang, Zhaolong Ling, and Xindong Wu. Causality-Based Feature Selection: Methods and Evaluations. *ACM Comput. Surv.*, 53(5), 9 2020.

[97] Xin Zhang, Cheng-Wei Wu, Philippe Fournier-Viger, Lan-Da Van, and Yu-Chee Tseng. Analyzing students' attention in class using wearable devices. In *2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pages 1–9, 2017.

[98] Zhongheng Zhang, Hwa Jung Kim, Guillaume Lonjon, and Yibing Zhu. Balance diagnostics after propensity score matching., 1 2019.